KU LEUVEN

ARENBERG DOCTORAL SCHOOL
Faculty of Engineering Science

Network analysis of drug responses

Griet Laenen

Supervisors:
Prof. dr. ir. Yves Moreau
Prof. dr. ir. Lieven Thorrez

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor in Engineering Science

January 2020

Network analysis of drug responses

Griet LAENEN

Examination committee:
Prof. dr. ir. Omer Van der Biest, chair
Prof. dr. ir. Yves Moreau, supervisor
Prof. dr. ir. Lieven Thorrez, co-supervisor
Prof. dr. Joris Winderickx
Prof. dr. ir. Luc De Raedt
Prof. dr. Vera van Noort
Prof. dr. ir. Julio Saez-Rodriguez

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor in Engineering Science

January 2020

© 2020 KU Leuven – Faculty of Engineering Science
Uitgegeven in eigen beheer, Griet Laenen, Kasteelpark Arenberg 10, bus 2446, B-3001 Leuven (Belgium)

Alle rechten voorbehouden. Niets uit deze uitgave mag worden vermenigvuldigd en/of openbaar gemaakt worden door middel van druk, fotokopie, microfilm, elektronisch of op welke andere wijze ook zonder voorafgaande schriftelijke toestemming van de uitgever.

All rights reserved. No part of the publication may be reproduced in any form by print, photoprint, microfilm, electronic or any other means without written permission from the publisher.

Preface

Daar waar ik tijdens mijn opleiding biochemie soms toch een beetje op mijn honger bleef zitten, was ik meteen gebeten door al wat ik leerde tijdens mijn master bio-informatica. Ik heb dan ook niet lang getwijfeld toen ik de kans kreeg een doctoraat in de bio-informatica aan te vatten en niet veel later werd dit tot mijn eigen grote verbazing een feit: ik ging doctoreren. Dat doctoreren werd een lange maar onvergetelijke reis, waarop heel wat mensen met verschillende achtergronden en de meest uiteenlopende interesses mijn pad kruisten, maar ik evengoed een aantal keer mezelf ben tegengekomen. Ik leerde doorheen de jaren enorm veel bij, zowel op wetenschappelijk als persoonlijk vlak. Het samenbrengen van al die jaren in één boekje bleek voor mij veruit de grootste uitdaging, maar mede dankzij de vele aanmoedigingen kan ik vandaag eindelijk de laatste en misschien wel belangrijkste woorden op papier zetten. Een enorme dankjewel aan iedereen die op welke manier dan ook mee getimmerd heeft aan mijn weg!

Eerst en vooral wil ik mijn promotor, Prof. Yves Moreau, van harte bedanken om mij de mogelijkheid te bieden als doctoraatsstudent binnen de BIOI-groep een nieuwe onderzoeksrichting in te slaan. Yves, bedankt voor het vertrouwen, voor de steun en voor de kansen die je me gaf om mijn eigen weg te vinden. Je stond altijd klaar om mee te denken en nieuwe ideeën aan te reiken. Je verliest daarbij nooit de persoon achter het onderzoek uit het oog. Verder ben ik ook mijn co-promotor, Prof. Lieven Thorrez, erg dankbaar voor zijn kritische input en onuitputtelijk enthousiasme. Lieven, jij wist mij steeds richting en raad te geven wanneer ik het noorden even kwijt was. Bedankt voor de vele tips en betrokkenheid bij mijn werk. Ook alle leden van de jury, Prof. Luc De Raedt, Prof. Joris Winderickx, Prof. Vera van Noort en Prof. Julio Saez-Rodriguez, wil ik graag bedanken voor het kritisch nalezen van dit manuscript, de leerrijke discussies en waardevolle suggesties.

Het Instituut voor Innovatie door Wetenschap en Technologie (IWT) ben ik zeer dankbaar voor het financieren van mijn onderzoek. Verder gaat een

speciaal woord van dank uit naar Ida, Elsy, John, Wim en Mimi voor alle administratieve en logistieke steun en naar Maarten en Liesbeth voor de geweldige IT-ondersteuning.

Doctoreren kan uiteraard niet zonder een stel geweldige collega's om te delen in alle ups en downs die een doctoraat rijk is. Ik wil dan ook iedereen die mijn jaren in 'de toren' tot een inspirerende, uitdagende, maar onvergetelijke tijd heeft gemaakt immens bedanken: Adam, Alejandro, Amin, Arnaud, Babis, Daniela, Daniele, Dusan, Georgios, Gorana, Inge, Jaak, Jan, Jiqiu, Leo, Marc, Nico, Peter, Pooya, Raf, Ryo, Sarah, Thomas, Toni, Tunde, Yousef en al degenen die ik, geheel onterecht, vergeten ben. Hierbij ook een bijzondere dankjewel aan Amin, voor het leuke samenwerken, maar zeker ook voor de vele hulp en het fijne gezelschap. Ook bedankt aan masterstudenten Sander, Volodimir en Oliver en de vele collega's van andere onderzoeksgroepen voor de vruchtbare en leerrijke samenwerkingen. Verder wil ik ook zeker mijn collega's bij Open Analytics bedanken om de voorbije drie jaren een onuitputtelijke bron van plezier, inspiratie en expertise te zijn en mijn beeld op de wereld verder te verruimen.

De Leuvense biochemiebende, bedankt voor de onvergetelijke studentenjaren met vele gezellige etentjes en swingende feestjes. Ook mijn Kempense vrienden, in het bijzonder de pony's, wil ik bedanken om op tijd mijn gedachten te verzetten, maar ook om een luisterend oor te zijn, om interesse te veinzen zonder enigszins te begrijpen waar ik mee bezig ben en om mij op tijd en stond die broodnodige peptalk te geven.

Moeke en vake, in de eerste plaats bedankt voor alle mogelijkheden die ik de voorbije jaren gekregen heb. Bedankt ook voor de vele aanmoedigingen en schouderklopjes, om erin te blijven geloven wanneer ik dit zelf niet meer deed en om aan te voelen wanneer het misschien toch beter was even niet te polsen achter de vorderingen van dat doctoraat. Bedankt voor de warme en gezellige thuis. Ook mijn zussen, An en Lies, bedankt om er voor mij te zijn, elk op jullie eigen manier.

Tot slot, Dries, bedankt om mij hoofdzaak van bijzaak te leren onderscheiden, om mij de nodige tijd en ruimte te geven, om mijn rots in de branding te zijn, om zowel de mooie als de minder mooie momenten met mij te willen delen. En nu op naar een nieuw avontuur...

Abstract

The pharmaceutical industry is facing unprecedented pressure to increase its productivity. Attrition rates in the later stages of development have risen sharply, with toxicity and lack of efficacy being the main bottlenecks. To address both these safety and efficacy issues, a better understanding of the complex biological response to drug treatment is vital. Although many drugs exert their therapeutic activities through the modulation of multiple targets, for most drugs and drug candidates these targets are largely unknown and identification among the thousands of gene products remains difficult. Furthermore, this polypharmacology often goes hand in hand with unintended off-target effects. A better knowledge of these drug-protein interactions, along with the molecular pathways involved and the associated diseases, could be of substantial value to drug development, in particular to predict side effects and explore potential drug repositioning.

In this thesis, we propose a computational method to support the identification of putative targets of a drug by means of a dual approach combining gene expression and protein interaction data with chemical structure similarity. The first component of our method tackles this target identification problem by the analysis of gene expression following drug treatment in the context of a functional protein association network. More specifically, genes are prioritized as potential targets based on the transcriptional response of functionally related genes. To this end, differential expression signals are diffused over the network either using a kernel-based random walk or on the basis of connectivity correlations between nodes. This diffusion strategy was evaluated on 235 publicly available gene expression data sets for treatment with bioactive molecules having a known target. AUC values up to 92% indicate the predictive power of integrating experimental gene expression data with prior knowledge on protein interactions to identify the targets of a drug.

As gene expression analysis is frequently employed to study the effects of drug compounds on cells, we have developed Galahad to facilitate the analysis

of this particular data type and enable an in-depth exploration of a drug's mode of effect. Galahad is a web-based application for the analysis of gene expression data from drug treatment versus control experiments, aimed at predicting a drug's molecular targets and biological effects. Galahad provides data quality assessment and exploratory analysis, as well as computation of differential expression. Based on the obtained differential expression values, drug target prioritization, as well as both pathway and disease enrichment, can be computed and visualized. All of the above functionalities are demonstrated on gene expression data for treatment with well-characterized drugs.

In addition to using gene expression, drug-protein interactions can also be predicted from structural information. Building on the similar property principle, the second component of our drug target prioritization method ranks proteins as potential drug targets based on the interaction with compounds structurally similar to the drug of interest. Compound-compound similarity scores are therefore combined with compound-protein interaction scores. Both our structure-based and expression-based approach produce a genome-wide ranking of potential targets that can eventually be fused to obtain a single ranking. This combined approach has been evaluated on a test set of 62 small molecule drugs with known targets. AUC values of up to 95% were obtained. These results indicate the predictive power of combining gene expression data and structural information for a drug of interest with known protein-protein and protein-compound interaction information respectively, to identify the targets of that drug. As such our method can aid in gaining a better knowledge of a candidate drug's mode of action and its off-target effects and thus be of value in the drug development process.

Beknopte samenvatting

De farmaceutische industrie staat voor ongekende druk om hun productiviteit te verhogen. De ontwikkeling van nieuwe geneesmiddelen wordt steeds vaker stopgezet in de latere stadia van het ontwikkelingsproces, met toxiciteit en gebrek aan werkzaamheid als belangrijkste knelpunten. Om deze veiligheids- en werkzaamheidsgerelateerde problemen aan te pakken, is een beter begrip van de complexe biologische reactie op de behandeling met geneesmiddelen van vitaal belang. Hoewel veel geneesmiddelen hun therapeutische activiteiten uitoefenen door middel van interactie met meerdere doelwitmoleculen, zijn deze doelwitten voor de meeste (kandidaat) geneesmiddelen grotendeels onbekend en blijft identificatie onder de duizenden genproducten moeilijk. Bovendien gaat deze polyfarmacologie vaak hand in hand met onbedoelde 'off-target' effecten. Een betere kennis van deze interacties tussen geneesmiddelen en eiwitten, tezamen met de betrokken moleculaire signaalwegen en bijhorende ziekteprocessen, kan van substantiële waarde zijn voor de ontwikkeling van geneesmiddelen, met name voor het voorspellen van bijwerkingen en het onderzoeken van de potentiële herpositionering van geneesmiddelen.

De eerste component van onze methode behandelt dit doelwitidentificatie-probleem door middel van analyse van genexpressie na behandeling met geneesmiddelen tezamen met een functioneel eiwitassociatienetwerk. Meer specifiek worden genen geprioritiseerd als potentiële doelwitten op basis van de transcriptierespons van functioneel gerelateerde genen. Daartoe worden differentiële expressiesignalen verspreid over het netwerk, ofwel met behulp van een kernel-gebaseerde 'random walk' of op basis van connectiviteitscorrelaties tussen knooppunten. Deze diffusiestrategie werd geëvalueerd aan de hand van 235 publiek beschikbare genexpressie datasets voor behandeling met bioactieve moleculen met een bekend doelwit. AUC-waarden tot 92% geven de voorspellende kracht aan bekomen door integratie van experimentele genexpressiegegevens met voorafgaande kennis omtrent eiwitinteracties om de doelwitmoleculen van een geneesmiddel te identificeren.

Daar genexpressie-analyse vaak gebruikt wordt om de effecten van geneesmiddelen op cellen te onderzoeken, hebben we Galahad ontwikkeld om de analyse van dit specifieke datatype te vergemakkelijken en studie van de werking van een geneesmiddel mogelijk te maken. Galahad is een webgebaseerde applicatie voor de analyse van genexpressiegegevens afkomstig van geneesmiddelbehandeling versus controle-experimenten, gericht op het voorspellen van de moleculaire doelwitten van een geneesmiddel en bijhorende biologische effecten. Galahad voorziet in een kwaliteitscontrole van de data, een verdere verkennende analyse, evenals in de berekening van differentiële expressie. Gebaseerd op de verkregen differentiële expressiewaarden, kan zowel prioritisatie van potentiële doelwitmoleculen als identificatie van betrokken signaalwegen en gerelateerde ziekten berekend en gevisualiseerd worden. Al de bovenstaande functionaliteiten worden tevens gedemonstreerd aan de hand van genexpressiedata voor behandeling met goed gekarakteriseerde geneesmiddelen.

Naast het gebruik van genexpressie kan ook structurele informatie aangewend worden voor het voorspellen van interacties tussen een geneesmiddel en proteïnes. Gebaseerd op het 'similar property principle' rangschikt het tweede onderdeel van onze doelwitidentificatiemethode potentiële doelwitmoleculen op basis van de interactie met verbindingen die structureel vergelijkbaar zijn met een zeker geneesmiddel. Hiertoe worden de similariteitsscores tussen verbindingen gecombineerd met scores voor de interacties met proteïnes. Zowel onze structuurgebaseerde als onze expressiegebaseerde methode resulteert in een genoomwijde rangschikking van potentiële doelwitmoleculen die uiteindelijk samengevoegd kunnen worden om een enkele rangschikking te verkrijgen. Deze gecombineerde aanpak werd geëvalueerd aan de hand van een testgroep van 62 geneesmiddelen met gekende doelwitten. AUC-waarden tot 95% werden verkregen. Deze resultaten geven de voorspellende kracht aan die bekomen kan worden door het combineren van genexpressiedata en structurele informatie voor een geneesmiddel met respectievelijk gekend proteïne-proteïne- en proteïne-compound interactieinformatie om de doelwitmoleculen van dat geneesmiddel te identificeren. Als zodanig kan onze methode bijdragen tot het verwerven van een betere kennis omtrent de werking van een kandidaatgeneesmiddel en diens 'off-target' effecten en dus van waarde zijn in het geneesmiddelontwikkelingsproces.

Abbreviations

2D	two-dimensional
ATC	Anatomical Therapeutic Chemical
AUC	area under the curve
BioGRID	Biological General Repository for Interaction Datasets
BLA	Biologic License Application
BLAST	Basic Local Alignment Search Tool
cAMP	cyclic adenosine monophosphate
CETP	cholesteryl ester transfer protein
cGMP	cyclic guanosine monophosphate
CMap	Connectivity Map
CSV	comma-separated values
CTD	Comparative Toxicogenomics Database
CYP	Cytochrome P450
CYP24A1	Cytochrome P450 Family 24 Subfamily A Member 1
CYP2C9	Cytochrome P450 Family 2 Subfamily C Member 9
CYP3A4	Cytochrome P450 Family 3 Subfamily A Member 4
DeMAND	Detecting Mechanism of Action by Network Dysregulation
DHCR7	7-Dehydrocholesterol Reductase
DILI	drug-induced liver injury
DIP	Database of Interacting Proteins
DNA	deoxyribonucleic acid
DUSP10	Dual Specificity Phosphatase 10
ECFP	Extended Connectivity FingerPrint
FDA	US Food and Drug Administration
FDPS	Farnesyl Diphosphate Synthase
FDR	false discovery rate
FOS	Fos Proto-Oncogene, AP-1 Transcription Factor Subunit
G6PD	Glucose-6-Phosphate Dehydrogenase
GAPDH	glyceraldehyde 3-phosphate dehydrogenase
GEO	Gene Expression Omnibus

GLIDA	GPCR-LIgand DAtabase
GPCR	G protein-coupled receptor
G protein	guanine nucleotide-binding protein
GRID	General Repository for Interaction Datasets
GSEA	gene set enrichment analysis
GTP	guanosine triphospate
HDL	high-density lipoprotein
hERG	human ether-a-go-go-related gene
HGNC	HUGO Gene Nomenclature Committee
HMGCR	3-hydroxy-3-methyl-glutaryl coenzyme A reductase
HMGRI	3-hydroxy-3-methyl-glutaryl coenzyme A reductase inhibitor
HMGCS1	3-Hydroxy-3-Methylglutaryl-CoA Synthase 1
HMGCS2	3-Hydroxy-3-Methylglutaryl-CoA Synthase 2
HPO	Human Phenotype Ontology
HPRD	Human Protein Reference Database
HT	hydroxytryptamine
IC_{50}	half maximal inhibitory concentration
ID	identifier
IID	Integrated Interactions Database
IL-6	Interleukin 6
InChI	International Chemical Identifier
INSIG1	Insulin Induced Gene 1
KEGG	Kyoto Encyclopedia of Genes and Genomes
LDL	low-density lipoprotein
MACCS	Molecular ACCess System
MANTRA	Mode of Action by NeTwoRk Analysis
MeSH	Medical Subject Heading
MGI	Mouse Genome Informatics
MINT	Molecular INTeraction
MKL	multiple kernel learning
MoA	mechanism of action
MPO	Mammalian Phenotype Ontology
MVD	Mevalonate Diphosphate Decarboxylase
MVK	Mevalonate Kinase
mRNA	messenger ribonucleic acid
NCBI	National Center for Biotechnology Information
NCI	US National Cancer Institute
NDA	New Drug Application
NIH-NCATS	National Institutes of Health – National Center for Advancing Translational Sciences
NME	new molecular entity

NMR	nuclear magnetic resonance
NUSE	normalized unscaled standard error
OMIM	Online Mendelian Inheritance in Man
PTH	Parathyroid Hormone
PCA	principal component analysis
PDB	Protein Data Bank
PDE5	phosphodiesterase 5
PDSP	Psychoactive Drug Screening Program
PharmGKB	Pharmacogenomics Knowledgebase
PhRMA	Pharmaceutical Research and Manufacturers of America
PID	Pathway Interaction Database
PINA	Protein Interaction Network Analysis
PPARA	peroxisome proliferator-activated receptor alpha
PPI	protein-protein interaction
PTH	parathyroid hormone
R&D	Research & Development
RDO	RGD Disease Ontology
RGD	Rat Genome Database
RLE	relative log expression
RMA	robust multi-array average
ROC	receiver operating characteristic
SAR	structure-activity relationship
SEA	Similarity Ensemble Approach
SIDER	Side Effect Resource
SIMAP	Similarity Matrix of Proteins
SMARTS	SMiles ARbitrary Target Specification
SMILES	simplified molecular-input line-entry system
SQL	structured query language
SQLE	Squalene Epoxidase
STITCH	Search Tool for Interacting Chemicals
TM7SF2	Transmembrane 7 Superfamily Member 2
TNFα	Search Tumor Necrosis Factor alpha
TLR	Toll-like receptor
TTD	Therapeutic Target Database
US(A)	United States (of America)
VDR	Vitamin D receptor

Contents

Abstract iii

Abbreviations vii

Contents xi

List of Figures xv

List of Tables xvii

1 Introduction 1

1.1 Drug discovery and development today 1

1.1.1 The productivity challenge and the curse of attrition . . 1

1.1.2 From systems-driven to target-driven drug discovery and back 6

1.1.3 Drug repositioning 8

1.2 Drug target prediction 10

1.2.1 Structure-based and ligand-based approaches 10

1.2.2 Phenotype-based approaches 12

1.2.3 Applications of drug target prediction 14

1.3 Network analysis for drug discovery 17

1.3.1 Protein-protein interaction networks 17
1.3.2 Drug-drug similarity networks 18
1.3.3 Drug-target networks 20
1.3.4 Propagating information through networks 22
1.4 Objective and outline of the thesis 23

2 Network diffusion of differential expression for drug target prediction 25
2.1 Introduction 26
2.2 Materials and methods 28
2.2.1 Functional protein association network 28
2.2.2 Drug target prioritization strategies 29
2.2.3 Test data 32
2.2.4 ROC analysis 33
2.3 Results and discussion 33
2.3.1 Performance of the drug target prioritization methods . 34
2.3.2 Impact of the network 36
2.3.3 Impact of the expression data 41
2.3.4 Comparison of the performance on different drug and target types 43
2.4 Conclusion 43

3 Galahad: a web server for drug effect analysis from gene expression 47
3.1 Introduction 48
3.2 Input 48
3.3 Methods 49
3.3.1 Quality control 49
3.3.2 Data exploration 50
3.3.3 Differential expression 50

3.3.4 Drug target prioritization 52
3.3.5 Enrichment 52
3.4 Output 52
3.4.1 Quality control and data exploration 53
3.4.2 Differential expression 53
3.4.3 Drug target prioritization 54
3.4.4 Enrichment 56
3.5 Implementation 58
3.6 Case studies 59
3.6.1 Calcitriol 59
3.6.2 Atorvastatin and rosuvastatin 61
3.7 Conclusion 64

4 A dual approach combining network diffusion of differential expression with structural similarity for improved drug target prediction 65

4.1 Introduction 66
4.2 Materials and methods 67
4.2.1 Ligand-based drug target prioritization 67
4.2.2 Expression-based drug target prioritization 69
4.2.3 Rank aggregation 70
4.2.4 ROC analysis 71
4.2.5 Training and test data for ligand-based drug target prioritization 72
4.2.6 Test data for ligand-based, expression-based, and combined drug target prioritization 72
4.3 Results and discussion 72
4.3.1 Protein scoring function 72
4.3.2 STITCH evidence types 73

4.3.3 Molecular fingerprint 74
4.3.4 Performance on test data and comparison with SEA . . 75
4.3.5 Effect of ligand similarity on the performance 76
4.3.6 Improvement of gene expression-based drug target prioritization 77
4.4 Conclusion 77

5 Discussion **80**

5.1 Representation of pharmacological space 80
5.2 Data integration 82
5.3 Drug target prediction 83
5.4 Applicability of the work in a pharmaceutical research context 85
5.5 Future perspectives 87
5.6 General conclusion 89

Bibliography **91**

List of publications **117**

List of Figures

1.1 The number of new drugs approved by the FDA per billion US dollars spent on R&D 2

1.2 The drug development process 3

1.3 Clinical phase success rates 4

1.4 Causes of failure in clinical phase attrition 5

1.5 Progression of the drug discovery process 7

1.6 An overview of the most common structure- and ligand-based approaches for drug target prediction 13

1.7 Giant component of the drug-target network of FDA-approved drugs constructed by Yildirim and co-workers 21

1.8 Structure of the thesis 24

2.1 Drug target prioritization strategies 31

2.2 ROC curves comparing the performance of the two STRING versions 40

2.3 ROC curves illustrating the effect of the experiment, drug, and target type on the performance 42

3.1 The Galahad submission page 50

3.2 Overview of the different Galahad analysis steps 51

3.3 ROC curves comparing the performance of network diffusion-based drug target prioritization and prioritization based on differential expression only 53

3.4 The quality control output 54

3.5 The data exploration output 55

3.6 The differential expression output 55

3.7 The drug target prioritization output 56

3.8 Interaction network for a potential target 57

3.9 The pathway (and disease) enrichment output 57

3.10 Pathway enrichment network 58

3.11 Interaction network for the calcitriol target VDR 62

3.12 Interaction network for the atorvastatin target HMGCR 64

4.1 Schematic overview of protein score computation, normalization and ranking 69

4.2 Distribution of the Tanimoto similarity for all compounds in the test data set with all compounds in STITCH, for each of the three fingerprints 76

4.3 ROC curves comparing the performance of expression-based, ligand-based, and combined drug target prioritization 78

5.1 Computational methods for target prediction based on network analysis of gene expression illustrating the growing popularity of this approach 85

List of Tables

2.1 AUC values for the different drug target prioritization strategies 37

2.2 AUC values for kernel diffusion ranking using different diffusion rates α 38

2.3 AUC values for correlation diffusion ranking using different correlation thresholds s 39

2.4 AUC values for the different protein association evidence types in STRING resulting from one-step symmetric normalized Laplacian kernel diffusion and correlation diffusion of $\log_2$ ratios 41

4.1 Number of proteins, ligands and interactions for each of the STITCH evidence types 68

4.2 AUC values for the different protein scoring functions and molecular fingerprints 73

4.3 AUC values for the different STITCH evidence types and molecular fingerprints 74

4.4 AUC values for the test data after removal of the most similar ligands for each query molecule 77

Chapter 1

Introduction

1.1 Drug discovery and development today

1.1.1 The productivity challenge and the curse of attrition

De novo drug development is a complex, time-consuming, and very costly process [1]. Recent data indicate that the cycle for introduction of a new molecular entity (NME) now can take more than fiteen years from project inception to market–with on average five years spent in preclinical research and around nine years required for a drug candidate to progress through the clinical phases of development [2, 3, 4]. The costs of development are also increasing [5]. DiMasi and co-workers reported a total capitalized cost per approved drug of almost 2.6 billion US dollars, based on data analysis for 1,442 investigational drugs from ten multinational pharmaceutical firms of varying sizes [6]. Although widely cited, this number should not be considered a gold standard, as published estimates vary strongly due to the lack of transparency of the data sources and analytical methods [7], as well as the lack of a standard methodology to assess attrition rates and investment costs [8]. It has even been suggested that these estimates of high costs for developing drugs are being published in part as justification for the cost of drugs to consumers [2, 9].

Nevertheless, the past decades have been a period of turbulence and change for the pharmaceutical industry, as they are facing unprecedented productivity challenges. To maintain income, rounds of mergers and acquisitions have occurred such that only eleven out of forty-two members of PhRMA existing in 1988 now survive [10]. Research and development (R&D) output in terms

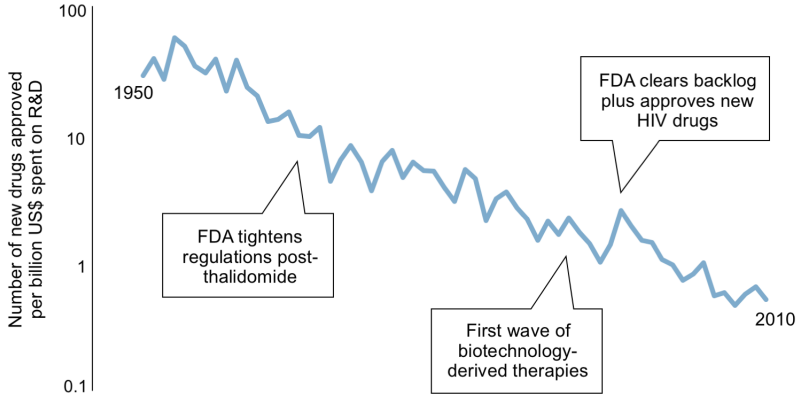


Figure 1.1: The number of new drugs approved by the FDA per billion US dollars spent on R&D. This number is halving roughly every nine years (adapted from [14]).

of the number of new drugs has remained relatively constant throughout this period, while R&D spending has increased dramatically [11, 12], to almost one hundred billion US dollars per year at present [13]. Figure 1.1 shows this decline in R&D productivity as measured by the number of new FDA-approved drugs per billion US dollars spent on R&D per year, halving roughly every nine years. This trend is called 'Eroom's Law', in contrast to the more familiar Moore's Law [14]. With the costs per new drug having soared into billions of dollar, this raises questions about the sustainability of the industry's R&D business model [15]. Several possible explanations for the divergence of R&D spending and new product approvals have been offered by professionals in the industry, such as the 'low-hanging fruit' problem, resulting from the progressive exploitation of drug targets that are more technically tractable, the 'better than the Beatles' problem, resulting from the ever-improving back catalog of approved medicines, and the 'cautious regulator' problem, resulting from the progressive lowering of the risk tolerance of drug regulatory agencies [14]. Other issues include loss of revenue due to patent expirations, increasingly cost-constrained healthcare systems, and commercial and financial decisions driving project termination [1, 16]. However, the biggest issue facing the pharmaceutical industry today is perhaps the increased complexity and cost of clinical trials and the related 'curse of attrition' [5, 16, 17].

Following the high-throughput screening of thousands of molecules during drug discovery and elaborate testing of the identified and optimized lead compounds, both *in vitro* and *in vivo*, during the preclinical phase, up to a handful of drug candidates is selected to be studied in humans during clinical trials (Figure 1.2).

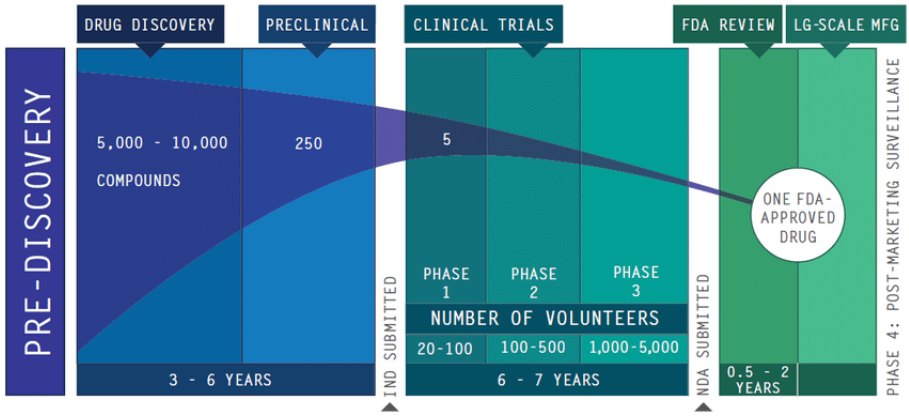


Figure 1.2: The drug development process. The complete drug development process, from drug discovery to marketing, can take up to fifteen years, with on average only one in 10,000 compounds tested eventually being approved for sale (reprinted from [22]).

Classically, these clinical trials are divided into Phases 1 through 4. In Phase 1 the candidate drug is tested in humans for the first time. In a small group of usually about twenty to one hundred healthy volunteers, researchers evaluate the treatment's safety, determine a safe dosage range, and identify short-term side effects. Once a safe dose or range of doses is determined, Phase 2 clinical trials seek further safety data, as well as preliminary evidence in support of biological effect, by administering the drug to several hundred patients with the disease or condition under study. Researchers study both the pharmacokinetics and pharmacodynamics of the drug during this phase and establish the optimal therapeutic dose range to be used in Phase 3. In this Phase 3, the drug is studied in a larger, more diverse patient population, usually involving between 1,000 and 3,000 patients, to assess the effectiveness of the drug candidate as well as its benefit/safety profile and thereby its value in clinical practice. Phase 3 trials are both the longest and costliest trials, carrying more than 60% of the overall trial costs [18]. Following the completion of Phases 1 through 3, an NDA (New Drug Application) or BLA (Biologic License Application) with all scientific information is filed with a regulatory authority such as the FDA, in order to request approval to market the drug. Studies conducted after regulatory approval and marketing of the drug are referred to as Phase 4 trials and are meant to evaluate rare but serious effects that cannot be assessed in the smaller Phase 3 studies [3, 19, 20, 21].

Thomas and co-workers analyzed clinical and regulatory phase success rates over a ten-year time frame, from 2006 to 2015, for 7,455 development programs

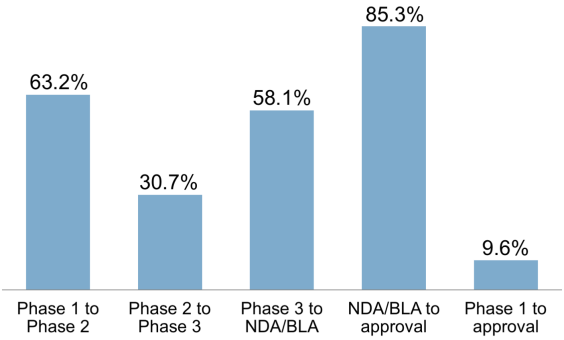


Figure 1.3: Clinical phase success rates. The probability that a drug will successfully advance to the next phase is depicted for Phase I up to approval (adapted from [23]).

across 1,103 companies. As can be deducted from Figure 1.3, only one out of ten development paths in Phase 1 eventually results into a marketed product. Put another way, 90% of new drug candidates entering human clinical trials fail. The weakest link in the chain is Phase 2, where proof-of concept is tested in humans for the first time: only about 30% of the compounds that enter Phase 2 proceed to full Phase 3 clinical trials [23]. Attrition in Phase 3 also remains an important determinant of overall R&D efficiency, as its low survival percentages are further exacerbated by the very high costs associated with attrition at such an advanced stage [5]. Breaking down success rates by therapeutic area shows substantial variation among disease with overall success rates ranging from as high as 26% for hematology down to a modest 5% for oncology, reflecting the complex and heterogeneous nature of cancer. Furthermore, drug candidates for rare diseases, often identified by specific genetic mutations, out-perform the average with an overall success rate of 25%. Finally, success rates for biologics are higher than for small molecules, while non-NMEs show success rates well above those of NMEs in every phase [23].

