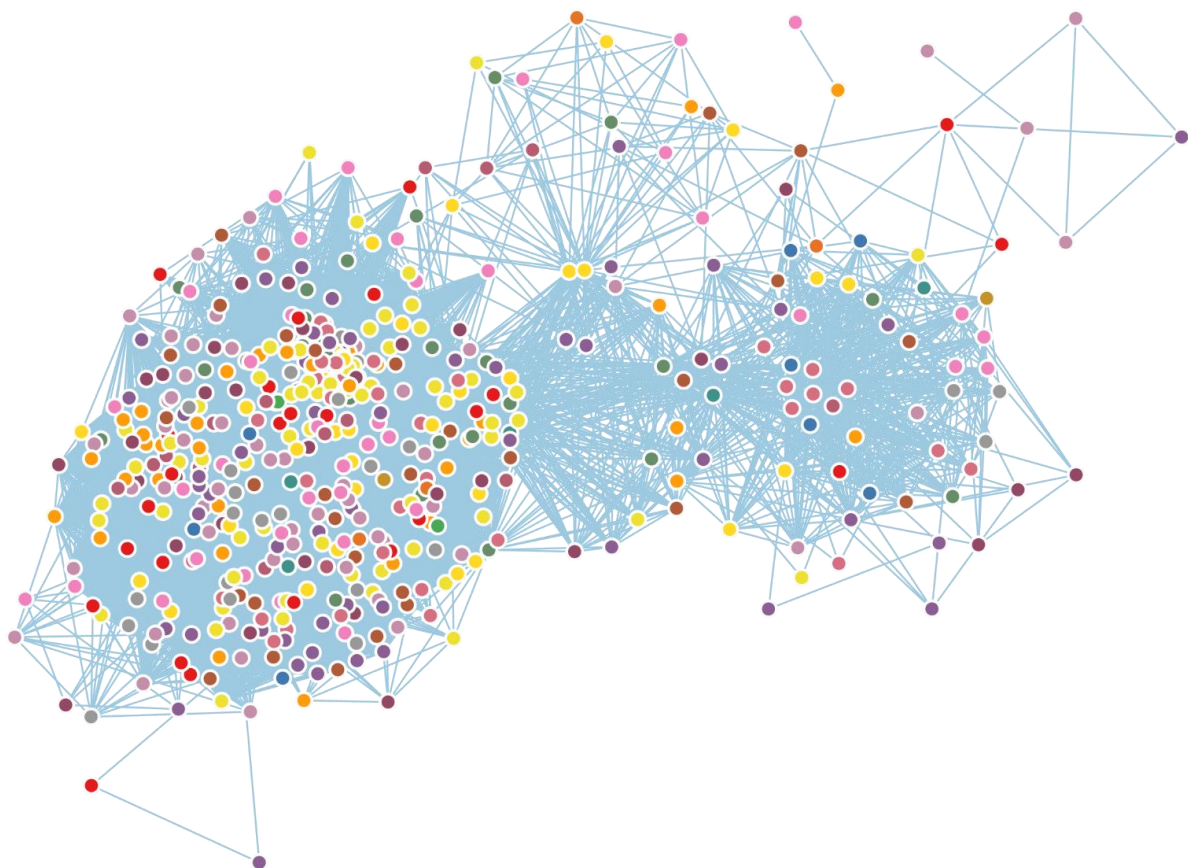


Exploring the Drug-Adverse Reaction and Drug-Target Landscape through Networks, Statistics and Machine Learning Approaches

Cristiano Galletti



Exploring the Drug-Adverse Reaction and Drug-Target Landscape through Networks, Statistics and Machine Learning Approaches

Cristiano Galletti

TESI DOCTORAL UVIC / 2022

THESIS DIRECTOR: Dr. Narcis Fernandez Fuentes

THESIS TUTOR: Dr. M. Luz Calle Rosingana

*Every day you may make progress.
Every step may be fruitful.
Yet there will stretch out before you an ever-lengthening,
ever-ascending, ever-improving path.
You know you will never get to the end of the journey.
But this, so far from discouraging,
only adds to the joy and glory of the climb.
—Winston Churchill*

ACKNOWLEDGMENTS

To be honest, I've never been particularly good at these things, but after four years in Spain for this PhD, it seems like the least I can do is thank everyone who has helped me in some way in carrying on this work.

First and foremost, I want to thank my advisor, Prof. **Narcis** Fernandez-Fuentes, for his unwavering support of my PhD studies and related research, as well as his patience, motivation, and insight.

Even when I was 1000 kilometers away or confined at home during the pandemic, his advice was invaluable throughout this entire research and writing of the thesis.

My sincere thanks also go to professors **Baldo** Oliva and **Malu** Luz Calle Rosingana for their insightful comments and encouragement throughout my research.

We weren't able to collaborate as closely as I would have liked, but I appreciated every minute of it.

I'd also want to thank the **Prous** Institute for Biomedical Research's Dr. **Joseph** Prous, Dr. **David** Prous, and Dr. **Ramon** Flores Pons for their assistance in the early stages of the project and for allowing me to witness how a successful firm operates from the inside.

Special thanks to my little Italian enclave of Barcelona researchers, **Leila**, **Andrea**, **Pat**, **Damiano**, and **Pippo**. We spent two years together in Bologna during the master's program, and I'm thrilled to have had the opportunity to continue our friendship here in Spain for the PhD.

When I wanted to enjoy Barcelona and good company, you were my safe harbor.

I would like to offer my special thanks also to my BETA Spanish friends, who welcomed me into their first PhD office before moving to the other side of Vic.

I hope I wasn't the reason for the relocation.

Another special thanks go to **Graeme**, or should I say **Dr. Dean**, who supported me spiritually throughout the PhD and, above all, for teaching me to play pool like a real Irishman does.

Last but not least, I'd like to thank my family, even if they're still confused about my field and what I'm doing.

To my **mother**, the best mother on the planet.

Thank you for always being there for me, worrying about me, and bestowing many of your qualities on me.

To my **father**, even if we don't see or talk as much as I'd like, I know you're proud of me.

To my **uncles, aunts, and grandmothers** who were always concerned about whether I was eating enough in Spain.

Finally, a special thanks to my girlfriend **Silvia**, my release valve and therapist.

This thesis is dedicated to all of you, and a little bit also to myself, always unsecure of what the future brings.

RINGRAZIAMENTI

Ad essere sincero, non sono mai stato particolarmente bravo in queste cose, ma dopo quattro anni in Spagna per questo dottorato, il minimo che posso fare è ringraziare tutti coloro che in qualche modo mi hanno aiutato a portare avanti questo lavoro.

Innanzitutto, voglio ringraziare il mio advisor, il Prof. **Narcis** Fernandez-Fuentes, per il suo incrollabile sostegno ai miei studi e nella ricerca, nonché per la sua pazienza, motivazione e saggezza.

Anche quando ero a 1000 chilometri di distanza o confinato a casa durante la pandemia, i suoi consigli sono stati preziosi per l'intera ricerca e la stesura di questa tesi.

I miei più sinceri ringraziamenti vanno anche ai professori **Baldo** Oliva e **Malu** Luz Calle Rosingana per i loro suggerimenti e l'incoraggiamento durante il mio corso di studi.

Non siamo stati in grado di collaborare strettamente come avrei voluto, ma ne ho apprezzato ogni minuto.

Vorrei anche ringraziare il Dr. **Joseph** Prous, il Dr. **David** Prous e il Dr. **Ramon** Flores Pons del **Prous** Institute for Biomedical Research per la loro assistenza nelle prime fasi del progetto e per avermi permesso di vedere come opera un'azienda di successo dall'interno.

Un ringraziamento speciale alla mia piccola enclave italiana di ricercatori a Barcellona, **Leila**, **Andrea**, **Pat**, **Damiano** e **Pippo**. Abbiamo trascorso due anni insieme a Bologna al Master e sono entusiasta di aver avuto l'opportunità di continuare la nostra amicizia qui in Spagna per il dottorato.

Quando volevo godermi Barcellona e una buona compagnia, voi eravate il mio porto sicuro.

Vorrei esprimere un ringraziamento speciale anche ai miei amici spagnoli del BETA, che mi hanno accolto nel loro primo ufficio dei PhD prima di trasferirsi dall'altra parte di Vic.

Spero di non essere stato io il motivo del trasferimento.

Un altro ringraziamento speciale va a **Graeme**, o dovrei dire al **Dr. Dean**, che mi ha supportato spiritualmente durante tutto il dottorato e, soprattutto, per avermi insegnato a giocare a biliardo come fa un vero irlandese.

Ultimo ma non meno importante, vorrei ringraziare la mia famiglia anche se non sono ancora sicuri di quale sia il mio campo e cosa stia facendo esattamente.

A mia **madre**, che è la migliore madre del mondo.

Grazie per essere sempre lì per me, preoccuparti per me e di avermi concesso molti dei tuoi pregi.

A mio **padre** che se anche non ci vediamo e parliamo spesso, sono sicuro che sei orgoglioso di me.

Ai miei **zii, zie e nonne** che erano sempre preoccupati se mangiassi abbastanza in Spagna.

Infine un grazie speciale alla mia ragazza **Silvia**, la mia valvola di sfogo e terapeuta.

Questa tesi è dedicata a tutti voi, e un pochino anche a me, sempre insicuro su quello che il futuro porti.

ABSTRACT

The development of a novel drug is a long and winded process plagued with challenges and pitfalls. Among these, the lack of toxicological or safety knowledge for targets is one of the most significant challenges in drug development [1]. In other words, it is very difficult to know *a priori* if the targeting of a protein by a drug will result in the so-called undesirable adverse drug reactions (ADRs). Indeed, clinical trials have a high incidence of drug attrition due to the severity of ADRs associated with toxicity, which drives up costs and limits the development of new therapies for emerging targets [2].

To reduce the risk associated with the development of novel drugs, various approaches, including the use of animal models and *in vitro* toxicology studies, have been used in past years [3] [1]. However, *in vitro* models have high maintenance costs and ethical concerns, not to mention that they are not always applicable to human biology [4]. As a result, researchers were forced to adapt to new strategies, and the vast majority of recent advances are built on computational frameworks.

The new methodologies applied include various examples of machine learning and deep learning, which have been used in target-based predictions, analyses of the underlying protein network and interactions, and quantitative structure–activity connections studies. The study of protein–protein interactions related to drug discoveries, in particular, has attracted important attention in recent years and piqued pharmaceutical companies' interest. Indeed, high-coverage protein interaction maps can be used to find feasible therapeutic targets from which to develop or repurpose medications (as in the case of the COVID-19 drug race), as well as to find specific interactions that may contribute to the beginning of drug toxicity as this thesis will focus. [5] [6] [7]

While methodologies and mechanisms for linking candidate drugs with ADRs are well established, the association of ADRs with protein targets is less so but still studied. Two recent examples of the latter are the ADReCS-Target database [6], a recent study on ADRs generated from clinical trials and post-marketing reports [8] and a peculiar work of Kuhn and colleagues [9].

The lack of a protein - focused method to assess drug toxicity is one of the main gaps that my thesis aims to fill, starting from standardizing the link between ADRs and protein targets, in the hope that this information may be used to cut the time and costs of pre-clinical studies. As previously stated, there is no clear methodology for obtaining ADR-target data; however, this information can be retrieved using drugs as a connecting element to identify the link between ADR and proteins. In theory, if drug X produces ADR Y and drug X interacts with protein Z, then protein Z is linked to ADR Y. This simple assertion, however, is incorrect.

As Kuhn and colleagues demonstrated, most drugs bind to groups of pharmacologically similar proteins, such as members of the same protein family [9]. While only one of the targets is likely to be responsible for a specific ADR, a direct Target-ADR relationship, such as the one used in this oversimplified model, would link each target to every possible ADR of the same drug, leading to false positives. To avoid this, the relationship must be statistically evaluated, and Kuhn et al. propose a method for identifying statistically significant links between ADR and proteins using drugs as connecting factors [9].

Using this prior knowledge, I created the T-ARDIS (Target-Adverse Reaction Database Integrated Search) database [10], which attempts to demystify the ADRs-protein targets landscape. Since T-ARDIS provides a direct link between proteins and ADRs, the question arose as to whether this information can be used to predict potential ADRs linked to proteins. The answer was the development of DocTOR (Direct fOreCast Target On Reaction - [11]), a target-centric prediction method that uses T-ARDIS information to train a combination of machine-learning classifiers to predict whether the modulation of a given protein is likely to result in ADR. In some way, all of the measurements used in DocTOR exploit network-based information, and thus include elements that are intrinsic not only to the protein but also to their associations.

The accuracy of the obtained models justified their use in identifying problematic protein targets at the individual ADR level as well as across a group of related ADRs aggregated into common system organ classes.

The development of DocTOR led naturally to the final part of my thesis that dealt with understanding the molecular basis of the relationship between ADRs using information of protein targets underlying such ADRs. The SONG (Side effect On Network Graph) analysis allowed the study of relationships between ADRs condensing the vast ADR-target landscape into a novel network called "Adverse Reactome." As the name might suggest, this network translates the ADRs identified by T-ARDIS as nodes and the protein shared by the latter as edges. Using a clustering method to extract the relevant association of nodes and targets, this approach may be able to shed light on the possible role of ADRs associations and the molecular basis of ADRs emergence by extrapolating the enriched functions of the identified cluster's proteins.

All of this work is devoted to assist researchers and pharmaceutical companies in their pursuit of safer and more effective drugs.

RESUM

El desenvolupament d'un nou fàrmac és un procés llarg i sinuós ple de reptes i esculls. Entre aquests, la manca de coneixements toxicològics o de seguretat de les proteïnes és un dels reptes més importants en el desenvolupament de fàrmacs [1]. En altres paraules, és molt difícil saber *a priori* si modulació d'una proteïna per part d'un fàrmac donarà lloc a les anomenades reaccions adverses als medicaments (RAM). De fet, és en etapes avançades dels assaigs clínics que deguda a la gravetat de les RAM part del fàrmac que s'estan investigant s'han de abandonar augmentant d'aquesta manera els costos i limitant el desenvolupament de noves teràpies [2].

Per reduir el risc associat al desenvolupament de nous fàrmacs s'utilitzen diferents estratègies com l'ús de models animals i estudis de toxicologia *in vitro* [3] [1]. Tanmateix, els models animals tenen uns costos de manteniment elevats i problemes de tipus ètics, sense oblidar que no sempre són aplicables a la biologia humana [4]. Com a resultat, els investigadors s'han vist obligats a adaptar-se a noves estratègies i la gran majoria dels avenços recents es construeixen en marcs de noves eines computacionals. Aquestes noves eines inclouen mètodes basats en intel·ligència artificial i s'han utilitzat en prediccions basades en anàlisis de la xarxa i interaccions de proteïnes subjacents així com estudis quantitius de estructura-activitat a fàrmacs. L'estudi de les interaccions proteïna-proteïna relacionades amb els descobriments de fàrmacs, en particular, ha tingut una atenció important en els últims anys i ha despertat l'interès de les empreses farmacèutiques. De fet, els mapes d'interacció de proteïnes es poden utilitzar per trobar dianes terapèutiques a partir dels quals desenvolupar o reutilitzar medicaments, com en el cas més recent durant la pandèmia de la COVID-19 . [5] [6] [7].

Tot i que les relacions entre els fàrmacs candidats amb les RAM s'han estudiat extensivament, l'associació de les RAM amb dianes terapèutica, es a dir les proteïnes, es quelcom que està menys desenvolupat. De fet hi ha molts pocs recursos disponibles comptant entre elles base de dades ADRCS-Target [6], un estudi recent sobre RAM generades a partir d'assaigs clínics i informes posteriors a la comercialització [6] i un

treball realitzat de Kuhn i col·legues [9]. La manca doncs d'un mètode centrat en proteïnes per avaluar la toxicitat dels fàrmacs és una de les principals qüestions i fites d'aquesta tesis. Hi ha tot un seguit de bases dades que tenen informació sobre la relació entre fàrmacs i RAM. Per altre banda hi han tot un seguit de recursos que classifiquen informació sobre proteïnes i els fàrmacs associats a elles. Per tant, indirectament i en teoria, si el fàrmac X produeix el RAM Y i el fàrmac X interacciona amb la proteïna Z, aleshores la proteïna Z està vinculada al RAM Y. Aquesta simple afirmació, però, és incorrecta. Com s'ha demostrat, la majoria de fàrmacs uneixen a grups de proteïnes farmacològicament similars, com ara membres de la mateixa família de proteïnes [9].

Tot i que és probable que només una ó poques proteïnes d'aquesta família sigui responsable de la RMA, una relació directa entre proteïna i RMA, com la que utilitza en aquest model simplificat, vincularia cada proteïna a totes les possibles ADR del mateix fàrmac, donant lloc a falsos positius. Per evitar-ho, la relació s'ha d'avaluar estadísticament, i Kuhn et al. Varem proposar un mètode per identificar associacions estadísticament significatius entre RMA i proteïnes utilitzant fàrmacs com a factors de connexió [9]. Utilitzant diferents recursos i aplicant mètodes estadístics de validació, vaig crear la base de dades T-ARDIS (Target-Adverse Reaction Database Integrated Search) [11]. T-ARDIS proporciona un enllaç directe entre proteïnes i RMA que han estat validades estadísticament.

La següent pregunta que vaig abordar a la meua tesis va ser si la informació continguda a T-ARDIS es podia utilitzar per predir associacions entre RMA i proteïnes. La resposta va ser el desenvolupament de DocTOR (Direct foreCast Target On Reaction – (Galletti et. Al – [11])). DocTOR està basat en intel·ligència artificial i prediu si la modulació de una proteïna pot donar lloc a RMA. La informació utilitzada per DocTOR utilitzen dades derivades de l'estudi de xarxes d'interacció entre proteïnes de manera que més enllà de utilitzar elements intrínsecs també utilitza elements sistèmics. La precisió dels prediccions obtingudes per DocTOR justifica el seu ús per identificar dianes de proteïnes problemàtiques a nivell RMA individual, així com en un grup de RAM agregades en classes d'òrgans del sistema cos humà.

Finalment, el desenvolupament de DocTOR va conduir naturalment a la part final de la meua tesi que tractava d'entendre les bases moleculars de les relacions entre RAMs utilitzant informació de proteïnes dianes subjacents a aquestes. L'anàlisi SONG (Side effect On Network Graph) va permetre l'estudi de les relacions entre RMAs condensant l'ampli ventall de tipus de RAM en una nova xarxa anomenada "*Adverse Reactome*". Com el seu propi nom podria indicar, el *Adverse Reactome* es una xarxa de RMA connectada per les proteïnes responsables d'aquest RMAs. L'estudi d'aquest network mitjançant estudis de clusterització ha permet extreure conjunts d'associacions rellevant entre RMAs i donar pistes per aclarir el possible paper de les associacions de RMAs i les bases moleculars de l'aparició de RMAs. A més a més, ha permet extrapolar les funcions enriquides d'aquest clústers situant-los en el context global cel·lular.

TABLE OF CONTENTS

ACKNOWLEDGMENTS.....	I
ABSTRACT	VII
RESUM.....	X
TABLE OF CONTENTS.....	XIV
LIST OF FIGURES	XVIII
LIST OF TABLES.....	XX
1 - INTRODUCTION	- 1 -
1.1 – Incipit.....	- 2 -
1.2 - The drug discovery process	- 3 -
1.2.1 - Adverse event definition.....	- 6 -
1.3 – Network Biology.....	- 8 -
1.3.1 - Definition of network	- 9 -
1.3.2 - Types of networks.....	- 9 -
1.3.3 - Network-based metrics	- 11 -
1.4 - Drug-related databases used in this thesis	- 20 -
1.4.1 - Adverse reaction terminology database	- 20 -
1.4.2 - Drug – Adverse Event Databases.....	- 22 -
1.4.3 - Drug – target databases.....	- 25 -
1.4.4 - Proteins and Functional annotation databases	- 27 -
1.5 - Statistical Methods.....	- 30 -
1.5.1 - Likelihood Ratio Test (LRT) Methodology	- 30 -
1.5.2 - Fisher’s exact test methodology for ADR-target pairs	- 33 -
1.6 - Machine Learning Theory	- 35 -
1.6.1 - Support Vector Machine.....	- 36 -
1.6.2 - Tree-based methods	- 42 -
1.6.3 - Neural Networks.....	- 45 -
1.6.4 - Evaluation metrics of binary classification models	- 47 -
1.7 - Summary.....	- 51 -
2 - OBJECTIVES.....	- 53 -
3 - IDENTIFYING ADVERSE REACTION - TARGET ASSOCIATIONS MINING DRUG-TARGET AND DRUG-ADR RESOURCES: T-ARDIS.....	- 57 -
3.1 - Abstract	- 59 -
3.2 - Mining and curation of Databases cleaning procedure	- 62 -

3.2.1 - FAERS and MEDEFECT entries standardization	- 62 -
3.2.2 - Self reporting Drug-ADRs validated associations.....	- 66 -
3.2.3 - SIDER and OFFSIDES databases.....	- 66 -
3.2.4 - Advanced Filtering procedures	- 66 -
3.2.5 - Drug – Target databases	- 69 -
3.2.6 - Combining different databases increases the coverage of associations	- 70 -
3.3 - Statistically validated associations ADR-proteins	- 73 -
3.4 - Examples of uncovered associations and T-ARDIS benchmarking	- 74 -
3.4.1 - Examples of uncovered associations	- 74 -
3.4.2 - T-ARDIS benchmark.....	- 75 -
3.5 - Accessing and querying T-ARDIS	- 78 -
3.6 - Summary.....	- 80 -
4 - PREDICTING TARGET-LIABILITIES USING NETWORK-BASED ANALYSES: DocTOR	- 82 -
.....
4.1 - Abstract	- 84 -
4.2 - Data Extrapolation and Features computation	- 88 -
4.2.1 - ADRs considered for model construction.....	- 88 -
4.2.2 - Features considered for prediction.	- 91 -
4.3 - Machine Learning implementation	- 97 -
4.3.1 - Positive and negative sets definitions	- 97 -
4.3.2 - The predictive performance of individual features in the Self-reporting and curated datasets	- 99 -
4.3.3 - Features vectorization and model construction	- 103 -
4.4 - Performance score implementation	- 108 -
4.5 - Single ML methods CV results	- 109 -
4.6 - Meta predictor implementation and details	- 111 -
4.7 - Single predictor vs Meta-Predictor	- 113 -
4.7.1 - Predicting at SOC level	- 114 -
4.8 - Discussion.....	- 116 -
4.8.1 - Classifiers performances	- 116 -
4.8.2 - Self-reporting vs curated Dataset results	- 118 -
4.8.3 - The DocTOR utility	- 120 -
4.9 - Summary.....	- 121 -
5 - LINKING AND IDENTIFYING THE MOLECULAR BASES OF ADRS THROUGH SHARED TARGETS: SONG	- 122 -

5.1 - Abstract	- 124 -
5.2 - The SONG Network	- 125 -
5.2.1 - Clustering application	- 127 -
5.3 - Functional data enrichment.....	- 130 -
5.3.1 - g:profiler	- 130 -
5.4 - Uncovered Associations and examples	- 132 -
5.4.1 - G-coupled serotonin receptor signaling pathway disruption causes multi-organ failure	- 132 -
5.4.2 - Perturbation of Smooth muscle Adaptation and NADPH binding inficiate multi-level biological functions	- 135 -
5.4.3 - Cyclooxygenase inhibition presents a multi-system impact	- 137 -
5.4.4 - The Cluster-1 Analysis	- 139 -
5.5 - Cluster's related drugs exploration.....	- 146 -
5.6 - Summary.....	- 150 -
6 - GENERAL DISCUSSION.....	- 152 -
6.1 - The T-ARDIS database: " <i>Allons-y</i> " towards the identification of ADR-Target relationships	- 156 -
6.2 - The DocTOR approach: Precise predictions from blue black-box methods....	- 157 -
6.3 - SONG: Echoes from the ADRs' choirs	- 160 -
6.4 - Future perspectives and implication in the field of Protein-ADRs relationships	- 162 -
7 - CONCLUSIONS.....	- 165 -
8 - BIBLIOGRAPHY	- 171 -
9 - PAPERS ANNEX.....	- 181 -

