KU LEUVEN

ARENBERG DOCTORAL SCHOOL
Faculty of Engineering Science

Data integration in drug-target interaction prediction in drug discovery and drug repositioning

Exploiting the potential of structure-based contribution in machine learning -based drug repositioning

Daniele Parisi

Supervisors:
Prof. Yves Moreau
Prof. Michael Schroeder
(TU-Dresden)

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Science (PhD): Electrical Engineering

March 2021

Data integration in drug-target interaction prediction in drug discovery and drug repositioning

Exploiting the potential of structure-based contribution in machine learning -based drug repositioning

Daniele PARISI

Examination committee:
Prof.Paul Sas, chair
Prof. Yves Moreau, supervisor
Prof. Michael Schroeder, supervisor
(TU-Dresden)
Prof.Arnout Voet
Prof.Ben Sprangers
Dr. Piet Herdewyn
Prof.Olga Kalinina
(Helmholtz Centre)
Dr. Vladimir Chupakhin
(The Janssen Pharmaceutical Companies)

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Science (PhD): Electrical Engineering

March 2021

© 2021 KU Leuven – Faculty of Engineering Science
Uitgegeven in eigen beheer, Daniele Parisi, Kasteelpark Arenberg 10, 3001 Leuven, Belgium, 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 and acknowledgment

My research stems from my passion for pharmacology and my interest in computer science. As the world moves further into the era of artificial intelligence, generating and storing a vast amount of biomedical data, there will be a greater need to make the right use of them. How will we take advantage of this content to faster and cheaper develop small-molecule drugs? How will different expertise be able to join forces to overcome the significant challenges of the present and future drug discovery? My ambition to become an interdisciplinary scientist and the dream of discovering different cultures around my continent pushed me to build the international and multidisciplinary network that made this work possible.

In truth, I could not have achieved my results without the support I received from all the people I met along this journey. First of all, I want to thank my parents, Nino and Gina, and my sister, Stefania, who supported me with love and understanding. I would have never got so far without their help. For years they have been dealing with airports, time zones, and video-calls waiting for me to be back. Like in most Italian families, I live surrounded by a dense galaxy of stars in the form of my cousins (Simona, Umberto, Isaura, Martina, Gianni, Vania, Ilaria, Alfonso e Dario), aunts (Giovanna, Rosaria, Rita, Anna e Marianna), uncles (Marco, Salavatore, Gennaro, Beniamino, and Giulio), my grandma (Anna), my niece (Chiara), bother- and sister-in-law (Mimmo and Alessandra) who are constantly heating me with their love-irradiating hearts. I especially want to dedicate this work to the stars of this galaxy that shine inside me with their bright lights after their expiration: Beniamino, Giovanna, Giovanni, Rachele, and Felice.

Then, I want to thank my friends spread all around the globe. They are small pieces of a pulsing puzzle that depicts my whole existence. Some of them watched me growing up and nurturing some of the dreams that led me to where I am now. Others are relatively new to my life, as a refreshing ocean wave in a torrid south European summer. To each one of you, thank you from the bottom

of my heart: "La Grande Famiglia", "Capracotta", "Mykonos", "ParisPD" and all the others in Italy (Ambra, Ciccio, Davide, Fefe, Donny, Mincia, Mirko, Silvia, Mostro, Francesca, I Tranfas, Pulcetta, Valerione, Sara, Squarcia, Princiotta, Ludo, Eleonora, La Fra, Matte, Kevin, Ale, Chiara, Paola, Marta, Roberta, Irene, Federica, Michela, Roberto, Rossella, Annamaria); "Leuven dream team", "Culitos" and all the others in Belgium (Valentina, Martijn, Andreas, Kaoru, Puya, Alessandra, Agnese, Mario, Franco, Melania, Elisabetta, il buon Lorenzo, Claudia, Tommaso, Gianluca, Gabriele, Martina, Bruno, Lars, Panteleimou, Marianna, Carlito, Julien, Lazaros, Manuelito, Nikos, Paco, Toni, Lazaros, Juan and Brunito); "BioinfoPT", "Cottaer Fische" and the others in Germany (Anna, Leo, Julia, Chris, Antigoni, Melissa, Aldo, Claudio, Greta, Andrea, Alesandro, Ceciel, Francesco, Anastasia, Flo E., Flo T., Fra, Matt, Max, Alex and Toni); "Meffrelab" and "Food" in the USA (Cammy, Elif, Niko, Nick, Priya, Miriam, Naz, Sally, Leslie, Ruchi, Rick, Sammy, Spencer, Oli, Dan, Chase, Zach, Silvie, Andrew, Mauro), as well as my current supervisor Eric Meffre.

A particular mention goes to someone who gave me the chance to start a new exciting chapter of my career and supported me through this tough last year far from home. Thank you, Niko. It would not be the same without you.

An extra thank you to Rick and Cammy, two special persons who helped me proof-reading this manuscript and supported me a lot during the last period of my Ph.D.

Last but not least, I want to thank my supervisors, prof. Yves Moreau and prof. Michael Schroeder, the members of my Ph.D. committee, the KULeuven and TU-Dresden secretaries (Ida, Elsy, Leen, Hanna, John, Wim, Anne, and Gloria), and all my colleagues that helped me develop the knowledge and the critical eye I needed to pursue this doctorate degree (Amin, Pooya, Haleh, Edward, Daniele, Nasim, Jaak, Griet, Melissa, Mahmood, Negin, Sarah, Ali, Florian, Joachim, Sebastian, Jörg, Gary).

I want to dedicate my thesis two prof. Roberta Misasi and prof. Fiorella De Vito who encouraged me to use my virtues to pursuit knowledge wherever it lies.

La Divina Comemdia, Inferno XXVI - Dante Alighieri

'O frati,' dissi, 'che per cento milia
perigli siete giunti a l'occidente,
a questa tanto picciola vigilia
d'i nostri sensi ch'è del rimanente
non vogliate negar l'esperïenza
di retro al sol, del mondo sanza gente.
Considerate la vostra semenza:
Fatti non foste a viver come bruti,
ma per seguir virtute e canoscenza.'

The Divine Comedy, Inferno XXVI - Dante Alighieri

'O brothers!' I began, 'woe to the west
'Through perils without number now have we reach'd;
'To this the short remaining watch, that yet
'Our senses have to wake, refuse not proof
'Of the unpeopled world, following the track
'Of Phoebus. Call to mind from whence ye sprang:
Ye were not form'd to live the life of brutes
'But virtue to pursue and knowledge high.'

Abstract

Identifying drug-target interactions is a crucial step in drug repositioning, intended as the process of suggesting new indications for known drugs. Since the 9000 FDA-approved and experimental small molecule drugs and more than 500.000 protein records available, performing in vitro experiments to check all the putative drug-target couples would be dramatically too expensive and time-consuming. Therefore computational techniques might help to predict compound biological activity (IC50) and suggest new putative medical indications for existing drugs. Machine learning techniques such as Bayesian matrix factorization and deep neural network can integrate structural information of drugs, proteins, and their binding to better predict the biological activity and suggest new drug-target interactions, with a significant impact on the drug discovery process. Different types of side information, such as chemical structures of the drugs and 3D structures of the protein targets, can help the prediction process. In this work, I analyse the contribution of different kinds of heterogeneous information in the prediction process, taking into account different modalities of validation and advantages and difficulties related to the application of each specific type of data.

Beknopte samenvatting

Goedgekeurde geneesmiddelen hebben favoriete of gevalideerd farmacokinetische eigenschappen en toxicologische profielen, en de herpositionering van bestaande geneesmiddelen voor nieuwe indicaties kunnen potentieel dure kosten in verband met een vroeg stadium het testen van de hit verbindingen te vermijden. In de afgelopen jaren de technologische vooruitgang in de virtuele screening methoden hebben toegestaan medicinale chemici om snel te screenen drug bibliotheek voor therapeutische werkzaamheid tegen nieuwe biomoleculaire doelen. Hoewel de combinatie van virtuele screening en herbestemming mogelijk is een zeer efficiënte en kosteneffectieve strategie in drug discovery echter geen perfecte rekensoftware voor voorspelling beschikbaar. In feite, elke benadering beschouwt slechts specifieke typologie van gegevens, zonder rekening te houden andere belangrijke informatie. Het doel van mijn project is om de snelheid van vals-negatieve en vals positieve te verminderen, en het verbeteren van de voorspelling kracht in drug herpositionering, door het fuseren van verschillende benaderingen. Gedurende de komende vier jaar ga ik om te studeren en structure-based, ligand-based en netwerk-gebaseerde technieken toe te passen voor het voorspellen van de interactie tussen bekende verbindingen en verschillende doelen, en ik zal alle resultaten met data-integratie methoden te verenigen, met het oog om nauwkeuriger resultaten te verkrijgen. Het protocol wordt toegepast op verschillende concrete gevallen, om nieuwe treffers interessante doelen, zoals sommige kinasen betrokken bij metastase en B-celactivering te vinden. Data-integratie is een wiskundig werktuig kan diepgaand inzicht in processen en kan worden uitgevoerd zoals meta-analyse heterogene gegevens uit verschillende experimenten integreren. In dit specifieke geval wordt matrixfactorizatie worden gebruikt voor het combineren van verschillende soort resultaten, verkregen uit drie verschillende voorspellende methoden: Ligand-based: ligand gelijkenis wordt bepaald door het vergelijken van 3D- en 2D-fingerprinting van de structuur van niet-verwante bekende verbindingen, met het oog op gemeenschappelijke kenmerken die kunnen leiden van de binding te vinden; Structuurgebaseerde: bestaat de vergelijking van vingerafdrukken

farmacofoor basis van de bindingsplaatsen, omdat de aanname dat soortgelijke bindingsplaatsen vergelijkbare liganden bindt; Netwerkgebaseerde: voorspelling op basis van de studie van het netwerk van interacties tussen eiwitten, verbindingen en genen, gebouwd op de bekende informatie, voorspellingen en differentiële expressie van functioneel gerelateerde genen.

List of abbreviations

3D-QSAR 3D Quantitative Structure-Activity Relationship

ABAT 4-aminobutyrate aminotransferase

AChE acetylcholinesterase

AD Alzheimer's Disease

ADMET Absorption, Distribution, Metabolism, Excretion and Toxicity

ALDH5A1 Succinate-semialdehyde dehydrogenase

ATC Anatomical Therapeutic Chemical

AUC Area Under the Curve

BACE1 beta secretasecleavage enzyme

BLM Bipartite Local Models

BPMF Bayesian Probabilistic Matrix Factorization

BuChE butyrylcholinesterase

CMap Connectivity Map

CoMFa Comparative Molecular Field Analysis

CTD Comparative Toxicogenomics Database

DBN Deep Belief Network

DL Deep Learning

DMSO Dimethylsulfoxide

DNILMF Dual-Network Integrated Logistic Matrix Factorization

DNN Deep Neural Network

DTI Drug Target Interaction

ECFP Extended-Connectivity Fingerprints

EENN End-to-End Neural Network

FDA Food and Drug Administration

GPCR G protein-coupled receptor

HDAC2 histone deacetylase 2

IBD Inflammatory Bowel Disease

KBMF2K Kernelized Bayesian Matrix Factorization with twin Kernels

KMDR Kernel Matrix Dimension Reduction Method

KronRLS Kronecker Regularized Least-Square

MAO monoamine oxidease

MDT Molecular Drug Targets

MeSH Medical Subject Headings

MF matrix factorization

MLR Mixed Lymphocyte Reaction

MSCMF Multiple Similarities Collaborative Matrix Factorization

NLM National Library of Medicine

NMDA N-methyl-D-aspartate receptor

NRLMF Neighborhood Regularized-based Logistic Matrix Factorization

NSCLC Non-Small Cell Lung Cancer

PBMC Peripheral Blood Mononuclear Cells

PDB Protein Data Bank

PLIP Protein-Ligand Interaction Profiler

PLS Partial Least Square methods

PMF Probabilistic Matrix Factorization

RDD Repurposed Drug Database

ReD Research and Development

RIN Residues Interaction Network

RLS Regularized Least Squares

RTK Receptor Tyrosine Kinases

SI Specificity Index

TI Therapeutic Index

Contents

Abstract v

Beknopte samenvatting vii

List of abbreviations ix

Contents xiii

List of Figures xvii

List of Tables xix

1 Introduction 1
- 1.1 Motivation . . . 1
  - 1.1.1 The negative trend in drug discovery . . . 1
  - 1.1.2 The incidence of new drug-target interaction discoveries in drug repositioning . . . 2
- 1.2 Aims . . . 2
- 1.3 Structure . . . 3
  - 1.3.1 Background . . . 3
  - 1.3.2 Analysis of the current drug repositioning . . . 3
  - 1.3.3 Example of PLIP-based drug repositioning . . . 3
  - 1.3.4 Data integration . . . 4
  - 1.3.5 Conclusions and future steps . . . 4

2 Background 5
- 2.1 Drug repositioning . . . 5
  - 2.1.1 Drug repositioning: an overview . . . 5
  - 2.1.2 Polypharmacology: beyond one-drug-one-target . . . 6
  - 2.1.3 Rational strategies to repurpose drugs . . . 7
- 2.2 Molecular-based approaches in computational drug repositioning 8

2.2.1 Ligand-based approaches in drug repositioning 8
2.2.2 Target-based approaches in drug repositioning 10
2.2.3 Interaction-based approaches in drug repositioning . . . 12
2.3 Ligand-protein interaction prediction: data-based approaches . 14
2.3.1 The big-data approach 14
2.3.2 Network-based methods in drug-repositioning 15
2.3.3 Machine learning in drug-repositioning 16
2.4 Online tools and databases for Drug-Target Interaction prediction 21
2.4.1 tools for Drug-Target Interaction prediction 21
2.4.2 tools for features extraction 21
2.4.3 Databases 21

3 Retrospective analysis of the current drug repositioning status 26
3.1 Drugs, targets and diseases connections in rational-based drug repositioning 26
3.2 Methodology 27
3.2.1 Identification of current drug repositioning cases 28
3.2.2 Retrieval of molecular drug targets data 28
3.2.3 Identification of targets involved in the repositioning cases 29
3.2.4 Filtering of non-small molecules and non-protein targets 29
3.2.5 Identification of disease-centric repositioning cases . . . 29
3.2.6 Target assignment to original and secondary indication . 29
3.2.7 Identification of target-centric repositioning cases 31
3.2.8 Identification of drug-centric repositioning cases 31
3.2.9 Structural data availability for drug-centric cases 31
3.3 Results 32
3.4 Discussion 35
3.4.1 More than one third of the cases does not fit the ”small molecule drug - protein target” definition 35
3.4.2 Drug repositioning is mostly disease- and target-centric 36
3.4.3 The role of drug-target interaction analysis in drug repositioning 36
3.4.4 Limit and potential of structure-based drug repositioning as drug-centric approach 37
3.4.5 Database selection and data availability 37
3.4.6 Ambiguous nomenclature of diseases 38
3.4.7 Issues of the classification process 38
3.5 Conclusion 39

4 Interaction profile similarity-based drug repositioning reveals new protein-ligand interactions for B-cell activation inhibition 44
4.1 PLIP-based drug repositioning approach for new kinase inhibitors 45
4.2 Target selection 45

4.3 Interaction profile similarity-based virtual screening 47
4.4 Kinase inhibition test confirmed part of the structure-based prediction 48
4.5 Binding affinity test revealed ibrutinib and suramin high affinity for VEGFR2 ATP binding pocket 49
4.6 Ibrutinib and suramin interaction profiles explained PLIP-based prediction 50
4.7 B-cell activation assay shows ibrutinib inhibits B-cell activation 50
4.8 Discussion on target selection process 52
4.9 Discussion on interactio-profile similarity based drug repositioning 53
4.10 Conclusion 55

5 Data integration for drug-protein interaction prediction 62
5.1 Biological activity for compound-protein couples 63
5.2 Side information: description and previous applications in drug-target interaction prediction 64
5.2.1 ECFP as ligand representation in machine learning . . . 64
5.2.2 Protein-based side information: amino acid frequency, sequence similarity, domain presence and residues interactions network (RIN) 66
5.2.3 PLIP for ligand-protein interaction profile representation 67
5.3 Algorithm description 68
5.4 Methodology 70
5.4.1 Software 70
5.4.2 ECFP-based binary fingerprints generation and validation 71
5.4.3 Protein-based side information: fingerprint generation . 72
5.4.4 PLIP-based binary fingerprint generation and validation 73
5.4.5 Data integration process, training and ranking 74
5.4.6 Parameter optimization 75
5.5 Results 76
5.5.1 Side information validation 76
5.5.2 Data fusion 76
5.6 Discussion 77
5.7 Conclusions on data fusion with structural side information . . 79

6 Conclusions 83

A Data fusion algorithm code 87

Bibliography 97

List of Figures

2.1 Drug-target-disease relationships in rational drug repositioning 9

3.1 Collection and classification of known repositioning cases 30
3.2 Frequency of repositioning cases among indication pairs 33

4.1 Overview of the study 46
4.2 Selection of targets and inhibitors 48
4.3 Chemical structures of known (A) and predicted (B) inhibitors 58
4.4 Ibrutinib and suramin are inhibitors of VEGFR2 59
4.5 Pairwise chemical similarity of known and predicted inhibitors . 61

5.1 Structural drug-repositioning with PLIP interaction fingerprints 69
5.2 Architecture of the deep-learning algorithm used for data fusion 70
5.3 Ligand-protein interaction system representation 71
5.4 Predictive performance using ECFP-based binary fingerprints and biological activity 77
5.5 Confusion matrices for biological activity prediction using PLIP fingerprints on each activity threshold 80
5.6 ROC curve for biological activity prediction using PLIP fingerprints on each activty threshold 81
5.7 Accuracy of ligand-protein interaction prediction 82

List of Tables

2.1 Examples of repurposed drugs 6
2.2 Examples of polypharmacological drugs 7
2.3 Examples of binding site comparison tools 12
2.4 Examples of binding site detection algorithms 13
2.5 List of tools for protein-ligand interaction profile analysis and comparison 15
2.6 List of network-based methods for drug-target interaction prediction 16
2.7 List of machine learning-based methods for drug-target interaction prediction 20
2.8 List of on-line tools for drug-target interaction prediction . . . 23
2.9 List of on-line tools for drug and target feature extraction . . . 24
2.10 List of on-line databases of ligand, protein and interaction features 25

3.1 Root MeSH tree key for the group of diseases and corresponding name 40
3.2 Disease-centric cases of repositioning 41
3.3 Taget-centric repositioning cases 42
3.4 Drug-centric repositioning cases 43

4.1 Targets identified in RNAi B-cell activation screen sorted by availability of structural data 56
4.2 List of new drug-target interactions predicted by interaction profile similarity and confirmed by experimental kinase activity assay 57
4.3 Potency on B-cell, B-cell Therapeutic Index TI, T-cell Therapeutic Index (TI), and Specificity Index (SI) 60

5.1 Summary of the analysed side information 64
5.2 List of machine learning-based works using ECFPs 66
5.3 Example of rows from PLIP database 74

5.4 Explanation of PLIP bins 75
5.5 Protein-ligand interaction features and fingerprints 75
5.6 List of optimized parameters for each side information category 76
5.7 List of best AUCs per each side information category 78
5.8 Comparison of algorithm performances with and without interaction-based side information 79

Chapter 1

Introduction

1.1 Motivation

1.1.1 The negative trend in drug discovery

Over the last years, the productivity crisis of the pharmaceutical Research and Development (ReD) has raised concerns about the future of the field (153). According to Eroom's Law, drug discovery efficiency is declining, notwithstanding impressive scientific and technological advances (176). Taking a new chemical entity to the market has an average estimated cost of $2 to 3 billion (147) and a success rate of only 2% (233). Many hypotheses have been formulated to explain this negative trend, such as the research shift towards more difficult targets (153), the necessity to meet the standard of quality set by previously marketed drugs, and the tendency to overestimate the ability of advances in basic research (176). One strategy that might overcome these difficulties is drug repositioning (or drug repurposing), which exploits new indications for marketed drugs. Repurposing drugs offers several advantages, such as a lower risk of failure, a reduced time to market, and an expected average cost lowered to $300,000 (159). However, repurposing pipelines are often based on random observations or obstructed by the scarce data availability about diseases, drugs, and targets. Therefore, to fully exploit drug repositioning advantages, new systematic, rational approaches must be designed.

1.1.2 The incidence of new drug-target interaction discoveries in drug repositioning

Rational drug repositioning can be pursuit by following either a disease-centric approach, building a close relationship between an old and new indication, a target-centric one, by linking a known target and its established drug to a new indication, or a drug-centric approach, which connects a known drug to a new target and its associated indication. The three approaches differ in complexity, potential, and limits. According to our recent study described in Chapter 3 the majority of repositioned cases, more than 60%, falls within the disease-centric definition, around 30% within the target-centric, and only less than 10% within the drug-centric. The scarce occurrence of new drug-target interactions in current repositioned cases seems to reflect a general trend of the whole drug discovery world, where the cheapest and easiest opportunity with highest expected returns is exploited first, and less attractive opportunities with lower returns are left behind (the so-called "drilling-oil" approach) (190). Investing in techniques focused on the study of subtle characteristics of drugs, targets, and diseases, as well as on the refinement of systematic and efficient repositioning approaches, might help both the pharmaceutical R&D by generating more profit and patients by bringing new therapeutic tools in a fast and cheap way.

1.2 Aims

This thesis aims to contribute to a systematic ligand-protein interaction prediction approach to maximize the potential of rational drug repurposing. Each drug-centric repositioning technique uses different data to infer new drug-target interactions, showing peculiar strong and weak points. Different kinds of information have been integrated with a new machine learning approach trained to predict ligand-protein biological activity, to overcome the weaknesses, and emphasize the strengths related to each class of data. Such an approach allowed me to exploit the potential of the diverse kinds of data currently available on drugs and targets and the impact of data fusion in drug-target interaction prediction.

1.3 Structure

1.3.1 Background

In Chapter 2 the state of the art of drug repositioning and its common approaches are explained. First, some general concepts such as drug promiscuity and polypharmacology as well as an introduction on the three main repositioning approaches (disease-centric, target-centric, drug-centric) are treated in Section 2.1. Following, Section 2.2 descries some of the main techniques used to predict drug-target interactions for drug-centric repositioning based on information about ligand, target or their interactions (molecular-based drug repositioning), while Section 2.3 focus on the application of networks and machine learning in ligand-protein interaction prediction (data-based approaches). A list of databases and tools for drug-target interaction prediction and features extraction follows in Section 2.4 to give an idea of the current state of data and models availability for computational drug-repositioning.

1.3.2 Analysis of the current drug repositioning

In Chapter 3, a retrospective analysis of the currently available drug repositioning cases is described. After an introduction (Section 3.1), the methodology in Section 3.2 describes the approach used to carry out the analysis. The results are showed in Section 3.3, the discussion in Section 3.4 and the conclusion in Section 3.5.

1.3.3 Example of PLIP-based drug repositioning

In Chapter 4, a case study of structure-based drug repositioning is presented. The work is briefly described in Section 4.1, while in Section 4.2 the target selection process is shown, including methodology and results. The virtual screening based on drug-target interaction profile similarity is described in Section 4.3, and in Section 4.4 the experimental validation of the prediction through kinase inhibition test is shown, and the binding affinity tests between the predicted drugs and the new targets are shown in Section 4.5. In Section 4.6 the screening results are connected with the binding profile similarity of the best new binders, and in the end, Section 4.7 shows the results of B-cell activation tests for the new hit compounds. Major discussion points about the target selection process (Section 4.8) and the Protein-Ligand Interaction Profiler (PLIP)-based virtual screening (Section 4.9) are addressed in the

end. In Section 4.10 important conclusion about interaction-profile based drug repositioning is drawn.

1.3.4 Data integration

Chapter 5 describes the process of data fusion and the different information used. The biological activity dataset and the generation of the sparse matrix for the training and test set are described in Section 5.1. Each side information is described in Section 5.2, while the deep learning algorithm is presented in Section 5.3. In Section 5.4, the methodology of fingerprint generation, side information validation, and integration are described Section 5.5 shows the results of the evaluation process and data integration. Discussion and conclusions follow in Section 5.6 and Section 5.7.

1.3.5 Conclusions and future steps

In sight of the results of this work and previous research conclusions on the role of structural information for machine learning-based drug repositioning are presented in Chapter 6.

Chapter 2

Background

2.1 Drug repositioning

2.1.1 Drug repositioning: an overview

Drug repositioning aims to find new indications for existing drugs to reduce drug discovery costs and time. Investigating the efficacy of approved or failed drugs for new indications can overcome some of the obstacles that are usually behind the failure of the new-molecular-entity approach, such as the necessity to meet the standard of quality set by previously marketed drugs (176). Drug repositioning also represents a reasonable opportunity to find pharmacological tools against rare diseases (159; 84), and make personalized treatments more affordable (197). Drugs targeting rare diseases encounter difficult commercial conditions due to the lack of pathophysiological characterizations (159; 84), and personalized treatments have enormous difficulties in reaching the market because of the small size of trial cohorts (197). Currently, numerous successful stories of drug repositioning have already been reported, and many drugs are available on the market, as shown in Table 2.1 (19). A well-known example is sildenafil, first meant to treat hypertension and successively commercialized against erectile dysfunction (19). Undesired effects can also represent starting points for different indications. Sadly, such is the case of thalidomide, where the intense antiangiogenic activity (cause of numerous teratogenic side effects) became useful to treat multiple myeloma (19).

Drug name	Original indication	Secondary indication
Aspirin	Analgesia	Colorectal cancer
Atomoxetine	Parkinson disease	ADHD
Celecoxib	Pain and inflammation	Familial adenomatous polyps
Dapoxetine	Analgesia and depression	Premature ejaculation
Duloxetine	Depression	Stress Urinary Incontinence
Fingolimod	Transplant rejection	Multiple sclerosis
Ketoconazole	Fungal infections	Cushing syndrome
Minoxidil	Hypertension	Hair loss
Raloxifene	Osteoporosis	Breast cancer
Rituximab	Various cancers	Rheumatoid arthritis
Thalidomide	Morning sickness	Multiple myeloma
Topiramate	Epilepsy	Obesity
Zidovudine	Cancer	HIV/AIDS

Table 2.1: **Examples of repurposed drugs.** Over the last years, many known drugs reached the market with a new therapeutic indication, drawing attention to his potentials in the drug discovery process.

2.1.2 Polypharmacology: beyond one-drug-one-target

The exploitation of repositioning techniques for drug discovery is a consequence of the paradigm shift from "one drug, one target" to "one drug, multiple targets". The observation that drugs are often promiscuous and bind multiple molecular targets, a concept coined as polypharmacology, was once viewed as a disadvantage however, most recently, it has been viewed more favorably as its intrinsic property often expresses its many advantages (79). Polypharmacology has been used to explain the intrinsic therapeutic effects of many drugs that work better than other molecules acting on the same target. This can happen when the off-target interactions fall within the same pathway (or a closely related one) of the desired target, playing a supporting effect to the original indication. Many recent studies show evidence that many of them can suppress more than one biomarker, leading to the conclusion that their efficacy might be related to the polypharmacological action on different targets and pathways (Table 2.2) (17). However, insights on the promiscuous behaviour that generates off-target actions offer precious opportunities to develop new therapeutic tools. Understanding polypharmacology can help discover new off-targets for existing drugs and lead to new therapeutic indications with less effort than a new molecular entity drug discovery approach. Nevertheless, screening all the existing drugs against the product of more than 20,000 human genes is an enormous effort. Therefore, a rational approach based on the systematic rational-based integration of available

information about drugs, targets, and diseases is necessary to develop a feasible drug repositioning pipeline based on polypharmacology (162).

Drug name	**Targets**
Imatinib	ABL1, KIT, PDGFRA, PDGFRB
Crizotinib	ALK, ROS1, MET
Afatinib	EGFR, HER2
Ceritinib	ALK, ROS1
Dasatinib	ABL1, DDR2, SRC
Erlotini	EGFR, JAK2, HER2, HER3
Nilotinib	ABL1, KIT
Ponatinib	ABL1, FLT3, FGFR2, RET

Table 2.2: **Examples of polypharmacological drugs.** 8 FDA approved anticancer agents can suppress different biomarkers, showing a polypharmacological action on different targets. This promiscuous behaviour might be responsible for their pharmacological efficacy.

2.1.3 Rational strategies to repurpose drugs

The potential of promiscuous molecules in pharmacology became clear after medicinal chemists started focusing on tuning specific chemical groups to enhance side effects and turn them into main pharmacological effects. By focusing on a limited set of carefully chosen, structurally diverse drug molecules with bioavailability and toxicity studies that have already been performed, the approach called Selective Optimization of Side Activities (SOSA) aimed to be a first rational alternative to the blind high throughput screening (214). However, since structural modifications are essential parts of the SOSA approach, this strategy cannot be adequately considered a sort of drug repositioning, which instead only relies on the polypharmacological properties of the selected molecule (86). As shown in Figure 2.1, a rational drug repositioning approach can infer a new indication for known drugs by focusing on each of these three levels: disease, target, or drug. A focus on the disease (so-called disease-centric) is the most direct approach since it relies on the hypothesis that a drug's use can be expanded from the original to a closely related indication. As an example, consider nilotinib, a tyrosine kinase inhibitor, approved for the treatment of imatinib-resistant chronic myelogenous leukemia (89). A few years later, Novartis proposed the reposition of nilotinib to gastrointestinal stromal tumors. However, despite such commonalities, repositioning may not always succeed, and actually, Novartis' efforts to expand nilotinib to gastrointestinal stromal tumors were abandoned after a negative phase III

trial (28). By focusing on a target, target-centric repurposing builds on a novel link between a new indication and an established target, with its original drugs' repositioning. For example, the protein tyrosine kinase ABL was recently suggested as a novel player in Parkinson's disease (90), and therefore its inhibitors, such as nilotinib, may be effective against this syndrome (124). Drug-centric repositioning occurs when a novel drug-target interaction is established. For example, valproic acid is used in bipolar disorder and seizures because of its ability to hit the mitochondrial enzymes Succinate-semialdehyde dehydrogenase (ALDH5A1) and 4-aminobutyrate aminotransferase (ABAT). However, since its off-target interaction with the histone deacetylase 2 (HDAC2), and the role of this protein in many types of cancers, valproic acid has been hypothesized to induce differentiation, growth arrest, and apoptosis in cancer cells, leading to the repositioning to the treatment of neoplastic conditions such as familial adenomatous polyposis (82). Precise identification of new drug-target interactions is challenging, and many computational techniques have been developed to understand the polypharmacological behaviour of many attractive therapeutic agents. Such computational approaches have been generally classified into ligand-based approaches, target-based approaches, and machine learning-based methods (227). However, several new techniques have been published over the last years, and new efforts of accurate classifications have been made. In this work, I divide techniques focused on the similarity of one or few molecular features, so-called molecular-based approaches, from the predictive models built on the analysis of large, complex, and variegated sets of data, defined as data-based approaches.

2.2 Molecular-based approaches in computational drug repositioning

According to the new paradigm, pharmacological research shifted from focusing on a selective single-target interaction to building and understanding a more global and comprehensive multi-targeted network for each compound. In this section, an overview of the most known and used ligand-based, target(or structural)-based, and interaction-based approaches, defined as molecular-based, are discussed.