Causes of failure change during the course of development. Early in the process, compounds fail primarily for safety reasons. Compounds that successfully navigate Phase 1 increasingly drop out due to lack of efficacy in Phase 2 and 3 [24]. Using data gathered from Thomson Reuters' Drugs of Today, Harrison analyzed a total of 218 drugs for major indications that were dropped between Phase 2 and submission during the period 2013-2015. The reasons of failure were disclosed for 174 of these (Figure 1.4 A). Insufficient efficacy—defined as undifferentiated from current standard care, no advantage as add-on therapy, or no advantage versus placebo—appeared to be the foremost reason (56%). When looked at by phase, the proportion of failures due to lack of efficacy was

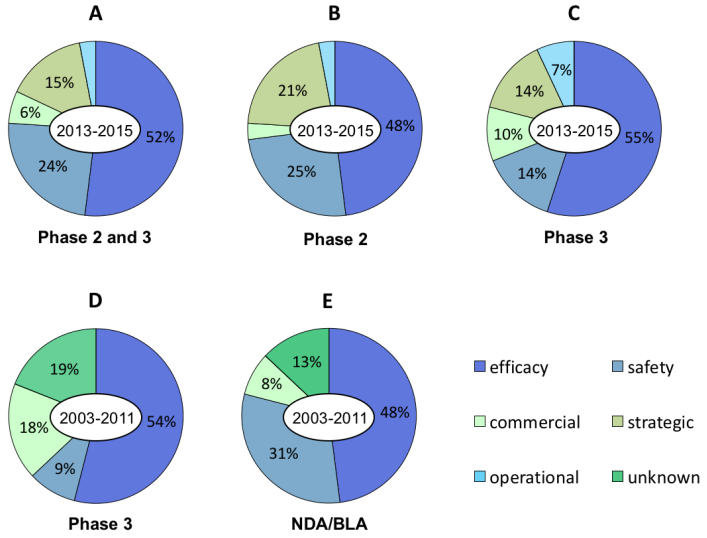


Figure 1.4: Causes of failure in clinical phase attrition. Figures 1.4 A-C are based on the work of Harrison on data for 174 drugs that failed during Phase 2 or 3 in the period 2013-2015. Figures 1.4 D and E are based on the work of Hay and co-workers on data for 359 drugs that failed in Phase 3 or after submission between 2003 and 2011 (adapted from [16, 26]).

already disturbingly high in Phase 2 (Figure 1.4 B), but at 55% even higher in Phase 3 (Figure 1.4 C) [25]. This number is consistent with the earlier findings of Hay and co-workers on 359 Phase 3 suspensions between 2003 and 2011 (Figure 1.4 D), with over half of them attributable to some measure of efficacy. With only 9%, safety was found to be the least likely cause for attrition in Phase 3, potentially due to identification of most adverse events earlier in the development process. In contrast, approximately one third of submission failures were attributable to safety concerns raised by regulators (Figure 1.4 E) [16].

While clinical success rates have been declining for more than a decade, recent data now show encouraging trends for the pharmaceutical industry. Smietana and co-workers tracked the progress of more than 9,200 compounds in development using a rolling three-year window between 1996 and 2014. A gradual increase in Phase 2 and Phase 3 rates was observed, resulting in a cumulative success rate of 11.6% in 2011-2014, up from a low of 7.5% in 2008-2011. Although a decreasing proportion of successful drugs was observed in Phase 1, this can be considered a positive trend, suggesting an increasingly thorough early evaluation preventing costly failures later on [27]. Data from

CMR International for the period 2010-2017 confirm increasing Phase 3 success rates, but show Phase 2 success rates to remain relatively static [28].

Increasing success rates can be partially attributed to the industry pipeline growing in successful areas. Hot spots of innovation are emerging in fields that have long been stagnant, such as recent successes of immunotherapy in oncology. Moreover, rare disease studies now account for nearly a quarter of all Phase 2 and Phase 3 trials combined and biologics currently constitute 40% of the overall pipeline, up from 20% about twenty years ago [27, 29]. Looking at the causes of failure, the proportion of both efficacy and safety failures in Phase 3 trials has decreased, while the proportion of failures for commercial and strategic reasons, including changes in therapeutic focus and projects being dropped owing to mergers and acquisition, increased substantially. Together with the increasing proportion of failures due to safety in Phase 2 trials, these observations suggest efforts to remove unsafe drug candidates earlier in development are starting to pay off. Though efficacy is still the number one reason for failure, this is expected to improve further in the coming years as trial populations are increasingly stratified using efficacy biomarkers and both understanding of disease biology and target identification advance [26].

1.1.2 From systems-driven to target-driven drug discovery and back

Since the dawn of the genomic era in the early 1990s, the main focus of drug discovery has been on the design of selective ligands, supposedly acting on a single disease target, typically a protein that appears to have a key role in disease pathogenesis. This target-centric approach, based on the 'one gene, one drug, one disease' paradigm, arose from a congruence between genetic reductionism and increasingly sophisticated molecular biology technologies enabling isolation and characterization of individual disease-causing genes. Although 'being blind' to other processes that are inevitably connected through complex networks, drug design at the single molecular target level provides a rational basis for the discovery of new medicines, which was proved successful by drugs such as imatinib (Gleevec) and gefitinib (Iressa) [30, 31, 32]. Before the introduction of hypothesis-driven approaches, drug discovery was driven primarily by the evaluation of compounds against phenotypes in assays preserving the cellular context of protein function, but often with limited knowledge of the molecular mechanisms of disease [31, 32]. With the recent advances in assay technology and the alarming attrition rates in late-stage clinical development, many research programs are increasingly returning to such cell- or organism-based phenotypic assays. This rather empirical approach offers the potential to uncover new therapeutic principles and molecular pathways for currently untreatable

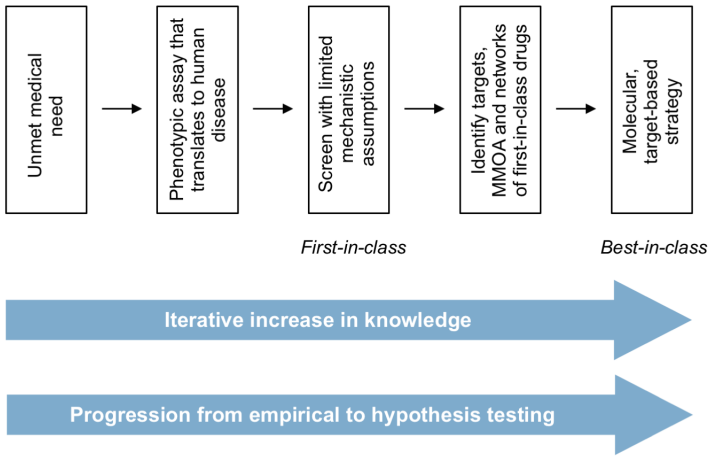


Figure 1.5: Progression of the drug discovery process. Progression from empirical to hypothesis testing, correlated with the iterative increase in knowledge to specifically address a certain phenotype related to an unmet medical need (adapted from [32]).

diseases [33]. The cost paid for this benefit is the often limited understanding of the precise functional protein targets and key mediators of the phenotypic change caused by the small molecule in a certain disease model, making target prediction a crucial step in present-day drug development [34]. In the end, rather than choosing between one or the other, systems and target approaches can complement each other: early in the process, knowledge can be achieved by empirical analysis, but the more knowledge is gained to link a specific molecular mechanism of action (MoA) to the desired phenotype, the more drug discovery can focus efforts addressing specific hypotheses of target-based therapeutic intervention and enable the discovery of follow-on drugs (Figure 1.5) [31, 32, 33].

The growing understanding on the role of networks in the phenotypic robustness of biological systems has moved drug research to a more systems-driven approach as well [30, 35]. Integrating drug discovery with systems biology shed light on the topology of drug-target interactions and several lines of evidence now suggest that most drugs interact with multiple proteins rather than with a single target [35, 36, 37, 38]. As this can lead to both beneficial and harmful outcomes, this concept of polypharmacology has significant implications for tackling the two major sources of attrition in drug development: efficacy and toxicity [30]. Focusing more and more on diseases involving multiple genes, identifying potential drugs has become even more complicated [39]. The therapeutic challenge of complex diseases often requires polypharmacology across several

disease-relevant targets to effectively reset the regulatory network altered in the disease state and thus perturb these often robust pathological imbalances [30]. In particular, many protein kinase inhibitors [40] and antipsychotics [41] appear to bind more than twenty proteins [42], with some 'off-target' activities being essential for efficacy [43]. Besides improving therapeutic efficacy, a drug active on multiple targets is expected to be less prone to the emergence of drug resistance mutations in oncology and infectious diseases, as well as to therapeutic target-related adverse effects [44]. Furthermore, multi-target activity offers possibilities for drug repositioning, a research area quickly gaining interest because of the proven shortcut between laboratory and clinic [35, 38].

1.1.3 Drug repositioning

Pharmaceutical companies have adopted a number of strategies to offset the problems caused by the fall in R&D productivity. One area that remains relatively underexploited, but is now rapidly emerging as an increasingly important strategy is drug repositioning, also called drug repurposing, drug reprofiling, indication switching, or indication expansion [1, 45, 46]. As its name implies, drug repositioning describes the process of identifying and developing new clinical indications for existing drugs. This concept can be traced back to the late 1950s [47] and the use of reticulose as a drug for post-radiation effects is among the earliest examples [48]. More recent efforts take inspiration from these classic success stories, the most famous undoubtedly being Pfizer's blockbuster Viagra (sildenafil). This phosphodiesterase 5 (PDE5) inhibitor was under development to treat pulmonary arterial hypertension and angina pectoris in the 1980s, but clinical trials got suspended when the desired cardiovascular effects were not observed in Phase 1 healthy volunteers. However, unexpected reporting of erectile side effects led to a large-scale study carried out on 3,700 men worldwide with erectile dysfunction. In 1998, the company won approval for and began marketing sildenafil as the first oral treatment for erectile dysfunction in the USA, with PDE5 being the molecular target responsible for its famous 'side effect' [45, 49].

Around 3,000 drugs have been approved by at least one country and thus represent a vast untapped resource for alternative indication discovery [46]. Furthermore, the increasing realization that many existing drugs do indeed provide opportunities for additional therapeutic indications made the industry alert for this potential within their pool of developmental drugs. As approximately 150 to 200 compounds fail in the various phases of clinical trials each year and the number of products sitting on big companies' shelves is estimated to range from 2,000 to 30,000, these developmental drugs that have failed in their original intended indication offer a huge playground for drug

repositioning as well. Such so-called drug rescue can become one of the major sources for revenue growth within the pharmaceutical industry [24, 50, 51, 52], especially because a repositioned compound can often benefit from shorter development times, reduced costs, and lower risks. Development risk is reduced when repositioning candidates have already been through several stages of clinical development resulting in well-known safety and pharmacokinetic profiles, substantially lowering the risk of producing dramatic side effects in later phases. Shorter routes to the clinic are also possible as *in vitro* and *in vivo* screening, chemical optimization, toxicology, bulk manufacturing, formulation development, and even early clinical development have, in many cases, already been completed and can therefore be bypassed, accelerating a compound's entry into clinical Phase 2. Repositioning a drug candidate can reduce the traditional time line and make the drug available for use in patients in three to twelve years [45, 53, 54]. Finally, both the seventeen neglected tropical diseases affecting up to 1 billion of the world's population and the more than 7,000 rare diseases affecting cumulatively six to seven percent of the population have a great prospect for drug repositioning. One example of such a repurposed drug is miltefosine, initially developed for breast cancer but now used for treating visceral leishmaniasis [55, 56].

The serendipity responsible for most earlier drug repositioning discoveries is giving way to systematic searches for candidates and branching drug development processes with several parallel indications. Three to four specialized drug repositioning companies are started every year and most large pharma companies, such as Pfizer, AstraZeneca, and GlaxoSmithKline, have created units dedicated to systematic scanning of repositioning opportunities [46, 56]. Moreover, drug repositioning has recently seen a number of initiatives driven by academic and non-profit organizations, often in partnership with pharmaceutical companies, such as the National Institutes of Health – National Center for Advancing Translational Sciences (NIH-NCATS) and AstraZeneca – MRC collaborations, paralleled by the availability of clinical compound collections such as the NIH and Johns Hopkins clinical collections. Independently of these collaborations, pharmaceutical companies are also getting more inclined to make certain compound libraries and assays available to the wider scientific community [57]. With the advances in technology and the accompanying accumulation of omics data, bioinformatics is starting to play a pivotal role in accelerating drug repurposing studies. New repositioning strategies and techniques are emerging, integrating data from various sources, like pharmacological, genetic, chemical and clinical data. These computational methods and predictive models usually generate a handful of interesting hits, for which experimental validation through both *in vitro* and *in vivo* models is required to select the candidate hits for preclinical drug evaluation [48, 58].

1.2 Drug target prediction

At the most basic level, drugs usually exert their effects on biological systems by binding with protein targets and affecting their downstream activity. For active compounds isolated from phenotypic screens, knowledge of these interactions is key toward understanding and predicting higher-level information, such as therapeutic and off-target mechanisms, as well as novel indications. Obtaining information on mode of action is essential for the development of drugs with superior pharmacological properties and fewer side effects [59, 60]. Both prior to and during clinical trials substantial research is now undertaken to get a more comprehensive view on this [39]. As target identification from phenotypic screens is expected to generate a spectrum of possible targets, the term 'target deconvolution' was coined to more accurately define this specific target prediction task [61]. Many efforts are, however, also being made to retrospectively predict and characterize the targets of both failed developmental and marketed drugs, as the multi-target interactions of most existing drugs are still largely unknown [43, 50, 62]. Experimental target identification remains both cost- and time-intensive, but with the wealth of biological and pharmacological data available today, computational methods can be of significant value here. Drug target prediction can be approached *in silico* by exploiting either molecular features of chemicals and proteins, or phenotypic effects following drug treatment [63, 34]. Combining both may provide an even more powerful toolbox for the characterization of on- and off-target effects and understanding of mechanisms of small molecule action [64].

1.2.1 Structure-based and ligand-based approaches

The traditional computational target prediction methods exploiting molecular structure information can be classified either as structure/target-based or as ligand-based. An overview of the most common approaches is given in Figure 1.6. Structure-based methods generally rely on the three-dimensional structure of proteins to predict their interaction with a given compound and mainly comprise docking studies (Figure 1.6 D) [34]. Molecular docking uses three-dimensional modeling and computer simulation to predict the binding orientation and affinity of a small molecule into a protein's binding pocket. More conventionally, ligands are identified as interaction partners for a given protein, but this concept has also been reverted to dock one small molecule against a panel of protein targets. This inverse docking approach has been successfully explored in recent years for target prediction. Besides identifying potential targets, molecular docking can also provide structural insights into the nature of a compound-protein interaction, enabling further optimization of the compound's

structure toward a higher binding affinity for its target. However, several drawbacks prohibit inverse docking from attracting a wider acceptance in the pharmaceutical industry. The main limitation is the lack of reliable three-dimensional protein structures coming from X-ray crystallography, NMR, or homology modeling, especially for membrane-bound proteins. Furthermore, the approach is computationally very intensive, limiting its feasibility for large-scale prediction tasks [38, 60, 65]. To circumvent these issues, certain methods relying on the analysis of structural drug-target interaction information have been developed, including pharmacophore-based methods (Figure 1.6 C) and binding site comparison (Figure 1.6 E). Pharmacophore-based methods make use of an abstract depiction of the molecular features necessary for a compound to be recognized by a molecular target and trigger a pharmacological effect. The consecutive fitting of a compound into pharmacophore models representing different targets is less computationally demanding with respect to docking. Target prediction by means of binding site comparison is based on the observation that interaction pattern similarity has a tight correlation with binding site similarity and relies on computational methods predicting, representing, and comparing binding pockets for specific ligands [38, 66].

While structure-based or target-based methods rely on protein structure, ligand-based methods focus on the chemical structure of the compound, often exploiting similarities with other small molecules to infer its activity. Analogous to the similar property principle, stating that molecules with similar structures have similar properties [67], the underlying principle of these ligand-based target prediction methods is that pairs of molecules acting on a common target tend to be similar, as supported by the work of Cleves and Jain [38, 68, 69]. The most straightforward approach makes target predictions based on pairwise comparisons between individual molecules. In this way, huge chemical libraries can be mined in a short time to identify the ligands most similar to a query compound and thus its most probable targets. To do this, molecular descriptors are generated from the two-dimensional representations of both the query compound and a set of target-annotated small molecules (Figure 1.6 A). Though a huge variety of molecular descriptors encoding different molecular features is available, circular fingerprints in their various definitions seem to have the best correlation with bioactivity. Alternatively, chemical similarity can be determined by aligning three-dimensional models of the compounds (Figure 1.6 B) [65, 70]. However, at present, none of the many different types of three-dimensional representations that have been described has clearly emerged as the most appropriate one for this task. More specialized chemical similarity methods have also been developed, taking the similarity of target proteins into account by predicting the binding of drugs to proteins related to known drug targets [63]. Finally, drug targets can also be predicted using data mining to perform compound classification in vast chemogenomic space. Machine learning

methods such as support vector machines, linear regression models, binary kernel discrimination, naive Bayesian classifiers, and artificial neural networks can all be employed here [38, 65]. Compared with structure-based approaches, ligand-based approaches do not depend on three-dimensional structure information of targets or complete knowledge of ligand-target interaction mechanisms. Furthermore, descriptor calculations are relatively fast, meaning predictions for a new compound are rapidly available. However, results depend strongly on the available data. Predictions for targets with little or no bioactive ligands are infeasible, while other target classes such as G-protein-coupled receptors and kinases are overrepresented in bioactivity databases [38, 66].

1.2.2 Phenotype-based approaches

Computational methods generating target hypotheses from phenotypic assay information generally apply a 'guilt by association' approach. Based on the principle that compounds with the same mechanism of action show similar behavior across different biological assays, target or mechanism inference relies on the profiling of molecules with known mechanisms of action, allowing new compounds to be assessed for similar patterns of performance [34]. Commonly used profile types include gene expression, cytotoxicity, and chemical-genetic profiles [63]. However, this logic can be extended to other profiling methods, such as high-content screening [71], side effect profiling [72], and even behavioral profiling [73]. Phenotype-based approaches are becoming more prominent because of the growing availability of phenotypic information on drugs in public databases, such as ChEMBL [74], DrugBank [75], PharmGKB [76], and SIDER [77]. A noteworthy development in this domain is the Connectivity Map (CMap) [78, 79], a large collection of gene expression profiles derived from treatment of human cell lines with over 1,000 different small molecule perturbagens. A simple pattern matching algorithm compares these expression profiles with the expression pattern for a specific mutation, disease, or treatment of interest in order to search for functional associations with compounds from CMap.

The major limitation of phenotypic profiling is that drug target inference stays limited to 'landmark' compounds with known mechanisms of action. Nevertheless, the group of Butte has reported the successful repositioning of two FDA-approved drugs by comparing the gene expression signatures for marketed drugs from CMap to signatures for diseases extracted from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO). Both the prediction of the antiulcer drug cimetidine as a candidate therapeutic in the treatment of lung adenocarcinoma [80] and of the anticonvulsant topiramate as a novel therapeutic option for inflammatory

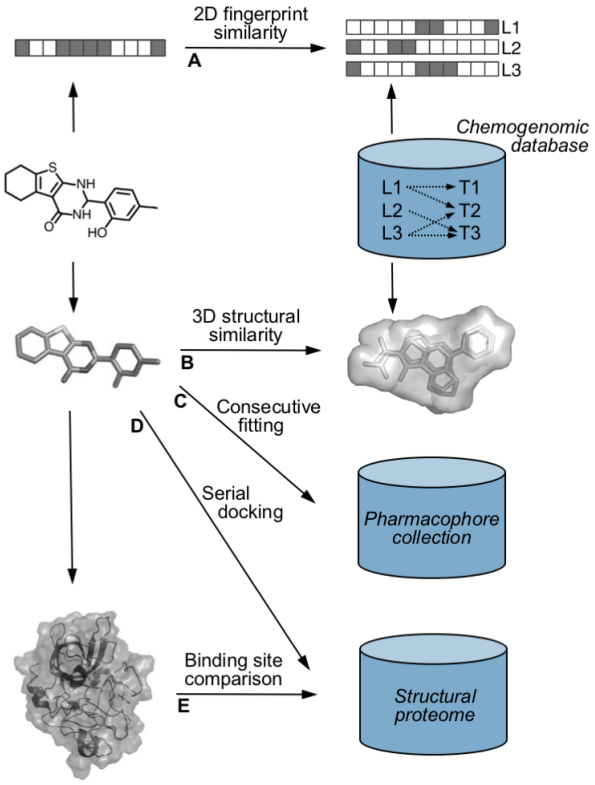


Figure 1.6: An overview of the most common structure- and ligand-based approaches for drug target prediction. (A-B) Ligand-based approaches exploiting structural similarity between the query compound and a set of target-annotated ligands, either based on (A) two-dimensional fingerprints or on (B) three-dimensional structures. (C) Pharmacophore-based approach screening the query compound against a set of pharmacophore models representing different targets. (D) Inverse docking of the query compound against a set of protein structures (E) Binding site comparison between a protein known to interact with the query compound and a set of protein structures (adapted from [38]).

bowel diseases [81] were experimentally validated. Genome-wide transcriptional profiling in both presence and absence of compound treatment can also play an important role in systems biology-based approaches analyzing differential expression at the level of a biological network. Information on the biological context of a protein (i.e., the relationship with other proteins) enables us to reveal insights into the action of drugs and the prediction of additional drug targets without the need for data from landmark compounds. Cosgrove and co-workers developed a sparse simultaneous equation model and associated inferential strategy for predicting directly perturbed genes from expression profiles. However, within this framework a directly targeted gene is interpreted as a gene that, adjusted for the regulatory effects of other genes, shows an elevated expression level. This interpretation led to an inconsistent performance for drug target prediction [82]. More recently Woo and co-workers developed DeMAND, a method elucidating compound MoA by assessing global dysregulation of a gene's regulon following compound perturbation. Changes in the joint gene expression probability density of a gene and each of its interaction partners in a regulatory network are measured, capturing both direct effects and, more importantly, modulation of the interaction partner's expression [83].

1.2.3 Applications of drug target prediction

Drug repositioning

One application of computational target prediction that has gained increasing attention is drug repositioning based on the biological role the predicted targets play in disease. A drug can be evaluated for repositioning to treatment of another disease, if one or more of its predicted protein targets are associated with a disease that is not among those for which the drug was originally indicated. Human genetics provides the opportunity to make these connections between specific genes and diseases: recent technological advances drastically reduced the cost of whole-exome sequencing and also genome-wide association studies offer a strong line of evidence for associating genetic variants and thus proximal genes with complex polygenic diseases [54].

A good example of a target-based drug repositioning approach was presented by Campillos and co-workers. In their work they make use of side effect similarity between marketed drugs to infer molecular activities that are not implied by chemical similarity. Proposing additional targets for existing drugs, potentially implicated in different therapeutic categories, this method has been applied successfully, leading to the patented repositioning of the antiemetic drug aprepitant for cancer treatment [72, 84]. Another example in this regard is the work of Kinnings and co-workers, who were able to reposition the

Parkinson's drug entacapone as a promising lead compound for a new class of antitubercular therapeutics using a chemical systems biology approach to identify off-targets [85]. By means of their similarity ensemble approach (SEA) based on drug and ligand structural similarities, Keiser and co-workers built a drug-target network and used it to predict novel targets for known drugs as well. Mebhydrolin for example, used since the 1950s as a symptomatic antihistamine but now being investigated for Alzheimer's disease, was found to interact with the 5-HT5A serotonin receptor, which may have implications for its role as an Alzheimer's disease therapeutic [43]. In a last example, Chavali and co-workers used a metabolic network-driven approach to link 254 FDA-approved drugs to fifteen metabolic genes and eight synthetic lethal double gene combinations, predicted to be relevant targets for the neglected tropical disease *Leishmania major*. *In vitro* validation against *L. major* promastigotes showed antileishmanial activity for the antimalarial drug halofantrine, as well as superadditivity for four combinations involving the drug disulfiram [86]. Finally, efforts have also been made to bring together relevant publicly available data scattered over different resources, facilitating a more comprehensive analysis. PROMISCUOUS, an exhaustive network-focused resource of protein-protein and protein-drug interactions enriched with side effects and structural information has been developed. This database enables exploration and understanding of the interplay between drugs and targets including off-target effects and thus can be used as a starting point for both indication finding and drug repositioning [87]. Likewise, Cockell and co-workers generated an integrated data set to visualize links between drugs, proteins, and diseases and facilitate the discovery of novel drug repositioning candidates using Ondex, a data integration and visualization platform for systems biology [88].

Side effect prediction

Information about a drug's target profile can also be employed for the *in silico* prediction of unwanted side effects. The current standard approach to safety screening is preclinical testing in animal disease models. However, such experiments are costly and results do not always reflect what will happen in patients, for instance due to genetic differences including small conservative changes in the binding pocket of a target. Computational approaches can help address this challenge by predicting side effects at earlier stages in the drug development pipeline, based on the off-target interactions the drug is predicted to have [60, 89]. Perturbing metabolic or signaling pathways, these interactions may be part of a larger set of mechanisms that can explain the molecular basis of certain side effects, such as dosage effects, insufficient metabolization, aggregation, or irreversible binding of proteins [90]. Relating targets to drugs is also relevant for gaining more insight into the adverse effects of approved

compounds, an important aim falling under the heading of pharmacovigilance. Side effects might be missed in clinical trials, for example due to the rarity of occurrence or a delay between the start of medication and the onset of symptoms [60].

Certain protein targets have been identified as causally implicated for undesirable effects and especially hydrophobic compounds have the tendency to interact with some of these well-known promiscuous proteins, such as the human ether-a-go-go-related gene (hERG) potassium channel, pregnane X-receptor, cytochrome P450s, P-glycoprotein, and some Phase 2 metabolizing enzymes. These interactions can have a big impact on cardiotoxicity (hERG), as well as on the metabolism and distribution of many drugs [60, 89]. As drug-target interaction information can be used to link drugs with such effects, Lounkine and co-workers used the SEA method to predict the activity of 656 marketed drugs on seventy-three side effect-associated proteins. Among several newly identified drug-target associations, one of the most compelling ones might be the experimentally validated interaction between synthetic oestrogen chlorotrianisene and cyclooxygenase-1, potentially explaining its abdominal pain side effect [91]. In their web tool called DR. PRODIS, Zhou and co-workers combined drug target prediction using their FINDSITE^{comb} method [92] based on ligand homology modeling with drug side effect prediction based on target-side effect relations inferred from SIDER. They also introduced a killing index, estimating the likelihood that a drug will cause a serious side effect such as death, stroke, cancer, or heart failure, and showed correlation with an approved drug's probability of being withdrawn [93]. Unlike the former two examples, which are based on established connections between proteins and side effects, new connections for elucidating the physiological mechanisms responsible for adverse drug effects may be hypothesized as well. This approach is exemplified by the work of Xie, Chang, and co-workers on the off-target effects of torcetrapib, a cholesteryl ester transfer protein (CETP) inhibitor developed to prevent cardiovascular disease by raising high-density lipoprotein (HDL) cholesterol, but withdrawn from Phase 3 clinical trial because of presumed off-target effects causing deadly hypertension. An algorithm for identifying and comparing ligand binding sites on protein structures was employed to predict drug off-target binding sites across the proteome. Results were evaluated in the context of a model of renal metabolism, as it has been suggested that the hypertensive side effect may result from the renal regulation of blood pressure via metabolite reabsorption and secretion. Eventually six off-targets were predicted to alter renal function, with literature support connecting two of them to hypertension, providing a molecular explanation for the observed serious adverse drug reaction following torcetrapib treatment [94, 95].

1.3 Network analysis for drug discovery

With the advances in high-throughput technologies and bioinformatics methods, a plethora of chemical and biological interactions and associations has been predicted and modeled by means of networks. Protein-protein interaction networks, drug-drug similarity networks, and drug-target networks are all being employed for computational drug discovery nowadays, mainly for the prediction and characterization of drug targets and the elucidation of drug MoA [58].

1.3.1 Protein-protein interaction networks

Protein-protein interactions (PPIs) play a key role in cellular functions, such as metabolism, cell cycle control, and signal transduction [96]. Hence networks representing this complex system of gene products linked by physical interaction can provide insight into protein function and facilitate the modeling of functional pathways to elucidate the molecular mechanisms of cellular processes [97]. The complete human interactome has been estimated to comprise from approximately 130,000 up to 650,000 protein interactions [98, 99], with an average of five interaction partners per protein [100]. However, only a subset of these has been experimentally identified [101]. A number of techniques has been developed for studying PPIs. Two widely established experimental high-throughput methods are the yeast two-hybrid system and tandem affinity purification coupled to mass spectrometry [96]. In addition, phage display and protein microarrays are also applied for PPI identification. Computational methods in their turn use these PPI data obtained from experimental methods to predict a wider network of PPIs based on genomic sequence, protein sequence, protein structure, domain data, and semantic similarity [102]. Public databases storing human protein interaction data are numerous and diverse. The vast majority, including BioGRID (Biological General Repository for Interaction Datasets) [103], HPRD (Human Protein Reference Database) [104], IntAct [105], MINT (Molecular INTeraction) [106], and DIP (Database of Interacting Proteins) [107], mainly contain curated PPI information extracted from literature. IID (Integrated Interactions Database) [108], ConsensusPathDB [109], Mentha [110], PINA (Protein Interaction Network Analysis) [111], and iRefWeb [112] combine interactions from multiple databases. IID also includes orthologous PPIs and computationally predicted PPIs. Moreover, as both genetic and functional interaction data can provide information on protein function as well, they are often integrated with protein interaction data [96]. BioGRID, for instance, combines physical protein interactions with genetic interactions, while STRING [113] integrates direct, physical protein interactions with indirect,

functional protein associations derived from genomic context, co-expression, and phylogenetic co-occurrence.

Despite being incomplete and error-prone, analysis of these PPI networks using various computational approaches can shed light on the global proteome organization. Moreover, placing individual proteins into a functional context, PPI networks have proved particularly important for providing clues on protein function. For example, neighboring proteins in the network are generally considered to share functions. Topological features reflect the functionality of interacting proteins and thus can be used for discovering novel protein function and identifying functionally coherent modules [96, 102, 114]. In addition, PPI networks have become essential and powerful tools for associating proteins with distinct phenotypes, hence unravelling the molecular basis of diseases. Using linkage methods, module-based methods, and diffusion methods, network properties of disease genes have been studied, new genes and proteins associated with a certain disease identified, and novel relationships between these proteins uncovered [115, 116]. This knowledge can then be used to identify disease sub-networks for potential pharmacological targeting, supporting the development of drugs focused on highly specific disease pathways [102]. Similar to disease-causing proteins, drug targets have also been characterized in terms of PPI network attributes, such as degree and centrality. Significant differences between drug targets and non-drug targets suggest that these network-based properties might be useful in predicting new targets [117]. Finally, PPI networks can be used to monitor the effects of drug-induced perturbations at the level of a molecular network, facilitating the understanding of the various drug effects by identifying potential targets and mechanisms of action [96].

1.3.2 Drug-drug similarity networks

In a drug-drug similarity network, nodes represent drugs and edges connecting two drugs represent pairwise similarity based on chemical or structural properties, or phenotypic profiles. In this way, the network provides an immediate visualization of chemical space and can be well characterized and compared using the highly efficient algorithms that exist for analyzing the properties of large networks. The major motivation for constructing drug-drug networks, however, is to understand the relationship between drugs and exploit the degree of similarity between them to identify previously unknown targets, potential side effects, and opportunities for drug repositioning [118, 119, 120].

The most straightforward drug-drug network is based upon numerical values for structural similarity between drugs, for instance by computing the Tanimoto coefficient between fingerprint descriptors and setting a certain Tanimoto

threshold. As an alternative, substructure-based similarity, which considers two molecules similar if they share a defined substructure, or hybrid metrics can be used [120]. Comparing a structural similarity network of ZINC [121] compounds, representing general chemical space, with a network of bioactive compounds extracted from ChEMBL [74], Zwierzyna and co-workers showed that the properties of both networks systematically differ. The bioactive compounds network appeared to be more informative with consistently higher clustering coefficients, assortativity, and modularity, in this way distinguishing different activity classes [122]. Clustering according to therapeutic activity can also be observed when constructing a network based on similarity in drug-induced gene expression, as illustrated by the work of Iorio and co-workers. For each drug in CMap, they generated a consensus synthetic profile summarizing the transcriptional effect of the drug on different cell lines and at different dosages. Using gene set enrichment analysis (GSEA), pairwise similarities between these profiles were computed and a drug-drug network consisting of 1,302 nodes and 41,047 edges was constructed. By applying network theory, the network was divided into communities, which appeared to be significantly enriched for compounds with similar mechanisms of action or acting on the same pathway. New compounds can thus be integrated into the network to predict their therapeutic and off-target effects. This approach has already proved successful by the correct MoA prediction of nine anti-cancer drugs, as well as the experimentally validated discovery of the previously unknown autophagy activity for the rho-kinase inhibitor Fasudil, making it a viable candidate for repositioning to treat several neurodegenerative disorders. Finally, a web tool called MANTRA (Mode of Action by NeTwoRk Analysis) was made available to explore the constructed drug-drug network and query it for classification of previously undescribed compounds [123]. Another phenotype-based similarity network, developed by Campillos and co-workers, relates drugs to each other based on their known side effects. Applying this approach to 746 marketed drugs, a network of 1,018 side effect-driven drug-drug relations became apparent, 261 of which are formed by chemically dissimilar drugs from different therapeutic indications. *In vitro* validation of 20 implied drug-target interactions resulted in confirmed binding activity for 13, of which 9 were tested and confirmed in cell assays [72].