LIST OF FIGURES

Figure 1. Drug development process	- 5 -
Figure 2. Side effect (SE), Adverse drug event (ADE), Adverse drug reaction (ADR) definition.	- 7 -
Figure 3. Graphical representation of a gene regulatory network vs. a protein-protein interaction network.....	- 10 -
Figure 4. A network in which the shortest path between the dark circled nodes has a length of four.....	- 11 -
Figure 5. Basic concepts of network centralities.	- 14 -
Figure 6. A sample PPIN is used to demonstrate the concept of disease modules. .-	15 -
Figure 7. Representation of the clustering coefficient and degree.	- 16 -
Figure 8. The MedDRA 5-level hierarchy demonstrated by using 'common cold' as an example (adapted from [[meddra]] - Figure 1).....	- 21 -
Figure 9. Number of adverse event reports received by FDA for drugs and therapeutic biologic products on 03/2022..	- 23 -
Figure 10. Schematic representation of Fisher exact test.	- 34 -
Figure 11. Representation of a linear SVM..	- 37 -
Figure 12. Representation of the application of kernel function	- 41 -
Figure 13. Schematic representation of a binary tree.....	- 43 -
Figure 14. Schematic representation of a perceptron.	- 46 -
Figure 15. Workflow followed to combine and derive statistical associations between proteins and ADR.....	- 60 -
Figure 16. Upset plot showing the overlap between the different databases compiling drug-ADR associations.....	- 71 -
Figure 17. Bubble plots showing the number of drugs per protein (X axis) vs number of statistically significant ADR per protein (Y axis).....	- 77 -
Figure 18. Snapshot of the result page example upon querying by drug "Aspirin". .-	78 -
Figure 19. Schematic depiction of feature extraction, training and testing procedures. .-	87 -
Figure 20. List of selected ADR by System Organ Class.	- 89 -
Figure 21. Network Feature extrapolation.....	- 92 -

Figure 22. DIAMOnD Distribution for the Negative sets in the case of T-ARDIS self-reporting (A) and T-ARDIS controlled (B) datasets.	- 98 -
Figure 23. Distribution of negative node shortest path.	- 99 -
Figure 24. Distribution plots of 8 different input variables used by classifiers.	- 102 -
Figure 25. Box- and violin plots of the cross-validation AUC results for the three different classifiers.....	- 110 -
Figure 26. Box- and violin plots for accuracy (ACC), precision (PREC), recall (REC), Receiver Operating Area Under Curve (ROC AUC) and Matthew Correlation Coefficient (MCC).....	- 113 -
Figure 27. Evaluation of adverse reaction-protein association predictions of the different classifiers at SOCs level.	- 115 -
Figure 28. Box- and violin plots for accuracy (ACC), precision (PREC), recall (REC), Receiver Operating Area Under Curve (ROC AUC) and MCC for the curated dataset. ...	- 118 -
Figure 29. Heatmap of predictions at SOCs for curated dataset.	- 119 -
Figure 30. Representation of the Adverse Reactome.....	- 126 -
Figure 31. Results of the clustering procedure on the Adverse Reactome.....	- 129 -
Figure 32. Example of g:GOSt method output.....	- 131 -
Figure 33. Cluster 2 resulting network.....	- 133 -
Figure 34. Cluster 4 resulting network.....	- 135 -
Figure 35. g:profiler results for cluster 4 analysis.....	- 136 -
Figure 36. Cluster 12 resulting network.....	- 137 -
Figure 37. g:profiler results for cluster 12 analysis.....	- 138 -
Figure 38. Significant molecular functions extracted from the functional enrichment procedure of the entire Cluster1's related proteins.....	- 139 -
Figure 39. Results of the clustering procedure on Cluster1.	- 140 -
Figure 40. Cluster 1-3 resulting network.....	- 141 -
Figure 41. g:profiler results for Cluster1-3 analysis.....	- 142 -
Figure 42. g:profiler results for Cluster1-8 analysis.....	- 144 -
Figure 43. Definition of "outra" and "intra" Tanimoto Scores.	- 147 -
Figure 44. Distribution of "intra" ed "outra" Tanimoto scores in the different cluster...-	- 148 -

LIST OF TABLES

Table 1. Representation of Drug - ADRs relationships extracted from self-reporting databases.	- 31 -
Table 2. Single Drug-ADR data extracted from table 1	- 31 -
Table 3. Representation of a confusion matrix or contingency table.	- 48 -
Table 4. List of retrieved LAERS/FAERS files.....	- 63 -
Table 5. Comparison of different datasets and T-ARDIS.....	- 76 -
Table 6. Vector representation of GO terms.....	- 95 -
Table 7. List of publications for the Cluster2 ADRs.....	- 134 -
Table 8. List of publications for Cluster1-3 ADRs.....	- 143 -

1 - INTRODUCTION

1.1 – Incipit

In this first chapter, I will present all of the fundamental concepts required to comprehend the underlying logic of the developed methods, as well as an extensive background and literature review necessary to comprehend all of the implications that a simple phrase like "*I took an Aspirin and now I have stomach ache*" may have.

First, I will outline the entire drug discovery process, highlighting the bottlenecks and difficulties of this lengthy and costly process. Then I'll focus on network biology, which will provide us with the foundation to understand that the proteins involved in our research are not single entities, but rather an intricate intertwined system. Following that, I will expose all of the databases used as the foundation of my work in this thesis, beginning with the drug-ADRs and drug-target sources. The final sections of this chapter will be devoted to the statistical theory supporting my methods, ranging from clustering and statistical validation approaches to the machine learning theory of the predictors used in this study.

1.2 - The drug discovery process

The process of bringing a new drug to market is complex and time-consuming, costing pharmaceutical companies an average of \$2.6 billion and ten years of R&D. [12]. This procedure is divided into stages, each with its own set of challenges, deadlines, and costs (figure 1). This section will provide a brief overview of the drug discovery process, beginning with the "target discovery" phase, which involves in-vitro research to identify characteristic molecules in specific diseases, such as nucleic acid sequences or proteins that regulate gene expression or intracellular signaling. As obvious as it may appear, not every condition-related protein can be chosen as a drug target, and extensive research must be conducted before deciding on which protein to focus to ensure that the chosen one is "druggable," or capable of being controlled by an external chemical. [13] (Figure 1 - point A).

After identifying a suitable target, the next step is to develop a compound that can interact with the chosen molecule. Conducting careful and precise target validation experiments is critical for the success of drug development during this stage, which is known as "Lead compound identification".[12] (Figure 1 - point B). This phase entails screening experiments to identify naturally occurring molecules that could be repurposed as drugs, as well as the development of synthetic compounds that can be specifically designed to target the selected molecule while not interfering with other cellular processes.

The next stage, or "Lead Optimization", involves preliminary safety tests performed in cell culture to test the drug's mechanism of action [14]. The pharmacokinetics and pharmacodynamics of the drug — how it is metabolized and how it affects various bodily functions — are also investigated at this step (figure 1 - point C). Following the identification of the candidate drug and preliminary testing of its mechanism, the latter's safety and efficacy must be improved while dealing with the major issue of off-target binding. Off-target binding is one of the most serious problems in drug development. It refers to the effects that can occur when a candidate drug interacts with

molecules other than those for which the drug was designed to bind. This can be avoided by computationally redesigning the candidate drugs so that they do not interact with molecules other than the target, but at the expense of increased research time. [15].

While investigating the event of off-target binding, the optimal dosage and administration strategy (oral, injectable) are also explored in this phase using two- and three-dimensional cell culture platforms and later integrated with preliminary in vivo testing to determine if the drug is safe for human trials and performs as expected [14] (Figure 1 - point D). To ensure the drug's potential success, preclinical trials must be as accurate as possible. At this point, companies have already spent an average of \$500 million on R&D, and a drug failure will cause significant economic harm. As a result, more precise toxicology research is conducted using animal models that mimic human conditions, such as knockouts or genetically engineered mice. [16] (Figure 1 - point E).

On positive pre-clinical results, before human testing can begin, an Investigational New Drug (IND) application must be submitted to the national medicine agency. This document typically contains critical information such as toxicity data, manufacturing process information, or clinical trial protocols that are being developed for the intended human trials.

Clinical trials may begin following the acceptance of the IND, [17] starting with the test of the new drug on 100 or fewer healthy patients to determine the medication's relative safety. Simultaneously, various carcinogenicity tests are carried out on Tg rasH2 mice [16]. In particular, this animal model is especially helpful in reducing the time required for carcinogenicity testing, cutting it from two years to six months. Following positive phase I results, the number of patients is increased to 100-500, and the effective efficacy of the drug is investigated.

Phase II is designed to assess a drug's efficacy on the illness together with the appropriate dosage and frequency of administration. The possibility of serious side effects is being closely monitored at this stage, as well as in the next phase, where the number of patients will be increased to 1,000-5,000 in order to collect statistically significant results at population level. Only about 12% of candidate drugs make it

through this stage, which is critical for determining the overall safety and efficacy of the new compound. Following the successful completion of clinical studies, a New Drug Application (NDA) is submitted to government agencies for review and possible approval. The purpose of this document is to demonstrate the drug's safety and efficacy based on clinical trial results. (Figure 1 - point F)

Once the NDA is approved, the novel drug is made available to patients, but it is still monitored in the general population for any side effects (Figure 1 - point G). This process is known as Pharmacovigilance. [17]. Unlike all the information acquired during clinical trials, pharmacovigilance data is obtained from patients and healthcare providers, as well as other sources such as medical literature and case studies. Pharmacovigilance is critical in monitoring drug efficacy and potential complications in the population, as seen in the well-known case of thalidomide. [18]

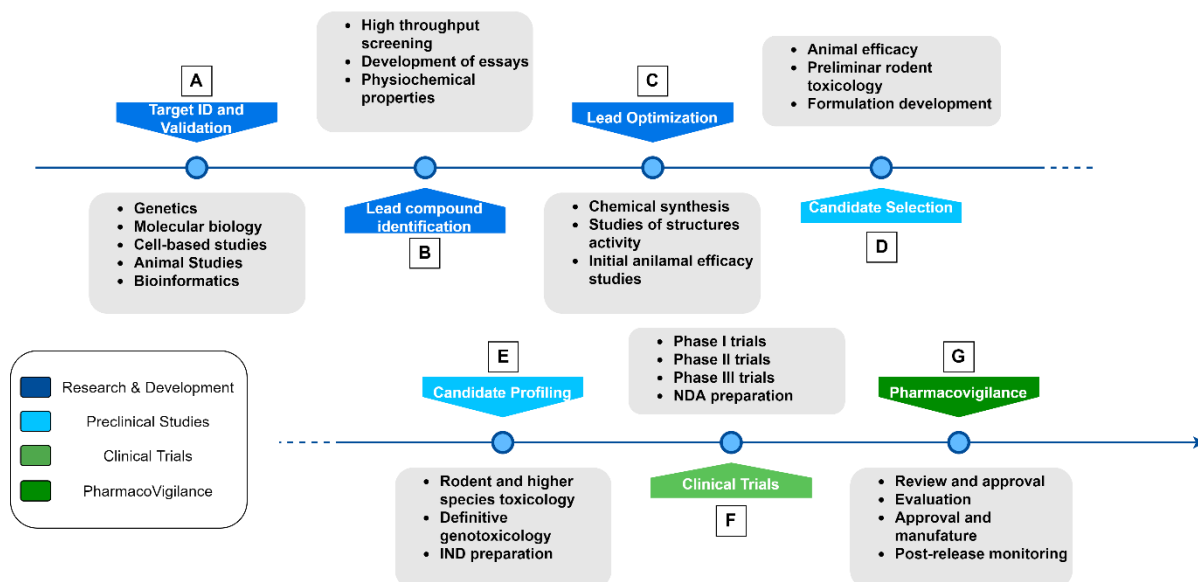


Figure 1. Drug development process - different steps of the whole drug discovery process, from identification of target to post release monitoring. The whole drug discovery process can be divided in three macro-steps, the first relates to the Research & Development part (dark blue, *points A-B-C*), the second consist in the pre-clinical trials (light blue, *points D-E*), the third step consist in the human clinical trials (light green, *point F*) and finally the post-marketing surveillance (dark green, *point G*).

1.2.1 - Adverse event definition

Adverse drug events (ADE), *side effects (SE)* and *adverse drug reactions (ADR)* are not the same thing [19]. In fact, these three words are used interchangeably, despite the fact that they have completely distinct connotations (figure 2). An *adverse drug event (ADE)* is defined as "a harm caused by the use of a drug." The term ADE, according to this definition, comprises both harm produced by the drug (adverse drug reactions-ADR and overdoses) and harm caused by the drug's use (dose reductions and drug therapy termination).

An *adverse drug reaction (ADR)* is defined as a "noxious and unanticipated response to a drug that occurs at therapeutic levels, diagnosis, or therapy, or for the alteration of physiologic function". In other words, an *adverse drug reaction* can be defined as "harm induced directly by a drug at regular doses and during normal use", indicating a causal relationship between drug and an *adverse drug reaction*.

Finally, a *side effect* is an unwanted consequence that occurs independently of the dose when a medicine is taken. Unlike *adverse drug events* or *adverse drug reactions*, *side effects* are usually anticipated by the physician, and the patient is informed of the potential side effects while on therapy. Some medications are even employed because of their negative effects, such as *Mirtazapine*, which is used in anorexic individuals since it has the potential to produce weight gain. In the following sections and discussion, we will always refer to *adverse drug reactions (ADRs)*.

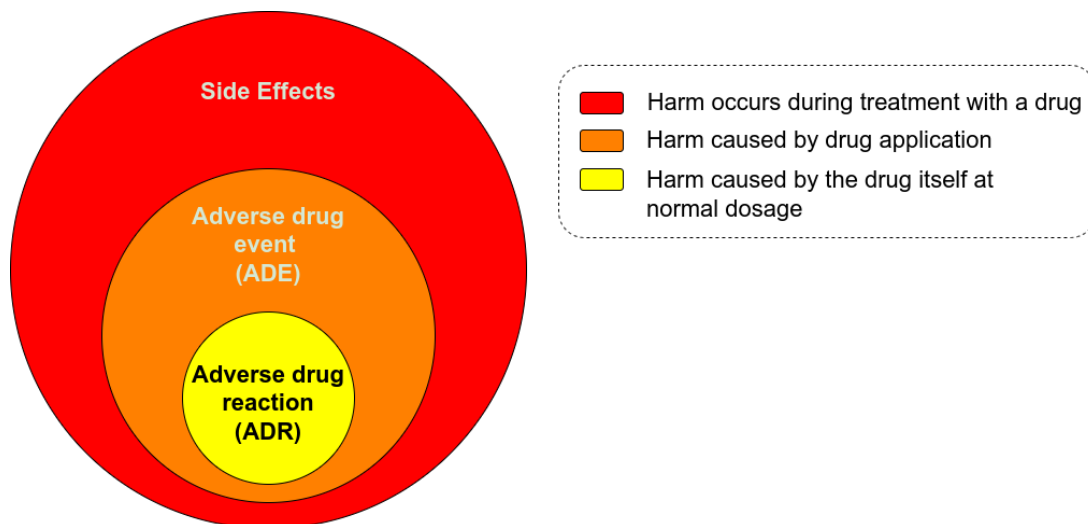


Figure 2. Side effect (SE), Adverse drug event (ADE), Adverse drug reaction (ADR) definition. Side Effects include all harmful events occurring during treatment with a drug without the necessity of a causal link between the drug and the reaction. If the use of medication is causal for the reaction, the condition is called an adverse drug event. A sub form of adverse drug events are adverse drug reactions that are triggered by the drug itself despite its appropriate dosage. (Adapted from the work of [19] - figure 1).

1.3 – Network Biology

A truly integrated framework investigating the interactions between all bio-molecules is essential to completely comprehend the functioning of the human organism. Since cellular functions are so intertwined, network analysis is well suited to exploring their molecular mechanisms. When applied to biochemical processes, the network's nodes may represent proteins, genes, or even illnesses, while the network's edges reflect the interactions between these biological entities.

There are several forms of network representations used to investigate human biology:

- (i) gene regulatory networks: in which nodes are transcription factors and genes, while edges represent regulatory connections (Figure 3 A)
- (ii) protein interaction networks: where nodes represent proteins and edges represent physical interactions (Figure 3 B);
- (iii) Metabolic networks: networks in which nodes represent metabolites and proteins and edges are metabolic activities and finally
- (iv) Disease networks: networks in which nodes represent illnesses and edges reflect different kinds of interactions such as genetic variants.

The study of network biology tries to precisely depict biological networks and analyze them in order to understand the behavior of a biological system. In the sections that follow, I will go through the features of networks and how they might help us better understand various biological systems.

1.3.1 - Definition of network

A network or graph (G) can be described as a pair $G = (V, E)$, where V is a set of nodes (or vertices) and E is a set of paired nodes, whose members are referred to as edges (or links) (figure 3). A network may also be defined as a structure that has a set of components, some of which are connected. The network's elements are known as nodes, and the connections between them are known as edges. As mentioned, in biological systems, nodes might represent proteins, genes, or even illnesses, while edges indicate the connections between these biological entities.

1.3.2 - Types of networks

Networks can have different classifications depending on the directionality of the relationship represented by the edges. When interactions in a network have a definite direction that travels from a source to a destination the network is said to be *directed* and the edges in this case are represented by arrows (Figure 3 A). In contrast, a network is *undirected* when the interactions do not have a definite direction. In this case edges are represented by lines.

Protein – protein interaction networks, for example, are usually *undirected* since their edges indicate relationships between proteins, which may not always follow a certain order (Figure 3 B). Metabolic networks, on the other hand, are directed since the edges reflect metabolic processes that begin with substrates and terminate with products. Finally, gene regulatory networks are also directed since they depict how the expression of one gene influences the expression of another.

Networks can also be weighted or unweighted based on whether or not additional information is assigned to the edges. The edges in an unweighted network are present only if a threshold of evidence for the association is met while weighted networks present edges in which the weight indicates a specific aspect of the association. Gene co-expression networks, for example, are networks in which the nodes are genes that are linked by their expression relationship. These networks can be weighted, displaying

the relationship between the expression of two genes in the edge weight; or they can be unweighted, displaying just the edges that meet a specified association cutoff point.

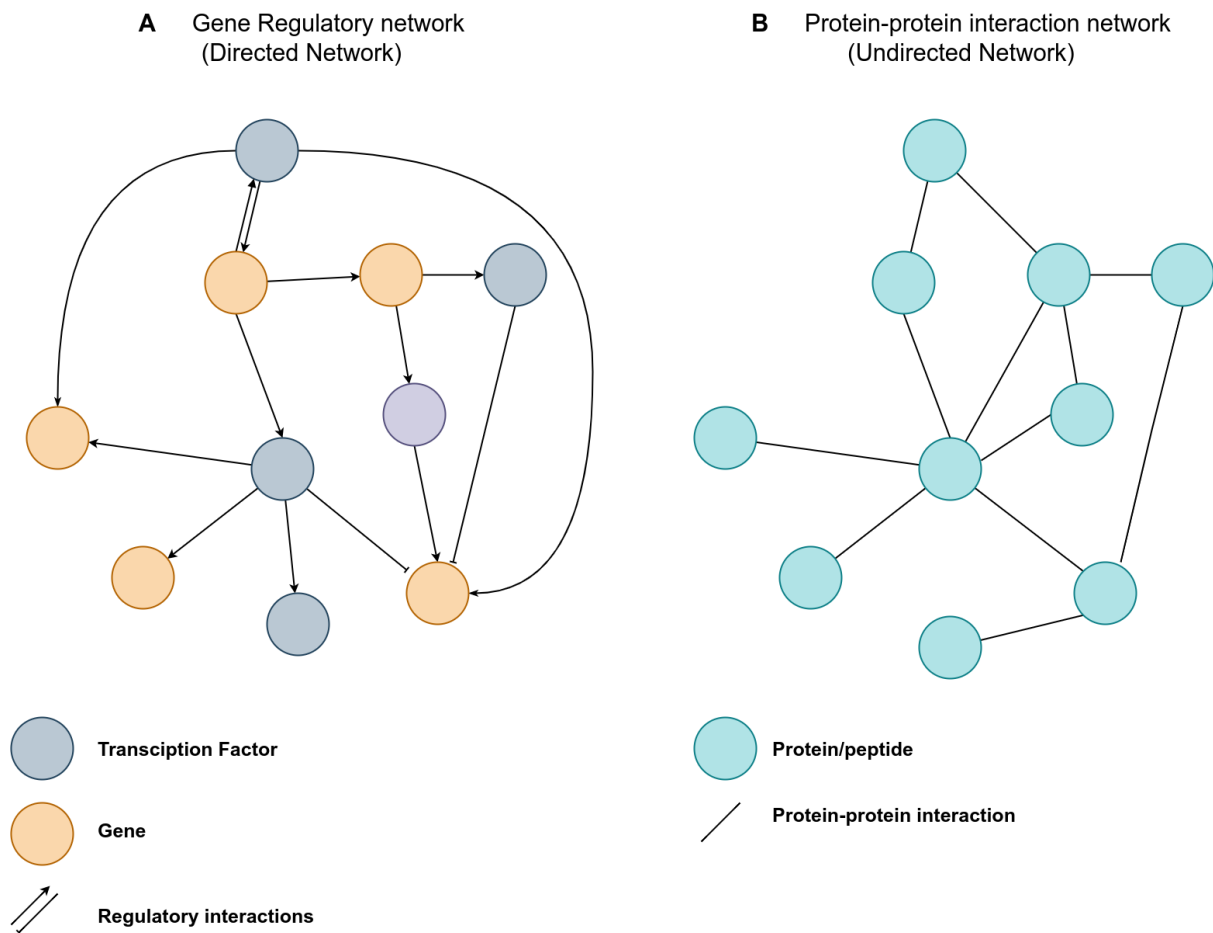


Figure 3. Graphical representation of a gene regulatory network vs. a protein-protein interaction network. (A) In a gene regulatory network, nodes represent genes or proteins and lines between them regulatory interactions. Regulatory networks can be defined as directed networks (B) In a protein-protein interaction (PPI) network nodes always represent proteins and the connecting lines of physical protein-protein interactions. Since the represented interaction is bivalent, a PPI network can be defined as undirected.

1.3.3 - Network-based metrics

1.3.3.1 - Definition of network path

A network path is a connection between two nodes that follows a set number of edges. The number of edges involved in the path determines the path's length. A path (P) in an undirected graph can be defined mathematically as a sequence of nodes (v):

$$P = (v_1, \dots, v_n) \quad \text{Eq. 1}$$

We can also define the concept of shortest path and characteristic path length. The first is the path with the fewest edges connecting them (Figure 4), the second equal to the mean shortest path length among all network nodes. The latter can be computed as:

$$a = \sum_{s,t \in V} \frac{d(s,t)}{n(n-1)} \quad \text{Eq. 2}$$

Where V is the set of nodes in the entire network of size n, and d(s, t) is the shortest path between nodes s and t.