2.2.1 Ligand-based approaches in drug repositioning

Ligand-based repositioning approaches are based on the concept that similar ligands bind similar targets, and therefore similar pharmacological profiles derive

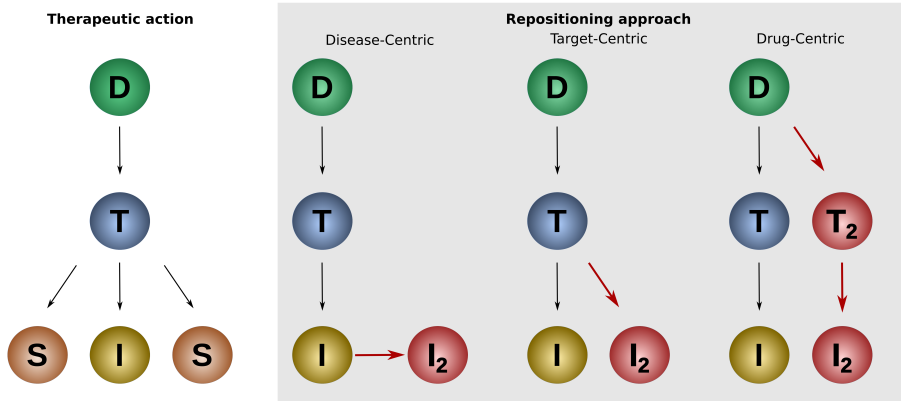


Figure 2.1: **Drug-target-disease relationships in rational drug repositioning.** Relations among drug (D), targets (T), and diseases/indications(I) according to the different rational drug repositioning concepts are shown. The pharmacological action of a drug (D) is usually explained through the interaction with a target (T), which might lead to a therapeutic indication (T) and several undesired side effects (S). In disease-centric, a drug application can be expanded from the original to a closely related indication. In target-centric, the identification of a new indication is linked to a well-established target, while in drug-centric, a new target is identified by linking the drug to a new indication.

from similarities in the chemical structure (130). In drug-target interaction prediction, ligand-based techniques are applied to predict the binding by comparing the candidate ligand with other compounds for which the activity on the putative protein target is known (225). Keiser et al. translated this assumption into the Similarity Ensemble Approach, a network of protein targets connected by the interactions with chemically similar ligands (93; 92). By increasing the complexity of ligand representation, the chemical entity can be described as a set of topological and electronic features responsible for binding to a particular target, which can subsequently be used to predict the binding of different compounds with similar features. The ensemble of features that characterizes a compound is called pharmacophore, and the selection process of candidates with similar pharmacophoric features is called pharmacophore screening. IUPAC defines pharmacophore as "an ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target and to trigger (or block) its biological response" (215). *Astolfi et al.* generated a ligand-based three dimensional pharmacophore model for inhibitors of the S. aureus protein NorA to repurpose drugs against this target of antibiotics resistance. The absence of structural information about

NorA and its interactions with ligands made it suitable for a pharmacophore-based investigation. As a result, three approved non-antibiotics drugs have been selected thanks to the 3D pharmacophore model generated on NorA ligands, and the in vitro test showed an ability to restore the effect of ciprofloxacin (20). Pharmacophoric features of small bioactive molecules can also be stored in databases, which are extensively used in drug repositioning to fish out compounds with different scaffolds, but that might show a similar binding profile. The Inte: Ligand's Pharmacophore Database (3) is among the first and most popular pharmacophore databases. It consists of a private-owned collection of 10,000 manually assembled pharmacophore models covering 300 clinically relevant pharmacological targets. It has been successfully used in ligand-protein interaction prediction to identify acetylcholinesterase, human rhinovirus coat protein, and cannabinoid receptors type 2 as novel targets for secondary metabolites from the medicinal plant Ruta graveolens (166). The advantages of ligand-based techniques are based on the ever-growing pools of chemical and biological information such as binding affinity, cellular activity, and Absorption, Distribution, Metabolism, Excretion and Toxicity (ADMET) data (130). The number of publicly accessible compound records is far greater than the number of deposited protein crystal structures, which makes ligand-based techniques suitable for a more considerable amount of cases (130). However, ligand-based techniques' performances strictly depend on the chemical space coverage of already known molecules (130). Moreover, a high overall chemical, pharmacophoric, or force field similarity does not necessarily guarantee activity on a secondary target since local structural divergences in chemical scaffolds can lead to "activity cliffs". (191)

2.2.2 Target-based approaches in drug repositioning

Protein-based drug repositioning approaches aim to link a relevant target with a biologically active compound based on similarity among 3D structures of different protein targets. Among the structure-based virtual screening approaches for drug repositioning, molecular docking is the most commonly used (107). It is applied to predict the binding affinity and the possible binding orientation of a stable ligand-target complex, starting from the single 3D structures of the target (obtained by crystallography or modeling) and the ligand (64). Many docking variants have been developed to optimize the process for all possible applications: rigid ligand and rigid receptor docking, flexible ligand and rigid receptor docking, flexible ligand and flexible receptor docking, ensemble docking, hybrid methods. The application of molecular docking in drug repositioning resulted in the identification of several drugs for stress incontinence, erectile dysfunction, obesity, psychosis, attention deficit disorder, cancer, irritable bowel

syndrome, smoking cessation, cardiovascular disease, Parkinson's disease, and Alzheimer disease (107). In particular, it is worth mentioning the case of the antipsychotic drug Benperidol, suggested to treat Alzheimer's Disease (AD) for its polypharmacological profile predicted by structure-based virtual screening. The candidate drug showed affinity to several important targets involved in AD: acetylcholinesterase (AChE), butyrylcholinesterase (BuChE), monoamine oxidease (MAO), N-methyl-D-aspartate receptor (NMDA), and beta secretasecleavage enzyme (BACE1) (108). Another case of docking-based drug repositioning is the anti-cancer sorafenib against the 5-HT2A receptor (65). Docking has also helped to discover many protein-protein interaction modulators, with a promising impact on cell proliferation, differentiation, and apoptosis, such as ezetimibe and darifenacin against TNF-a receptor interaction, which was subsequently confirmed in vitro (198). Despite the mentioned successful cases and the enormous popularity, molecular docking can suffer from drawbacks, just like all the other repositioning approaches. First and most important, docking requires 3D structural data, which might be available on specific databases such as the Protein Data Bank (PDB). Unfortunately, many pharmaceutically important human proteins are not available on PDB, such as most of the G protein-coupled receptor (GPCR) and transmembrane receptors (72). Moreover, many false positives are generated, and meticulous manual work from an expert hand is necessary to evaluate and select the right predictions (102). In the end, it can be challenging to automatically recognize a binding site, especially when the protein structure was crystallized without the ligand. A faster protein-based approach for drug repositioning, which does not require all the manual work and computational power necessary for the docking, is based on protein binding site similarity. As shown by *Haput et al.* (2013), drug promiscuity appears to correlate with structural binding site similarity of the different protein targets (71), and therefore identifying similarities among binding sites of different targets can generate hypotheses for new ligand-protein complexes. Binding site comparison approaches can be generally distinguished in alignment methods, based on the alignment of 3D structures or amino acid sequences, and alignment-free methods, which uses fingerprints or graphs to represent the binding pocket (71). Examples of binding site comparison tools are shown in Table 2.3. Repositioning pipelines based on docking or binding site similarity approaches have a common step: the binding site identification. A binding site can be either described by the presence of a ligand co-crystallized in a 3D complex or predicted according to specific criteria. Such criteria characterize different binding site identification methods; these are roughly divided into two major categories: geometric algorithms and energy-based methods (161). Some algorithms use a mixed geometric/energy-based approach, while others use a genomic scoring function to rank predicted pockets based on the degree of conservation (161). A representative list of tools for

binding site detection is shown in Table 2.4. An interesting example of drug-target interaction identified via protein-binding site similarity was described by *De Franchi et al.* (2010). Using a binding site three-dimensional comparison method, they identified the similarity between synapsin I, a protein involved in the regulation of neurotransmitter release, and the kinase product of the proto-oncogene Pim-1. After the repositioning hypothesis was generated and Pim-1 inhibitors were tested on synapsin I, staurosporine, and other three kinase inhibitors, they have been identified to bind the new target with nanomolar affinities (46). Difficulties in this approach are also general limitations of the other structure-based drug repositioning tools and come from pocket flexibility and structural data availability (161). To obviate those obstacles, homology modeling and manual selection of conformers could help, but those approaches are still not optimized for a large amount of data.

Name and Reference	**Year**	**Comparison Method**
FINDSITE (32)	2008	Combined evolution and structure-based
IsoCleft (143)	2008	Alignment-free graph-matching based
MultiBind (183)	2008	Multiple local alignment of spatial conformations
Park and Kim (154)	2008	Alignment-free graph-matching based
PROSURFER (136)	2008	Structural alignment
SiteAlign (178)	2008	Structural alignment
CPASS (158)	2006	Structural and sequence alignment
Query3d (21)	2005	3D functional annotation and structural comparison
SuMo (83)	2005	Alignment-free graph-matching based
SiteBase (30)	2004	Alignment-free graph-matching based
SiteEngine (182)	2004	Alignment-free, hashing triangles of physico-chemical properties
CSC (135)	2003	Local structure features alignment
CavBase (179)	2002	Alignment-free graph-matching based
eF-seek (96)	2002	Alignment-free, based on the triangulated solvent accessible surface

Table 2.3: **Examples of binding site comparison tools.** This figure shows a representative list of binding site comparison tools showing the approach used for the binding pocket identification. The names are sorted by year of publication.

2.2.3 Interaction-based approaches in drug repositioning

Another way to bridge protein-based methods and ligand-based methods is to compare the physicochemical interactions between a protein target and a small molecule ligand when bound together in a complex. By simplifying structural information contained within the description of ligand-protein interaction behavior, it is possible to perform a fast comparison to support virtual screening and drug repositioning (50). Notwithstanding the diversity existing among tools to encode and compare interactions (195), there are standard features for all these algorithms. First, the 3D structure of the protein-ligand complex

Name and Reference	Year	Identification Method
Fpocket (115)	2009	Geometric
AutoLigand (70)	2008	Energy based
PocketDepth (87)	2008	Geometric
ProSurfer (136)	2008	Geometric
SiteAlign (178)	2008	Geometric
PocketPicker (212)	2007	Geometric
SiteMap (68)	2007	Geometric and energy based
Q-SiteFinder (112)	2005	Energy based
SitesBase (30)	2004	Geometric
SiteEngine (182)	2004	Geometric
CavBase (179)	2002	Geometric
eF-Site (96)	2002	Geometric
CAST (119)	1998	Geometric
APROPOS (156)	1996	Geometric
SURFNET-ConSurf (111)	1995	Geometric with genomic ranking

Table 2.4: **Examples of binding site detection algorithms.** This figure shows a representative list of binding site prediction tools showing the approach used for the binding-pocket identification. The names are sorted by year of publication.

has to be provided, and the binding pocket where the ligand lies has to be identified. After that, the atomic coordinates' structural information at the interactive interface has to be simplified and converted into a format suitable for comparison. Those last steps are what characterizes the different tools. For most of the protein-ligand interaction methods, the encoding step utilizes pharmacophore models or fingerprints Table 2.5. In the pharmacophore-based drug-target interaction tools, the features that enable a ligand to interact with a target in a specific binding mode are represented in their spatial arrangements. Lately, pharmacophoric maps of protein-ligand interactions have been stored in databases such as LigandScout (217), PharmaTargetDB (121), and web servers such as PharmMapper (121) and ZINCPharmer (99). These databases have been developed to compare a query set of interactions with the interaction profile previously stored in order to find similar binding behaviors. Regarding the fingerprint-based tools, the presence of each type of interaction between a target residue and the ligand is encoded in a binary format, resulting in a fixed-length series of numbers, called fingerprints. Fingerprints are easy to generate, manipulate and compare (using, for example, the Tanimoto coefficient), enabling a systematic analysis of large data set (50). For example, PLIP is the first web service to provide comprehensive analysis and visualization of protein-ligand interactions (170). It is a versatile tool that allows both manual inspection and the computational processing of interaction data. It takes an input of a 3D protein-ligand complex, and without any structure preparation, returns a

list of detected interactions among seven interaction types (hydrogen bonds, hydrophobic contacts, pi-stacking, pi-cation interactions, salt bridges, water bridges, and halogen bonds), including distance or angle constraints between arrangements of atoms. A final filtering step aims to eliminate redundant or overlapping interactions and favour the most biologically relevant contacts which can be translated into a machine-readable series of integers (171). More about PLIP characteristics is available in Section 4.2. PLIP has been applied to perform virtual screening in drug repositioning pipelines, identifying new drug-target couples such as amodiaquine-HSP27 (170) and ibrutinib-VEGFR2 (12). Interaction-based drug repositioning allows overcoming some crucial issues of both ligand-based techniques and target-based methods. For example, many comparison-based repositioning strategies struggle in reaching a certain level of chemical diversity when suggesting new drugs for particular targets or genetic diversity when repurposing a specific drug for different genes. PLIP-based repositioning pipelines instead showed the ability of drug-target interaction profile approaches to perform scaffold hopping for the new ligands predicted to bind the protein target (170; 12). Similarly, it is possible to suggest a drug to a new target not belonging to the same protein family. Furthermore, ligand-protein interaction profile similarity tools do not require either tedious preparation steps such as ligand or protein alignment nor manual results evaluation (such as in docking). However, ligand-protein interaction similarity techniques require 3D structures of ligand-protein complexes. That data is obtainable either with crystallography, which can be very hard for many pharmacologically interesting human targets such as GPCR proteins, or coupling homology modeling, docking, and molecular dynamics, which require high expertise, computational power, and not always generate models suitable for drug discovery purposes.

2.3 Ligand-protein interaction prediction: data-based approaches

2.3.1 The big-data approach

Notwithstanding the popularity of molecular-based approaches, the low availability of 3D structures and target-specific active ligands represent severe limitations for those drug-target interaction prediction techniques. Omics approaches such as genomics, transcriptomics, proteomics, and metabolomics generate biological data at an unprecedented level, and algorithms are developed to parse the data and make a prediction to infer new drug-target interactions. Those data-based approaches which try to address the limitations of molecular-

Name and Reference	Year	Binding Representation Method
PLIP (171)	2015	Fingerprint-based
PLIF (80)	2012	Fingerprint-based
Pharm-IF (175)	2010	Pharmacophore-based
AIF (23)	2010	Fingerprint-based
w-SIFt (144)	2009	Fingerprint-based
APIF (160)	2009	Fingerprint-based
IF-SFP (196)	2009	Fingerprint-based
Pharmacophore featured FP (211)	2009	Pharmacophore-based
IASF (45)	2008	Fingerprint-based
FLAP (22)	2007	Pharmacophore-based
Molecular IFP (204)	2007	Fingerprint-based
Knowledge-based IFP (140)	2006	Fingerprint-based
r-SIFt (48)	2006	Fingerprint-based
LigandScout (217)	2005	Pharmacophore-based
p-SIFt (42)	2005	Fingerprint-based
SIFt (47)	2004	Fingerprint-based
Expanded IFP (95)	2004	Fingerprint-based
Four-point pharmacophore method (131)	1999	Pharmacophore-based

Table 2.5: **List of tools for protein-ligand interaction profile analysis and comparison.** This figure shows a representative list of ligand-target analysis tools showing the approach used to describe the binding. The names are sorted by year of publication.

based methods are divided into network-based approaches and machine learning-based approaches (237).

2.3.2 Network-based methods in drug-repositioning

Biological network-based methods can take advantage of the massive amount of available -omics data to identify a new target for a drug. In drug-target networks, the nodes represent proteins, drugs, genes, or diseases, while the edges are used to show an existing association built on an aggregation of multiple types of pharmacological and clinically relevant pieces of evidence. That network allows inferring new informative associations via statistical and computational methods (125). Biological networks have already gained interest in drug discovery and drug repositioning (177), and many different network-based methods have been developed to predict drug-target interactions. Some of those methods are extended by applying similarity among the biological entities, outperforming the previous models (125). In contrast, others have the potential to unveil hidden or unknown drug-disease relationships based on the guilt-by-association principle by integrating heterogeneous data such as side-effects, transcriptional data, and signaling pathway data. (232). *Cheng*

et al. compared different approaches for drug-target interaction prediction and showed that the network-based method performed better than similarity-drug and-target based approaches (41). Table 2.6 shows some examples of network-based methods to predict drug-target interaction for drug repositioning. Limitations for network-based approaches are related to the current knowledge about the molecular interactomes. Furthermore, these interactomes provide only static snapshots of the biological systems which, in reality, are more like dynamic systems (125).

Short Name	Full Name	Reference
DTINet	Drug-target interaction network	Luo et al., 2017 (126)
NRWRH	Network-based Random Walk with Restart on the Heterogeneous network	Chen et al., 2012 (39)
DBSI	Drug-based similarity inference	Cheng et al., 2012 (41)
TBSI	Target-based similarity inference	Cheng et al., 2012 (41)
NBI	Network-based inference	Cheng et al., 2012 (41)
DSSI	Drug side-effect similarity-based inference method	Campillos et al., 2008 (34)
MTOI	multiple target optimal intervention solutions	Yang et al., 2008 (230)

Table 2.6: **List of network-based methods for drug-target interaction prediction.** This figure shows a representative list of tools for protein-ligand interaction prediction with network-based approaches. The names are sorted by year of publication.

2.3.3 Machine learning in drug-repositioning

Different algorithms, called machine learning, have been developed to learn from data and make predictions. The performance of such algorithms are related to the quantity and quality of data available and therefore are applied to solve problems for which a large amount of data are available (199). Regarding drug repositioning, machine learning can explore the known drug-target interactions and potentially repurpose approved molecules to new targets by using several public and private databases and different models. For example, machine learning approaches led to the identification and experimental validation of the anticonvulsant drug topiramate as a new agent to treat Inflammatory Bowel Disease (IBD) (53) and also suggested that the application of imipramine, bepridil, promethazine, and cimetidine inhibit the growth of Non-Small Cell Lung Cancer (NSCLC) (186).

Machine learning in drug repositioning: algorithms

The pool of machine learning algorithms is vibrant and continuously growing. Based on the level of inputs and the quality of the output produced, it's

possible to draw a general classification and divide supervised from unsupervised learning. Supervised methods are applied to predict future values of data categories or continuous variables by first training on known input and output data relationships. Unsupervised methods are instead used to identify hidden patterns that have not been specified by the user (199). Examples of machine learning algorithms for drug-target interaction prediction are: Bipartite Local Models (BLM), Regularized Least Squares (RLS), matrix factorization (MF), Deep Learning (DL), and other methods (237) Table 2.7. BLM consists in an extension of the concept of local models (29) to the bipartite network problem, and it's used in drug repositioning to learn and predict protein-compound interaction networks (33). In these supervised methods, the predictions are based on the proximity of drugs and targets in the interaction space called pharmacological space, which is generated by the integration of chemical similarity data about compounds (chemical space) and sequence similarity data about proteins (genomic space) (223). BLM shows that the integration of independent ligand-based prediction and target-based prediction has a better performance than the single approach (29). However, this approach is unable to learn in the absence of training data, and hence, it does not provide a reasonable prediction for drug/target candidates with no current known interactions (132). RLS is a framework for data-dependent regularization that exploits the geometry of the probability distribution (25). For drug-target interaction prediction, a semi-supervised natural extension of that frame, Laplacian Regularized Least Squares (220), has been developed to utilize both labeled and unlabeled data. It works by incorporating information from different biological spaces, such as chemical space, genomic space, and drug-protein interaction network space, by a multi-task kernel. Notwithstanding the good predictive performance, it is important to specify that the superiority of this model is clear only when most of the data are unlabeled (220). RLS has also been combined with Gaussian Interaction Profile kernel (203), based on the assumption that a drug, which shows a similar interaction pattern or non-interaction pattern with targets in a known drug-target interaction network, is likely to exhibit similar interacting behavior when finding new targets for the drug. Although this approach is applied to detect new interactions for a target or a drug for which at least one interaction has already been established, it is not suitable for discovering interactions for a new drug or target. (203). Successively, a Kronecker Regularized Least-Square (KronRLS)-based method has been developed as an extension of classical RLS to predict new drug-target interactions for unknown drugs or targets. For example, if a new drug is a query, the KronRLS is reduced to a QSAR model, in which no information about target similarities are needed (151). A further improvement of the RLS approach for drug-target interaction prediction is the Kernel Matrix Dimension Reduction Method (KMDR) and is based on the assumption that two similar entities tend to link similar nodes

to each other (103). About the third class, given a sparsely filled drug-target interaction matrix Y, F methods can predict the missing values and complete the matrix. (237). Based on Kernelized Bayesian Matrix Factorization with twin Kernels (KBMF2K), which uses one kernel function for chemical similarity and another kernel function calculated on protein sequences for genomic similarity, a fully probabilistic formulation has been developed considering drug-target interaction prediction a binary classification problem. It uses the only chemical similarity between drug compounds and genomic similarity between target proteins (66). Another method predicts drug-target interaction, grouping drugs according to their therapeutic effects, irrespective of their 3D shape similarity. Defined as the Probabilistic Matrix Factorization Probabilistic Matrix Factorization (PMF) method, the approach is based on a collaborative filtering algorithm and uses a probabilistic model with Gaussian noise (44). PMF has been successively extended to Bayesian Probabilistic Matrix Factorization (BPMF) by introducing common multi-variate Gaussian priors for the latent variables, one for the rows and one for the columns. Based on BPMF, a new method called Macau for factorization of heterogeneous data has been presented. It enables the incorporation of features (side information) for any entity and any relation. Macau has been applied on data-driven drug-target interaction prediction, showing the impact of circular fingerprints as side information on the prediction accuracy (184). Another formulation named Multiple Similarities Collaborative Matrix Factorization (MSCMF) is based on the weighted low-rank approximation for predicting drug-target interactions. It integrates relevant biological information, such as chemical structures and Anatomical Therapeutic Chemical (ATC) classification of drugs and genomic sequence, GO and protein-protein interaction network of targets, in the shape of multiple similarity matrices (235). Another approach called Neighborhood Regularized-based Logistic Matrix Factorization (NRLMF) method (122) integrates logistic matrix factorization with neighborhood regularization to predict the probability that a drug would interact with a target by representing drugs and targets as two latent vectors in the shared dimensional latent space. For each drug-target pair, the interaction probability is modeled by a logistic function of the drug-specific and target-specific latent vectors. Unlike the previous approaches, NRLMF assigns higher importance to positive observations of drug-target interaction since those are experimentally validated (122). NRLMF has also been extended with a Dual-Network Integrated Logistic Matrix Factorization (DNILMF) method to consider the drug-target profile information when building the model. Prediction resulted improved by employing the nonlinear diffusion technique among similarity matrices, assuming that similar drugs (or targets) may contribute to the accuracy of the predictions for their neighbors, following the "Social Trust Ensemble" idea borrowed from the social network's algorithms (69). Deep learning is probably one of the most promising and quickly growing approaches in machine

learning. With the dramatic increase of biological data, new powerful machine learning methods have been developed to deal with complex, heterogeneous, and high-dimensional data sets. Those algorithms, known as deep learning, have already been used in various pharmaceutical applications for drug-target interaction prediction (237). Deep learning is also referred to as Deep Neural Network (DNN), which is inspired by biological neural networks and consists of multiple processing interconnected layers. DNNs are composed of an input layer, an output layer, and at least one but usually several intermediate hidden layers. The parameters for each stage, called weights, are optimized via the backpropagation algorithm to capture high- or low-level transformed features of the original data (172). The peculiar characteristics of DNN allow exploiting large data sets to discover intricate structures that can improve drug-target interaction prediction performance (232). Deep Belief Network (DBN) is the first example of deep learning applied to drug-target interaction prediction. It infers potential interactions without classifying each target into different classes, resulting in relatively high prediction performance for drug repositioning (213). Another approach described by *Gao at al., 2018* has been designed to predict drug-target interactions directly from raw chemical structures and amino acids sequences (56). This End-to-End Neural Network (EENN) uses long short term memory recurrent neural networks and graph-based convolutional neural networks to project proteins and drug structures into dense vector space, generating highly interpretable results (56). In another case, a novel model has been designed, combining stacked autoencodings in deep learning models. This method infers potential unknown drug-target interactions on a genome-wide scale by integrating the sequence of the drug's target and the drug's molecular structure (208).

Weak points and Challenges

Notwithstanding the strong potential of machine learning approaches in drug-target interaction prediction, many limits have been identified, and some of them represent new challenges for improving the prediction process. The first limitation for all machine learning methods is the dependency on the training data. Since drug-target interaction databases can be sparse and imbalanced (237) bias can be present and inherited by the model, allowing for a negative impact on the prediction (203). Moreover, it seems very difficult for methods based on those training data sets to predict interactions for drugs and targets which do not have other recorded interactions. Therefore, it is difficult to imagine machine-learning being able to unlock alternative pharmacological paradigms (125). Also, due to the absence of experimentally-validated negative drug-target interactions, it is impossible to train the model with negative samples (237). A right approach against this problem might

Category	Short Name	Full Name	Reference
BLM	KRM	Kernel Regression Method	Yamanishi et al., 2008 (224)
	BLM	Bipartite Local Model	Bleakley et al., 2009 (29)
	BLM-NII	Neighbor-based Interaction-profile Inferring model BLM	Mei et al., 2012 (133)
RLS	LapRLS	Laplacian regularized least squares	Xia et al., 2010 (220)
	RLS-GIP	RLS combined with Gaussian Interaction Profile kernel	Tvan et al., 2011 (203)
	WNN	Weighted nearest neighbor	Tvan et al., 2013 (202)
	Kron-RLS	Kronecker Regularized Least-Square-based method	Pahikkala et al., 2014 (151)
	KMDR	Kernel matrix dimension reduction method	Kuang et al., 2017 (103)
MF	KBMF2K	ernelizedBayesian Matrix Factorization with twin Kernels	Gönen et al., 2017 (66)
	PMF	Probabilistic matrix factorization method	Cobanoglu et al., 2013 (44)
	MSCMF	Multiple Similarities Collaborative Matrix Factorization	Zheng et al., 2013 (235)
	NRLMF	Neighborhood Regularized-based Logistic Matrix Factorization	Liu et al., 2017 (122)
	DNILMF	Dual-Network integrated Logistic MatrixFactorization method	Hao et al., 2017 (69)
	Macau	Macau	Simm et al., 2015 (184)
DL	DeepDTIs	Deep Belief Network for DTI	Wen et al., 2017 (213)
	EENN	End-to-End Neural Network	Gao et al., 2018 (56)
	Stacked Autoencoder	Stacked Autoencoder	Wang et al., 2018 (208)

Table 2.7: **List of machine learning-based methods for drug-target interaction prediction.** This figure shows a representative list of algorithms for drug-target interaction prediction using machine learning approaches. For each category (bipartite local model BLM, regularized least squared RLS, matrix factorization MF and deep learning DL), the names are sorted by year of publication from the most recent.

be integrating heterogeneous information about drugs, targets, and diseases, using different points of view for the same phenomenon. Even though using heterogeneous data might improve the prediction power, most of the current machine-learning methods can only work optimally on homogeneous data (237). Another issue lies within the oversimplification, which often characterizes prediction models. Drug-target interactions are often reduced to a binary phenomenon, not considering other critical factors like quantitative affinities and molecule concentrations, which leads to over-fitting. Therefore, drug-target interaction prediction should be seen as a regression problem more than a binary problem (236). Another problem for many machine learning methods is poor interpretability. Difficulties in understanding how the model generates specific results is a recurrent issue, mostly in deep learning. Without any clear idea about the molecular mechanisms behind the predicted interactions, it might be challenging to convince interlocutors about the potential validity of the predicted drug-target couple (199).

2.4 Online tools and databases for Drug-Target Interaction prediction

2.4.1 tools for Drug-Target Interaction prediction

Based on the aforementioned molecular-based approaches or data-driven models, many automatic online tools and web-services have been designed to predict Drug Target Interaction (DTI) and very easily accessed (237). Some of them are based on 2D or 3D protein similarity (i.e., SwissTargetPrediction). Others use machine learning approaches (i.e., DrugE-Rank) or network analysis (i.e., DINIES). The choice of which tool to use is related to the user's experience and each specific pipeline's condition. A summary overview of some computational software for drug-target interaction prediction is presented here (Table 2.8).

2.4.2 tools for features extraction

Biochemical and pharmacological information about drugs, targets, and physicochemical interactions are essential features that can be used by molecular-based and data-driven prediction models. For that reason, several tools have been developed to extract from complex information such as biological assays and crystal structures, more machine-readable data, suitable for different predictive models (237). A list of some tools useful to generate data about drugs, targets, or interactions is presented here (Table 2.9).

2.4.3 Databases

The predictive power of the drug-target interaction prediction tools is extremely dependent on the availability and the quality of data. Expert curation is extremely valuable for the prediction process; it is always required (55). Several kinds of information are stored in the different databases available online, as shown in Table 2.10. One of those data types is experimental and predicted ligand-protein interactions with or without quantitative values such as binding affinities. In ChEMBL, for example, almost 14,000,000 activities of almost 1,500,000 compounds against more than 11,000 targets are stored (88). The Comparative Toxicogenomics Database (CTD) also contains nearly 10,000,000 gene-disease associations (63). Other databases are focused on a specific class of targets such as kinases, GPCRs, or ion channels. For example, BRENDA is the main enzyme and enzyme-ligand platform, storing information about 83,000 enzymes and 206,000 enzyme ligands (157). Another important type

of data stored in databases is the 3D structures of targets. For example, PDB stores more than 120,000 biological macromolecular structures such as proteins, DNA and RNA, obtained with different techniques (i.e., NMR or X-ray crystallography), in complex with or without a ligand bound to the structure (27). A subset of PDB, called sc-PDB, has been created to group protein binding sites in complex with ligand stored in PDB to facilitate the study of drug-target interaction and the drug discovery process. It contains more than 16,000 binding site-ligand complexes, obtained by combining more than 4700 proteins and 6300 ligands (49). A different kind of data useful for drug repositioning purposes is phenotypic information. Gene expression profiles of cells treated with drugs have been collected with the Connectivity Map (CMap) project (109). This repository contains 6100 gene expression profiles from almost 13,500 human genes and more than 13,000 bioactive small compounds. Also, adverse effects of drugs are powerful phenotypic information that might be used to predict new drug-target interactions. The database MetaADEDB contains more than 527,000 drug-side effect pairs used to build a drug-protein-phenotype network or train machine learning algorithms (40).

Name and reference	Description	Website
DINIES (226)	Uses supervised network inference to integrate heterogeneous biological data and predict unknown drug–target interaction networks	www.genome.jp/tools/dinies/
Drug2Gene (165)	It integrates the relational, compound, and genetic entities from source databases by merging actually identical entries	www.drug2gene.com
DrugE-Rank (234)	Uses machine learning combining feature-based and similarity-based methods with ensemble learning	http://datamining-iip.fudan.edu.cn/service/DrugE-Rank
DT-Web (14)	Based on bipartite network projection implementing resources transfer within the network, it uses domain-specific knowledge expressing drugs and targets similarity	http://alpha.dmi.unict.it/dtweb/
DTome (193)	Based on network-based approach that integrates drugs, drug primary targets, drug-associated genes, and target/gene functional associated proteins	http://bioinfo.mc.vanderbilt.edu/DTome
iGPCR-Drug (222)	A prediction engine for GPCR operated by fuzzy K-nearest neighbour algorithm, which takes 2D fingerprint of compounds and pseudo amino acid composition of the target	www.jci-bioinfo.cn/iGPCR-Drug/
PharmMapper (209)	Matches the query compound against an in-house pharmacophore model database	http://lilab.ecust.edu.cn/pharmmapper/
SDTNBI (218)	Network-based approach that incorporates network and chemoinformatics to bridge the gap between new chemical entities and known DTI network	http://lmmd.ecust.edu.cn/methods/sdtnbi/
SynSysNet (205)	Platform that creates a comprehensive 4D network of synaptic interactions for drug and targets.	http://bioinformatics.charite.de/synsysnet
SwissTargetPrediction (60)	Combines both 2D and 3D similarity measures with known ligands, provides results in five different species, and allows users to map predictions between and within organisms based on target homology	www.swisstargetprediction.ch
TargetNet (231)	Perform predictions based only on molecular structures, using Naïve Bayes models and multi-target SAR methodology.	http://targetnet.scbdd.com

Table 2.8: **List of on-line tools for drug-target interaction prediction.** Here A representative list of online web-servers and tools available for protein-ligand interaction prediction based on the combination of various aforementioned computational approaches is shown.

Name and reference	Description	Website
RDKit (110)	An open-source cheminformatics software that generates molecular descriptors for drugs to be used in machine learning, such as SMILES, 2D depiction, fingerprinting, chemical reactions, molecular serialization, similarity/diversity picking	www.rdkit.org/
ChemDes (52)	Integrates multiple state-of-the-art package to compute molecular descriptors and several different types of fingerprints	www.scbdd.com/chemdes/
OpenBabel (148)	An open chemical toolbox that can read and generate data from various areas including molecular modeling and biochemistry	http://openbabel.org/wiki/Main$_{P}age$
Rchemcpp (97)	It provides various important applications for drug development, such as finding structural analogs in Drugbank, ChEMBL, and Connectivity Map	http://shiny.bioinf.jku.at/Analoging/
PyDPI (36)	It computes features of proteins and drugs from amino acid sequences and chemical structures	https://sourceforge.net/projects/pydpicao/
Rcpi (35)	Used to represent complex molecules from proteins and drugs and complex interactions and to compute abundant physicochemical and structural features	http://bioconductor.org/packages/release/bioc/html/Rcpi.html
KeBABS (152)	An R package to analyze biological sequences including amino acid, DNA, and RNA sequences	http://www.bioinf.jku.at/software/kebabs/
PROFEAT (118)	It can extract protein features from network properties of protein-protein interaction network and amino acid sequences	http://bidd2.nus.edu.sg/cgi-bin/profeat2016/main.cgi
Pse-in-One (120)	It effectively generates key features of DNA, RNA or protein from its sequence	http://bioinformatics.hitsz.edu.cn/Pse-in-One/download/
ProtrWeb (221)	A R package providing various numerical representation of proteins and peptides from amino acid sequences	http://protrweb.scbdd.com/

Table 2.9: **List of on-line tools for drug and target feature extraction.** /hlHere a representative list of online web-servers and tools available for protein and ligand features extraction /hlis shown.

Category	Name and reference	Website
Drug-target interactions and binding affinities	BindingDB (61)	http://bindingdb.org/
	DrugBank (114)	www.drugbank.ca/
	ChEMBL (26)	www.ebi.ac.uk/chembldb/
	KEGG (88)	www.kegg.jp/kegg/
	ComparativeToxicogenomicsDatabase (CTD) (63)	http://ctdbase.org/
	Guide to PHARMACOLOGY (187)	www.guidetopharmacology.org/
	PDSP Ki database (168)	http://kidbdev.med.unc.edu/databases/pdsp.php
	TherapeuticTargetDatabase (TTD) (228)	http://bidd.nus.edu.sg/group/cjttd/
	STITCH (106)	http://stitch.embl.de/
	SuperTarget (73)	http://insilico.charite.de/supertarget/
	PROMISCUOUS (206)	http://bioinformatics.charite.de/promiscuous/
	NCGC Pharmaceutical Collection (81)	https://tripod.nih.gov/npc/
	ChemProt (101)	http://potentia.cbs.dtu.dk/ChemProt/
	BRENDA (157)	http://www.brenda-enzymes.org/
	PubChem BioAssay (210)	https://ncbi.nlm.nih.gov/pcassay/
	MATADOR (67)	http://matador.embl.de
3D structures	Protein Data Bank(PDB) (27)	www.rcsb.org/pdb/home/home.do
	MMDB (129)	www.ncbi.nlm.nih.gov/structure
	PDBbind (123)	www.pdbbind-cn.org/
	Binding MOAD (13)	http://bindingmoad.org/
	sc-PDB (94; 49)	http://bioinfo-pharma.u-strasbg.fr/scPDB/
Phenotypic information	SIDER (104)	http://sideeffects.embl.de/
	MetaADEDB (40)	http://lmmd.ecust.edu.cn/online$_s$*ervices*/*metaadedb*/
	Connectivity Map(CMap) (109)	www.broadinstitute.org/connectivity-map-cmap

Table 2.10: **List of on-line databases of ligand, protein and interaction features.** A representative list of online databases storing information about ligand, protein, and interaction features is shown here.

Chapter 3

Retrospective analysis of the current drug repositioning status

This chapter is based on the published work "Drug repositioning or target repositioning: A structural perspective of drug-target-indication relationship for available repurposed drugs", D Parisi, MF Adasme, A Sveshnikova, SN Bolz, Y Moreau, M Schroeder; Computational and Structural Biotechnology Journal 18, 1043, 2020. Parisi and Adasme have been in charge of managing the project as well as the writing part. Sveshnikova conducted the experiments of database mapping and network analysis. Bolz contributed on the revision of the manuscript while Moreau and Schroeder supervised the project. Schroeder conceived the experiments.

3.1 Drugs, targets and diseases connections in rational-based drug repositioning

Repositioning opportunities can be uncovered systematically by following either a disease-centric approach, as a result of a close relationship between an old and a new indication, a target-centric one, which links a known target and its established drug to a new indication, or a drug-centric approach, which connects a known drug to a new target and its associated indication. The three approaches showed in Figure 2.1 differ in their complexity, potential,

limitations, and, above all, in the required start information and computing power. At first glance, disease-centric repositioning may appear faster and more linear than target- and drug-centric repositioning. Indeed, a disease-centric repositioning hypothesis is based on a direct connection between drug and indication, thereby allegedly avoiding a deeper understanding of the physicochemical interactions between drug and therapeutic target. However, if there were always a direct connection, one cancer drug would cure all cancer forms. Instead, disease-centric approaches require a detailed understanding of the disease phenotype and the underlying molecular processes to seek novel indications. Also, disease-centric approaches may be affected by patents as the repositioning candidate, and patent claims usually protect the corresponding old indication. Hence, the commercial exploitation of a disease-centric repositioning needs to be closely coordinated with the related patent claims. Systematic approaches to disease-centric repositioning typically define numeric similarities of diseases. These approaches include comprehensive, computational comparisons of disease phenotypes and drug side effects (34; 105) and comparisons of gene expression profiles (194). In contrast to disease-centric approaches, target-centric repositioning approaches only search drugs where the old and new indications are of a different type. Therefore, it becomes less likely that patents of the drug already cover the new indication. However, a novel link from the target to a new indication is a rare finding. Consequently, these approaches are limited by the technologies available to uncover new target-disease associations. Besides screening methods such as deep sequencing, micro-arrays, and RNAi, which can provide clues to candidate targets, the target-centric approach requires a deep understanding of the molecular relationship between target and disease. Drug-centric repositioning can be considered the least direct approach because the drug is only linked to a novel indication via the discovery of a target already established for this indication. Since each repositioning approach has advantages and disadvantages, we performed a retrospective analysis to examine their distribution to the real cases of successful drug repositioning and to study the role of drug-target interaction prediction in drug repositioning. To address the question of which approach prevails in current drug discovery and what this implies for future developments, we systematically evaluated over 100 drugs, 200 target structures, and over 300 indications from the Drug Repositioning Database, and we classified each of them as belonging to one of the three repositioning approaches.