Taken together these results show that drug-drug networks based on different measures of drug similarity can be effective for gaining insight into drug MoA. Nevertheless, currently only one exhaustive public tool exists that combines chemical structure similarity with phenotypic information on a large scale to infer relationships between compounds, namely STITCH (Search Tool for Interacting Chemicals). Relationships between compounds are deducted from similar structures and similar activities, as well as literature associations and reactions in pathway databases. Activity similarity is based on inhibition

responses from the US National Cancer Institute (NCI) anti-cancer drug screens on sixty human tumour cell lines (NCI60), pharmacological actions assigned to chemicals in the Medical Subject Headings (MeSH), and gene expression profiles in CMap. The latest version integrates these disparate data sources for 430,000 compounds in a single, searchable network [124, 125, 126].

1.3.3 Drug-target networks

A drug-target network consists of both drug and protein nodes, which are connected to each other if the protein is a known target of the drug. The resulting bipartite graph has been shown to contain a giant connected component, suggesting the involvement of most drugs in targeting more than one protein [119, 127]. Yildirim and co-workers, for example, built a network composed out of 890 FDA-approved drugs linked to 394 protein targets, extracted from DrugBank. Most drugs in this network are connected into a highly interlinked giant component, with strong local clustering of the drugs according to the major Anatomical Therapeutic Chemical (ATC) classes, as illustrated in Figure 1.7 [128]. Likewise, Mestres and co-workers combined data from seven annotated chemical databases, including DrugBank, with ligand-based drug-target interaction predictions to assemble a drug-target network consisting of 4,767 interactions between 802 drugs and 480 targets. From this network, it could be observed that small hydrophobic drugs are significantly more promiscuous than large hydrophilic ones, and that G protein-coupled receptors (GPCRs) are significantly more connected than any other target family, reflecting how phylogenetic relationships can result into cross-pharmacology [62]. The topology of a drug-target network can also be exploited to predict new drug-target interactions. This can, for example, be done by training a binary classifier with topological features of known drug-target pairs, or by supervised bipartite graph inference [119], as was done by Yamanishi and co-workers through learning the correlation of drug structure similarity and target sequence similarity with drug-target interaction network topology [129]. Apart from being an important resource for drug-drug similarity, STITCH is also one of the most comprehensive public databases centered around drug-target interactions. By aggregating manually curated data sets, such as DrugBank with high-throughput experimental data and both structure and text mining predictions into one global network, many aspects of how drugs function in a cellular context can be captured [124, 126].

Starting from a drug-target network, two biologically relevant decompositions can be generated: a drug-drug network, linking two drugs if they share a certain number of targets, and a target-target network, comprising links between targets interacting with the same set of drugs [127]. Paolini and co-workers

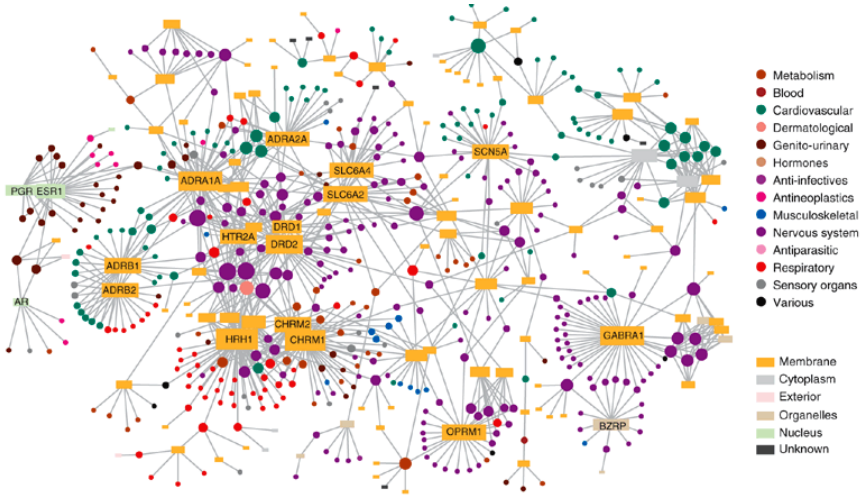


Figure 1.7: Giant component of the drug-target network of FDA-approved drugs constructed by Yildirim and co-workers. Circles represent drugs and are colored according to Anatomical Therapeutic Chemical (ATC) classification, while rectangles represent protein targets and are colored according to Gene Ontology (GO) cellular component (reprinted from [128]).

integrated structure-activity relationship (SAR) data from several sources to explore the global relationships between chemical structures and biological targets. Thereupon all interactions were mapped to a target-target network outlining polypharmacology relationships between targets, with aminergic GPCRs and protein kinases showing the highest promiscuity [37]. Network analysis tools to identify topological features in these drug-drug and target-target networks can also be applied to discern the global trends in drug discovery. Comparing the giant component size of the target-target network for both FDA-approved drugs and experimental drugs with a randomized network, Yildirim and co-workers showed the initial tendency of pharmaceutical industry to target already validated proteins, leading to an increase in the number of 'me-too' drugs, followed by a trend toward more functionally diverse targets improving polypharmacology. This phenomenon was also confirmed by the changes of the degree distribution and clustering coefficient after addition of the experimental drugs to the drug-drug and target-target network of FDA-approved drugs [128].

1.3.4 Propagating information through networks

As discussed above, both chemicals and proteins can be viewed in the context of a large interacting network to obtain a better understanding of the molecular mechanisms of disease, drug action, and associated adverse effects [90]. Over the last years, much work has been devoted to the development of network-based methods for predicting protein function and identifying unknown disease-associated genes. Combining network information with prior knowledge, these methods generally work by propagating information such as functional annotations across network edges [130, 131]. Growing evidence is now suggesting that these methods may also be successfully applied in drug discovery, such as for predicting new drug targets associated with disease genes [130]. Current approaches for gene prioritization mostly rely on the 'guilt by association' principle assuming that genes or proteins that are close to each other in a functional or interaction network tend to be involved in the same biological pathways and have a similar effect on phenotypes. Hence, the key question is how to measure the proximity of candidate genes to annotated genes within these interactome networks and different scoring strategies have been developed for this purpose [131].

One of the most straightforward approaches, termed direct neighbor counting, assesses whether a candidate gene has a direct interaction to a gene with a certain annotation. The more direct neighbors with this annotation the candidate gene has, the more likely the gene is to share this annotation [131, 132]. However, since two proteins can be involved in the same biological pathway without a physical interaction, much research has been focused on integrating additional local information to the prioritization by exploiting the network neighborhood. As a result many methods predicting functions based on PPI network structure are using the simple shortest path algorithm quantifying the closeness of proteins based on the shortest path between them [130, 131, 133]. Still neither of the two aforementioned local distance measures captures the important knowledge encoded in the surrounding network structure. Two nodes with many low-degree neighbors in common should, for instance, be considered more similar than nodes with few low-degree neighbors in common. Likewise, these nodes should be thought of as more similar than two nodes whose common neighbors have a high degree. Global distance measures such as random walk with restart, kernel diffusion, and network propagation can catch these differences by taking the overall network structure into account [131, 133]. A random walk on a network is defined as the transition of a random iterative walker through the network, which in each iteration extends its path from the current node to a randomly selected neighboring node with equal probability. Additionally, the walker can have a certain probability of terminating and restarting from its starting node [130, 134]. The random walk with restart can be used as a ranking

algorithm, where all nodes in the network are ranked by the probability of the walker reaching this node [135]. The diffusion kernel can be considered a lazy, continuous time random walk consisting of an infinite number of infinitesimally small transitions to one of each of the current node's neighbors with probability β, whereby the walker remains at the current node i with a probability of $1-d_i\beta$, with d_i being the degree of node i [134, 136]. Network propagation is similar to random walks in thought as well. Each node in the network is assigned a positive score when known to be related to a phenotype of interest and zero otherwise. Based on this prior knowledge, nodes can be ranked by means of an information flow through the network, where in each iteration information is propagated to neighboring nodes until a steady state is reached [130, 137].

1.4 Objective and outline of the thesis

The growing awareness that networks are key to understand diseases and find treatments, together with the increasing availability of genome-wide data, including protein-protein interaction, drug-target interaction, and gene expression data, opens new roads for predicting the effects of drugs on human biology. Although individual analysis of pharmacological data undoubtedly results into interesting findings, it is only by integrating them with the expert knowledge captured in networks that we can gain global insights into the cellular behavior of drug compounds. In this thesis, we therefore aim to develop novel methods that leverage knowledge of biological networks for a better understanding of the complex biological response to drug treatment.

The thesis text is organized as indicated in Figure 1.8. In the introductory Chapter 1 we describe the current state of drug discovery and development and expand on computational approaches for target prediction, as well as on the use of network analysis in drug discovery. In Chapter 2, we propose a new method tackling this target identification problem by analyzing gene expression following drug treatment in the context of a functional protein association network. To further facilitate the analysis of this particular data type and enable an in-depth exploration of a drug's mode of effect, we have developed the web server Galahad, which is presented in Chapter 3. Chapter 4 focusses on a further improvement of drug target prediction through the use of structural information combined with known protein-compound interactions. Finally, in Chapter 5, we reconsider our work in the light of modern computational pharmacology, focusing on the representation and integration of various types of pharmacological data and their application in drug target prediction. We conclude by highlighting the applicability or our work in a pharmaceutical research context and outlining some future perspectives.

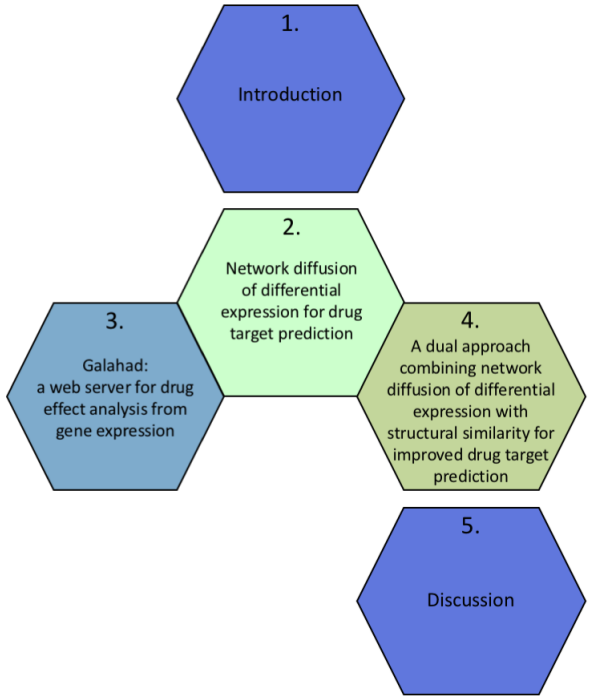


Figure 1.8: Structure of the thesis.

Chapter 2

Network diffusion of differential expression for drug target prediction

This chapter has been previously published as

Laenen G., Thorrez L., Börnigen D. & Moreau Y. (2013) **Finding the targets of a drug by integration of gene expression data with a protein interaction network.** *Molecular BioSystems* **9**, 1676-1685.

PhD candidate's contributions: Griet Laenen contributed to the development of the methodology, collected the test data, and performed all data analyses. She both drafted and revised the paper.

Abstract

Polypharmacology, which focuses on designing drugs that bind efficiently to multiple targets, has emerged as a new strategic trend in today's drug discovery research. Many successful drugs achieve their effects via multi-target interactions. However, these targets are largely unknown for both marketed drugs and drugs in development. A better knowledge of a drug's mode of action could be of substantial value to future drug development, in particular for side effect prediction and drug repositioning. We propose a network-based computational

method for drug target prediction, applicable on a genome-wide scale. Our approach relies on the analysis of gene expression following drug treatment in the context of a functional protein association network. By diffusing differential expression signals to neighboring or correlated nodes in the network, genes are prioritized as potential targets based on the transcriptional response of functionally related genes. Different diffusion strategies were evaluated on 235 publicly available gene expression data sets for treatment with bioactive molecules having a known target. AUC values of up to more than 90% demonstrate the effectiveness of our approach and indicate the predictive power of integrating experimental gene expression data with prior knowledge from protein association networks.

2.1 Introduction

For many years, the development of highly selective compounds interacting with a single target has been a pragmatic criterion in drug design. However, this reductionist approach is now being questioned, as its application has been concurrent with an increased rate of drugs failing in the late clinical development stages. Moreover, several lines of evidence suggest that many effective drugs act through the modulation of multiple targets and that this 'polypharmacology' can even be a therapeutic requirement [30]. While drug interaction with a set of targets seems to be beneficial for efficacy in many cases, there are also proteins one would rather avoid modulating [65], for unintended drug activity at off-targets is one of the leading causes of unexpected side effects [138]. Off-target toxicity accounts for about one third of the failures in clinical development [30] and even causes withdrawal of approved drugs. Fenfluramine for instance, was used in combination with phentermine as the popular appetite suppressant Fen-Phen, but was withdrawn from the market in 1997 because of causing valvular heart disease, later attributed to off-target activation of the serotonin receptor 2B (HTR2B) [139]. A better knowledge of drugs' off-target interactions could prevent such serious adverse drug reactions and allow safer drug candidates to be prioritized for preclinical development [91]. Although usually unwanted and harmful, occasionally off-target effects can be beneficial and lead to new therapeutic indications. For example mifepristone, a progesterone receptor antagonist originally developed as an abortifacient, was repositioned as a treatment for hyperglycemia in adults with Cushing's Syndrome on the basis of its interaction with the glucocorticoid receptor [140]. Repurposing administration-approved drugs cannot only speed up drug discovery and reduce costs through the elimination of toxicological and pharmacokinetic assessments [139], it also holds promise for rare and neglected diseases [141], as illustrated by mifepristone. Therefore, target prediction of approved drugs is

highly valued in drug discovery nowadays.

The experimental testing of drug-target interactions remains cost- and time-intensive, and thus there is a strong incentive to develop *in silico* methods addressing this issue [45]. Several computational approaches have been developed over the years [63] exploiting molecular features of chemicals and proteins [43, 142, 143, 144, 145, 146, 147, 148] as well as phenotypic effects of drug treatment [72, 78, 79, 80, 81, 82, 123, 148, 149, 150]. As gene expression data allow for a straightforward determination of which genes are active, this type of data is often being used to hint at a drug's mode of effect. Differential expression between drug-treated cells or tissues and non-treated controls pinpoints the main transcriptional effects of a drug. Treatment does, however, not necessarily result in a significant increase or decrease of the targets' mRNA level. Nevertheless we believe that genes functionally related to a target do tend to be up- or down-regulated after treatment, since they are often involved in the same pathways and may undergo regulatory effects from both the target and molecules functioning downstream. On the basis of this assumption we propose to predict a drug's targets by considering gene expression following treatment in the view of a functional protein association network. More specifically, we prioritize genes as potential targets based on the differential expression of functionally related genes in the surrounding network module. Nitsch and co-workers [151] have applied a similar approach to select the most promising genes from a list of candidate disease genes in human genetics. They evaluated three distinct random walk-based machine learning techniques that diffuse differential expression values over a protein interaction or functional association network in order to score and prioritize candidate genes based on their differentially expressed neighborhoods. The three strategies, being kernel ridge regression ranking, heat kernel diffusion ranking, and Arnoldi diffusion ranking, are all based on the exponential diffusion kernel, but differ in their exact methodologies. The best results were obtained by iterative heat kernel diffusion of RMA-preprocessed expression data over the functional STRING network [152]. In our study we extended this kernel diffusion methodology and assessed its performance on a genome-wide drug target prioritization problem. Additionally, we also introduce a new diffusion method, based on connectivity correlations in the protein network. Drug target prioritization by means of both diffusion strategies was evaluated on a large benchmark consisting of 235 publicly available gene expression data sets for treatment with bioactive molecules that have a known target. For these data AUC values of up to 92% could be obtained.

2.2 Materials and methods

2.2.1 Functional protein association network

A functional protein association network can be modeled as an undirected graph $G = (V, E)$, where V is the set of vertices or nodes and E is the set of edges. Two nodes, representing proteins, are connected by an undirected edge if there is a functional association between them. Each edge can be assigned a weight resembling the strength of evidence for the association.

Such a functional protein association network for human was obtained from STRING [152], a database of known as well as predicted protein-protein associations, including both direct (physical) and indirect (functional) interactions. STRING integrates prior knowledge from OMIM [153] and PubMed abstracts [154] with associations from interaction databases such as MINT [155], HPRD [156], BIND [157], DIP [107], BioGRID [158], IntAct [105], KEGG [159], Reactome [160], NCI/Nature PID [161], and Gene Ontology [162]. These known interactions are further complemented with newly predicted interactions based on gene fusion events, conserved genomic neighborhoods, phylogenetic co-occurrence, and co-expression. Lastly, interactions are also transferred between organisms by means of exhaustive cross-genome homology searches. As of version 9.0, these protein-protein similarity data are imported from the Similarity Matrix of Proteins (SIMAP) Project [163]. All transferred interactions as well as all predicted and imported associations are provided with a probabilistic confidence score, derived by benchmarking the associations against the functional grouping of proteins as annotated at KEGG. The associations are scored independently for the various major data sources, after which a combined score is computed indicating the higher confidence when more types of information support a given association.

We extracted a functional association network from two STRING releases: 8.3 and 9.0. The human protein network obtained from STRING 8.3 consists of 17,369 proteins and 1,288,886 interactions; the network extracted from STRING 9.0 consists of 18,600 proteins and 1,640,707 interactions. The protein Ensembl IDs from both these STRING networks were mapped to their encoding HGNC gene symbols.

2.2.2 Drug target prioritization strategies

Genome-wide knowledge on functional relations between proteins can help us achieve an understanding of the system as a whole. We evaluated two strategies integrating this information with gene expression data to establish a ranking of genes as potential drug targets and compared them to a simple expression ranking strategy, prioritizing genes based on expression data only. Both integrative methods assess the relevance of a gene as a drug target by considering the differential expression of its surrounding subnetwork of functionally related genes. The genes surrounding a drug target are expected to show an altered expression after drug administration in response to the perturbation of the target itself, while other genes, not functionally related to a target, are assumed not to be part of such a disrupted expression module. The subnetwork around each gene is selected either by means of a kernel-based random walk through the STRING network, or by means of computing connectivity correlations between node pairs.

Simple expression ranking

Ranking genes based on their own differential expression following treatment (Figure 2.1 A) can be a simple strategy to predict a drug's targets using only expression data. For each gene the expression level difference between the drug-treated and untreated samples is assessed in terms of $\log_2$ ratios or moderated t-statistics, after which the genes can be ranked accordingly. This simple expression ranking can be considered a baseline for performance assessment of the proposed network-based drug target prioritization strategies set out below.

Kernel diffusion ranking

For both the STRING 8.3 and STRING 9.0 network, represented as undirected graphs with symmetric weights wij between nodes i and j, the adjacency matrix A has entries $a_{ij} = w_{ij}$ if the nodes i and j are connected and $a_{ij} = 0$ otherwise. The combinatorial Laplacian matrix is given by $D - A$ where D is a diagonal matrix with $d_{ii} = \sum_j a_{ij}$. The asymmetric normalized version of the Laplacian is defined as $L_1 = I - D^{-1}A$, whereas the symmetric normalized version of the Laplacian, called the regularized Laplacian, is defined as $L_2 = I - D^{-1/2}AD^{-1/2}$. Both these normalized Laplacians are used in a drug target prioritization method that diffuses the differential expression values of all genes over the STRING network based on the confidence scores of the neighboring functional associations. This diffusion process can be formulated using a Laplacian exponential diffusion

kernel, defined by Kondor and Lafferty [164] as

$$K = e^{-\alpha L} = \lim_{n\to\infty}(I - \frac{\alpha}{n}L)^n,$$

which originates from the differential equation describing heat diffusion in material.
To obtain a genome-wide drug target prioritization, a scoring vector p_α is computed by multiplying the differential expression values p_0 with this heat kernel:

$$p_\alpha = p_0 e^{-\alpha L},$$

with L being one of the normalized Laplacian matrices. We have applied a discrete approximation of this heat kernel introduced by Yang and co-workers [165] and recently used by Nitsch and co-workers [151] to prioritize candidate disease genes:

$$p_\alpha = p_0(I - \frac{\alpha}{N}L)^N,$$

with N being the number of iterations and α the diffusion rate. By means of this iterative diffusion process the differential expression value for each gene is spread to all neighboring nodes reached in no more than N steps (Figure 2.1 B). All possible paths of length N or less are thereby taken into account, resulting in a faster diffusion between strongly interconnected nodes, which are assumed to have a more reliable functional relationship. This partial diffusion strategy can be considered comparable to the functional flow method proposed by Nabieva and co-workers [166] for protein function prediction. However, instead of applying a 'guilt by association' principle requiring prior knowledge, our method relies on experimental gene expression data. Furthermore, we introduce a diffusion parameter α controlling the diffusion rate. A value close to one causes rapid diffusion, giving a larger weight to genes located further away, while a value close to zero promotes a slower diffusion process, giving a larger weight to nearby genes. As a result of this diffusion each node gets a new score, no longer just representing its own differential expression, but also including information on the differential expression of its close neighborhood. Drug targets are thus assumed to get a high score, since these genes tend to be the central nodes in subnetworks showing significant transcriptional changes following treatment. In order to attenuate the systematic bias of the obtained scores toward highly connected genes, p-values indicating a gene's significance as potential drug target are computed by means of a permutation procedure. The differential expression values are randomized across the network a 1,000 times, after which a p-value is calculated for each gene as the fraction of permutations resulting in a score at least as extreme as the real score.

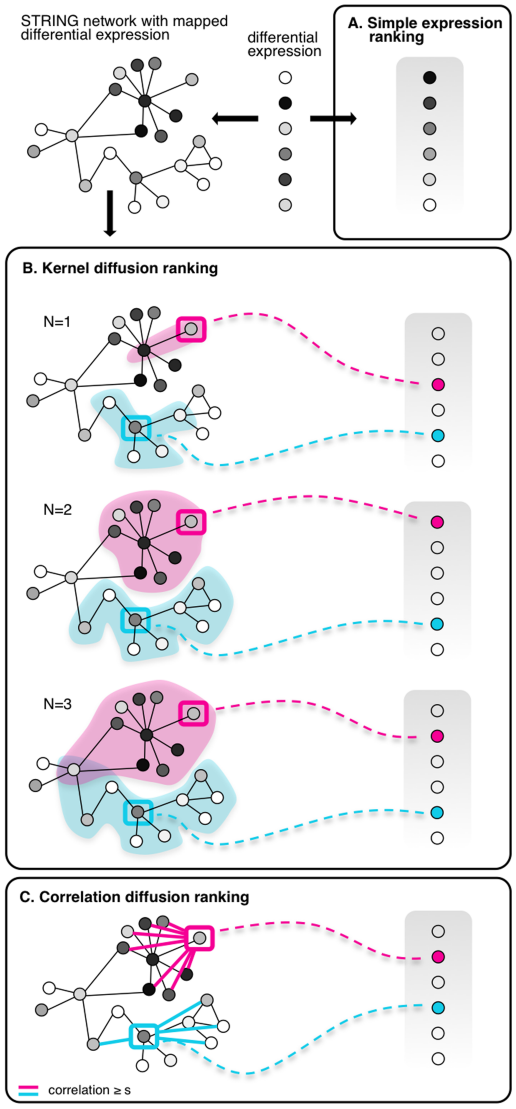


Figure 2.1: The drug target prioritization strategies. A treatment versus control differential expression level is determined for every gene. (A) Based on these a simple expression ranking is set up as a baseline. To integrate knowledge on functional associations between proteins, the differential expression values are mapped to the STRING network. (B) Using kernel diffusion, genes are ranked based on the differential expression of their closest neighbors in this network. The number of steps through the network determines the size of the neighborhood taken into account. (C) The correlation diffusion method ranks the genes based on the differential expression of genes with a strong connectivity correlation.

Correlation diffusion ranking

An alternative to the heat kernel approach, which diffuses differential expression values to neighboring nodes in the STRING network, is a correlation-based approach diffusing the differential expression values to genes/proteins considered functionally related based on the correlation of their weighted connectivity structure (Figure 2.1 C). To this end, a Pearson correlation coefficient ρ is computed for each pair of nodes:

$$\rho_{ij} = \frac{\mathrm{cov}(w_i, w_j)}{\sigma_{w_i}\sigma_{w_j}},$$

with w_i the vector containing weights for the interactions between node i and all other nodes. By filtering out all connectivity correlation coefficients below a certain threshold s and normalizing the coefficients based on the remaining correlations per gene, the elements of a normalized correlation matrix C can be defined as

$$c_{ij} = \begin{cases} \dfrac{\rho_{ij}}{\sum_j \rho_{ij}}, & \text{for all } \rho_{ij} \geq s \\ 0, & \text{otherwise.} \end{cases}.$$

Multiplication of the differential expression values with this matrix results in a ranking score for each gene:

$$p_\alpha = p_0 C.$$

The obtained scores are again corrected for the number of nodes taken into account by means of the randomization procedure described above.

2.2.3 Test data

Our drug target prioritization strategies were evaluated on 235 publicly available gene expression data sets where treatment with a bioactive molecule was tested against a control on one of three different Affymetrix Human GeneChips: HG U133A, HG U133 Plus 2.0, or HuGene 1.0 ST. The raw CEL files were downloaded from GEO and preprocessed with the RMA function of the R/Bioconductor affy [167] or oligo [168] package. After preprocessing, we computed three different measures of differential expression for each gene using the R/Bioconductor limma [169] package: the $\log_2$ ratio, the moderated t-statistic, and the significant $\log_2$ ratio. This latter statistic was obtained by setting the $\log_2$ ratio of all genes with a moderated t-statistic-derived p-value larger than 0.05 to zero.

The 235 data sets account for 130 unique treatment molecules that do not only include small molecules, but also antibodies, other proteins, and a nucleic acid. For each of these molecules we identified the primary target from literature and databases such as PubChem [170], ChEMBL [171], and BindingDB [172], after which we retrieved its position in the corresponding genome-wide ranking lists resulting from application of our prioritization methods. An overview of the data can be accessed through http://www.rsc.org/suppdata/mb/c3/c3mb25438k/c3mb25438k.pdf.

2.2.4 ROC analysis

The performance of the different drug target prioritization approaches was assessed in terms of the sensitivity and specificity with which the primary targets appear at the top of the ranking lists. In an ROC (Receiver Operating Characteristic) curve, the true positive rate (sensitivity) was plotted against the false positive rate (1 – specificity) by imposing all possible cut offs on the prioritization lists, each time dividing the genes into positive and negative ranked genes. True positives are the primary targets ranked above the cut off in the corresponding prioritization list, whereas false positives are all other genes ranked above. False negatives are the primary targets below the cut off and true negatives all other genes below.

As an evaluation measure, we computed the AUC (Area Under the ROC Curve) which is equivalent to the probability that a classifier ranks a randomly chosen positive instance higher than a randomly chosen negative instance, according to Hanley and McNeil [173]. In our case this means the AUC would be 100% if all primary targets ranked in first position, while it would be only 50% if the prioritization was no better than ranking the genes randomly.

2.3 Results and discussion

In this section, we evaluate the ability of our proposed prioritization strategies to predict a drug's targets by combining information on differential gene expression following treatment with knowledge on functional associations between proteins. The performance is assessed by means of a ROC analysis on genome-wide target prioritizations, obtained by these methods on 235 gene expression data sets for bioactive molecules having a known target. The corresponding AUC values are presented in Table 2.1. Apart from the prioritization strategies themselves, also the two integrated data sources, being the functional protein association

network and the gene expression data set, are investigated for their impact on the performance level.

2.3.1 Performance of the drug target prioritization methods

Simple expression ranking

Upon target binding, drug treatment can give rise to a plethora of complex cellular responses, including a positive or negative feedback loop adjusting the expression level of the respective target mRNA. Hence genes are ranked according to their own differential expression following treatment in order to establish a baseline for comparison with the suggested drug target prioritization strategies. Only genes present in the STRING network are taken up in these rankings, so as to make a direct comparison between the different prioritization methods possible. Table 2.1 shows that this simple strategy is not sufficient to predict a drug's targets: AUC values vary between 64% and 66%. This is consistent with the findings of Iskar and co-workers [174] concerning the quantification of drug-induced alterations of target mRNA levels. They extracted gene expression data for drug treatment from CMap [78, 79] for 466 drugs; only 8% of the targets appeared to be subject to regulations at the mRNA level. This low percentage may, however, in part be attributed to the relative short time of six hours between the addition of the drug and gene expression analysis in the CMap approach, during which there is not sufficient time for effects via feedback loops. In the 235 data sets used for our analysis, time between drug addition and gene expression analysis is more variable and mostly longer. Still, the simple expression ranking results indicate that drug treatment mostly leads to relatively small differences in mRNA expression of the direct targets, hence the need to consider the neighborhood in order to find the drug target based on gene expression data.

Kernel diffusion ranking

The kernel diffusion strategy does not restrict itself to the actual differential expression of genes to set up their ranking as potential target. That is to say, it also takes the differential expression of close neighbors in the STRING network into account. To what extent the contribution of these neighbors to the ranking score decreases with an increasing distance in the network, is determined by the diffusion rate α (see Materials and methods). Table 2.2 shows the performance of kernel diffusion on the test data using different values for this diffusion parameter. Results show a decrease in performance with the decrease of α. Since drug targets are not always differentially expressed following treatment,

as shown by the simple expression ranking results, the contribution of the gene for which a score is computed has to be kept small in comparison with the contribution of its neighboring nodes. This demands a rapid diffusion and so a higher diffusion parameter of 0.9 is required.

Besides the degree of diffusion, the edge weights between neighboring nodes also play an important role in the contribution of each gene. Weights are extracted from the STRING network, after which they are normalized by computing one of the Laplacians discussed in the Materials and methods section. The asymmetric normalized Laplacian L_1 assures that the weights of all edges leaving a node, including the self loop, sum up to one. In this way the weight used for computing the contribution of a node j to the score of a neighboring node i is inversely proportional to the node degree of node j, defined as the sum of the weights of all adjacent nodes. Thus a lower confidence is given to highly connected nodes and edges get different weights for both directions. Using the symmetric normalized Laplacian L_2 edge weights are considered the same in both directions. The contribution of a node j to the score of a neighboring node i is now divided by the square root of the degree of node j multiplied by the degree of node i.

Table 2.1 depicts the performance level of the kernel diffusion strategy for both Laplacians. The performance has improved significantly in comparison to simple expression ranking, with AUC values ranging from 76% up to 91%. Results are comparable for both Laplacians when taking two or three steps in the network, but the symmetric normalized Laplacian clearly outperforms the asymmetric normalized Laplacian when limiting the scores to the differential expression values of direct neighbors (N=1). The higher AUC values for such a one-step diffusion based on the symmetric normalized Laplacian may reflect some typical properties of the STRING network. First of all STRING is not limited to known, physical interactions, but also includes predicted, functional relations. This may substantially reduce the reliability of indirect links between second and third level neighbors. Moreover, STRING is a scale-free network, with most proteins participating in only a few associations and a few hubs interacting with thousands of proteins. These hubs make the network suffer from the small world effect, that is that most nodes can be reached from every other node through a few steps only. Most likely too many nodes are reached in two or three steps, causing dilution of the differential expression signals of direct neighbors. In addition, network hubs tend to be essential proteins regulated through complex mechanisms, making them less suitable to capture a transcriptional response focused around a single gene. Thus the confidence of such highly connected nodes needs to be corrected accordingly, as is done by normalization of the

Laplacian. Results suggest downweighting by the square root of the node degree using the symmetric normalized Laplacian performs best.

Correlation diffusion ranking

Table 2.3 illustrates the results obtained by applying correlation diffusion ranking with different correlation thresholds s (see Materials and methods). The AUC values clearly show that a lower threshold of 0.1 is required to obtain adequate rankings, presumably because of the reduced correlation matrix sparsity.

Although the kernel diffusion method achieved the best performance when taking into account direct neighbors only, results obtained for the correlation diffusion method with a correlation threshold of 0.1 indicate that drug target prioritization can benefit from the inclusion of differential expression measures for nodes lying further away in the network. Indeed, Table 2.1 shows the best AUC value (91.9%) was derived by the application of correlation diffusion, which includes the differential expression of all nodes with a correlated connectivity structure for computing ranking scores. Nodes with a high degree of correlation can be interpreted here as nodes sharing a large part of their direct neighbors. The good performance when diffusing the expression signals to such correlated nodes could be explained by proteins associated with the same set of proteins having greater likelihood of sharing functional characteristics. For example, Chua and co-workers [175] observed that a substantial number of proteins in the *Saccharomyces cerevisiae* GRID interaction network share functions with second level but not first level neighbors. This may explain the higher reliability of functional associations based on connectivity correlations in comparison to associations derived by taking short paths through the network, as demonstrated by the higher AUC values for correlation diffusion ranking in comparison with two- or three-step kernel diffusion.