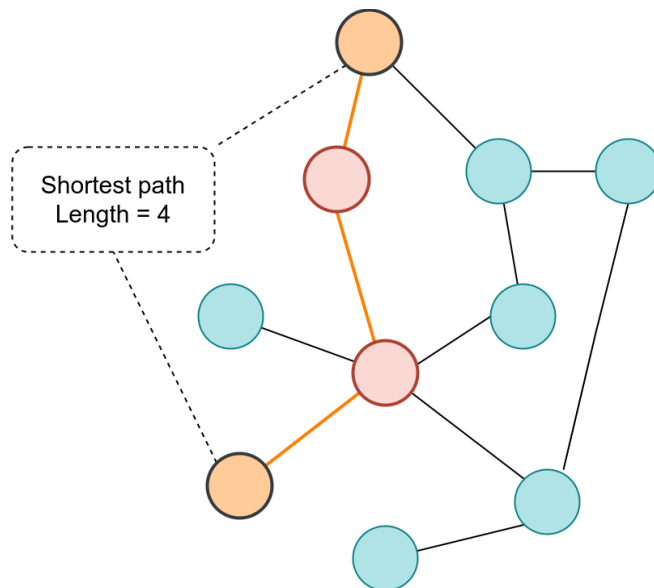


Figure 4. A network in which the shortest path between the dark circled nodes has a length of four.

The shortest path provides valuable information on the relationship between a specific node and the rest of the network, making this measurement critical in network research. One of the main applications of this measurement relates to identification of disease-associated proteins. As I'll explain in the next chapters, disease-associated proteins tend to cluster in topological proximity of the network forming the so-called disease modules [20]

1.3.3.2 - Centrality measurements

By attributing scores to nodes and edges, centrality measurements provide insight about their relevance. Centrality metrics are employed in systems biology to determine nodes that play critical roles in biological processes. There are several sorts of metrics that account for network centralities and provide varying degrees of priority to the highest scoring nodes such as Degree centrality, closeness and betweenness centrality. Nevertheless, different metrics of centrality tend to be directly proportional with one another, and it has also been demonstrated that hubs tend to have high centrality tending to be associated with highly conserved biological functions. [21]

Degree centrality (Figure 5 A, B) refers to the number of edges associated with a node. It is defined as

$$C_D(v) = deg(v) \quad \text{Eq .3}$$

Where $deg(v)$ is the degree of the node v . The degree centrality can be normalized by dividing the maximum possible degree in a graph by $n - 1$ where n is the number of nodes in the network under analysis.

The *Closeness centrality* (Figure 5 C) is computed by calculating the shortest-path distance between nodes. In particular, this measure relates how close a node is to the rest of the network's nodes. It can be described as follows:

$$C(u) = \frac{n-1}{\sum_{v=1}^{n-1} d(v,u)} \quad \text{Eq. 4}$$

Where the shortest-path distance between nodes v and u is defined $d(v, u)$, and the total number of nodes in the network is n . This measure is of particular importance since it represents the efficiency with which the network's nodes exchange information.

Finally, the *betweenness centrality* (Figure 5 D) shows the frequency with which the node appears in the network's collection of shortest paths. A node's betweenness centrality v can be evaluated as:

$$C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \quad \text{Eq. 5}$$

Where σ_{st} is the total number of shortest paths from node s to node t and $\sigma_{st}(v)$ is the number of these shortest paths that passes through v . In other words, it counts the number of times the node of interest appears among all pairs of nodes' shortest pathways. Betweenness centrality can be a powerful measure for predicting "bridge" or "link" nodes that connect various network modules.

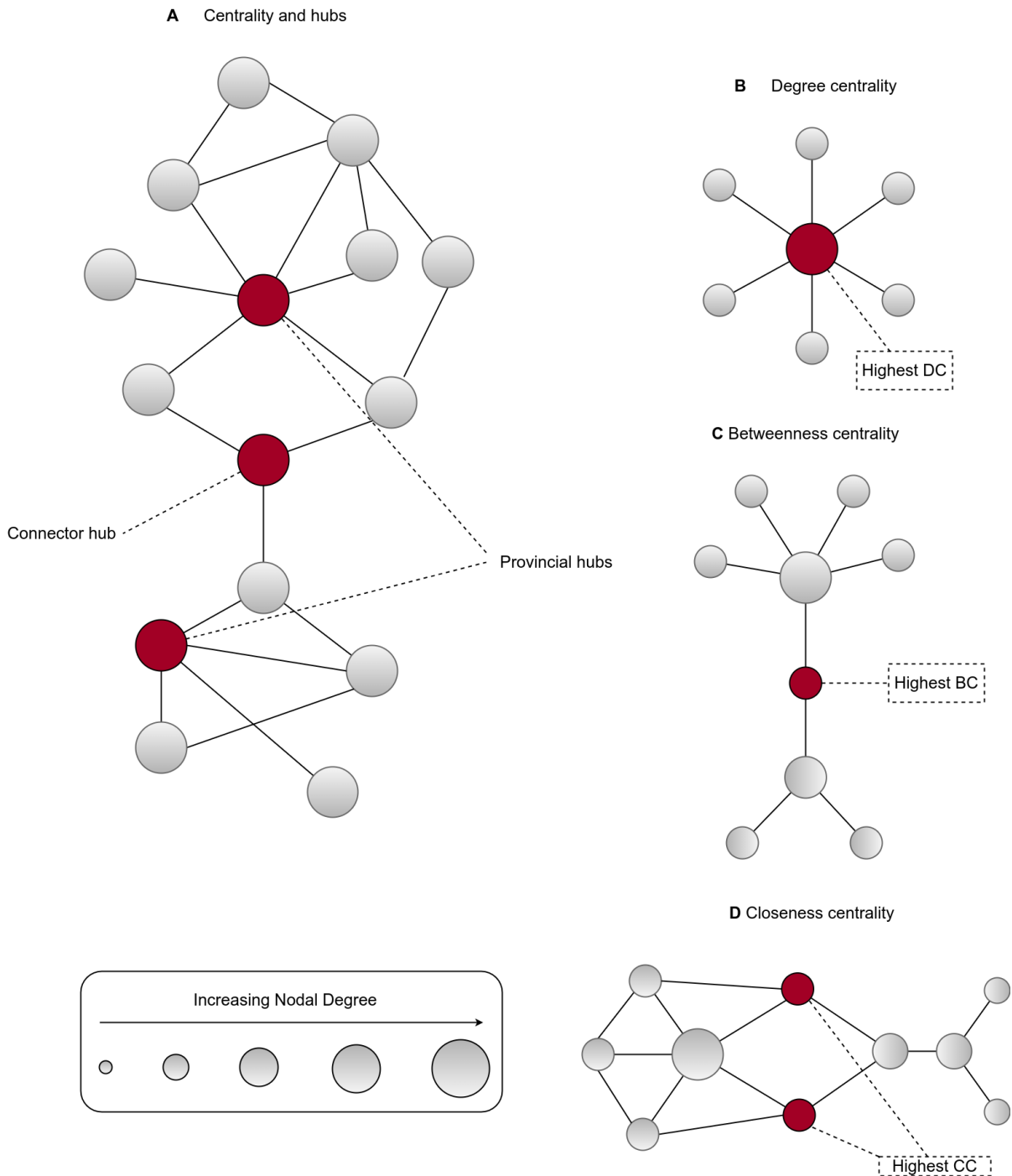


Figure 5. Basic concepts of network centralities. (A) Hubs (connector or provincial) are nodes that have a high nodal centrality and can be identified using various metrics. (B) The number of node neighbors is used to calculate degree centrality. (C) By determining the ratio of all shortest paths in the network that incorporate a given node, the betweenness centrality quantifies the node's role as a bridge between disparate clusters. (D) Closeness centrality measures how quickly a node in a linked graph can access all other nodes; the closer a node is to all other nodes, the more central it is.

1.3.3.3 - Network modules

Many forms of biological networks have nodes with similar roles or functions interacting with one another, forming so-called modules or communities. This concept has been expanded to include proteins which relates to diseases development. [22]. Within the interactome, three types of modules can be identified (Figure 6). (A) the functional module, a neighborhood of nodes with similar or related functions; (B) the disease module, a neighborhood of nodes that contribute to cellular functions whose disruption results in a specific disease; and (C) the topological module, a locally dense neighborhood that clustering algorithms can identify.

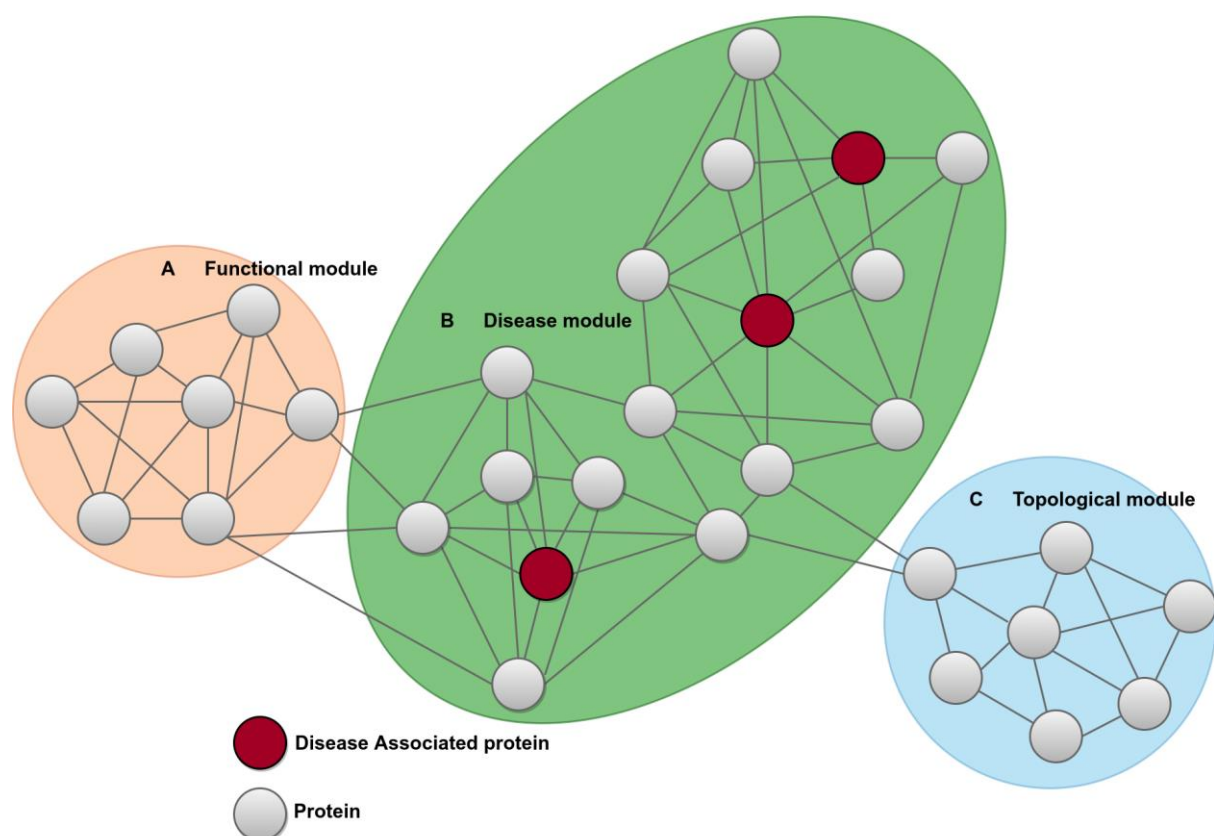


Figure 6. A sample PPIN is used to demonstrate the concept of disease modules. Proteins involved in similar biological processes form functional modules in one or more topological modules (A, C). A disease module (B) is a protein subnetwork enriched with disease-relevant proteins, such as known disease-associated proteins.

As a result, identifying modules in a network can be critical in acquiring a deeper understanding of module members' biological significance. There are a number of metrics and techniques that may be used to determine the degree of clustering of a node.

The *clustering coefficient* (figure 7), for example, reflects the likelihood that two nodes that are linked to each other are also directly connected between them (forming a triangle) (Figure 6). The proportion of feasible triangles in a node's neighborhood is measured by its local clustering coefficient, which may be computed as:

$$C_i(v) = \frac{2L_i}{k_i(k_i-1)} \quad \text{Eq. 6}$$

Where i is the node with degree k_i and L_i is the number of connections between node i 's neighbors.

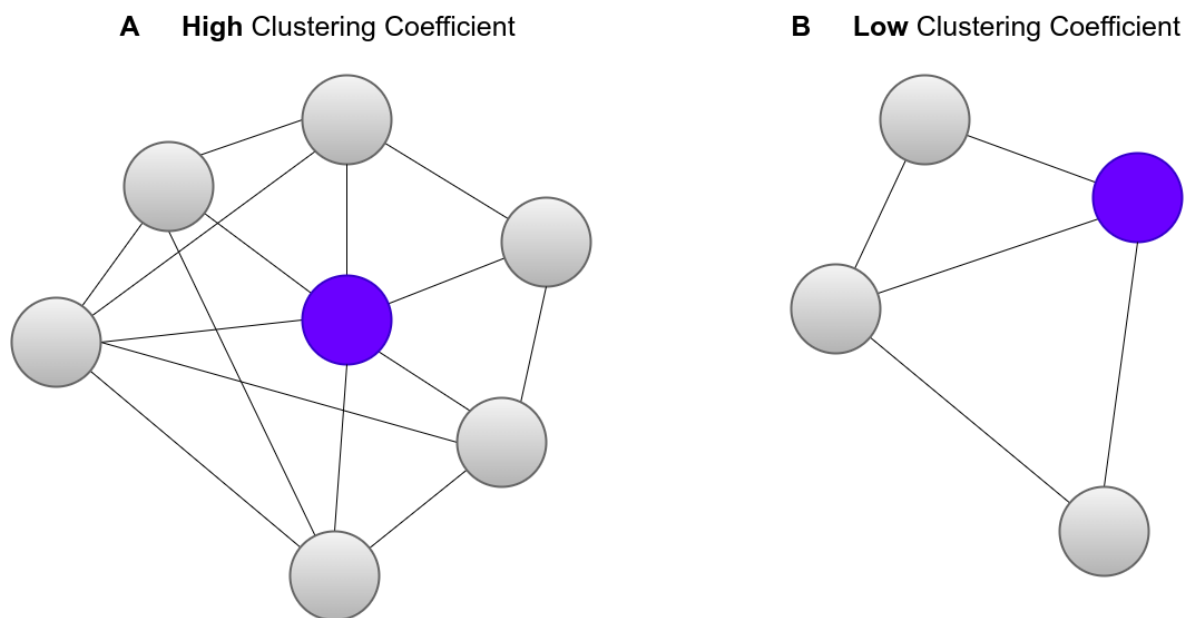


Figure 7. Representation of the clustering coefficient and degree. Networks have nodes with different values of degree and clustering coefficient. (A) Purple node presents both a high degree and clustering coefficients. (B) Purple node with low degree and clustering coefficient

There are also several approaches for automatically identifying modules in a network. These algorithms are often divided into two types. The first group comprises algorithms that make use of a priori information about nodes, such as connections that's also shared by network modules. These nodes are referred to as seeds. These techniques are based on the localization of nodes in the surroundings of the defined seeds, in other words topologically closer nodes. Methods that identify modules utilizing community structure detection algorithms fall into the second group. These approaches examine the network's topology and discover regions that have features attributable to modules, such as a high within-edge density of connections.

These module identification methods are frequently used to identify disease modules, which are groups of proteins related with the same disease. A disease module can also be defined as a cluster of nodes that contribute to cellular functions and whose disruption causes a specific disease (Figure 7 B). The identification of disease modules has become critical for achieving a comprehensive molecular understanding behind diseases [20] [23] [24]. Several approaches have been developed to this end in order to identify such modules. They are roughly divided into two categories: methods based on prior knowledge and ab-initio methods.

Prior-knowledge methods as its name indicates covers approaches that make use of pre-existing knowledge about disease-related genes (also known as seed genes). In a nutshell, these strategies seek proteins that are topologically close to those encoded by the seed genes.

This group is further subdivided into three categories: diffusion-based methods, community-finding methods, and network neighbor methods. Diffusion-based methods are based on the "message passing" theory, releasing signals (known as "random walkers") from the selected seed nodes to the rest of the nodes of the network. The nodes closest to the seeds are more often visited by the signals and so are enriched with a higher score. One example of this technology can be found in the GUILD software package [25]. The algorithm transmit a signal from the seeds to the rest of the network and rank each node based on how quickly the message reaches them while taking many network features into consideration.

The second group of methods, community-finding, is based on the identification of sub-networks formed by the selected seeds. If the sub-networks found are statistically significant, the algorithms proceed to rank each other node in the network trying to find new protein candidates that are topologically and functionally connected to the rest of the module of proteins and could be included to the sub-network in consideration. An example of these methods is the DIAMOnD algorithm [26]. This method utilizes an iterative search pattern to determine the relevance of protein interactions in the vicinity of the selected seeds (i.e., if the number of interactions is higher than a random expectation). Finally, the last type of methods, also called linkage method, assumes that proteins that directly interact with other proteins linked to a certain disease are more likely to be linked to the same disease themselves.

Ab-initio methods include module identification algorithms that do not rely on prior knowledge, such as previously identified disease-associated proteins. These methods rely on community structure detection algorithms, such as algorithms based on the maximum clique enumeration problem, to find protein regions with a high within-edge density of connections. [27]. Identifying disease modules with high accuracy remains a challenge, despite the fact that this is a highly-active field of research.

As part of a community effort to develop in this research area, the Synapse platform hosted a DREAM competition in 2018 focused on the blind prediction of disease modules from various types of networks [28]. Different types of methods (diffusion state distance, kernel clustering, modularity optimization, random-walk-based, and local methodologies) were among the top performers in this contest, indicating that no single approach was superior to the others. One of the top performers was the diffusion state distance approach (DSD), which is an enhanced measure of network closeness between pairs of nodes and demonstrates the importance of considering the entire topology of the network rather than just the local region. In a nutshell the DSD metric is used to define a pairwise distance matrix between all nodes, which is then used by a spectral clustering algorithm. Using standard graph techniques, dense bipartite subgraphs are identified in parallel and combined into a single set [29].

1.4 - Drug-related databases used in this thesis

In chapter 3, I will introduce the T-ARDIS database, a statistical validated compendium of drug's target – Adverse reaction association. The creation of this database relied on the exploitation of different pharmacovigilance resources. Pharmacovigilance, as previously explained, is the study of the effects of therapeutic products after they have been made accessible and marketed to the general public, with the objective of finding potentially unreported or under-reported adverse events. Only after a drug has passed phase III and is used by the general public in a non-clinical context, the most comprehensive profile of its side effects can be generated [30]. One of the major reporting bodies for such adverse reactions is the FAERS database (FDA Adverse Event Reporting System) [31] together with its Canadian sister database MEDEFECT [32] and the European EMA [33]. T-ARDIS will retrieve drug-ADRs information from such repositories together with more reliable resources such as SIDER [34] and OFFSIDES [35]. At the same time, databases that compile information on drug – protein associations will be mined to extract drugs affinities and targets. In particular, this category includes STITCH [36] and drug-target commons [37]. In the next subchapter I will explain in detail the databases used and the type of information acquired.

1.4.1 - Adverse reaction terminology database

The Medical Dictionary for Regulatory Activities (MedDRA) is a vocabulary that contains over 10,000 medical terms organized in a hierarchical structure. The MedDRA nomenclature is characterized by five layers from the most specific (LLT) to more abstract concept (SOC) (Figure 8) [38]. MedDRA is constantly updated as new medical concepts are introduced or modified. The version adopted in this thesis is the 25.0, released on 05/2022.

The highest-level layer of MedDRA, 'System Organ Class' (SOC), comprised 27 terms in this version, whereas the lowest level layer, 'Lowest Level Term' (LLT), contained almost

80,000 terms. The highest-level layer (SOC) contains the most generic concepts, while the terms become more specialized with each layer. The 'Preferred term' (PT) is the term typically used in all the drug-ADRs databases to label adverse reactions, mapping almost 24,000 conditions. The most specific term (LLT) may include synonyms or alternate spellings of the PT, as well as the PT itself. Every PT is primarily assigned to one SOC but may also be assigned to numerous additional SOCs on a secondary basis (e.g., the PT "Asthma" is identified under its primary SOC "Respiratory, thoracic, and mediastinal illnesses" (SOC), but also as a secondary SOC "Immune system disorders" (SOC)). Before using this data, a couple of mapping procedures are applied to avoid redundancy, such as using only PT as discerning elements for the ADRs and considering only primary SOCs associated with the single PT in analysis.

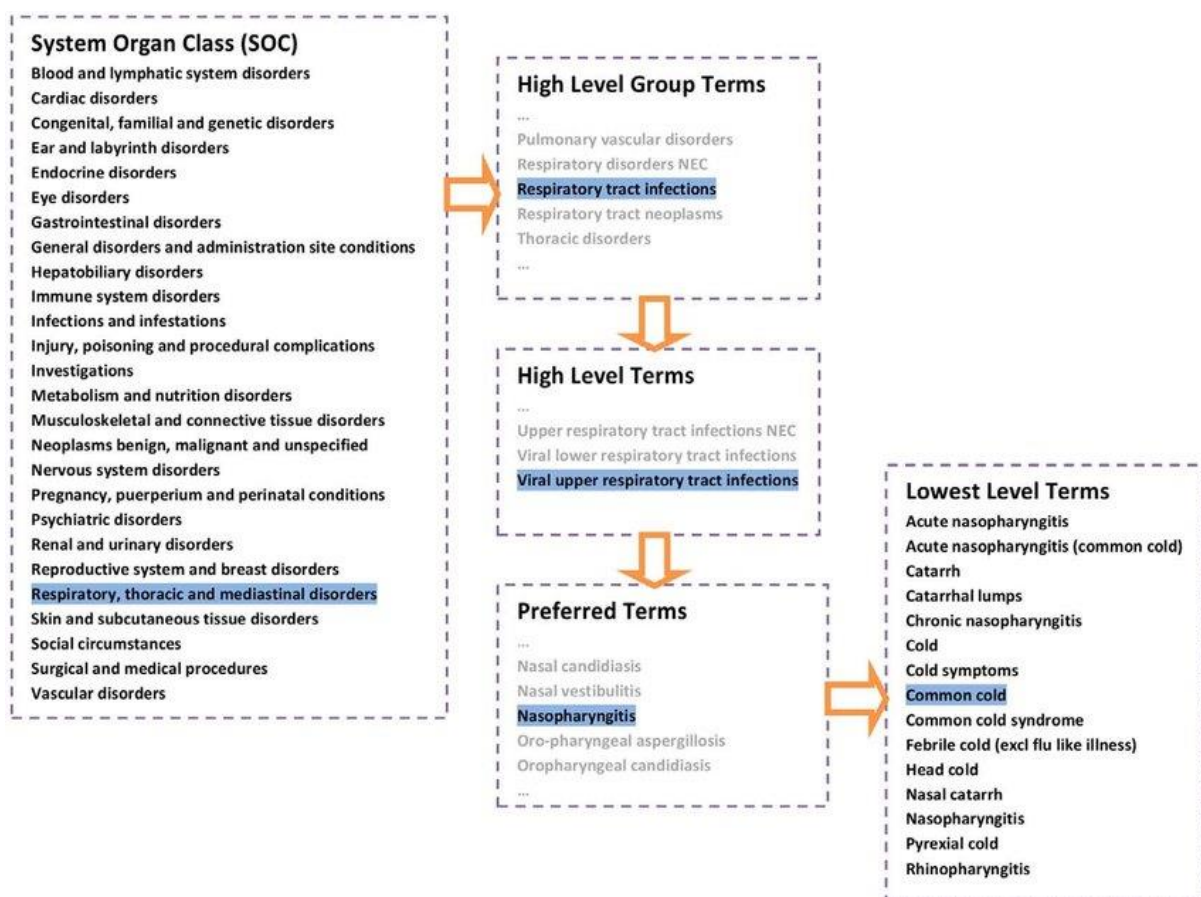


Figure 8. The MedDRA 5-level hierarchy demonstrated by using 'common cold' as an example (adapted from [38] - Figure 1).