3.2 Methodology

The data presented above are the product of a retrospective analysis carried out on a list of repurposed drugs, their indications, and protein targets from

the Repurposed Drug Database (RDD), Molecular Drug Targets (MDT) and literature. The results are the product of a classification process explained in detail below, consisting of a stepwise division of the repurposed cases into a same-disease group and a same-target group.

3.2.1 Identification of current drug repositioning cases

The RDD was downloaded from www.drugrepurposingportal.com in January 2017. The database contains 438 drugs that have been repurposed and provides information about the drugs with their original and new indications. This includes drugs that have failed initially, approved, or experimentally associated with another original indication and have later been approved for a new indication. Given the above, it was necessary to integrate this data into another source providing information about the protein targets involved in both indications.

3.2.2 Retrieval of molecular drug targets data

Drug targets are often poorly defined in the literature, both for launched drugs and potential therapeutic agents in discovery and development. Several databases provide data of drug-target interactions, each with different scopes and focus, e.g. the Therapeutic Targets Database (150), Drugbank (216), SuperTarget (67), and the IUPHAR/BPS Guide (155), among others. Despite the variety of valuable online resources, it is still a challenge to retrieve a consistent and comprehensive view of the targets of approved drugs with their molecular efficacy and organized by therapeutic use. Especially when mapping the target concept to specific genes and gene products. However, *Santos et al* (174) presented a solution to the problem. They carried out a "comprehensive map of molecular drug targets" (MDT), where they curated 893 human and pathogen-derived biomolecules, which are affected by 1,578 US FDA-approved drugs. Santos data set was downloaded in .php format in January 2017, and a list of MDT was retrieved from it. Importantly, it was defined by *Santos et al.* that" biomolecules that the drug may also bind to, or be metabolized by, but which are not known to be responsible for its therapeutic effect, are not defined as targets." Moreover, the work defines a drug as any therapeutic agent, including small molecules and other biological agents that are used to enhance health.

3.2.3 Identification of targets involved in the repositioning cases

Data in RDD were merged with data in MDT (drug targets) based on the commonly known drug name, resulting in 196 drugs matched by their common names, 263 unique targets, and 333 unique indications (Figure 3.1). The merging of data was performed via a Python 2.7 script in April 2017 and complemented with other biologically relevant databases to enrich the analysis. In some cases, no protein target was stored for a drug in MDT, and a manual literature search was necessary.

3.2.4 Filtering of non-small molecules and non-protein targets

This study only considered proteins as biomolecule targets and small molecules as drugs. Other cases, such as antibodies as drugs and other biomolecules (e.g., enzymes or unknown) than proteins as targets, were removed. Out of the 196 drugs both in RDD and MDT, a total of 68 were removed under this filtering criteria.

3.2.5 Identification of disease-centric repositioning cases

A repositioning case is considered disease-centric when the repurposing is based on the similarity of the disease phenotype. For each repositioning case, the MeSH tree root key was assigned to each indication (see Table 3.1). The frequency of repurposed cases among diseases was calculated and visualized with the Matplotlib python library, and the number of cases for each pair of root MeSH keys was plotted in R with the ggplot2 package (see Figure 3.1). For each pair, all targets mentioned at the intersection of the same root MeSH keys were collected, and the number of targets at the intersection calculated. Literature evidence about the applicability of these targets for the group was also collected. Drugs linked to the same MeSH key were classified as cases of possible disease-centric repositioning.

3.2.6 Target assignment to original and secondary indication

To divide the targets into initial and secondary indications, we used literature evidence in PubMed (see Figure 3.1). Target-indication connections were retrieved manually from literature by searching for direct pieces of evidence (the improvement in a condition upon treatment with a particular drug due to action

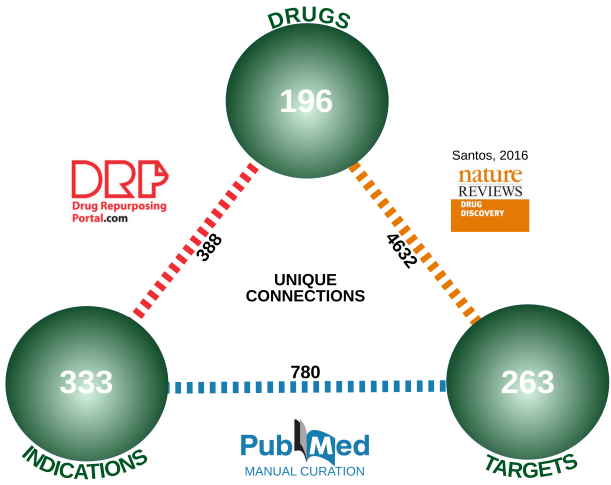


Figure 3.1: **Collection and classification of known repositioning cases.** The merging process of Repositioned Drug Database (RDD), containing 196 drugs and 333 indications linked through 388 connections, the Molecular Target Database from Santos (MTD), containing 4632 links between 196 drugs and 263 targets, and PubMed, which allowed to find 780 different links between 263 targets and 333 indications is shown here.

on the protein target); indirect evidence (the correlation of a condition with a particular target activity and a disease); or generalized evidence (single evidence of a target-condition link found even in the absence of a strong correlation). The validation for the previous manual-curation was done via text-mining using the ensemble biclustering algorithm (EBC), which allows the extraction of the connections from a natural text in a machine-processable form. The text mining dataset used in this work consists of 2 parts. Part I connects dependency paths to labels, or' themes.' They were introduced in this dataset to label what kind of interaction exists between two terms, e.g., whether a casual mutation has a role in the pathogenesis or promotes the disease's progression. The second part of the dataset contains information about drug-target-indication associations. To validate manual PubMed curation and estimate how applicable text-mining is for this aim in general, target-indication associations in this dataset were used. To extract the correct information, UniProt IDs were turned into gene IDs using UniProt API service, and MeSH on demand was used to assign indications IDs to the textual description of diseases. Afterward, all the records containing the genes of the proteins classified as therapeutic targets for drug repositioning were collected from the dataset, and it was checked whether the pair "gene ID - MeSH ID" is present. Part I and Part II of the text mining dataset were linked to identify a relationship between the gene and the drug target. In this way,

targets were again divided into initial and secondary indication. The resulting distribution was compared to manual distribution.

3.2.7 Identification of target-centric repositioning cases

A repositioning case is considered target-centric when the same protein target is exploited but in different pathological contexts. The remaining cases that were not classified as disease-centric repositioning were analyzed in order to determine whether the drug has two different protein targets related to the initial and secondary indications or not. Therefore, the cases for which the UniProt IDs were the same for the two indications were marked as target-centric repositionings. Even though some targets have different Uniprot IDs, they can be homologous targets with the same function in different organisms (ortholog). To perform a correct evaluation, target sequence identities were determined using Clustal Omega alignment (https://www.ebi.ac.uk/Tools/msa/clustalo), where the number of similar and identical positions was summarized and divided by the length of the alignment. According to the accuracy cutoff established by *Rost et al.* (167), targets at a sequence identity above 30% were classified as homologous targets and the repositioning case as target-centric.

3.2.8 Identification of drug-centric repositioning cases

A repositioning case is considered drug-centric when it exploits the drug's chemical properties. First, the cases that were not classified as disease-centric or target-centric were considered potentially drug-centric. Cases were also marked as drug-centric if there was a piece of literature evidence that the targets are different, but the comprehensive map did not mention the secondary target.

3.2.9 Structural data availability for drug-centric cases

We finally evaluated how feasible it is for the structure-based drug repositioning approach to identify the selected drug-centric cases as repositioning stories. Therefore we performed a mapping from Uniprot to PDB ID to search all available structures in Protein Data Bank, describing the binding between the drugs and their corresponding targets (associated with both indications). Also, structures of a drug with a target not described in the MDT were considered and evaluated under the terms of structure-based drug repositioning.

3.3 Results

Which of the three approaches dominates the field of drug repositioning? Is drug-target interaction prediction a driving force of drug repositioning? To address these questions, we analyzed all the repositioned small-molecule drugs that are active against a protein target and present in the RDD (2). We performed a classification of repurposed drugs according to the criteria specified in the Methods section. By this means, we determined the number of repositioning cases assigned to the drug-centric approach. It should be noted that other classification criteria could be applied to shed light on different characteristics, which would generate different results from the same dataset.

Current Drug repositioning set contains 196 known cases The merging of RDD with the Molecular Drug Targets (MDT) data led to a compiled report comprising 196 drug repositioning cases, 263 unique targets, and 333 unique indications. After removing the non-small molecules or non-protein targets cases, 128 repositioning cases constituted our classification's starting point.

The majority of repositioning cases (59%) were discovered via a disease-centric approach. To identify and characterize diseases that are susceptible to drug repositioning, we determined the number of repurposed drugs for each type of disease (results displayed in Figure 3.2). To this end, diseases were distinguished by the root Medical Subject Headings (MeSH) term key, which is a comprehensive, controlled vocabulary that provides a consistent way to retrieve information that may be described by variable terminology, thus facilitating indexing and searching. The MeSH vocabulary is organized into groups, of which one group are diseases (group C). The most common MeSH disease categories in the RDD are various neoplasms, immune system diseases, pathological signs and symptoms (clinical manifestations that can be either objective when observed by a physician or subjective when perceived by the patient), and nervous system diseases. Noteworthy, repositioning cases not only exist in MeSH group C (diseases) but also in groups E01 (Diagnosis), F02 (Psychological Phenomena and Processes), F03 (Mental Disorders), G08 (Reproductive and Urinary Physiological Phenomena), and G11 (Musculoskeletal and Neural Physiological Phenomena). The analysis of the "original indication - secondary indication" pairs for the small-molecule drugs (see Figure 3.2) revealed that the most exciting repositioning cases are combinations of bacterial infections (C01) and parasitic diseases (C02 and C03) with other types of diseases. In these cases, the repositioning occurred either to a homolog protein or to a completely different protein target. An example of repositioning to a homologous protein is the antimycotic drug ketoconazole, repositioned from a fungal target to the human

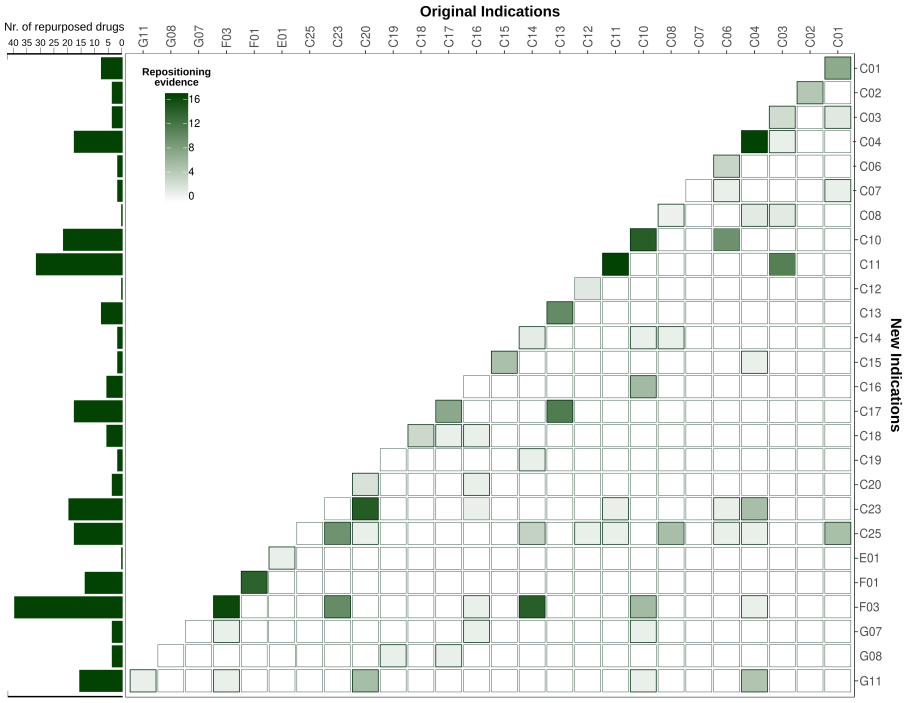


Figure 3.2: **Frequency of repositioning cases among indication pairs**. The disease classes have been plotted on both axes. The squares' color intensity expresses the number of repurposed drugs from one disease class to another. The darkest squares lay on the central diagonal, showing that most successful repositioning cases were discovered within the same disease class. On the left side, the number of repositioned drugs for the new indication is displayed.

homolog (Cytochrome P450) to treat cyclosporine-induced nephrotoxicity. On the other hand, Doxycycline is an example of repositioning to a distinct target as it has been repurposed from an inhibitor of bacterial 30S ribosomal proteins S4 and S9 to an inhibitor of human metalloproteinase to treat stomatognathic disease. However, such repositioning cases are rare. Strikingly, the main diagonal of the" original indication - secondary indication" heatmap (Figure 3.2) is the most populated, meaning that most drugs were repositioned within the same disease class. In total, 76 out of the 128 cases belong to the disease-centric repositionings. Most drugs (16 drugs) were repositioned from one type of neoplasm to another (C04 - C04). For instance, the kinase inhibitor nilotinib has been repurposed from the treatment of Philadelphia chromosome-positive chronic myelogenous leukemia to gastrointestinal stromal tumors (Table 3.2). The repositioning within immune system diseases (C20 - C20) is also very

common, with a total of 13 cases. For example, the steroid beclomethasone has been repositioned from the treatment of rhinitis to the intestinal graft-versus-host disease (Table 3.2). Among nervous system diseases (C10 - C10), eight repositioning cases were detected, including the repositioning of intravenous midazolam hydrochloride from preoperative sedation to epileptic seizure activity (Table 3.2). Five cases were identified within pathological conditions, signs and symptoms (C23 - C23), for example, the repurposing of aminocaproic acid from enhancing hemostasis to topical treatment of traumatic hyphema of the eye (Table 3.2). Based on phenotypical and handling similarities, the groups C10, C23, and F03 (mental disorders) could be combined to a large group of brain-related diseases and perception modification, which would form the most frequent disease category in disease-centric drug repositioning. The overall high proportion of disease-centric repositioning cases could be explained by therapeutic targets that play a crucial role in treating multiple similar diseases.

36% of the repositioning cases fall into the target-centric category. The drug targets were linked to original and secondary indications using data mining and literature information for the remaining 52 repositioned molecules (128 total small molecules minus 76 disease-centric cases). If the targets were the same for both indications or showed a protein sequence identity of at least 30% (167), the drug repositioning was classified as target-centric. Five target-centric repositioning cases were based on the binding of the drug to two homologous targets with the same function (ortholog) (see Table 3.3). The protein targets' calculated sequence identities were 49% for ketoconazole, 57% for eflornithine, 60% for dapsone, 65% for atovaquone, and 66% for trimetrexate. These values are much higher than the herein defined 30% similarity threshold. In total, 45 drug repositioning cases were identified as target-centric (Table 3.3). An example is chlorpromazine, whose interaction with the serotonin receptor HTR2A is involved in both the antiemetic/antihistamine indication (pathological conditions, signs, and symptoms) and the non-sedating tranquilizer action (mental disorder) (Table 3.3). As an inhibitor of cyclooxygenase-2, the non-steroidal anti-inflammatory drug celecoxib has originally been approved for the treatment of osteoarthritis and adult rheumatoid arthritis (immune system diseases) and has subsequently been repurposed to familial adenomatous polyposis (congenital, hereditary, and neonatal diseases) (Table 3.3).

5% of the repositioning cases were classified as drug-centric The remaining seven cases that could not be assigned to disease- or target-centric repositioning were classified as drug-centric repositioning. In these cases, the primary and secondary indications are linked to distinct protein targets. For instance, valproic acid, which has originally been developed to treat episodes in bipolar

disorder and seizures (nervous system diseases) by hitting the mitochondrial enzymes ALDH5A1 and ABAT, has been repurposed to the treatment of familial adenomatous polyposis (congenital, hereditary, and neonatal diseases) based on its interaction with HDAC2, as shown in Table 3.4. Allopurinol has been repurposed from a human protein target (Xanthine dehydrogenase/oxidase) to a parasitic target (Hypoxanthine phosphoribosyltransferase of Trypanosoma Cruzi) with low sequence similarity (6%).

None of the drug-centric repositioning cases had sufficient structural data to perform a drug-target interaction prediction Finally, we wished to assess the putative impact of structure-based drug-target interaction prediction on drug repositioning. Therefore, we searched the PDB for structures of the drug-centric repositioned drugs both in complex with the target linked to the original and the target linked to the secondary indication. However, in none of the seven cases, drug-target complex structures were available for both targets.

3.4 Discussion

3.4.1 More than one third of the cases does not fit the ”small molecule drug - protein target” definition

We focused our analysis exclusively on small-molecule drugs because they show more promiscuous binding behavior and form more defined interactions with their targets. Moreover, we only considered drug targets that are proteins. Altogether, 68 out of 196 repositioning cases were excluded from the analysis as they did not fit into the "small molecule drug-protein target" scheme. Non-small molecule drugs were typically antibodies, while non-protein targets included RNA, DNA, and other non-protein biomolecules. Examples of repositioned therapeutic antibodies are infliximab and adalimumab, used in both Crohn’s disease and juvenile rheumatoid arthritis. The database also included therapeutic proteins such as somatropin, which is used to treat children with growth disorders and induce ovulation in infertile women. An example of drugs that have a non-protein drug target is melphalan, which binds DNA and is applied in multiple myeloma and metastatic melanoma. DNA is also the target of cladribine, which is used to treat hairy cell leukemia and chronic lymphocytic leukemia.

3.4.2 Drug repositioning is mostly disease- and target-centric

TThe retrospective analysis of the repositioning cases present in the RDD database gave us an exciting picture of the current state of drug repositioning. Sixty percent of the repositioned drugs analyzed (76 cases out of 128) have been redirected to the same disease family. This tendency was particularly pronounced within two categories of therapeutic indications: neoplasms and immune system disorders (Figure 3.2), which also have the highest number of repositioning cases in the database. Thirty percent of the analyzed drugs (46 out of 128) have been repurposed to a different indication but the same protein target, as indicated by an identical Uniprot ID. Only 10% (seven cases) have been repositioned to a different disease and a different target. The situation described here seems to reflect a general trend in drug discovery, which is considered one reason for the structural crisis in pharmaceutical R&D mentioned in the introduction. The current pharmaceutical R&D situation has been compared to a drilling-oil process, where the cheapest and easiest options with highest expected returns are exploited first, and less attractive options are left behind (190). In terms of our results, this could mean that certain disease classes and rapid repositioning approaches within the same disease and target family are prioritized, leaving a big pool of drug-target-disease connections virtually unexplored. Therefore it might be crucial in the future to invest in drug repositioning techniques that focus on the subtle characteristics of drugs, targets, and diseases (drug-centric approaches), thereby overcoming the barriers defined by a disease category or target. A systematic and efficient repositioning approach that connects unrelated diseases and targets might benefit both the pharmaceutical R&D and the patients by increasing the profits and delivering novel therapeutic agents in a fast and cost-effective way.

3.4.3 The role of drug-target interaction analysis in drug repositioning

Although computational drug repositioning has lately developed many strategies for predicting drug-target interactions, our analysis shows that most of the actual repositioning cases could result from a disease-centric or target-centric approach. As shown in Figure 2.1, disease-centric drug repositioning can directly link a drug to a pathological condition with no need for assessing target similarity or analyzing drug-target interactions. Target-centric drug repositioning, on the other hand, requires a firm connection between target and indication. Platforms like Open Targets (173) and Beagle (18) facilitate the identification of such connections. However, the target-indication link is often not so direct and clear.

Finally, the drug-centric repositioning cases are the only ones that could have benefited from drug-target interaction prediction methods.

3.4.4 Limit and potential of structure-based drug repositioning as drug-centric approach

Structure-based drug repositioning techniques are examples of drug-centric approaches. They can be applied to infer new interactions between drugs, targets, and indications by considering information about the structure of drugs and targets and their interactions. Although structure-based drug repositioning has great potential for repurposing known drugs to different targets and indications (128; 170), several limitations make it less easy to apply this approach in a relevant and systematic way. Structural data for both the original drug and its target and the potential new target, possibly its ligands, are required. Lack of this information considerably limits the searching space for drug repositioning. Actually, for none of the cases classified as drug-centric repositioning, both the original and the repositioned drug-target complex were available in the PDB, confirming the barriers of structure-based drug repositioning. Various techniques such as homology modeling and molecular docking can be used to predict the structure of a protein and its interaction with a query drug to tackle the scarce data availability. However, generating reliable data requires considerable expertise and computational power.

3.4.5 Database selection and data availability

By the time we started this work (2016), RDD was the only comprehensive drug repositioning database. Later, other data sets of repurposed drugs were published. For instance, in 2017, a new gold standard database of successful and failed repositioning cases, repoDB, was released (31). Despite differences in size between RDD and repoDB, we have decided to present our analysis as a proof-of-concept for the present drug repositioning situation. We are currently unaware of previous work based on RDD, but the database is manually curated and validated. We used the MDT database to retrieve drug-target connections, although many others, such as DrugBank, were available for this scope. The MDT database results from a comprehensive and accurate annotation that considers several sources of targets for FDA approved drugs (174). However, the combination of RDD and MDT has resulted in a data reduction that may have compromised our analysis's quality.

3.4.6 Ambiguous nomenclature of diseases

The nomenclature of diseases in medical research tends to be ambiguous. This is mainly because humans, albeit experts in their fields, assign the corresponding names to the diseases they study or mention in their work. Since these experts are not the same persons for every research work, and humans naturally differ in their opinions, perceptions, and approaches, disease names may vary. A systematic literature search is an essential tool of evidence-based medicine and, as such unavoidable in almost every part of medical research (134). Therefore, a consistent disease nomenclature is needed. Several databases that support the correct assignment of publications to the corresponding disease exist. Examples are the ICD-10, a medical classification list of the World Health Organization (WHO) (11); the Elsevier Life Science thesaurus Emtree (10), which is integrated into the Embase database; and the MeSH of the United States National Library of Medicine (NLM) (8), which is currently used by the MEDLINE/PubMed article database and by NLM's catalog of book holdings. MeSH and Emtree are the most commonly used databases, with MeSH usually preferred due to free access. MeSH consists of sets of descriptors arranged in alphabetical order and hierarchical structure (8). Emtree is organized similarly and was modeled in 1988 based on MeSH. It includes several terms for drugs, diseases, medical devices, and essential life science concepts. MeSH has extensive history notes that can track early literature predating the introduction of particular terms. Moreover, it exhibits comprehensive scope notes (9) and extensive nursing, veterinary medicine, and dentistry terminology. All of these terms can also be found in Emtree, but often as synonyms (7). Furthermore, MeSH and Emtree differ in a temporal component: While Emtree is used for all articles dating back to 1988, MeSH terms are also indexed for even older publications. We decided to use MeSH terms because of the above points and because the article indexing in MeSH differs from the indexing in other databases and, therefore, sometimes gives better results for a given search strategy.

3.4.7 Issues of the classification process

This study aimed to assess the impact of new drug-target interactions on the analyzed drug repositioning cases. Our work is a retrospective analysis based on the final results of different successful repositioning processes and does not consider the specific techniques used to repurpose the respective drugs (i.e., ligand-based, target-based, network-based, and machine-learning-based). The assignment of specific methods or hypotheses to each case of repositioning would have required a tremendous manual effort and would not have added essential information to our analysis. Besides, many repositioning processes require the

application of multiple techniques, so that a clear and unambiguous classification would have been difficult. Furthermore, we suppose that many repurposed drugs that we have assigned to disease-centric repositioning because of a similarity between their original and secondary indication could also be a product of a target-centric or drug-centric approach. For example, following target affinity experiments, the drug nilotinib, which we have classified as disease-centric repositioning, has been repurposed to treat gastrointestinal stromal tumors by affecting a different protein target (159). Moreover, the classification process was carried out step by step, beginning with the cases where the indications had an identical MeSH code (disease-centric) and continuing with those where the target was identical (target-centric). If a case was labeled as disease-centric, no further examination of the target was performed, thereby possibly missing overlaps between the different drug repositioning categories. For the reasons given above, a deeper molecular understanding of drug repositioning databases could lead to more detailed analysis and provide a clearer picture of the true state of repurposing.

3.5 Conclusion

Drug-target interaction prediction is an essential part of most of the rational drug repositioning approaches. Different biochemical, physical, and mathematical techniques have been designed and optimized to infer links between ligands and proteins accurately. In this work, we analyzed many successful drug-repositioning cases and, based on the similarity between old and new indication and old and new target. We evaluated the actual impact of drug-target interaction prediction on these cases. Dividing all the cases containing small molecules and protein targets in disease-centric, target-centric, and drug-centric drug repositioning, we conclude that only seven out of 128 cases would have required drug-target interaction prediction to initiate repurposing rationally. This unexpectedly small number could potentially be explained by the higher amount of information, time, and money required for drug-target interaction prediction compared to target- and disease-centric approaches. A more detailed analysis of the drug-target complexes present in the PDB revealed that there is currently not enough structural data available for any repositioning cases classified as drug-centric to identify new drug-target interactions using structure-based techniques such as interaction profile similarity. On the other hand, these results highlight a big unexplored niche for drug-target interaction prediction in drug repositioning. This great potential will increasingly be used as the techniques for detecting new links between ligands and protein targets, such as structure-based drug repositioning, are continually evolving.

MeSH Category	Disease group	MeSH Category	Disease group
C01	Bacterial Infections and Mycoses	C16	Congenital, Hereditary, and Neonatal Diseases and Abnormalities
C02	Virus Diseases	C17	Skin and Connective Tissue Diseases
C03	Parasitic Diseases	C18	Nutritional and Metabolic Diseases
C04	Neoplasms	C19	Endocrine System Diseases
C06	Digestive System Diseases	C20	Immune System Diseases
C07	Stomatognathic Diseases	C23	Pathological Conditions, Signs and Symptoms
C08	Respiratory Tract Diseases	C25	Chemically-Induced Disorders
C10	Nervous System Diseases	E01	Diagnosis
C11	Eye Diseases	F01	Behaviour Mechanisms
C12	Male Urogenital Diseases	F03	Mental Disorders
C13	Female Urogenital Diseases and Pregnancy Complications	G07	Physiological processes
C14	Cardiovascular Diseases	G08	Reproductive and Urinary Physiological Phenomena
C15	Hemic and Lymphatic Diseases	G11	Musculoskeletal and Neural Physiological Phenomena

Table 3.1: **Root MeSH tree key for the group of diseases and corresponding name**. MeSH category codes (left column) and common names (right column) for all types of diseases and conditions, for which drug repositioning was previously applied according to RDD.

N	MeSH Category	Drug Names	Number of Drugs
1	Neoplasms	Alitretinoin, Arsenic trioxide, Clofarabine, Daunorubicin liposomal, Doxorubicin, Erlotinib hydrochloride, Floxuridine, Idarubicin, Lapatinib, Nilotinib, Paclitaxel, Paclitaxel aqueous gel, Paclitaxel protein-bound particles for injection suspension, Pazopanib, Sorafenib, Toremifene	16
2	Immune System Diseases	Azathioprine, Beclomethasone 17,21-diproprionate, Fludarabine phosphate, Leflunomide, Lenalidomide, Mesalamine, Mycophenolate mofetil, Nabumetone, Nevirapine, Pentostatin, Pralatrexate, Thalidomide, Vorinostat	13
3	Nervous System Diseases	Apomorphine, Clonazepam, Gabapentin, Galantamine, Midazolam HCl, Memantine, Riluzole, Tetrabenazine	8
4	Bacterial infections and Mycoses	Aztreonam, Clindamycin, Doripenem, Levofloxacin, Rifabutin	5
5	Digestive System Diseases	Fluorouracil, Nitazoxanide, Nitisinone, Synthetic human secretin, Ursodiol,	5
6	Mental Disorders	Aripiprazole, Atomoxetine hydrochloride, Fluoxetine, Milnacipran, Pramipexole	5
7	Pathological Conditions, Signs and Symptoms	Aminocaproic acid, Bupivacaine, Medroxyprogesterone acetate, Midazolam nasal spray, Tramadol hydrochloride	5
8	Respiratory Tract Diseases	Ambrisentan, Bosentan, Mifepristone, Nitric oxide, Tiotropium bromide	5
9	Virus Diseases	Disoproxil fumarate, Ribavirin, Tenofovir	3
10	Cardiovascular Diseases	Bethanidine Sulfate, Nitroprusside	2
11	Female Urogenital Diseases and Pregnancy Complications	Progesterone, Testosterone propionate	2
12	Hemic and Lymphatic Diseases	Anagrelide, Decitabine	2
13	Behavior and Behavior Mechanisms	Bupropion	1
14	Diagnosis	Synthetic porcine secretin	1
15	Eye Diseases	Brimonidine	1
16	Musculoskeletal and Neural Physiological Phenomena	Mepivacaine	1
17	Nutritional and Metabolic Diseases	Miglustat	1
18	Parasitic Diseases	Praziquantel	1

Table 3.2: **Disease-centric cases of repositioning.** This table shows the 76 disease-centric repositioned drugs grouped by indication category according to the MeSH tree classification. Since the original and secondary therapeutic indications fall within the same MeSH category, no further analyses on the targets were carried out. The MeSH indication names are ordered by the quantity of repositioned drugs.

N	Drug name	Original Indication		Secondary Indication	
		MeSH Category	**Gene Target**	**MeSH Category**	**Gene Target**
1	Adenosine	Congenital, Hereditary and Neonatal Diseases	ADORA1/2A/2B/3	Nervous System Diseases	ADORA1/2A/2B/3
2	Albuterol	Respiratory Tract Diseases	ADRB2	Pathological Conditions, Signs and Symptoms	ADRB2
3	Alfetanil	Musculoskeletal and Neural Physiological Phenomena	OPRM1	Nervous System Diseases	OPRM1
4	Alprostadil	Mental Disorder	PTGER1/PTGER2	Cardiovascular Diseases	PTGER1/PTGER2
5	Amiloride	Nutritional and Metabolic Diseases	SCNN1A	Congenital, Hereditary and Neonatal Diseases	SCNN1A
6	Atovaquone	Pneumonia	Cytochrome b *(Pneumocystis carinii)*	Toxoplasma gondii encephalitis	Cytochrome b *(Toxoplasma gondii)*
7	Azacitidine	Hemic and Lymphatic Diseases	DNMT1/3A	Neoplasms	DNMT1/3A
8	Buprenorphine	Eye Diseases	OPRK1,OPRM1,OPRD1	Mental Disorder	OPRK1,OPRM1,OPRD1
9	Capsaicin	Pathological Conditions, Signs and Symptoms	TRPV1	Cardiovascular Diseases	TRPV1
10	Celecoxib	Immune System Diseases	PTGS2	Congenital, Hereditary and Neonatal Diseases	PTGS2
11	Chlorpromazine	Pathological Conditions, Signs and Symptoms	DRD2-4,HTR2A/2C,HRH1/4	Mental Disorder	HTR2A
12	Dapsone	Dermatitis herpetiformis	Dihydropteroate synthase *(Mycobacterium lepmae)*	Toxoplasmosis	Dihydropteroate synthase *(Toxoplasma gondii)*
13	Desmethylmifepristone	Respiratory Tract Diseases	NR3C1	Endocrine System Diseases	NR3C1
14	Dexamethasone	Eye Diseases	NR3C1	Immune System Disorders	NR3C1
15	Difluprednate	Pathological Conditions, Signs and Symptoms	NR3C1	Eye Diseases	NR3C1
16	Dihydrodigitoxin	Cardiovascular Diseases	ATP1A1-4	Endocrine System Diseases	ATP1A and more (127)
17	Dimethylstilberstrol	Female urogenital Diseases and Pregnancy Complications	KEAP1	Skin and Connective Tissue Diseases	KEAP1
18	Duloxetine	Mental Disorder	SLC6A,SLC6A4	Pathological Conditions, Signs and Symptoms	SLC6A4
19	Effornithine	African trypanosomiasis	Ornithine decarboxylate *(Trypanosoma Brucei)*	Pneumocystis carinii pneumonia	Ornithine decarboxylate *(Pneumocystis carinii)*
20	Epoprostenol Sodium	Respiratory Tract Diseases	PTGIR,PTGER1	Pathological Conditions, Signs and Symptoms	PTGIR,PTGER1
21	Ethinyl Estradiol	Respiratory Tract Diseases	ESR1	Skin and Connective Tissue Diseases	ESR1
22	Everolimus	Immune System Diseases	FKBP1A	Digestive System Diseases	FKBP1A
23	Finasteride	Male Urogenital Diseases	SRD5A1/2	Pathological Conditions, Signs and Symptoms	SRD5A1/2
24	Glycopyrrolate Bromide	Digestive System Diseases	CHRM1-5	Stomatognathic Diseases	CHRM1-5
25	Guanethidine	Cardiovascular Diseases	SLC6A2	Nervous System Diseases	SLC6A2
26	Guanfacine	Mental Disorder	ADRA2A/2B/2C	Congenital, Hereditary and Neonatal Diseases	ADRA2A/2B/2C
27	Histamine	Immune System Diseases	HRH1	Neoplasms	HRH1
28	Iloprost	Respiratory Tract Diseases	PTGIR	Cardiovascular Diseases	PTGIR
29	Ketokonazole	Fungal infection	Cytochrome P450 *(Candida albicans)*	Nephrotoxicity induced by cyclosporine	Cytochrome P450 *(Homo sapiens)*
30	Levomilnacipran	Mental Disorder	SLC6A2/4,	Nervous System Diseases	SLC6A2
31	Mecamylamine Hydrochloride	Cardiovascular Diseases	CHRNA3/B4	Mental Disorder	CHRNA3/B4
32	Metyrosine	Neoplasms	TH	Mental Disorder	TH
33	Minoxidil	Cardiovascular Diseases	ABCC9	Pathological Conditions, Signs and Symptoms	ABCC9
34	Misoprostol	Digestive System Diseases	PTGER3	Pathological Conditions, Signs and Symptoms	PTGER3
35	Oxandrolone	Physiological Phenomena	AR	Congenital, Hereditary and Neonatal Diseases	AR
36	Phentolamine	Cardiovascular Diseases	ADRA1A/1B/1D/2A/2B/2C	Mental Disorder	ADRA1A/1B/1D/2A/2B/2C
37	Propranolol	Pathological Conditions, Signs and Symptoms	ADRB1-3	Neoplasms	ADRB1-3
38	Raloxifene	Nutritional and Metabolic Diseases	ESR1/2	Skin and Connective Tissue Diseases	ESR1/2
39	Ropinirole	Cardiovascular Diseases	DRD2-4	Mental Disorder	DRD2-4
40	Sibutramine	Mental Disorder	SLC6A2-4	Physiological Phenomena	SLC6A2-4
41	Sildenafil	Pathological Conditions, Signs and Symptoms	PDE5A	Mental Disorder	PDE5A
42	Tadalafil	Cardiovascular Diseases	PDE5A	Mental Disorder	PDE5A
43	Tranexamic Acid	Pathological Conditions, Signs and Symptoms	PLG	Immune System Disorders	PLG
44	Tretinoin	Neoplasms	RARA/B/G	Musculoskeletal and Neural Physiological Phenomena	RARA/B/G
45	Trimetrexate	Pneumonia	Dihydropholate reductase *(Pneumocystis carinii)*	Metastatic carcinoma of the head and neck	Dihydropholate reductase *(Homo sapiens)*

Table 3.3: **Taget-centric repositioning cases.** Disease (MeSH category) and protein target (gene name or Uniprot ID) are shown for both the primary and the secondary indication. For the 46 cases of target-centric repositioning, the target UniprotID is identical for original and secondary indication. The drugs are listed in alphabetical order. The references of target-disease associations retrieved from PubMed are given in the respective cells. Five drugs have been repurposed to/from a non-human ortholog target with a sequence similarity higher than 30%.