2.3.2 Impact of the network

An important bottleneck for the performance of both diffusion methods is the quality and coverage of the functional network around the actual drug targets. We have chosen to extract our network from the STRING database, since it combines multiple heterogeneous data sources, consequently mitigating sparsity and boosting reliability. Furthermore, the STRING network has already proven to be effective for similar prioritization problems as described by Nitsch and co-workers [151]. Our drug target prioritization methods were evaluated on the current STRING 9.0 network as well as on its previous release STRING 8.3, in

Table 2.1: AUC values for the different drug target prioritization strategies.

				STRING 8.3	STRING 9.0	randomized network
Simple expression ranking			$\log_2$ ratio	0.659	0.657	0.659
			t-statistic	0.646	0.644	0.646
Kernel diffusion ranking	L_1	$N = 1$	$\log_2$ ratio	0.835	0.785	0.646
			t-statistic	0.828	0.767	0.642
			sign. $\log_2$ ratio	0.821	0.780	0.626
		$N = 2$	$\log_2$ ratio	0.829	0.801	0.663
			t-statistic	0.822	0.792	0.651
			sign. $\log_2$ ratio	0.830	0.805	0.644
		$N = 3$	$\log_2$ ratio	0.820	0.797	0.662
			t-statistic	0.812	0.784	0.651
			sign. $\log_2$ ratio	0.825	0.801	0.644
	L_2	$N = 1$	$\log_2$ ratio	0.909	0.903	0.646
			t-statistic	0.911	0.901	0.641
			sign. $\log_2$ ratio	0.887	0.876	0.625
		$N = 2$	$\log_2$ ratio	0.828	0.822	0.662
			t-statistic	0.814	0.806	0.650
			sign. $\log_2$ ratio	0.846	0.840	0.642
		$N = 3$	$\log_2$ ratio	0.812	0.805	0.663
			t-statistic	0.795	0.792	0.651
			sign. $\log_2$ ratio	0.840	0.834	0.644
Correlation diffusion ranking			$\log_2$ ratio	0.919	0.893	0.657
			t-statistic	0.919	0.891	0.608
			sign. $\log_2$ ratio	0.918	0.896	0.645

Table 2.2: AUC values for kernel diffusion ranking using different diffusion rates α.

				STRING 8.3	STRING 9.0
$\alpha = 0.9$	L_1	$N = 1$	$\log_2$ ratio	0.835	0.785
			t-statistic	0.828	0.767
			sign. $\log_2$ ratio	0.821	0.780
		$N = 2$	$\log_2$ ratio	0.829	0.801
			t-statistic	0.822	0.792
			sign. $\log_2$ ratio	0.830	0.805
		$N = 3$	$\log_2$ ratio	0.820	0.797
			t-statistic	0.812	0.784
			sign. $\log_2$ ratio	0.825	0.801
	L_2	$N = 1$	$\log_2$ ratio	0.909	0.903
			t-statistic	0.911	0.901
			sign. $\log_2$ ratio	0.887	0.876
		$N = 2$	$\log_2$ ratio	0.828	0.822
			t-statistic	0.814	0.806
			sign. $\log_2$ ratio	0.846	0.840
		$N = 3$	$\log_2$ ratio	0.812	0.805
			t-statistic	0.795	0.792
			sign. $\log_2$ ratio	0.840	0.834
$\alpha = 0.7$	L_1	$N = 1$	$\log_2$ ratio	0.825	0.788
			t-statistic	0.817	0.778
			sign. $\log_2$ ratio	0.813	0.782
		$N = 2$	$\log_2$ ratio	0.805	0.782
			t-statistic	0.796	0.770
			sign. $\log_2$ ratio	0.816	0.792
		$N = 3$	$\log_2$ ratio	0.800	0.778
			t-statistic	0.789	0.764
			sign. $\log_2$ ratio	0.814	0.794
	L_2	$N = 1$	$\log_2$ ratio	0.835	0.826
			t-statistic	0.822	0.816
			sign. $\log_2$ ratio	0.841	0.834
		$N = 2$	$\log_2$ ratio	0.790	0.783
			t-statistic	0.772	0.767
			sign. $\log_2$ ratio	0.826	0.823
		$N = 3$	$\log_2$ ratio	0.780	0.772
			t-statistic	0.760	0.756
			sign. $\log_2$ ratio	0.824	0.819
$\alpha = 0.5$	L_1	$N = 1$	$\log_2$ ratio	0.792	0.768
			t-statistic	0.782	0.754
			sign. $\log_2$ ratio	0.794	0.772
		$N = 2$	$\log_2$ ratio	0.775	0.760
			t-statistic	0.763	0.742
			sign. $\log_2$ ratio	0.800	0.781
		$N = 3$	$\log_2$ ratio	0.771	0.755
			t-statistic	0.757	0.739
			sign. $\log_2$ ratio	0.799	0.781
	L_2	$N = 1$	$\log_2$ ratio	0.769	0.764
			t-statistic	0.752	0.746
			sign. $\log_2$ ratio	0.811	0.806
		$N = 2$	$\log_2$ ratio	0.751	0.745
			t-statistic	0.729	0.727
			sign. $\log_2$ ratio	0.814	0.806
		$N = 3$	$\log_2$ ratio	0.744	0.739
			t-statistic	0.725	0.723
			sign. $\log_2$ ratio	0.812	0.808

Table 2.3: AUC values for correlation diffusion ranking using different correlation thresholds s.

		STRING 8.3	STRING 9.0
$s = 0.1$	$\log_2$ ratio	0.919	0.893
	t-statistic	0.919	0.891
	sign. $\log_2$ ratio	0.918	0.896
$s = 0.3$	$\log_2$ ratio	0.842	0.861
	t-statistic	0.868	0.860
	sign. $\log_2$ ratio	0.772	0.795
$s = 0.5$	$\log_2$ ratio	0.730	0.729
	t-statistic	0.729	0.732
	sign. $\log_2$ ratio	0.611	0.617

order to investigate the effect of small network differences on the effectiveness of the proposed methods. Table 2.1 shows a direct comparison between both STRING releases and also includes results obtained with a randomly permuted STRING network. The performance of this last network is not completely random; AUC values compare to those obtained by simple expression ranking. This could be explained by the fact that the edges are permuted but that the initial expression value on each node remains the same. Therefore, the informative contribution of the expression level of a gene itself remains preserved during prioritization, such that the score of each gene is determined by its own differential expression plus the noise caused by the differential expression of its randomly assigned neighbors. When comparing the two STRING releases, the updated version 9.0 results in slightly worse AUC values regardless of the applied method. This observation indicates that although coverage tends to be an important property, it may yield a larger portion of false positive associations, lowering the contribution of the truly associated genes. Figure 2.2 illustrates ROC curves for both versions of the STRING network, derived from correlation diffusion rankings and symmetric as well as asymmetric normalized Laplacian kernel diffusion rankings based on $\log_2$ ratios. For all three methods compared, AUC was 1-5% better when the STRING 8.3 version was used.

Last, we considered the dependency of the results on different protein association evidence types integrated in the STRING network: co-expression, association in a curated database, experimental data, co-occurrence across genomes, co-mentioning in PubMed abstracts, occurrence in close neighborhood across genomes, and gene fusion. The two best performing ranking methods, one-step symmetric normalized Laplacian kernel diffusion and correlation diffusion,

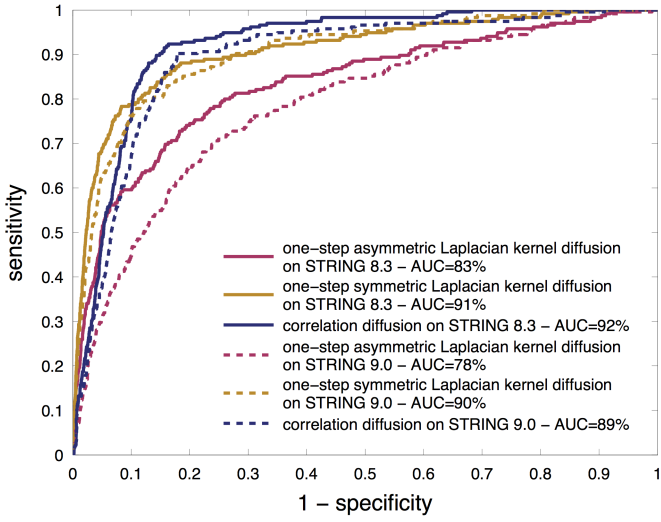


Figure 2.2: ROC curves comparing the performance of the two STRING versions. STRING 8.3 (solid line) performs slightly better than STRING 9.0 (dotted line) regardless of which prioritization method is applied to the $\log_2$ ratio statistics.

were applied to seven variants of the STRING 8.3 network, each comprising a single evidence type. The obtained results are depicted in Table 2.4. As the AUC values for each type of evidence are only based on those data sets for which the drug target is present in the relevant network variant, not all results are equally representative. Nevertheless, text mining clearly outperforms all other kinds of functional associations. This is partially due to the larger number of genes and interactions included in the text mining network compared to the often small or sparse networks representing the other evidence types. The gene fusion network consists for example of only 3,300 genes connected through only 4,017 interactions of which most have a very low weight. As a consequence only a minimal diffusion can take place, resulting in AUC values comparable to those obtained by simple expression ranking. Integrating the different association types resolves such sparsity issues and provides us with a more complete and hence robust network for our diffusion approach. Moreover, the higher AUC values for text mining can also be partially attributed to a performance assessment based on the prediction of drug targets of which many are well studied disease genes abundant in the scientific literature. Therefore, integration of all other evidence types is vital for our method to also perform well in a more prospective research setting.

Table 2.4: AUC values for the different protein association evidence types in STRING resulting from one-step symmetric normalized Laplacian kernel diffusion and correlation diffusion of $\log_2$ ratios.

Protein association evidence type	Number of data sets tested	AUC for kernel diffusion ranking	AUC for correlation diffusion ranking
all	235	0.91	0.92
co-expression	30	0.66	0.70
database	230	0.77	0.77
experimental	235	0.74	0.79
co-occurrence	37	0.69	0.73
text mining	235	0.91	0.92
neighborhood	13	0.82	0.71
fusion	111	0.61	0.67

2.3.3 Impact of the expression data

Just like the functional network, the gene expression data, measured before and after treatment with a drug of interest, also have an important impact on the performance of the target prioritization methods. The quality of the measurements and the experimental set up (*e.g.* relevant cell type) can strongly influence the results. Hence divergent results were sometimes observed for the same drug tested in different experiments. Moreover, differential expression values can be passed to the methods in various manners. We compared rankings based on $\log_2$ ratios (for all genes as well as for the significant genes only) and moderated t-statistics. Table 2.1 shows that the performance differences between these expression measures are almost negligible. We also examined the influence of the type of experiment on the performance level, by splitting up the test data in *in vivo* and *in vitro* experiments. In the 27 *in vivo* experiments a patient group was treated with the drug of interest, in the 202 *in vitro* experiments the drug was administered to a cell line or biopsy, the 6 data sets derived from xenografts were not included in the analysis. Figure 2.3 A shows the ROC curves for a single-step symmetric normalized Laplacian kernel diffusion applied to the $\log_2$ ratios of both data subsets. With a difference of only 2% between both AUC values, the performance of our method seems quite robust for the type of experiment.

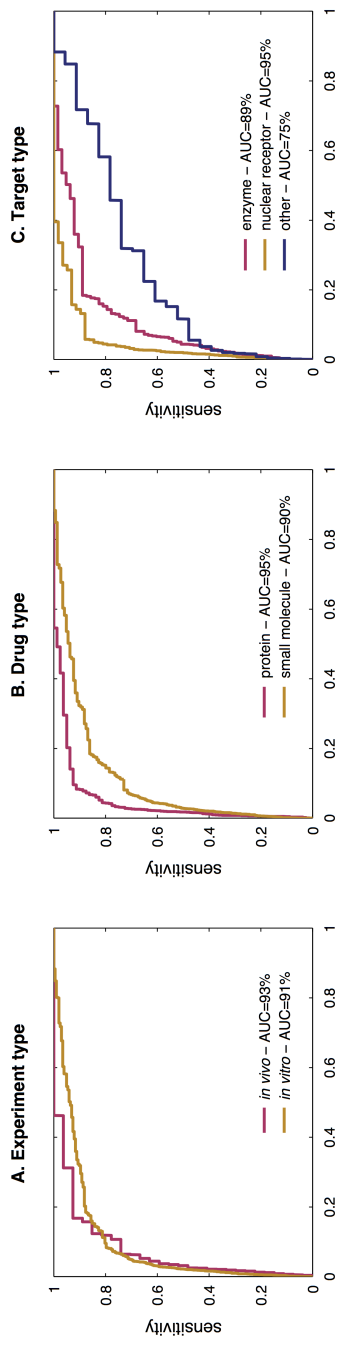


Figure 2.3: ROC curves illustrating the effect of the experiment, drug, and target type on the performance. When applying a one-step symmetric normalized Laplacian kernel diffusion of $\log_2$ ratios, performance appears to be slightly better for data coming from *in vivo* experiments compared to *in vitro* experiments, for protein drugs compared to small molecule drugs and for the prediction of nuclear receptor targets compared to enzyme targets.

2.3.4 Comparison of the performance on different drug and target types

Together with the network characteristics and the quality of the expression data, also the prioritization problem in itself can affect the performance level of the diffusion methods. Specifically, the methods may not be equally effective for each type of drug and each type of target. Since most of our test data sets originate from treatment with either a protein (eighty, antibodies not included) or a small molecule (144), a simple partition of the two can elucidate whether the diffusion of differential expression measures over a functional protein network is more effective in predicting protein targets or small molecule targets. Figure 2.3 B shows that, with an AUC value of 95%, the one-step symmetric normalized Laplacian kernel diffusion performs slightly better for protein drugs, as compared to an AUC value of 90% for the small molecules. This may be explained by the fact that these proteins often are natural ligands of their targets, exhibiting both high specificity and high affinity.

Our prioritization methods are based on the assumption that drug treatment induces differential expression of genes that are functionally related to the actual drug target. In a last analysis we examine for which type of targets this assumption is most correct, by partitioning our small molecule test data sets according to target class. We distinguish between enzymes (sixty three), nuclear receptors (fifty eight), and others (twenty three), the latter being a heterogeneous group consisting of, among others, membrane receptors, structural proteins, and secreted proteins. Figure 2.3 C shows that the performance of our one-step symmetric normalized Laplacian kernel diffusion is significantly lower for this last class in comparison to the other two classes. This may be because interference with the function of some of these proteins does not result in a direct, focused transcriptional change but in a delayed, more diffuse signal that is not picked up by our method. The nuclear receptors perform best with an AUC value of 95% and the enzyme class obtains an in-between AUC of 89%. As nuclear receptors are transcription factors causing gene expression changes directly downstream, it is not surprising that our method is most effective for this target type. By normalizing scores based on protein classification we might be able to correct for this target type bias.

2.4 Conclusion

To explain and prevent serious adverse drug reactions, to allow for safer drug candidates to be prioritized for preclinical development, and to explore the possibilities for drug repositioning, a better knowledge of drug-target interactions

is vital. Several computational approaches tackling this problem have been described. We propose the use of gene expression data following treatment, despite the fact that direct targets are not necessarily retrieved by a standard differential expression analysis. Our approach considers the expression data in the context of a molecular network, thereby taking regulatory interactions between the proteins into account. More specifically, differential expression values following treatment are diffused over a functional subnetwork around each protein, resulting in a new statistic based on which a genome-wide ranking of potential targets can be established. We have explored two strategies for diffusion: a kernel-based random walk, employing a symmetric or asymmetric normalized Laplacian, and diffusion to genes with a strong connectivity structure correlation. Both methods were evaluated on 235 publicly available gene expression data sets for bioactive molecules with a known target. With AUC values of 91% and 92% respectively, the best results were obtained by single-step symmetric normalized Laplacian kernel diffusion and by correlation diffusion, both over the STRING 8.3 network. Results show no significant performance differences for the type of experiment from which gene expression data were derived, but performance does seem to depend on the type of drug, with protein drugs performing slightly better than small molecules. Our results also suggest performance differences between different target classes, depending on how concentrated the drug-induced transcriptional response downstream of the target protein is. Still we believe that both proposed diffusion strategies show promise for future drug discovery research, in particular for drug target prioritization on a genome-wide scale.

Postscript

Since the publication of this chapter, several computational approaches fusing heterogeneous information through network diffusion toward drug-target interaction prediction have been introduced. First of all, our approach was picked up by Isik and co-workers, who evaluated it against some state-of-the-art network measures and further improved on it by means of a new network measure called local radiality, which describes the reachability of a target protein via the shortest path to deregulated proteins [176]. Seal and co-workers integrated existing drug-target interactions with drug-drug similarity and target-target similarity into one coherent framework. The random walk with restart network diffusion algorithm was applied, with a drug of interest and its known targets as seed nodes, to predict novel targets for the drug by capturing indirect paths to potential targets [177]. Wang and co-workers constructed a network consisting of three different types of nodes, diseases, drugs, and drug targets, with disease-drug and drug-target relationships constructed based on prior

knowledge and disease-disease, drug-drug, and target-target relationships based on their similarities. Again, an iterative updating algorithm that propagates information across the network was applied to solve the missing edge prediction problem [178]. Finally, Luo and co-workers integrated heterogeneous data for drugs, proteins, diseases, and side effects by combining the random walk with restart with a dimensionality reduction scheme. They subsequently used the resulting low-dimensional but informative vector representations of both drugs and proteins to identify the best projection from drug space onto protein space, based on which new drug-target interactions were predicted [179].

Chapter 3

Galahad: a web server for drug effect analysis from gene expression

Part of this chapter has been previously published as

Laenen G.*, Ardeshirdavani A.*, Moreau Y. & Thorrez L. (2015) **Galahad: a web server for drug effect analysis from gene expression.** *Nucleic Acid Research* **43(W1)**, W208-W212.

**Equally contributing first authors: Griet Laenen developed the pipeline, implemented all analyses, documented the software, and created examples. She took the lead in writing the paper.*

Abstract

Galahad (https://galahad.esat.kuleuven.be) is a web-based application for analysis of drug effects. It provides an intuitive interface to be used by anybody interested in leveraging microarray data to gain insights into the pharmacological effects of a drug – mainly identification of candidate targets, elucidation of mode of action, and understanding of off-target effects. The core of Galahad is a network-based analysis method of gene expression. As an input, Galahad takes raw Affymetrix microarray data from treatment versus control experiments for

human, mouse, or rat and provides quality control and data exploration tools, as well as computation of differential expression. Alternatively, differential expression values can be uploaded directly. Using these differential expression values, drug target prioritization and both pathway and disease enrichment can be calculated and visualized. Drug target prioritization is based on the integration of the gene expression data with a functional protein association network. The web site is free and open to all and there is no login requirement.

3.1 Introduction

DNA microarray technology enables genome-wide analysis of the transcriptional response to a compound treatment, which is frequently employed to study the effects of small molecules on cells [80]. Gene expression patterns that describe the perturbation of a biological system by a drug compound can provide valuable information for identifying compound-protein interactions and resulting effects [174] prior to clinical trials. In addition, this information may also be useful for already marketed drugs, in the light of drug repositioning [54]. This approach has proven to be successful as demonstrated by the many applications of the Connectivity Map [78, 79, 180, 181].

We have developed an easy-to-use web server called Galahad, for the in-depth exploration of a drug's mode of effect based on gene expression changes following treatment. Our software provides multiple tools needed for gaining new insights into the biological effects of a drug by combining Affymetrix gene expression data preprocessing, quality assessment and exploratory analysis, genome-wide drug target prioritization, differential expression analysis, and pathway, as well as disease phenotype enrichment. Drug target prioritization relies on the integration of the obtained differential expression values with prior knowledge on functional protein associations. By means of a network neighborhood analysis the functional relation between proteins is taken into account, which significantly improves the expression-based prediction of drug-protein interactions.

3.2 Input

The submission page (Figure 3.1) is easy to find and help is available by mouse-over where input is required from the user. The main input for *Galahad* are raw DNA microarray data derived from both untreated control samples and samples treated with a drug of interest, with one of the the following common

Affymetrix GeneChips: HG-U133A, HG-U133 Plus 2.0, or HuGene 1.0 ST for human, Mouse 430 2.0 or MoGene 1.0 ST for mouse, and Rat 230 2.0 or RaGene 1.0 ST for rat. Both the control and the treatment data need to be uploaded as a zipped file containing CEL files for at least two samples. The total size of both these zip files is currently limited to 250 MB. Once the data are uploaded completely, Galahad can be launched by clicking the submit button. A typical workflow then starts with the preprocessing of these uploaded expression data, either with the RMA [182] or the MAS 5.0 algorithm, as indicated by the user. Next, the desired analyses can be selected; these may include quality assessment and exploratory analysis of the data, differential expression analysis, drug target prioritization, and both pathway and disease enrichment. Drug target prioritization and enrichment analysis can also be performed on differential expression data derived from other platforms through the submission of a plain text file containing HGNC gene symbols, $\log_2$ ratios, and corresponding p-values. The input interface provides the possibility to enter a study name, as well as an email address to get notified when all computations are finished.

3.3 Methods

An overview of the different analysis steps is depicted in Figure 3.2.

3.3.1 Quality control

To assess the quality of the different arrays within the microarray data set, uncover experimental problems, and help in deciding whether certain arrays need to be considered as outlying, several quality metrics can be generated. This quality control is mainly based on the R/Bioconductor ArrayTools package, but also relies on functionalities from the affy [167], oligo [168], oligoClasses, genefilter, simpleaffy [183], affyPLM [184], and RColorBrewer libraries. The following plots are produced: side-by-side boxplots of the $\log_2$ intensities to assess the overall signal comparability, probe level model-based pseudo-images to detect spatial biases, and a NUSE and RLE plot to check for probe set homogeneity. In addition, for HG-U133A, HG-U133 Plus 2.0, Mouse 430 2.0, and Rat 230 2.0 arrays, the average background, the percentage of genes called present, and the intensities of the spike-in hybridization controls give insight into the hybridization quality. Scale factors provide information on signal comparability and 3'/5' ratios for β-actin and GAPDH can be used to assess sample quality. Finally, an RNA degradation plot is generated for these array types.

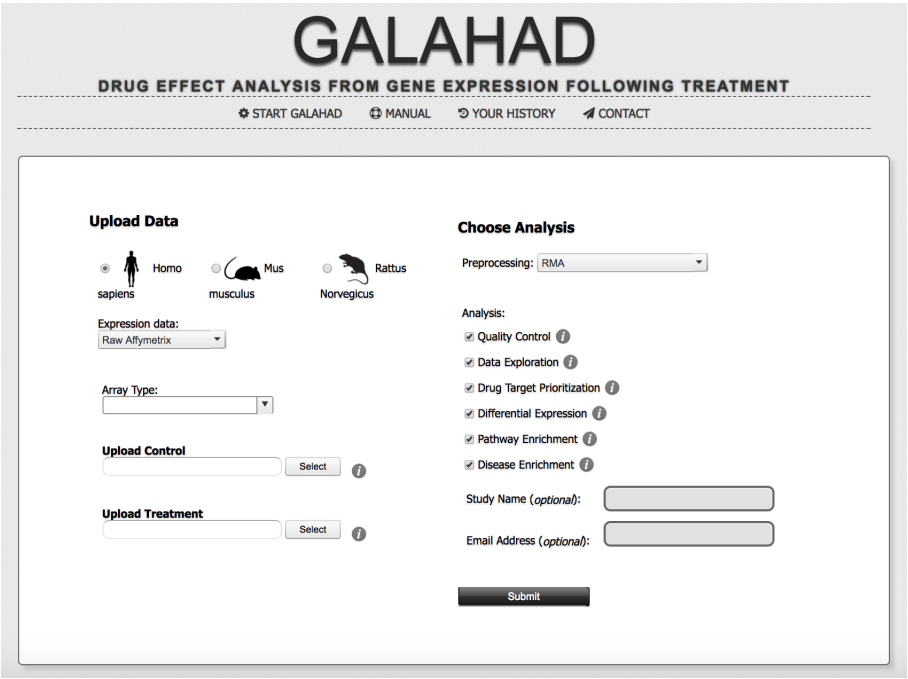


Figure 3.1: The Galahad submission page. The right organism and data type need to be chosen before uploading the data in zip format. The desired preprocessing method and analyses can be selected as well. Entering a study name and/or email address is optional.

3.3.2 Data exploration

To assess the correlation and grouping of samples three more plots are generated on the preprocessed data: a PCA plot, a hierarchical clustering, and an intensity correlation heat map.

3.3.3 Differential expression

Differential expression analysis to determine the significance of gene up- and downregulation following drug treatment is based on the R/Bioconductor limma [185] package. A linear model is fitted to the expression data for each gene and by computing a global variance estimator using empirical Bayes moderated t-statistics and corresponding p-values are obtained. These p-values are then corrected for multiple testing by the method of Benjamini and Hochberg [186], by controlling the false discovery rate (FDR).

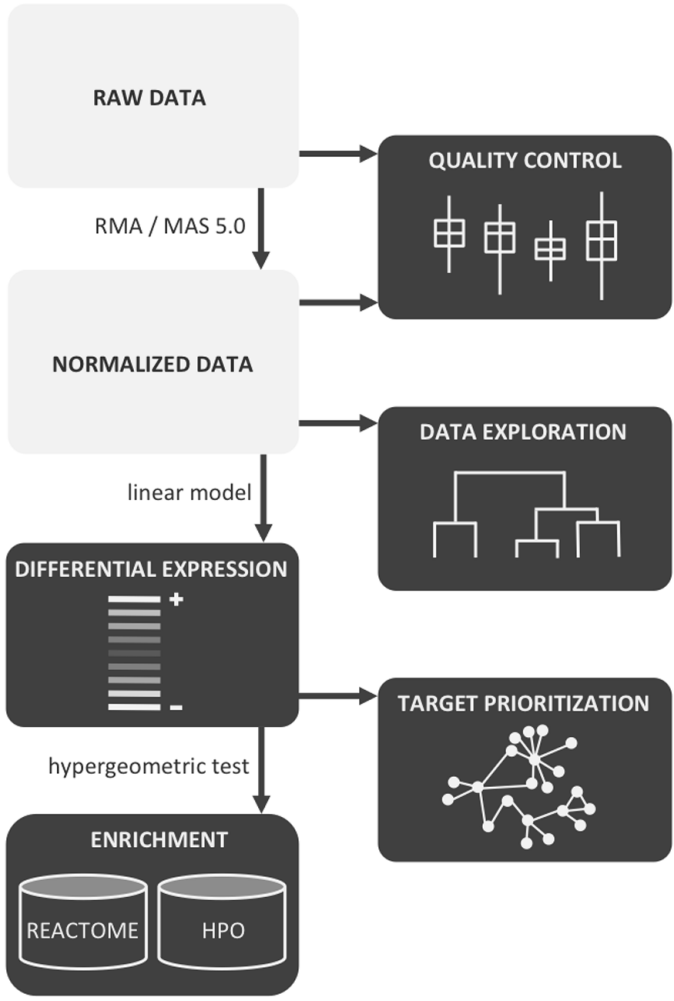


Figure 3.2: Overview of the different Galahad analysis steps. Quality control is performed on both the raw and MAS 5.0 normalized data. After normalization some further exploratory analysis is performed and differential expression is computed. The obtained differential expression values are then used for both pathway and disease phenotype enrichment as well as for network-based drug target prioritization.

3.3.4 Drug target prioritization

Genome-wide drug target prioritization is performed by means of an in-house developed algorithm for network neighborhood analysis [187] integrating the expression data with functional protein association information derived from STRING [152] (see Chapter 2 for details). More specifically, this approach prioritizes potential targets by diffusing the differential expression signals to neighboring network nodes based on the confidence scores of these functional associations. *P*-values are obtained by random reassignment of the expression values to the network nodes and computation of the corresponding randomized scores for all genes. The method has now been validated on a new test set consisting of thirty five publicly available Affymetrix human gene expression data sets where treatment with a chemical compound was tested against a control. In ChEMBL [74], we identified all proteins interacting with an IC_{50} below 10 μM with one of these compounds, resulting in a total of eighty five distinct targets. A table with the thirty five GEO accession numbers, together with the drugs of treatment and their targets as derived from ChEMBL can be accessed through https://academic.oup.com/nar/article/43/W1/W208/2467909#supplementary-data. Based on the positions of these known targets in the protein rankings obtained with our method for each of the interacting compounds, an AUC of 87% was achieved (Figure 3.3). In comparison, ranking the genes based on differential expression data only resulted in an AUC of 60%.

3.3.5 Enrichment

Functional enrichment of differentially expressed genes is performed by means of a hypergeometric test, followed by an FDR-based multiple testing correction of the obtained *p*-values. In this way, Reactome [188] (human) or KEGG [189] (mouse and rat) pathways involved in the drug's mode of effect are predicted, as well as associated disease phenotypes from the Human Phenotype Ontology (HPO) [190] (human), Mammalian Phenotype Ontology (MPO) [191] (mouse), or Rat Genome Database (RGD) Disease Ontology (RDO) [192] (rat) enabling side effect prediction and drug repositioning.

3.4 Output

After submission the user is automatically redirected to the results page, where the job details are provided. When the job is completed, the results appear on this page, and an email containing a hyperlink to the page is sent to the user if

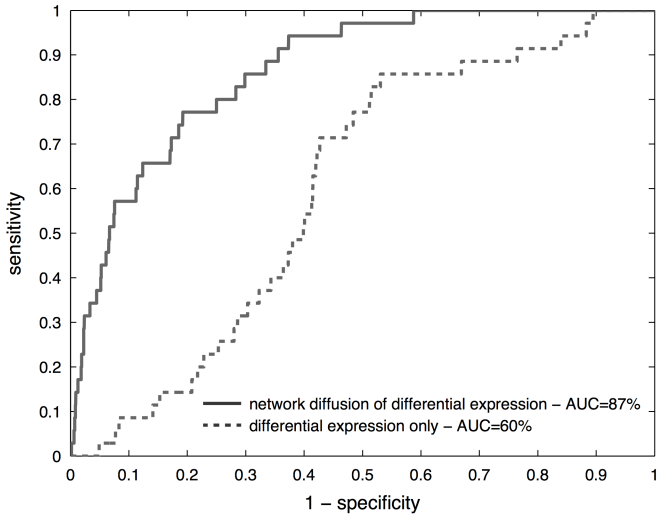


Figure 3.3: ROC curves comparing the performance of network diffusion-based drug target prioritization and prioritization based on differential expression only. Both methods were applied to a new validation data set consisting of thirty five drug treatments corresponding to eighty five known targets.

a valid email address was provided. If cookies are enabled, a hyperlink to the result page (study name may help in case of submitting several jobs) is also stored in a personal history for twenty days. The output is displayed in a series of tabs corresponding to the different analyses selected by the user.

3.4.1 Quality control and data exploration

In the Quality Control (Figure 3.4) and Data Exploration (Figure 3.5) tabs, the different plots are displayed along with a table in which each sample is assigned a new index as used in the plots. By clicking on a plot a larger image is shown, together with a short explanation.