1.4.2 - Drug – Adverse Event Databases

1.4.2.1 - FAERS

As a global reporting organization, the Spontaneous Reporting System is a tangle of nomenclature and protocols from many hospitals and countries around the world. The FDA is severely hampered by the requirement to standardize and sanitize adverse event data due to the variability of the data [31]. Since the majority of reports are filed in free-text format, text mining became a major issue in collecting and parsing this information. Hiring developers or purchasing products to perform text mining can significantly increase the cost of a project, especially when combined with the limited nature of data and the need to condense report information to no more than one-word responses across multiple variables such as sex, date, product, and so on. Furthermore, without a defined list of labels shared with all event reporters, drug names can be reported in a variety of formats, including generic, brand, abbreviation, and non-standard nouns. The openFDA API has addressed this issue by introducing "openfda" fields, which provide a standardized version of these field variables [39].

The openFDA makes use of a multifaceted URL, which may be studied more directly from the site. It is possible to get specific endpoint information such as Drug Product Labeling, Device Adverse Events, and the NDC Directory, with other relevant data. Unfortunately, only one drug at a time may be processed, making the use of this utility not suitable for my research.

T-ARDIS, on the other hand, relied on the direct interaction of FAERS quarterly data which the organization collected from 2004. The quarterly data files, which are available in ASCII or SGML formats, contain different information such as demographic and administrative data, drug bureaucratic information, adverse reactions information, patient outcome and finally information on the source of the reports. In total FAERS reports more than 67,000 labeling for drugs currently on the market (over-the-counter and prescription drugs in the United States, including biological therapeutics and generic drugs) and more than 24 million reports on negative side effects since 2003

(Figure 8 - shorturl.at/dnrsL). Direct access to this large wealth of information is difficult, necessitating several filtering procedures explained in the T-ARDIS chapter.

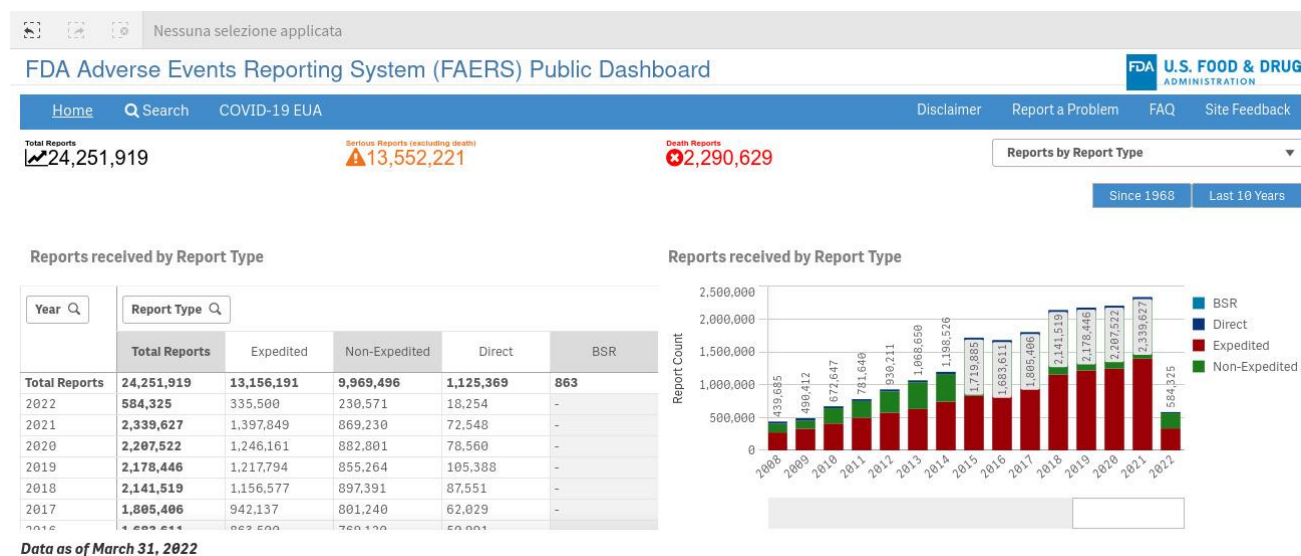


Figure 9. Number of adverse event reports received by FDA for drugs and therapeutic biologic products on 03/2022. This data includes the direct voluntary reports submitted through the MedWatch program by consumers and the mandatory reports from manufacturers. At the moment more than 24 million reports are present.

1.4.2.2 - MEDEFECT

The Canada Vigilance Program is a post-market surveillance program that gathers and evaluates reports of suspected adverse drug reactions (ADRs) to health products sold in Canada. This program allows Federal Regulators to track the safety profile of health products after they've been approved for sale, ensuring that the benefits continue to outweigh the risks.

Since 1965, the Canada Vigilance Program has been collecting complaints of suspected ADRs. Health consumers and healthcare professionals (HCPs), as well as market authorization holders (MAHs) who are required to submit reports under the Food and Drugs Regulations, voluntarily submit these reports to Health Canada [32]. The Canada Vigilance Program Online Database, MEDEFECT, contains approximately 225,000 suspected adverse reaction reports that have occurred in Canada since 1965, allowing health consumers, HCPs, and MAHs to see the sorts of adverse responses that have been

reported to Health Canada [32] The database is updated four times a year to incorporate new data which contains suspected adverse reactions to Canadian-marketed health items that occur in Canada or North America [32].

1.4.2.3 - SIDER

The SIDER database offers information on marketed drugs as well as adverse drug reactions that have been identified. The data is collected from public records and product information leaflet [34]. The sample for this research is being gathered from three files: "meddra all se.tsv.gz," "meddra all indications.tsv.gz," and "drug names.tsv." The names of side effects and indications are copied from the Medical Dictionary for Regulatory Activities (MedDRA) [38]. MedDRA refers to the clinical terminology and diagnoses that a physician will provide to a patient. To increase the chances of text matching potential, these terms are included in both the lower-level term and the preferred term. Natural language processing techniques were used to collect adverse drug responses and drug pairings from biomedical literature and package inserts for the SIDER database. SIDER has been tagged with PubChem and MedDRA identifiers so that drug side effects and indications may be tracked immediately. The current version (SIDER 4.1) was released on October 21, 2015, matching 139756 Adverse Reaction – Drugs pairs.

1.4.2.4 - OFFSIDES

OFFSIDES [35] is a manually curated drug adverse reactions database available at <http://tatonettilab.org/resources/nsides/>. The database contains 438,801 off-label effects (those not included on the FDA's official drug label) originating from 1332 pharmaceutical compounds, as well as 10,097 adverse reactions. On average, 69 "on-label" adverse events are listed on a drug label. Each medicine had an average of 329 high-confidence off-label adverse occurrences. For example, the SIDER database, which was compiled from medication package inserts, contains 48,577 drug-event

relationships for 620 medicines and 1092 adverse events, all of which are included in the data mining. OFFSIDES recovers 38.8% of SIDER associations (18,842 drug-event associations) from adverse event reports. As a result, OFFSIDES reports connections that differ from those found in clinical trials prior to drug approval.

1.4.3 - Drug – target databases

Working with drug-target relationships has proven to be difficult, despite the abundance of information available. This is primarily due to a lack of overlap among various sources. The data, in fact, is dispersed across multiple databases and repositories. This is directly correlated to the fact that drug-target relationships may be determined using a plethora of qualitative and quantitative criteria requiring different approaches. In vitro experiments have traditionally been used to identify drug-target associations. The two main types of experimental procedures usually comprise genetic interaction methods, which are based on monitoring gene expression after drug application, and direct biochemical and biophysical methods, which are based on determining the binding affinity between the target and the drug.

Other techniques may involve different measurements used to determine the intensity of the interaction between a medication and its target. Between them, the most applied are the inhibition constant (K_i), the dissociation constant (K_d), the half-maximal inhibitory concentration (IC_{50}), and half-maximal effective concentration (EC_{50}). Recently, also computational methods such as molecular docking-based methods, pharmacophore-based approaches, and machine learning/network-based methods have seen widespread application.

Given this huge number of methods and, consequently, of different repositories, the necessity of comprehensive databases arose. Among the most use are the Drugbank [40], Matador [41], Therapeutic target database [42], STITCH 5.0 [36] and Drug-target commons 2.0 [37] database. The research presented in Chapter 2 relied on Drug-target commons 2.0 and STITCH 5.0 that are explained in more detail in the following sections.

1.4.3.1 - Drug-target commons

Drug Target Commons (DTC) is a community-driven bioactivity data integration and standardized online platform for through mapping, reuse, and analysis of compound–target interaction profiles (<https://drugtargetcommons.fimm.fi/>) [37]. End users can utilize an application programmable interface (API), database dump, or tab-delimited text download options to search, upload, amend, annotate, and export expert-curated bioactivity data for further research. DTC version 2.0 offers updated clinical development information for the drugs and target gene–disease connections, as well as cancer-type indications for mutant protein targets, which are crucial for precision medicine. The gene–disease relationships contained in DTC are derived from DisGeNET and there are currently 1573 genes linked to 4123 disorders, with 331 514 references supporting the connections. Clinical data supports the cancer-type indications for 185 mutant protein targets gathered from Cancer Genome Interpreter (CGI). The major source of bioactivity data in DTC is presently ChEMBL, which is further confirmed by the DTC curation team and annotated using the μ BAO annotations [37]. Furthermore, around 60 000 completely annotated bioactivity values were taken directly from scientific publications. The annotation of 204 901 bioactivity data points among 4276 chemical substances and 1007 different protein targets may finally be found in DTC [37].

1.4.3.2 - STITCH

STITCH ('search tool for chemical interactions') combines data from metabolic pathways, crystal structures, binding assays, and drug–target correlations to create a comprehensive picture of chemical interactions [36]. Chemical relationships are assessed using extrapolated information from phenotypic outcomes, text mining, and chemical structure similarity. STITCH also helps to explore the network of chemical relationships, as well as associated binding proteins. The original data sources may be traced back to each potential relationship. The database, which shares protein space with STRING v10, contains around 9 600 000 proteins and 430 000 chemicals from 2031 eukaryotic and prokaryotic genomes [36].

Manually curated datasets such as DrugBank [40], GPCR-ligand database (GLIDA) [43], Matador [41], the Therapeutic Targets Database (TTD) [42], and the Comparative Toxicogenomics Database (CTD) [44], as well as several pathway databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) [45], NCI/Nature Pathway Interaction Database [46], Reactome [47], and BioCyc [48], provide a large portion of the known interactions stored in STITCH. The datasets of experimentally confirmed interactions, such as ChEMBL [49], PDSP Ki Database [50], and Protein Data Bank (PDB) [51], are also included in STITCH.

1.4.4 - Proteins and Functional annotation databases

2.4.4.1 - Gene Ontology

The rapid growth of genomic data has prompted the creation of tools to aid in the representation and processing of information on genes, their products, and their roles. The Gene Ontology (GO) is one of the most important of these tools. Within the scope of the umbrella project OBO (open biological ontologies), GO is being developed in combination with a number of biological databases such as FlyBase (Drosophila), the Saccharomyces Genome Database (SGD), and the Mouse Genome Database (MGD). [52]

The scope of GO is to provide a standardized vocabulary to describe cellular components, molecular functions, and biological processes. Actually, GO contains over 43 thousand GO terms associated with more than 7 million annotations and 1 million gene products distributed for over 5 thousand species [52]. Terms are stored in a hierarchical structure; in such a manner it is possible to distinguish if one term is more general than another or whether the entity defined by one term is a portion of the entity denoted by another allowing in deep understanding of functional synergy. GO is divided into three disjoint term hierarchies: (i) the cellular component, (ii) the molecular function, and (iii) the biological process ontologies.

The cellular component terms in GO are the counterpart of anatomy within the medical framework. It's designed to help biologists keep track of the physical structure that a gene or gene product is linked to. Both the extracellular environment of cells and the cells themselves are included in the GO vocabulary. Molecular Function Ontology describes the action characteristic of a gene product; terms like ice nucleation, binding, or protein stabilization are part of this hierarchy. Finally, biological processes ontology describes all those phenomena which are marked by changes that lead to a specific result, mediated by one or more gene products. Biological process terms tend to be quite specific (i.e., glycolysis) or very general (i.e., death). As one might expect, molecular function and biological process terms are strongly intertwined. [52]. The GO terms will be widely used during this research as they represent key information on the function and relationship of the analyzed proteins.

1.4.4.2 - The KEGG database

The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a 1995 knowledge-based method [45]. KEGG, in particular, employs graphical representation to help understand high-order systematic behavior of cells and organisms based on genomic and molecular system exploration. KEGG applications range on the different levels of human interactome and reactome with particular attention on pathways. Different methods are available for each aspect of human biology:

- KEGG GENES: group of gene categories for all fully sequenced genomes and some partial genomes, with up-to-date gene function annotations
- KEGG LIGAND: chemical building block system information for endogenous and exogenous chemicals.
- KEGG PATHWAYS: used to represent molecular relationships and reaction networks, such as genetic information processing, environmental information processing (signaling), cellular processes and also human diseases.

2.4.4.3 The Uniprot database

The Uniprot database [53] is the largest and most complete database for protein sequences. The UniProt Knowledgebase, or UniProtKB, is the result of the merger of two historical sequence databases: Swiss-Prot and TrEMBL [54]. The first comprises "curated" entries, which are protein sequences that have been tested experimentally; the second is generated automatically and contains proteins inferred from genomic data. Swiss-Prot currently has roughly 550,000 entries, while TrEMBL has around 195,000,000 (updated at 05/2022).

1.5 - Statistical Methods

As the drug – target databases can rely on multiple sources and methods to validate the relationship, in the case of drug – adverse reactions it is less clear and thus requires statistical methods to assess the significance of associations. Indeed, it is quite critical for public health to identify safety signals utilizing huge datasets like FAERS and MEDEFECT. One way to take advantage of this huge mole of data is to statistically validate the drug-ADR pairs found. There are several statistical methods available for this type of signal detection, including the reporting odds ratio (ROR), the proportional reporting ratio (PRR), the multi-gamma Poisson shrinker (MGPS), the Bayesian confidence propagation neural network (BCPNN), a Bayesian method based on a new information component (IC), the simplified Bayes (sB), or hierarchical models based on the Conway–Maxwell–Poisson distribution. Each one of them however is subject to different types of confounding factors.

1.5.1 - Likelihood Ratio Test (LRT) Methodology

Among the proposed techniques, the likelihood ratio test-based method (LRT) proved to be the most versatile in handling this type of data, managing to control Type-I error and false discovery rate (FDR). The LRT approach was firstly proposed by Huang et al. [55] and assumes that the number of reports for a drug-adverse event pair follows a Poisson distribution. This method was especially designed to find ADR signals for a single drug or drug signals for a specific ADR [56]. A likelihood ratio test is a statistical test that compares the fit of two models, one of which is the null model and the other the alternative model. The likelihood ratio represents how many times more likely the data are under one model than the other. This likelihood ratio, or its logarithm, may then be used to calculate a p-value or compared to a critical value to determine if the null model should be rejected in favor of the alternative model.

Thousands of pharmaceuticals products and ADRs are often included in big drug safety databases like FAERS and MEDEFECT, which may be shown as a data matrix with I rows (ADRs) and J columns (drugs). The number of instances reported is defined as n_{ij} for each ADR-drug combination (cell (i, j)) in the data matrix. Proceeding with this data abstraction, the marginal number of reports for the i^{th} ADR and j^{th} drug may be defined as $n_{i.}$ and $n_{.j}$. At this point the grand total number of reports will be defined as $n_{..}$. The resulting table is shown below in table 1.

Table 1. Representation of Drug-ADRs relationships extracted from self-reporting databases. Each column represents a drug, while each row an ADR. The cells contain the number of reports for each drug-ADR pair.

ADRs	Drugs						Row total
	1	...	j	...	J		
1	n_{11}	...	n_{1j}	...	n_{1J}	$n_{1.}$	
2	n_{21}	...	n_{2j}	...	n_{2J}	$n_{2.}$	
...	
i	n_{i1}	...	n_{ij}	...	n_{iJ}	$n_{i.}$	
...	
I	n_{I1}	...	n_{Ij}	...	n_{IJ}	$n_{I.}$	
Column total	$n_{.1}$...	$n_{.j}$...	$n_{.J}$	$n_{..}$	

The $I \times J$ data-matrix may be flattened into 2×2 tables, each corresponding to a single ADR, for a single drug J of interest as shown in table 2.

Table 2. Single Drug-ADR data extracted from table 1

	Drug j	Other Drugs	Row total
ADR i	$n_{ij} = a$	$(n_{i.} - n_{ij}) = b$	$n_{i.} = a + b$
Other ADRs	$(n_{.j} - n_{ij}) = c$	$(n_{..} - n_{i.} - n_{.j} + n_{ij}) = d$	$(n_{..} - n_{i.}) = c + d$
Column total	$n_{.j} = a + c$	$n_{..} - n_{.j} = b + d$	$n_{..} = a + b + c + d$

Given the definitions on table 2, the LR_{ij} can be defined as:

$$LR_{ij} = \frac{\left(\frac{n_{ij}}{n_{i.}}\right)^{n_{ij}} \left(\frac{n_{.j}-n_{ij}}{n_{..}-n_{i.}}\right)^{n_{.j}-n_{ij}}}{\left(\frac{n_{.j}}{n_{..}}\right)^{n_{.j}}} \quad \text{Eq. 5}$$

This function can be easily converted in logarithmic scale using the definition of table 2.2 as follow:

$$\log LR_{ij} = a * [\log(a) - \log(a + b)] + c * [\log(c) - \log(c + d)] - (a + c) * [\log(a + c) - \log(a + b + c + d)] \quad \text{Eq. 6}$$

At this point we can define the maximum likelihood ratio (MLR) as $\max_i LR_{ij}$, where the maximum is taken over i. However, result more convenient work with its logarithm defined as:

$$MLLR = \max_i (\log LR_{ij}). \quad \text{Eq. 7}$$

The distribution of the MLLR test statistic under the null hypothesis is intractable, hence an empirical distribution is obtained using a Monte Carlo approach. Considering now the conditional distribution of (n_{1j}, \dots, n_{ij}) under the sum $n_{.j}$, we can define it also as a multinomial distribution with parameters $n_{.j}$ and probabilities $\left(\frac{n_{1.}}{n_{..}}, \dots, \frac{n_{I.}}{n_{..}}\right)$ or in other words as:

$$(n_{1j}, \dots, n_{Ij}) | n_{.j} \sim \text{Mult} \left(n_{.j}, \left(\frac{n_{1.}}{n_{..}}, \dots, \frac{n_{I.}}{n_{..}} \right) \right). \quad \text{Eq. 8}$$

The empirical distribution of MLLR under the null hypothesis may now be derived by utilizing this multinomial distribution to generate a large number of Monte Carlo samples. If the MLLR based on the observed data, $MLLR_{data}$, is larger than the threshold value of $MLLR_{0.05}$ (the upper 5th percentile points of the empirical distribution) the null hypothesis is rejected with an $\alpha=0.05$. The ADR associated with MLLR is then the most significant signal detected, or in other words the ADR with the largest log LR value. In our study, all that was necessary for a drug to be linked to an ADR was for the MLLR

to be greater than the threshold. All non-significant associations detected in FAERS or MEDEFECT were discarded.

1.5.2 - Fisher's exact test methodology for ADR-target pairs

One of the main goals of this research work was to associate drugs target and drugs ADR as will be explained in the following chapters. This objective, a part relying on extensive data curation, is based also on specific statistical validation of the ADR-target pair identified. This statistical significance was calculated following the method proposed by Kuhn and colleagues [9]. In a nutshell, the approach computes a contingency matrix for each ADR-protein combination and uses Fisher's exact test to get the p-value. The method proposed is drug-centric, counting how many drugs present or not the ADR or target.

The contingency matrix computed contains the following elements (Figure 10):

- (i) the number of drugs that presents the given ADR
- (ii) the number of drugs that binds to the given protein
- (iii) the number of drugs that presents the given ADR and binds to the given protein;
- (iv) the number of drugs that neither presents the ADR nor binds to the given protein

Given the high number of relationships, p-values were corrected for multiple testing using the 'q-value' module contained in the python package 'MultyPy'. An ADR-protein relationship is accepted if the computed q-value is equal or smaller than 0.05.

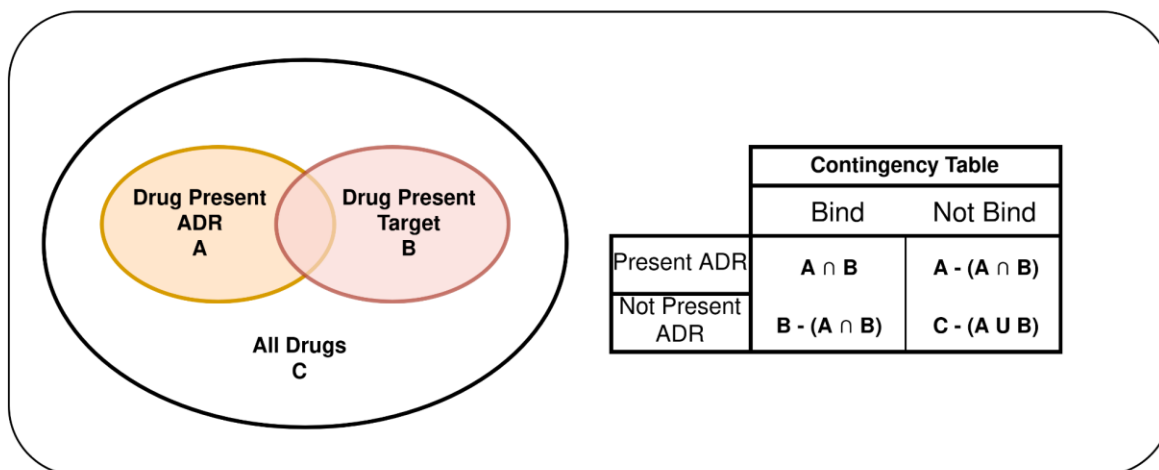


Figure 10. Schematic representation of Fisher's exact test - Set and contingency table representation of drugs used for the computation of Fisher exact test. For each Target-ADR pair identified the method count: the number of drugs that present the given ADR; the number of drugs that bind to the given protein; the number of drugs that present the given ADR and bind to the given protein; and the number of drugs that neither present the ADR nor bind to the given protein. A p-value is then computed for each contingency table.

1.6 - Machine Learning Theory

The broad goal of independent automated processes is artificial intelligence (AI). Humans are becoming increasingly reliant on Artificial Intelligence in modern life, such as quick language processing, malware detection for email, and financial forecast. Machine Learning (ML) is a mere tool for developing Artificial intelligence but it's based on a variety of complex and deep analytical methodologies. Prior to the advent of machine learning, bioinformatics algorithms and energy functions had to be defined manually, which proved difficult for problems such as protein structure prediction given the high number of possible variables. [57] [58]. Deep learning and other machine learning techniques rapidly became the main tools for such problems, giving their ability to learn features of data sets rather than requiring the programmer to define them individually. When properly trained, this multi-layered approach allows such systems to make sophisticated predictions by combining low-level features to create more abstract features, resulting in a more generalized model.

In bioinformatics, machine learning algorithms can be used for prediction, classification, and feature selection. Classification and prediction tasks aim to create models that describe and distinguish classes or concepts in order to predict possible outcomes. Examples of this application are cancer data image recognition or stroke diagnosis. [59] [60]. We can define three different types of machine learning based on data processing: supervised learning, unsupervised learning, and reinforcement learning.