N	Drug name	Original Indication		Secondary Indication	
		MeSH Category	Gene Target	Category	Gene Target
1	Allopurinol	Neoplasm	Xanthine dehydrogenase/oxidase *(Homo sapiens)*	Parasitic Diseases	Hypoxanthine phosphoribosyltransferase *(Trypanosoma Cruzi (141))*
2	Doxycycline	Bacterial Infection and Mycoses	*rpsD,*rpsI	Stomatognathic Diseases	MMP1/7/8/13
3	Lidocaine	Musculoskeletal and Neural Physiological Phenomena	SCN1A/2A/3A/4A/5A/7A/8A/9A/10A/11A	Immune System Disorders	various/not specified (cytokines release) (137)
4	Mazindol	Stomatognathic Diseases	SLC6A2-4	Congenital, Hereditary and Neonatal Diseases	various/not specified (growth hormone release) (43)
5	Topiramate	Nervous System Diseases	GABR(+),GRIK1-5,GRIA1-4,SCN(+)	Stomatognathic Diseases	CA2/4
6	Valproic acid	Nervous System Diseases	ALDH5A1,ABAT	Congenital, Hereditary and Neonatal Diseases	HDAC2
7	Zidovudine	Neoplasms	HIV1 Reverse transcriptase (138)	Immune System Disorders	Human DNA polymerase (146)

Table 3.4: **Drug-centric repositioning cases**. For each drug-centric case, therapeutic indications (MeSH category) and protein targets (gene name or Uniprot ID (*: Non-human, (+): multiple subunits)) for both the original and the second indications are shown. According to our drug-centric definition, these seven cases must have a different MeSH code and protein target for the primary and secondary indication. The seven cases are most interesting for applying drug-target interaction prediction techniques because they have the highest target and indication diversity. The drugs are listed in alphabetical order. The references of target-disease associations retrieved from PubMed are given in the respective cells.

Chapter 4

Interaction profile similarity-based drug repositioning reveals new protein-ligand interactions for B-cell activation inhibition

This chapter is based on the published work "Structure-based drug repositioning explains ibrutinib as VEGFR2 inhibitor", MF Adasme, D Parisi, K Van Belle, S Salentin, VJ Haupt, GS Jennings, Jörg-Christian Heinrich, Jean Herman, Ben Sprangers, Thierry Louat, Yves Moreau, Michael Schroeder; Plos one 15 (5), e0233089, 2020. Adasme and Parisi developed the theory, performed the computations and wrote the manuscript. Van Belle executed the experimental part. Salentin, Haupt and Jennings verified the theory and computational analyses. Heinrich contributed to sample preparation. Herman, Sprangers and Louat supervised the experimental part. Moreau and Schroeder supervised the project.

4.1 PLIP-based drug repositioning approach for new kinase inhibitors

We applied PLIP-based interaction profile similarity drug repositioning to identify new kinase inhibitors. We studied drug promiscuity at the structural level to screen vast amounts of drug-target interactions to identify drugs inactivating B-cells, whose dysregulation can function as a driver of autoimmune diseases. Screening over 500 kinases, we identified 22 whose knock out impeded the activation of B-cells. Among these 22 is the gene KDR, whose gene product VEGFR2 is a prominent cancer target with anti-VEGFR2 drugs on the market for over a decade. This study's main result is that structure-based drug repositioning for the identified kinase targets revealed the cancer drug ibrutinib as a micromolar VEGFR2 inhibitor with a very high therapeutic index in B-cell inactivation. These findings suggest that ibrutinib is not only acting on Bruton's tyrosine kinase BTK, against which it was designed. Instead, it may be a polypharmacological drug, which additionally targets angiogenesis via inhibition of VEGFR2. Therefore ibrutinib carries the potential to treat other VEGFR2 associated diseases. Structure-based drug repositioning explains ibrutinib's anti VEGFR2 action through a specific pattern of interactions shared between BTK and VEGFR2. Overall, structure-based drug repositioning predicted these findings at a fraction of the time and cost of a conventional screen. Our drug repositioning approach has been summarized on Figure 4.1.

4.2 Target selection

To identify protein targets relevant for B-cell activation, a target selection based on a previously established knock-out assay has been performed (200). Burkitt lymphoma Namalwa cells stimulated with ODN2006 for 24 hours have been used as a model of activated B-cells. An shRNA library (MISSION LentiExpress TM shRNA library, Sigma-Aldrich, Diegem, Belgium) containing 3 - 10 shRNA clones per almost every human kinase (501 in total) has been used. Cells were incubated with fluorescent antibodies for surface markers CD70 and CD80. After FACS sorting, expression was measured by the amount of fluorescence. Then, the y-value was calculated as follows:

$$y = \frac{SC - SS}{SC - NSC}$$

where SC is the expression of stimulated control, SS the stimulated sample, and NSC the non-stimulated control. Concretely, it is the fraction of expression

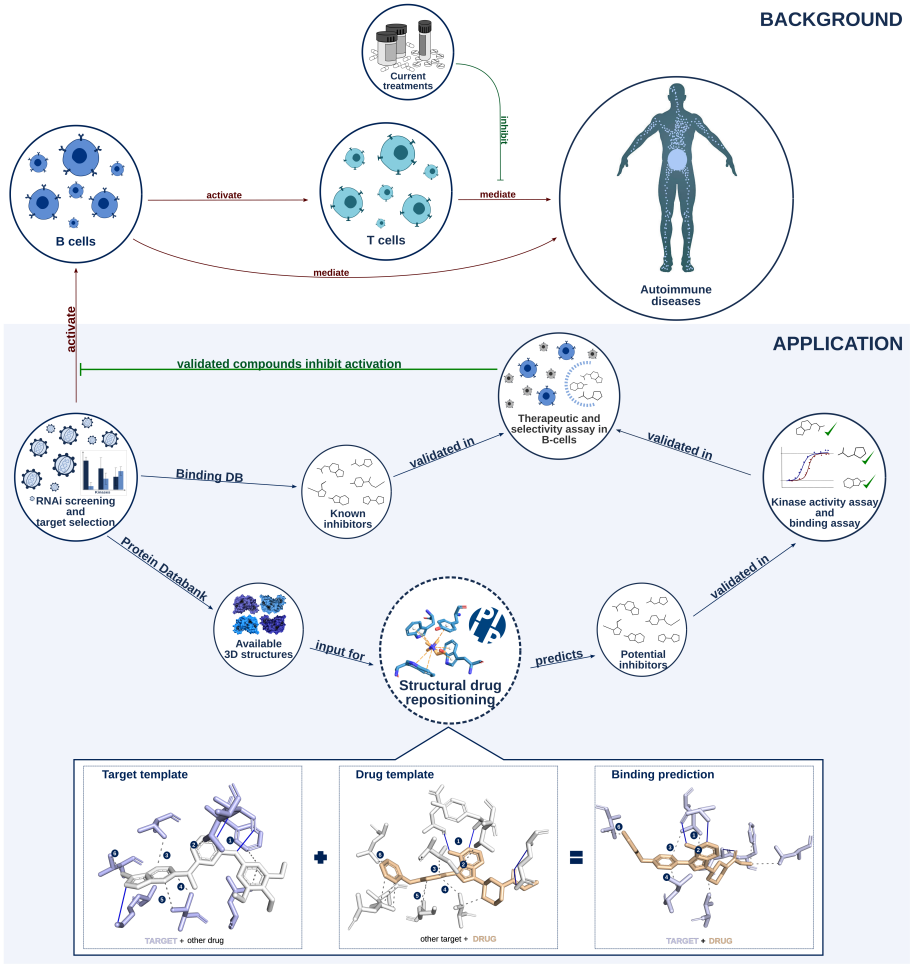


Figure 4.1: **Overview of the study. Background.** B- and T-cells play an important role in autoimmune diseases. Current treatments focus on the inhibition of T-cell activation. Novel or improved treatments may emerge by inhibiting B-cells. **Application.** Through an RNAi screen, we identified kinases that play a role in B-cell activation and are therefore suitable autoimmune targets. There are two routes to validate inhibitors against these kinases. First, known inhibitors may exist and second, structural drug repositioning may yield novel inhibitors after *in vitro* validation. Finally, all validated compounds are checked *in vitro* for selectivity and efficacy of B-cell inactivation.

level of activated control minus knock out and expression level of activated control minus non-activated control. A y-value close to zero indicates that the knock out does not influence B-cell's activation, a value higher than one means that the knock out potentially inhibits activation of B-cells. It must be considered that without stimulation, CD70 and CD80 expression in B-cells is low; upon stimulation, the markers are upregulated, and finally, upon knock out of a kinase essential in B-cell activation, expression of the markers decreases. It is not sufficient to compare expression levels of non-activated controls to activated knockouts, but activated controls have to be factored in. Based on a close observation of the data, we observed that silencing of "hit" kinases gave a more potent inhibition of marker CD70 compared to marker CD80, clones with $y > 0.8$ for CD70 and $y > 0.5$ for CD80 were selected and resulted in the 22 kinase targets. The process has been summarized on Figure 4.2. Besides the definition of the y-value, the quality of knock out is important to consider. Per kinase to be silenced, there are multiple RNA constructs, hitting different target gene regions, which may suffer from different off-targets. There were between one and ten clones across the 22 kinases with varying degrees of effectiveness in silencing. Subsequently, kinases were considered suitable targets for B-cell inactivation/modulation if at least one clone shows a y-value greater than 0.5 for CD80 and greater 0.8 for CD70. Four hundred targets resulted in being non-lethal and 22 inactivating B-cells, leading to a downregulation of the cell surface markers CD70 and CD80. Table 4.1 shows these fractions for the best clone of each of the 22 kinases. Those 22 kinases were the starting point for the interaction profile similarity-based virtual screening.

4.3 Interaction profile similarity-based virtual screening

From the set of kinase targets for which at least one structure in complex with a small molecule ligand was available on PDB (May 2018), a fingerprint of the interaction profile was generated with PLIP. Those vectors have been compared with interaction fingerprints of all the other ligand-protein couples stored on PDB (over 300.000 complexes) by computing their Tanimoto score. The significance of similarity scores has been expressed by p-value. Only those with a p-value lower than 10 3 have been retained. We found available 3D structures only for 14 of the 22 kinase targets. This number has been reduced to 10 after removing structures without interacting compounds or binding ions or biologically non-relevant small molecules, according to BioLiP (zhanglab.ccmb.med.umich.edu/BioLiP). For these ten kinases, there were a total of 141 drug-target templates on PDB, while regarding candidate drugs to

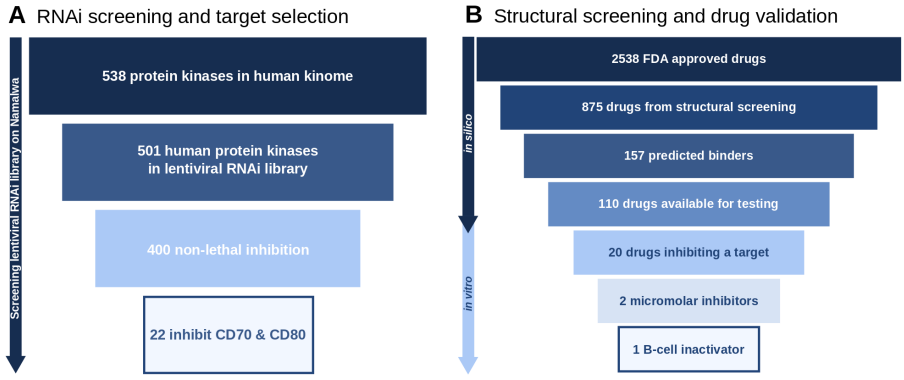


Figure 4.2: **Selection of targets and inhibitors. A.**From 538 human kinases to 22 involved in B-cell activation in Namalwa cells. Nearly all human kinases (501) are present in the RNAi library, and for most of them (400), the knockout is not lethal. Of those 400, 22 inhibit expression of both B-cell activation markers CD70 and CD80. **B.**From 2538 FDA approved drugs to 1μM B-cell inactivator. Out of the 875 drugs with available structures, the *in silico* screening yielded 157 positive binders. From the 110 drugs available for *in vitro* testing, 20 were validated as target inhibitors, two validated in kinase binding assay as positive micromolar inhibitors, and finally, one validated drug with positive effects in B-cells inactivation.

be repurposed, PDB covers relevant structural data for 875 out of some 2500 FDA approved drug. The structural drug repositioning screening predicted 157 approved drugs as putative ligands for our targets. Those predictions are not equally distributed across the ten targets, but they correlate with the amount of structural data available, e.g., for VEGFR2 with its 36 structures, there are 53 resulting predictions, while MLK4 with one structure has three predicted ligands.

4.4 Kinase inhibition test confirmed part of the structure-based prediction

Before testing the predicted binders for their activity against the targets, the compounds predicted with structure-based virtual screening have been filtered using DrugBank to exclude nutraceuticals and already known binders. Also, six compounds not available from Sigma-Aldrich have been discarded. Dasatinib, a strong binder of all targets, was added as a control. The 111 resulting compounds

were bought from Sigma-Aldrich and Selleckchem. A kinase inhibition test in duplicate at 1M and 10M concentrations has been carried by ReactionBiology, resulting in 888 assays involving the ten targets showed on the top part of Table 4.1. Data were normalized to percentage Enzyme Activity relative to DMSO controls. Compounds inhibiting more than 50 percent of the enzyme activity have been defined as potent inhibitors and inhibition between 30 and 50 percent as weak inhibition. Only one compound, Ibrutinib, displayed potent inhibition at both 10 and 1 M against one of the targets (VEGFR2), while nine other compounds showed potent inhibition at ten M only. The activity of ibrutinib on VEGFR2 is also confirmed by other experimental results in the LINCS kinomescan database (91). Ten additional compounds show weak inhibition at ten M, as shown on Table 4.2.

Thus, considering 22 known binders predicted by PLIP and confirmed by BindingDB, and 20 new ligands validated with the kinase assay, our interaction profile similarity-based approach highlighted 42 FDA approved drugs for their potential to inactivate B-cells. Names, structures and primary therapeutic indication of those drugs are shown in Figure 4.3.

4.5 Binding affinity test revealed ibrutinib and suramin high affinity for VEGFR2 ATP binding pocket

Since the used repositioning approach is based on the similarity of binding behavior, we wanted to prove that the potent kinase inhibition recorded with suramin and ibrutinib is related to the interaction with the analyzed binding pocket and not to other allosteric phenomena. Thus we performed a binding affinity test between our predicted binders and the ATP binding pocket of VEGFR2. Dose-response curves for ibrutinib and suramin against VEGFR2 were determined with the KINOMEscan TM by DisvcoverX, a competition binding assay that quantitatively measures the ability of a compound to compete with an immobilized, active-site directed ligand. It is measured via quantitative PCR of the DNA tag linked to the kinase. An 11-point 3-fold serial dilution of each test compound was prepared in 100% DMSO at 100x of the final test concentration and subsequently diluted to 1x in the assay (final DMSO concentration = 1%). Most Kds were determined using a compound top concentration of 30000 nM. If the initial Kd determined appeared to be lower than 0.5 nM (the lowest concentration tested), the measurement was repeated with a serial dilution starting at a lower top concentration. A Kd value reported as 40000 nM

indicates that the Kd was higher than 30000 nM. Binding constants (Kds) were calculated with a standard dose-response curve using the Hill equation:

$$Response = Background + \frac{Signal - Background}{1 + (KdHillSlope/DoseHillSlope)}$$

The Hill Slope was set to -1. Curves were fitted using a non-linear least square fit with the Levenberg-Marquardt algorithm. The resulting curve Figure 4.4. A Ibrutinib binds to VEGFR2 at a Kd of 2M and suramin at a Kd of 25M.

4.6 Ibrutinib and suramin interaction profiles explained PLIP-based prediction

To better understand how ibrutinib and suramin achieve this high affinity against VEGFR2, we checked their interaction profiles. Interactions were obtained from PLIP v1.4.0 applied to PDB IDs 3cjg, 4ifg, 3c7q, 3gan. The prediction of ibrutinib was based on ibrutinib's interaction with CDPK1 (PDB 4ifg), which is very similar to pazopanib's interaction with VEGFR2 (PDB 3cjg). This is a remarkable similarity, as CDPK1 is from the parasite Toxoplasma gondii and VEGFR2 from human. Despite two targets from completely different species, Figure 4.4. B shows that both interactions comprise a double hydrogen bond (1) and five hydrophobic interactions in a similar position (2), (3), (4), (5), and (6). Similarly, a suramin's binding to RDM1 (PDB 3gan) is similar to nintedanib's interaction with VEGFR2 (PDB 3c7q). Again, this similarity is very remarkable, as RDM1 is a regulator of DNA methylation in Arabidopsis thaliana and VEGFR2 is from human. Nonetheless, both comprise in their ligand interaction a double hydrogen bond (1), a simple hydrogen bond (2), three hydrophobic interactions (3), (4), (5), and a salt bridge (6). These examples document scaffold hopping from pazopanib to ibrutinib and from nindetadib to suramin and target hopping from parasitic CDPK1 to human VEGFR2 and from plant RDM1 to human VEGFR2.

4.7 B-cell activation assay shows ibrutinib inhibits B-cell activation

In order to test the drugs' ability to impede B-cell's activation (drug efficacy) and not affecting T-cells (drug selectivity) for the selected hit compounds, four assays were performed: B-cell assay, Mixed Lymphocyte Reaction (MLR), WST-1 assay

on RPMI1788 cells (B-cell line) and cytotoxicity WST-1 assay on Jurkat cells (T-cell line). To express such characteristics, we use the therapeutic index for B- and T-cells, which relates toxicity to inactivation. From this, we compute the specificity index as the ratio of B-cell to T-cell therapeutic index. Compounds were prepared as 10 mM stock solutions in Dimethylsulfoxide (DMSO)and were tested at 50-10-1-0.1-0.01 and 0.001 M in the four assays. Blood samples of healthy volunteers were collected at the Red Cross of Mechelen, Belgium. Each donor consents to the use of his blood for research purposes. Human Peripheral Blood Mononuclear Cells (PBMC) were obtained by density gradient centrifugation of the heparinized venous blood over Lymphoprep (Axis Shield PoC AS; density 1.077 0.001 g/mL). The human B cell line RPMI1788 (American Type Culture Collection, USA) and the human T cell line Jurkat (European Collection of Cell Cultures, ECACC, England) were maintained in RPMI1640 culture medium (BioWhittaker, Lonza, Verviers, Belgium) containing 10% fetal calf serum (FCS, HyClone Thermo Scientific, United Kingdom) and 5 g/mL gentamicin sulfate (BioWhittaker, Lonza, Verviers, Belgium) at 37C and 5% CO2. Highly purified naive peripheral human B cells were separated from fresh human PBMCs on magnetic columns by positive selection using CD19 magnetic beads according to the manufacturer's instructions (MACS Miltenyi Biotech, Leiden, The Netherlands). The purity of the isolated naive B cells was 95% as analyzed by flow cytometry. Freshly isolated human CD19+ B cells were then plated at 25 000 cells per well in a 384-well plate (Perkin Elmer, Zaventem, Belgium) in 55 L in DMEM medium containing 10% FCS and 5 g/mL gentamicin sulfate. After seven days of stimulation with 0.1 M ODN2006 (InvivoGen, Toulouse, France), the supernatant was taken to analyze IgG using the AlphaLISA human IgG kit according to the manufacturer's instructions (Perkin Elmer, Zaventem, Belgium). The analysis was performed with the EnVisionTM 2103 Multilabel Reader (Perkin Elmer, Zaventem, Belgium). MLR assay (T cell assay): Freshly isolated human PBMCs and (responder cells) were resuspended in RPMI1640 medium containing 10% FCS and 5ng/mL gentamicin sulfate. RPMI1788 cells (stimulator cells) were inactivated by treatment with mitomycin C (Kyowa, Takeda Belgium, Brussels, Belgium) for 20 min at 37 C, washed, and finally suspended in the culture medium. An amount of 100 L of each cell suspension was mixed with 20 L of the diluted compound. The mixed cells were cultured at 37 C for six days in 5% CO2. DNA synthesis was assayed by the addition of 10 Ci (methyl-3H) thymidine (Perkin Elmer, Zaventem, Belgium) per well during the last 18 h in culture. After that, the cells were harvested on glass filter paper, and the counts per minute were determined in a liquid scintillation counter (TopCount, Perkin Elmer, Zaventem, Belgium). WST-1 assays on RPMI-1788 cells and Jurkat cells were performed as follows: 20 L of test compound dilution were added to each well containing exponential growing cells, and the plates were incubated for 48 h at 37 C, 5%

CO2. Untreated cells and positive control (1% triton X-100, for the last 15 min) served as the reference for maximum and minimum viability. At the end of incubation, 100 L of supernatant were removed and replaced by 10 L of WST-1 solution (Cell Proliferation Reagent WST-1, Roche Applied Science). After 3 to 5 h incubation at 37C, 5% CO2, optical density was measured at 450nm with the EnVisionTM 2103 Multilabel Reader. B cell Therapeutic Index (TI) was calculated as the ratio IC50 WST-1 assay RPMI1788/IC50 B cell assay; T cell TI referred to the ratio IC50 WST-1 assay Jurkat/IC50 B CELL assay; while the Specificity Index (SI) was calculated as the ratio IC50 MLR assay/IC50 B cell assay. Our goal is to uncover compounds with high B-cell TI and high specificity index. Indices are computed from duplicates. As a main result, ibrutinib has a very high B-cell therapeutic index of over 40000, confirming ibrutinib's ability to inactivate B-cells at low toxicity. A SI of over 5000 also means that ibrutinib interferes only little with T-cells. Table 4.3 summarizes these indices, including 15 other compounds. For suramin, the therapeutic and SIs are close to one, and hence suramin is not a B-cell modulator, although it is a micromolar VEGFR2 inhibitor. Tofacitinib, which can be seen as a control due to its approval for autoimmune indications, has a good B-cell TI but lacks specificity. The full selection process has been summarized on Figure 4.2.

4.8 Discussion on target selection process

The RNAi screen on the human B-cell line Namalwa identified novel B-cell targets using CD70and CD80 expression as a read-out for B-cell activity. We must clarify that the role of CD70 and CD80 is not exclusive of B-cells activity, and therefore their inhibition might also be significant in contrasting T-lymphocytes over-activation with an advantageous pleiotropic effect in immunosuppressive therapy. Among the various stimuli investigated for B-cell activation, in vitro stimulation with TLR9 agonist ODN2006 resulted in the broadest phenotypic change profile in human polyclonal B-cells. Compared with an assay of human B-cell IgG, which checks the IgG production level by primary B-cells after seven days incubation with the same stimulation agent ODN2006, the Namalwa-ODN2006 protocol takes only 24 hours, making it ideal for fast first-line screening. Furthermore, the heterogeneity between the different human blood donors, the heterogeneity after isolation and purification from peripheral blood, the limited yield of the purification, and the short longevity make primary B-cells less optimal for repeatable assays. A monoclonal population of dividing cells such as the human B-cell line Namalwa stabilizes the assay by overcoming the limitations mentioned above. VEGFR2 is a Receptor Tyrosine Kinases (RTK) with a critical role in regulating angiogenesis and lymphangiogenesis, which makes it an attractive target for antiangiogenic and proangiogenic

therapies (181). Many VEGFR2 neutralizing agents have been developed for cancer treatment, such as the human antibody bevacizumab (Avastin) (181), the kinase inhibitors sunitinib, sorafenib, axitinib, and vandetanib. This kinase already showed an exciting connection with the autoimmune disease psoriasis (74), and two of its inhibitors have been repositioned from the anti-cancer indication to immunomodulation and are currently in clinical trials for the treatment of psoriasis (pazopanib) and systemic scleroderma (Nintedanib). The RNAi screen generated another 21 novel B-cell targets overall. For some, there is a link to autoimmunity, and for a few, there are drugs known to inhibit them. However, future work should fully explore all candidates regarding their relation to disease, a challenging thing to do for any target and any drug: even if a drug and a target can be linked to a disease, it still does not necessarily mean that the drug acts through the target. As an example, the RNAi screen suggests that silencing of VEGFR2 leads to B-cell inactivation. The inhibition of VEGFR2 with ibrutinib confirmed this effect, whereas inhibition with suramin did not. Generally, besides a methodological difference in silencing and inhibition, ibrutinib may have a more favorable polypharmacological profile for B-cell inactivation than suramin has.

4.9 Discussion on interactio-profile similarity based drug repositioning

Drug promiscuity is a common phenomenon, and we have exploited it here in a structure-based drug repositioning screen to identify ibrutinib, a novel VEGFR2 inhibitor, and a B-cell inactivator. Moreover, last but not least, the structure-based drug repositioning comprises the prediction and a transparent physio-chemical explanation for the prediction. Thus, giving confidence and a starting point for subsequent optimization steps. A prerequisite for these successes is the availability of structures available in the PDB. While there are far fewer structures than sequences in public databases, the growth of structural data is none the less impressive: PDB has more than doubled in size over the last seven years. Today, it contains 3D structures of over 1200 different drug targets, and more than 60% of all PDB structures contain proteins in complexes with biologically relevant ligands. For the present screen, PDB contained suitable structures for about half of the targets. Supplementary Figure S1 shows how these structures became available over time. It reveals that 15 years ago, structure-based drug repositioning would not have been possible for the identified 22 targets. If we assume a continuing linear growth, the figure suggests that in another 15 years, the complete set of targets can be surveyed. Thus, the potential of structure-based drug repositioning is visible

today and will further increase over time. Technological advance is also likely to speed up progress since improvements in structure prediction, and modeling may soon reach the maturity to close the structural gap faster. The structural drug repositioning screening, as carried out here, considered two aspects: the ligand-target interaction and superposition of binding site and ligand. Generally, many previous approaches in the past have focused either on the ligand (see, e.g., (47)) or on the target (geometrical or chemical similarities of the target proteins (78)). Our approach combined both views. The quality of such an integrated approach can be measured by the hit rates achieved. Here, we considered 110 compounds after the prediction and filtering steps. For 15 predictions, there was experimental evidence of high-affinity binding in BindingDB, and for other 20 unknown binders, our kinase inhibition assay provided confirmation, which corresponds to a very high hit rate of 32%. Such hit rates are unprecedented in traditional high-throughput screening, but they come at a cost. There is a large compound library in classical high-throughput screening, which may be already focused on a specific compound type. In contrast, the structural screen is limited by the compounds in PDB. Overall, there are over 30000, which is substantial. However, compared to the millions of compounds available in databases such as PubChem, it is small. This gap in size is reduced if one adds additional selection criteria. In drug repositioning, this can be, e.g., the approval of the compounds. There are some 2500 approved drugs, and for the structural screening, some 800 are available in PDB. So, the orders of magnitude in size difference reduce to a factor of three. A possible disadvantage of a smaller library may be the inability to cover a large chemical space. However, this problem does not arise in the structural screen as carried out here since the focus on drug-target interaction supported scaffold hopping. It ensured that huge steps in chemical space were possible and were indeed observed in the resulting data, as shown in the chemical similarity matrix (Figure 4.5). Finally, in silico screening implies low cost and time. Instead of 8750 validation experiments (875 approved drugs in PDB times ten targets), there were only some hundreds, a cost reduction by nearly two orders of magnitude. In summary, the combination of a small in silico library with the algorithmic approach ultimately leads to a very high hit rate over a large chemical space at a low screening cost. Other computational approaches also pursue in silico drug-target prediction by using text mining, machine learning, or chemical similarities, among other methods. When comparing our method to the three most cited approaches (54; 60; 113), in order to check whether they can predict the ibrutinib-VEGFR2 connection or not, we came out with the following: SwissTargetPrediction (129 citations) is based on similarity measures of the chemical structure and molecular shape. Performing a target-based screening for Ibrutinib, it predicted a long list of TYR kinases, including EGFR, but no specific link to VEGFR2. On the other hand, SuperPred (98 citations) is based on structural fingerprint similarities and

text mining for target prediction. It identifies no known targets for Ibrutinib, and while it can predict 5 TYR kinases with good E values, none of them is a direct link to VEGFR2. Finally, Hitpick (45 citations) combines two methods (1-Nearest-Neighbour and machine learning) to calculate 2D similarity. For ibrutinib, it predicted only 2 TYR kinases as targets, but none of them VEGFR2. Although they all successfully predicted targets from the protein tyrosine kinase class, which is the same as VEGFR2, none of them made the direct link as our method did. Furthermore, all of the mentioned approaches are based on 2D/3D similarity measures of the compounds and available assay data from databases. Therefore, their screening results often in similar scaffolds and protein classes, being unable to make such a specific prediction for distant biochemical entities.

4.10 Conclusion

B-cells play an important role in diseases ranging from auto-immunity to cancer. There is an unmet need for small molecules inactivating B-cells. Here, we present ibrutinib as such a candidate immuno-modulator. First, we ran an RNAi screen with a recently introduced drug screening assay for B-cell inactivation. We found 22 kinases whose silencing inactivates B-cells. Many of these targets are already drug targets with approved drugs, which will benefit future drug repositioning efforts. We focused specifically on VEGFR2, as it is already an established cancer drug target, and there is evidence for a role in autoimmunity, too. For lead identification, we pursued two approaches. We evaluated existing drugs affecting the identified targets. As we found mostly chemically similar cancer drugs with a polypharmacological design hitting many of our targets, we wanted to expand the scaffold space. We used structural drug repositioning to predict novel inhibitors, which we then validated in enzyme activity and critical study. We identified the cancer drug ibrutinib and the anti-infective suramin as micromolar VEGFR2 inhibitors. In a final step, we tested B-cell inactivation and found ibrutinib as a very effective inhibitor with a very high therapeutic index and reasonable specificity to B-cells rather than T-cells. The discovery is particularly remarkable as the underlying scaffold hopping is tied to target hopping. The source of the prediction is similarities in the binding sites of human VEGFR2 to a parasitic and a plant protein. These similarities are remarkable as they are unlikely of an evolutionary origin since family boundaries are crossed and species boundaries. Instead, they document the large but limited number of binding sites, as postulated in (16; 57). Overall, the in silico structural screen produced results at a very high hit rate, translating into low cost and time. It was able to predict chemically distant compounds, thus covering a large chemical space.

	Gene name	Protein name	PDB structures	CD70 y-value	CD80 y-value	% Cell survival
1	DAPK1	Death-associated protein kinase 1	39	1.21	0.87	100
2	**KDR**	**Vascular endothelial growth factor receptor 2 (VEGFR2)**	**36**	**1.27**	**0.98**	**75.6**
3	NEK2	Serine/threonine-protein kinase Nek2	24	1.37	0.79	100
4	CSNK1D	Casein kinase I isoform delta (CK1d)	17	0.97	1.28	100
5	EPHB4	Ephrin type-B receptor 4	14	1.95	0.74	100
6	CDC7	Cell division cycle 7-related protein kinase	4	1.88	0.83	100
7	ILK	Integrin-linked protein kinase	2	1.28	0.95	100
8	MST-4	Serine/threonine-protein kinase 26	2	2.63	0.73	100
9	MAP2K2	Dual specificity mitogen-activated protein kinase kinase 2	2	1.32	0.67	82.9
10	MLK4	Mitogen-activated protein kinase kinase kinase 21	1	1.81	0.54	99.7
11	PXK	PX domain-containing protein kinase-like protein	0	1.16	0.54	94.4
12	PTK7	Inactive tyrosine-protein kinase 7	0	1.07	0.71	86.2
13	PAK3	Serine/threonine-protein kinase PAK 3	0	2.83	0.98	86.8
14	BRSK1	Serine/threonine-protein kinase	0	1.76	0.83	100
15	FER	Tyrosine-protein kinase Fer	0	1.60	0.95	100
16	IRAK-3	Interleukin-1 receptor-associated kinase 3	0	1.70	1.02	93.3
17	MAP3K10	Mitogen-activated protein kinase kinase kinase 10	0	1.65	0.93	87.0
18	PRKAG2	5'-AMP-activated protein kinase subunit gamma-2	0	1.57	0.74	100
19	SCYL1	N-terminal kinase-like protein	0	1.18	0.68	84.6
20	STK33	Serine/threonine-protein kinase 33	0	1.43	0.92	100
21	TXK	Tyrosine-protein kinase TXK	0	1.49	1.06	79.9
22	WNK2	Serine/threonine-protein kinase WNK2	0	1.47	0.86	100

Table 4.1: **Targets identified in RNAi B-cell activation screen sorted by availability of structural data.** The top ten targets are suitable for structure-based drug repositioning screening. For ten targets (number 11 to 20), no structural were available, and thus these targets could not be used for structure-based drug repositioning. Silencing of all listed targets leads to strong downregulation of CD70 and CD80 expressions in the B cell Namalwa cell line activated by the TLR9 agonist ODN2006 without causing cell death (column cell survival).

Compound	Target	% Enzyme activity at 1μM	% Enzyme activity at 10μM
Hexachlorophene	NEK2	84.92	00.02
Ibrutinib	**VEGFR2**	**19.52**	**01.50**
Hexachlorophene	VEGFR2	109.5	05.14
Suramin	**VEGFR2**	**72.67**	**06.48**
Toremifene	DAPK1	90.42	14.77
Ibrutinib	DAPK1	95.59	23.72
Saquinavir	DAPK1	87.09	27.81
Tolcapone	CK1d	91.69	35.58
Ritonavir	DAPK1	92.58	38.46
Etravirine	VEGFR2	108.4	41.08
Thyroxine	NEK2	90.16	48.11
Lopinavir	DAPK1	90.43	49.89
Adenosine	CK1d	87.72	50.22
Diethyldtilbestrol	NEK2	97.92	50.73
Proflavine	NEK2	104.1	51.57
Triiodothyronine	NEK2	111.9	55.67
Adenosine	NEK2	96.13	58.99
Triclosan	NEK2	99.17	64.62
Triclosan	CDC7	70.45	65.43
Propofol	EPHB4	66.96	68.44
Prochlorperazine	NEK2	70.10	68.73
Estradiol	DAPK1	110.6	68.98
Amodiaquine	CK1d	107.2	69.07
Niacinamide	EPHB4	69.30	69.74

Table 4.2: **List of new drug-target interactions predicted by interaction profile similarity and confirmed by experimental kinase activity assay**. Enzyme activity was measured at a compound concentration of 1μM and 10 μM. Drugs able to lower the enzyme activity to less than 50% (above dashed line) at 10 μM were defined as potent inhibitors while drugs that lowered kinase activity at 10μM between 50% and 70% (below the dashed line) were classified as weak inhibitors. Discarding the disinfectant Hexachlorophene because of its low medical interest, two compounds have been highlighted for showing inhibition at both 1μM and 10μM, ibrutinib and suramin.