3.4.2 Differential expression

The Differential Expression tab (Figure 3.6) contains a table listing all genes together with their $\log_2$ ratios and both the naive and adjusted p-values for differential expression. Genes are ranked according to the adjusted p-



Control	
Array Name	Array Index
GSM423010.CEL.gz	C.1
GSM423013.CEL.gz	C.2
GSM423015.CEL.gz	C.3
GSM423017.CEL.gz	C.4
GSM423019.CEL.gz	C.5
GSM423021.CEL.gz	C.6
GSM423023.CEL.gz	C.7
GSM423025.CEL.gz	C.8
GSM423027.CEL.gz	C.9
GSM423029.CEL.gz	C.10
GSM423031.CEL.gz	C.11

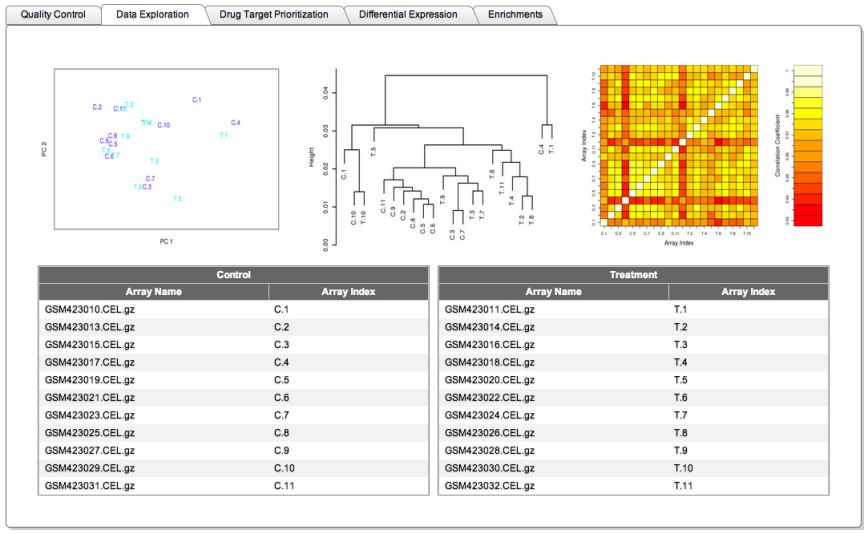
Treatment	
Array Name	Array Index
GSM423011.CEL.gz	T.1
GSM423014.CEL.gz	T.2
GSM423016.CEL.gz	T.3
GSM423018.CEL.gz	T.4
GSM423020.CEL.gz	T.5
GSM423022.CEL.gz	T.6
GSM423024.CEL.gz	T.7
GSM423026.CEL.gz	T.8
GSM423028.CEL.gz	T.9
GSM423030.CEL.gz	T.10
GSM423032.CEL.gz	T.11

Figure 3.4: The quality control output. In the Quality Control tab the different quality assessment plots are displayed along with a table in which each sample is assigned a new index as used in the plots.

value, but can for example also be ranked according to $\log_2$ ratio by simply clicking on the respective column header. Furthermore, different filters can be applied. The complete table can be exported in comma delimited CSV format, which can be imported by any spreadsheet. By clicking on a gene name, the user is brought to the corresponding GeneCards [193] (human), MGI (Mouse Genome Informatics) [191] (mouse), or RGD [192] (rat) section. To select the differentially expressed genes used for Enrichment Analysis, a p-value threshold can be entered at the top of the page. By default an adjusted p-value of 0.01 is used.

3.4.3 Drug target prioritization

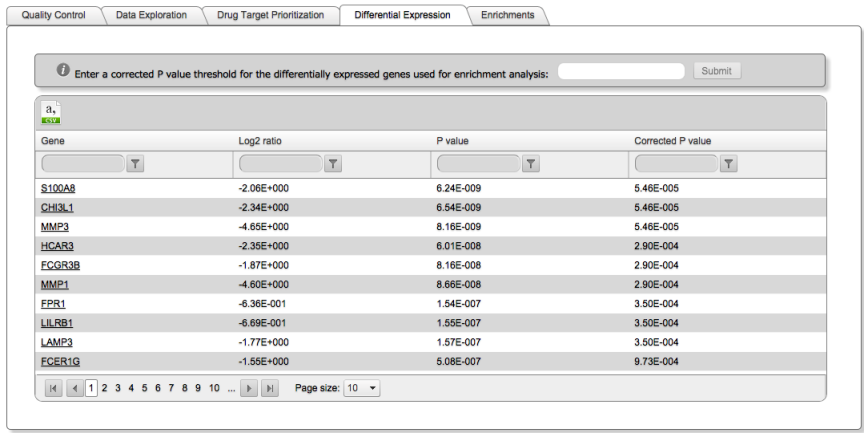
In the Drug Target Prioritization (Figure 3.7) tab, a ranked list of genes as potential targets of the drug can be found together with the network diffusion-based scores and both the naive and adjusted p-values for prioritization. Genes



Control	
Array Name	**Array Index**
GSM423010.CEL.gz	C.1
GSM423013.CEL.gz	C.2
GSM423015.CEL.gz	C.3
GSM423017.CEL.gz	C.4
GSM423019.CEL.gz	C.5
GSM423021.CEL.gz	C.6
GSM423023.CEL.gz	C.7
GSM423025.CEL.gz	C.8
GSM423027.CEL.gz	C.9
GSM423029.CEL.gz	C.10
GSM423031.CEL.gz	C.11

Treatment	
Array Name	**Array Index**
GSM423011.CEL.gz	T.1
GSM423014.CEL.gz	T.2
GSM423016.CEL.gz	T.3
GSM423018.CEL.gz	T.4
GSM423020.CEL.gz	T.5
GSM423022.CEL.gz	T.6
GSM423024.CEL.gz	T.7
GSM423026.CEL.gz	T.8
GSM423028.CEL.gz	T.9
GSM423030.CEL.gz	T.10
GSM423032.CEL.gz	T.11

Figure 3.5: The data exploration output. The Data Exploration tab is constructed in exactly the same manner as the Quality Control tab.



Gene	Log2 ratio	P value	Corrected P value
S100A8	-2.06E+000	6.24E-009	5.46E-005
CHI3L1	-2.34E+000	6.54E-009	5.46E-005
MMP3	-4.65E+000	8.16E-009	5.46E-005
HCAR3	-2.35E+000	6.01E-008	2.90E-004
FCGR3B	-1.87E+000	8.16E-008	2.90E-004
MMP1	-4.60E+000	8.66E-008	2.90E-004
FPR1	-6.36E-001	1.54E-007	3.50E-004
LILRB1	-6.69E-001	1.55E-007	3.50E-004
LAMP3	-1.77E+000	1.57E-007	3.50E-004
FCER1G	-1.55E+000	5.08E-007	9.73E-004

Figure 3.6: The differential expression output. In the Differential Expression tab all genes are displayed together with their $\log_2$ ratio, p-value, and adjusted p-value.

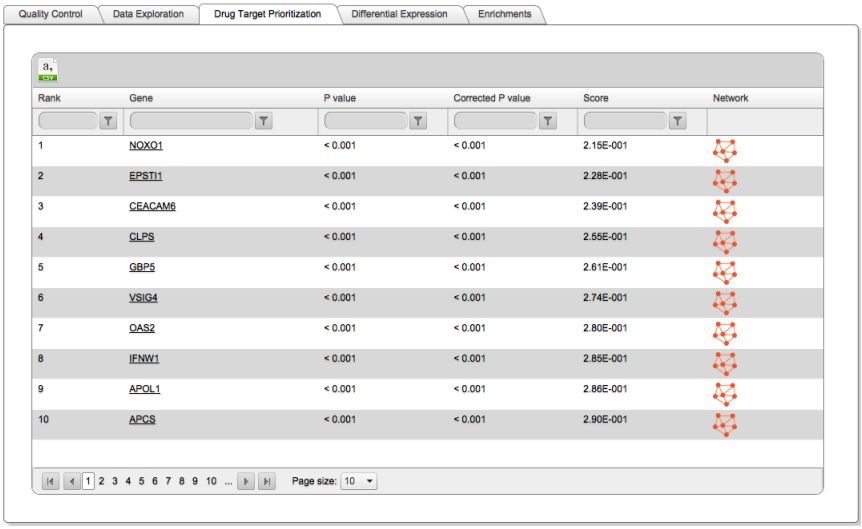


Figure 3.7: The drug target prioritization output. In the Drug Target Prioritization tab a ranked list of genes as potential targets of the drug can be found together with the network diffusion-based scores and both the naive and adjusted p-values for prioritization.

of interest can be detected in this list with the aid of the filter function. The complete table can again be exported as a CSV file and clicking on a gene name opens the corresponding GeneCards (human), MGI (mouse), or RGD (rat) section. In addition, a network-based visualization (Figure 3.8) is available for each gene, showing the ten interaction partners contributing most to the gene's ranking. Genes in this network are colored according to fold change and sized according to the significance (t-statistic) of this differential expression. Clicking on a node or edge opens the corresponding gene or interaction in STRING.

3.4.4 Enrichment

After selecting the differential expression threshold on which the gene set used for enrichment analysis is based, two separate tabs display tables summarizing the enrichment results for Pathway and Disease Enrichment (Figure 3.9). These tables contain Reactome/KEGG or HPO/MPO/RDO IDs and names, together with the number of differentially expressed genes in the corresponding gene sets and the accompanying p-values and adjusted p-values. The differentially expressed genes for each pathway or disease phenotype are accessible through dropdown tables. Clicking on an ID brings the user to the pathway diagram or ontology page. The enrichment tables can again be sorted, filtered, and

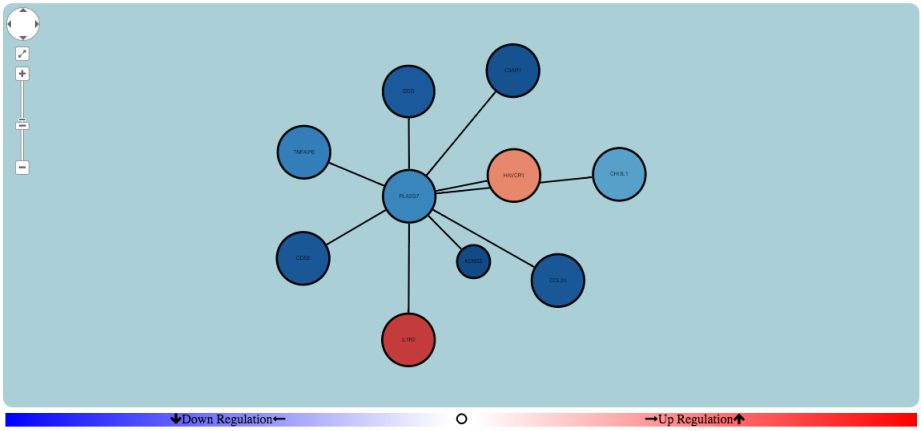


Figure 3.8: Interaction network for a potential target. A network-based visualization displays the ten interaction partners contributing most to a gene's ranking as potential drug target.

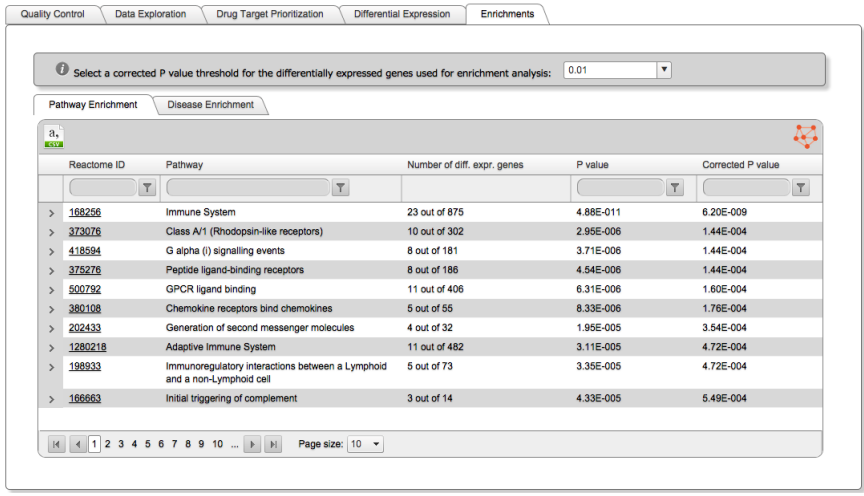


Figure 3.9: The pathway (and disease) enrichment output. In the Pathway and Disease Enrichment tabs a ranked list of pathways or phenotypes can be found, together with the number or differentially expressed genes per pathway or phenotype and both the naive and adjusted p-values for enrichment.

exported as CSV files. In addition, network graphs (Figure 3.10) are available, consisting of the top ten most significant pathways or disease phenotypes, along with their associated genes colored according to fold change.

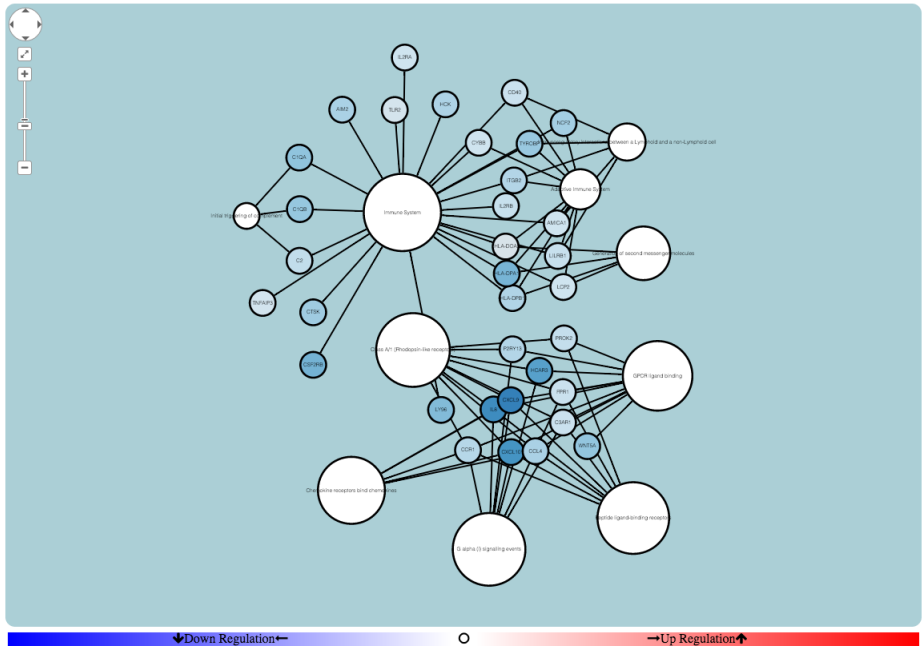


Figure 3.10: Pathway enrichment network. A network graph visualizes the top ten most significantly enriched, along with their associated genes colored according to fold change.

3.5 Implementation

The Galahad server has three modules: a user interface, a job controller, and an analyzer. The user interface provides a database-related service based on a three-tier architecture using Microsoft .Net technology. At the presentation layer on top, we developed a user-friendly and rich interface. To speed up interaction between the clients and the server, we added AJAX functionalities to retrieve large amounts of data as fast as possible. All the procedures for importing data into and retrieving data from the database are located in the data layer. The business layer represents the main link between the presentation layer and the data layer and thus contains all the rules for processing information.

The job controller is a simple Java application responsible for communication between the web application and the analyzer. After submission of a new request the job controller transfers the uploaded control and treatment file to a Linux server and runs the analyzer to process the request. When the results

are ready the job controller transfers them back to the web server and makes them available for the user.

The analyzer is the main module responsible for analysis of the user's request. This module is completely developed in R. The time needed for a job to finish is depending on the data size, the requested analyses, and the workload of our server. Data produced by the analyzer module are stored on a Microsoft SQL server. Galahad can be accessed with all main web browsers. Uploaded data and the results of analysis are kept private and not viewable by other users through the use of a secure server and are deleted after two weeks. The web site is free and open to all and there is no login requirement.

3.6 Case studies

To demonstrate the applicability of Galahad as an easy-to-use analysis suite for functional characterization of drugs, two data sets from our initial test data were subjected to analysis with our tool. GSE15947 containing Affymetrix gene experession data for immortalized human prostate epithelial cells treated with calcitriol and GSE24188 containing Affymetrix gene expression data for human primary hepatocytes treated with either atorvastatin or rosuvastatin.

3.6.1 Calcitriol

Calcitriol or 1,25-dihydroxycholecalciferol is the biologically active form of cholecalciferol or vitamin D3. This natural form of vitamin D is produced in the skin from 7-dehydrocholesterol upon irradiation. Apart from that, it can also be taken in the diet, although present in only a few foods including fortified dairy products and fish oils. After transportation to the liver vitamin D3 is hydroxylated to produce 25-hydroxycholecalciferol or calcidiol, the major circulating form of vitamin D3. Calcidiol is then transported further to the kidney where it is filtered by the glomerulus and in the proximal renal tubule is again hydroxylated, resulting in the formation of 1,25-dihydroxycholecalciferol or calcitriol, the functional, hormonally active form of vitamin D3 responsible for most of the biologic actions of vitamin D [194]. Calcitriol acts as a steroid hormone through activation of the nuclear vitamin D receptor (VDR) resulting in transcriptional regulation. Additionally, calcitriol may initiate the synthesis of second messengers such as cAMP, cGMP, inositol, triphosphate, diacylglycerol, and calcium through interaction with membrane VDRs. Both these membrane and nuclear VDRs may participate in the classical effects of calcitriol comprising

regulation of intestinal calcium and phosphate homeostasis, as well as calcium absorption in the distal tubes by working in concert with PTH [194, 195]. Furthermore, calcium homeostasis is highly intertwined with bone formation, as calcium is a major bone constituent hence the largest store of calcium in the body. Osteoblast VDR signaling participates in calcium metabolism primarily during a negative calcium balance by stimulating bone resorption. However, calcitriol can also inhibit bone matrix mineralization, thereby preserving normal serum calcium levels. Besides maintaining calcium and phosphate homeostasis, calcitriol regulates multiple cellular processes with effects on both normal and malignant cell growth and differentiation, on the innate and adaptive immune function, on cardiovascular function, and on the complex interplay with other hormones [194].

Calcitriol is commonly used as a medication in the treatment of vitamin D deficiency or insufficiency, refractory rickets, familial hypophosphatemia and hypoparathyroidism, and in the management of hypocalcemia and renal osteodystrophy in patients with chronic renal failure undergoing dialysis. Furthermore, it is used in conjunction with calcium in the management and prevention of primary or corticosteroid-induced osteoporosis and as treatment for psoriasis. Calcitriol is marketed under various trade names including Rocaltrol® (Roche), Calcijex® (Abbott), Decostriol® (Mibe, Jesalis) and Silkis®/Vectical® (Galderma) [76, 196]. Numerous studies also have shown beneficial effects of calcitriol on the malignant behavior of cancer cells, either by arresting cells in the G0/G1 phase of the cell cycle, inducing their differentiation, or by the induction of apoptotic cell death. Moreover calcitriol can influence angiogenesis, alter cell adhesion and migration, and reduce the invasiveness of cancer cells [194].

Although vitamin D signaling mainly affects cancer progression rather than carcinogenesis itself, epidemiological evidence suggests that high vitamin D status can also protect men from prostate cancer. It is however not clear whether this chemopreventative effect of high vitamin D status in the normal, healthy prostate is mediated by the same mechanisms. In this context Kovalenko and co-workers examined calcitriol-induced changes in the transcriptome of the phenotypically normal, immortalized human prostate epithelial cell line RWPE1, to provide new insights into the mechanisms by which vitamin D may prevent the development of human prostate cancer. Cells were treated with medium containing 100 nM of calcitriol or vehicle for six, twenty four or fourty eight hours and gene expression was measured using the Affymetrix HU133 plus 2.0 GeneChip [197]. To showcase the potential of Galahad for in-depth exploration of a drug's mechanism of action we analyzed the samples treated for six hours with our tool. VDR, the only target known to interact with calcitriol with an IC_{50} below 10 μM, was ranked sixteenth out of 13,465 with a p-value smaller than 0.001. The network depicting the ten genes contributing most to

the ranking score of VDR (Figure 3.8) shows a high upregulation of *CYP24A1*, known to be responsible for hydroxylation of calcitriol in order to limit the amount of circulating calcitriol in the kidneys, as well as of *CD14*, which is produced by prostate epithelial cells sensing the early signs of infection [194] and potentially contributes to the protection against breast cancer by stimulation of the innate immune system [198]. Other genes induced by calcitriol include *FOS*, a well-known proto-oncogene, *G6PD*, controlled through a VDRE located in the first intron and protecting RWPE1 cells against H_2O_2-induced apoptosis as shown by Bao and co-workers [199], and *DUSP10*, inhibiting TNF*alpha*-induced IL-6 production as shown by Nonn and co-workers [200]. Finally, both *VDR* itself and *CYP27B1*, encoding the enzyme responsible for the hydroxylation of calcidiol to calcitriol in the proximal renal tubule, are significantly downregulated following calcitriol treatment. The top enriched Reactome pathways mainly include Vitamin D metabolism, pathways related to the innate immune system, in particular Toll-like receptors (TLRs) cascades, regulation of lipid metabolism by peroxisome proliferator-activated receptor alpha (PPARA), and signaling by Rho GTPases, all of which have been linked to calcium signaling or more specifically to calcitriol treatment before [170, 194, 201]. The enriched phenotype ontologies comprise hypercalcemia and dehydration, both listed as important side effects in the calcitriol package inserts, as well as different forms of anemia, suggested to be prevented by maintenance of sufficient vitamin D status [202], and osteolysis, present in bone metastases of androgen-independent prostate cancer and treatable with calcitriol, as it has a direct effect on both the skeleton and the invading metastatic cells [203].

3.6.2 Atorvastatin and rosuvastatin

Statins are competitive inhibitors of 3-hydroxy-3-methyl-glutaryl coenzyme A reductase (HMGCR), a membrane-bound, endoplasmatic reticulum-localized enzyme catalyzing the two-step conversion of (S)-3-hydroxy-3-methylglutaryl-coenzyme A to 3-(R)-mevalonic acid through a putative hemithioacetal, the rate-limiting and first committed step in the biosynthesis of cholesterol from acetyl-coenzyme A. This tightly controlled pathway is regulated by the levels of low-density lipoprotein receptors on liver cells as the means of ensuring cholesterol homeostasis. Inhibition of biosynthesis reduces the amount of circulating atherogenic lipoproteins and thus is an effective therapy for lowering plasma cholesterol. Statins are among the most commonly used type of drugs worldwide in patients with hypercholesterolemia and other risk factors for cardiovascular disease, both in primary and secondary prevention [204, 205, 206].

Lovastatin, a fungal metabolite, was the first HMGCR inhibitor (HMGRI) approved by the FDA for treatment of hypercholesterolemia and was marketed

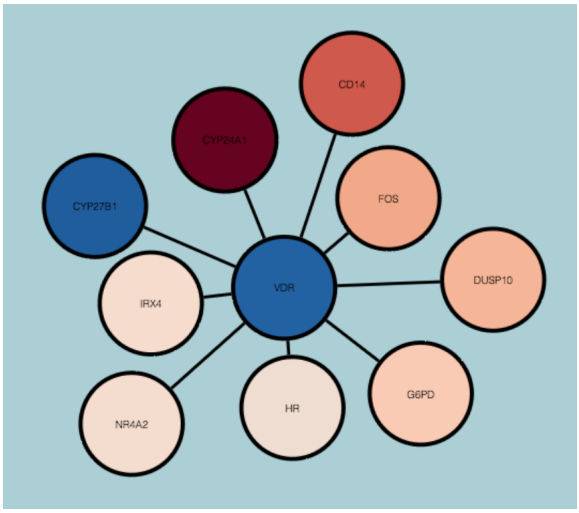


Figure 3.11: Interaction network for the calcitriol target VDR. Differential expression of VDR and its interaction partners contributing most to the drug target prioritization score are visualized. A red color indicates upregulation, while a blue color indicates downregulation.

under the trade name Mevacor®. Two other potent fungal metabolite HMGRIs would ultimately become marketed drugs, pravastatin (Pravachol®) and simvastatin (Zocor®). Nevertheless, due to concern over the effects of these compounds on cell viability, there was a perceived need for structurally novel HMGRIs such that any non-mechanism related toxicity would be avoided. In 1996, atorvastatin, a fully synthetic second generation statin got approved by the FDA and was marketed in the United States as Lipitor®. By the time atorvastatin got approved statins Mevacor®, Pravachol®, and Zocor® had all been approved for marketing by the FDA as well. However, these statins differ in affinity for the active site of HMGCR, entry rate into hepatic and non-hepatic tissues, systemic bioavailability, and elimination half-life and as atorvastatin had shown outstanding potency and efficacy at lowering total cholesterol, LDL-cholesterol, and triglycerides it became one of the most widely prescribed pharmaceuticals in the world [206]. In terms of structural binding characteristics atorvastatin is most similar to rosuvastatin, a third generation and the most potent statin approved for use in humans. Rosuvastatin, marketed as Crestor®, has a higher degree of selectivity for liver cells and undergoes only minor metabolism by the hepatic cytochrome P450 (CYP) system, primarily by CYP2C9, while atorvastatin is undergoing extensive CYP3A4 metabolism. Hence rosuvastatin has a longer elimination half-life,

compared with atorvastatin [205].

Statin-associated adverse effects cannot be overlooked as millions of patients are under statin treatment worldwide. However, clinical trials are often underpowered to detect uncommon side effects such as drug-induced liver injury (DILI), often only detected in the post-marketing phase. Large prospective randomized trials found a positive relationship between elevated liver enzymes and statin doses and in recent years, three prospective DILI studies have been undertaken in Spain, the USA, and Iceland. From these studies atorvastatin appeared to be the most frequently implicated statin in all of the cases of statin-induced hepatotoxicity [204]. To gain more insight into the interactions between cholesterol homeostasis and drug metabolism following statin treatment, Hafner and co-workers have investigated the transcriptome of human primary hepatocytes, a system similar to the human liver, treated by the two commonly prescribed cholesterol lowering drugs atorvastatin and rosuvasatin. Hepatocytes from human donors were treated for 6, 24, or 48 hours and gene expression was measured using the Affymetrix HU133 plus 2.0 GeneChip [205]. After quality assessment of the different sample groups with Galahad, we selected the samples treated for 24 hours for a further study on how both statins may influence the endogenous metabolic pathways in human primary hepatocytes. HMGCR, for both atorvastatin and rosuvastatin the only target known to interact with an IC_{50} below 10 μM, was ranked at position 77 for atorvastatin and at position 63 for rosuvastatin, two times with a p-value smaller than 0.001. From the HMGCR network for atorvastatin (Figure 3.12) it immediately becomes clear that not only *HMGCR* is strongly upregulated following treatment, but also a plethora of other genes involved in cholesterol biosynthesis including *HMGCS1*, *HMGCS2*, *TM7SF2*, *DHCR7*, *INSIG1*, *FDPS*, *MVK*, *MVD*, and *SQLE*. The HMGCR network for rosuvastatin (not displayed) is almost identical. Moreover, the top enriched Reactome pathways mainly include metabolism of lipids, in particular steroid, triglyceride, fatty acid, and ketone body metabolism and also the highest ranked phenotype ontologies are abnormality of lipid and cholesterol metabolism. As no enriched phenotype ontologies relate to liver toxicity, the most commonly tested liver enzymes were checked for differential expression, but neither glutamate oxaloacetate transaminase, glutamate pyruvate transaminase, alkaline phosphatase, or gamma-glutamyl transpeptidase was significantly altered following treatment with both atorvastatin and rosuvastatin. Finally, *CYP3A4* was not differentially expressed following rosuvastatin treatment, but significantly upregulated following atorvastatin treatment, suggesting rosuvastatin may be less prone to adverse drug interactions with potent inhibitors of *CYP3A4*. *CYP2C9* was not differentially expressed following either treatment.



Figure 3.12: Interaction network for the atorvastatin target HMGCR. Differential expression of HMGCR and its interaction partners contributing most to the drug target prioritization score are visualized. All eleven genes appear to be upregulated.

3.7 Conclusion

Galahad is an open and free platform to analyze microarray data for the characterization of drug effects. It provides an easy interface where in an integrated way quality control, data exploration, differential expression, and both pathway and disease enrichment can be calculated and visualized. In addition, drug target prioritization provides a ranked list of target candidates.

Chapter 4

A dual approach combining network diffusion of differential expression with structural similarity for improved drug target prediction

PhD candidate's contributions: Griet Laenen contributed to the method development, collected the training and test data, and performed all data analyses. She has written this chapter.

Part of the method development has been carried out by Sander Willems for his MSc thesis (September 2013).

Abstract

With the emergence of large-scale genomic, chemical, and pharmacological data, computational prediction of a drug's targets has become of significant importance in the drug discovery and repositioning process. Several strategies to integrate

these heterogeneous data sources in view of drug target prediction have been introduced. In Chapter 2, we integrated experimental gene expression data with prior knowledge on protein interactions to support the identification of putative targets of drugs. Drug-target interactions can, however, also be predicted from structural information. Building on the similar property principle, we therefore propose a second method that prioritizes proteins as drug targets based on the interaction with compounds structurally similar to the drug of interest, by combining compound-protein interaction scores with compound-compound similarity scores. Our method has been evaluated on a test set of small molecule drugs for which the known targets were derived from ChEMBL. AUC values of up to 96% were obtained. Moreover, as both this ligand-based and our expression-based approach produce a genome-wide ranking of potential targets, these two methods can be easily integrated by means of rank aggregation. In this way the results from our expression-based method could be significantly improved, demonstrating the added value of integrating multiple data sources in support of drug-target interaction prediction.

4.1 Introduction

The growing diversity of publicly accessible biological and pharmacological data provides a strong basis for data-driven target prediction and has created many possibilities for integrative approaches. Computational methods mining and integrating data from these heterogeneous sources hold great promise for gaining a global understanding of drug action [64]. Starting from this idea of data fusion for target prediction, we developed a straightforward ligand-based target prioritization, which can be smoothly integrated with our previously developed gene expression-based target prioritization. Our method applies the structural similarity principle onto the enormous amount of drug-protein interaction information integrated in the STITCH database [207], obtaining a score for each protein that allows for easy fusion with the expression-based ranking of potential targets. We performed extensive validation of the method and show that combining this ligand-based approach with our expression-based method leads to improved performance.

4.2 Materials and methods

4.2.1 Ligand-based drug target prioritization

Molecular fingerprints

To compare molecular structures on a large scale, some level of simplification or abstraction is required. Many such techniques for representing chemical structures are available and have been used widely for similarity searching. The abstractions most commonly applied are 2D molecular fingerprints, which involve turning the 2D representation of a chemical into a sequence of bits corresponding to the presence or absence of certain predefined features within the molecule. The two major types of 2D fingerprints are substructure key-based fingerprints and topological or path-based fingerprints [69, 208]. In addition, circular fingerprints have shown to perform well in virtual screening studies also [209]. We concerned ourselves with one well-established fingerprint for each of these three types. The different fingerprints were generated from the molecules' SMILES representation [210, 211, 212] using Open Babel [213].

The MACCS fingerprint is a substructure key-based fingerprint consisting of 166 bits that monitors the presence or absence of certain substructures or features from a given list of 166 predefined structural keys based on SMARTS patterns [214]. The path-based FP2 fingerprint identifies all linear and ring substructures in the molecule of lengths one to seven by mapping them onto a bit string of length 1,024 using a hash function [213]. The more complex circular Molprint2D descriptor encodes the environment of each atom by means of a vector varying in size, with elements being counts of atom types at a given distance from the central atom [215, 216].

Similarity search in STITCH

Using these molecular fingerprint representations, the potential targets of a drug compound can be predicted based on the 2D similarity between the drug and a set of ligands associated to their respective targets. The required ligand-target interaction information is available on a genome-wide scale in STITCH [207], a database combining chemical-protein interactions derived from multiple data sources into a unified network. STITCH integrates experimental information from ChEMBL [74], PDSP Ki Database [217], PDB [218], and two large-scale studies on kinase-ligand interactions [219, 220] with information from the pathway databases KEGG [221], PID [161], Reactome [188], and BioCyc [222] and the drug-target databases DrugBank [223], GLIDA [224],

Table 4.1: Number of proteins, ligands and interactions for each of the STITCH evidence types.

STITCH evidence type	Number of proteins	Number of ligands	Number of interactions
all	17,142	323,988	2,119,937
experimental	8,270	256,111	625,812
database	14,272	9,178	154,805
text mining	16,196	61,364	742,514
prediction	862	208,212	772,512

Matador [225], TTD [226], and CTD [227]. Moreover, STITCH contains predicted interactions based on text mining, as well as chemical structure. Interactions are also transferred to orthologous proteins in other organisms. We extracted the complete chemical-protein interaction network for human from STITCH version 4.0, as well as four subnetworks comprising a single interaction evidence type. The obtained protein Ensembl IDs were mapped to their encoding HGNC gene symbols. Table 4.1 provides an overview of the five networks.

Our structure-based drug target prioritization method takes into consideration the similarity between a drug compound of interest and all chemicals in this extracted network. Therefore, both this drug of interest and all chemicals from STITCH need to be encoded using the same fingerprint representation. The widely used Tanimoto coefficient remains the method of choice for assessment of 2D fingerprint-based similarity [228] and is defined as follows:

$$T_{A,B} = \frac{c}{a + b - c},$$

where fingerprints a and b are the number of bits set to one for compound A and B respectively, and c is the number of bits set to one for both compounds.