From mapping inputs exploiting a conditional density estimation $p(y_i|x_i,H)$, the *supervised learning* aims to learn the desired output label. In this case the training set, the data on which the machine learning will acquire information, can be defined as $H = \{(x_i,y_i)\}, i = 1,2,\dots,N$, where i define the N training sample, x_i is a N -dimensional vector storing the characteristics, or features, of targets such as the size of an image in the case of image recognition or analytical values in the case of our biological problem, and finally y_i is a nominal or certain variable. Regression is used to handle simple

variable operational issues, whereas classification and pattern recognition are used to tackle some variable questions (label prediction of unknown data). *Unsupervised learning* employs unconditional density estimation to develop a model of a set of data samples $H = \{x_i\}, i = 1, 2, \dots, N$, which can be used for decision making, communicating, and logic, among other things e.g., clustering methodologies. Finally, *Reinforcement learning* is an effective method for learning behavior in response to reward or punishment signals.

Regression, clustering, and classification are the three most common subjects in Machine Learning, but in this chapter, we will focus on three different machine learning technologies framed in the classification problem. Classification, as previously stated, is a supervised learning approach for modeling and predicting categorical variables. In a nutshell, the purpose of these algorithms is to learn an x to y mapping form. We can distinguish between binary classification and multi-class classification based on the number of labels y . For this project we implemented three different binary classifiers, Support Vector Machine (SVM), Random Forest (RF) and Neural Networks (NN).

1.6.1 - Support Vector Machine

SVM is a classification algorithm for datasets. An SVM can be described as a hyperplane that can reliably distinguish between multiple cases with a maximum margin (Figure 11). In the case of binary classification, we can distinguish between positive and negative labels. A margin can be defined as the distance between the hyperplane and the nearest positive and negative sample.

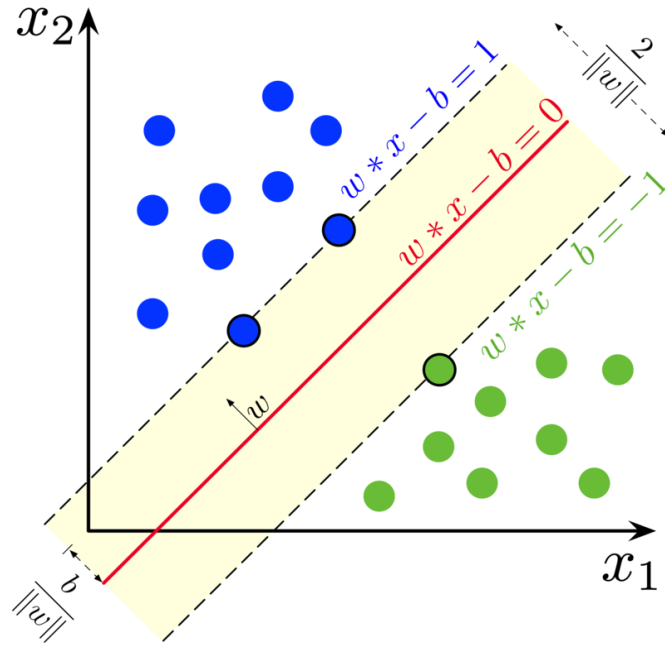


Figure 11. Representation of a linear SVM. In the figure are visible the maximum-margin hyperplane and margins for an SVM trained on two classes of samples. The support vectors are samples on the margin.

Mathematically, the binary classification output of a SVM for a linearly separable dataset is:

$$u = \vec{w} \cdot \vec{x} - b, u = \pm 1 \quad \text{Eq. 9}$$

where \vec{w} is a support vector and \vec{x} is an input vector.

H_1 and H_2 , the nearest points located on the hyperplanes, can be defined as follows:

$$H_1: \vec{w} \cdot \vec{x} - b = 1 \quad \text{Eq. 10}$$

$$H_2: \vec{w} \cdot \vec{x} - b = -1 \quad \text{Eq. 11}$$

Given the distance formula:

$$Distance = \frac{|ax_1 + bx_2 + c|}{\sqrt{a^2 + b^2}} \quad \text{Eq. 12}$$

the margin m can be defined as follows:

$$m = \frac{|\vec{w} \cdot \vec{x} - b|}{\|\vec{w}\|_2} = \frac{1}{\|\vec{w}\|_2} \quad \text{Eq. 13}$$

Given the purpose of the binary classification problem, one way to separate the positives and negatives is maximizing the margin m as follow:

$$\max_{\vec{w}} \frac{1}{\|\vec{w}\|_2} \quad \text{Eq. 14}$$

given the defined H_1 (Eq. 10) and H_2 (Eq. 11) the relative constraints are $\vec{w} \cdot \vec{x}_i - b \geq 1$, for $y_i = 1$ for positive examples and $\vec{w} \cdot \vec{x}_i - b \leq -1$, for $y_i = -1$ for the negative ones. From the combination of the latter, we obtain

$$y_i(\vec{w} \cdot \vec{x}_i - b) \geq 1, y \in \{-1, 1\}. \quad \text{Eq. 15}$$

As a result, the maximizing margin can be obtained as follows:

$$\min_{\vec{w}, b} \frac{1}{2} \|\vec{w}\|^2 \text{ subject to } y_i(\vec{w} \cdot \vec{x}_i - b) \geq 1, y \in \{-1, 1\}, \forall i \quad \text{Eq. 16}$$

where \vec{x}_i is the i th training example and y_i is the label of SVM. At this point, we turn the primal function into a Lagrange function using the Lagrange method,

$$L(\vec{w}, b, \alpha_i) = \frac{1}{2} \|\vec{w}\|^2 + \sum_{i=1}^n \alpha_i [1 - y_i(\vec{w} \cdot \vec{x}_i - b)] \quad \text{Eq. 17}$$

Taking in consideration Eq. 16 we may obtain the definition of the Primal Problem as $\min_{\vec{w}, b} L(\vec{w}, b, \alpha_i)$ concerning \vec{w} and b . However, we must turn this primal problem into a dual problem due to the intricacy of the constraints and the uncertainty of the input variables. Giving the definition of the Dual function as $g(\alpha_i) = \min_{\vec{w}, b} L(\vec{w}, b, \alpha_i)$ we can expose the Dual Problem as follows:

$$\max_{\alpha_i \geq 0} [g(\alpha_i)] = \max_{\alpha_i \geq 0} \left\{ \min_{\vec{w}, b} \left\{ \frac{1}{2} \|\vec{w}\|^2 + \sum_{i=1}^n \alpha_i [1 - y_i (\vec{w} \cdot \vec{x}_i - b)] \right\} \right\} \quad \text{Eq. 18}$$

The above function meets the Karush-Kuhn-Tucker (KKT) necessary condition at point $(\vec{w}^*, b^*, \alpha_i^*)$:

$$\nabla_{\vec{w}} L(\vec{w}^*, b^*, \alpha_i^*) = \vec{w}^* - \sum_{i=1}^n x_i \alpha_i^* y_i = 0 \rightarrow \vec{w}^* = \sum_{i=1}^n \alpha_i^* x_i y_i \quad \text{Eq. 19}$$

$$\nabla_b L(\vec{w}^*, b^*, \alpha_i^*) = -\sum_{i=1}^n \alpha_i^* y_i = 0 \rightarrow \sum_{i=1}^n \alpha_i^* y_i = 0 \quad \text{Eq. 20}$$

$$\text{Primal Feasibility: } y_i (\vec{w} \cdot \vec{x}_i - b) \geq 1, y \in \{-1, 1\} \quad \text{Eq. 21}$$

$$\text{Dual Feasibility: } \alpha_i^* \geq 0 \quad \text{Eq. 22}$$

$$\text{Complementary Slackness: } y_i (\vec{w} \cdot \vec{x}_i - b) \geq 1, y \in \{-1, 1\} \quad \text{Eq. 23}$$

Taking in consideration Eq. 18 and Eq. 19, the primal problem can be transformed as follows:

$$L(\vec{w}, b, \alpha_i) = \frac{1}{2} \|\vec{w}\|^2 + \sum_{i=1}^n \alpha_i [1 - y_i (\vec{w} \cdot \vec{x}_i - b)] \quad \text{Eq. 24}$$

$$\min_{\vec{w}, b} L(\vec{w}, b, \alpha) = \min_{\vec{w}, b} \left\{ \frac{1}{2} \|\vec{w}\|^2 + \sum_{i=1}^n \alpha_i [1 - y_i (\vec{w} \cdot \vec{x}_i - b)] \right\} \quad \text{Eq. 25}$$

$$\min_{\vec{w}, b} L(\vec{w}, b, \alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \|\vec{w}\|^2 \quad \text{Eq. 26}$$

from Eq. 18 we obtain:

$$\vec{w}^* = \sum_{i=1}^n \alpha_i^* y_i \vec{x}_i \quad \text{Eq. 27}$$

So:

$$\begin{aligned} \min_{\vec{w}, b, \xi} L(\vec{w}, b, \xi_i, \alpha_i, \beta_i) = \\ \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n y_i y_j (\vec{x}_i \cdot \vec{x}_j) \alpha_i \alpha_j \end{aligned} \quad \text{Eq. 28}$$

$$\begin{aligned} \max_{\alpha_i \geq 0} \left[\min_{\vec{w}, b, \xi} L(\vec{w}, b, \xi_i, \alpha_i, \beta_i) \right] = \\ \max_{\alpha_i \geq 0} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n y_i y_j (\vec{x}_i \cdot \vec{x}_j) \alpha_i \alpha_j \end{aligned} \quad \text{Eq. 29}$$

As a result, the reduced dual (QP) issue based on the $\vec{\alpha}$ -dependent objective function ψ is solved:

$$\min_{\vec{\alpha}} \psi(\vec{\alpha}) = \min_{\vec{\alpha}} \left[\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n y_i y_j (\vec{x}_i \cdot \vec{x}_j) \alpha_i \alpha_j - \sum_{i=1}^n \alpha_i \right]$$

$$\text{subject to } 0 \leq \alpha_i \leq C \forall i \quad \text{Eq. 30}$$

$$\sum_{i=1}^n \alpha_i y_i = 0 \quad \text{Eq. 31}$$

The slack variable ξ_i will not appear in the QP function. In the case of non-linear classifiers SVM, the output is computed from Lagrange multipliers:

$$u = \sum_{i=1}^n \alpha_i y_i K(\vec{x}_j, \vec{x}) - b \quad \text{Eq. 32}$$

where $K(\vec{x}_j, \vec{x})$ is a kernel function. The kernel function is a useful tool for dealing with non-linear classifiers. It converts non-linear sample features into a high-dimensional space where the relevant features can be linearly distinguished. The linearly separable scenario is used to derive the output of non-linear classifiers.

$$u = \vec{w} \cdot \vec{x} - b, \text{ and } \vec{w} = \sum_{j=1}^n \alpha_j x_j y_j \quad \text{Eq. 33}$$

Other SVM techniques require data in raw input to be transformed into feature vector representation, whereas kernel approaches just require a user-specified kernel (Figure 12). The polynomial kernel, Gaussian kernel, and neural network non-linearities are all examples of kernel functions. In this research work we exploit the *Radial Basis Function kernel*. On the two samples \vec{x}_j and \vec{x} represented as feature vectors in some input space, the RBF kernel is defined as:

$$K(\vec{x}_j, \vec{x}) = \exp\left(-\frac{\|\vec{x}_j - \vec{x}\|^2}{2\sigma^2}\right) \quad \text{Eq. 34}$$

In particular the RBF kernel has a straightforward interpretation as a similarity measure since its value decreases with distance and ranges between zero (in the limit) and one (when $\vec{x}_j = \vec{x}$).

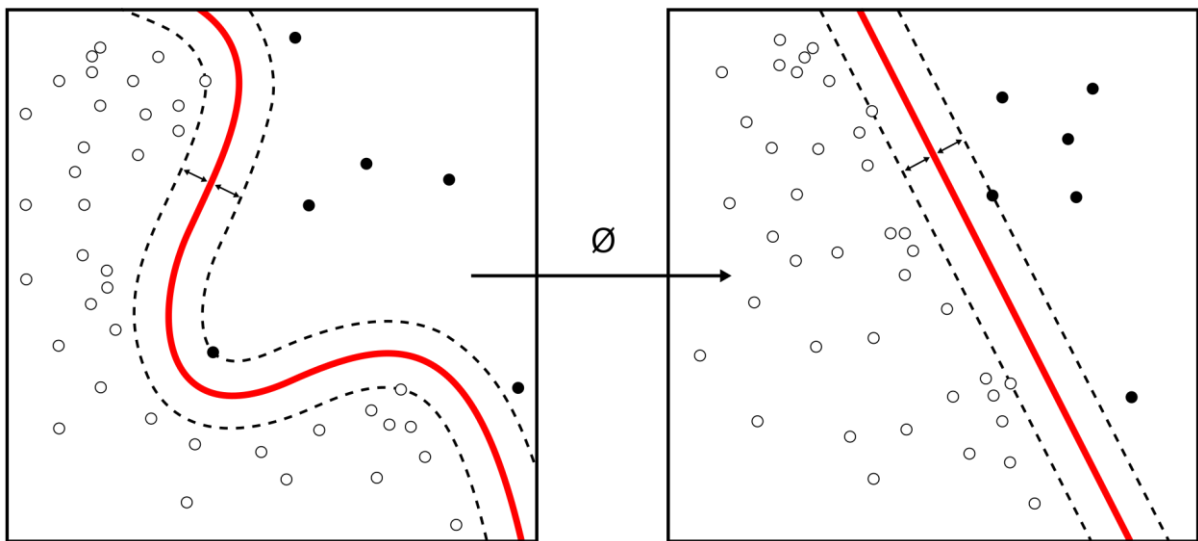


Figure 12. Representation of the application of kernel function

1.6.2 - Tree-based methods

Decision trees can be defined as hierarchical learners made out of a group of basic (binary) choices. Breiman et al. [61] developed the tree model for classification and regression applications in 1984. Following that, decision trees grew in popularity and were frequently used in a variety of machine learning techniques. One explanation for their success could be that they have a number of advantages: they are simple to use, rapid and adaptable to huge datasets, and they can be constructed in a probabilistic manner to account for variability. Instead of attempting to optimize a single complicated tree, Ho [61] proposed constructing an ensemble of "weak" decision trees, namely random forests, as a result of the formation of ensemble learning. The authors of these papers propose injecting randomness into the learning process to build decorrelated trees. Acquiring better generalization, the new method demonstrated superior accuracy by averaging their predictions. Since then, random forests have been successfully used in a wide range of applications, the majority of which have been phrased as classification problems.

1.6.2.1 - Decision tree

As already mentioned, the classification problem in a supervised learning environment can be defined as a maximum *a posteriori* probability function $p(y_i|x_i, H)$ and given the training set $H = \{(x_i, y_i)\}, i = 1, 2, \dots, N$ we aim at learning the posterior $P(x|y)$. Finding and developing a good model for this posterior throughout the entire feature space X is a challenging task. A decision tree uses a "divide" and "conquer" technique to solve this problem: (a) it uses a series of decisions to form a partition over the input feature space, and (b) it calculates $P(Y|X)$ within every layer of this environment. The concept behind decision trees is to make predictions using a series of basic selections. In fact, a decision tree model is made up of a set of (binary) decisions that are arranged in a hierarchical order (Figure 13).

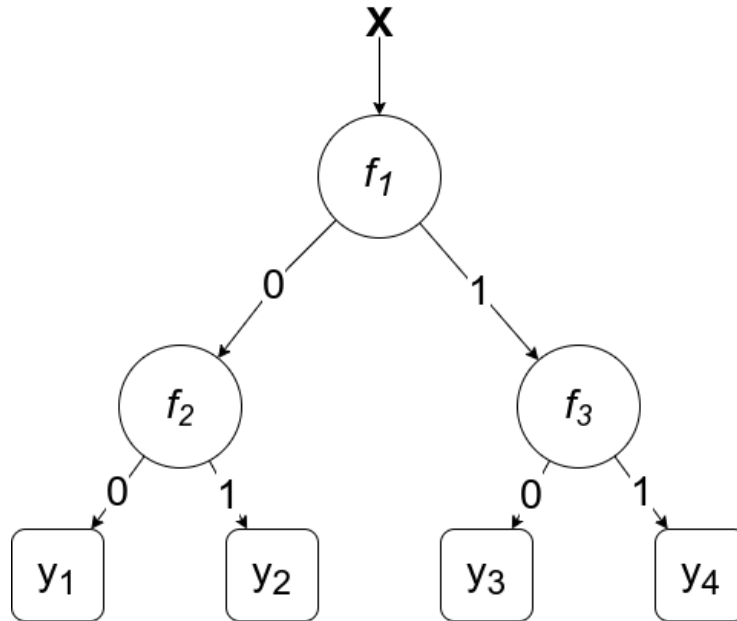


Figure 13. Schematic representation of a binary tree

To perform a binary decision, a node N_i from the set N of a tree is equipped with a so-called splitting function f_i whose role is to split incoming observations in two subsets denoted as S_i^{left} and S_i^{right} . These two subsets represent respectively the left and right child of N_i . The splitting function can be represented as follows:

$$\{f_i(x) = 0, x \text{ is sent to left } f_i(x) = 1, x \text{ is sent to the right} \quad \text{Eq. 35}$$

Tree learning can be characterized as an iterative node optimization and splitting. Indeed, at each node, a suitable splitting function must be selected first, and then the training data must be split and distributed to the left and right children. Several parameters must be selected depending on the function class used and during training the decision functions at each node are optimized to that end until a stopping criterion has been reached. The iterative splitting of the training data stops once the bottom of the tree is reached, and the current node becomes a leaf node.

Usually, the three most common stopping criteria are the (a) maximum tree depth, (b) the minimum population per leaf, and (c) the target function's minimum variation. The first approach just analyzes the hierarchy's depth, and once that depth is achieved, the

iterative splitting ends. The second is based on the amount of training examples arriving at a node, with the splitting stopping if the population of training points falls below a specified threshold. The last one is about the optimized objective function. If the variation is less than a given threshold, it is assumed that splitting the training cases yielded no additional information.

1.6.2.2 - Random Forests

Several randomization approaches have been developed to generate decorrelated or independent trees. Breiman coined the term "bagging," which is a mix of the terms "bootstrap" and "aggregating." Given a training set H , a bootstrap is essentially a fraction H_p of the whole training set, with or without replacement, where each element has been randomly selected using a uniform distribution. The ensemble's trees are then trained using a new bootstrap H_p . Finally, averaging is used to combine the predictions from all of the separate trees. Integrating randomness in tree training has several advantages: firstly, increasing the levels of randomness reduces the correlation between the different trees, resulting in greater generalization; secondly, if the total number of features is constrained to be sparse, it enables implicit feature selection from each tree; and thirdly, it allows independence from the training set, — in other words, it add robustness to noisy data.

Random forests provide a very versatile architecture that allows for the creation of task-specific objective functions, various splitting functions, and Bayesian models. Furthermore, they only have a few hyperparameters: (1) the number of trees and (2) the tree depth which are the two most essential random forests' hyperparameters. Increasing the number of trees allows noisy predictions to be averaged out, resulting in a monotonic decrease in prediction error. The maximum permissible depth of the tree is a critical parameter that must be optimized since it has a direct impact on each tree's generalization capacity. Furthermore, while a small tree's prediction will be unreliable due to the large amount of heterogeneous data in its leaves, a very deep tree's leaves will have very little training data to compute sound statistics. This is directly related to over-

fitting the training data, such as fitting noisy features, and suffers from poor generalization. To avoid this scenario, it's best to optimize the prediction error curve until it reaches a minimum. This minimum corresponds to the optimal tree depth, allowing for good modeling and generalization of the observations.

1.6.3 - Neural Networks

Serious efforts were invested in establishing mathematical representations of cognitive processing in biological systems in the last 30 years. [62]. This research resulted in the development of the neural network methodology. While this goal is still a long way off, (artificial) neural networks have demonstrated to be among the most effective approach in a wide range of problems [63] [64] [65]. A neural network is usually made up of numerous units, also known as neurons, that are mathematically described as:

$$h_j(x) = \sigma\{w_j + \sum_{i=1}^n w_{ij}x_i\} \quad \text{Eq. 36}$$

where σ is a non-linear activation function, such as the sign, sigmoid, or softmax activation function. The activation functions of interest for this research are the following:

The *logistic sigmoid function*:

$$\sigma(x) = \frac{1}{1+\exp(-x)} \quad \text{Eq. 37}$$

The *Logistic sigmoid function* monotonically maps real numbers to the [0, 1] range, making it suitable for modeling binary classification;

The *rectified linear function* (ReLU):

$$\sigma(x) = \max(x, 0)$$

Eq. 38

The *rectified linear function* (ReLU) currently is one of the most popular activations (owing to its biological plausibility, sparsity, lack of vanishing gradients, and computational efficiency).

In the majority of cases, these components are organized into layers, with the outputs of one layer directed to the inputs of the next layer by weighted connections, known as synapses. Figure 14 represents a three-layered neural network.

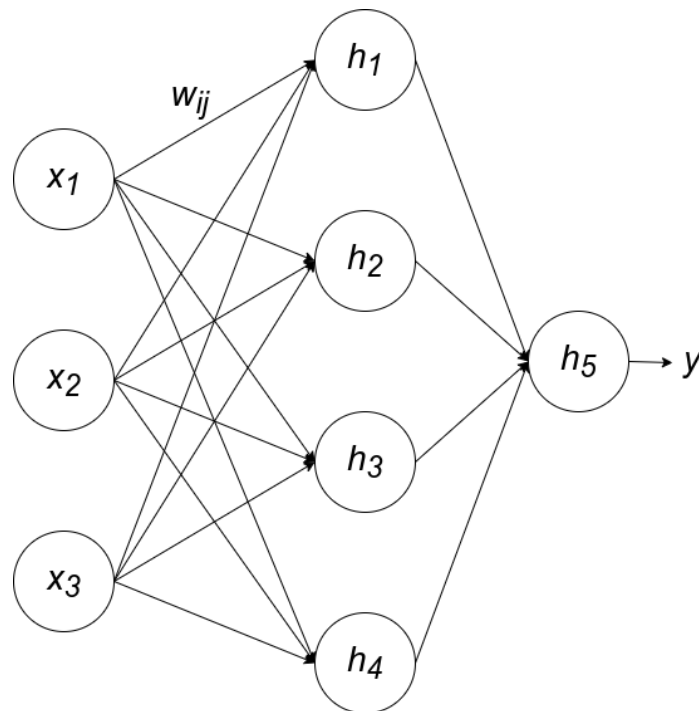


Figure 14. Schematic representation of a perceptron.

The first is known as input layer, which transmits the initial values $x = \{x_1, \dots, x_n\}$ to the second layer. The second layer consists of activation units h_n , which take the weighted values from the first layer as inputs and output nonlinear transformed values. The predicted value y is produced by the third layer, which is made up of a single activation unit that takes the weighted outputs of the second layer as inputs. The learning procedure of a neural network entails estimating the weights w_{ij} that minimize a certain loss function using a specific optimization procedure. The *backpropagation* algorithm is the most well-known of all of them. Recent breakthroughs in neural networks, commonly called "deep learning," have demonstrated that these models can learn high-level and very effective data representations on their own. On a range of difficult tasks, such as picture classification and speech recognition, neural networks have proven to outperform both human operators and state-of-the-art technologies.

1.6.4 - Evaluation metrics of binary classification models

In this section, we define several evaluation metrics in the case of a binary classification problem. These measures will be one of the stepping stones of the empirical analyses in the following chapters. In this context, the term *classifier* will be used to denote the model inferred by supervised learning, and replace the target variable y by the term *class* $\in \{0, 1\}$.