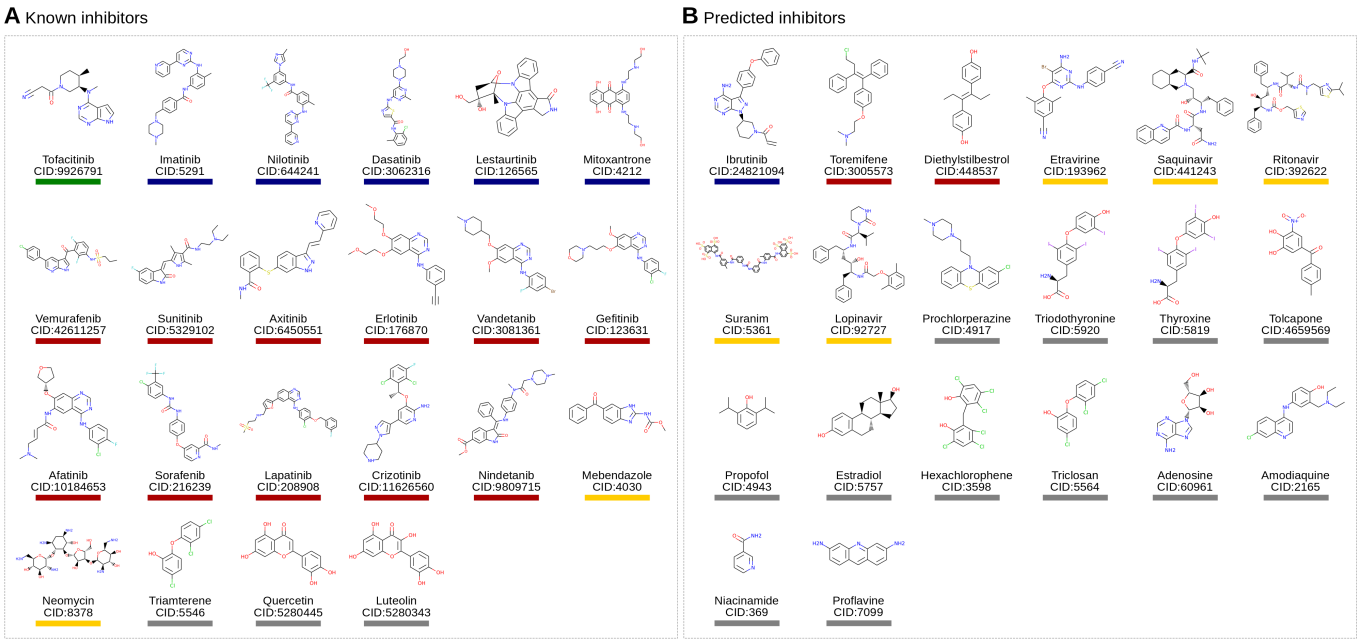


Figure 4.3: **Chemical structures of known (A) and predicted (B) inhibitors.** Here, drugs with drug name and compound ID (CID) are grouped by therapeutic indication: Autoimmune diseases(green), Cancer immune system (blue), Cancer (red), Infections (yellow), and others (gray).

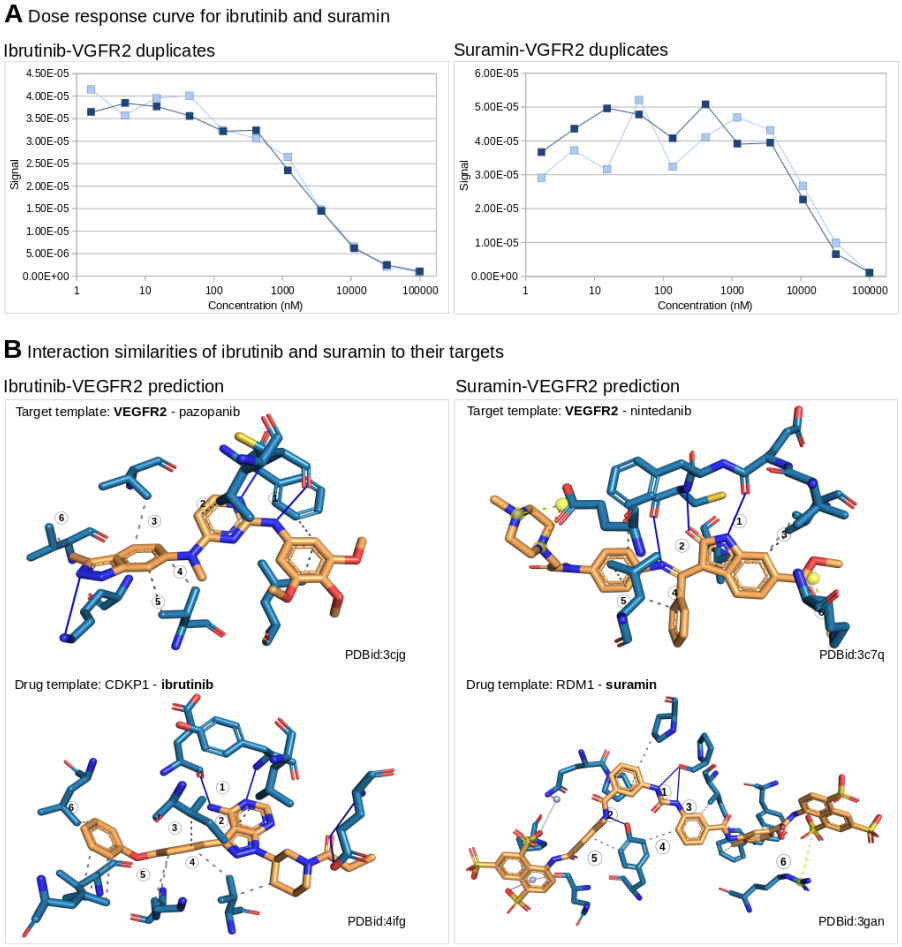


Figure 4.4: **Ibrutinib and suramin are inhibitors of VEGFR2. A.**Dose response curves for ibrutinib (KD value of 2μM) and suramin (Kd value of 25μM) from the competition binding assay determined with the KINOMEscan TM by DisvcoverX prove that ibrutinib and suramin are μM inhibitors of VEGFR2. **B.** Structural repositioning based in interaction of inhibitors (orange) to their targets (blue), predicts ibrutinib as VEGFR2 inhibitor since ibrutinib's interaction to CDPK1 is similar to pazopanib's interaction to VEGFR2.Both have (1) a double hydrogen bond (blue lines) and (2)(3)(4)(5)(6) hydrophobic interactions (gray dashed lines) in common. Similarly, suramin is predicted as VEGFR2 inhibitor. They have (1) a double hydrogen bond, (2) a simple hydrogen bond, (3)(4)(5) hydrophobic interactions and (6) a salt bridge (yellow dashed lines) in common. Subsequent in vitro validation proves both predictions correct.

Drug	B-cell IC50 (μM)	B-cell TI	T-cell TI	T/B-cell SI
Ibrutinib	**0.0002**	**3856**	**81090**	**5530**
Dasatinib	0.046	0,71	160,0	0,07
Tofacitinib	**0.432**	**>115,6**	**>115,6**	**0,61**
Proflavine	0.59	5,14	8,55	7,50
Amodiaquin	0.66	42,33	39,43	14,73
Hexachlorophene	5.34	0,79	1,10	0,27
Etravirine	5.90	1,35	5,15	1,23
Prochlorperazine	6.12	1,67	3,52	0,76
Triclosan	20.88	1.32	1,25	1.64
Diethylstilbestrol	31.07	1,01	0,22	0,59
Estradiol	33.91	1,06	0,58	0,62
Thyroxine	34.10	>1,47	>1,47	>1,47
Suramin	**35.20**	**>1,42**	**>1,42**	**0,84**
Thyronine	36.82	>1,36	>1,36	1,20
Adenosine	>50	1,00	1,00	1,00
Nicotinamide	>50	1,00	1,00	1,00

Table 4.3: **Potency on B-cell, B-cell Therapeutic Index TI, T-cell Therapeutic Index (TI), and Specificity Index (SI).** The B cell IC50 is given in μM and is the mean of 4 independent experiments. The TI relates toxicity to activity. B cell Therapeutic index (B cell TI) is the ratio between the toxic concentration and active concentration on B-cell and was calculated as IC50 WST-1 assay RPMI1788/IC50 B-cell assay. T cell Therapeutic index (T cell TI) is the ratio between toxic and active concentrations on T-cells and was calculated as IC50 WST-1 assay Jurkat/IC50 B-cell assay. The selectivity index is the ration between active concentrations on T-cells and B-cells and was calculated as IC50 MLR assay/IC50 B-cell assay. A high value of Therapeutic Index corresponds to high activity at low toxicity and is hence desirable. A high value of Selectivity Index SI indicates high specificity to B-cells rather than T-cells. Ibrutinib has high B-cell TI and SI.

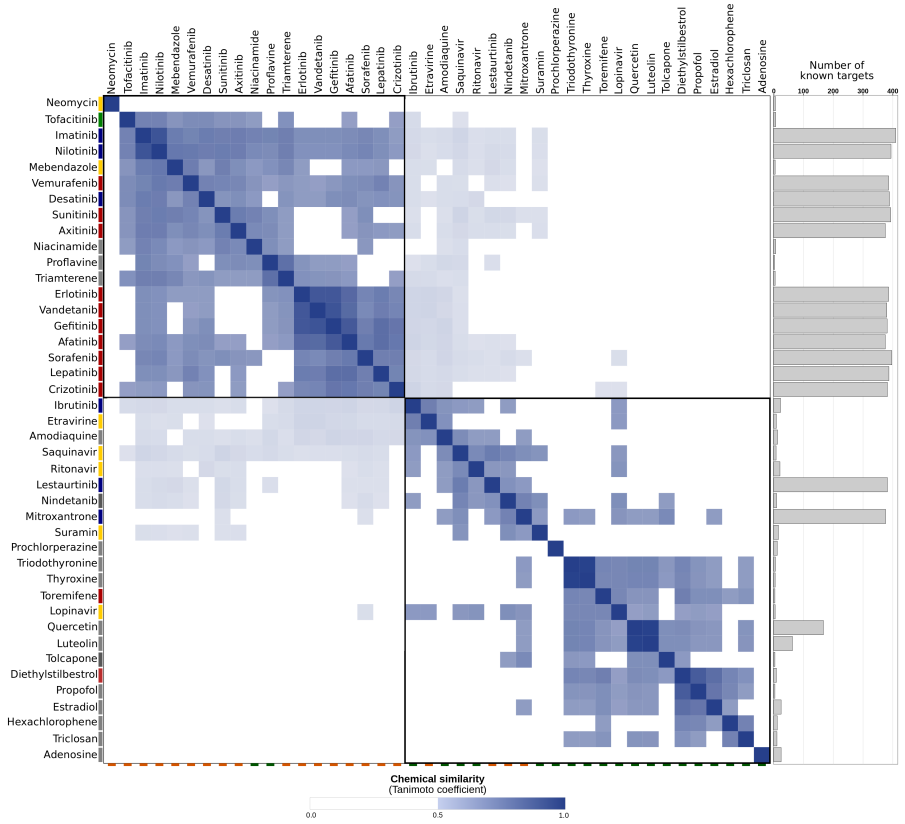


Figure 4.5: **Pairwise chemical similarity of known and predicted inhibitors.** Known inhibitors labeled with orange tag on bottom line and predicted ones with a green tag on bottom line. Dark blue boxes represent high chemical similarity of inhibitors (see Methods). Bars on the left are therapeutic indication (red=cancer, blue=cancer of the immune system, green=autoimmune disease, yellow=infection, and gray=other) and gray bars on the right represent the number of known targets (size of bar). The top left quadrant are mostly known inhibitors, which are chemically similar to each other and are mostly linked to cancer. The bottom right quadrant are new, predicted inhibitors, which have different scaffolds and are hence not derived from the known inhibitors. Consequently, these inhibitors have other therapeutic indications.

Chapter 5

Data integration for drug-protein interaction prediction

In this chapter, the different categories of side information and the process of data integration for biological activity prediction are described. My goal is to assess the contribution of different side information, especially structured-based ones, in ligand-protein interaction prediction using a machine-learning approach. Notwithstanding the results obtained applying structure-based approaches in drug repositioning, as shown in Chapter 4, the impact of structure-based information in machine learning for drug-target interaction prediction is still debatable, and its application remains less spread than other kinds of information easier to acquire and integrate. To achieve my goal, I used a data fusion algorithm previously validated and applied to similar cases in order to predict biological activities. Finally, I analyzed the ligand-protein interaction prediction accuracy as AUC values resulting from different combinations of side information to determine the impact of structure-based side information such as Residues Interaction Network (RIN) for the protein target and PLIP for the ligand-protein interaction. For this chapter, the authors' contribution is the following: Parisi conceived the idea, developed the theory, and performed the computations. Arany verified the analytical methods as well as supervised the work. Simm and Arany developed and modified the algorithm. Schroeder supervised Salentin to generate the interaction profile-based side information. Kalinina supervised Keller to generate the protein-based structural side information. Moreau helped supervise the project.

5.1 Biological activity for compound-protein couples

The biological data describing ligand-protein interaction, which we split into train set and test set, have been obtained from ChEMBL. ChEMBL is a manually curated, open, and large-scale database of bioactive molecules that collects chemical, bioactivity, and genomic data (58). I downloaded the version 23 from May 2017 containing 14,675,320 activities from 1,735,442 compounds interacting with 11,538 targets. After selecting only the biological activity expressed in IC50 (concentration at which half of the maximal registered biological effect is recorded) of small molecule ligands and protein targets, the new data set consisted of 295,820 compounds and 702 protein targets. The remaining interaction data have been transformed from continuous molar concentration values to categorical activity thresholds To help the prediction process. Converting IC50 concentration values to the pIC50 scale, four different activity thresholds have been generated: 5.5, 6.5, 7.5, and 8.5. The higher the threshold, the lower is the active concentration, therefore, the compounds are more active. The active concentration has been considered for each ligand-protein couple, and the negative logarithm has been calculated. When this number appeared to be between two thresholds, 1 was assigned to the lower threshold and all the thresholds below, indicating the activity range for which the compound is active against a specific target. The value -1 was assigned to the higher thresholds, representing an absence of activity for a concentration lower than the active one. As a result, a vector of 4 numbers was generated for each ligand-protein complex. For example, a compound with an activity of 10^{-6} has a negative logarithm of 6, and therefore its activity fingerprint is 1, -1, -1, -1, expressing an active concentration between the first (5.5) and the second (6.5) activity threshold. Data from ChEMBL have been merged with data from PDB through the PLIP database to study the contribution of structural data to biological activity prediction, which reduced the dataset's dimension to 2595 compounds and 314 targets. A matrix has been generated using the 2595 compounds as rows and 1256 columns (314 targets time four activity thresholds), storing 13,528 points of activity. Subsequently, the activity matrix has been converted to a sparse matrix to save the space occupied by all the '0' elements.

5.2 Side information: description and previous applications in drug-target interaction prediction

In this work, heterogeneous data called side information would be integrated using a deep neural network algorithm to understand the impact of different data on bioactivity prediction. In this section, the different categories of side information will be described in detail, focusing on previous applications on drug-target interaction prediction. A summary of those data is shown in Table 5.1.

Category	Name	Num of entities	Num of features	Application
Ligand	ECFP	2595	43,151	Table 5.2
Protein	Amino acid frequency	314	20	(116)
	Sequence similarity	314	314	(117)
	Domain composition	314	540	(15)
	Residues Interaction Network	314	2048, 8192	(71)
Interactions	PLIP	2744	1155	(170; 12)

Table 5.1: **Summary of the analysed side information.** List of the side information integrated with the deep learning algorithm divided into the three categories "Ligand," "Protein," and "Interactions." Per each type of side information, the number of entities (number of ligands, number of proteins, and number of ligand-protein couples) and number of features (number of bits of the corresponding fingerprint) is showed. One or more articles where those data are linked to ligand-protein interaction prediction, drug repositioning, or polypharmacology have been cited in the "Application" column.

5.2.1 ECFP as ligand representation in machine learning

The first category of side information to be described is the ligand-based chemical description of the drugs. Such data are easy and relatively fast to generate, and they have been used for many years as support for drug discovery and repositioning. Machine learning, and most of all, DNN can boost predictive power using small molecules' properties to infer their biological activities. Scalar experimental and calculated properties such as molecular weight, solubility, charge, number of rotatable bonds, atom types, topological polar surface area are one-dimensional attributes that don't take into account direct molecule's bonds patterns. 2D-descriptors focus on covalent and aromatic bonds arrangement, regardless of the spatial coordinates, which are instead represented by 3D descriptors. Converting chemical information into data suitable for machine

learning is a multi-layer computational process crucial for chemoinformatics. A molecule's connectivity can be represented in different ways, such as ASCII character strings called SMILES ("Simplified molecular-input line-entry system") or a fixed-length descriptor vector called fingerprint. Although SMILES are reasonably human-readable, they lack uniqueness since there are usually several equivalent ways to write the same compound. Among the fingerprint-based ways of representing molecules, the circular fingerprints or Extended-Connectivity Fingerprints (ECFP) are the most used. ECFPs are representations of small molecule structures developed to express chemical features relevant to the biological activity. ECFPs are designed using a modified version of the Morgan algorithm, an iterative process that assigns unique identifiers to the atoms (139). For ECFP, the computationally expensive step of creating identifiers to describe in a unique way the atom's environment is replaced by a fast-hashing scheme. Such modification allows for generating identifiers that are comparable across molecules (164). The ECFP generation process follows three different steps: first, an initial assignment stage for an integer identifier and each atom; second, an array is created attaching to each atom's identifier the integer identifies its neighborhood. In the last step, a hash function is applied to map back to an integer from the concatenated neighborhood string, with subsequent duplicate removal, which creates a new fingerprint set of single representatives of 2048 bits (164). ECFPs are easy and fast to generate, they contain both positive and negative information about substructures that are there and that are not, and they can be applied to describe novel structures. Those strengths make them useful for a vast number of applications, some of which have been listed in Table 5.2 such as virtual screening, structure-activity relationship modeling, and compound library analysis. In processing noisy high-volume HTS data, ECFP has proven to be effective at categorizing active from non-active compounds. For example, Rogers et al. (163) coupled ECFPs with Laplacian-modified Bayesian analysis to screen 50,000 compounds against *Escherichia coli* dihydrofolate reductase, showing that, compared to 3D approaches, such method offers surprisingly competitive results with a lower computational cost. In conclusion, ECFPs high dimensionality can be handled (and potentially be advantageous) by some approaches based on Bayesian analysis (77) or Tanimoto similarity (37), in contrast with other methods such as linear regression or neural networks, where the high number of features might cause overfitting of the data (164). Other ways to describe ligands, differently than ECFP, are the Coulomb matrix (169), which combines the nuclear charges of a molecule and their coordinates, the Grid featurizer (219), which includes structural features of both the ligand and the target protein as well as their physicochemical interactions, the Symmetry function (24), which uses a series of radial and angular symmetry functions to describe the distance between atom pairs and the on angles formed within triplets of atoms, the Graph convolution (219),

which represents the local chemical environment of an atom using an initial feature vector and a list for each neighbor atom, and the Weave (219), which forms a feature matrix from feature vectors calculated by each pair of atoms in the molecule and their connectivity.

Application	Method	Reference
Virtual Screening	Support vector machines	Glick et al., 2004 (62)
	Naive Bayes	Rogers et al., 2005 (163)
	Tanimoto similarity using nearest neighbor	Hert et al., 2005 (76)
	Multiple machine learning methods	Chen et al., 2007 (38)
	Multicategory Naivve Bayes	van Hoorn et al.' 2009 (201)
Structure Activity Relationship Modeling	Naive Bayes	Yang et al., 2003 (229)
	PLS regression	Gedeck et al., 2006 (59)
	Gaussian kernel weightedk-nearest neighbor	Jensen et al,. 2007 (85)
	Similarity ensemble approach	Hert et al., 2008 (75)
	Support vector machines	Wale et al., 2009 (207)
Compound Library Analysis	Naive Bayes	Klon et al., 2004 (98)
	Scaffold self-similarity	Krier et al., 2006 (100)
	Maximal dissimilarity clustering	Steindl et al., 2007 (189)
	Clustering with Tanimoto similarity	Schuffenhauer et al., 2007 (180)
	PLS regression	Sullivan2008 et al., 2008 (192)

Table 5.2: **List of machine learning-based works using ECFPs.** A list of works on drug-target interaction prediction using ECFPs as molecular descriptors for machine learning. For each category, the names are sorted by year of publication.

5.2.2 Protein-based side information: amino acid frequency, sequence similarity, domain presence and residues interactions network (RIN)

Different kinds of information have been used to describe the target space. Genetic information about the protein sequence and topological information about the protein structure have been alternatively used to check their impact on the ligand-protein interaction prediction process. In regard to the amino acid frequency, amino acid composition has already been used as side information for drug-target interaction prediction. *Li et al.* used the frequency of amino acids to describe the protein space in a Bayesian additive regression trees-based model with positive results (116). However, this type of data doesn't consider any information about the order and connectivity of the protein sequence, discarding highly informative characteristics that might help the algorithm to learn about protein behavior with ligands. Another way to describe the target space is the degree of similarity of the different proteins. Assuming that two proteins with a similar sequence might have similar binding behavior and might show an overlap of the pharmacological profile, it's possible to make repositioning hypotheses

based on sequence similarity. *Li et al.* used the Needleman-Wunsch algorithm (145) to align protein sequences and predict drug-target interactions with the 'guilt-by-association' principle and topology features from interactome network (117). The third way used to describe the targets is domain presence. Protein structures are functionally organized in domains, intended as modular portions of the proteins essential for their functions. Domains are pretty conserved, and they have their evolutionary history. Protein binding sites are usually conserved domains or exist in the intersection between two domains. *Alaimo at al.* used domain-dependent knowledge to improve drug-target interaction predictions through a method called DT-Hybrid (15). For this reason, starting from the assumption that the composition in domains of the targets might contain useful information on the ability to bind specific ligands, the presence or absence of domains has been used as side information also in our case. Besides biological and genetic information related to the protein sequence, structural information about the 3D arrangement of the protein can be used as a source of knowledge. *Haupt et al.* demonstrated that similarity of the structures of binding sites strongly correlates with drug promiscuity and that 3D structural data about protein active sites should be included in drug repositioning pipelines (71). For that reasons, I used a RIN generated by Olga Kalinina's group at Helmholtz Institute for Pharmaceutical Research Saarland to detect structural motifs inside each binding sites, and we used binary fingerprints to describe the absence or presence of such motifs in the target structure.

5.2.3 PLIP for ligand-protein interaction profile representation

The ligand-protein interaction profile has been used to describe each ligand-protein interaction, corresponding to a cell of our sparse matrix of ligand-protein couples. Such data contains structural information about the biochemistry of the specific binding environment but requires available 3D structures of complexes. Starting from a 3D protein-ligand complex, PLIP can detect, visualize, and describe relevant non-covalent protein-ligand contacts. The process is based on four automated steps: structure preparation, functional characterization, rule-based matching, and filtering of interactions (171). In the first step, hydrogen atoms are added to the input structure and ligands extracted along with their binding sites, using OpenBabel (149) for internal representation of molecules and other chemoinformatics calculations. A specific list, available on the PLIP website (5) is used to exclude preparation artifacts, modified residues, ions, and solvent molecules to retain only valid binding small molecules. After the preparation step, a functional characterization to identify the different types of bonds is performed. Hydrogen and halogen bonds rely on the detection

of hydrophobic atoms and acceptors/donors, while for p-stacking, p-cation interactions, or salt bridges, aromatic rings and charge centers in protein-ligand are searched. Geometric criteria (such as distance or angle constraints between sets of atoms) are applied to match putative interactive groups based on knowledge-based thresholds retrieved from large sets of high-quality protein structures present in literature. In the end, the filtering step aims to eliminate redundant, overlapping interactions. As a result, the interaction fingerprint is generated as binary vectors, where each feature represents a combination of interactions within an angle and distance, as summarized in Figure 5.1. PLIP authors performed a test with 30 literature-validated examples checking whether the software could detect all interactions reported in the corresponding paper (171). Later, PLIP has been used to perform a structure-based virtual screening to find an inhibitor for the chaperone Hsp27. According to the similarity of the target ligand's brivudine interaction profile, the antimalaria drug amodiaquine has been identified as a new hit compound, highlighting the repositioning approach's potential (170).

5.3 Algorithm description

The data fusion algorithm is based on the SparsChem codebase, an open-source code developed in our lab and available on https://github.com/melloddy/SparseChem. It belongs to the family of deep learning algorithms, and it has been explicitly designed for biochemical applications. It mainly supports very high-dimensional models with sparse inputs, e.g., millions of features and millions of compounds, with specific extensions for protein-compound couple features such as PLIP interaction profiles. Matrix factorization algorithms with side information such as Macau have inspired it (185). The algorithm was designed to predict biological activity from ECFP using efficient sparse matrix computation, uses a ReLU activation function (142), using weight decay and dropout (188). The algorithm's architecture is described in Figure 5.2 and the code has been reported in Appendix A. Such architecture results in a fast and accurate machine learning method able to support very high-dimensional models with sparse inputs, e.g., millions of features and millions of compounds, with no need to make ECFPs smaller by cutting or folding them. Notwithstanding its performance, the network is relatively simple and easy to describe. The drug-target interaction system is described as a matrix where each row refers to a single ligand and each column to one of the four activity thresholds for one of the protein targets. Each of the elements gets connected to its related ligand-based or protein-based side information, while the interaction-based information describes single cells inside the matrix referring to the specific ligand-protein couple. How the algorithm integrates the different information

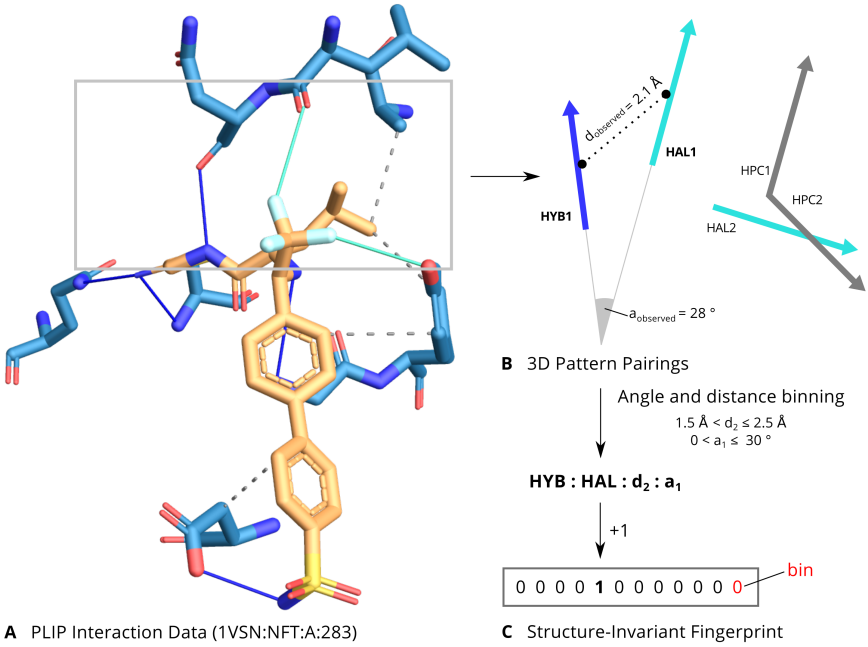


Figure 5.1: **Structural drug-repositioning with PLIP interaction fingerprints. A.**Schematic for the PLIP interaction fingerprints. Raw interaction data from PLIP are calculated for a complex. **B.**For each pair of interaction vectors, their distances and angles are calculated. **C.**Finally, the geometric measures are binned and combined with the types of pairing into a string representation. Each such combination corresponds to a fingerprint feature, which is incremented for each observation of the encoded pairing. The resulting fingerprint is structure-invariant since it only considers the interaction vectors.

and handles the gap between structural and non-structural side information is shown in Figure 5.3. This algorithm faces some limitations, such as handling missing values in compound and protein descriptors. Even though this code was developed specifically for this application, a related code has been tested already on the ChEMBL database to predict biological activity from ligand features. It is now used in the MELLODDY project, an IMI project aiming to enhance machine learning models' performance on decentralized data of 10 pharmaceutical companies without exposing proprietary information.

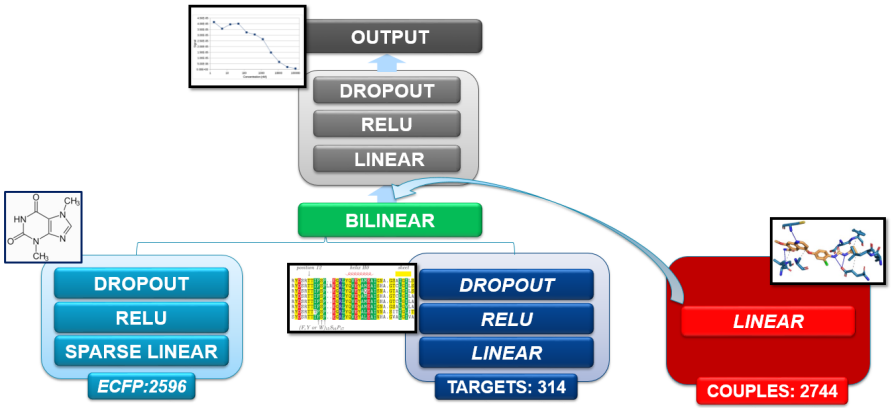


Figure 5.2: **Architecture of the deep-learning algorithm used for data fusion.** The structure of the algorithm is described in this figure. Each side information enters as input (bottom side) inside each block of code separately. Ligand side information (light blue) and target side-information (dark blue) are processed first separately by a (sparse) linear model coupled followed by a ReLu activation function and a dropout layer. The result enters a bilinear layer, and the output is concatenated with the interaction profile side information (red), previously processed by a linear function. Altogether, the results of learning from ligand, protein, and interaction profile side feature become inputs for a final block of linear-ReLu-Dropout functions, which will deliver the final prediction. The numbers of features per each side information have been reported on the bottom of each block.

5.4 Methodology

5.4.1 Software

The side information validation and analyses and the algorithm were run using Python 3.6.8 (Dec 30, 2018) on Linux Ubuntu 16.04 LTS (2019). The data analysis and representation were done using JupiterLab (2019) and the following Python libraries: PyTorch for the machine learning; pandas, pumpy, scipy, and sparse for the data analysis; sciKitLearn for the features extraction; matplotlib for the data visualization. The algorithm has been run with different sets of ligands, proteins, and protein-ligand side information.

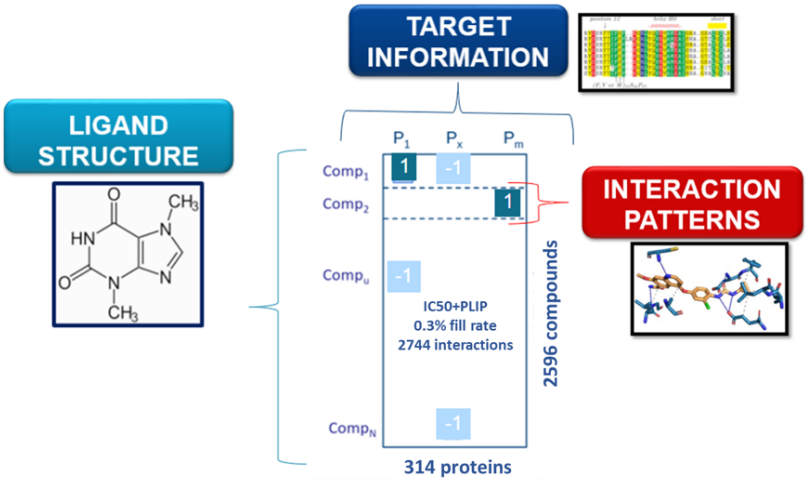


Figure 5.3: **Ligand-protein interaction system representation.** The figure shows how the data fusion algorithm represents the different ligands, targets, and their interactions. Every row of the matrix represents a compound, the protein targets occupy the columns, and the interaction pattern structural information describes several cells of the matrix.

5.4.2 ECFP-based binary fingerprints generation and validation

Using Python, the list of ligands with biological activity available on ChEMBL (version 23) and interaction patterns stored on PLIP (December 2018) has been generated. The 2595 compounds were loaded into a data frame using the Python's library for data manipulation and analysis Pandas, the module rdkit.Chem.PandasTools specific for chemoinformatics and the command LoadSDF(). From the compound data frame, the list of SMILES has been selected and used as input to generate the ECFPs. The circular fingerprints were generated using AllChem.GetMorganFingerprint(), specifying 3 as radius. From the total amount of 43,151 features generated for all the compounds, a binary fingerprint has been created describing the presence (1) or absence (0) of a certain feature in a single molecule. Those binary fingerprints have been successively used as one of the side information for the biological activity prediction. To confirm the presence of information useful to predict biological activity against a target stored within the binary fingerprints, we evaluate the predictive power of ECFP using a sparse flow model. First, a sparse matrix has been used to store the binary fingerprints and the biological activity of the same

ligands against 314 targets. Similar to what concerned the ligands, those targets have been selected because of the availability of experimental information from ChEMBL and interaction profiles on PLIP. Even though the biological activity on ChEMBL is expressed as a continuous value named IC50, concentration for which half of the maximum biological effect has been registered, to improve the prediction performance, we decided to turn it into a classification problem. We defined four thresholds of activities and tried to predict whether a ligand might be active in that range of concentration (1) or not (-1). Zeros were stored in case of the absence of measurements.Subsequently, the SparseFlow (6) model has been trained on our data and tested to predict each threshold's binary value. After a manual optimization, the best hyperparameters have been plotted with boxes showing the best five repeats.