Protein score computation and normalization

By assessing the similarity of a drug D to all chemicals i in STITCH, a score S representing the best drug-protein link can be computed for each protein P as follows:

$$S_P = \max_i C_{P,i}.T_{D,i},$$

where $C_{P,i}$ is the confidence score for the interaction between protein P and chemical i, and $T_{D,i}$ is the Tanimoto similarity between drug D and chemical i. To place progressively greater weight on highly similar chemical pairs compared to those that are less similar, a power function can be applied to the Tanimoto

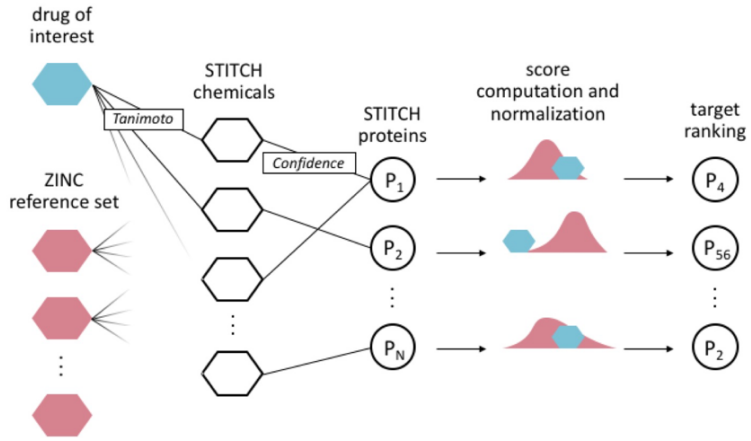


Figure 4.1: Schematic overview of protein score computation, normalization and ranking. By combining the Tanimoto similarity between a drug of interest and the ligands interacting with a protein on one hand, and the confidence score for these ligand-protein interactions on the other hand, a drug-protein interaction score is computed for each protein. Next, scores are normalized based on a set of reference compounds, after which proteins can be ranked as potential targets.

coefficient.

By taking the maximum, this raw score can become biased toward hub proteins. To correct for this, the raw score for each protein is compared to a reference distribution based on a reference set of 6,000 compounds obtained from the ZINC database, a free database of commercially available compounds for virtual screening [121]. Ranking all proteins based on the obtained normalized scores finally results in a genome-wide prioritization of potential targets based on chemical structure. A schematic representation of this protein score computation, normalization and ranking is given in Figure 4.1.

4.2.2 Expression-based drug target prioritization

As extensively discussed in Chapter 2, drug-induced gene expression responses can assist in the process of drug target prediction, especially when taking the differential expression of genes functionally related to potential drug targets into account. For this reason, we analyze gene expression following drug treatment in the context of the functional protein association network STRING

[152]. Differential expression signals are diffused to neighboring nodes in the network, such that genes can be prioritized as potential targets based on the transcriptional response of their surrounding network module.

The diffusion process applied here makes us of a Laplacian exponential diffusion kernel. By multiplying the differential expression values p_0 with a discrete approximation of this heat kernel, a scoring vector p_α and thus genome-wide drug target prioritization is obtained:

$$p_\alpha = p_0(I - \frac{\alpha}{N}L)^N.$$

Here N is the number of iterations, corresponding to the number of steps taken in the network; α is the diffusion rate and determines to what extent the contribution of a node to the ranking score decreases with an increasing distance in the network. Based on our previous results for 235 publicly available drug-induced gene expression data sets, we set N to 1 and α to 0.9. The regularized or symmetric normalized Laplacian L is computed from the weighted STRING network as follows:

$$L = I - D^{-1/2}AD^{-1/2},$$

with A being the adjacency matrix and D a diagonal matrix with $d_{ii} = \sum_j a_{ij}$.

Lastly p-values indicating a gene's significance for encoding a potential drug target are computed using a permutation test. The differential expression values are randomized across the network 1,000 times and a p-value is calculated for each gene as the fraction of permutations resulting in a score at least as extreme as the real score in p_α. In this way, the scores get corrected for the systematic bias toward highly connected nodes.

4.2.3 Rank aggregation

Both our ligand-based and expression-based approach result in a genome-wide ranking of potential targets, which should be combined to obtain a single, global ranking. Furthermore, when multiple gene expression data sets are available for a single drug, and thus multiple expression-based rankings are obtained for that drug, these need to be fused into a single ranking as well. To do so, we applied the robust rank aggregation method developed by Kolde and co-workers [229], as it is robust to outliers, noise, and errors and facilitates the calculation of significance probabilities for all elements in the final ranking. The method is based on order statistics as first described by Stuart and co-workers [230] and further optimized by Aerts and co-workers [231]. For each gene, the algorithm looks at how it is positioned in the obtained rankings compared to a null model

that assumes randomly ordered gene lists. Based on this a p-value can be assigned to each gene, used both for reordering the genes and deciding their significance. Taking only the best ranks into account for each gene, the method is robust, as is needed when using high-throughput gene expression data.

First, for each vector p_α , the p-values corresponding to its scores are converted into ranks. Using these ranks, a vector is generated for each gene, containing the ranks from the different p_α vectors. Dividing these ranks by the total number of genes in the corresponding p_α vector results in normalized ranks with a maximal value of 1. For such a normalized rank vector r, let $r_{(1)}, ..., r_{(n)}$ be a reordering of r such that $r_{(1)} \leq ... \leq r_{(n)}$. Then for rank vector $\hat{r}$, sampled from a uniform distribution, the probability $\beta_{k,n}(r)$ that $\hat{r}_{(k)} < r_{(k)}$ can be defined under this null model as

$$\beta_{k,n}(r) = \sum_{i=k}^{n} \binom{n}{i} r^i (1-r)^{n-i},$$

and the final score of rank vector r is defined as

$$\rho(r) = \min_{k=1,...,n} \beta_{k,n}(r).$$

Genes can now be ranked according to these rank vector ρ scores, resulting in a single ranked list of potential drug targets. When elements of the different rankings p_α do not overlap perfectly, for example when results from different microarray platforms are being aggregated, missing values are replaced with the maximal relative rank 1.

4.2.4 ROC analysis

The performance of expression-based, ligand-based, and combined drug target prioritization was assessed in terms of sensitivity and specificity by means of an ROC analysis. Imposing a cut off on the obtained rankings, genes were divided into positives and negatives. Known targets ranked above the cut off were considered true positives, whereas all other genes ranked above were considered false positives. Likewise, known targets ranked below the cut off were considered false negatives, while all other genes below were considered true negatives. The true positive rate (sensitivity) was plotted against the false positive rate (1 – specificity) for all possible cut offs and the area under this curve, known as the AUC, was computed. According to Hanley and McNeil [173], this evaluation measure is equivalent to the probability that a classifier ranks a randomly chosen positive instance higher than a randomly chosen negative instance, which in our case means the AUC would be 100% if all targets ranked at the top, while it would be only 50% if the prioritization was no better than ranking the genes randomly.

4.2.5 Training and test data for ligand-based drug target prioritization

For training and testing our method, known ligand-target interactions were extracted from the ChEMBL database [74]. All interactions with an IC_{50} value below 10 μM confirmed in at least three assays were identified, resulting in a total of 1,284 compounds interacting with 367 proteins, accounting for a total of 1,579 drug-protein interactions. For all of these compounds canonical SMILES were downloaded from ChEMBL and converted into molecular fingerprints as described above. The different proposed fingerprints and protein scoring functions were then evaluated on a training set consisting of 70% of the compounds; the remaining 30% was used for testing.

4.2.6 Test data for ligand-based, expression-based, and combined drug target prioritization

Our ligand-based and expression-based drug target prioritization strategy, as well as the combination of both, were eventually all evaluated on a set of 62 drug compounds, corresponding to the small molecule drugs used for evaluation of expression-based primary target prediction desribed in Chapter 2, as well as for the re-evaluation of expression-based target prediction described in Chapter 3. Gene expression data were downloaded as raw CEL files from GEO [232] and preprocessed with the RMA function of the Bioconductor affy [167] or oligo package [168]. After preprocessing, $\log_2$ ratio differential expression values were computed for each gene using the Bioconductor limma package [169]. Canonical SMILES were downloaded from ChEMBL and converted to molecular fingerprints as described above. For each compound, all protein targets binding with an IC_{50} value below 10 μM were identified. This resulted in a total of 142 proteins, accounting for a total of 526 drug-protein interactions.

4.3 Results and discussion

4.3.1 Protein scoring function

Proteins are scored as potential targets of a certain compound based on the similarity of their known ligands with that compound. More specifically, each protein gets scored by identifying its best scoring ligand, taking into account both the confidence score for ligand-protein interaction and the similarity between ligand and compound. Using a power function on this similarity, greater weight

Table 4.2: AUC values for the different protein scoring functions and molecular fingerprints.

		Molecular fingerprint		
		MACCS	FP2	Molprint2D
Protein scoring function	c.T	0.921	0.938	0.953
	$c.T^2$	0.937	0.950	0.957
	$c.T^3$	0.945	0.949	0.954
	$c.T^4$	0.945	0.942	0.946

can be placed on those ligands that are highly similar to the compound of interest. To select the optimal power function, all compounds in the training set were first removed from the STITCH database. Then, for each of these compounds protein scores were computed using different power functions for all three molecular fingerprints. Ranking the proteins based on the obtained scores and retreiving the positions of the known targets in the corresponding ranking, an AUC value was computed for each power function-fingerprint combination. Results are depicted in Table 4.2. AUC values range from 92 to 96%, with the squared Tanimoto performing best overall. Only for the MACCS fingerprint a higher power of the Tanimoto similarity performs better, which is due to the Tanimoto coefficients being higher on average for this fingerprint as shown in Figure 4.2.

4.3.2 STITCH evidence types

Next, we considered the dependency of the results on different interaction evidence types integrated in the STITCH network: experimental, database, text mining, and prediction based on chemical structure. The best performing protein scoring method, making use of a squared Tanimoto similarity for both FP2 and Molprint2D fingerprints and a third power of the Tanimoto similarity for MACCS fingerprints, was applied to four variants of the STITCH network, each comprising a single evidence type. Results obtained on the training data are depicted in Table 4.3. The AUC values for each type of evidence are only based on those compound-target interactions in the training set for which the target is present in the relevant network (i.e., 1,083 for experimental, 1,073 for database, 1,091 for text mining, and 942 for prediction, out of a total of 1,091 interactions for the network comprising all four evidence types). Nevertheless, experimental and prediction evidence clearly outperform both database and text mining evidence, which most likely can be attributed to the smaller number of compounds included in the two latter networks. Integrating the different

Table 4.3: AUC values for the different STITCH evidence types and molecular fingerprints.

		Molecular fingerprint			
		MACCS	FP2	Molprint2D	aggregated
STITCH evidence type	all	0.945	0.950	0.957	0.961
	experimental	0.952	0.958	0.960	0.967
	database	0.759	0.746	0.752	0.839
	text mining	0.730	0.757	0.822	0.863
	prediction	0.879	0.926	0.926	0.933

evidence types resolves this issue, providing us with a more complete and hence robust network for protein scoring. Moreover, by integrating all four evidence types a better protein coverage is achieved.

4.3.3 Molecular fingerprint

To identify the best performing fingerprint for our 2D similarity comparison, the MACCS, FP2, and Molprint2D fingerprints were compared based on the training set. Although results do not show remarkable differences (see Table 4.3), the circular Molprint2D fingerprint seems to exhibit a slightly better performance. Previous work aiming to compare different molecular fingerprints confirm these findings. For instance, based on the evaluation of 37 molecular fingerprints by calculating retrieval rates of active molecules, Bender and co-workers concluded that circular fingerprints such as Molprint2D perform best, regardless of diameter [233].

Next, the rankings obtained with each of the three fingerprints were aggregated to obtain a single consensus ranking for each compound in the training set. Despite the fact that these diverse fingerprints capture different aspects of molecular structure, this fusion of multiple fingerprints only marginally improved the results obtained with a single fingerprint, as was also observed by Bender and co-workers [233]. However, by combining fingerprints capturing different chemical information, the highest-ranked targets will be those that are highly ranked by several fingerprints, reducing the number of artifacts that would be introduced by a single fingerprint [208].

4.3.4 Performance on test data and comparison with SEA

Based on the results described above, the performance of our method was eventually evaluated on the test data. We made use of all STITCH evidence types and the best performing protein scoring function for each fingerprint, as was determined on the training data. After aggregation over the three fingerprints an AUC of 96.2% was obtained, corresponding exactly to the result obtained for the training data, with separate AUC values of 94.5%, 95.2% and 95.9% for respectively MACCS, FP2, and Molprint2D fingerprints.

Next, the performance of our method was compared to that of the Similarity Ensemble Approach (SEA) developed by Keiser and co-workers, a widely used and highly predictive method for target fishing based on ligand set similarity. Representing each target solely by the structures of its set of known ligands, SEA scores shared patterns of chemical functionality between a query compound and a ligand set by summing up all pairwise Tanimoto similarities above a certain threshold. This score is then normalized to an expectation value according to a null model of random similarity adapting the BLAST algorithm [143]. The authors have applied SEA to screen 3,665 drugs against hundreds of targets and confirmed 23 out of 30 predicted interactions that were tested experimentally [43]. SEA was also used to predict the activity of 656 marketed drugs on 73 unintended side-effect targets and approximately half of the predictions were confirmed, either from proprietary databases or experimentally [91]. Moreover, SEA was successfully applied to de-orphanize seven FDA-approved drugs without a known target [234]. Lastly, SEA has also been used to predict targets of 586 compounds that were active in a zebrafish behavioral assay. Out of 20 predictions tested experimentally, 11 were confirmed [235].

First, SEA was applied for target prioritization on all compounds in the test set by making use of the original setup where proteins are quantitavely related based on the chemical similarity of ligands present in ChEMBL20, represented by ECFP4 fingerprints. An AUC value of 81.8% was obtained. As STITCH is not limited to experimental data, but combines these data coming from, amongst others, ChEMBL itself with information from pathway and drug-target databases, and even predicted interactions based on text mining and chemical structure, we also applied the SEA approach using STITCH for training. To do so we either limited the training library to interactions with a confidence score higher than 0.7, or divided the interactions in affinity classes based on these confidence scores, as is also done by the original SEA making use of the ChEMBL20 library. In the first case an AUC value of 88.9% was obtained, while in the latter case the AUC value was 94.8%. This higher score obtained using the complete STITCH database is comparable with the results obtained with our own method on the same data and indicates that the predictive power

strongly depends on how comprehensive the library of known interactions is, rather than on the exact methodology used for similarity computation and protein scoring.

4.3.5 Effect of ligand similarity on the performance

The applicability of a ligand-based method is limited by the availability, number, and diversity of the known ligands for each target. Furthermore, our method by design only finds targets interacting with high confidence with already known ligands, hence performance depends strongly on the query compound. To evaluate the effect of the presence or absence of ligands highly similar to this compound of interest on the performance of our method, we applied it again to each compound in the test set after removal of its most similar ligands from the STICH network. Rather than filtering on Tanimoto similarity, a fixed percentage of compounds was held out, since different fingerprints can result in a completely different Tanimoto similarity for the same pair of molecules, as depicted in Figure 4.2.

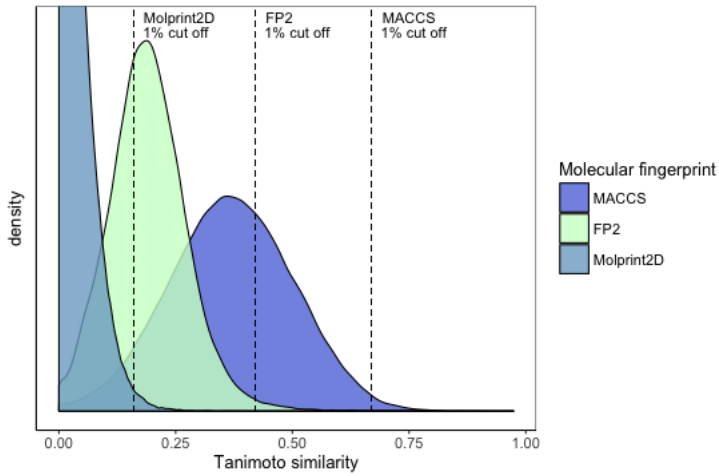


Figure 4.2: Distribution of the Tanimoto similarity for all compounds in the test data set with all compounds in STITCH, for each of the three fingerprints. Tanimoto similarities based on MACCS fingerprints are higher on average, while the use of FP2 and especially Molprint2D fingerprints results in lower Tanimoto similarities. To address this discrepancy, data cut offs are made using percentiles rather than Tanimoto similarity cut offs.

AUC values for removal of 0.01, 0.1, and 1% of the ligands most similar to each query compound in the test set are shown in Table 4.4. While still above 90%

Table 4.4: AUC values for the test data after removal of the most similar ligands for each query molecule.

		Tanimoto similarity cut off			AUC		
		MACCS	FP2	Molprint2D	MACCS	FP2	Molprint2D
Percentage	0.01%	<0.87	<0.80	<0.39	0.920	0.939	0.941
deleted	0.1%	<0.76	<0.59	<0.24	0.860	0.895	0.855
compounds	1%	<0.67	<0.42	<0.16	0.715	0.673	0.561

for a 0.01% cut off, AUC drops significantly for a 0.1% and especially a 1% cut off. The MACCS fingerprints appears to be the least sensitive to the absence of highly similar compounds, most likely due to a higher Tanimoto similarity on average for this fingerprint.

4.3.6 Improvement of gene expression-based drug target prioritization

Finally, we evaluated how our ligand-based approach performs when combined with our expression-based approach prioritizing proteins as potential targets of a drug by integrating experimental gene expression data with prior knowledge on protein interactions. As both methods result in a genome-wide ranked list of potential targets, the two methods can be combined by simply fusing their proteins rankings. Both drug target prioritization strategies were evaluated separately on a data set consisting of 62 compounds and 142 proteins, accounting for 526 drug-protein interactions in total. An AUC value of 79.5% was achieved using the expression-based method, while an AUC value of 91.8% was achieved with the ligand-based method, after combining the MACCS, FP2 and Molprint2D rankings (separately resulting in AUC values of respectively 89.2, 89.9, and 91.8%). Although the AUC value resulting from expression-based drug target prioritization is slightly lower than expected from results obtained before, aggregation of both the ligand-based and expression-based ranking resulted in an AUC value of 94.4%. ROC curves for both approaches, as well as for the combined result, are depicted in Figure 4.3.

4.4 Conclusion

To address the efficacy and safety issues accounting for the high attrition rates in the later stages of drug development, a better understanding of the complex

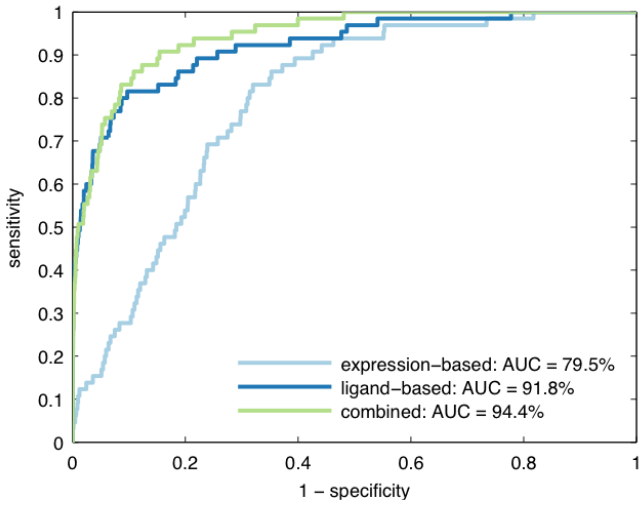


Figure 4.3: ROC curves comparing the performance of expression-based, ligand-based, and combined drug target prioritization. While our ligand-based method clearly outperforms our expression-based method on the test data set, performance is best when combining both.

biological response to drug treatment and more specifically the on- an off-targets interacting with a drug is vital. Complementing our gene expression-based drug target prioritization method relying on a kernel-based random walk through the STRING network and discussed in detail in Chapter 2, we propose a ligand-based approach further improving the prediction of a drug's targets. Building on the similar property principle, this second method prioritizes proteins as drug targets based on the protein interactions of compounds structurally similar to a drug of interest. For this, we make use of STITCH, a database built on the STRING data containing both known and predicted chemical-protein interactions. By means of a simple and computationally inexpensive prioritization strategy, AUC values up to 96% could be obtained on the test set. The three molecular fingerprints used in our study did not show significant differences in performance, but combining all three seems to result in a slightly better target ranking. With respect to the source of known target-ligand interactions, the results for the separate STITCH evidence types, as well as for application of the SEA methodology using either ChEMBL or STITCH, indicate the greater importance of interaction coverage compared to interaction confidence. Finally, when compared to our expression-based approach, structure-based drug target prioritization shows a higher predictive value. Structural similarity, however, tends to predict interactions of a more trivial origin, compared to the more complex associations

that can be captured by gene expression analysis. Nevertheless, combining both showed to further improve performance, which indicates the predictive power of integrating gene expression data and structural information for a drug of interest with known protein-protein and protein-compound interaction information respectively, to identify the targets of that drug. As such this dual method can aid in gaining a better knowledge of a candidate drug's mode of action and its off-target effects and thus be of great value in the drug development process. However, like most other ligand-based approaches our ligand-based method is not designed to identify targets of entirely new binders with new structures and binding modes, as suggested by the performance of our method after applying increasing Tanimoto cut offs to the STITCH data. In contrast, for compounds structurally similar to known binders our ligand-based method has proven to perform well, and thus can aid in improving their properties in terms of affinity and selectivity, as well as in predicting side effects and in drug repositioning.

Chapter 5

Discussion

Pharmacological and biological data are accumulating at ever increasing rates, presenting new opportunities in drug discovery and development. However, as these data are often complex, high-dimensional, and noisy, the challenge now lies in the development of *in silico* techniques to assimilate these data into testable hypotheses and actionable insights. Hence, modern, data-driven drug discovery requires sophisticated methods for both representing and integrating this variety of data. Here we look at our work and the added value of biological networks, in the light of this representation of pharmacological space and the integration of heterogeneous data types, as well as the application of these data towards drug target prediction. Finally, we shortly discuss the applicability of our work in a pharmaceutical research context and conclude with some future perspectives.

5.1 Representation of pharmacological space

Drug and drug-like compounds can be described or quantified in a variety of ways, enabling various downstream analyses and predictions. First of all, we can quantify the physicochemical properties of a drug such as melting point or hydrophobicity, but also chemical structure. There are different approaches to represent the chemical structure of small molecule compounds including molecular graphs and line notations, such as InChI and SMILES. While these character strings can be analyzed algorithmically, they are both variable-length and non-numeric, making them less straightforward to work with. For this reason, fixed-length binary fingerprints have been developed, in which each bit

corresponds to the presence or absence of a certain atom or substructure [60]. In Chapter 4 we make use of two types of such molecular fingerprints, MACCS and FP2, the first one being a substructure key fingerprint based on a list of 166 predefined structural keys, while the second one is a path-based fingerprint that identifies all linear and ring substructures in the molecule of lengths up to seven atoms. Besides these, we also included the more complex circular Molprint2D fingerprint. Although our ligand-based method did not show any significant differences in performance for these three fingerprints, several methods do heavily rely on the type of fingerprint used. The circular ECFP4 fingerprint, for example, is still one of the most widely used fingeprints because of its great performance in ligand-based virtual screening efforts [236].

Besides quantifying properties of a compound itself, we can also quantify downstream biological perturbations, such as gene expression. By simultaneously measuring the transcriptional effect of a compound on multiple genes, gene expression profiling gives a comprehensive snapshot of the biological state of a living system in response to drug treatment. As these transcriptomic changes can nowadays be measured in high-throughput, screening of many compounds in multiple cell lines is often used as a means to catalog the biological responses to a large number of diverse drug candidates at relatively low cost. The obtained genome-wide mRNA expression levels can be used as a proxy to measure transcriptional perturbations of cellular state by comparing expression in cellular samples with and without administration of the chemical compound. Each compound can then be represented by means of an expression profile, where each gene is assigned a value corresponding to the degree of up- or downregulation relative to the control state [60, 237]. In Chapter 2, we propose a method for drug target prioritization, starting from such gene expression profiles. This use of transcriptomic data for characterizing small molecules has become increasingly popular since the advent of the Connectivity Map, containing expression measurements for hundreds of *in vitro* drug perturbations in multiple human cell lines [78, 79]. However, to obtain good results a large number of parameters should be optimized for each perturbation, including cell type, concentration, and treatment duration. For this reason, to benchmark our methods, we opted for data coming from GEO, a public gene expression repository with over one million samples to date and covering a wide variety of experiments including drug perturbations.

Finally, drugs can be represented using metadata, such as disease indications, side effects, and known physiological interactions with other drugs. Similarly, we can also represent a drug by means of its interactions with biological targets on various levels going from binary level information simply indicating the presence or absence of an interaction to detailed, experimentally determined binding data. As such, a numeric interaction profile vector with length equal to

the total number of potential targets can be constructed for each drug. Taken together, these vectors can be translated into a network that may be used to explore the context of chemicals and proteins [60]. The known drug-target interactions used in Chapter 4 were extracted from STITCH, a network making this type of information available on a genome-wide scale.

5.2 Data integration

Pharmacological space comprises a variety of data types providing diverse information and a multi-view perspective on drug action. Over the last years, several computational strategies have been introduced that integrate these heterogeneous data sources to clarify the story surrounding a particular drug and its potential effects on the human body. Combining multiple data dimensions in the chemical, biological, and phenotypic space helps building models with improved predictive performance as some component of the noise contained in each data modality is independent, and hence, combining these modalities lessens the obfuscating effects of such noise. The structure of the data and specific research questions largely dictate the methods used for integration [60]. For both our expression-based and ligand-based drug target prediction method, results are represented as prioritized lists of genes, making a simple rank aggregation a useful and general solution for the integration task. We applied the robust rank aggregation method developed by Kolde and co-workers, which is robust to outliers, noise, and errors [229].

Networks provide an intuitive framework for integrating a wide variety of information sources and capturing both quantitative and qualitative relationships between entities, such as gene expression correlation or the presence or absence of protein interactions [60]. STITCH, for example, became one of the most comprehensive public databases of drug-target interactions, by integrating manually curated data sets with high-throughput experimental data and predicted interactions based on both text mining and chemical structure into one unified network. Both individual likelihood scores per evidence type and an overall confidence score are assigned to each interaction [126]. In Chapter 4, these confidence scores were extracted from the network for all ligands of each protein and used in combination with structural information on the ligands to score each protein as potential target for a drug of interest. The functional protein-protein association network STRING, sister project of STITCH, integrates prior knowledge from text mining with known interactions extracted from various databases and newly predicted interactions based on gene fusion events, conserved genomic neighborhoods, phylogenetic co-occurrence, and co-expression [152]. This makes STRING a perfect network scaffold for

studies with large omics data sets, connecting experimental results to external information, as we show for gene expression data in Chapter 2. An important drawback of many network approaches, however, is the tendency to be somewhat *ad hoc* in nature, having many tuning parameters [60]. Ideally, a systematic exploration of various parameter settings should be performed, as was done in Chapter 2 for the diffusion rate and the number of iterations in kernel diffusion ranking, as well as for the correlation threshold in correlation diffusion ranking, and in Chapter 4 for the power function on the Tanimoto similarity in the protein score computation.

5.3 Drug target prediction

As drugs exert their effects on biological systems by binding protein targets and affecting their downstream activity, knowledge of these interactions is of great importance toward understanding therapeutic mechanisms, as well as predicting higher-level responses including side effects and disease indications. However, for most drugs there are still many gaps in this knowledge and filling these gaps by means of the available wet lab techniques is currently infeasible in terms of both expenditure and effort. As a consequence, considering the tremendous growth of pharmacological data, the use of computational methods to predict potential interactions has become increasingly important in recent years. The most widely used computational methods for drug target identification can be classified in virtual screening on the one hand, both ligand-based and target-based, and phenotype-based methods on the other [60, 238]. A further distinction can be made between *de novo* target prediction (based only on pharmacological information for a compound itself) and similarity-based approaches, making use of information on compounds somehow related to the compound of interest to generate new predictions.

De novo target-based drug-target interaction prediction by means of molecular docking is advantageous in that it can provide structural insights into the nature of the interaction, but depends on the existence of reliable three-dimensional protein structures and is computationally too demanding for large-scale interaction prediction. By contrast, *de novo* drug-target interaction prediction by means of pharmacophore modeling is computationally much more efficient and is therefore often used to screen large compound libraries. Rather than screening compounds with a pharmacophore query, sometimes the compound itself is being used as the query to identify the pharmacophore model most likely fitting the compound. Performance is, however, highly dependent on the quality of the model and the scoring metric [60, 239]. Hence we opted for a similarity-based approach in Chapter 4, where we tried to improve our expression-

based drug target prioritization method by using structural information. Our method employs a 'guilt by association' approach based on the similar property principle, that is, chemical structure similarity between drugs tends to present similar targeting activities, meaning that if there are similar compounds known to interact with a certain target, this information can be used for predicting new interactions. While *de novo* target prediction has the potential to explore new areas of pharmacological space, our method has the potential to improve in accuracy as more data become available. Moreover, it is straightforward and computationally inexpensive.

Though compound structural similarity is likely to be the most natural and well-supported similarity metric used for drug-target interaction prediction, approaches based on phenotype similarity have also found success, including disease indication, side effect, and gene expression similarity. Gene expression profiling provides a genome-scale readout of cellular state following biological perturbation from drug treatment. Hence it enables quantitative molecular comparisons between drug compounds, with mechanisms hypothesized to be shared between compounds with similar profiles, an approach gaining ground with the publication of the Connectivity Map. Nevertheless, one needs to question the effect of tissue- or cell type-specificity of these transcriptional perturbations when using this approach [60]. Using a *de novo*-like approach on this type of data, a substantial focus has been placed on differentially regulated genes for new target identification. However, this focus did not account for the fact that many drug targets are not differentially expressed following treatment. Hence more recent approaches organize gene expression data along with other information into networks, allowing for a better functional interpretation of a transcriptional response to drug treatment, not considering a target as a single, isolated molecule, but rather as a component of a more complex functional module [240, 241]. In Chapter 2, we followed this path when prioritizing potential targets of a drug based on gene expression, by diffusing the differential expression signals over the functional protein association network STRING, either using a kernel-based random walk or on the basis of connectivity correlations between nodes. This network-based analysis of gene expression data is still gaining popularity, as illustrated in Figure 5.1, and our approach was picked up by Isik and co-workers, who further improved on it by means of a new network measure called local radiality, describing the reachability of a target protein via the shortest path to deregulated genes. Both methods were compared to eleven other state of the art methods for drug target prediction that make use of drug perturbation data and/or functional interaction networks, by applying them on the CMap data. Known drug-target interactions were extracted from the STITCH database. Local radiality and our symmetric kernel diffusion came out as the top predictors, with 153 out of 11,787 targets shared between the first percentile of both prioritizations, while respectively 153 and 79

targets were found only by local radiality or kernel diffusion [176]. Later, Woo and co-workers introduced a similar approach called DeMAND, assessing the global dysregulation of molecular interactions following compound perturbation for each gene based on the joint gene expression probability density for each of its regulon genes [83].

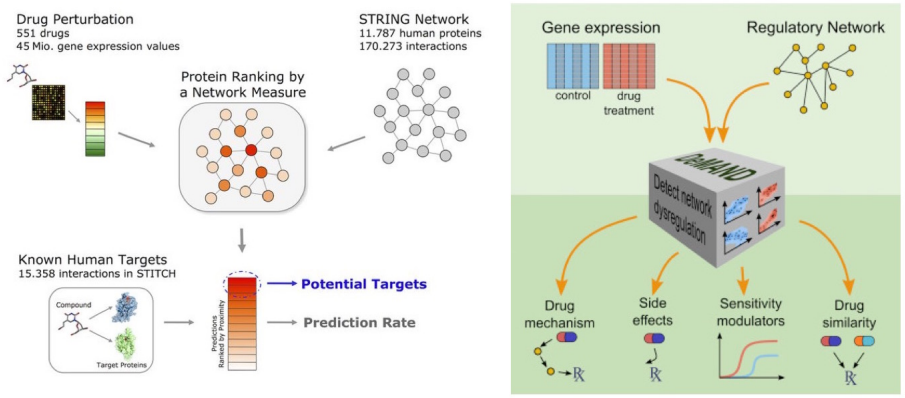


Figure 5.1: Computational methods for target prediction based on network analysis of gene expression illustrating the growing popularity of this approach. The left scheme depicts the methodology as described by Isik and co-workers, while the right scheme depicts the DeMAND approach proposed by Woo and co-workers (reprinted from [176, 83]).

5.4 Applicability of the work in a pharmaceutical research context

With both biological and technological advances, cell-based phenotypic assays have gained new momentum in pharmaceutical research. By addressing several (potentially novel) targets in parallel, phenotypic drug screening holds great promise for delivering first-in-class drugs for more complex, multifactorial diseases and diseases without a validated molecular target. Furthermore, by preserving the cellular context of protein function, phenotypic drug screening allows for the prioritization of compounds more likely to translate to *in vivo* and clinical efficacy. The cost paid for these benefits is that the precise protein targets and mechanism of action underlying the observed phenotype remain to be determined, which can be a complex endeavor [34, 242, 243].