Once a classifier C has been trained and built, evaluating its predictive capability can be achieved using an independent test set T of size n (this test set is typically the part of the database that was not used to create the model).

$$T = \{sample_j\} = \{(x_j, y_j)\}, \quad j = 1, \dots, n' \quad \text{Eq. 39}$$

The most common and straightforward evaluation metric is the *accuracy*, which is equal to the ratio of the number of correctly identified items to the size of the test set:

$$Accuracy = \frac{\#\{sample_j: C(x_j) = class_j, j=1, \dots, n'\}}{n'} \quad \text{Eq. 40}$$

This measure, however, presents two main drawbacks. First of all, it's strictly dependable by the number of objects of each class represented in the test set i.e., one class, for example, may be significantly over-represented, resulting in an average mistake rate that mostly reflects the rate of properly categorizing objects from this latter class. Secondly, the most common output of a trained model is a class-probability $[0, 1]$ for each input vector of features. A threshold must be specified in order to convert this into a class prediction. For example, in a binary classification problem, the most typical decision is to choose a threshold of 0.5, however this may not always be the best option. However, also the case of incorrect prediction has to be taken in consideration, such as in the event of negative prediction for positive samples or positive prediction for negative samples. In this case we can distinguish between Type I (false positive) and Type II (false negative) errors. Given these prediction characteristics, it is then possible to derive a contingency table (also known as confusion matrix) (Table 3) to have an estimate of this type of error and be able to compute supplementary evaluation metrics.

Table 3. Representation of a confusion matrix or contingency table. Given a classifier and a predicted point, there are four possible outcomes. If the point is positive and it is classified as positive, it is counted as a true positive; if it is classified as negative, it is counted as a false negative. If the point is negative and it is classified as negative, it is counted as a true negative; if it is classified as positive, it is counted as a false positive. P and N represent the total number of positives points and negative points respectively (Adapted from [66] - Figure 1)

		<u>True class</u>	
		p	n
<u>Hypothesized class</u>	Y	True Positives	False Positives
	N	False Negatives	True Negatives
Column Totals:		P	N

In the contingency matrix we can define different classes of elements based on the trueness of prediction. In this thesis to evaluate the model's performance we relied on this particular measurement.

- Accuracy, as stated before;
- Precision;
- Recall;
- Receiver operating characteristic area under curve (ROC AUC);
- Matthew Correlation Coefficient

The *precision*, or positive predictive value, is defined as:

$$PREC = \frac{TP}{(TP+FP)} \quad \text{Eq. 41}$$

where TP is the number of true positives (or the positive labels correctly predicted) and FP the number of false positives (the number of labels predicted as positives but in reality, negatives). The precision is intuitively the ability of the classifier not to label as positive a sample that is negative.

The *recall*, or sensitivity, represent the ability of the classifier to find all the positive cases. It can be defined as:

$$REC = \frac{TP}{(TP+FN)} \quad \text{Eq. 42}$$

where FN is the number of false negatives (the number of labels predicted as negatives but in reality, positives).

A *receiver operating characteristic curve*, or ROC curve, is a visual representation that shows how a binary classifier system's performance changes as the discrimination threshold change. The ROC curve is created by plotting the recall against the false positive rate (FPR) at various threshold settings. The FPR is defined as:

$$FPR = \frac{FP}{FP+TN} \quad \text{Eq. 43}$$

where TN represents the number of true negatives (the negative labels correctly predicted). The AUC can also be interpreted as the probability that the classifier will assign a higher score to a randomly chosen positive sample than to a randomly chosen negative one.

Finally, the *Matthew Correlation Coefficient* is a metric for assessing the quality of binary and multi-class classifications in machine learning. It accounts for true and false positives and negatives, and is widely recognized as a fair metric that can be applied even when the classes are imbalanced. The MCC is essentially a -1 to +1 correlation coefficient number. A perfect prediction has a coefficient of +1, an average random prediction has a coefficient of zero and an inverse prediction has a coefficient of -1. Mathematically can be defined as:

$$MCC = \frac{TP*TN-FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad \text{Eq. 44}$$

1.7 - Summary

In this chapter, we introduced the biological concept of network, drug discovery, statistical methods, supervised machine learning, and the possible measures of evaluation of such models. It constitutes the basis of our research and we strongly believe that the overall methodology “package” is well suited to the biological problem at hand. Indeed, there is a strong overlap between the two fields in the tasks they try to achieve. The first one is the identification of association between the different drug targets and related drug ADRs and the second one is the ability to exploit the underlying target biological information to estimate the probability of an unknown protein to elicit an adverse reaction when targeted by a drug. In chapter 4, I will discuss the details of how to obtain reliable Proteins-Adverse event relationships and how we’re going to exploit this information in the supervised-learning framework to obtain reliable predictions in chapter 5. Chapter 6 will exploit the acquired knowledge to investigate the molecular relationship between the different ADRs and protein targets.

2 - OBJECTIVES

The underlying theme of this thesis has been the development of novel computational tools to aid in the process of drug discovery. Particularly, I have focused my research in the identification of target liabilities and early detection of associated adverse reactions. Ultimately, the long-term goal is to aid in the development of safer and more effective drugs while shortening the time required for preclinical studies.

The first aspect of my thesis has been the mining and uncovering of protein-ADRs associations using the vast landscape of pharmacovigilance resources and drug-protein research. I hope to advance our understanding of Adverse Reaction-Protein relationships by providing resources to organize and analyze both systems. Therefore, the first objective of the thesis is:

1. Create a publicly accessible database of statistically validated Adverse reaction-drug target pairs. At the same time, provide all of the tools for reproducing the results obtained locally, allowing for the incorporation of new information and data as new database releases are produced. In this way, the database can be kept up-to-date and abreast of new data.

This objective involves the following sub-objectives:

- Obtain information on proteins, drugs, and adverse reactions from a variety of sources. Linked to this was the identification, study and selection of specific databases from which to collect data.
- Statistically validate the entries from self-reporting databases, extracting only meaningful reports. This is particularly important given the heterogeneity and patchy quality of these resources.
- Compile and integrate the data from drug-target and drug-adverse event databases
- Statistically validate the ADR-target pair identified following the method of Kuhn and colleagues [9].

- Creating a publicly accessible database to access and query the data, T-ARDIS, as well as creating a public repository for users to replicate the database is needed.

Due to unforeseen adverse drug reactions, drug discovery attrition rates remain astoundingly high, particularly in late clinical trial phases. As a result, not only recognizing but also anticipating negative adverse reactions prior to clinical trials would help to create safer medicines while avoiding economic losses. In this sense, a predictive approach that can foresee potential issues with the development of a novel drug to target a given protein will be of special interest. An so, the second objective of my thesis is:

2. Develop a method for predicting whether the modulation of a specific protein will result in a specific Adverse Reaction.

This objective involves the following sub-objectives:

- Extract proteins linked to a subset of T-ARDIS' adverse reactions
- Extract relevant network-based information on associated proteins for each of the Adverse reactions chosen (this includes also studying the relationship and the identified proteins' interactome neighborhood via GUILDify).
- Convert the obtained data into an appropriate format and develop a predictive model based on three different machine learning approaches for each of the selected ADRs.
- Propose three different consensus scoring functions to combine the models obtained from the machine learning methods.
- Develop an easy-to-use application for accessing and using all the models obtained for the research community.

While we can predict and associate proteins with their putative adverse reactions, the actual molecular mechanisms of the latter and the phenomenon of associated adverse reaction still eludes us. The third objective of my thesis is:

3. Investigate the molecular basis of identified adverse reactions studying the underlying set shared protein targets from a network perspective. Namely, connect different adverse reactions (nodes) if sharing common proteins (edges) and perform a number of network-based studies.

This objective involves the following sub-objectives:

- Develop a novel network to integrate all the information of ADR-target relationships extracted from T-ARDIS. The resulting network was named "Adverse Reactome" to emphasize its nature.
- Applying different network-based clustering to extrapolate meaningful Adverse-reaction subsets, defined by the protein shared between them.
- For each cluster, investigate the molecular function of the identified proteins.
- Examine the scientific literature for evidence that the disruption of the enriched functions can cause the cluster's distinctive adverse reactions.
- Look into the chemical similarities between the drugs that target the cluster-associated proteins with the view of identifying any similarities.

**3 - IDENTIFYING ADVERSE
REACTION - TARGET
ASSOCIATIONS MINING DRUG-
TARGET AND DRUG-ADR
RESOURCES: T-ARDIS**

3.1 - Abstract

As previously stated, unexpected adverse drug reactions (ADRs), associated with drug candidates, are a well-known cause of the high dropout rate in drug development. The ability to predict adverse reactions when modifying specific protein targets would aid in the growth of safer medications without considering the far-reaching commercial implications. Still, how to reach this goal is a matter of discussion and an active field of research. Indeed, while many databases collect information about drug–target interactions for research purposes, and many public resources collect information about drugs and adverse drug reactions in the context of pharmacovigilance, databases that directly link the relationship between adverse reactions and protein targets are quite scarce. Despite this, given the large amount of raw data available, it appears possible to link targets and ADRs by using drug entities as connecting components.

In this chapter, I will discuss T-ARDIS (Target—Adverse Reaction Database Integrated Search), a freely accessible database designed to store information about the drug-adverse reaction relationship. By integrating publicly available databases I created a new resource that links statistically significant associations between known ADRs and protein targets. This innovative database emerges in response to a growing interest in identifying problematic protein targets early in the drug development pipeline, which modulation could result in ADRs. The chapter is divided in three parts and major steps are described in figure 15.

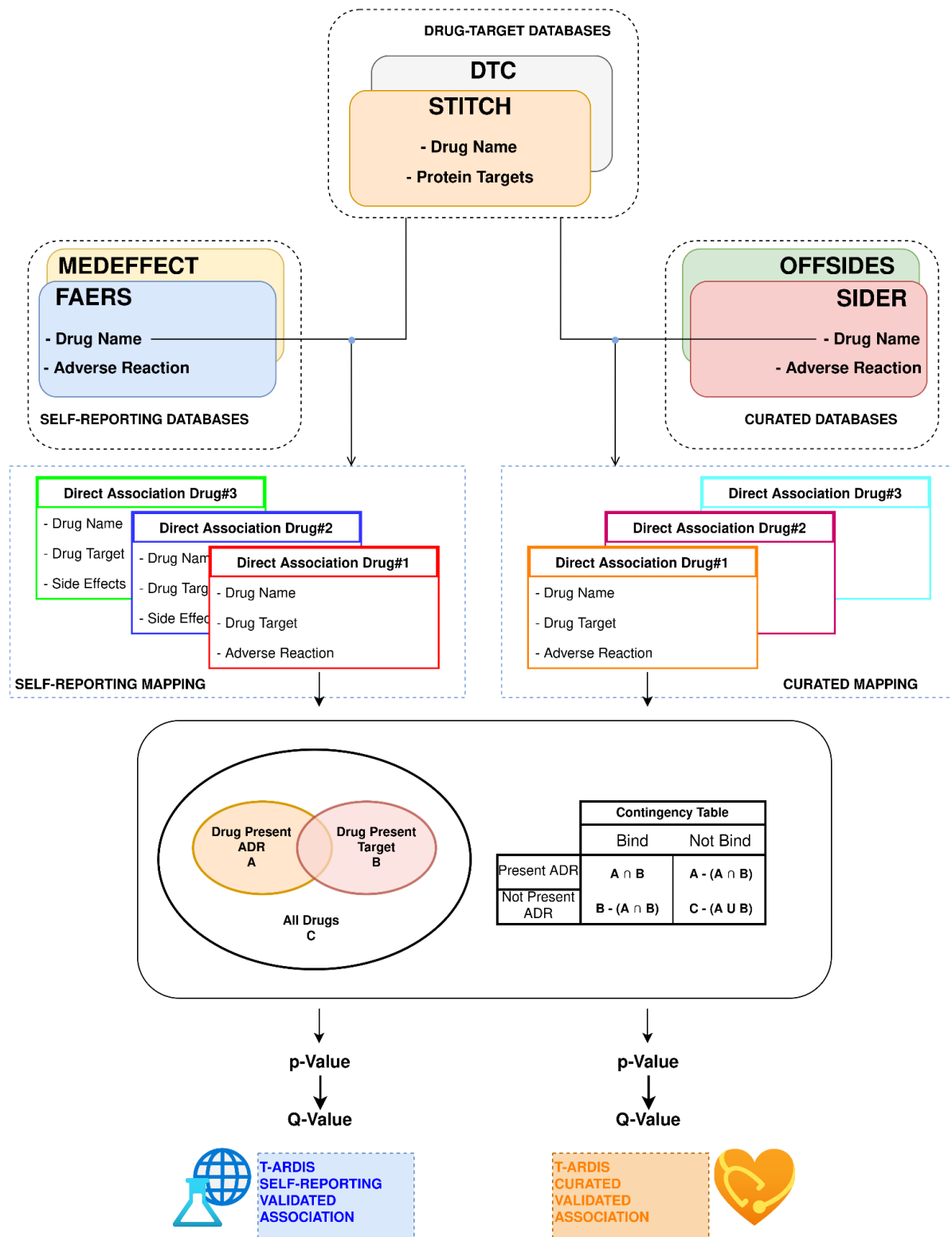


Figure 15. Workflow followed to combine and derive statistical associations between proteins and ADR. Drug-ADR and drug-target associations are retrieved from relevant databases. Subsequently, statistical association between proteins and ADRs is computed as described by Kuhn et al. [9].

The first part will introduce the selection and mining of the databases that are at the basis of T-ARDIS development, explaining all the process of filtering, quality-control and standardization of entries for both drug-ADR and drug-target databases. The second section, which is the heart of T-ARDIS, will focus on statistical validation of the identified protein-ADR pairs using the method developed by Kuhn et al [9]. The last part will instead describe the relevant ADRs-proteins association discovered in parallel with the possible clinical implication. T-ARDIS currently contains about 3000 ADR and 248 targets, totaling around 17 000 pairwise interactions. Each entry can be found using a variety of search criteria, including the target Uniprot ID, gene name, adverse effect, and drug name. Furthermore, the database can be redeployed locally via a convenient git-hub repository or simply accessed via the available web-service.

3.2 - Mining and curation of Databases cleaning procedure

The primary sources of information of T-ARDIS comes from many different public repositories. The databases used for this study can be divided into three main categories:

- (i) Drug-ADR self-reporting databases: FAERS [31] and MEDEFECT [32],
- (ii) Drug-ADR curated data: OFFSIDES [35] and SIDER [34],
- and finally (iii) Drug-Target databases: STITCH 5.0 [36] and Drug-Target commons2.0 [37].

As I will be explained in the following sections, each of these datasets underwent specific filtering and quality-control procedures to standardize and repurpose the information for further analysis. The Drug – ADR self-reporting category, in particular, has been subjected to extensive data curation and validation, first to standardize the drug names contained in FAERS and MEDEFECT, and then to statistically validate the drug – ADR pairs obtained.

3.2.1 - FAERS and MEDEFECT entries standardization

The original code implemented for the quality-control of the FAERS database was developed by Banda and col. [67] in the AEOLUS and OHDSI initiative [68], managing to obtain reliable data until the 2015 FAERS release. In the context of T-ARDIS, this code has been updated and modified to accept the new format of FAERS and MEDEFECT data until 2021. The drug – ADR standardization process applied is divided into several cleaning steps with the primary goal of eliminating duplicate case records and implementing standardized vocabularies with drug names mapped to RxNorm concepts [69]. All cleaning stages will be explained in detail in the following sections, beginning with the cleaning and standardization of FAERS data.

As one might expect, the first step is to collect the raw information, which is available on the FDA's website in two formats (XML and Comma Separated Values files (CSV)) (<http://goo.gl/9Lcc65>). The files are divided by year, starting from 2004. The data coming from 2004 to 2012, in particular, are referred to as LAERS. This distinction is due to a significantly different format than the one adopted since September 2012, referred in this chapter simply as FAERS. As stated in the introduction chapter, the data stored in the LAERS/FAERS data-sets consists of adverse events and medication errors voluntarily reported in the United States by healthcare professionals (pharmacists, nurses, physicians) and consumers (patients, lawyers, family members). The FDA is responsible for compiling all of these reports into a single resource, which given its self-reporting and unsupervised nature, creates a number of challenges assessing the quality of the data, for instances in case of duplications, typological errors, etc.

3.2.2.1 - Data Download and handling

Each of the quarterly FAERS/LAERS data files stored at the FDA website are separated into seven independent tables (table 4). While the format of files remains stable, the most significant distinction between LAERS and FAERS datasets are some of the essential fields such as *isr* and *case* and *primaryid* and *caseid*. To efficiently merge and compare the information stored in the various files, these fields must be mapped to a single unique identifier so that all of the primary report information is preserved, which means that any information retrieved can be traced back to its original source. Table 4 reports the main files available for download from the FDA website as in the original paper of Banda et al. [67]. The nomenclature suggests the report's information type, the year (indicated as *yy*) and year's quartile (indicated as *Qq*). Among the different files provided, *DEMOyyQq* tables, which contain patient administrative information, are crucial in the missing value imputation and case de-duplication processes.

Table 4. List of retrieved LAERS/FAERS files - Each of the presented files contain different report information useful for case deduplication and statistics. *yyQq* indicates the report year and year's quartile.

File name	Description
<i>DEMOyyQq</i>	Contains patient demographic and administrative information, each row represents an individual event report
<i>DRUGyyQq</i>	Contains drug information for all medications reported for the event report (1 or more rows per report)
<i>INDIyyQq</i>	Contains all MedDRA terms for the indications of use for the reported drugs (0 or more per drug per event)
<i>OUTCyyQq</i>	Contains patient outcomes for the event report (0 or more rows per report)
<i>REACyyQq</i>	Contains all MedDRA terms related to the adverse event report (1 or more rows per report)
<i>RPSRyyQq</i>	Contains the source of the event report (0 or more rows per report)
<i>THERyyQq</i>	Contains drug therapy start dates and end dates for the reported drugs (0 or more rows per report)

3.2.1.2 - Missing values and cases de-duplication

Banda et al. perform a very peculiar analysis in order to identify and remove duplicate entries and cases [67]. The amount of redundancy in FAERS is directly proportional to the number of reports submitted to the FAERS database by patients, medical doctors, and pharmaceutical companies. Furthermore, many reports may include multiple versions, such as the initial case version, additional follow-up case versions, or even exist in the older LAERS dataset as well. As a result, and in order to eliminate redundancy

and not miss any single piece of information that could have been added in the various reports, the case deduplication algorithm must account for all of these multiple case versions. The DEMO files come at hand in this first mapping procedure, containing useful information such as the unique report row keys (*isr* in LAERS and *primaryid* in FAERS), as well as the various report country codes and administrative information. In particular, all four "important" demographic data fields (event date, age, gender, and reporter country) must be fully compiled in at least one version of the report case, or be completely discarded. With the elimination of report case redundancy, which inflates the number of total reports associating a drug with an ADR, the algorithm can now focus on the drug's name standardization.

Given the self-reporting nature of FAERS and MEDEFECT, there is a significant discrepancy in the drug's label names in many reports. This disparity may manifest as typographical errors, mistyping, or the use of non-standard drug names. To provide a standardized framework, the drug names present in the database are extracted and mapped into the RxNorm vocabulary using in combination the OHDSI Vocabulary version 5 and regular expressions [67]. A second round of standardization of drugs' names is also performed using the FDA's orange book of NDA ingredients, checking the report for New Drug Applications (NDAs) codes. Finally, mapped drug names are linked to their respective adverse events using the unique report identifiers while unmapped drugs are definitely discarded.

MEDEFECT, despite its different file structure, uses the same cleaning logic. In this case, redundancy is removed by modifying the files to look like the FAERS ones and checking the unique MEDEFECT identifier. In a nutshell, the columns in the MEDEFECT files are renamed and ordered to match those in the FAERS files. Once this preliminary passage is completed, the MEDEFECT reports drug's names are extrapolated using the RxNorm and regular expression mapping as described before. MEDEFECT and FAERS data have been treated independently until now in order to minimize involuntary redundancy due the report's unique id. After standardizing FAERS and MEDEFECT entries, the data from both databases can be combined and subjected to the drug-ADR statistical validation process outlined below.

3.2.2 - Selfreporting Drug-ADRs validated associations

Following the standardization of drug names and case de-duplication for the FAERS and MEDEFECT databases, statistical validation of the drug-ADR entries retrieved is required. For this purpose, the approach proposed by Huang et al. [56], which is already outlined in the statistical methods introduction sub-chapter, is used. Only drug-ADR associations that are statistically significant, that is, have a likelihood ratio value greater than the 5th percentile of the multinomial distribution and are present in both FAERS and MEDEFECT, were kept. Following the Advanced Filtering procedure (see below) FAERS yielded approximately 4 million pairwise interactions originating from over 9000 chemicals and around 17 000 distinct ADR as a result of the curation methodology. From a total of over 4000 and 12 000 drugs and ADR occurrences reported respectively in the database, 1.5 million drug-ADR connections were discovered instead for MEDEFECT.

3.2.3 - SIDER and OFFSIDES databases

The information stored in other drug -ADR datasets such as SIDER and OFFSIDES were used without any filtering procedures apart from the advanced ones (see below). The reason is both databases are already curated and thus do not require any quality check actions. On the one hand, over 108 000 pairwise interactions were mined for a total of 1344 distinct medicines and 2303 ADRs in the SIDER analysis. On the other hand, OFFSIDES produced a huge number of pairwise drug-ADR associations: 1.5 million from a total of 2708 and 4368 distinct drugs and ADRs, respectively.

3.2.4 - Advanced Filtering procedures

Another major concern with drug-adverse-reaction databases is that certain ADRs are too broad to be specific to body regions, tissues, or basic human biology. As a result, any ADR belonging to the following SOCs was rejected, as suggested in the article of Ietswaart et al. [70]

- *General disorders and administration site conditions.*

As the name implies, this SOC contains concepts that do not easily fit into any one SOC's hierarchy or are non-specific illnesses that affect multiple body systems or places. It should be noted that including PTs in this SOC in each possible secondary SOC would result in an excessive number of redundancies. As a result, the majority of the PTs in this SOC are primarily associated with SOC General disorders and administration site conditions, with only minor representation in secondary SOCs (e.g., PT Injection site atrophy is primarily associated with SOC General disorders and administration site conditions, with only minor representation in secondary SOCs).

- *Injury, poisoning and procedural complications*

This SOC categorizes medical concepts in which there is a major injury, poisoning, procedural, or device complication in the medical event being reported. In general, all of the events in this SOC appear to be directly attributable to trauma, poisoning, and procedural complications, in other words, all of the occurrences that are due to an external cause.

- *Investigations.*

A clinical laboratory test idea (including biopsies), radiologic test concept, physical examination parameter, and physiologic test concept (e.g., pulmonary function test) are all considered investigations by MedDRA. This SOC only had PTs that represented investigation techniques and qualitative results (e.g., PT blood sodium decreased, PT blood glucose normal). PT hyperosmolar state, PT haemosiderosis, PT orthostatic proteinuria, and PT renal glycosuria are excluded from this SOC and can be found in the related 'disorder' SOCs (e.g., PT hyperosmolar state, PT haemosiderosis, PT orthostatic proteinuria, and PT renal glycosuria).

- *Neoplasms benign, malignant and unspecified (incl. cysts and polyps)*

This SOC is physically classified, with pathologic sub-categories for staging benign and malignant neoplasms.

- *Product issues*

This glossary defines terminology related to product quality, gadgets, manufacturing quality systems, product supply and distribution, and counterfeit goods.

- *Social circumstances*

The objective of this SOC is to provide a logical grouping for those aspects that may provide insight into personal concerns that may have an impact on the reported occurrence. This SOC, in essence, contains information about the individual, not the unfavorable occurrence. This SOC contains phrases like PT drug addict and PT death of family, for example.

- *Surgical and medical procedures*

Only terms related to surgical or medical procedures are included in this SOC. This SOC is more of a “support” SOC for recording case information and developing inquiries due to its nature.

- *Infections and infestations*

This SOC only gives location-based information on infectious diseases, not specific targets.