5.4.3 Protein-based side information: fingerprint generation

To describe the amino acid frequency of our 314 unique protein targets obtained by mapping the ChEMBL and PLIP database, a matrix showing per each target the level of presence of each of the 20 amino acids has been generated. For the sequence similarity of our 314 targets, a similarity matrix has been generated using the Needleman-Wunsch algorithm. The similarity values generated have been normalized as $(X - \mu)/\sigma$, where X is the similarity value, μ represents the mean and σ the standard deviation. The final 314x314 matrix containing the normalized similarity scores has been generated with Python and used as protein side information. Regarding the domain presence, the descriptive matrix has been generated in two steps. First, the web tool Pfamscan (4) has been used on each protein target to make a global list of domains identified comparing the amino acid sequence with a list of functional regions (Pfam). Each protein was then described with a fingerprint expressing the presence (1) or absence (0) of any domain among the total 540 detected in the complete list of targets. The RIN fingerprint generation process is longer than the previous ones because of the structural information's higher complexity. First, a set of resolved protein structures that corresponds to the Uniprot IDs of the 314 targets were determined. This set should cover as much of the UniProt sequence as possible while avoiding large overlaps of different structures for the same sequence. After that, mutated residues found in the structures are reconverted to the residues corresponding to the UniProt sequence. Subsequently, binding pockets in each structure are determined using the web tool Concavity (1). Individual residues are assigned with a score and then filtered according to a specific threshold. For each of these structures, a network of the residue interactions was calculated using RINerator (51). Only the RIN vertices corresponding to the previously identified binding pockets and residues in contact with those of

the binding pockets were selected. These graphs from the different structures corresponding to a single UniProt entry were combined into a single disconnected graph. A database of all the graphs from all the targets was created, and the discriminative subgraphs were determined. In this way, many subgraphs are collected since every subgraph of an included pattern will be included as well. Also, such subgraphs will be highly correlated with the graphs they occur. Therefore, clustering has been applied based on the Jaccard distance of each subgraph's occurrence vector to select only the most representative subgraph of each cluster. Finally, the fingerprints have been generated for each UniProt entry based on the presence or absence of the subgraphs chosen in the previous step. Based on the selection thresholds, many subgraphs can be identified, impacting the fingerprint length. To understand the impact of such threshold on the information stored within the data we generated two sets of fingerprints with different length for each target: one consisting of 2048 bits generated by applying a higher threshold and considering only high-resolution structures, and an 8192 bits-long fingerprint by including also lower resolution structures. A logistic regression model has been used to find a putative correlation between PLIP fingerprints and IC50 for the ligand-protein couples.

5.4.4 PLIP-based binary fingerprint generation and validation

Protein-ligand interaction fingerprints for machine learning have been obtained by selecting protein-ligand interaction features from the PLIP database (November 2018). This repository contains information about 188.633 protein-ligand complexes from PDB in the form of bins associated with the biochemistry and geometry of the interactions. An example of rows from the PLIP database file is shown in Table 5.3, and the detailed explanation of the bins is shown in **??** The protein-ligand couples from PLIP have been mapped to ChEMBL IDs, and only those for which a biological interaction is present in the ChEMBL database (version 23) have been retained. For the resulting 2744 couples, a list of binary fingerprints has been generated, describing the presence or absence of any of the 1155 feature bins shown in Table 5.5. A sparse matrix has been created, consisting of 2595 rows (number of ligands) and 362.670 columns (314 targets x 4 activity thresholds x 1155 fingerprint bits compressed in order not to store the columns made by all zeros) with 62076 descriptive elements stored for 2744 ligand-protein complexes. Each cell of this matrix will be used as structural side information to describe every single cell of the protein-ligand sparse matrix containing the ChEMBL biological activity. The PLIP fingerprints containing structural information about protein-compound interactions have been tested for predictive power concerning ChEMBL biological activity (version 23) using logistic regression (LogisticRegression from sklearn.linear_model).

We classified the ligand-protein couples according to four categories of biological activity values: compounds active at a concentration lower or equal to $10^{-5.5}$ M (threshold 5.5), those active at a concentration lower or equal to $10^{-6.5}$ M (threshold 6.5), compounds active at a concentration lower or equal to $10^{-7.5}$ M (threshold 7.5), and the ones active at a concentration lower or equal to $10^{-8.5}$ M (threshold 8.5). For the first threshold, 5262 structural fingerprints from the PLIP database have been analyzed; for the second threshold, 5272 ligand-target couples from the PLIP database have been analyzed; for the third threshold, 5267 PLIP interaction profiles have been analyzed, and for the last threshold 5228. The total number of right and wrong predictions have been plotted on a confusion matrix per each threshold, as shown in Figure 5.5, and the predictive performance has been expressed with ROC curves and AUC in Figure 5.6. Subsequently, the same logistic regression model has been applied to the sparse matrix of fingerprints for each couple ligand-protein previously generated. An AUC score of 0.70937 was obtained, confirming a positive signal in PLIP data, which allowed us to use that structural information as side information in the deep learning-based data fusion.

Target	**Ligand**	**Chain**	**Residue**	**Bin**
1JIP	HEM	A	410	1
5B2Y	W10	B	502	2
2P5G	DG	H	101	64

Table 5.3: **Example of rows from PLIP database.** Each row in the database contains a bin that describes specific characteristics of the protein-ligand interaction profile. Each bin is related to a specific amino acid ('Residue') from a specific protein chain ('Chain') where the compound ('Ligand') interacts with the protein ('Target'). PLIP uses 3D data from PDB to extract interaction features, and therefore PDB IDs are used for targets and ligands.

5.4.5 Data integration process, training and ranking

The list of ligand ECFP was transformed into a sparse matrix using scipy.sparse.coo_matrix() and fed into the algorithm as .mtx file using the command scipy.io.mmread() (see Appendix A line 93). The same process was applied to the IC50 matrix (see Appendix A line 94) and PLIP side information matrix (Appendix A line 91). The protein-based side information consisted of two tsv files defining the list of targets (Appendix A line 84) and the specific target property (Appendix A line 83). ECFP and IC50 matrices are essential to run the algorithm, while protein-based and PLIP files can be omitted to try a

Bin	Target atom interaction type	Ligand atom interaction type	Minimal distance	Maximal distance	Minimal angle	Maximal angle
1	Halogen	Halogen	0	3	0	60
2	Halogen	Halogen	0	3	60	120
64	HB donor	HB donor	0	3	0	60

Table 5.4: **Explanation of PLIP bins.** Each bin describes the biochemistry and geometry of the interaction uniquely, carrying important 3D structural information. It is associated with a combination of two-atom types related to the target ('Target atom interaction type') and the ligand ('Ligand atom interaction type') and a specific geometric configuration. The spatial information is described by a range of angles ('Minimal angle' and 'Maximal angle') and distance ('Minimal distance' and 'Maximal distance').

Target	Ligand	Features	Fingerprint
CHEMBL2474	CHEMBL1738934	85, 88, 89, 232, 235, 236, 256, 257, 261, 274...	0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
CHEMBL3117	CHEMBL3763993	148, 177, 421, 425, 428, 463, 466, 467, 470, ...	0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
CHEMBL3717	CHEMBL2170968	153, 232, 236, 256, 257, 341, 345, 421, 425, ...	0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...

Table 5.5: **Protein-ligand interaction features and fingerprints.** From the list of features stored on the PLIP database ('Features') for each protein-compound complex, a binary fingerprint of 1155 bits has been generated ('Fingerprint'). Each number in the fingerprint describes the presence (1) or absence (0) of one of the 1155 PLIP features.

different combination of side information by turning into False the parameters prot_features (Appendix A line 86) and use_plip (Appendix A line 120). For the training process, information can be found in the code in Appendix A. The machine learning approach used is a classifier method that provides for each run a list of AUC values as output, subsequently used to generate a rank.

5.4.6 Parameter optimization

For each combination of features, the algorithm's parameters have been optimized manually to reach the highest AUC according to the size and density of the fingerprints. The parameters affecting the algorithm performance are listed in Appendix A from line 109 to line 121. The optimized values are shown in Table 5.6. After the optimization process, the algorithm was run five times, and the best AUC from each run was selected. The resulting 5 AUCs were represented with a boxplot to describe the prediction's quality of side information contribution.

Side information category	Hidden size c	Hidden size p	Hidden size bi	Hidden size plip	Learning rate	number epochs	dropout 1	dropout bi	Weight decay	Min. num. positive	Min. num. negative
ECFP	700	300	200	/	5e-3	50	0.6	0.4	2e-2	10	10
ECFP + Sequence similarity	900	700	200	1	5e-3	50	0.8	0.5	0.1	10	10
ECFP + Amino acid frequency	900	600	300	1	5e-3	50	0.8	0.5	2e-2	10	10
ECFP + Domain presence	900	900	200	1	5e-3	50	0.8	0.5	0.05	10	10
ECFP + RIN 2048	900	700	200	1	5e-3	50	0.8	0.8	0.1	10	10
ECFP + RIN 8192	900	600	200	1	5e-3	50	0.8	0.6	0.1	10	10
ECFP + PLIP	900	35	500	900	5e-3	50	0.7	0.8	2e-2	10	10
ECFP + Seq. sim. + PLIP	900	400	400	800	5e-3	50	0.5	0.5	2e-2	10	10
ECFP + AA freq. + PLIP	900	500	500	900	5e-3	50	0.5	0.9	2e-2	10	10
ECFP + Dom. pre. + PLIP	900	500	500	900	5e-3	50	0.5	0.9	0.05	10	10
ECFP + RIN 2048 + PLIP	900	800	500	900	5e-3	50	0.5	0.7	0.05	10	10
ECFP + RIN 8192 + PLIP	900	500	600	900	5e-3	50	0.5	0.7	0.1	10	10

Table 5.6: **List of optimized parameters for each side information category**. Manual optimization consisted of running the algorithms with a different set of parameters and tuning each value to modify the resulting AUCs' trend. The best set of parameters listed here have been subsequently used to check data fusion performance for each side information group.

5.5 Results

5.5.1 Side information validation

After generating the ECFP fingerprints for the small molecule ligands, their correlation with IC50 has been tested. Figure 5.4 shows a high correlation (higher than 90% for the best five points) between our fingerprints and the biological activity. Regarding the protein-ligand interaction profile validation, the total number of right and wrong IC50 predictions using PLIP have been plotted on a confusion matrix per each threshold, as shown in Figure 5.5, and the predictive performance has been expressed with ROC curves and AUC in Figure 5.6. Subsequently, the same logistic regression model has been applied on the sparse matrix of fingerprints for each couple ligand-protein previously generated. An AUC score of 0.70937 was obtained, confirming a positive signal in PLIP data, which allowed us to use that structural information as side information in the deep learning-based data fusion.

5.5.2 Data fusion

For each different combination of ligand-, protein- and interaction-based profile side information, the data fusion algorithm has been run five times with the optimized parameters to predict biological activity. The best 5 AUCs per

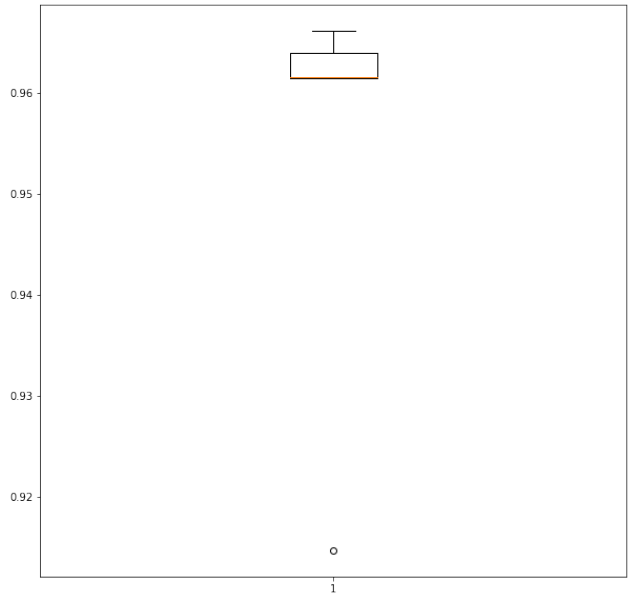


Figure 5.4: **Predictive performance using ECFP-based binary fingerprints and biological activity.** The boxplot shows the best Area Under the Curve Area Under the Curve (AUC) from five runs of the sparse flow model for predicting the biological activity of our set of ligand-protein couples using ECFP-based fingerprints.

combination of side information are listed in Table 5.7. Figure 5.7 summarizes those results with boxplots.

5.6 Discussion

The results plotted in Figure 5.7 show a very mild gap among the different side information categories, corresponding to a delta of 0.0524 between the lowest and the highest values from the set of best AUCs listed in Table 5.7. However, a weak trend is visible and worthy of discussion. In fact, beyond the blue horizontal line, the box plots representing prediction accuracy using structure-based side information show inferior ability in improving the ligand-based prediction chosen as the baseline (1). Genetic-based information (mostly amino acid frequency (3)) and evolutionary-based data (protein domain presence (4)) seem to improve the performance of prediction when coupled with ECFP, while the addition of

Features			Best AUC					Median
Ligand	**Protein**	**Interaction**	**Run n.1**	**Run n.2**	**Run n.3**	**Run n.4**	**Run n.5**	
ECFP	Domain presence	/	0.7663	0.7676	0.7712	0.7747	0.7761	0.7712
ECFP	Amino acid frequency	/	0.7618	0.7646	0.7666	0.769	0.7817	0.7666
ECFP	Amino acid frequency	**PLIP**	0.7468	0.7546	0.7547	0.765	0.7678	0.7547
ECFP	**RIN 8192**	/	0.7419	0.7467	0.7516	0.7546	0.7645	0.7516
ECFP	/	/	0.7488	0.7494	0.7511	0.7542	0.7570	0.7511
ECFP	**RIN 2048**	/	0.7479	0.7500	0.7508	0.7526	0.7540	0.7508
ECFP	Sequence similarity	/	0.7447	0.7461	0.7496	0.7527	0.7552	0.7496
ECFP	**RIN2048**	**PLIP**	0.7481	0.7491	0.7495	0.7511	0.7591	0.7495
ECFP	/	**PLIP**	0.7410	0.7450	0.7470	0.7473	0.7558	0.7470
ECFP	Domain presence	**PLIP**	0.7443	0.7452	0.7463	0.7473	0.7519	0.7463
ECFP	**RIN8192**	**PLIP**	0.7343	0.7366	0.7421	0.7455	0.7571	0.7421
ECFP	Sequence similarity	**PLIP**	0.7293	0.7298	0.7321	0.7400	0.7405	0.7321

Table 5.7: **List of best AUCs per each side information category**. The side information combinations, their five best AUCs, and their corresponding median are listed here. The entries have been ordered according to the values of the median in a descending way. The algorithm was run with the parameters listed in Table 5.6 and the ability to predict ChEMBL values of ligand-protein interaction have been recorded as AUC plotted in Figure 5.7. Structure-based features have been highlighted in bold.

structure-based features in general decreases the prediction accuracy. In fact, notwithstanding the excellent correlation between PLIP fingerprints and IC50 showed in Figure 5.5 and Figure 5.6, it appears that adding interaction-based data as PLIP fingerprints the best AUC always remains lower than the same ligand-based and protein-based data combination without PLIP as shown in Table 5.8. This behavior could be explained by the complexity of structure-based data, which gives the algorithm more properties to learn (since the fingerprint describes a 3D space and various biochemical interactions) without providing enough information to learn from. It must be said that The side information considered in this work is meant to cover different characteristics of the ligand-protein interaction system, and they have been chosen not only for the quality of data provided but also for the ability to be easily turned into a machine-readable fingerprint. This is mostly valid for the structural side information about proteins and interactions, for which many other tools could have been used. For example, PLIP represents a perfect candidate for the machine-learning approach for the availability of ligand-protein interaction in machine-readable easily retrievable from its database. That does not imply that PLIP might be the only suitable tool for this side information. Future works exploring different structural side information and different algorithms are necessary further to highlight the potential of structure-based data in machine learning to improve drug-target interaction prediction and help drug repositioning. Another point of discussion concerns the output format. The result of each run is a list of probability values in the form of AUC, and even though the ROC curve can be more informative about the overall performance, the algorithm has been designed to predict drug-target interactions for a large number of points. For that reason, it does not allow to generate a single ROC curve for each protein

target considered here.

Features		Median	
Ligand	**Protein**	**Without PLIP**	**With PLIP**
ECFP	Domain presence	0.7712	0.7463
ECFP	Amino acid frequency	0.7666	0.7547
ECFP	RIN 8192	0.7516	0.7421
ECFP	/	0.7511	0.7470
ECFP	RIN 2048	0.7508	0.7495
ECFP	Sequence similarity	0.7496	0.7321

Table 5.8: **Comparison of algorithm performances with and without interaction-based side information**. In this table, the results of ligand-protein interaction prediction with and without PLIP fingerprints' contribution are shown. For each combination of side information, the addition of biochemical interaction description leads to a lower median. The entries are sorted for the Median without PLIP in descending order.

5.7 Conclusions on data fusion with structural side information

Over the last years, machine learning techniques have supported computational ligand-protein interaction prediction for drug repositioning. Using data fusion and deep learning, scientists have used genetic and pharmacological data on both ligands, proteins, and their interaction space. However, the contribution of structural information seems to be minimal at this moment. With this work, for the first time, a comparative assessment of the impact of structural data on the prediction accuracy and machine learning algorithm performance has been conducted by exploring the contribution of different kinds of information in predicting small molecules' biological activity. The results showed in this chapter confirm that, indeed, an in-depth neural network approach designed for non-structural information integration struggles in taking advantage of the richness of 3D data. Since the data validation showed a clear correlation between the ligand-protein interaction-profile similarity and the biological activity, it has been hypothesized that the causes of this adverse effect might be linked to the complexity of the data and the scarce availability of 3D information. The quick progression of computational and experimental techniques in structural biology will soon impose considerable pressure on computational drug repositioning, demanding specific algorithms to accommodate a large amount of structural information.

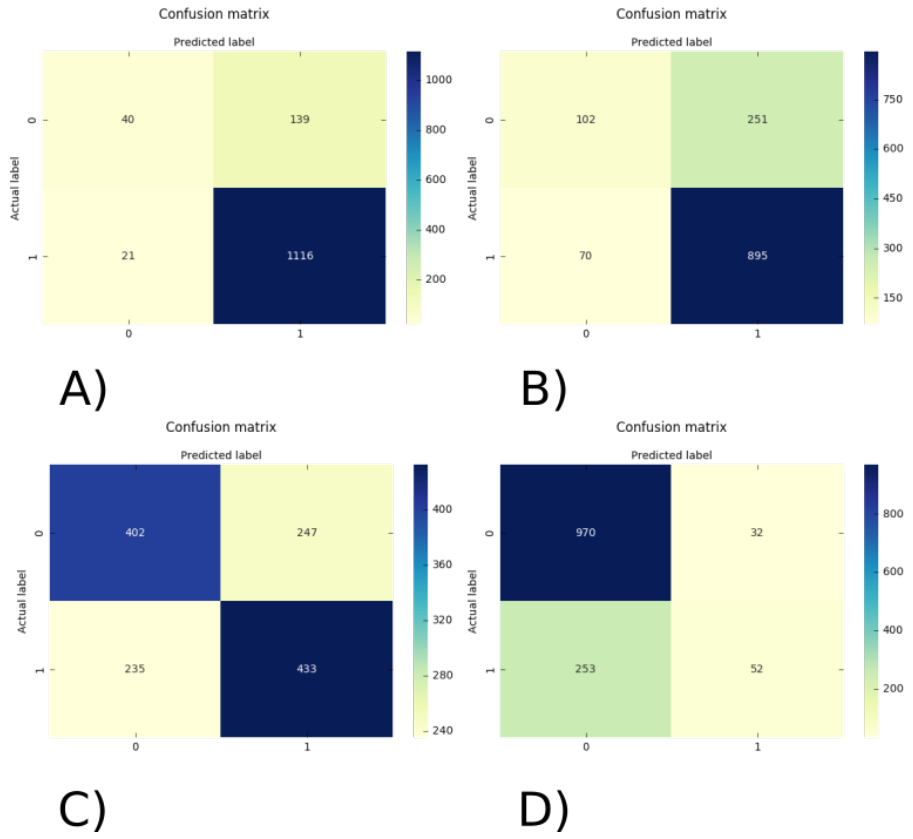


Figure 5.5: **Confusion matrices for biological activity prediction using PLIP fingerprints on each activity threshold.** A confusion matrix is used to describe a classification model's performance on a set of test data for which the actual values are known. It compares the predicted values with the actual data. The rows are labeled with real data (negative on the top row, positive on the bottom one), while the columns are labeled with predicted outcomes (negative on the left and positive on the right). Their combinations generate four different cells representing real negative cases (top left), false positive (top right), false negative (bottom left), and true positive (bottom right). A good classification model maximizes true negative and true positive while aiming to reduce as much as possible the false negative. The application of logistic regression for PLIP to predict biological activity for each IC50 threshold is represented here: **A.** shows the confusion matrix for compounds active at the threshold 5.5, **B.** shows the confusion matrix for compounds active at the threshold 6.5, **C.** shows the confusion matrix for compounds active at the threshold 7.5, and **D.** shows the confusion matrix for compounds active at the threshold 8.5. The true positive or true negative is higher than the false positive and negative values for all thresholds.

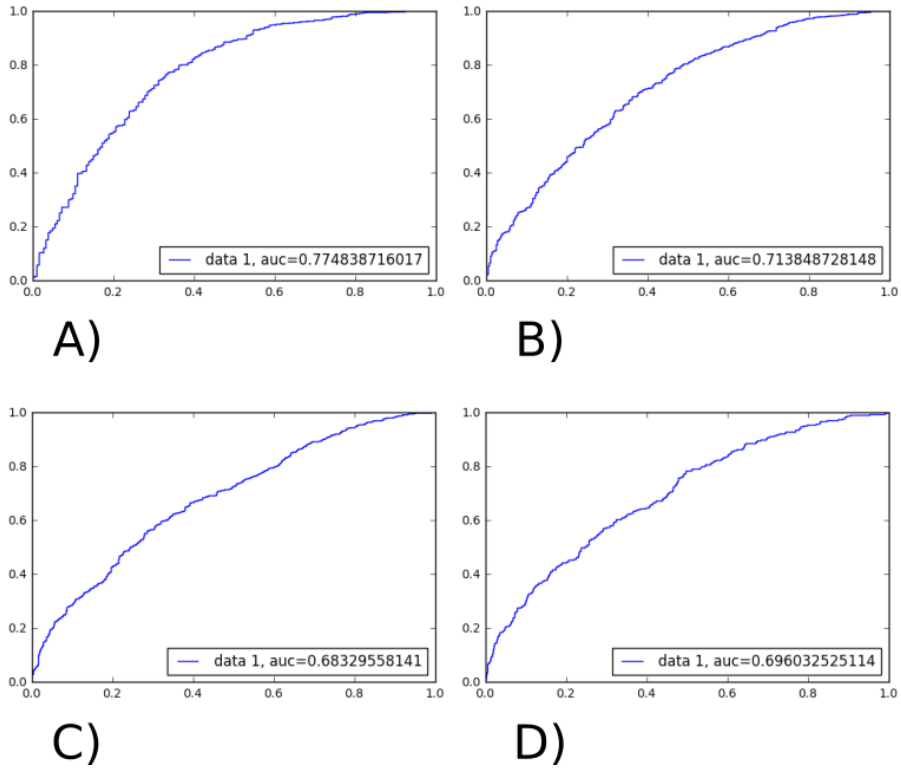


Figure 5.6: **ROC curve for biological activity prediction using PLIP fingerprints on each activty threshold.** The ROC curve is generated by plotting the true positive rate (y-axis) against the false positive rate (x-axis), and its area (AUC) expresses the rate of successful classification. **A.** shows the ROC curve for predictions at the threshold 5.5, **B.** refers to the threshold 6.5, **C.** to the threshold 7.5, and **D.** for compounds active at the threshold 8.5. For all the data sets, the AUC is higher than 0.5, confirming that useful information to predict biological activity is present in PLIP fingerprints at all the level of biological activity.

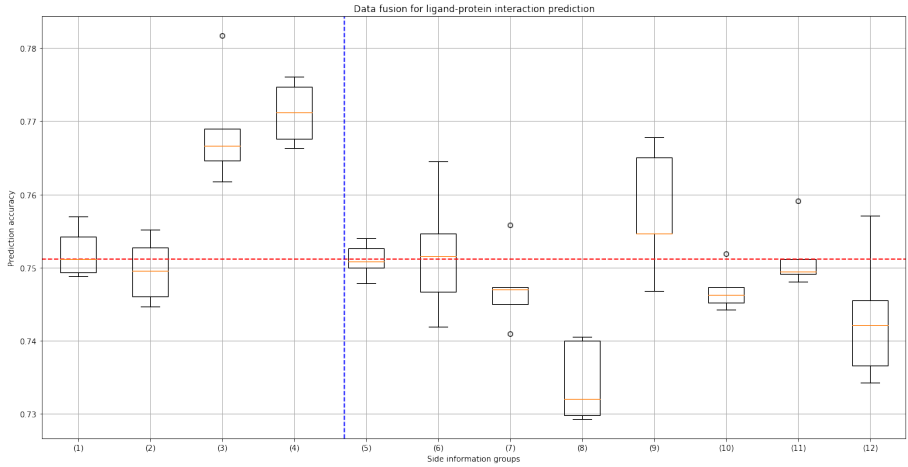


Figure 5.7: **Accuracy of ligand-protein interaction prediction.** The data fusion algorithm was run five times per each side information combination (listed on the x ax), and the best AUCs have been used to generate boxplots to describe the accuracy of ligand-target interaction prediction (y ax). The side information categories from left to right are (1) ECFP; (2) ECFP and sequence similarity; (3) ECFP and amino acid frequency; (4) ECFP and domain presence ; (5) ECFP and residues interaction network 2048 features; (6) ECFP and residues interaction network 8192 features ; (7) ECFP and PLIP; (8) ECFP, sequence similarity and PLIP ; (9) ECFP, amino acid frequence and PLIP; (10) ECFP, domain presence and PLIP; (11) ECFP, residues interaction network 2048 features and PLIP; (12) ECFP, residues interaction network 8192 features and PLIP. Boxplots on the right of the blue dashed line ((5), (6), (7), (8), (9), (10), (11), (12)) have been generated using one or more structure-based features as side information. The red dashed line corresponds to the median of the ECFP only (1) prediction.

Chapter 6

Conclusions

Machine learning is an ensemble of mathematical and statistical approaches able to learn from big data sets and make accurate predictions. Machine learning techniques have been applied to several different fields, including the pharmaceutical one. For example, in drug repositioning, machine learning algorithms have been used to predict drug-protein interactions with many successful results. Moreover, as shown in Chapter 3, the current status of drug repositioning witnesses an unbalanced inclination for target-repositioning and disease-centric repositioning over drug-repositioning, suggesting the need to explore more advanced ways to infer novel drug-target interactions. However, the absence of data sets storing negative information about drug-target interactions, the difficulties in transforming a large amount of complex pharmacological information in machine-readable formats (such is the case of structure-based features), and the difficulties for non-experts to interpret the results represent major limitations in the application of machine learning in drug-repositioning. In this work, I explored drug-target interaction prediction performances when learning from different types of ligand-based, protein-based, and interaction-based information. Also, I assessed the contribution of side information describing 3D structural features of the system's entities. The data considered in this work described the drug-target interaction space, focusing on different characteristics: ECFPs delineate the 2D chemical space of the ligands, the frequency of amino acids and the sequence similarity describe the genetic of the proteins, the presence of specific domains classify the targets based on their hypothetical pharmacological behavior, the residues interaction network come into play by delineating the 3D space of the binding site, and PLIP fingerprints describe the physicochemical characteristics of the interaction profile. The main research question is if using structure-based data the deep neural network would

perform better in the prediction process. As shown in Chapter 4, interaction-profile similarity contains a proper amount of structural information that can be successfully used in drug repositioning. However, data fusion results show a decline in the AUC when PLIP side information is added, suggesting a situation where the algorithm overfits due to data describing every single drug-target couple's details in their specific biochemical interaction profile. Concerning structure-based protein side information RIN when no data on interaction profile is added, they show a similar score than ECFP alone. Another piece of binding site-related information concerning the mere presence or absence of a specific domain in a protein seems to outperform the RINs; this suggests that this algorithm might suffer depending on the size of information provided to learn from when structural data are added. This work showed a detailed retrospective analysis of drug-target interaction prediction in drug repositioning and represented, to the best of my knowledge, the first systematic attempt of assessing the impact of structural information in machine learning-based drug repositioning. However, more recent and comprehensive databases of repurposed drugs, and a more precise link between the repositioning cases and the repositioning techniques, would give a better, more accurate, picture of the current status of drug repositioning. Regarding the data fusion process, the work is based on the prediction accuracy of only one algorithm, previously designed for integrating non-structural data. Future analysis would most likely benefit from the comparison of different algorithms, possibly optimized for 3D information. in the same way, using different kinds of structural data to describe the protein targets and the interaction profile with ligands might help to assess the weight of each specific side information on the prediction accuracy. One of the biggest problems faced in this work is the little overlap between biological activity and ligand-protein couples' structural information, which considerably reduces the algorithm's training test. A way to reduce the negative impact on the algorithm performance could be increasing the overlap of biological and structural data by generating models of the ligand-protein binding mode with homology modeling and molecular docking. Moreover, the degree of similarity among many of the dataset compounds might also affect the prediction performance by over-representing chemical structures that might be very similar. It might be beneficial to assess the problem above to cluster highly similar compounds into one entity. Another important point of discussion can be raised on the biological link between the IC50 values and the PLIP fingerprints describing the ligand-protein interaction profile. While the 3D information about the interaction environment describes the mere biochemical phenomenon of binding (binding affinity), the IC50 expresses an alteration in the target's biological activity, whatever that might be. Once a small molecule binds a protein in its binding site, the target undergoes a conformational change that can either increase or decrease its activity. In the first case, the binding molecule

is defined as an agonist, while in the second case, it gets the name of antagonist. The nature of the ligand-protein binding and the direction the target activity goes after the interaction is not described by the structural information, while the IC50 value expresses it. Also, the interaction pattern-IC50 correlation does not reflect ligand-protein binding strength, disregarding the interaction's avidity. Because of this scenario's complexity, it would help develop algorithms that can take into account more types of experimental measurements other than the IC50, such as the Kd, which better correlates with the biochemical affinity. Kd and IC50 together might generate a rank based on biochemical affinity and biological activity that could better describe the phenomenon. Another characteristic of this complex phenomenon is its dynamic behavior. While the 3D information reflects only one specific conformational status of the target, in reality, the proteins exist in several shapes that usually reflect their active/inactive status. Therefore, structure-based fingerprints only provide a static snapshot of the real situation. In an extreme case, the same target might bind different ligands in different conformations, resulting in a different structural fingerprint. In terms of structural side information, each protein conformation should be considered a different target to create the best conditions for the algorithm to learn even the smallest difference and predict the right ligand-protein association. However, this approach would create problems in merging other kinds of information unrelated to the protein structure, such as the protein sequence or differential gene expression data. Also, in this case, clustering similar conformations might be an option to reduce the dimension of the problem. In conclusion, the 3D structural data's detailed information seems to encounter several problems when coupled with different data types to feed a machine learning model. Future works are needed to exploit all the specific issues to prepare the field of artificial intelligence ready to welcome the future explosion of structure-based data in pharmacology.