Although a plethora of wet lab techniques for target identification are available, these are often time- and resource-intensive and usually require high-affinity interactions, making them more suitable for chemically optimized compounds than for primary phenotypic hits. Therefore, as discussed in Section 5.3, *in silico* methods for generating drug-target interaction hypotheses are quickly gaining ground. Over recent years, extensive information has been gathered connecting small molecules with biological activity and probable target proteins [243]. Hence, structural similarity searches, such as our ligand-based similarity method presented in Chapter 4, can be easily applied to derive hypotheses from known activities of similar compounds. However, this type of similarity-based target prediction is still limited by the availability, number, and diversity of the known ligands for each target. By means of our expression-based target prioritization method presented in Chapter 2, on the other hand, hypotheses regarding the mechanism of action of a compound can be generated *de novo*. Technological developments in transcriptomics coupled with the availability of chemical libraries and cell lines as models for numerous human diseases have made gene expression profiling a prominent technology for measuring the biological effects of compounds and many pharmaceutical companies now have a back catalog of compound-induced expression data at their disposal [237]. As our expression-based drug target prioritization method relies on the multiplication of a differential expression vector for a compound of interest with a kernel or correlation matrix, this approach can easily be applied on a big scale using matrix multiplication. This means results from both our ligand- and expression-based approach can be generated and thereby integrated early in the development process, preferably at the start of lead optimization.

After all, the retrospective identification of the full target spectrum of active compounds, termed target deconvolution, is not only important for elucidating the biological mechanisms of disease and the resulting development of target-based assays during phenotypic drug discovery, but has become an indispensable step in rational drug design, both for lead optimization and preclinical studies. Identification of the therapeutic targets allows for efficient structure–activity relationship studies to be carried out during chemical optimization, while information on interactions with 'off-target' proteins can address target-specific toxicity and side effects, as certain proteins have been associated with specific safety concerns. Finally, though side effects usually carry negative connotations, these unintended consequences can sometimes offer clues toward other therapeutic indications [60, 244, 245]. These indications can, however, also be identified directly based on the biological role a drug target plays in disease. Our tool Galahad, presented in Chapter 3, does not only include drug target prioritization, but also comes with gene set enrichment analysis to identify pathways affected by a drug as well as associated diseases. This enables both side effect identification and drug repurposing, making Galahad relevant

for post-market studies as well.

5.5 Future perspectives

The results in Chapter 4 demonstrate the predictive power of combining both our expression diffusion and structural similarity approaches, showing again that the target identification problem is not generally solved by a single method but rather by the integration of multiple, complementary approaches. It is therefore desirable to also make our ligand-based method available as part of our tool Galahad. Furthermore, as drug candidates are usually tested at different doses, Galahad would benefit from supporting the analysis of more complex experimental setups including dose-response experiments and time series. Lastly, extension of the tool to more organisms used in drug discovery (e.g. yeast) and other transcriptomic platforms would be beneficial as well.

After all, with both technological and analytical advances, more types of omics data are becoming available on a large scale, providing us with new, valuable resources to further expand our network diffusion approach. Without a doubt, one of the most profound developments of the past years is the enormous scale-up of the Connectivity Map, by means of the new, low-cost, high-throughput, reduced representation expression profiling method L1000. Combining ligation-mediated amplification with fluorescently addressed microspheres in a multiplex reaction, gene expression for about a thousand highly informative 'landmark genes' was measured following perturbation of a variable number of human cell lines with more than 20,000 distinct small molecules [246]. Moreover, as L1000 allows for thousands of compounds to be screened per day at relatively low cost, prominent pharmaceutical companies, such as Merck and Johnson & Johnson, have been reported to make use of this platform to generate large-scale transcriptomics data for compound profiling [247]. Since more than 80% of the remaining transcriptome can be inferred from the landmark gene expression measurements with high fidelity, our kernel and correlation diffusion methods could be applied to the combined set of measured and inferred expression values in exactly the same way as described in Chapter 2. However, to reduce noise introduced by this inference, it is also worth exploring diffusion of the directly measured expression values only. In this case, to obtain a genome-wide ranking of potential targets, the diffusion parameter α should be set to one and the optimal number of steps taken in the network needs to be re-evaluated. Finally, also other omics data such as relative quantitative proteomics and even epigenomics provide further avenues of development for the near future.

While in Chapter 2 a single kernel matrix was computed from the STRING

network and applied for network diffusion, the key advantage of kernels probably lies within the mining and integrating of data from different heterogeneous sources including biological networks, but also chemical structures, biological assays, and transcriptomics data. Data collected for a set of objects, such as genes, proteins, compounds, or even phenotypes, can be replaced by a kernel matrix that measures the similarity between any pair of objects across these data. To this end, the original data, which describes these objects either through a simple vector of measurements or through a more complex description such as a gene sequence, a protein structure, or the graph of a small molecule, are transformed into a similarity matrix by means of a so-called kernel function with appropriate mathematical properties. Once these kernel similarity matrices have been obtained, the data can be integrated by a class of data fusion methods called multiple kernel learning (MKL) and a large number of machine learning procedures for, among other, dimensionality reduction, classification, clustering, and ranking becomes available to apply to the data [248, 249]. Using this approach, it would be possible to extend our similarity-based method from Chapter 4 with other similarity metrics, for example based on physicochemical properties or phenotypic information such as gene expression, high-content imaging, or side effects. Carrying it one step further, we could apply a pairwise kernel function, measuring the similarity between drug-target pairs by combining a drug and a protein base kernel via a kernel product. However, in this case traditional MKL methods might be subject to memory limitations imposed by the pairwise kernel making them unable to perform predictions in the complete drug-protein space. Therefore, more recent methods integrating similarity measures for drugs and targets often rely on matrix factorization. Besides its use in multi-scale integration, matrix factorization also allows for data imputation via matrix completions, where missing entries in a matrix are filled in based on the observed entries [60]. Recently our group has developed Macau, a method for Bayesian matrix factorization extended with high-dimensional side information [250]. Macau has already been applied for gene prioritization through gene-phenotype matrix completion using multiple sources of genomic and phenotypic side information [251], for biological activity prediction through compound-assay matrix completion using cellular images as side information [252], and for linking drug targets to signaling pathway activation through compound-cell line matrix computation using protein targets as drug side information and both gene expression and pathway activity as cell line side information [253].

5.6 General conclusion

Though the pharmaceutical industry is still struggling to counteract stagnating R&D output, the ever-increasing amount of both biological and pharmacological data provides them with a whole range of new opportunities. Hence, computational pharmacology has become a burgeoning area of research enabling a systematic exploration of drug space and the generation of novel hypotheses surrounding drug action. With new sources of drug-related information at hand and the generation of high-throughput data sets becoming more routine, several new and exciting directions can now be explored. Omics technologies, such as transcriptomics, facilitate the application of a system biology strategy toward a better comprehension of disease mechanisms and therapeutic intervention, by considering a therapeutic target a component of a complex functional module, rather than an independent, isolated molecule. Network pharmacology approaches taking this biological complexity into account do not only contribute to a system-level understanding of drug action, they also enhance hypothesis generation by providing a comprehensive framework for organizing scientific questions and systematic experimentation. Furthermore, networks can be employed as scaffolds to connect experimental results to prior knowledge and integrate multiple sources of data, providing a more complete picture of a biological system or pharmacological space. As such, network pharmacology can help improve drug design efficiency, the success rate of new drugs in the clinic, the finding of new uses for existing drugs, and even the prediction of drug effects in individual patients enabling personalized medicine.

Bibliography

[1] E. Lekka, S. N. Deftereos, A. Persidis, A. Persidis, and C. Andronis. Literature analysis for systematic drug repurposing : a case study from Biovista. *Drug Discovery Today: Therapeutic Strategies*, 8(3-4):103–108, 2011.

[2] A. A. Ciociola, L. B. Cohen, and P. Kulkarni. How drugs are developed and approved by the FDA : current process and future directions. *The American Journal of Gastroenterology*, 109(5):620–623, 2014.

[3] Pharmaceutical Research and Manufacturers of America. 2016 biopharmaceutical research industry profile. `http://phrma-docs.phrma.org/sites/default/files/pdf/biopharmaceutical-industry-profile.pdf`, 2016. Accessed: June, 2018.

[4] KMR Group. 14 years to create a new drug. `http://kmrgroup.com/wp-content/uploads/2015/09/2015_09_21_KMRGroup_RandD_CycleTime_Release.pdf`, 2015. Accessed: June, 2018.

[5] M. E. Bunnage. Getting pharmaceutical R&D back on target. *Nature Chemical Biology*, 7(6):335–339, 2011.

[6] J. A. Dimasi, H. G. Grabowski, and R. W. Hansen. Innovation in the pharmaceutical industry: new estimates of R&D costs. *Journal of Health Economics*, 47:20–33, 2016.

[7] S. Morgan, P. Grootendorst, J. Lexchin, C. Cunningham, and D. Greyson. The cost of drug development: a systematic review. *Health policy*, 100(1):4–17, 2011.

[8] E. S. Pronker, T. C. Weenen, H. R. Commandeur, A. D. M. E. Osterhaus, and H. J. H. M. Claassen. The gold industry standard for risk and cost of drug and vaccine development revisited. *Vaccine*, 29(35):5846–5849, 2011.

[9] R. G. Frank. New estimates of drug development costs. *Journal of Health Economics*, 22:325–330, 2003.

[10] J. L. Lamattina. The impact of mergers on pharmaceutical R&D. *Nature Reviews Drug Discovery*, 10(8):559–560, 2011.

[11] D. Lendrem, S. J. Senn, B. C. Lendrem, and J. D. Isaacs. R&D productivity rides again? *Pharmaceutical Statistics*, 14(1):1–3, 2015.

[12] U. Schulze, M. Baedeker, Y. T. Chen, and D. Greber. R&D productivity: on the comeback trail. *Nature Reviews Drug Discovery*, 13(5):331–332, 2014.

[13] Pharmaceutical Research and Manufacturers of America. 2019 profile biopharmaceutical research industry. `https://www.phrma.org/-/media/Project/PhRMA/PhRMA-Org/PhRMA-Org/PDF/2019-Profile-Booklet_FINAL_NoBleeds.pdf`, 2019. Accessed: December 1, 2019.

[14] J. W. Scannell, A. Blanckley, H. Boldon, and B. Warrington. Diagnosing the decline in pharmaceutical R&D efficiency. *Nature Reviews Drug Discovery*, 11(3):191–200, 2012.

[15] B. Munos. Lessons from 60 years of pharmaceutical innovation. *Nature Reviews Drug Discovery*, 8(12):959–968, 2009.

[16] M. Hay, D. W. Thomas, J. L. Craighead, C. Economides, and J. Rosenthal. Clinical development success rates for investigational drugs. *Nature Biotechnology*, 32(1):40–51, 2014.

[17] S. M. Paul, D. S. Mytelka, C. T. Dunwiddie, C. C. Persinger, B. H. Munos, S. R. Lindborg, and A. L. Schacht. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery*, 9(3):203–214, 2010.

[18] L. Martin, M. Hutchens, C. Hawkins, and A. Radnov. How much do clinical trials cost? *Nature Reviews Drug Discovery*, 16:381–382, 2017.

[19] *Fundamentals of Clinical Trials*. Springer International Publishing, New York, New York, 2015.

[20] US Food and Drug Administration. The Drug Development Process Step 3: Clinical Research. `https://www.fda.gov/patients/drug-development-process/step-3-clinical-research#phases`, 2019. Accessed: November, 2019.

[21] *The Prevention and Treatment of Missing Data in Clinical Trials*. National Academies Press, Washington, D.C., 2010.

[22] Pharmaceutical Research and Manufacturers of America. Alzheimer's Disease 2013 Report. https://https://http://phrma-docs.phrma.org/sites/default/files/Alzheimer's%202013.pdf, 2013. Accessed: November, 2019.

[23] D.W. Thomas, J. Burns, J. Audette, A. Carroll, C. Dow-Hygelund, and M. Hay. Clinical Development Success Rates 2006-2015. https://www.bio.org/sites/default/files/Clinical%20Development%20Success%20Rates%202006-2015%20-%20BIO,%20Biomedtracker,%20Amplion%202016.pdf, 2016. Accessed: November, 2019.

[24] J. Arrowsmith and R. Harrison. Drug repositioning: the business case and current strategies to repurpose shelved candidates and marketed drugs. In Michael J. Barrat and Donald E. Frail, editors, *Drug repositioning: bringing new life to shelved assets and existing drugs*, chapter 1, pages 9–32. John Wiley & Sons, Inc., Hoboken, New Jersey, 2012.

[25] J. Arrowsmith and P. Miller. Trial watch: phase II and phase III attrition rates 2011-2012. *Nature Reviews Drug Discovery*, 12(8):569, 2013.

[26] K.H. Harrison. Phase ii and phase iii failures: 2013-2015. *Nature Reviews Drug Discovery*, 15(12):817–818, 2016.

[27] K. Smietana, M. Siatkowski, and M. Møller. Trends in clinical success rates. *Nature Reviews Drug Discovery*, 15(6):379–380, 2016.

[28] H. Dowden and J. Munro. Trends in clinical success rates and therapeutic focus. *Nature Reviews Drug Discovery*, 18(7):495–496, 2019.

[29] I. Lloyd. Pharma R&D Annual Review 2019. https://https://pharmaintelligence.informa.com/~/media/informa-shop-window/pharma/2019/files/whitepapers/pharma-rd-review-2019-whitepaper.pdf, 2019. Accessed: November, 2019.

[30] A. L. Hopkins. Network pharmacology: the next paradigm in drug discovery. *Nature Chemical Biology*, 4(11):682–690, 2008.

[31] D. C. Swinney and J. Anthony. How were new medicines discovered? *Nature Reviews Drug Discovery*, 10(7):507–519.

[32] D. C. Swinney. Phenotypic vs. target-based drug discovery for first-in-class medicines. *Clinical Pharmacology & Therapeutics*, 93(4):299–301, 2013.

[33] J. Eder, R. Sedrani, and C. Wiesmann. The discovery of first-in-class drugs: origins and evolution. *Nature Reviews Drug Discovery*, 13(8):577–587, 2014.

[34] M. Schenone, V. Dančík, B. K. Wagner, and P. A. Clemons. Target identification and mechanism of action in chemical biology and drug discovery. *Nature Chemical Biology*, 9(4), 2013.

[35] Y. Hu. Compound promiscuity: what can we learn from current data? *Drug Discovery Today*, 18(13–14):644–650, 2013.

[36] R. Morphy, C. Kay, Z. Rankovic, and R. Morphy. From magic bullets to designed multiple ligands. *Drug Discovery Today*, 9(15):641–651, 2004.

[37] G. V. Paolini, R. H. B. Shapland, W. P. Van Hoorn, J. S. Mason, and A. L. Hopkins. Global mapping of pharmacological space. *Nature Biotechnology*, 24(7):805–815, 2006.

[38] A. Lavecchia and C. Cerchia. In silico methods to address polypharmacology: current status , applications and future perspectives. *Drug Discovery Today*, 21(2):288–298, 2016.

[39] A. D. W. Boran and R. Iyengar. Systems approaches to polypharmacology and drug discovery. *Current Opinion in Drug Discovery and Development*, 13(3):297–309, 2010.

[40] A. Petrelli. Polypharmacological kinase inhibitors: new hopes for cancer therapy. In Peters J.-U., editor, *Polypharmacology in drug discovery*, chapter 8, pages 149–165. John Wiley & Sons, Inc., Hoboken, New Jersey, 2012.

[41] W.K. Kroeze and B.L. Roth. Polypharmacological drugs: “magic shotguns” for psychiatric diseases. In J.-U. Peters, editor, *Polypharmacology in Drug Discovery*, chapter 7, pages 135–148. John Wiley & Sons, Inc., Hoboken, New Jersey, 2012.

[42] J.-U. Peters. Polypharmacology - foe or friend? *Journal of Medical Chemistry*, 56(22):8955–8971, 2013.

[43] 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. 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.

[44] A. Anighoro, J. Bajorath, and G. Rastelli. Polypharmacology: challenges and opportunities in drug discovery. *Journal of Medical Chemistry*, 57(19):7874–7887, 2014.

[45] T. T. Ashburn and K. B. Thor. Drug repositioning: identifying and developing new uses for existing drugs. *Nature reviews Drug discovery*, 3(8):673–683, 2004.

[46] N. Nosengo. New tricks for old drugs. *Nature*, 534(7607):314–316, 2016.

[47] F. F. Yonkmann. New drugs for old uses and new uses for old drugs. *Journal of the Michigan State Medical Society*, 58(6):913–918, 1959.

[48] K. Shameer, B. Readhead, and J. T. Dudley. Computational and experimental advances in drug repositioning for accelerated therapeutic stratification. *Current Topics in Medicinal Chemistry*, 15(1):5–20, 2015.

[49] M. Boolell, M. J. Allen, S. A. Ballard, S. Gepi-Attee, G. J. Muirhead, A. M. Naylor, I. H. Osterloh, and C. Gingell. Sildenafil: an orally active type 5 cyclic GMP-specific phosphodiesterase inhibitor for the treatment of penile erectile dysfunction. *Nature Chemical Biology*, 8(2):47–52, 1996.

[50] D. Cavalla and C. Singal. Retrospective clinical analysis for drug rescue: for new indications or stratified patient groups. *Drug Discovery Today*, 17(3-4):104–109, 2012.

[51] L. M. Jarvis. Teaching an old drug new tricks. *Chemical & Engineering News*, 84(7):52–55, 2006.

[52] M. Wadman. New cures sought from old drugs. *Nature*, 490(7418):15, 2013.

[53] M.-C. Baek, B. Jung, H. Kang, H.-S. Lee, and J.-S. Bae. Novel insight into drug repositioning: methylthiouracil as a case in point. *Pharmacological Research*, 99:185–193, 2015.

[54] M. R. Hurle, L. Yang, Q. Xie, D. K. Rajpal, P. Sanseau, and P. Agarwal. Computational drug repositioning: from data to therapeutics. *Clinical Pharmacology & Therapeutics*, 93(4):335–341, 2013.

[55] C. R. Chong and D. J. Sullivan. New uses for old drugs. *Nature*, 448(7154):645–646, 2007.

[56] N. Novac. Challenges and opportunities of drug repositioning. *Trends in Pharmacological Sciences*, 34(5):267–272, 2013.

[57] G. F. Wilkinson and K. Pritchard. In vitro screening for drug repositioning. *Journal of Biomolecular Screening*, 20(2):167–179, 2015.

[58] J. Li, S. Zheng, B. Chen, A. J. Butte, S. J. Swamidass, and Z. Lu. A survey of current trends in computational drug repositioning. *Briefings in Bioinformatics*, 17(1):2–12, 2016.

[59] J. H. Jung, J. Ho, and J. Kwon. Target deconvolution of bioactive small molecules : the heart of chemical biology and drug discovery. *Archives of Pharmacal Research*, 38(9):1627–1641, 2015.

[60] R. A. Hodos, B. A. Kidd, K. Shameer, B. P. Readhead, and J. T. Dudley. In silico methods for drug repurposing and pharmacology. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 8(3):186–210, 2016.

[61] J. Lee and M. Bogyo. Target deconvolution techniques in modern phenotypic profiling. *Current Opinion in Drug Discovery and Development*, 17(1):118–126, 2013.

[62] J. Mestres, E. Gregori-Puigjané, S. Valverde, and R. V. Solé. The topology of drug-target interaction networks: implicit dependence on drug properties and target families. *Molecular BioSystems*, 5(9):1051–1057, 2009.

[63] M. Kuhn, M. Campillos, P. González, L. J. Jensen, and P. Bork. Large-scale prediction of drug-target relationships. *FEBS Letters*, 582(8):1283–1290, 2008.

[64] M. Iskar, G. Zeller, X.-M. Zhao, V. Van Noort, and P. Bork. Drug discovery in the age of systems biology: the rise of computational approaches for data integration. *Current Opinion in Biotechnology*, 23(4):609–616, 2012.

[65] A. Koutsoukas, B. Simms, J. Kirchmair, P. J. Bond, A. V. Whitmore, S. Zimmer, M. P. Young, J. L. Jenkins, M. Glick, R. C. Glen, and A. Bender. From in silico target prediction to multi-target drug design: current databases, methods and applications. *Journal of Proteomics*, 74(12):2554–2574, 2011.

[66] X. Liu, Y. Xu, S. Li, Y. Wang, J. Peng, C. Luo, X. Luo, M. Zheng, K. Chen, and H. Jiang. In silico target fishing: addressing a 'Big Data' problem by ligand-based similarity rankings with data fusion. *Journal of Cheminformatics*, 6:33, 2014.

[67] M. E. Johnson and G. M. Maggiora, editors. *Concepts and applications of molecular similarity*. John Wiley & Sons, Inc., Hoboken, New Jersey, 1990.

[68] A. E. Cleves and A. N. Jain. Robust ligand-based modeling of the biological targets of known drugs. *Journal of Medical Chemistry*, 49(10):2921–2938, 2006.

[69] P. Willett. Similarity searching using 2D structural fingerprints. In J. Bajorath, editor, *Chemoinformatics and computational chemical biology*, chapter 5, pages 133–158. 2011.

[70] A. Peón, C. C. Dang, and P. J. Ballester. How reliable are ligand-centric methods for target fishing? *Frontiers in Chemistry*, 4:15, 2016.

[71] D. W. Young, A. Bender, J. Hoyt, E. Mcwhinnie, G.-W. Chirn, C. Y. Tao, J. A. Tallarico, M. Labow, J. L. Jenkins, T. J. Mitchison, and Y. Feng. Integrating high-content screening and ligand-target prediction to identify mechanism of action. *Nature Chemical Biology*, 4(1):59–68, 2008.

[72] M. Campillos, M. Kuhn, A.-C. Gavin, L. K. Jensen, and P. Bork. Drug target identification using side-effect similarity. *Science*, 321(5886):263–266, 2008.

[73] J. Rihel, D. A. Prober, A. Arvanites, K. Lam, S. Zimmerman, S. Jang, S. J. Haggarty, D. Kokel, L. L. Rubin, R. T. Peterson, and A. F. Schier. Zebrafish behavioral profiling links drugs to biological targets and rest/wake regulation. *Science*, 327(5963):348–351, 2010.

[74] A. P. Bento, A. Gaulton, A. Hersey, L. J. Bellis, J. Chambers, M. Davies, F. A. Kru, Y. Light, L. Mak, S. McGlinchey, M. Nowotka, G. Papadatos, R. Santos, and J. P. Overington. The ChEMBL bioactivity database: an update. *Nucleic Acids Research*, 42(Database issue):1083–1090, 2014.

[75] C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, Y. Djoumbou, R. Eisner, A. C. Guo, and D. S. Wishart. DrugBank 3.0: a comprehensive resource for 'Omics' research on drugs. *Nucleic Acids Research*, 39(Database issue):1035–1041, 2011.

[76] 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. *Clinical Pharmacology and Therapeutics*, 92(4):414–7, 2012.

[77] M. Kuhn, M. Campillos, I. Letunic, L. J. Jensen, and P. Bork. A side effect resource to capture phenotypic effects of drugs. *Molecular Systems Biology*, 6(343):343, 2010.

[78] 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, and S. A. Carr. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313(5795):1929–1935, 2006.

[79] J. Lamb. The Connectivity Map: a new tool for biomedical research. *Nature Reviews Cancer*, 7(1):54–60, 2007.

[80] 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. *Science Translational Medicine*, 3(96):96ra77, 2011.

[81] 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. *Science Translational Medicine*, 3(96):96ra76, 2011.

[82] E. J. Cosgrove, Y. Zhou, T. S. Gardner, and E. D. Kolaczyk. Predicting gene targets of perturbations via network-based filtering of mRNA expression compendia. *Bioinformatics*, 24(21):2482–2490, 2008.

[83] J. H. Woo, Y. Shimoni, S. Yang, B. R. Stockwell, M. Bansal, and A. Califano. Elucidating compound mechanism of action by network perturbation analysis. *Cell*, 162(2):441–451, 2015.

[84] M. Kuhn, M. Campillos, P. Bork, L. J. Jensen, A.-C. Gavin, M. P. Costi, R. Luciani, R. Preissner, H. Fan, J. Hossbach, H. Jing, and J. Ahmed. Use of aprepitant and derivatives thereof for the treatment of cancer. WO/2009/124756; PCT/EP2009/002621, 10 2009.

[85] 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 Computational Biology*, 5(7):e1000423, 2009.

[86] A. K. Chavali, A. S. Blazier, J. L. Tlaxca, P. A. Jensen, R. D. Pearson, and J. A. Papin. Metabolic network analysis predicts efficacy of FDA-approved drugs targeting the causative agent of a neglected tropical disease. *BMC Systems Biology*, 6(27):1–16, 2012.

[87] J. Von Eichborn, M. S. Murgueitio, M. Dunkel, S. Koerner, P. E. Bourne, and R. Preissner. PROMISCUOUS: a database for network-based drug-repositioning. 39(Database issue):1060–1066, 2011.

[88] S. J. Cockell, J. Weile, P. Lord, C. Wipat, and D. Andriychenko. An integrated dataset for in silico drug discovery. *Journal of Integrative Bioinformatics*, 7(3):1–13, 2010.

[89] A. Merino, A. K. Bronowska, D. B. Jackson, and D. J. Cahill. Drug profiling: knowing where it hits. *Drug Discovery Today*, 15(17-18):749–756, 2010.

[90] L. Brouwers, M. Iskar, G. Zeller, V. Van Noort, and P. Bork. Network neighbors of drug targets contribute to drug side-effect similarity. *PLoS One*, 6(7):e22187, 2011.

[91] E. Lounkine, M. J. Keiser, S. Whitebread, D. Mikhailov, J. Hamon, J. L. Jenkins, P. Lavan, E. Weber, A. K. Doak, S. Côté, B. K. Shoichet, and L. Urban. Large-scale prediction and testing of drug activity on side-effect targets. *Nature*, 486(7403):361–367, 2012.

[92] H. Zhou and J. Skolnick. FINDSITE(comb): a threading/structure-based, proteomic-scale virtual ligand screening approach. *Journal of Chemical Information and Modeling*, 53:230–240, 2013.

[93] H. Zhou, M. Gao, and J. Skolnick. Comprehensive prediction of drug-protein interactions and side effects for the human proteome. *Scientific Reports*, 5(11090):1–13.

[94] R. L. Chang, L. Xie, L. Xie, P. E. Bourne, and B. Ø. Palsson. Drug off-target effects predicted using structural analysis in the context of a metabolic network model. *PLoS Computational Biology*, 6(9), 2010.

[95] L. Xie, J. Li, L. Xie, and P. E. Bourne. Drug discovery using chemical systems biology: identification of the protein-ligand binding network to explain the side effects of CETP inhibitors. *PLoS Computational Biology*, 5(5):e1000387, 2009.

[96] S. Jaeger and P. Aloy. From protein interaction networks to novel therapeutic strategies. *IUBMB Life*, 64(6):529–537, 2012.

[97] A. Zhang. *Protein interaction networks computational analysis*. Cambridge University Press, New York, New York, 2009.

[98] M. P. H. Stumpf, T. Thorne, E. De Silva, R. Stewart, H. J. An, and M. Lappe. Estimating the size of the human interactome. *Proceedings of the National Academy of Sciences*, 105(19):6959–6964, 2008.

[99] K. Venkatesan, J.-F. Rual, A. Vazquez, U. Stelzl, I. Lemmens, T. Hirozane-Kishikawa, T. Hao, M. Zenkner, X. Xin, K.-I. Goh, M. A. Yildirim,

N. Simonis, K. Heinzmann, F. Gebreab, J. M. Sahalie, S. Cevik, C. Simon, A.-S. De Smet, E. Dann, N. Klitgord, R. R. Murray, C. Lin, M. Lalowski, J. Timm, K. Rau, C. Boone, P. Braun, M. E. Cusick, F. P. Roth, D. E. Hill, J. Tavernier, M. Vidal, and E. E. Wanker. An empirical framework for binary interactome mapping. *Nature Methods*, 6(1):83–90, 2009.

[100] J. Piehler. New methodologies for measuring protein interactions in vivo and in vitro. *Current Opinion in Structural Biology*, 15(1):4–14, 2005.

[101] L. I. Furlong. Human diseases through the lens of network biology. *Trends in Genetics*, 29(3):150–159, 2013.

[102] E. D. Coelho, J. P. Arrais, and J. Luís. From protein-protein interactions to rational drug design: are computational methods up to the challenge? *Current Topics in Medicinal Chemistry*, 13(5):602–618, 2013.

[103] A. Chatr-Aryamontri, B.-J. Breitkreutz, R. Oughtred, L. Boucher, S. Heinicke, D. Chen, C. Stark, A. Breitkreutz, N. Kolas, L. O. Donnell, T. Reguly, J. Nixon, L. Ramage, A. Winter, A. Sellam, C. Chang, J. Hirschman, C. Theesfeld, J. Rust, M. S. Livstone, K. Dolinski, M. Tyers, and D. Montr. The BioGRID interaction database : 2015 update. 43(November 2014):470–478, 2015.

[104] T. S. K. Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, C. J. H. Kishore, S. Kanth, M. Ahmed, M. K. Kashyap, R. Mohmood, Y. L. Ramachandra, V. Krishna, B. A. Rahiman, S. Mohan, P. Ranganathan, and S. Ramabadran. Human Protein Reference Database - 2009 update. 37(November 2008):767–772, 2009.

[105] S. Kerrien, B. Aranda, L. Breuza, A. Bridge, F. Broackes-Carter, C. Chen, M. Duesbury, M. Dumousseau, M. Feuermann, U. Hinz, C. Jandrasits, R. C. Jimenez, J. Khadake, U. Mahadevan, P. Masson, I. Pedruzzi, E. Pfeiffenberger, P. Porras, A. Raghunath, B. Roechert, S. Orchard, and H. Hermjakob. The IntAct molecular interaction database in 2012. 40(Database issue):841–846, 2012.

[106] L. Licata, L. Briganti, D. Peluso, L. Perfetto, M. Iannuccelli, E. Galeota, F. Sacco, A. Palma, A. P. Nardozza, E. Santonico, L. Castagnoli, and G. Cesareni. MINT the molecular interaction database : 2012 update. 40(Database issue):857–861, 2012.

[107] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg. The Database of Interacting Proteins: 2004 update. *Nucleic acids research*, 32(Database issue):D449–51, 2004.

[108] M. Kotlyar, C. Pastrello, N. Sheahan, and I. Jurisica. Integrated interactions database : tissue-specific view of the human and model organism interactomes. *Nucleic Acids Research*, 44(Database issue):536–541, 2016.

[109] A. Kamburov, U. Stelzl, H. Lehrach, and R. Herwig. The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Research*, 41(Database issue):793–800, 2013.

[110] H. Calderone, L. Castagnoli, and G. Cesareni. Mentha: a resource for browsing integrated protein-interaction networks. *Nature Methods*, 10(8):690–691, 2013.

[111] M. J. Cowley, M. Pinese, K. S. Kassahn, N. Waddell, J. V. Pearson, S. M. Grimmond, A. V. Biankin, S. Hautaniemi, and J. Wu. PINA v2.0: mining interactome modules. *Nucleic Acids Research*, 40(Database issue):862–865, 2012.

[112] B. Turner, S. Razick, A. L. Turinsky, J. Vlasblom, E. K. Crowdy, E. Cho, K. Morrison, I. M. Donaldson, and S. J. Wodak. iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. *Database*, 2010:baq023, 2010.

[113] D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos, K. P. Tsafou, M. Kuhn, P. Bork, L. J. Jensen, and C. Von Mering. STRING v10: protein–protein interaction networks, integrated over the tree of life. *Database issue*, 43(Database issue):447–452, 2015.

[114] E. Chautard, N. Thierry-Mieg, and S. Ricard-Blum. Interaction networks: from protein functions to drug discovery. A review. *Pathologie Biologie*, 57(4):324–333, 2009.

[115] A.-L. Barabási, N. Gulbahce, and J. Loscalzo. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68, 2011.

[116] T. Ideker and R. Sharan. Protein networks in disease. *Genome Research*, 18(4):644–652, 2008.

[117] A. Mora and I. M. Donaldson. Effects of protein interaction data integration, representation and reliability on the use of network properties for drug target prediction. *BMC Bioinformatics*, 13:294, 2012.

[118] S. I. Berger and R. Iyengar. Network analyses in systems pharmacology. *Bioinformatics*, 25(19):2466–2472, 2009.

[119] P. Li, Y. Fu, and Y. Wang. Network based approach to drug discovery: a mini review. *Mini-Reviews in Medicinal Chemistry*, 15(8):687–695, 2015.

[120] M. Vogt, D. Stumpfe, and G. M. Maggiora. Lessons learned from the design of chemical space networks and opportunities for new applications. *Journal of Computer-Aided Molecular Design*, 30(3):191–208, 2016.

[121] John J Irwin, Teague Sterling, Michael M Mysinger, Erin S Bolstad, and Ryan G Coleman. ZINC: a free tool to discover chemistry for biology. *Journal of Chemical Information and Modeling*, 52(7):1757–1768, 2012.

[122] M. Zwierzyna and M. Vogt. Design and characterization of chemical space networks for different compound data sets. *Journal of Computer-Aided Molecular Design*, 29(2):113–125, 2015.