- *Psychiatric disorders*

Due to being too wide and/or broad, the following high-level general terms and high-level terms were removed from this specific SOC. Depressed mood disorders and disturbances, eating disorders and disturbances, impulse control disorders not elsewhere classified (NEC), manic and bipolar mood disorders and disturbances, personality disorders and disturbances in behavior, psychiatric disorders NEC, suicidal and self-injurious behavior NEC, paraphilias and paraphilic disorders, and sexual and gender identity disorders NEC were among the terms used.

3.2.5 - Drug – Target databases

The data regarding drug–protein relationships is mined from two specific databases: Drug-Target Commons (DTC) database (<https://drugtargetcommons.fimm.fi>) [37] and STITCH 5.0 [36]. As already mentioned in the introduction chapter, DTC's goal is to provide an open-data platform for a community-driven crowd-sourcing effort to annotate drug–target associations and provide bioactivity information for medications such as IC₅₀, EC₅₀, and potency values. DTC's T-ARDIS version was acquired from <https://drugtargetcommons.fimm.fi> in April 2021. STITCH, the second database considered, contains the majority of drug–target relationships information available, combining multiple sources of data into a composite scoring function. T-ARDIS includes STITCH version 5, which can be found at <http://stitch.embl.de>.

The initial databases were subjected to two filtering procedures to ensure that biologically/therapeutically relevant relationships were identified and that redundant items from the same drug were renamed. At the same time, the Uniprot ID was used to ensure that the targets in consideration were the same in both databases. Although DTC already provides the Uniprot ID for each drug–target pair stored, this information for STITCH entries is obtained programmatically from the Uniprot database [53] using the STRING [29] identifying code.

Only drug–protein interactions with an IC₅₀ (or EC₅₀) of 100 nM or less were taken into account from DTC, obtaining nearly 10,000 drug–target relationships, accounting for 5007 and 1075 different drug and chemical compounds and proteins (as determined by Uniprot IDs). In the case of the drug protein pairs in STITCH are extracted using the provided database scoring function with a cut-off of 0.8. Only associations with scores greater than the threshold were considered, yielding over 6 million from over 42 000 chemical compounds (including licensed drugs) and 7264 distinct proteins.

To avoid duplication, the identified drug entries from both databases are unified using the InChiKey hash descriptors and the drug's standard name. Furthermore, two other filtering procedures were used to increase the reliability of the discovered drug–target association: the first easily removed all targets enriched with the GO terms "drug catabolic" and "drug anabolic" processes; the second, on the other hand, involved the computation of a Tanimoto similarity index for all the drugs targeting the same protein and the removal of one of them above a certain threshold. This filtering step helps in the event of the same drugs with different names but will have direct consequences during the SONG method development in chapter 6.

3.2.6 - Combining different databases increases the coverage of associations

As detailed in the preceding sections, the nature, purpose, and level of curation of the databases used vary greatly. However, once cleaned and standardized, the obtained data can be integrated, allowing access to a massive core of knowledge. The most striking example comes from drug-ADR databases, where data integration revealed a common set of drugs shared by all databases (figure 16).

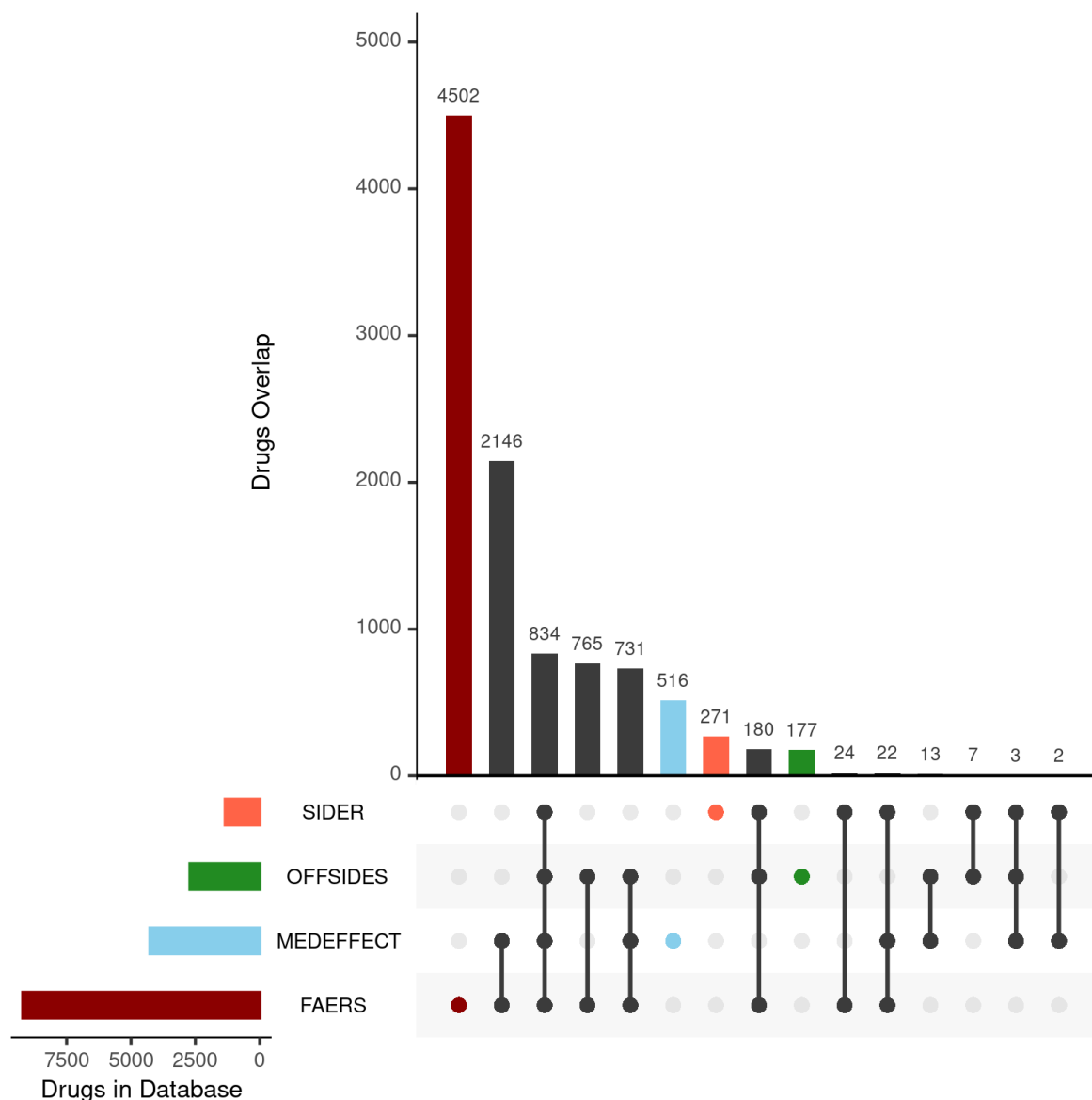


Figure 16. Upset plot showing the overlap between the different databases compiling drug-ADR associations. FAERS, MEDEFECT, OFFSIDES and SIDER are represented as dark red, light blue, green and orange, respectively.

As drug-ADR relationships annotated in OFFSIDES are subsequently added to FAERS on new releases, there is a lot of overlap between the two databases. FAERS and MEDEFECT rely on a variety of sources and real-time reporting systems, and they have the most drugs-ADRs associations and the highest percentage of unique entries. Over 4 million pairwise interactions originating from over 9000 compounds and around 17 000 unique ADR were collected from FAERS through the developed curation approach. From a total of over 4000 and 12 000 drugs and ADR occurrences documented in the

database, 1.5 million drug–ADR associations were discovered in MEDEFECT. As already stated, SIDER and OFFSIDES, unlike FAERS and MEDEFECT, feature manually selected drug and ADR relationships. When compared to the spontaneous reporting databases FAERS and MEDEFECT, these databases offer fewer associations (between 1 and 2 orders of magnitude less). Over 108 000 pairwise interactions were mined for a total of 1344 distinct drugs and 2303 ADRs in the SIDER analysis. OFFSIDES produced a huge number of pairwise drug–ADR associations: 1.5 million from a total of 2708 and 4368 distinct medicines and ADRs, respectively. FAERS and MEDEFECT have a higher percentage of shared medications amongst the databases in terms of uniqueness of information (Figure 16).

The databases describing drug–protein target associations, such as DTC [37] and STITCH [36], were the second set of resources examined. The nature of the two databases is quite different, which is reflected in the number of associations collected from each. After applying the filter outlined in the previous section to DTC, nearly 10,000 drug–target relationships were found, accounting for 5007 and 1075 different drug and chemical compounds and proteins (as per Uniprot IDs), respectively. The number of associations in STITCH was substantially higher: more than 6 million from over 42 000 chemical compounds (including licensed medicines) and 7264 distinct proteins. In terms of shared drugs, the overlap between the two databases was roughly 1600.

3.3 - Statistically validated associations ADR-proteins

Following the standardization and review of all data from the various drug-ADR and drug-target databases, the next step is to integrate it using the shared drug name as a linking element. This will yield a large number of ADR-protein associations, which, as previously stated, contain a large number of non-existent relationships requiring an extensive statistical validation. This problem has already been successfully addressed by Kuhn and colleagues [9], from which this thesis adapted the method.

In a nutshell, the approach computes a contingency matrix for each ADR-protein pair and uses Fisher's exact test to compute the *p-value*. As already explained, the contingency matrix contains the following elements: (i) the number of drugs that present the given ADR; (ii) the number of drugs that bind to the given protein; (iii) the number of drugs that both present the given ADR and bind to the given protein; and (iv) the number of drugs that neither present the ADR nor bind to the given protein. P-values were corrected for multiple testing using the '*q-value*' module in the python package 'MultyPy' due to the large number of relationships. If the computed q-value is equal to or less than 0.05, the ADR-protein association was accepted (see figure 10).

It's essential to note that self-reporting (FAERS and MEDEFECT) and curated (OFFSIDES and SIDERS) drug-ADR data were considered separately. In the case of protein-ADR associations discovered by combining drug-target and drug-ADR (self-reporting), a total of 998 drugs were unequivocally identified on both sets (i.e., drug-target, drug-ADR), producing over 100k statistically significant (i.e., q-value 0.05) protein-ADR associations accounting for approximately 3k and 211 different ADRs and proteins, respectively. In the second set of drugs-ADR databases, the curated set (or non-self-reporting), i.e., SIDER and OFFSIDES, a total of 1135 common drug entities were identified between drug-target, yielding approximately 40k statistically significant protein-ADR associations, including 537 and 194 ADRs and proteins, respectively.

3.4 - Examples of uncovered associations and T-ARDIS benchmarking

T-ARDIS, as I've presented, is a large-scale mining exercise that relies on a fully automated pipeline that explores any potential correlations between proteins and ADRs. Through different methods it has been possible to statistically find and validate significant relationships between protein and ADR utilizing drugs as linking components by combining public databases on drug-protein and drug-ADR associations. In the next section I will present some of the statistical validated interactions found in T-ARDIS and their clinical implication, then I will study T-ARDIS results in the optic of protein-ADRs databases ecosystem.

3.4.1 - Examples of uncovered associations

In vitro studies and literature have both corroborated many of the associations stored in T-ARDIS. Some of the more eye-catching examples are provided below.

- It is known that the anti-inflammatory drug aspirin (acetylsalicylic acid) inhibits the cyclo-oxygenase 2 enzyme found in the stomach mucosa (COX-2 or PTGS₂; Uniprot ID: P35354). [71] Aspirin also inhibits the prostaglandin G/H synthase 1 enzyme (COX-1 or PTGS₁; Uniprot ID: P23219). [71] [72] These secondary interactions have been linked to gastritis and bleeding ulcers in several articles dating back to 1955. [73] [74] Both the PTGS₁ and PTGS₂ proteins have substantial low q-values when it comes to Peptic ulcer and Peptic ulcer hemorrhage ADRs, according to our findings.
- The serotonin norepinephrine reuptake inhibitor Venlafaxine inhibits the sodium-dependent serotonin transporter (SLC6A4; Uniprot ID P31645) [75], which has been linked to sexual dysfunction [75]. SLC6A4 appears to be highly

significantly associated (i.e., q-value 0.05) with a variety of sexual dysfunctions (e.g., ejaculation failure and female sexual dysfunction) in our analyses.

- Another example is the interaction between Budesonide and the glucocorticoid receptor (Uniprot ID: P04150). ADRs related to Budesonide treatment have included respiratory infections, coughs, and headaches in the inhaled form [76], and weariness, vomiting, and joint pains in the oral form [76]. A much rarer condition, Adrenal insufficiency, has also been identified in the case of the long-term use of the oral form of budesonide [77] without a specific etiology. T-ARDIS, on the other hand, associates this ADR with the previously reported glucocorticoid receptor with a highly significant q-value. Furthermore, given its importance, the relationship between glucocorticoids and Adrenal insufficiency is a hot topic in the current relevant literature. [78]
- The activation of the 5-hydroxytryptamine receptor family by zolmitriptan (HTR_{1A}, HTR_{1B}, and HTR_{1E}; Uniprot IDs: P08909, P28222, and P28566, respectively) has been linked to hyperesthesia [79]. The association between these proteins and hyperesthesia was all significant in our study, with q-values of 0.0001, 0.006, and 0.02 for HTR_{1A}, HTR_{1B}, and HTR_{1E}, respectively. It is worth noting that Kuhn et al. discovered and validated this association in vitro [9].

3.4.2 - T-ARDIS benchmark

In a more general framework, the degree of congruence and complementarity between ADRs and proteins discovered in T- ARDIS were compared to prior studies. Four different datasets have been used to compare the connections discovered in T-ARDIS. The first group came from the ADReCS-Target database [6] and consisted of 1710 protein-ADR top scoring interactions. The second collection comes from Smit et al.'s recent work [80], which used an older release of SIDER (ver.3) to extract around 2000 protein-ADR interactions. The third set is based on a set of 225 pairwise interactions that were validated in Kuhn et al.'s work [9]. Finally, the fourth group, which comprises

816 protein–ADR relationships, was manually curated from scholarly literature and reported in the study by Kuhn et al. [9]. This analysis proved that regardless of whether significant or not, the total representation of target–ADR relationships described is low (table 5).

Table 5. Comparison of different datasets and T-ARDIS

SET	# Associations	Self-reporting ^a	Curated ^b
Associations mined from the literature in Kuhn et al. [9]	224	27 (4)	17 (6)
Associations validated in vivo in Kuhn et al. [9]	2170	115 (69)	113 (85)
Associations described in Smit et al. [81]	2153	340 (48)	297 (167)
Associations from ADReCS database [6]	816	171 (14)	87 (11)

^a Associations present in the self-reporting set of T-ARDIS; significant associations shown within parentheses (q-values < 0.05).

^b Associations present in the curated set of T-ARDIS; significant associations shown within parentheses (q-values < 0.05).

For example, in the self-reporting and curated sets of T-ARDIS, only 12 percent and 8% of the target–ADR connections mined from the literature are reported, respectively. Overall, the self-reporting set's values vary from 20% to 5%, while the curated set's values range from 8% to 5%. There are possible explanations for these low results. On the one hand, the lack of target–ADR correlations in T-ARDIS could be attributed to the fact that no safety concerns have been identified in self-reporting (FAERS, MEDEFECT) or curated databases (OFFSIDES, SIDER). It's also possible that no link exists between the specified medicine and the target in either of the two databases used in this study: DTC and STITCH. Finally, when collecting and merging the databases required to compute T-ARDIS, one of the possible reasons could be the robust and stringent filtering method applied, as explained in the previous section. As a result, the drug–ADR and/or drug–target associations may exist but fail to pass the filtering

processes. In any case, these findings demonstrate T-ARDIS' complementary nature to other current resources in the field, allowing for a more thorough and fuller picture of target-ADR relationships (figure 17). Indeed, these analyses revealed some intriguing findings, such as the particular trend of ADRs associated with single proteins. Figure 17 clearly demonstrates that the number of ADR associated with a given protein target varies, but in most cases, the number of ADR associated with proteins is low, both in self-reporting and curated datasets. The number of ADRs associated with a given target is proportional to the number of drugs identified to target the given protein; as the number increases, so does the number of ADRs, with a clearer trend in the case of the curated dataset-ARDIS also allowed the identification of unusual proteins linked to a large number of ADRs. Interleukin-8 (Uniprot ID: P10145), endothelin-1 (Uniprot ID: P05305), and leptin (Uniprot ID: P41159) were shown to be related with 1532, 933, and 717 ADRs, respectively, in the protein-ADR connections discovered from the self-reporting dataset. The 5-hydroxytryptamine receptor 2C (Uniprot ID P28335), the 5-hydroxytryptamine receptor 1A (Uniprot ID: P08908), and the alpha-2A Adrenergic receptor (Uniprot ID: P08913) are the top three proteins in the curated dataset, with 119, 104, and 98 linked ADRs, respectively (Figure 17). This high number can be explained by the biological role that these proteins play. For example, leptin is linked to more than 150 biological processes (according to GO classification), spanning from signal transduction (GO:0007165) to autophagy regulation (GO:0010507).

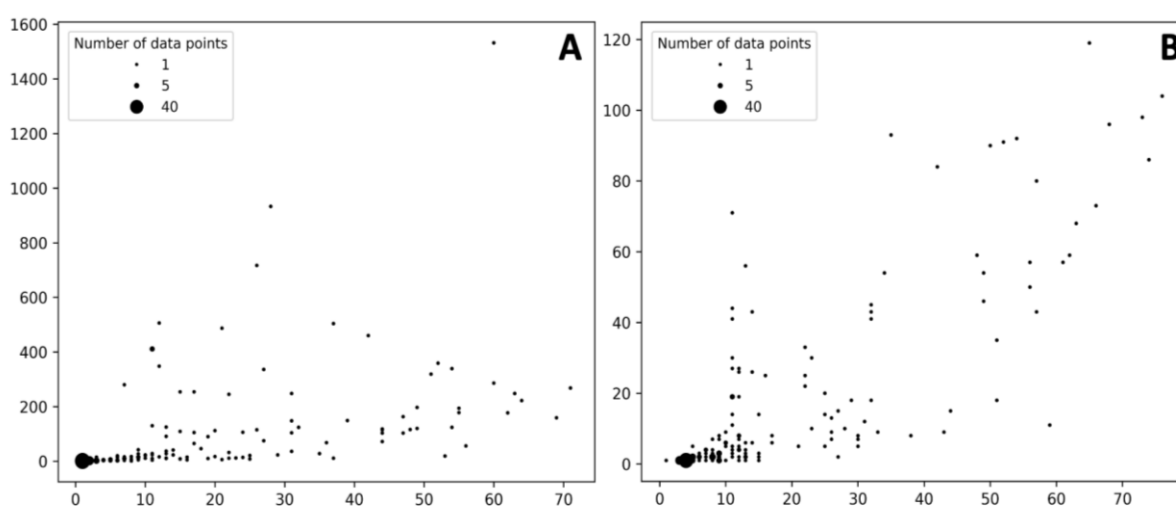
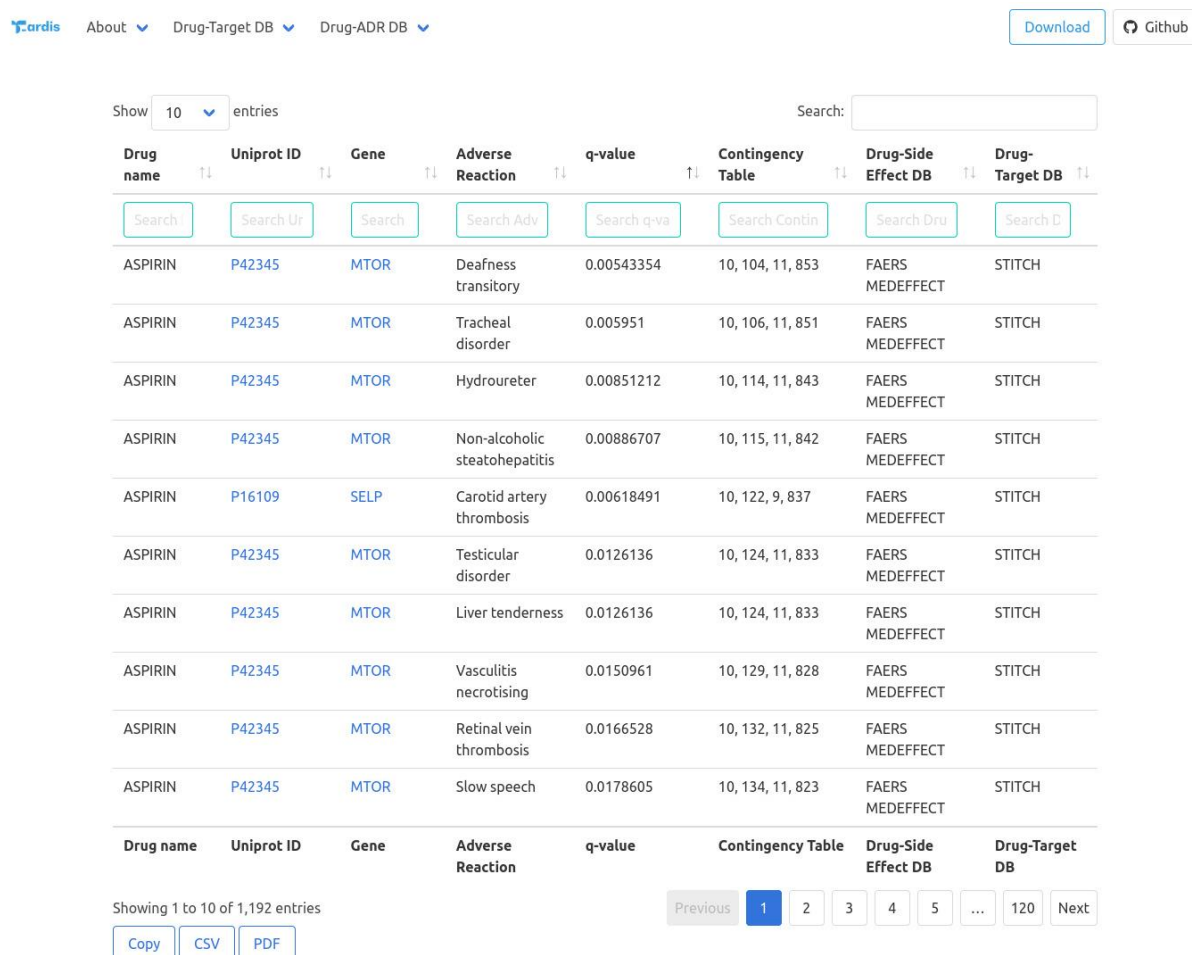


Figure 17. Bubble plots showing the number of drugs per protein (X axis) vs number of statistically significant ADR per protein (Y axis). (A) Distribution of the self-reporting set; (B) distribution of the curate set.

3.5 - Accessing and querying T-ARDIS

All associations between drugs and proteins, including the original sources, have been deposited and compiled in a biological database: T-ARDIS. T-ARDIS can be found at <http://bioinsilico.org/T-ARDIS> providing a quick and easy access to information, including the ability to search and filter associations based on customized queries. The database can be searched by protein name (Uniprot ID or gene name), drug name, or ADR name. The tables that result provide information on the protein-ADR association as well as the q-value of the association and parent databases, both drug-protein and drug-ADR (Figure 18).