Appendix A

Data fusion algorithm code

```
1 import chembl_data
2 import sparselinear
3 import utils
4 import time
5 import torch
6 import numpy as np
7 import pandas as pd
8 import scipy.io
9 import logger
10 from torch import nn
11 from tqdm import tqdm

12 import logging
13 logging.basicConfig(level=logging.DEBUG,
format='%(asctime)s
-%(levelname)s-%(message)s')

14 cuda = False

15 def split(y, ratio, min_num_pos, min_num_neg):
16     y = np.array(y.todense())
17     numpos=(y==1.0).sum(0)
18     numneg=(y==-1.0).sum(0)
19     good_prot = ((numneg.reshape((-1, 4))>= min_num_neg)
&(numpos.reshape((-1, 4))>= min_num_pos)).sum(1)>0
20     num_good_prot = good_prot.sum()
```

```
21      validation_set_size = int(np.around
(num_good_prot * ratio))
22      pos_good_prot = np.where(good_prot)
23      valid_prot = np.zeros(len(good_prot), dtype=np.bool)
24      valid_prot[pos_good_prot[0][list(np.random.permutation
(num_good_prot)[:validation_set_size])]] = True
25      return (valid_prot)

26 class SparseBilinearNet(nn.Module):
27      def __init__(
28              self,
29              input_size_c,
30              input_prot,
31              hidden_size_c,
32              hidden_size_p,
33              hidden_size_bi,
34              dropout1,
35              dropout_i,
36              dropout_bi,
37              nlevels,
38              input_size_i=0,
39              hidden_size_i=0):


40          super(SparseBilinearNet, self).__init__()

        ## storying protein matrix with the module
41          self.input_prot = input_prot

42          self.fc = nn.Sequential(
43              sparselinear.SparseLinear
(input_size_c, hidden_size_c),
44              nn.ReLU(),
45              nn.Dropout(dropout1),
46          )
47          self.fp = nn.Sequential(
48              nn.Linear(input_prot.shape[1], hidden_size_p),
49              nn.ReLU(),
50              nn.Dropout(dropout1),
51          )
52          if input_size_i > 0:
53              self.fi = nn.Linear(input_size_i,
```

```
hidden_size_i)

54          else:
55              self.fi=None

56          self.bi = nn.Bilinear(
57              in1_features = hidden_size_c,
58              in2_features = hidden_size_p,
59              out_features = hidden_size_bi,
60          )
61          if self.fi is None:
62              out_in_size = hidden_size_bi
63          else:
64              out_in_size = hidden_size_bi+hidden_size_i

65          self.out = nn.Sequential(
66              nn.ReLU(),
67              nn.Dropout(dropout_bi),
68              nn.Linear(out_in_size, nlevels),
69          )

70      def forward(self, x, y_cmpds, y_prot, plip=None):
71          assert (plip is None) == (self.fi is None),
"Inconsistent use of interaction feature"
71          h_cmpds = self.fc(x)
72          h_prots = self.fp(self.input_prot)

73          batch_cmpds = h_cmpds[y_cmpds]
74          batch_prots = h_prots[y_prots]

75          pairs = self.bi(batch_cmpds, batch_prots)

76          if plip is None:
77              out = self.out(pairs)

78          else:
79              h_plip=self.fi(plip)
80              out = self.out(torch.cat([pairs, h_plip],
dim=-1))

          ## adding channel of 0s for CrossEntropyLoss
81          return torch.stack([torch.zeros_like(out), out],
```

```
dim=1)

## creating a data loader
82 logging.info("Loading␣ECFP␣and␣pIC50␣data.")
83 proteins = pd.read_csv("./data/plip_data/
chembl_23_targets_map_plip.tsv", sep="\t")
84 prot_df = pd.read_csv("./data/plip_data/
matrix_domains.csv", index_col=0)
85 assert (proteins.uniprot == prot_df.index).all()

86 prot_features = True

87 if prot_features:
88     prot = prot_df.values
89 else:
90     prot = np.ones((len(prot_df), 1))

91 plip =scipy.io.mmread("./data/plip_data/plip_sparse.mtx")
.tocsr()

92 dataset=chembl_data.ProteinChEMBLDataset(
93     x=scipy.io.mmread("./data/plip_data/ecfp_sparse.mtx")
.tocsr(),
94     y=scipy.io.mmread("./data/plip_data/ic50_sparse.mtx")
.tocsr(),
95     plip = plip,
96     nlevels = 4,
97     prot = prot
98 )


###CHECKING HOW MANY BITS IN PLIP

99 assert plip.shape[1] % len(prot_df) == 0
100 plip_dim = plip.shape[1] // len(prot_df)

101 logging.info("Data␣loaded.")
102 logging.info(f"Num␣compounds:␣{dataset.x.shape[0]}")
103 logging.info(f"Num␣targets:␣␣␣{dataset.prot.shape[0]}")
104 logging.info(f"Compound␣feat:␣{dataset.x.shape[1]}")
105 logging.info(f"Target␣feat:␣␣␣{dataset.prot.shape[1]}")
```

```
106 logging.info("Setting up pytorch model.")
107 input_size      = dataset.x.shape[1]
108 num_tasks       = dataset.y.shape[1]
109 hidden_size_c   = 900
110 hidden_size_p   = 700
111 hidden_size_bi  = 700
112 learning_rate   = 5e-3
113 num_epochs      = 50
114 dropout1        = 0.8
115 dropout_i       = 0
116 dropout_bi      = 0.5
117 weight_decay    = 0.1
118 min_num_pos     = 10
119 min_num_neg     = 10
120 use_plip        = True
121 hidden_size_plip = 700

122 if not use_plip:
123     plip_dim = 0

124 np.random.seed(12345)

125 valid_prot = split(dataset.y, 0.3, min_num_pos,
min_num_neg)

126 train_loader, valid_loader =
chembl_data.get_chembl_proteochem_loaders(
127     dataset     = dataset,
128     batch_size  = 100,
129     valid_prot  = valid_prot,
130     num_workers = 4,
131     pin_memory  = True,
132 )
133 logging.info("Loaders created.")


134 device = "gpu:0"

135 input_prot = torch.from_numpy(dataset.prot.astype
```

```
(np.float32))
136 if cuda:
137     input_prot = input_prot.to(device)

138 model  = SparseBilinearNet(
139     input_size_c    = input_size,
140     input_prot      = input_prot,
141     hidden_size_c   = hidden_size_c,
142     hidden_size_p   = hidden_size_p,
143     hidden_size_bi  = hidden_size_bi,
144     dropout1        = dropout1,
145     dropout_i           = dropout_i,
146     dropout_bi      = dropout_bi,
147     nlevels         = dataset.nlevels,
148     input_size_i    = plip_dim,
149     hidden_size_i   = hidden_size_plip
150 )
151 if cuda:
152     model = model.to(device)

## initializing weights:
153 def init_weights(m):
154     if type(m) == torch.nn.Linear and m.bias is not None:
155         m.bias.data.fill_(0.1)
156     if type(m) == sparselinear.SparseLinear and
m.bias is not None:
157         m.bias.data.fill_(0.1)

158 model.apply(init_weights)

159 col_counts = pd.DataFrame({
160     'task':     np.arange(num_tasks),
161     'protein': np.arange(num_tasks) // dataset.nlevels,
162     'level':    np.arange(num_tasks)  % dataset.nlevels,
163     'num_pos': np.array((dataset.y == +1.0)
.sum(0)).flatten(),
164     'num_neg': np.array((dataset.y == -1.0)
.sum(0)).flatten(),
})
165 col_counts = col_counts.assign(
166     valid     = valid_prot[col_counts.protein],
167    has_data = (col_counts.num_pos >= min_num_pos) &
```

```
(col_counts.num_neg >= min_num_neg),)
168 valid_cols    = col_counts.query("has_data and valid")
.task.values
169 logging.info(f"Validation AUC is calculated as mean of
{valid_cols.shape[0]} tasks.")

170 criterion = torch.nn.CrossEntropyLoss(ignore_index=-1,
reduction="sum")
171 optimizer = torch.optim.Adam(model.parameters(),
lr=learning_rate,weight_decay=weight_decay)
172 scheduler = torch.optim.lr_scheduler.MultiStepLR
(optimizer, milestones=[20, 23], gamma=0.3)

173 logging.info("Model setup complete.")

174 def select_plip(y_cmpds, y_prots, plip_dim):
175     if plip_dim == 0:
176         return None
177     real_size = y_cmpds.shape[0]
178     plip_x = np.repeat(y_cmpds, plip_dim)
179     plip_y = np.repeat(y_prots.numpy()
*plip_dim, plip_dim)
+ np.tile(np.arange(plip_dim), real_size)
180     return torch.Tensor(plip[plip_x, plip_y].
todense().reshape((real_size,-1)))


181 for epoch in range(num_epochs):
182     scheduler.step()
183     logloss_train = torch.FloatTensor([0.0])
184     nnz_train     = torch.LongTensor([0])
185     if cuda:
186         logloss_train = logloss_train.to(device)
187         nnz_train     = nnz_train.to(device)

    ## train loop
188     model.train()

189     for i, batch in tqdm(enumerate(train_loader)):

190         if cuda:
191             xi, xv, y_cmpds, y_prots, y_acts, y_plip=
```

```
[t.to(device) for t in batch[0:-1]]
192           else:
193               xi, xv, y_cmpds, y_prots, y_acts, y_plip=
batch[0:-1]
194           num_cmpds = batch[-1]

195           plip_batch = y_plip

196           if cuda:
197               x = torch.cuda.sparse.FloatTensor
(xi, xv, torch.Size([num_cmpds, input_size]))
198           else:
199               x = torch.sparse.FloatTensor
(xi, xv, torch.Size([num_cmpds, input_size]))

200           outputs  = model(x,y_cmpds,y_prots,
plip=plip_batch)


201           loss = criterion(input=outputs, target=y_acts)
202           loss_avg = loss / num_cmpds

          ## averaging all losses
203           logloss_train += loss.detach()
204           nnz_train     += (y_acts >= 0).sum().detach()

          ## Backward and optimize
205           optimizer.zero_grad()
206           loss_avg.backward()
207           optimizer.step()


      ## validation loop
208       logloss_valid = torch.FloatTensor([0.0])
209       nnz_valid     = torch.LongTensor([0])
210       if cuda:
211           logloss_valid = logloss_valid.to(device)
212           nnz_valid     = nnz_valid.to(device)

213       ypred_list = []
214       ycol_list  = []
215       ytrue_list = []
```

```
216     model.eval()
217     with torch.no_grad():
218         for i, batch in enumerate(valid_loader):
219             if cuda:
220                 xi, xv, y_cmpds, y_prots, y_acts,
y_plip =[t.to(device) for t in batch[0:-1]]
221             else:
222                 xi, xv, y_cmpds, y_prots, y_acts,
y_plip =[t for t in batch[0:-1]]

223             num_cmpds = batch[-1]
224             plip_batch = y_plip

225             if cuda:
226                 x = torch.cuda.sparse.FloatTensor
(xi, xv, torch.Size([num_cmpds, input_size]))
227             else:
228                 x = torch.sparse.FloatTensor
(xi, xv, torch.Size([num_cmpds, input_size]))
229             outputs = model(x, y_cmpds, y_prots,
plip=plip_batch)

230             loss = criterion(input=outputs, target=y_acts)
231             loss_avg = loss / num_cmpds

            ## averaging all losses
232             logloss_valid += loss.detach()
233             nnz_valid     += (y_acts >= 0).sum().detach()

            ## task id is (prot_id * nlevels + level)
234             prot_id  = np.repeat(batch[3].numpy(),
repeats=dataset.nlevels)
235             level_id = np.tile(np.arange
(dataset.nlevels), reps=batch[3].shape[0])

236             valid_col  = prot_id * dataset.nlevels+
level_id
237             valid_pred = outputs[:,1,:].detach()
.flatten()
238             valid_true = batch[4].numpy().flatten()
```

```
239                 ypred_list.append( valid_pred )
240                 ycol_list.append(  valid_col )
241                 ytrue_list.append( valid_true )


    ## computing AUCs
242         ypred = torch.cat(ypred_list).cpu().numpy()
243         ycol  = np.concatenate(ycol_list)
244         ytrue = np.concatenate(ytrue_list)
245         mask  = (ytrue >= 0)
246         aucs  = utils.comp_aucs_pd(cols=ycol[mask],
ypred=ypred[mask], ytrue=ytrue[mask]).loc[valid_cols]

247         logloss_tr = logloss_train.cpu().numpy()[0] /
nnz_train.cpu().numpy()[0]
248         logloss_va = logloss_valid.cpu().numpy()[0] /
nnz_valid.cpu().numpy()[0]
249         print(f"{epoch}. logloss_tr = {logloss_tr:.4f},
logloss_va = {logloss_va:.4f}, aucs_va = {aucs.mean():.4f}")
```

Bibliography

[1] Concavity, https://compbio.cs.princeton.edu/concavity/.

[2] Drug repurposing portal, http://www.drugrepurposingportal.com/repurposed-drug-database.php.

[3] inteligand, http://www.inteligand.com/pharmdb/.

[4] Pfamscan, https://www.ebi.ac.uk/tools/pfa/pfamscan/.

[5] Plip, atprojects.biotec.tu-dresden.de/plip-web.

[6] Sparseflow directory, https://gitlab.com/biolearning/sparseflow.

[7] Elsevier r&d solutions: A comparison of emtree and mesh, 2015.

[8] United states national library of medicine: Fact sheet medical subject headings (mesh), 2015.

[9] Alexandra grünbauer (clinfo): Databases for systematic literature searches, 2016.

[10] Elsevier: Emtree fact sheet, 2019.

[11] World health organization: Icd-10, 2019.

[12] Adasme, M. F., Parisi, D., Van Belle, K., Salentin, S., Haupt, V. J., Jennings, G. S., Heinrich, J.-C., Herman, J., Sprangers, B., Louat, T., Moreau, Y., and Schroeder, M. Structure-based drug repositioning explains ibrutinib as vegfr2 inhibitor. *PLOS ONE 15*, 5 (05 2020), 1–26.

[13] Ahmed, A., Smith, R. D., Clark, J. J., Dunbar, James B., J., and Carlson, H. A. Recent improvements to Binding MOAD: a resource for protein–ligand binding affinities and structures. *Nucleic Acids Research 43*, D1 (11 2014), D465–D469.

[14] Alaimo, S., Bonnici, V., Cancemi, D., Ferro, A., Giugno, R., and Pulvirenti, A. Dt-web: a web-based application for drug-target interaction and drug combination prediction through domain-tuned network-based inference. *BMC Systems Biology 9*, 3 (2015), S4.

[15] Alaimo, S., Pulvirenti, A., Giugno, R., and Ferro, A. Drug–target interaction prediction through domain-tuned network-based inference. *Bioinformatics 29*, 16 (05 2013), 2004–2008.

[16] Aloy, P., and Russell, R. B. Ten thousand interactions for the molecular biologist. *Nature Biotechnology 22*, 10 (2004), 1317–1321.

[17] Antolin, A. A., Workman, P., Mestres, J., and Al-Lazikani, B. Polypharmacology in precision oncology: Current applications and future prospects. *Current pharmaceutical design 22*, 46 (2016), 6935–6945. 27669965[pmid].

[18] Ardeshirdavani, A., Moreau, Y., ElShal, S., Tranchevent, L.-C., Sifrim, A., and Davis, J. Beegle: from literature mining to disease-gene discovery. *Nucleic Acids Research 44*, 2 (09 2015), e18–e18.

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

[20] Astolfi, A., Felicetti, T., Iraci, N., Manfroni, G., Massari, S., Pietrella, D., Tabarrini, O., Kaatz, G. W., Barreca, M. L., Sabatini, S., and Cecchetti, V. Pharmacophore-based repositioning of approved drugs as novel staphylococcus aureus nora efflux pump inhibitors. *Journal of Medicinal Chemistry 60*, 4 (2017), 1598–1604. PMID: 28117588.

[21] Ausiello, G., Via, A., and Helmer-Citterich, M. Query3d: a new method for high-throughput analysis of functional residues in protein structures. *BMC bioinformatics 6 Suppl 4*, Suppl 4 (Dec 2005), S5–S5. 16351754[pmid].

[22] Baroni, M., Cruciani, G., Sciabola, S., Perruccio, F., and Mason, J. S. A common reference framework for analyzing/comparing proteins and ligands. fingerprints for ligands and proteins (flap): theory and application. *Journal of Chemical Information and Modeling 47*, 2 (2007), 279–294. PMID: 17381166.

[23] Batista, J., Tan, L., and Bajorath, J. Atom-centered interacting fragments and similarity search applications. *Journal of Chemical Information and Modeling 50*, 1 (2010), 79–86. PMID: 20030298.

[24] Behler, J., and Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett. 98* (Apr 2007), 146401.

[25] Belkin, Mikhail, Partha, Sindhwani, and Vikas. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples, Jan 1970.

[26] Bento, A. P., Gaulton, A., Hersey, A., Bellis, L. J., Chambers, J., Davies, M., Krüger, F. A., Light, Y., Mak, L., McGlinchey, S., Nowotka, M., Papadatos, G., Santos, R., and Overington, J. P. The ChEMBL bioactivity database: an update. *Nucleic Acids Research 42*, D1 (11 2013), D1083–D1090.

[27] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. The Protein Data Bank. *Nucleic Acids Research 28*, 1 (01 2000), 235–242.

[28] Blay, J. Y., Shen, L., Kang, Y. K., Rutkowski, P., Qin, S., Nosov, D., Wan, D., Trent, J., Srimuninnimit, V., Papai, Z., Le Cesne, A., Novick, S., Taningco, L., Mo, S., Green, S., Reichardt, P., and Demetri, G. D. Nilotinib versus imatinib as first-line therapy for patients with unresectable or metastatic gastrointestinal stromal tumours (ENESTg1): a randomised phase 3 trial. *Lancet Oncol. 16*, 5 (May 2015), 550–560.

[29] Bleakley, K., and Yamanishi, Y. Supervised prediction of drug–target interactions using bipartite local models. *Bioinformatics 25*, 18 (07 2009), 2397–2403.

[30] Brakoulias, A., and Jackson, R. M. Towards a structural classification of phosphate binding sites in protein–nucleotide complexes: An automated all-against-all structural comparison using geometric matching. *Proteins: Structure, Function, and Bioinformatics 56*, 2 (2004), 250–260.

[31] Brown, A. S., and Patel, C. J. A standard database for drug repositioning. *Scientific Data 4*, 1 (2017), 170029.

[32] Brylinski, M., and Skolnick, J. A threading-based method (findsite) for ligand-binding site prediction and functional annotation. *Proceedings of the National Academy of Sciences of the United States of America 105*, 1 (Jan 2008), 129–134. 18165317[pmid].

[33] Buza, K., and Peška, L. Drug–target interaction prediction with bipartite local models and hubness-aware regression. *Neurocomputing 260* (2017), 284 – 293.

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

[35] Cao, D.-S., Xiao, N., Xu, Q.-S., and Chen, A. F. Rcpi: R/Bioconductor package to generate various descriptors of proteins, compounds and their interactions. *Bioinformatics 31*, 2 (09 2014), 279–281.

[36] Cao, D.-S., Xu, Q.-S., and Liang, Y.-Z. propy: a tool to generate various modes of Chou's PseAAC. *Bioinformatics 29*, 7 (02 2013), 960–962.

[37] Carhart, R. E., Smith, D. H., and Venkataraghavan, R. Atom pairs as molecular features in structure-activity studies: definition and applications. *Journal of Chemical Information and Computer Sciences 25*, 2 (1985), 64–73.

[38] Chen, B., Harrison, R. F., Papadatos, G., Willett, P., Wood, D. J., Lewell, X. Q., Greenidge, P., and Stiefl, N. Evaluation of machine-learning methods for ligand-based virtual screening. *Journal of Computer-Aided Molecular Design 21*, 1 (Jan 2007), 53–62.

[39] Chen, X., Liu, M.-X., and Yan, G.-Y. Drug–target interaction prediction by random walk on the heterogeneous network. *Mol. BioSyst. 8* (2012), 1970–1978.

[40] Cheng, F., Li, W., Wang, X., Zhou, Y., Wu, Z., Shen, J., and Tang, Y. Adverse drug events: Database construction and in silico prediction. *Journal of Chemical Information and Modeling 53*, 4 (2013), 744–752. PMID: 23521697.

[41] Cheng, F., Liu, C., Jiang, J., Lu, W., Li, W., Liu, G., Zhou, W., Huang, J., and Tang, Y. Prediction of drug-target interactions and drug repositioning via network-based inference. *PLOS Computational Biology 8*, 5 (05 2012), 1–12.

[42] Chuaqui, C., Deng, Z., and Singh, J. Interaction profiles of protein kinaseinhibitor complexes and their application to virtual screening. *Journal of Medicinal Chemistry 48*, 1 (2005), 121–133. PMID: 15634006.

[43] Coakley, J. H., Moorcraft, J., Hipkin, L. J., Smith, C. S., Griffiths, R. D., and Edwards, R. H. The effect of mazindol on growth hormone secretion in boys with duchenne muscular dystrophy. *Journal of Neurology, Neurosurgery & Psychiatry 51*, 12 (1988), 1551–1557.

[44] Cobanoglu, M. C., Liu, C., Hu, F., Oltvai, Z. N., and Bahar, I. Predicting drug–target interactions using probabilistic matrix factorization. *Journal of Chemical Information and Modeling 53*, 12 (2013), 3399–3409. PMID: 24289468.

[45] Crisman, T. J., Sisay, M. T., and Bajorath, J. Ligand-target interaction-based weighting of substructures for virtual screening. *Journal of Chemical Information and Modeling 48*, 10 (2008), 1955–1964. PMID: 18821751.

[46] De Franchi, E., Schalon, C., Messa, M., Onofri, F., Benfenati, F., and Rognan, D. Binding of protein kinase inhibitors to synapsin i inferred from pair-wise binding site similarity measurements. *PLOS ONE 5*, 8 (08 2010), 1–11.

[47] Deng, Z., Chuaqui, C., and Singh, J. Structural interaction fingerprint (sift): a novel method for analyzing three-dimensional proteinligand binding interactions. *Journal of Medicinal Chemistry 47*, 2 (2004), 337–344. PMID: 14711306.

[48] Deng, Z., Chuaqui, C., and Singh, J. Knowledge-based design of target-focused libraries using proteinligand interaction constraints. *Journal of Medicinal Chemistry 49*, 2 (2006), 490–500. PMID: 16420036.

[49] Desaphy, J., Bret, G., Rognan, D., and Kellenberger, E. sc-PDB: a 3D-database of ligandable binding sites—10 years on. *Nucleic Acids Research 43*, D1 (10 2014), D399–D404.

[50] Desaphy, J., Raimbaud, E., Ducrot, P., and Rognan, D. Encoding protein–ligand interaction patterns in fingerprints and graphs. *Journal of Chemical Information and Modeling 53*, 3 (2013), 623–637. PMID: 23432543.

[51] Doncheva, N. T., Klein, K., Domingues, F. S., and Albrecht, M. Analyzing and visualizing residue networks of protein structures. *Trends in Biochemical Sciences 36*, 4 (Apr 2011), 179–182.

[52] Dong, J., Cao, D.-S., Miao, H.-Y., Liu, S., Deng, B.-C., Yun, Y.-H., Wang, N.-N., Lu, A.-P., Zeng, W.-B., and Chen, A. F. Chemdes: an integrated web-based platform for molecular descriptor and fingerprint computation. *Journal of Cheminformatics 7*, 1 (2015), 60.

[53] Dudley, J. T., Sirota, M., Shenoy, M., Pai, R. K., Roedder, S., Chiang, A. P., Morgan, A. A., Sarwal, M. M., Pasricha, P. J., and Butte, A. J. Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease. *Science Translational Medicine 3*, 96 (2011), 96ra76–96ra76.

[54] Dunkel, M., Günther, S., Ahmed, J., Wittig, B., and Preissner, R. SuperPred: drug classification and target prediction. *Nucleic Acids Research 36*, suppl$_2$(052008), $W55 - -W59$.

[55] Ekins, S., Puhl, A. C., Zorn, K. M., Lane, T. R., Russo, D. P., Klein, J. J., Hickey, A. J., and Clark, A. M. Exploiting machine learning for end-to-end drug discovery and development. *Nature Materials 18*, 5 (2019), 435–441.

[56] Gao, K. Y., Fokoue, A., Luo, H., Iyengar, A., Dey, S., and Zhang, P. Interpretable drug target prediction using deep neural representation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence* (2018), IJCAI'18, AAAI Press, pp. 3371–3377.

[57] Gao, M., and Skolnick, J. Structural space of protein–protein interfaces is degenerate, close to complete, and highly connected. *Proceedings of the National Academy of Sciences 107*, 52 (2010), 22517–22522.

[58] Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L. J., Cibrián-Uhalte, E., Davies, M., Dedman, N., Karlsson, A., Magariños, M. P., Overington, J. P., Papadatos, G., Smit, I., and Leach, A. R. The chembl database in 2017. *Nucleic acids research 45*, D1 (Jan 2017), D945–D954. 27899562[pmid].

[59] Gedeck, P., Rohde, B., and Bartels, C. Qsar how good is it in practice? comparison of descriptor sets on an unbiased cross section of corporate data sets. *Journal of Chemical Information and Modeling 46*, 5 (2006), 1924–1936. PMID: 16995723.

[60] Gfeller, D., Grosdidier, A., Wirth, M., Daina, A., Michielin, O., and Zoete, V. SwissTargetPrediction: a web server for target prediction of bioactive small molecules. *Nucleic Acids Research 42*, W1 (05 2014), W32–W38.

[61] Gilson, M. K., Liu, T., Baitaluk, M., Nicola, G., Hwang, L., and Chong, J. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Research 44*, D1 (10 2015), D1045–D1053.

[62] Glick, M., Klon, A. E., Acklin, P., and Davies, J. W. Enrichment of extremely noisy high-throughput screening data using a naïve bayes classifier. *Journal of Biomolecular Screening 9*, 1 (2004), 32–36. PMID: 15006146.

[63] Grondin, C. J., Davis, A. P., Wiegers, T. C., King, B. L., Wiegers, J. A., Reif, D. M., Hoppin, J. A., and Mattingly, C. J. Advancing exposure science through chemical data curation and integration in the

comparative toxicogenomics database. *Environmental Health Perspectives 124*, 10 (2016), 1592–1599.

[64] Guedes, I. A., de Magalhães, C. S., and Dardenne, L. E. Receptor–ligand molecular docking. *Biophysical Reviews 6*, 1 (Mar 2014), 75–87.

[65] Guimarães, C. R. W., and Cardozo, M. Mm-gb/sa rescoring of docking poses in structure-based lead optimization. *Journal of Chemical Information and Modeling 48*, 5 (2008), 958–970. PMID: 18422307.

[66] Gönen, M. Predicting drug–target interactions from chemical and genomic kernels using Bayesian matrix factorization. *Bioinformatics 28*, 18 (06 2012), 2304–2310.

[67] Günther, S., Kuhn, M., Dunkel, M., Campillos, M., Senger, C., Petsalaki, E., Ahmed, J., Urdiales, E. G., Gewiess, A., Jensen, L. J., Schneider, R., Skoblo, R., Russell, R. B., Bourne, P. E., Bork, P., and Preissner, R. SuperTarget and Matador: resources for exploring drug-target relationships. *Nucleic Acids Research 36*, $\text{suppl}_1(102007), D919--D922$.

[68] Halgren, T. New method for fast and accurate binding-site identification and analysis. *Chemical Biology & Drug Design 69*, 2 (2007), 146–148.

[69] Hao, M., Bryant, S. H., and Wang, Y. Predicting drug-target interactions by dual-network integrated logistic matrix factorization. *Scientific Reports 7* (Jan 2017), 40376 EP –. Article.

[70] Harris, R., Olson, A. J., and Goodsell, D. S. Automated prediction of ligand-binding sites in proteins. *Proteins: Structure, Function, and Bioinformatics 70*, 4 (2008), 1506–1517.

[71] Haupt, V. J., Daminelli, S., and Schroeder, M. Drug promiscuity in pdb: Protein binding site similarity is key. *PLOS ONE 8*, 6 (06 2013), 1–15.

[72] Haupt, V. J., and Schroeder, M. Old friends in new guise: repositioning of known drugs with structural bioinformatics. *Briefings in Bioinformatics 12*, 4 (03 2011), 312–326.

[73] Hecker, N., Ahmed, J., von Eichborn, J., Dunkel, M., Macha, K., Eckert, A., Gilson, M. K., Bourne, P. E., and Preissner, R. Supertarget goes quantitative: update on drug-target interactions. *Nucleic acids research 40*, Database issue (Jan 2012), D1113–D1117. 22067455[pmid].

[74] Heidenreich, R., Röcken, M., and Ghoreschi, K. Angiogenesis drives psoriasis pathogenesis. *International Journal of Experimental Pathology 90*, 3 (2009), 232–248.

[75] Hert, J., Keiser, M. J., Irwin, J. J., Oprea, T. I., and Shoichet, B. K. Quantifying the relationships among drug classes. *Journal of Chemical Information and Modeling 48*, 4 (2008), 755–765. PMID: 18335977.

[76] Hert, J., Willett, P., Wilton, D. J., Acklin, P., Azzaoui, K., Jacoby, E., and Schuffenhauer, A. Enhancing the effectiveness of similarity-based virtual screening using nearest-neighbor information. *Journal of Medicinal Chemistry 48*, 22 (2005), 7049–7054. PMID: 16250664.

[77] Hert, J., Willett, P., Wilton, D. J., Acklin, P., Azzaoui, K., Jacoby, E., and Schuffenhauer, A. New methods for ligand-based virtual screening: use of data fusion and machine learning to enhance the effectiveness of similarity searching. *Journal of Chemical Information and Modeling 46*, 2 (2006), 462–470. PMID: 16562973.

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

[79] Hopkins, A. L. Predicting promiscuity. *Nature 462*, 7270 (Nov 2009), 167–168.

[80] Huang, Q., Jin, H., Liu, Q., Wu, Q., Kang, H., Cao, Z., and Zhu, R. Proteochemometric modeling of the bioactivity spectra of hiv-1 protease inhibitors by introducing protein-ligand interaction fingerprint. *PLOS ONE 7*, 7 (07 2012), 1–8.

[81] Huang, R., Southall, N., Wang, Y., Yasgar, A., Shinn, P., Jadhav, A., Nguyen, D.-T., and Austin, C. P. The ncgc pharmaceutical collection: a comprehensive resource of clinically approved drugs enabling repurposing and chemical genomics. *Science translational medicine 3*, 80 (Apr 2011), 80ps16–80ps16. 21525397[pmid].

[82] Huang, X., and Guo, B. Adenomatous polyposis coli determines sensitivity to histone deacetylase inhibitor–induced apoptosis in colon cancer cells. *Cancer Research 66*, 18 (2006), 9245–9251.

[83] Jambon, M., Andrieu, O., Combet, C., Deléage, G., Delfaud, F., and Geourjon, C. The SuMo server: 3D search for protein functional sites. *Bioinformatics 21*, 20 (09 2005), 3929–3930.

[84] Jegga, A. G., Zhu, C., Sardana, D., Yang, L., Zhang, M., and Gudivada, R. C. Drug repositioning for orphan diseases. *Briefings in Bioinformatics 12*, 4 (04 2011), 346–356.

[85] Jensen, B. F., Vind, C., , Brockhoff, P. B., and Refsgaard, H. H. F. In silico prediction of cytochrome p450 2d6 and 3a4 inhibition using gaussian kernel weighted k-nearest neighbor and extended connectivity fingerprints, including

structural fragment analysis of inhibitors versus noninhibitors. *Journal of Medicinal Chemistry 50*, 3 (2007), 501–511. PMID: 17266202.

[86] Jourdan, J.-P., Bureau, R., Rochais, C., and Dallemagne, P. Drug repositioning: a brief overview. *Journal of Pharmacy and Pharmacology 72*, 9 (2020), 1145–1151.

[87] Kalidas, Y., and Chandra, N. Pocketdepth: A new depth based algorithm for identification of ligand binding sites in proteins. *Journal of Structural Biology 161*, 1 (2008), 31 – 42.

[88] Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research 45*, D1 (11 2016), D353–D361.

[89] Kantarjian, H., Giles, F., Wunderle, L., Bhalla, K., O'Brien, S., Wassmann, B., Tanaka, C., Manley, P., Rae, P., Mietlowski, W., Bochinski, K., Hochhaus, A., Griffin, J. D., Hoelzer, D., Albitar, M., Dugan, M., Cortes, J., Alland, L., and Ottmann, O. G. Nilotinib in imatinib-resistant cml and philadelphia chromosome–positive all. *New England Journal of Medicine 354*, 24 (2006), 2542–2551.

[90] Karuppagounder, S. S., Brahmachari, S., Lee, Y., Dawson, V. L., Dawson, T. M., and Ko, H. S. The c-Abl inhibitor, nilotinib, protects dopaminergic neurons in a preclinical animal model of Parkinson's disease. *Sci Rep 4* (May 2014), 4874.

[91] Keenan, A. B., Jenkins, S. L., Jagodnik, K. M., Koplev, S., He, E., Torre, D., Wang, Z., Dohlman, A. B., Silverstein, M. C., Lachmann, A., Kuleshov, M. V., Ma'ayan, A., Stathias, V., Terryn, R., Cooper, D., Forlin, M., Koleti, A., Vidovic, D., Chung, C., Schürer, S. C., Vasiliauskas, J., Pilarczyk, M., Shamsaei, B., Fazel, M., Ren, Y., Niu, W., Clark, N. A., White, S., Mahi, N., Zhang, L., Kouril, M., Reichard, J. F., Sivaganesan, S., Medvedovic, M., Meller, J., Koch, R. J., Birtwistle, M. R., Iyengar, R., Sobie, E. A., Azeloglu, E. U., Kaye, J., Osterloh, J., Haston, K., Kalra, J., Finkbiener, S., Li, J., Milani, P., Adam, M., Escalante-Chong, R., Sachs, K., Lenail, A., Ramamoorthy, D., Fraenkel, E., Daigle, G., Hussain, U., Coye, A., Rothstein, J., Sareen, D., Ornelas, L., Banuelos, M., Mandefro, B., Ho, R., Svendsen, C. N., Lim, R. G., Stocksdale, J., Casale, M. S., Thompson, T. G., Wu, J., Thompson, L. M., Dardov, V., Venkatraman, V., Matlock, A., Van Eyk, J. E., Jaffe, J. D., Papanastasiou, M., Subramanian, A., Golub, T. R., Erickson, S. D., Fallahi-Sichani, M., Hafner, M., Gray, N. S., Lin, J.-R., Mills, C. E., Muhlich, J. L., Niepel, M., Shamu, C. E., Williams, E. H., Wrobel, D., Sorger, P. K.,

Heiser, L. M., Gray, J. W., Korkola, J. E., Mills, G. B., LaBarge, M., Feiler, H. S., Dane, M. A., Bucher, E., Nederlof, M., Sudar, D., Gross, S., Kilburn, D. F., Smith, R., Devlin, K., Margolis, R., Derr, L., Lee, A., and Pillai, A. The library of integrated network-based cellular signatures nih program: System-level cataloging of human cells response to perturbations. *Cell systems 6*, 1 (Jan 2018), 13–24. 29199020[pmid].

[92] Keiser, M. J., and Hert, J. *Off-Target Networks Derived from Ligand Set Similarity.* Humana Press, Totowa, NJ, 2009, pp. 195–205.

[93] Keiser, Michael J, R. B. L. A. B. N. E. P. I. J. J. S. B. K.

[94] Kellenberger, E., Muller, P., Schalon, C., Bret, G., Foata, N., and Rognan, D. sc-pdb: an annotated database of druggable binding sites from the protein data bank. *Journal of Chemical Information and Modeling 46*, 2 (2006), 717–727. PMID: 16563002.

[95] Kelly, M. D., and Mancera, R. L. Expanded interaction fingerprint method for analyzing ligand binding modes in docking and structure-based drug design. *Journal of Chemical Information and Computer Sciences 44*, 6 (2004), 1942–1951. PMID: 15554663.

[96] Kinoshita, K., Furui, J., and Nakamura, H. Identification of protein functions from a molecular surface database, ef-site. *Journal of Structural and Functional Genomics 2*, 1 (Mar 2002), 9–22.

[97] Klambauer, G., Wischenbart, M., Mahr, M., Unterthiner, T., Mayr, A., and Hochreiter, S. Rchemcpp: a web service for structural analoging in ChEMBL, Drugbank and the Connectivity Map. *Bioinformatics 31*, 20 (06 2015), 3392–3394.

[98] Klon, A. E., Glick, M., Thoma, M., Acklin, P., and Davies, J. W. Finding more needles in the haystack: a simple and efficient method for improving high-throughput docking results. *Journal of Medicinal Chemistry 47*, 11 (2004), 2743–2749. PMID: 15139752.

[99] Koes, D. R., and Camacho, C. J. ZINCPharmer: pharmacophore search of the ZINC database. *Nucleic Acids Research 40*, W1 (05 2012), W409–W414.

[100] Krier, M., Bret, G., and Rognan, D. Assessing the scaffold diversity of screening libraries. *Journal of Chemical Information and Modeling 46*, 2 (2006), 512–524. PMID: 16562979.

[101] Kringelum, J., Kjaerulff, S. K., Brunak, S., Lund, O., Oprea, T. I., and Taboureau, O. ChemProt-3.0: a global chemical biology diseases mapping. *Database 2016* (02 2016). bav123.

[102] Kruger, F. A., Rostom, R., and Overington, J. P. Mapping small molecule binding data to structural domains. *BMC Bioinformatics 13*, 17 (December 2012).

[103] Kuang, Q., Li, Y., Wu, Y., Li, R., Dong, Y., Li, Y., Xiong, Q., Huang, Z., and Li, M. A kernel matrix dimension reduction method for predicting drug-target interaction. *Chemometrics and Intelligent Laboratory Systems 162*, C (2017), 104–110.

[104] Kuhn, M., Letunic, I., Jensen, L. J., and Bork, P. The SIDER database of drugs and side effects. *Nucleic Acids Research 44*, D1 (10 2015), D1075–D1079.

[105] Kuhn, M., Letunic, I., Jensen, L. J., and Bork, P. The SIDER database of drugs and side effects. *Nucleic Acids Res. 44*, D1 (Jan 2016), D1075–1079.

[106] Kuhn, M., Szklarczyk, D., Pletscher-Frankild, S., Blicher, T. H., von Mering, C., Jensen, L. J., and Bork, P. STITCH 4: integration of protein–chemical interactions with user data. *Nucleic Acids Research 42*, D1 (11 2013), D401–D407.

[107] Kumar, S., and Kumar, S. Chapter 6 - molecular docking: A structure-based approach for drug repurposing. In *In Silico Drug Design*, K. Roy, Ed. Academic Press, 2019, pp. 161 – 189.

[108] Kumar, S., Ojha, A. K., Ahmed, B., Kumar, A., Das, J., and Materny, A. Tunable (violet to green) emission by high-yield graphene quantum dots and exploiting its unique properties towards sun-light-driven photocatalysis and supercapacitor electrode materials. *Materials Today Communications 11* (2017), 76 – 86.