[123] F. Iorio, R. Bosotti, E. Scacheri, V. Belcastro, P. Mithbaokar, R. Ferriero, L. Murino, R. Tagliaferri, N. Brunetti-Pierri, A. Isacchi, and D. di Bernardo. Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proceedings of the National Academy of Sciences of the United States of America*, 107(33):14621–14626, 2010.

[124] M. Kuhn, C. Von Mering, M. Campillos, and L. J. Jensen. STITCH: interaction networks of chemicals and proteins. *Nucleic Acids Research*, 36(Database issue):684–688, 2008.

[125] M. Kuhn, D. Szklarczyk, A. Franceschini, C. Von Mering, L. J. Jensen, and P. Bork. STITCH 3: zooming in on protein-chemical interactions. 40(Database issue):876–880, 2012.

[126] D. Szklarczyk, A. Santos, C. Von Mering, L. J. Jensen, P. Bork, and M. Kuhn. STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data. 44(Database issue):380–384, 2016.

[127] S. C. Janga and A. Tzakos. Structure and organization of drug-target networks: insights from genomic approaches for drug discovery. *Molecular BioSystems*, 5(12):1536–1548, 2009.

[128] M. A. Yıldırım, K.-I. Goh., M. E. Cusick, A.-L. Barabási, and M. Vidal. Drug-target network. *Nature Biotechnology*, 25(10):1119–1126, 2007.

[129] Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda, and M. Kanehisa. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, 24(13):232–240.

[130] D. Emig, A. Ivliev, O. Pustovalova, L. Lancashire, S. Bureeva, Y. Nikolsky, and M. Bessarabova. Drug target prediction and repositioning using an integrated network-based approach. *PLoS One*, 8(4):e60618, 2013.

[131] X. Wang and N. Gulbahce. Network-based methods for human disease gene prediction. *Briefings in Functional Genomics*, 10(5):280–293, 2011.

[132] P. I. Wang and E. M. Marcotte. It's the machine that matters: predicting gene function and phenotype from protein networks. *Journal of Proteomics*, 73(11):2277–2289, 2010.

[133] M. Cao, H. Zhang, J. Park, N. M. Daniels, M. E. Crovella, L. J. Cowen, and B. Hescott. Going the distance for protein function prediction: a new distance metric for protein interaction networks. *PLoS One*, 8(10):e76339, 2013.

[134] S. Kohler, S. Bauer, D. Horn, and P. N. Robinson. Walking the interactome for prioritization of candidate disease genes. *American Journal of Human Genetics*, 82(4):949–958, 2008.

[135] Y. Li and J. C. Patra. Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. *Bioinformatics*, 26(9):1219–1224, 2010.

[136] R. Kondor and J.-P. Vert. Diffusion kernels. In Tsuda K. Schölkopf R. and Vert J.-P., editors, *Kernel methods in computational biology*, chapter 1, pages 171–191. MIT Press, Cambridge, UK, 2004.

[137] O. Vanunu, O. Magger, E. Ruppin, T. Shlomi, and R. Sharan. Associating genes and protein complexes with disease via network propagation. *PLoS Computational Biology*, 6(1):e1000641, 2010.

[138] S. Whitebread, J. Hamon, D. Bojanic, and L. Urban. Keynote review: in vitro safety pharmacology profiling: an essential tool for successful drug development. *Drug Discovery Today*, 10(21):1421–1433, 2005.

[139] R. B. Rothman, M. H. Baumann, J. E. Savage, L. Rauser, A. McBride, S. J. Hufeisen, and B. L. Roth. Evidence for possible involvement of 5-HT(2B) receptors in the cardiac valvulopathy associated with fenfluramine and other serotonergic medications. *Circulation*, 102:2836–2841, 2000.

[140] M. S. Saporito, Lipinski C. A., and Reaume A. G. Phenotypic in vivo screening to identify new, unpredicted indications for existing drugs and drug candidates. In Barratt M. J. and Frail D. E., editors, *Drug repositioning: bringing new life to shelved assets and existing drugs*, chapter 9, pages 253–290. John Wiley & Sons, Inc., Hoboken, New Jersey, 2012.

[141] S. Ekins, A. J. Williams, M. D. Krasowski, and J. S. Freundlich. In silico repositioning of approved drugs for rare and neglected diseases. *Drug Discovery Today*, 16(7-8):298–310, 2011.

[142] A. Bender, J. Scheiber, M. Glick, J. W. Davies, K. Azzaoui, J. Hamon, L. Urban, S. Whitebread, and J. L. Jenkins. Analysis of pharmacology data and the prediction of adverse drug reactions and off-target effects from chemical structure. *ChemMedChem*, 2(6):861–873, 2007.

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

[144] Nidhi, M. Glick, J. W. Davies, and J. L. Jenkins. Prediction of biological targets for compounds using multiple-category bayesian models trained on chemogenomics databases. *Journal of Chemical Information and Modeling*, 46(3):1124–1133, 2006.

[145] Y. Z. Chen and D. G. Zhi. Ligand-protein inverse docking and its potential use in the computer search of protein targets of a small molecule. *Proteins: Structure, Function and Genetics*, 43(2):217–226, 2001.

[146] E. Kellenberger, N. Foata, and D. Rognan. Ranking targets in structure-based virtual screening of three-dimensional protein libraries: methods and problems. *Journal of Chemical Information and Modeling*, 48(5):1014–1025, 2008.

[147] Y. Y. Li, J. An, and S. J. M. Jones. The emergence of effective leaders: an experimental and computational approach. *E:CO Emergence: Complexity and Organization*, 8(4):36–51, 2006.

[148] S. Zhao and S. Li. Network-based relating pharmacological and genomic spaces for drug target identification. *PLoS ONE*, 5(7), 2010.

[149] F. Iorio, R. Tagliaferri, and D. di Bernardo. Identifying network of drug mode of action by gene expression profiling. *Journal of Computational Biology*, 16(2):241–251, 2009.

[150] G. Hu and P. Agarwal. Human disease-drug network based on genomic expression profiles. *PLoS ONE*, 4(8), 2009.

[151] D. Nitsch, J. P. Gonçalves, F. Ojeda, B. de Moor, and Y. Moreau. Candidate gene prioritization by network analysis of differential expression using machine learning approaches. *BMC Bioinformatics*, 11(1):460, 2010.

[152] D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M.l Stark, J. Muller, P. Bork, L. J. Jensen, and C. Von Mering. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Research*, 39(Database issue):561–568, 2011.

[153] J. Amberger, C. Bocchini, and A. Hamosh. A new face and new challenges for Online Mendelian Inheritance in Man (OMIM®). *Human Mutation*, 32(5):564–567, 2011.

[154] Z. Lu. PubMed and beyond: a survey of web tools for searching biomedical literature. *Database*, 2011:baq036, 2011.

[155] A. Ceol, A. C. Aryamontri, L. Licata, D. Peluso, L. Briganti, L. Perfetto, L. Castagnoli, and G. Cesareni. MINT, the molecular interaction database: 2009 update. *Nucleic Acids Research*, 38(Database issue):532–539, 2009.

[156] T. S. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, C. J. Harrys Kishore, S. Kanth, M. Ahmed, M. K. Kashyap, R. Mohmood, Y. I. Ramachandra, V. Krishna, B. A. Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady, and A. Pandey. Human Protein Reference Database - 2009 update. *Nucleic Acids Research*, 37(Database issue):767–772, 2009.

[157] C. Alfarano, C. E. Andrade, K. Anthony, N. Bahroos, M. Bajec, K. Bantoft, D. Betel, B. Bobechko, K. Boutilier, E. Burgess, K. Buzadzija, R. Cavero, C. D'Abreo, I. Donaldson, D. Dorairajoo, M. J. Dumontier, M. R. Dumontier, V. Earles, R. Farrall, H. Feldman, E. Garderman, Y. Gong, R. Gonzaga, V. Grytsan, E. Gryz, V. Gu, E. Haldorsen, A. Halupa, R. Haw, A. Hrvojic, L. Hurrell, R. Isserlin, F. Jack, F. Juma, A. Khan, T. Kon, S. Konopinsky, V. Le, E. Lee, S. Ling, M. Magidin, J. Moniakis, J. Montojo, S. Moore, B. Muskat, I. Ng, J. P. Paraiso, B. Parker, G. Pintilie, R. Pirone, J. J. Salama, S. Sgro, T. Shan, Y. Shu, J. Siew, D. Skinner, K. Snyder, R. Stasiuk, D. Strumpf, B. Tuekam, S. Tao, Z. Wang, M. White, R. Willis, C. Wolting, S. Wong, A. Wrong, C. Xin, R. Yao, B. Yates, S. Zhang, K. Zheng, T. Pawson, B. F F Ouellette, and C. W V Hogue. The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Research*, 33(Database issue.):418–424, 2005.

[158] C. Stark, B. J. Breitkreutz, A. Chatr-Aryamontri, L. Boucher, R. Oughtred, M. S. Livstone, J. Nixon, K. Van Auken, X. Wang, X. Shi, T. Reguly, J. M. Rust, A. Winter, K. Dolinski, and M. Tyers. The

BioGRID interaction database: 2011 update. *Nucleic Acids Research*, 39(Database issue):698–704, 2011.

[159] M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, and Y. Yamanishi. KEGG for linking genomes to life and the environment. *Nucleic Acids Research*, 36(Database issue):480–484, 2008.

[160] 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. MayMahajan, 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 Research*, 39(Database issue):691–697, 2011.

[161] C. F. Schaefer, K. Anthony, S. Krupa, J. Buchoff, M. Day, T. Hannay, and K. H. Buetow. PID: the pathway interaction database. *Nucleic Acids Research*, 37(Database issue):674–679, 2009.

[162] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25(1):25–29, 2000.

[163] T. Rattei, P. Tischler, S. Götz, M. A. Jehl, J. Hoser, R. Arnold, A. Conesa, and H. W. Mewes. SIMAP-A comprehensive database of pre-calculated protein sequence similarities, domains, annotations and clusters. *Nucleic Acids Research*, 38(SUPPL.1):223–226, 2009.

[164] R. I. Kondor and J. Lafferty. Diffusion Kernels on Graphs and Other Discrete Input Spaces. *ICML '02 Proceedings of the Nineteenth International Conference on Machine Learning*, pages 315–322, 2002.

[165] H. Yang, I. King, and M. R. Lyu. DiffusionRank: A Possible Penicillin for Web Spamming. *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 431–438, 2007.

[166] E. Nabieva, K. Jim, A. Agarwal, B. Chazelle, and M. Singh. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics*, 21(Suppl. 1):302–310, 2005.

[167] 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.

[168] B. S. Carvalho and R. A. Irizarry. A framework for oligonucleotide microarray preprocessing. *Bioinformatics*, 26(19):2363–2367, 2010.

[169] G. K. Smyth. limma: linear models for microarray data. In R. Gentleman, V. Carey, W. Huber, R. Irizarry, and S. Dudoit, editors, *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, number 2005, chapter 23, pages 397–420. Springer International Publishing, New York, New York, 2005.

[170] Y. Wang, J. Xiao, T. O. Suzek, J. Zhang, J. Wang, Z. Zhou, L. Han, K. Karapetyan, S. Dracheva, B. A. Shoemaker, E. Bolton, A. Gindulyte, and S. H. Bryant. PubChem's BioAssay database. *Nucleic Acids Research*, 40(Database issue), 2012.

[171] A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, and J. P. Overington. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, 40(Database issue):1100–1107, 2012.

[172] T. Liu, Y. Lin, X. Wen, R. N. Jorissen, and M. K. Gilson. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Research*, 35(Database issue):198–201, 2007.

[173] J. A. Hanley and B. J. McNeil. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148(3):839–43, 1983.

[174] M. Iskar, M. Campillos, M. Kuhn, L. J. Jensen, V. Van Noort, and P. Bork. Drug-induced regulation of target expression. *PLoS Computational Biology*, 6(9):e1000925, 2010.

[175] H. N. Chua, W. K. Sung, and L. Wong. Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics*, 22(13):1623–1630, 2006.

[176] Zerrin Isik, Christoph Baldow, Carlo Vittorio Cannistraci, and Michael Schroeder. Drug target prioritization by perturbed gene expression and network information. *Nature Publishing Group*, pages 1–13, 2015.

[177] A. Seal, Y.-Y. Ahn, and D. J. Wild. Optimizing drug-target interaction prediction based on random walk on heterogeneous networks. *Journal of Cheminformatics*, 7:40, 2015.

[178] W. Wang, S. Yang, X. Zhang, and J. Li. Drug repositioning by integrating target information through a heterogeneous network model. *Bioinformatics*, 30(20):2923–2930, 2014.

[179] Y. Luo, X. Zhao, J. Zhou, J. Yang, Y. Zhang, W. Kuang, J. Peng, L. Chen, and J. Zeng. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nature Communications*, 8(1):573, 2017.

[180] T. Montero-Melendez and M. Perretti. Connections in pharmacology: innovation serving translational medicine. *Drug Discovery Today*, 19(7):820–823, 2014.

[181] X. A. Qu and D. K. Rajpal. Applications of Connectivity Map in drug discovery and development. *Drug Discovery Today*, 17(23-24):1289–1298, 2012.

[182] R. A. Irizarry, B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs, and T. P. Speed. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research*, 31(4):e15, 2003.

[183] C. L. Wilson and C. J. Miller. Simpleaffy : a BioConductor package for Affymetrix Quality Control and data analysis. 21(18):3683–3685, 2005.

[184] B. M. Bolstad, Collin F., Brettschneider J., Simpson K., and Cope L. Quality assessment of affymetrix genechip data. In R. Gentleman, V. Carey, W. Huber, R. Irizarry, and S. Dudoit, editors, *Bioinformatics and computational biology solutions using R and Bioconductor*, chapter 3, pages 33–47. Springer International Publishing, New York, New York, 2005.

[185] Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47, 2015.

[186] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57(1):289–300, 1995.

[187] G. Laenen, L. Thorrez, D. Börnigen, and Y. Moreau. Finding the targets of a drug by integration of gene expression data with a protein interaction network. *Molecular BioSystems*, 9(7):1676–1685, 2013.

[188] D. Croft, A. F. Mundo, R. Haw, M. Milacic, J. Weiser, G. Wu, M. Caudy, P. Garapati, M. Gillespie, M. R. Kamdar, B. Jassal, S. Jupe, L. Matthews, B. May, S. Palatnik, K. Rothfels, V. Shamovsky, H. Song, M. Williams, E. Birney, H. Hermjakob, L. Stein, and P. D. Eustachio. The Reactome pathway knowledgebase. 42(November 2013):472–477, 2014.

[189] Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. KEGG: new perspectives on genomes, pathways , diseases and drugs. *Nucleic Acids Research*, 45(Database issue):353–361, 2018.

[190] S. Köhler, S. C. Doelken, C. J. Mungall, S. Bauer, H. V. Firth, I. Bailleul-Forestier, G. C. M. Black, D. L. Brown, L. G. Jackson, A. M. Kelly, J. A. Hurst, J. Ja, D. H. Ledbetter, S. Mansour, C. L. Martin, C. Moss, R. H. Scott, S. Sisodiya, S. Van Vooren, R. J. Wapner, N. De Leeuw, B. B. A. De Vries, N. L. Washingthon, and C. L. Smith. The Human Phenotype Ontology project : linking molecular biology and disease through phenotype data. 42(November 2013):966–974, 2014.

[191] J. T. Eppig, J. A. Blake, C. J. Bult, J. A. Kadin, and J. E. Richardson. The Mouse Genome Database (MGD): facilitating mouse as a model for human biology and disease. *Nucleic Acids Research*, 43(Database issue):726–736, 2015.

[192] M. Shimoyama, J. De Pons, G. T. Hayman, S. J. F. Laulederkind, W. Liu, R. Nigam, V. Petri, J. R. Smith, M. Tutaj, S.-J. Wang, E. Worthey, M. Dwinell, and H. Jacob. The Rat Genome Database 2015: genomic, phenotypic and environmental variations and disease. *Nucleic Acids Research*, 43(Database issue):743–750, 2015.

[193] G. Stelzer, N. Rosen, I. Plaschkes, S. Zimmerman, M. Twik, S. Fishilevich, T. I. Stein, R. Nudel, I. Lieder, Y. Mazor, S. Kaplan, D. Dahary, D. Warshawsky, Y. Guan-Golan, A. Kohn, N. Rappaport, M. Safran, and D. Lancet. The GeneCards suite: from gene data mining to disease genome sequence analyses. *Currents Protocols in Bioinformatics*, 54:1.30.1–1.30.33., 2016.

[194] S. Christakos, P. Dhawan, A. Verstuyf, L. Verlinden, and G. Carmeliet. Vitamin D: metabolism, molecular mechanism of action, and pleiotropic effects. *Physiological Reviews*, 96(1):365–408, 2016.

[195] L. Díaz, M. Díaz-muñoz, A. C. García-gaytán, and I. Méndez. Mechanistic effects of calcitriol in cancer biology. *Nutrients*, 7(6):5020–5050, 2015.

[196] D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, 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 Research*, 46(Database issue):1074–1082, 2018.

[197] P. L. Kovalenko, Z Zhang, M. Cui, S. K. Clinton, and J. C. Fleet. 1,25 dihydroxyvitamin d-mediated orchestration of anticancer, transcript-level

effects in the immortalized, non-transformed prostate epithelial cell line, rwpe1. *BMC Genomics*, 11:26, 2010.

[198] K. M. Simmons, S. G. Beaudin, C. J. Narvaez, and J. Welsh. Gene signatures of 1,25-dihydroxyvitamin D3 exposure in normal and transformed mammary cells. *Journal of Cellular Biochemistry*, 116(February):1693–1711, 2015.

[199] B.-Y. Bao, S.-D. Yeh, and Y.-F. Lee. 1alpha,25-dihydroxyvitamin D3 inhibits prostate cancer cell invasion via modulation of selective proteases. *Carciogenesis*, 27(1):32–42, 2006.

[200] A. Giangreco, A. A .and Vaishnav, D. Wagner, A. Finelli, N. Fleshner, T. Van Der Kwast, R. Vieth, and L. Nonn. Tumor suppressor microRNAs, miR-100 and -125b, are regulated by 1,25-dihydroxyvitamin D in primary prostate cells and in patient tissue. *Cancer Prevention Research*, 6(5):483–495, 2013.

[201] P. Aspenström. Integration of signalling pathways regulated by small GTPases and calcium. *Biochimica et Biophysica Acta*, 1742(1–3):51–58, 2004.

[202] Ellen M Smith and Vin Tangpricha. HHS Public Access. (404):1–14, 2016.

[203] S. Peleg. Responsiveness of osteoblastic and osteolytic bone metastases to vitamin D analogs. *Critical Reviews in Eukariotic Gene Expression*, 17(2):149–158, 2007.

[204] E. S. Björnsson. Hepatotoxicity of statins and other lipid- lowering agents. *Liver International*, 37(2):173–178, 2017.

[205] M. Hafner, P. Juvan, T. Rezen, and K. Monostory. The human primary hepatocyte transcriptome reveals novel insights into atorvastatin and rosuvastatin action. *Pharmacogenetics and Genomics*, 21(11):741–750, 2011.

[206] B. D. Roth. The discovery and development of atorvastatin, a potent novel hypolipidemic agent. *Progress in Medicinal Chemistry*, 40:1–22, 2002.

[207] M. Kuhn, D. Szklarczyk, S. Pletscher-Frankild, T. H. Blicher, C. Von Mering, L. J. Jensen, and P. Bork. STITCH 4: integration of protein-chemical interactions with user data. 42(Database issue):401–407, 2014.

[208] A. Cereto-massagué, M. José, C. Valls, M. Mulero, S. Garcia-vallvé, and G. Pujadas. Molecular fingerprint similarity search in virtual screening. *Methods*, 71:58–63, 2015.

[209] R. C. Glen, A. Bender, C. H. Arnby, L. Carlsson, and J. Smith. Circular fingerprints: flexible molecular descriptors with applications from physical chemistry to ADME. *IDrugs*, 9(3):199–204, 2006.

[210] David Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28:31–36, 1988.

[211] D. Weininger, A. Weininger, and J. L. Weininger. SMILES. 2. Algorithm for generation of unique SMILES notation. *Journal of Chemical Information and Computer Sciences*, 29:97–101, 1989.

[212] David Weininger. Smiles. 3. Depict. Graphical depiction of chemical structures. *Journal of Chemical Information and Computer Sciences*, 30:237–243, 1990.

[213] N. M. O. Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison. Open Babel: an open chemical toolbox. *Journal of Cheminformatics*, 3:33, 2011.

[214] J. L. Durant, B. A. Leland, D. R. Henry, and J. G. Nourse. Reoptimization of MDL keys for use in drug discovery. *Journal of Chemical Information and Computer Sciences*, 42(6):1273–1280, 2002.

[215] A. Bender, H. Y. Mussa, R. C. Glen, and S. Reiling. Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): evaluation of performance. *Journal of Chemical Information and Computer Sciences*, 44(5):1708–1718, 2004.

[216] A. Bender, H. Y. Mussa, R. C. Glen, and S. Reiling. Molecular similarity searching using atom environments, information-based feature selection, and a naïve Bayesian classifier. *Journal of Chemical Information and Computer Sciences*, 44:170–178, 2004.

[217] B. L. Roth, E. Lopez, S. Patel, and W. K. Kroeze. The multiplicity of serotonin receptors: uselessly diverse molecules or an embarrassment of riches? *The Neuroscientist*, 6(4):252–262, 2016.

[218] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

[219] T. Anastassiadis, S. W. Deacon, K. Devarajan, H. Ma, and J. R. Peterson. Comprehensive assay of kinase catalytic activity reveals features of kinase inhibitor selectivity. *Nature Biotechnology*, 29(11):1039–1045, 2011.

[220] M. I. Davis, J. P. Hunt, S. Herrgard, P. Ciceri, L. M. Wodicka, G. Pallares, M. Hocker, D. K. Treiber, and P. P. Zarrinkar. Comprehensive analysis of kinase inhibitor selectivity. *Nature Biotechnology*, 29(11):1046–1051, 2011.

[221] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe. KEGG for integration and interpretation of large-scale molecular data sets. *Database issue*, 40(Database issue):109–114, 2012.

[222] R. Caspi, T. Altman, R. Billington, K. Dreher, H. Foerster, C. A. Fulcher, T. A. Holland, I. M. Keseler, A. Kothari, A. Kubo, M. Krummenacker, M. Latendresse, L. A. Mueller, Q. Ong, S. Paley, P. Subhraveti, D. S. Weaver, D. Weerasinghe, P. Zhang, and P. D. Karp. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Research*, 42(Database issue):459–471, 2014.

[223] D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, 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 Research*, 46(Database issue):1074–1082, 2018.

[224] Y. Okuno, A. Tamon, H. Yabuuchi, S. Niijima, Y. Minowa, K. Tonomura, R. Kunimoto, and C. Feng. GLIDA: GPCR-ligand database for chemical genomics drug discovery - database and tools update. *Nucleic Acids Research*, 36(Database issue):907–912, 2008.

[225] S. Günther, M. Kuhn, M. Dunkel, M. Campillos, C. Senger, E. Petsalaki, J. Ahmed, E. G. Urdiales, A. Gewiess, L. J. Jensen, R. Schneider, R. Skoblo, R. B. Russell, P. E. Bourne, P. Bork, and R. Preissner. SuperTarget and Matador: resources for exploring drug-target relationships. *Nucleic Acids Research*, 36(Database issue):919–922, 2008.

[226] F. Zhu, Z. Shi, C. Qin, L. Tao, X. Liu, F. Xu, L. Zhang, Y. Song, X. Liu, J. Zhang, B. Han, and P. Zhang. Therapeutic Target Database update 2012: a resource for facilitating target-oriented drug discovery. *Nucleic Acids Research*, 40(Database issue):1128–1136, 2012.

[227] A. P. Davis, C. J. Grondin, K. Lennon-Hopkins, C. Saraceni-Richards, D. Sciaky, B. L. King, T. C. Wiegers, and C. J. Mattingly. The Comparative Toxicogenomics Database's 10th year anniversary: update 2015. *Nucleic Acids Research*, 43(Database issue):914–920, 2015.

[228] P. Willett. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today*, 11(23–24):1046–1053, 2006.

[229] R. Kolde, S. Laur, P. Adler, and J. Vilo. Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics*, 28(4):573–580, 2012.

[230] J. M. Stuart, E. Segal, D. Koller, and S. K. Kim. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302(5643):249–255, 2003.

[231] S. Aerts, D. Lambrechts, S. Maity, P. Van Loo, B. Coessens, F. De Smet, L.-C. Tranchevent, B. De Moor, P. Marynen, B. Hassan, P. Carmeliet, and Y. Moreau. Gene prioritization through genomic data fusion. *Nature Biotechnology*, 24(5):537–544, 2006.

[232] T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I. F. Kim, A. Soboleva, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, R. N. Muertter, and R. Edgar. NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Research*, 37(Database issue):885–890, 2009.

[233] A. Bender, J. L. Jenkins, J. Scheiber, S. C. K. Sukuru, M. Glick, and J. W. Davies. How similar are similarity searching methods? A principal component analysis of molecular descriptor space. *Journal of Chemical Information and Modeling*, 49(1):108–119, 2009.

[234] E. Gregori-puigjané, V. Setola, J. Hert, B. A. Crews, J. J. Irwin, and E. Lounkine. Identifying mechanism-of-action targets for drugs and probes. *Proceedings of the National Academy of Sciences*, 109(28):11178–11183, 2012.

[235] C. Laggner, D. Kokel, V. Setola, A. Tolia, H. Lin, J. J. Irwin, M. J. Keiser, C. Y. J. Cheung, D. L. Jr. Minor, B. L. Roth, R. T. Peterson, and B. K. Shoichet. Chemical informatics and target identification in a zebrafish phenotypic screen. *Nature Chemical Biology*, 8(2):144–146, 2011.

[236] N. M. O'Boyle and R. A. Sayle. Comparing structural fingerprints using a literature-based similarity benchmark. *Journal of Cheminformatics*, 8:36, 2016.

[237] B. Verbist, G. Klambauer, L. Vervoort, W. Talloen, QSTAR Consortium, Z. Shkedy, O. Thas, A. Bender, H. W. Göhlmann, and S. Hochreiter. Using transcriptomics to guide lead optimization in drug discovery projects: lessons learned from the QSTAR project. *Drug Discovery Today*, 20(5):505–513, 2015.

[238] W. Zhou, Y. Wang, A. Lu, and G. Zhang. Systems pharmacology in small molecular drug discovery. *International Journal of Molecular Sciences*, 17(2):246.

[239] Xiaoyu Qing. Pharmacophore modeling: advances, limitations, and current utility in drug discovery. *Journal of Receptor, Ligand and Channel Research*, 2014(7):81–92, 2014.

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

[241] Joaquin Dopazo. Genomics and transcriptomics in drug discovery. *Drug Discovery Today*, 19(2):126–132, 2014.

[242] J.G. Moffat, F. Vincent, A.L. Jonathan, J. Eder, and F. Prunotto. Opportunities and challenges in phenotypic drug discovery: an industry perspective. *Nature Reviews Drug Discovery*, 16(8):531–543, 2017.

[243] R. Heilker, U. Lessel, and D. Bisschof. The power of combining phenotypic and target-focused drug discovery. *Drug Discovery Today*, 24(2):526–532, 2019.

[244] G.C. Terstappen, C. Schlupen, R. Raggiaschi, and G. Gaviraghi. Target deconvolution strategies in drug discovery. *Nature Reviews Drug Discovery*, 6(11):891–903, 2007.

[245] R. Noberini and T. Bonaldi. Epigenetic drug target deconvolution by mass spectrometry–based technologies. *Nature Structural & Molecular Biology*, 26:854–857, 2019.

[246] 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, 2017.

[247] L. David, J. Arús-Pous, J. Karlsson, O. Engkvist, E.J. Bjerrum, T. Kogej, J.M. Kriegl, B. Beck, and H. Chen. Applications of deep-learning in exploiting large-scale and heterogeneous compound data in industrial pharmaceutical research. *Frontiers in Pharmacology*, 10:1303, 2019.

[248] T. De Bie, L. C. Tranchevent, L. M. van Oeffelen, and Y. Moreau. Kernel-based data fusion for gene prioritization. *Bioinformatics*, 23(13):125–132, 2007.

[249] G. R. G. Lanckriet, P. Bartlett, and M. I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5:27–72, 2004.

[250] J. Simm, A. Arany, P. Zakeri, T. Haber, J. K. Wegner, V. Chupakhin, and H. Ceulemans. Macau: scalable Bayesian multi-relational factorization with side information using MCMC. *2017 IEEE International Workshop on Machine Learning For Signal Processing*, 2017.

[251] P. Zakeri, J. Simm, A. Arany, S. Elshal, and Y. Moreau. Gene prioritization using Bayesian matrix factorization with genomic and phenotypic side information. *Bioinformatics*, 34(13):447–456, 2018.

[252] J. Simm, G. Klambauer, Á Arany, M. Steijaert, J. K. Wegner, E. Gustin, V. Chupakhin, Y. T. Chong, J. Vialard, P. Buijnsters, I. Velter, A. Vapirev, S. Singh, A. E. Carpenter, R. Wuyts, S. Hochreiter, Y. Moreau, and H. Ceulemans. Repurposing high-throughput image assays enables biological activity prediction for drug discovery. *Cell Chemical Biology*, 25(5):611–618, 2018.

[253] M. Yang, J. Simm, C. C. Lam, P. Zakeri, G. J. P. Van Westen, Y. Moreau, and J. Saez-Rodriguez. Linking drug target and pathway activation for effective therapy using multi-task learning. *Scientific Reports*, 8(1):8322, 2018.

List of publications

International journal papers (published)

G. Laenen, L. Thorrez, D. Börnigen, and Y. Moreau. Finding the targets of a drug by integration of gene expression data with a protein interaction network. *Molecular BioSystems*, 9(7):1676-1685, 2013.

G. Laenen, A. Ardeshirdavani, Y. Moreau, and L. Thorrez. Galahad: a web server for drug effect analysis from gene expression. *Nucleic Acids Research*, 43(W1):W208-212, 2015.

A. Rochtus, R. Winand, G. Laenen, E. Vangeel, B. Izzi, C. Wittevrongel, Y. Moreau, C. Verpoorten, K. Jansen, C Van Geet, and K. Freson. Methylome analysis for spina bifida shows *SOX18* hypomethylation as a risk factor with evidence for a complex (epi)genetic interplay to affect neural tube development. *Clinical epigenetics*, 8:108, 2016.

A. Yakubova, L. Thorrez, D. Svetlichnyy, L. Zwarts, V. Vulsteke, G. Laenen, W. Oosterlinck, Y. Moreau, L. Dehaspe, J. Van Houdt, Á. Cortés-Calabuig, B. De Moor, P. Callaerts, and P. Herijgers. ACE-inhibition induces a cardioprotective transcriptional response in the metabolic syndrome heart. *Scientific Reports*, 8(1):16169, 2017.

B. Verbist, E. Adriaensen, V. Keersmaekers, D. Putri, M. Crabbe, M. Derks, R. Bagdziunas, G. Laenen, and H. De Wolf. Analyzing magnetic bead QuantiGene® Plex 2.0 gene expression data in high throughput mode using QGprofiler. *BMC Bioinformatics*, 20(1):378, 2019.

M. Vanhaverbeke, M. Vausort, D. Veltman, L. Zhang, M. Wu, G. Laenen, H. Gillijns, Y. Moreau, J. Bartunek, F. Van De Werf, Y. Delvaux, S. Janssens, and P. R. Sinnaeve. Peripheral Blood RNA Levels of *QSOX1* and *PLBD1* are New Independent Predictors of Left Ventricular Dysfunction after Acute Myocardial

Infarction. *Circulation: Genomic and Precision Medicine*, published online ahead of print, 2019.

International journal papers (submitted)

A. Yakubova, L. Thorrez, D. Svetlichnyy, G. Van der Mieren, W. Oosterlinck, L. Zwarts, G. Laenen, Y. Moreau, L. Dehaspe, J. Van Houdt, B. De Moor, P. Callaerts, and P. Herijgers. Pathways Underlying Impaired Cardiac Preconditioning Potential in Subjects with the Metabolic Syndrome, and its Restoration by ACE-Inhibition. Under review with *Scientific Reports*.

International conference papers

C. Moschopoulos, G. Laenen, G. Kritikos, and Y. Moreau. Applying kernel methods on protein complexes detection problem In: C. Jayne, S. Yue, and L. Iliadis (eds) *Engineering Applications of Neural Networks, EANN 2012, Communications in Computer and Information Science*, vol 311, Springer, Berlin - Heidelberg, 2012.



FACULTY OF ENGINEERING SCIENCE
DEPARTMENT OF ELECTRICAL ENGINEERING
STADIUS CENTER FOR DYNAMICAL SYSTEMS, SIGNAL PROCESSING AND DATA ANALYTICS
Kasteelpark Arenberg 10, bus 2446
B-3001 Leuven
http://www.esat.kuleuven.be/STADIUS