[109] Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., Lerner, J., Brunet, J.-P., Subramanian, A., Ross, K. N., Reich, M., Hieronymus, H., Wei, G., Armstrong, S. A., Haggarty, S. J., Clemons, P. A., Wei, R., Carr, S. A., Lander, E. S., and Golub, T. R. The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science 313*, 5795 (2006), 1929–1935.

[110] Landrum, G. Rdkit: Open-source cheminformatics software.

[111] Laskowski, R. A. Surfnet: A program for visualizing molecular surfaces, cavities, and intermolecular interactions. *Journal of Molecular Graphics 13*, 5 (1995), 323 – 330.

[112] Laurie, A. T. R., and Jackson, R. M. Q-SiteFinder: an energy-based method for the prediction of protein–ligand binding sites. *Bioinformatics 21*, 9 (02 2005), 1908–1916.

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

[114] Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A. C., Liu, Y., Maciejewski, A., Arndt, D., Wilson, M., Neveu, V., Tang, A., Gabriel, G., Ly, C., Adamjee, S., Dame, Z. T., Han, B., Zhou, Y., and Wishart, D. S. DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Research 42*, D1 (11 2013), D1091–D1097.

[115] Le Guilloux, V., Schmidtke, P., and Tuffery, P. Fpocket: An open source platform for ligand pocket detection. *BMC Bioinformatics 10*, 1 (2009), 168.

[116] Li, L., Koh, C. C., Reker, D., Brown, J. B., Wang, H., Lee, N. K., Liow, H.-h., Dai, H., Fan, H.-M., Chen, L., and Wei, D.-Q. Predicting protein-ligand interactions based on bow-pharmacological space and bayesian additive regression trees. *Scientific Reports 9*, 1 (2019), 7703.

[117] Li, Z.-C., Huang, M.-H., Zhong, W.-Q., Liu, Z.-Q., Xie, Y., Dai, Z., and Zou, X.-Y. Identification of drug–target interaction from interactome network with 'guilt-by-association' principle and topology features. *Bioinformatics 32*, 7 (11 2015), 1057–1064.

[118] Li, Z. R., Lin, H. H., Han, L. Y., Jiang, L., Chen, X., and Chen, Y. Z. PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Research 34*, $\text{suppl}_2(072006), W32 - -W37$.

[119] Liang, J., Edelsbrunner, H., and Woodward, C. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein science : a publication of the Protein Society 7*, 9 (Sep 1998), 1884–1897. 9761470[pmid].

[120] Liu, B., Liu, F., Wang, X., Chen, J., Fang, L., and Chou, K.-C. Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Research 43*, W1 (05 2015), W65–W71.

[121] Liu, X., Ouyang, S., Yu, B., Liu, Y., Huang, K., Gong, J., Zheng, S., Li, Z., Li, H., and Jiang, H. PharmMapper server: a web server for potential drug target identification using pharmacophore mapping approach. *Nucleic Acids Research 38*, $\text{suppl}_2(042010), W609 - -W614$.

[122] Liu, Y., Wu, M., Miao, C., Zhao, P., and Li, X.-L. Neighborhood regularized logistic matrix factorization for drug-target interaction prediction. *PLOS Computational Biology 12*, 2 (02 2016), 1–26.

[123] Liu, Z., Li, Y., Han, L., Li, J., Liu, J., Zhao, Z., Nie, W., Liu, Y., and Wang, R. PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics 31*, 3 (10 2014), 405–412.

[124] Lonskaya, I., Hebron, M. L., and Moussa, C. E.-H. Nilotinib reverses loss of dopamine neurons and improves motor behavior via autophagic degradation of -synuclein in Parkinson's disease models. *Human Molecular Genetics 22*, 16 (05 2013), 3315–3328.

[125] Lotfi Shahreza, M., Ghadiri, N., Mousavi, S. R., Varshosaz, J., and Green, J. R. A review of network-based approaches to drug repositioning. *Briefings in Bioinformatics 19*, 5 (02 2017), 878–892.

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

[127] López-Lázaro, M. Dual role of hydrogen peroxide in cancer: Possible relevance to cancer chemoprevention and therapy. *Cancer Letters 252*, 1 (2007), 1–8.

[128] Ma, D.-L., Chan, D. S.-H., and Leung, C.-H. Drug repositioning by structure-based virtual screening. *Chem. Soc. Rev. 42* (2013), 2130–2141.

[129] Madej, T., Lanczycki, C. J., Zhang, D., Thiessen, P. A., Geer, R. C., Marchler-Bauer, A., and Bryant, S. H. MMDB and VAST+: tracking structural similarities between macromolecular complexes. *Nucleic Acids Research 42*, D1 (12 2013), D297–D303.

[130] March-Vila, E., Pinzi, L., Sturm, N., Tinivella, A., Engkvist, O., Chen, H., and Rastelli, G. On the integration of in silico drug design methods for drug repurposing. *Frontiers in Pharmacology 8* (2017), 298.

[131] Mason, J. S., Morize, I., Menard, P. R., Cheney, D. L., Hulme, C., and Labaudiniere, R. F. New 4-point pharmacophore method for molecular similarity and diversity applications: overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. *Journal of Medicinal Chemistry 42*, 17 (1999), 3251–3264. PMID: 10464012.

[132] Mei, J.-P., Kwoh, C.-K., Yang, P., Li, X.-L., and Zheng, J. Drug–target interaction prediction by learning from local information and neighbors. *Bioinformatics 29*, 2 (11 2012), 238–245.

[133] Mei, J.-P., Kwoh, C.-K., Yang, P., Li, X.-L., and Zheng, J. Drug–target interaction prediction by learning from local information and neighbors. *Bioinformatics 29*, 2 (11 2012), 238–245.

[134] Michaleff, Z. A., Costa, L. O., Moseley, A. M., Maher, C. G., Elkins, M. R., Herbert, R. D., and Sherrington, C. CENTRAL, PEDro, PubMed, and EMBASE Are the Most Comprehensive Databases Indexing Randomized Controlled Trials of Physical Therapy Interventions. *Physical Therapy 91*, 2 (02 2011), 190–197.

[135] Milik, M., Szalma, S., and Olszewski, K. A. Common Structural Cliques: a tool for protein structure and function analysis. *Protein Engineering, Design and Selection 16*, 8 (08 2003), 543–552.

[136] Minai, R., Matsuo, Y., Onuki, H., and Hirota, H. Method for comparing the structures of protein ligand-binding sites and application for predicting protein–drug interactions. *Proteins: Structure, Function, and Bioinformatics 72*, 1 (2008), 367–381.

[137] Mitsuya, H., Weinhold, K. J., Furman, P. A., St Clair, M. H., Lehrman, S. N., Gallo, R. C., Bolognesi, D., Barry, D. W., and Broder, S. 3'-azido-3'-deoxythymidine (bw a509u): an antiviral agent that inhibits the infectivity and cytopathic effect of human t-lymphotropic virus type iii/lymphadenopathy-associated virus in vitro. *Proceedings of the National Academy of Sciences 82*, 20 (1985), 7096–7100.

[138] Mitsuya, H., Weinhold, K. J., Furman, P. A., St Clair, M. H., Lehrman, S. N., Gallo, R. C., Bolognesi, D., Barry, D. W., and Broder, S. 3'-azido-3'-deoxythymidine (bw a509u): an antiviral agent that inhibits the infectivity and cytopathic effect of human t-lymphotropic virus type iii/lymphadenopathy-associated virus in vitro. *Proceedings of the National Academy of Sciences 82*, 20 (1985), 7096–7100.

[139] Morgan, H. L. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of Chemical Documentation 5*, 2 (1965), 107–113.

[140] Mpamhanga, C. P., Chen, B., McLay, I. M., and Willett, P. Knowledge-based interaction fingerprint scoring: a simple method for improving the effectiveness of fast scoring functions. *Journal of Chemical Information and Modeling 46*, 2 (2006), 686–698. PMID: 16562999.

[141] Muranjan, M., Wang, Q., Li, Y. L., Hamilton, E., Otieno-Omondi, F. P., Wang, J., Van Praagh, A., Grootenhuis, J. G., and Black, S. J. The trypanocidal cape buffalo serum protein is xanthine oxidase. *Infection and Immunity 65*, 9 (1997), 3806–3814.

[142] Nair, V., and Hinton, G. E. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on*

International Conference on Machine Learning (Madison, WI, USA, 2010), ICML'10, Omnipress, p. 807–814.

[143] Najmanovich, R., Kurbatova, N., and Thornton, J. Detection of 3D atomic similarities and their use in the discrimination of small molecule protein-binding sites. *Bioinformatics 24*, 16 (08 2008), i105–i111.

[144] Nandigam, R. K., Kim, S., Singh, J., and Chuaqui, C. Position specific interaction dependent scoring technique for virtual screening based on weighted proteinligand interaction fingerprint profiles. *Journal of Chemical Information and Modeling 49*, 5 (2009), 1185–1192. PMID: 19415918.

[145] Needleman, S. B., and Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology 48*, 3 (1970), 443 – 453.

[146] Nickel, W., Austermann, S., Bialek, G., and Grosse, F. Interactions of azidothymidine triphosphate with the cellular dna polymerases alpha, delta, and epsilon and with dna primase. *Journal of Biological Chemistry 267*, 2 (1992), 848–854.

[147] Nosengo, N. Can you teach old drugs new tricks? *Nature 534*, 7607 (June 2016), 314—316.

[148] O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., and Hutchison, G. R. Open babel: An open chemical toolbox. *Journal of Cheminformatics 3*, 1 (2011), 33.

[149] O'Boyle, N. M., Morley, C., and Hutchison, G. R. Pybel: a python wrapper for the openbabel cheminformatics toolkit. *Chemistry Central journal 2* (Mar 2008), 5–5. 18328109[pmid].

[150] Overington, J. P., Al-Lazikani, B., and Hopkins, A. L. How many drug targets are there? *Nature Reviews Drug Discovery 5*, 12 (2006), 993–996.

[151] Pahikkala, T., Airola, A., Pietilä, S., Shakyawar, S., Szwajda, A., Tang, J., and Aittokallio, T. Toward more realistic drug–target interaction predictions. *Briefings in Bioinformatics 16*, 2 (04 2014), 325–337.

[152] Palme, J., Hochreiter, S., and Bodenhofer, U. KeBABS: an R package for kernel-based analysis of biological sequences. *Bioinformatics 31*, 15 (03 2015), 2574–2576.

[153] Pammolli Fabio, Magazzini Laura, R. M. The productivity crisis in pharmaceutical Ramp;D. *Nature Reviews Drug Discovery 10* (2011), 428.

[154] Park, K., and Kim, D. Binding similarity network of ligand. *Proteins: Structure, Function, and Bioinformatics 71*, 2 (2008), 960–971.

[155] Pawson, A. J., Sharman, J. L., Benson, H. E., Faccenda, E., Alexander, S. P., Buneman, O. P., Davenport, A. P., McGrath, J. C., Peters, J. A., Southan, C., Spedding, M., Yu, W., Harmar, A. J., and NC-IUPHAR. The IUPHAR/BPS Guide to PHARMACOLOGY: an expert-driven knowledgebase of drug targets and their ligands. *Nucleic Acids Research 42*, D1 (11 2013), D1098–D1106.

[156] Peters, K. P., Fauck, J., and Frömmel, C. The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. *Journal of Molecular Biology 256*, 1 (1996), 201 – 213.

[157] Placzek, S., Schomburg, I., Chang, A., Jeske, L., Ulbrich, M., Tillack, J., and Schomburg, D. BRENDA in 2017: new perspectives and new tools in BRENDA. *Nucleic Acids Research 45*, D1 (10 2016), D380–D388.

[158] Powers, R., Copeland, J. C., Germer, K., Mercier, K. A., Ramanathan, V., and Revesz, P. Comparison of protein active site structures for functional annotation of proteins and drug design. *Proteins: Structure, Function, and Bioinformatics 65*, 1 (2006), 124–135.

[159] Pushpakom, S., Iorio, F., Eyers, P. A., Escott, K. J., Hopper, S., Wells, A., Doig, A., Guilliams, T., Latimer, J., McNamee, C., Norris, A., Sanseau, P., Cavalla, D., and Pirmohamed, M. Drug repurposing: progress, challenges and recommendations. *Nature Reviews Drug Discovery 18* (Oct 2018), 41 EP –.

[160] Pérez-Nueno, V. I., Rabal, O., Borrell, J. I., and Teixidó, J. Apif: A new interaction fingerprint based on atom pairs and its application to virtual screening. *Journal of Chemical Information and Modeling 49*, 5 (2009), 1245–1260. PMID: 19364101.

[161] Pérot, S., Sperandio, O., Miteva, M. A., Camproux, A.-C., and Villoutreix, B. O. Druggable pockets and binding site centric chemical space: a paradigm shift in drug discovery. *Drug Discovery Today 15*, 15 (2010), 656 – 667.

[162] Reddy, A. S., and Zhang, S. Polypharmacology: drug discovery for the future. *Expert Review of Clinical Pharmacology 6*, 1 (2013), 41–47. PMID: 23272792.

[163] Rogers, D., Brown, R. D., and Hahn, M. Using extended-connectivity fingerprints with laplacian-modified bayesian analysis in high-throughput

screening follow-up. *Journal of Biomolecular Screening 10*, 7 (2005), 682–686. PMID: 16170046.

[164] Rogers, D., and Hahn, M. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling 50*, 5 (2010), 742–754. PMID: 20426451.

[165] Roider, H. G., Pavlova, N., Kirov, I., Slavov, S., Slavov, T., Uzunov, Z., and Weiss, B. Drug2gene: an exhaustive resource to explore effectively the drug-target relation network. *BMC Bioinformatics 15*, 1 (2014), 68.

[166] Rollinger, J. M., Schuster, D., Danzl, B., Schwaiger, S., Markt, P., Schmidtke, M., Gertsch, J., Raduner, S., Wolber, G., Langer, T., and Stuppner, H. In silico Target Fishing for Rationalized Ligand Discovery Exemplified on Constituents of Ruta graveolens. *Planta Med* (2009).

[167] Rost, B. Twilight zone of protein sequence alignments. *Protein Engineering, Design and Selection 12*, 2 (02 1999), 85–94.

[168] Roth, B. L., Lopez, E., Patel, S., and Kroeze, W. K. The multiplicity of serotonin receptors: Uselessly diverse molecules or an embarrassment of riches? *The Neuroscientist 6*, 4 (2000), 252–262.

[169] Rupp, M., Tkatchenko, A., Müller, K.-R., and von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett. 108* (Jan 2012), 058301.

[170] Salentin, S., Adasme, M. F., Heinrich, J. C., Haupt, V. J., Daminelli, S., Zhang, Y., and Schroeder, M. From malaria to cancer: Computational drug repositioning of amodiaquine using PLIP interaction patterns. *Sci Rep 7*, 1 (09 2017), 11401.

[171] Salentin, S., Schreiber, S., Haupt, V. J., Adasme, M. F., and Schroeder, M. PLIP: fully automated protein–ligand interaction profiler. *Nucleic Acids Research 43*, W1 (04 2015), W443–W447.

[172] Sanchez-Lengeling, B., and Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science 361*, 6400 (2018), 360–365.

[173] Sanseau, P., Hasan, S., Koscielny, G., Papa, E., An, P., Wankar, P., Gasparyan, R., Platt, T., Bento, A. P., Munoz-Pomer Fuentes, A., Pierleoni, A., Gaulton, A., Hersey, A., Fabregat, A., Palka, B., Vaughan, B., Leroy, C., O'Donovan, C., Gonzalez, C. Y., Carvalho-Silva, D., Turner, E., Birney, E., Rowland, F., Lopez, F.-J., Saunders, G., Parkinson, H., Papatheodorou, I., Malone, J., Cham, J. A., McEntyre, J., Paschall, J., Sidiropoulos, K., Fumis, L., Keays,

M., Magarinos, M. P., Martin, M.-J., Maguire, M., Pignatelli, M., Pratanwanich, N., Karamanis, N., Vrousgou, O., Stegle, O., Petryszak, R., Kafkas, , Sarntivijal, S., Jupe, S., Smith, T., Burdett, T., Watkins, X., Tang, Y. A., Dunham, I., Barrett, J., Hermjakob, H., Forbes, S., Sondka, Z., and Vamathevan, J. Open Targets: a platform for therapeutic target identification and validation. *Nucleic Acids Research 45*, D1 (12 2016), D985–D994.

[174] Santos, R., Ursu, O., Gaulton, A., Bento, A. P., Donadi, R. S., Bologa, C. G., Karlsson, A., Al-Lazikani, B., Hersey, A., Oprea, T. I., and Overington, J. P. A comprehensive map of molecular drug targets. *Nat Rev Drug Discov 16*, 1 (01 2017), 19–34.

[175] Sato, T., Honma, T., and Yokoyama, S. Combining machine learning and pharmacophore-based interaction fingerprint for in silico screening. *Journal of Chemical Information and Modeling 50*, 1 (2010), 170–185. PMID: 20038188.

[176] Scannell Jack W., Blanckley Alex, B. H. W. B. Diagnosing the decline in pharmaceutical Ramp;D efficiency. *Nature Reviews Drug Discovery 11* (2012), 191.

[177] Schadt, E. E. Molecular networks as sensors and drivers of common human diseases. *Nature 461*, 7261 (2009), 218–223.

[178] Schalon, C., Surgand, J.-S., Kellenberger, E., and Rognan, D. A simple and fuzzy method to align and compare druggable ligand-binding sites. *Proteins: Structure, Function, and Bioinformatics 71*, 4 (2008), 1755–1778.

[179] Schmitt, S., Kuhn, D., and Klebe, G. A new method to detect related function among proteins independent of sequence and fold homology. *Journal of Molecular Biology 323*, 2 (2002), 387 – 406.

[180] Schuffenhauer, A., Brown, N., Ertl, P., Jenkins, J. L., Selzer, P., and Hamon, J. Clustering and rule-based classifications of chemical structures evaluated in the biological activity space. *Journal of Chemical Information and Modeling 47*, 2 (2007), 325–336. PMID: 17286395.

[181] Shibuya, M. Vascular endothelial growth factor and its receptor system: physiological functions in angiogenesis and pathological roles in various diseases. *The Journal of Biochemistry 153*, 1 (11 2012), 13–19.

[182] Shulman-Peleg, A., Nussinov, R., and Wolfson, H. J. Recognition of functional sites in protein structures. *Journal of Molecular Biology 339*, 3 (2004), 607 – 633.

[183] Shulman-Peleg, A., Shatsky, M., Nussinov, R., and Wolfson, H. J. Multibind and mappis: webservers for multiple alignment of protein 3d-binding sites and their interactions. *Nucleic acids research 36*, Web Server issue (Jul 2008), W260–W264. 18467424[pmid].

[184] Simm, J., Arany, A., Zakeri, P., Haber, T., Wegner, J. K., Chupakhin, V., Ceulemans, H., and Moreau, Y. Macau: Scalable bayesian multi-relational factorization with side information using mcmc, 2015.

[185] Simm, J., Arany, A., Zakeri, P., Haber, T., Wegner, J. K., Chupakhin, V., Ceulemans, H., and Moreau, Y. Macau: Scalable bayesian factorization with high-dimensional side information using mcmc. In *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)* (2017), pp. 1–6.

[186] Sirota, M., Dudley, J. T., Kim, J., Chiang, A. P., Morgan, A. A., Sweet-Cordero, A., Sage, J., and Butte, A. J. Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Science Translational Medicine 3*, 96 (2011), 96ra77–96ra77.

[187] Southan, C., Sharman, J. L., Benson, H. E., Faccenda, E., Pawson, A. J., Alexander, S., Buneman, O. P., Davenport, A. P., McGrath, J. C., Peters, J. A., Spedding, M., Catterall, W. A., Fabbro, D., Davies, J. A., and NC-IUPHAR. The IUPHAR/BPS Guide to PHARMACOLOGY in 2016: towards curated quantitative interactions between 1300 protein targets and 6000 ligands. *Nucleic Acids Research 44*, D1 (10 2015), D1054–D1068.

[188] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res. 15*, 1 (Jan. 2014), 1929–1958.

[189] Steindl, T. M., Schuster, D., Laggner, C., Chuang, K., Hoffmann, R. D., and Langer, T. Parallel screening and activity profiling with hiv protease inhibitor pharmacophore models. *Journal of Chemical Information and Modeling 47*, 2 (2007), 563–571. PMID: 17381173.

[190] Stott. Pharma's broken business model: An industry on the brink of terminal decline. `https://endpts.com/pharmas-broken-business-model-an-industry-on-the-brink-of-terminal-decline/`, 2017.

[191] Stumpfe, D., and Bajorath, J. Exploring activity cliffs in medicinal chemistry. *Journal of Medicinal Chemistry 55*, 7 (2012), 2932–2942. PMID: 22236250.

[192] Sullivan, D. C., and Martin, E. J. Exploiting structure–activity relationships in docking. *Journal of Chemical Information and Modeling 48*, 4 (2008), 817–830. PMID: 18399649.

[193] Sun, J., Wu, Y., Xu, H., and Zhao, Z. Dtome: a web-based tool for drug-target interactome construction. *BMC bioinformatics 13 Suppl 9*, Suppl 9 (Jun 2012), S7–S7. 22901092[pmid].

[194] Suthram, S., Dudley, J. T., Chiang, A. P., Chen, R., Hastie, T. J., and Butte, A. J. Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. *PLOS Computational Biology 6*, 2 (02 2010), 1–10.

[195] Tan, L., Batista, J., and Bajorath, J. Computational methodologies for compound database searching that utilize experimental protein–ligand interaction information. *Chemical Biology & Drug Design 76*, 3 (2010), 191–200.

[196] Tan, L., Vogt, M., and Bajorath, J. Three-dimensional protein–ligand interaction scaling of two-dimensional fingerprints. *Chemical Biology & Drug Design 74*, 5 (2009), 449–456.

[197] Tiriveedhi, V. Impact of precision medicine on drug repositioning and pricing: A too small to thrive crisis. *Journal of Personalized Medicine 8*, 4 (2018).

[198] Tsou, L. K., Cheng, Y., and Cheng, Y.-C. Therapeutic development in targeting protein–protein interactions with synthetic topological mimetics. *Current Opinion in Pharmacology 12*, 4 (2012), 403 – 407. Cancer • Immunomodulation.

[199] Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., and Zhao, S. Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery 18*, 6 (2019), 463–477.

[200] Van Belle, K., Herman, J., Boon, L., Waer, M., Sprangers, B., and Louat, T. Comparative in vitro immune stimulation analysis of primary human b cells and b cell lines. *Journal of immunology research 2016* (2016), 5281823–5281823. 28116319[pmid].

[201] van Hoorn, W. P., and Bell, A. S. Searching chemical space with the bayesian idea generator. *Journal of Chemical Information and Modeling 49*, 10 (2009), 2211–2220. PMID: 19788263.

[202] van Laarhoven, T., and Marchiori, E. Predicting drug-target interactions for new drug compounds using a weighted nearest neighbor profile. *PLOS ONE 8*, 6 (06 2013), 1–6.

[203] van Laarhoven, T., Nabuurs, S. B., and Marchiori, E. Gaussian interaction profile kernels for predicting drug–target interaction. *Bioinformatics 27*, 21 (09 2011), 3036–3043.

[204] Venhorst, J., Núñez, S., Terpstra, J. W., and Kruse, C. G. Assessment of scaffold hopping efficiency by use of molecular interaction fingerprints. *Journal of Medicinal Chemistry 51*, 11 (2008), 3222–3229. PMID: 18447336.

[205] von Eichborn, J., Dunkel, M., Gohlke, B. O., Preissner, S. C., Hoffmann, M. F., Bauer, J. M. J., Armstrong, J. D., Schaefer, M. H., Andrade-Navarro, M. A., Le Novere, N., Croning, M. D. R., Grant, S. G. N., van Nierop, P., Smit, A. B., and Preissner, R. SynSysNet: integration of experimental data on synaptic protein–protein interactions with drug-target relations. *Nucleic Acids Research 41*, D1 (11 2012), D834–D840.

[206] von Eichborn, J., Murgueitio, M. S., Dunkel, M., Koerner, S., Bourne, P. E., and Preissner, R. PROMISCUOUS: a database for network-based drug-repositioning. *Nucleic Acids Research 39*, $\text{suppl}_1(112010), D1060 - -D1066$.

[207] Wale, N., and Karypis, G. Target fishing for chemical compounds using target-ligand activity data and ranking based methods. *Journal of Chemical Information and Modeling 49*, 10 (2009), 2190–2201. PMID: 19764745.

[208] Wang, L., You, Z.-H., Chen, X., Xia, S.-X., Liu, F., Yan, X., Zhou, Y., and Song, K.-J. A computational-based method for predicting drug–target interactions by using stacked autoencoder deep neural network. *Journal of Computational Biology 25*, 3 (2018), 361–373. PMID: 28891684.

[209] Wang, X., Shen, Y., Wang, S., Li, S., Zhang, W., Liu, X., Lai, L., Pei, J., and Li, H. PharmMapper 2017 update: a web server for potential drug target identification with a comprehensive target pharmacophore database. *Nucleic Acids Research 45*, W1 (05 2017), W356–W360.

[210] Wang, Y., Bryant, S. H., Cheng, T., Wang, J., Gindulyte, A., Shoemaker, B. A., Thiessen, P. A., He, S., and Zhang, J. PubChem BioAssay: 2017 update. *Nucleic Acids Research 45*, D1 (11 2016), D955–D963.

[211] Weill, N., and Rognan, D. Development and validation of a novel proteinligand fingerprint to mine chemogenomic space: Application to g protein-coupled receptors and their ligands. *Journal of Chemical Information and Modeling 49*, 4 (2009), 1049–1062. PMID: 19301874.

[212] Weisel, M., Proschak, E., and Schneider, G. Pocketpicker: analysis of ligand binding-sites with shape descriptors. *Chemistry Central Journal 1*, 1 (2007), 7.

[213] Wen, M., Zhang, Z., Niu, S., Sha, H., Yang, R., Yun, Y., and Lu, H. Deep-learning-based drug–target interaction prediction. *Journal of Proteome Research 16*, 4 (2017), 1401–1409. PMID: 28264154.

[214] Wermuth, C. G. The sosa approach: An alternative to high-throughput screening. *Medicinal Chemistry Research 10* (2001), 431–439.

[215] Wermuth, C. G., Ganellin, C. R., Lindberg, P., and Mitscher, L. A. Glossary of terms used in medicinal chemistry (iupac recommendations 1998). *Pure and Applied Chemistry 70*, 5 (01 Jan. 1998), 1129 – 1143.

[216] Wishart, D. S., Knox, C., Guo, A. C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., and Woolsey, J. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research 34*, $\text{suppl}_1(012006), D668 - -D672$.

[217] Wolber, G., and Langer, T. Ligandscout: 3-d pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *Journal of Chemical Information and Modeling 45*, 1 (2005), 160–169. PMID: 15667141.

[218] Wu, Z., Cheng, F., Li, J., Li, W., Liu, G., and Tang, Y. SDTNBI: an integrated network and chemoinformatics tool for systematic prediction of drug–target interactions and drug repositioning. *Briefings in Bioinformatics 18*, 2 (03 2016), 333–347.

[219] Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., and Pande, V. Moleculenet: a benchmark for molecular machine learning. *Chemical science 9*, 2 (Oct 2017), 513–530. 29629118[pmid].

[220] Xia, Z., Wu, L.-Y., Zhou, X., and Wong, S. T. C. Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. *BMC systems biology 4 Suppl 2*, Suppl 2 (Sep 2010), S6–S6. 20840733[pmid].

[221] Xiao, N., Cao, D.-S., Zhu, M.-F., and Xu, Q.-S. protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. *Bioinformatics 31*, 11 (01 2015), 1857–1859.

[222] Xiao, X., Min, J.-L., Wang, P., and Chou, K.-C. igpcr-drug: A web server for predicting interaction between gpcrs and drugs in cellular networking. *PLOS ONE 8*, 8 (08 2013), 1–10.

[223] Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W., and Kanehisa, M. Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics 24*, 13 (07 2008), i232–i240.

[224] Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W., and Kanehisa, M. Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics 24*, 13 (07 2008), i232–i240.

[225] Yamanishi, Y., Kotera, M., Kanehisa, M., and Goto, S. Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics 26*, 12 (Jun 2010), i246–254.

[226] Yamanishi, Y., Kotera, M., Moriya, Y., Sawada, R., Kanehisa, M., and Goto, S. DINIES: drug–target interaction network inference engine based on supervised analysis. *Nucleic Acids Research 42*, W1 (05 2014), W39–W45.

[227] Yan, X.-Y., Zhang, S.-W., and He, C.-R. Prediction of drug-target interaction by integrating diverse heterogeneous information source with multiple kernel learning and clustering methods. *Computational Biology and Chemistry 78* (2019), 460 – 467.

[228] Yang, H., Qin, C., Li, Y. H., Tao, L., Zhou, J., Yu, C. Y., Xu, F., Chen, Z., Zhu, F., and Chen, Y. Z. Therapeutic target database update 2016: enriched resource for bench to clinical drug target and targeted pathway information. *Nucleic Acids Research 44*, D1 (11 2015), D1069–D1074.

[229] Yang, H., Shelat, A. A., Guy, R. K., Gopinath, V. S., Ma, T., Du, K., Lukacs, G. L., Taddei, A., Folli, C., Pedemonte, N., Galietta, L. J. V., and Verkman, A. S. Nanomolar affinity small molecule correctors of defective f508-cftr chloride channel gating. *Journal of Biological Chemistry 278*, 37 (2003), 35079–35085.

[230] Yang, K., Bai, H., Ouyang, Q., Lai, L., and Tang, C. Finding multiple target optimal intervention in disease-related molecular network. *Molecular Systems Biology 4*, 1 (2008), 228.

[231] Yao, Z.-J., Dong, J., Che, Y.-J., Zhu, M.-F., Wen, M., Wang, N.-N., Wang, S., Lu, A.-P., and Cao, D.-S. Targetnet: a web service for predicting potential drug–target interaction profiling via multi-target sar models. *Journal of Computer-Aided Molecular Design 30*, 5 (May 2016), 413–424.

[232] Yella, J. K., Yaddanapudi, S., Wang, Y., and Jegga, A. G. Changing trends in computational drug repositioning. *Pharmaceuticals 11*, 2 (2018).

[233] Yeu, Y., Yoon, Y., and Park, S. Protein localization vector propagation: a method for improving the accuracy of drug repositioning. *Mol. BioSyst. 11* (2015), 2096–2102.

[234] Yuan, Q., Gao, J., Wu, D., Zhang, S., Mamitsuka, H., and Zhu, S. DrugE-Rank: improving drug–target interaction prediction of new candidate

drugs or targets by ensemble learning to rank. *Bioinformatics 32*, 12 (06 2016), i18–i27.

[235] Zheng, X., Ding, H., Mamitsuka, H., and Zhu, S. Collaborative matrix factorization with multiple similarities for predicting drug-target interactions. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, 2013), KDD '13, ACM, pp. 1025–1033.

[236] Zhou, H., Wang, C., Deng, T., Tao, R., and Li, W. Novel urushiol derivatives as hdac8 inhibitors: rational design, virtual screening, molecular docking and molecular dynamics studies. *Journal of Biomolecular Structure and Dynamics 36*, 8 (2018), 1966–1978. PMID: 28632421.

[237] Zhou, L., Li, Z., Yang, J., Tian, G., Liu, F., Wen, H., Peng, L., Chen, M., Xiang, J., and Peng, L. Revealing drug-target interactions with computational models and algorithms. *Molecules 24*, 9 (2019).

Curriculum



Daniele Parisi was born in 1991 in a small city on the southern Italian coast (Salerno). He got a Bachelor in Biotechnology at *La Sapienza, University of Rome* with a thesis on drug delivery at *King's College London*. After that, he accomplished an MSc in Pharmaceutical Biotechnology at *La Sapienza, University of Rome* with a thesis on structure-based drug repositioning at *KULeuven*. He moved to Belgium to continue his study on computational drug repositioning with a Ph.D. program at *KULeuven*. During his Ph.D. he spent an exchange year at *TU-Dresden* to study structural bioinformatics and structure-based drug repositioning. At the moment, He continues his academic career at the *Yale Medical School*, where he is a postdoc associate at the department of immunobiology.

List of publications

- *Drug repositioning or target repositioning: A structural perspective of drug-target-indication relationship for available repurposed drugs*; **D Parisi**, MF Adasme, A Sveshnikova, SN Bolz, Y Moreau, M Schroeder; Computational and Structural Biotechnology Journal 18, 1043, 2020.
- *Structure-based Drug repositioning: Potential and Limits*; MF Adasme, **D Parisi**, A Sveshnikova, M Schroeder; Seminars in Cancer Biology,2020.
- *Structure-based drug repositioning explains ibrutinib as VEGFR2 inhibitor* MF Adasme, **D Parisi**, K Van Belle, S Salentin, VJ Haupt, GS Jennings, Jörg-Christian Heinrich, Jean Herman, Ben Sprangers, Thierry Louat, Yves Moreau, Michael Schroeder; Plos one 15 (5), e0233089, 2020.
- *Nurturing tomorrow's leaders: The ISCB Student Council Symposia in 2018*; **D Parisi**, GJ Olguín-Orellana, EJ Draizen, NDR Coimbra, , Nikolaos Papadopoulos, Susanne Kirchen, Yvonne Saara Gladbach, Numrah Fadra, Nazeefa Fatima, Aishwarya Alex Namasivayam, Sayane Shome, Dan DeBlasio, Alexander M Monzon, Farzana Rahman, R Gonzalo Parra; F1000Research 8, 2019.
- *Global network of computational biology communities: ISCB's Regional Student Groups breaking barriers*; S Shome, RG Parra, N Fatima, AM Monzon, B Cuypers, Y Moosa, Nilson Da Rocha Coimbra, Juliana Assis, Carla Giner-Delgado, Handan Melike Dönertaş, Yesid Cuesta-Astroz, Geetha Saarunya, Imane Allali, Shruti Gupta, Ambuj Srivastava, Manisha Kalsan, Catalina Valdivia, Gabriel J Olguin-Orellana, Sofia Papadimitriou, **Daniele Parisi**, Nikolaj Pagh Kristensen, Leonor Rib, Marouen Ben Guebila, Eugen Bauer, Gaia Zaffaroni, Amel Bekkar, Efejiro Ashano, Lisanna Paladin, Marco Necci, Nicolás N Moreyra, Martin Rydén, Jordan Villalobos-Solís, Nikolaos Papadopoulos, Candice Rafael, Tülay Karakulak, Yasin Kaya, Yvonne Gladbach, Sandeep Kumar Dhanda, Nikolina Šoštarić, Aishwarya Alex, Dan DeBlasio, Farzana Rahman; F1000Research 8, 2019.

FACULTY OF ENGINEERING SCIENCE
DEPARTMENT OF ELECTRICAL ENGINEERING
STADIUS
Kasteelpark Arenberg 10, 3001 Leuven, Belgium
B-3001 Leuven