Search results for Aspirin:

Drug name	Uniprot ID	Gene	Adverse Reaction	q-value	Contingency Table	Drug-Side Effect DB	Drug-Target DB
ASPIRIN	P42345	MTOR	Deafness transitory	0.00543354	10, 104, 11, 853	FAERS MEDEFFECT	STITCH
ASPIRIN	P42345	MTOR	Tracheal disorder	0.005951	10, 106, 11, 851	FAERS MEDEFFECT	STITCH
ASPIRIN	P42345	MTOR	Hydroureter	0.00851212	10, 114, 11, 843	FAERS MEDEFFECT	STITCH
ASPIRIN	P42345	MTOR	Non-alcoholic steatohepatitis	0.00886707	10, 115, 11, 842	FAERS MEDEFFECT	STITCH
ASPIRIN	P16109	SELP	Carotid artery thrombosis	0.00618491	10, 122, 9, 837	FAERS MEDEFFECT	STITCH
ASPIRIN	P42345	MTOR	Testicular disorder	0.0126136	10, 124, 11, 833	FAERS MEDEFFECT	STITCH
ASPIRIN	P42345	MTOR	Liver tenderness	0.0126136	10, 124, 11, 833	FAERS MEDEFFECT	STITCH
ASPIRIN	P42345	MTOR	Vasculitis necrotising	0.0150961	10, 129, 11, 828	FAERS MEDEFFECT	STITCH
ASPIRIN	P42345	MTOR	Retinal vein thrombosis	0.0166528	10, 132, 11, 825	FAERS MEDEFFECT	STITCH
ASPIRIN	P42345	MTOR	Slow speech	0.0178605	10, 134, 11, 823	FAERS MEDEFFECT	STITCH

Showing 1 to 10 of 1,192 entries

Navigation: Previous 1 2 3 4 5 ... 120 Next

Download options: Copy CSV PDF

Figure 18. Snapshot of the result page example upon querying by drug “Aspirin”.

The webservice also provides external links to native drug-target or drug-ADR databases, as well as protein-related repositories. Users can further filter the resulting

table by querying specific drug, ADR, or parent databases (e.g., filtering those associations resulting from FAERS). The table can be sorted by q-values to display the most significant associations first and downloaded in a variety of formats (simple copy, CSV or PDF). Finally, from the home page links, bulk downloads of the database and associated scripts to recreate the database are available.

3.6 - Summary

This chapter presented an approach to link adverse drug reactions to drug targets by utilizing various publicly available databases, a thorough curating methodology, and extensive statistical validation. T-ARDIS is a resource that will be useful to drug development researchers in both academia and industry, and, as will be highlighted in the following chapters, will also aid in the advancement and expansion of the underlying theory between proteins and adverse reactions, facilitating the development of a machine learning-based predictor based on such a relationship. This new resource will be known as DocTOR (Direct fOreCast Target On Reaction) and will be thoroughly explained in Chapter 5. SONG (Side Effect On Network Graph), a study on the inter-relationship and co-morbidities on an ADR-ADR network, will be described further on in Chapter 6.

4 - PREDICTING TARGET- LIABILITIES USING NETWORK- BASED ANALYSES: DocTOR

4.1 - Abstract

As presented in the previous chapters, Drug discovery attrition rates, particularly at advanced clinical trial stages, are high due to unexpected adverse drug reactions (ADRs) elicited by novel drug candidates. So not only identifying, as with T-ARDIS, but also predicting undesirable ADR produced by the modulation of certain protein targets would contribute to developing safer drugs, thus reducing economic losses associated with high attrition rates. In this chapter I will describe a target-centric approach to predict relationships between protein targets and ADR, rather than the more usual drug-centric methods, named DocTOR (Direct fOreCast Target On Reaction).

To this purpose, various machine learning classifiers such as the Support Vector Machine (SVM), Random Forest (RF), and Neural Networks (NN) were evaluated. It should be noted that different classifiers were developed for each adverse reaction. The classifiers, in particular, are not generic predictors of a protein eliciting any ADR, but rather a specific ADR. As a result, the predictions are tailored to each individual adverse reaction and thus have unique properties. Given this, all of the models developed based their training data on the T-ARDIS database, which was used to extract the highly significant connections between proteins and ADR.

The features used to train and predict are eight different topological-based features:

- I) The GUILDify network diffusion-based score.
- II) Several network-based clustering algorithms.
- III) A functional similarity index
- IV) Network distance to proteins used in preclinical drug development safety panels
- V) Network descriptors in the form of degree and betweenness centrality measurements and conservation.

In some way, all of the measures rely on network-based data, and so include elements that are fundamental not only to the protein, but also to the network. As a result, the proteins are framed within the interactome, and the impact of modifications on nearby proteins is evaluated. Specific models were created for each individual adverse reaction as well as clusters of ADR within the same system organ class (SOC), allowing the analysis to be expanded to a more general anatomical or physiological system, according to the MEDDRA nomenclature.

To assess prediction's reliability, the obtained models were tested against independent datasets, including manually curated sources obtained from literature and data submitted to the Critical Assessment of Massive Data Analysis (CAMDA) competition, in addition to the corpora derived from T-ARDIS benchmarking. Finally, the accuracy of a meta-predictor that integrates the predictions of each unique classifier is investigated. Based on how the predictions were combined, three different meta-predictors were developed and evaluated: (I) *a jury vote* system, (ii) consensus method, and (iii) *red flag* method.

In the next sections I will expose in detail the data extrapolation and feature computation, how the different machine learning methods were implemented to obtain the highest reliability possible on the predictions and how I combined the results to develop the meta-predictors methods. A schematic representation of the overall process is depicted in Figure 19

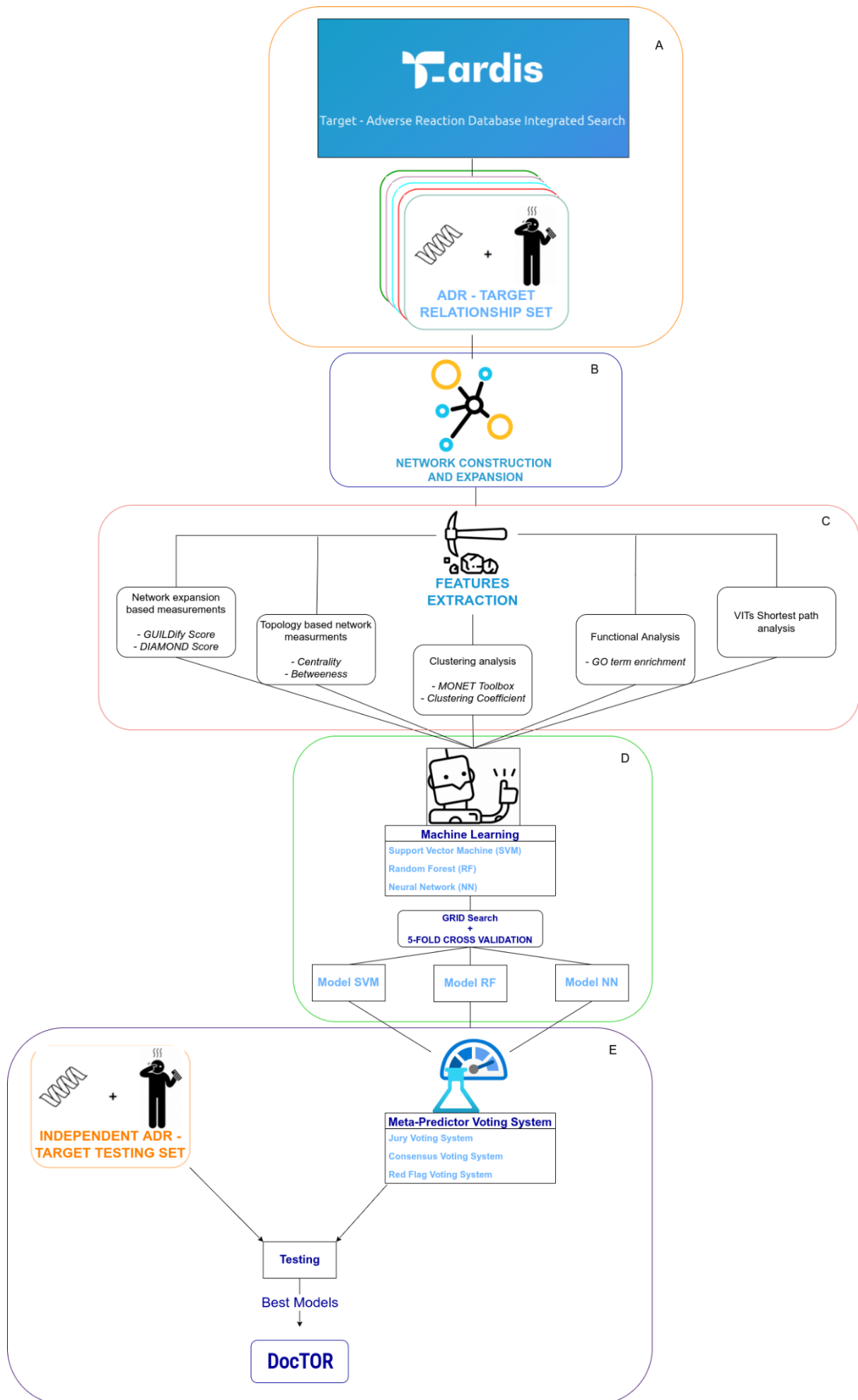


Figure 19. Schematic depiction of feature extraction, training and testing procedures. Panel (A) indicates the process of extraction of the training dataset from T-ARDIS [11]. Panel (B) indicates the process of network expansion of targets extracted in (A) using GUILDify [25]. Panel (C) summarizes the process of computation of different input features. Panel (D) represents the development of machine-learning classifiers. Finally, Panel (E) illustrates the development of the meta-predictors together with the testing of the classifiers and consensus functions on the independent dataset.

4.2 - Data Extrapolation and Features computation

4.2.1 - ADRs considered for model construction

T-ARDIS represents the cornerstone of this project, providing the data set for training and cross-validating the models. As exposed in chapter 4, T-ARDIS is a database that compiles statistically significant proteins-adverse reaction interactions. The data contained in the database is divided into two distinct groups: relationships derived from self-reporting databases, such as FAERS [31] and MEDEFFECT [32] which contains around 17k paired protein-adverse reaction interactions, and relationships derived from curated databases, such as SIDER [34] and OFFSIDES [35], which reports around ~3k pairwise associations. Given this vast body of knowledge, creating single models for each adverse reaction in the database is an impossibly time-consuming and computationally demanding task.

As a result, 84 distinct ADR were selected as a representative subset, covering the entire spectrum of SOC classes and yielding 434 unique relationships (figure 20).

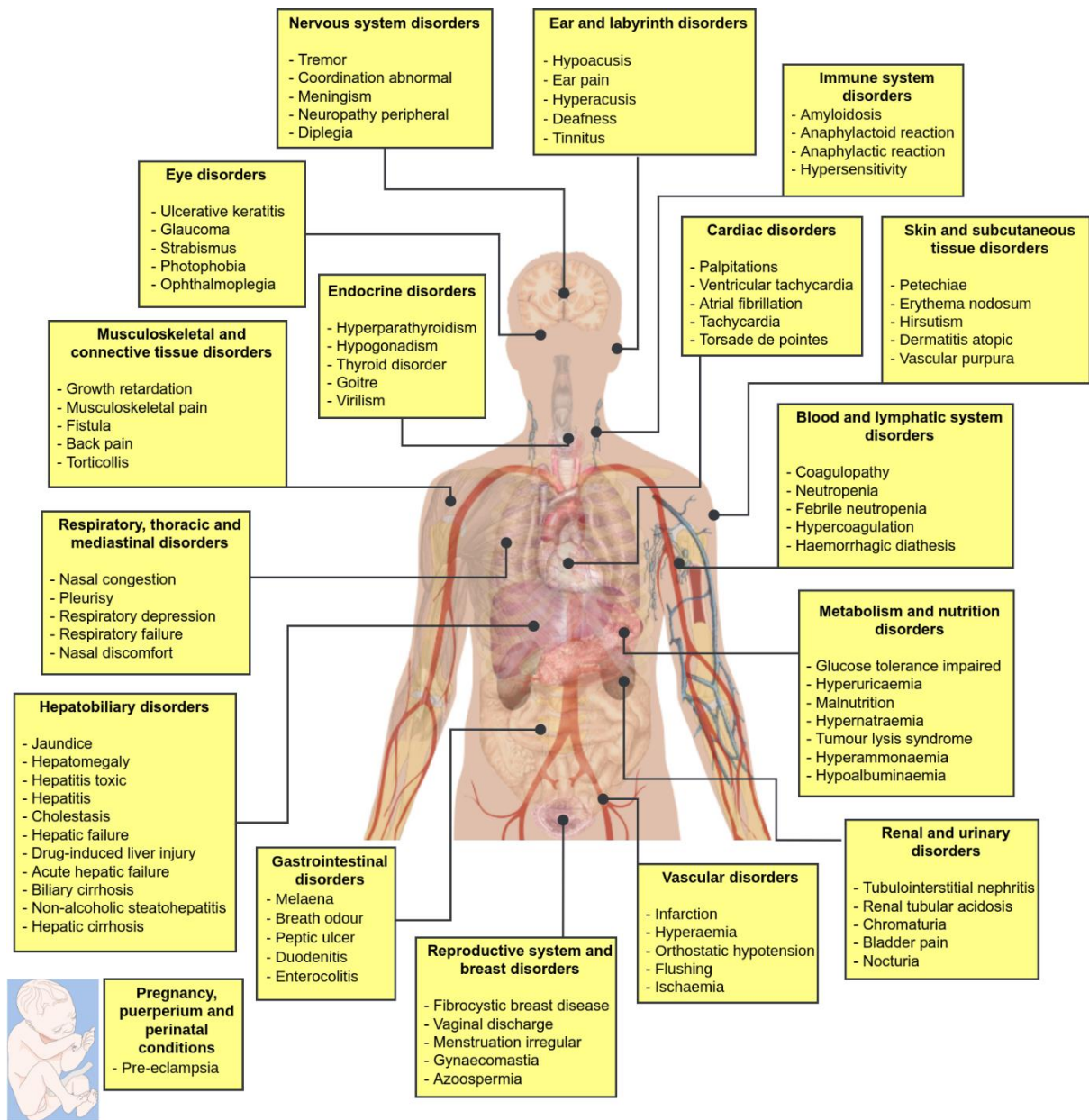


Figure 20. List of selected ADR by System Organ Class.

The initial plan set to select at least 5 ADR for each SOC registered in T-ARDIS, but this was not always possible, such in the case of Congenital, Familial, and Genetic Disorders and Pregnancy, Puerperium, and Perinatal Conditions SOCs (Figure 20). The low incidence of ADR in these two specific SOCs is directly related to their rarity in T-ARDIS. As one might expect, ADR being part of these classes can be extremely dangerous, and drugs that might cause them are unlikely to survive phase III [82]. On the other hand, given the wealth of independent information and recent clinical interest regarding hepatobiliary disorders (see below), it was reasonable to expand the ADR under investigation of this specific SOC.

The ADR have been selected based on their number of proteins association, coverage for SOC and presence in external adverse reaction-target databases.

The total number and types of ADR selected was also determined by the information available in external independent protein-adverse reaction databases. This is especially relevant since the data mined from independent sources will be required for later benchmarking. The databases used to determine the ADR subset are identical to those used for T-ARDIS benchmarking. As previously stated, these include the in vitro interactions derived from Kuhn et al. [9], the dataset from Smit et al. [81], Sayaka et al. [83], the ADRECs-Target database [6], and, as a novel source of information, the DisGeNet Drug-induced Liver Injury dataset. The latter, in particular, addresses a subset of drug-induced liver injuries comprising 12 distinct MEDDRA-defined events ranging from "Acute hepatic failure" to "Non-Alcoholic Steatohepatitis." A total of 15k interactions were mined by integrating the over 600 different adverse events and 428 proteins from this compendium. The final ADR were chosen solely on the basis of scientific and clinical interest, yielding the 84 ADR already mentioned, which are associated with 188 proteins not found in T-ARDIS and thus conforming to an ideal independent test-set.

Defining these associations is only the tip of the iceberg. Aside from their critical importance in drug discovery, single adverse reaction-protein relationships may not contain a lot of information. Indeed, the onset of ADR and diseases, like functions, is

mediated by a series of molecular interactions between proteins, rather than a single one [23].

As a result, a broader network perspective is required. By mapping the ADR associations discovered on the human PPIN (protein-protein interaction network), the protein's topological characteristics, that may define the onset of a specific adverse reaction, can be identified. As a result, the primary focus of this study becomes the extrapolation of adverse reaction-related protein features via network analysis.

It is critical at this point to define the protein network that will be used to map and analyze such information. BIANA [84] and GUILDFifyv2 [25] were used to integrate the human interactome data used in this study. The data contained in BIANA, in particular, were obtained by integrating interactome data from the IntAct [85], DIP [80], HPRD [86], BioGrid [87], MPACT [88], and MINT [89] databases, resulting in one of the most complete maps of the PPI landscape [84]. The most recent version included 13,090 proteins (or nodes) and 320,337 connections (or edge).

4.2.2 - Features considered for prediction.

The extensive network analysis of the human PPIN yielded eight distinct features that were used to characterize the associated proteins of the 84 ADR chosen. The selected features try to exploit all the possible topological characteristics, from message passing to shortest path between nodes (Figure 21).

Namely the features employed are:

- GUILDFify Score
- Centrality degree and Betweenness centrality measures
- Clustering based algorithms
- A functional conservation score
- Shortest path to in vitro ADR identified proteins (VIT - Very Important Targets)

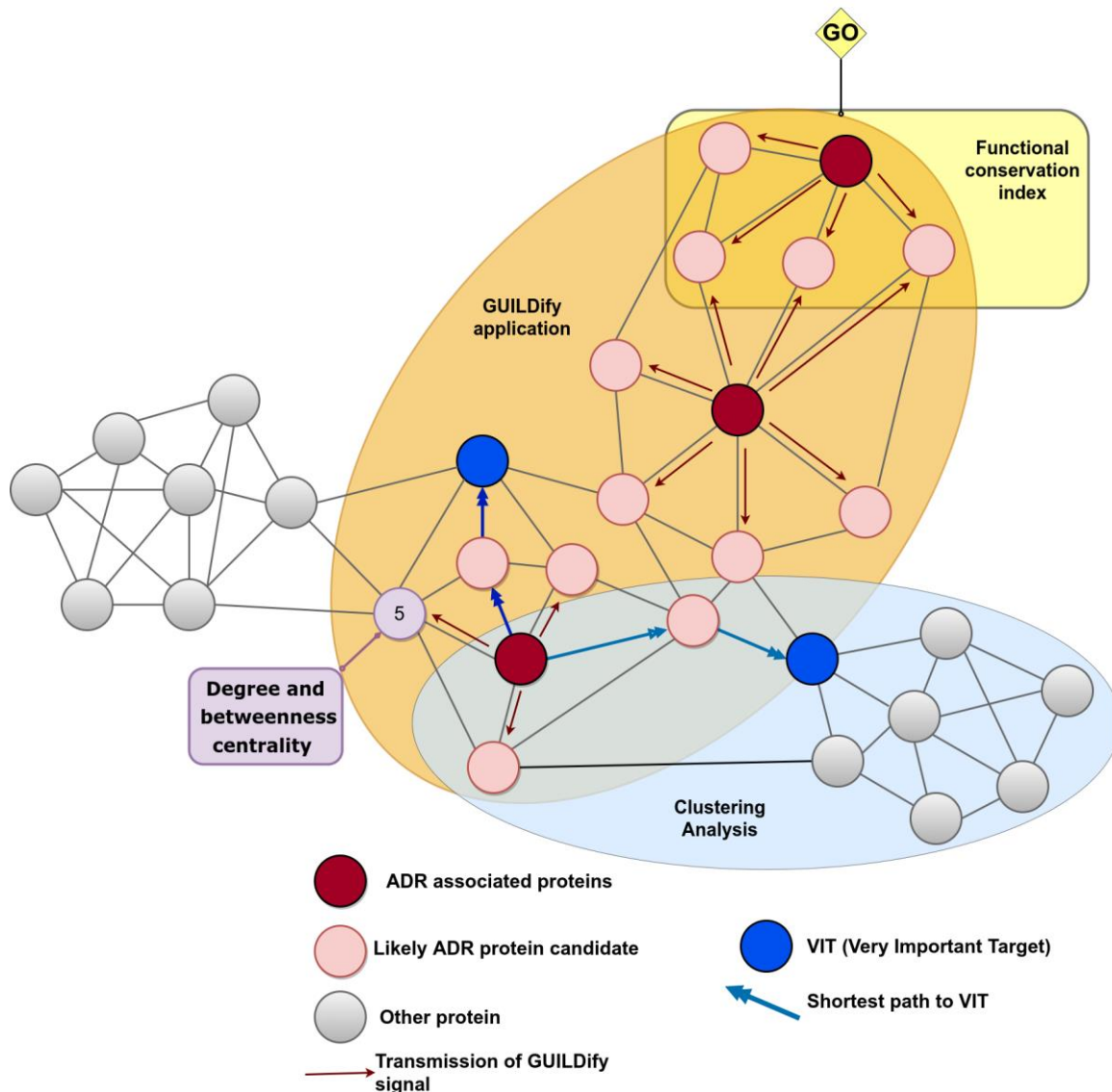


Figure 21. Network Feature extrapolation. Visualization of network feature extrapolation. application of GUILDify message passing method (red nodes and arrows), study of clustering analysis (cyan), degree centrality and betweenness centrality computation (purple), GO enrichment analysis (yellow), shortest path to VITs (blue nodes and arrows).

4.2.2.1 - GUILDify score

GUILDify is a web service that hosts network diffusion-based algorithms that can be utilized in a variety of network medicine applications [25]. GUILDify message-passing algorithms [25] send a signal from a collection of proteins linked with a phenotype or drug (known as seeds) to the rest of the network nodes and grade them based on how

quickly the message reaches them while taking many network features into account. GUILDiFy was originally created to prioritize gene-disease correlations and discover disease modules [25] [90], but it has lately been used to uncover disease comorbidities and medication repurposing possibilities [91]. In this study, GUILDiFy was used to predict protein-adverse reaction correlations. A GUILD score was assigned to each protein in the interactome upon expansion based on the adverse reaction's associated protein used as seed. The higher the score, the more likely there is a connection between the protein and the seeds used to expand it.

4.2.2.2 - Degree and betweenness centrality.

Degree and betweenness centrality are two of the network analysis metrics used as features in this research. As already mentioned in chapter two, the betweenness centrality can be defined as the number of times a node acts as a bridge along the shortest path between two others, whereas the degree centrality is defined as the number of edges connecting to a node. In terms of the interactions between proteins, both measurements indicate how significant a node is within a network, as well as how likely a protein is to be part of a signal cascade and engage in the same biological process. The degree and betweenness centrality values were calculated using NetworkX [92].

4.2.2.3 - Clustering-based algorithms

A further representation of the "guilt-by-association" theory is the interpretation of "disease module," which is a neighborhood of a molecular network whose components are all associated with one or more diseases or risk factors. Disease modules, as demonstrated, can be used to identify proteins/genes associated with specific diseases [23]. The assumption in the context of ADR is that proteins linked to the same ADR will cluster in local regions of the interactome, forming adverse reaction modules. Two different clustering algorithms were used to identify these modules.

The first method was the K₁ clustering algorithm based on the Diffusion State Distance (DSD) metric [29]. The DSD metric is used to define a pairwise distance matrix between all nodes, which is then used by a spectral clustering algorithm. Using standard graph techniques, dense bipartite subgraphs are identified in parallel. Finally, the results are combined into a single set of 858 non-overlapping clusters.

The second clustering method is based on the work of Lefebvre and colleagues [93], and is based on modularity optimization, assigning and removing nodes recursively to the modules discovered, each time evaluating the loss or gain of modularity. 46 unique modules were extracted from the interactome through this methodology. Finally, using the NetworkX utility, the "clustering coefficient" for each node in combination with the clustering approaches mentioned above was computed [92].

4.2.2.4 - A function conservation index

The use of Fisher's exact test to identify enriched Gene Ontology (GO) functions among top ranking proteins is a new feature in the current version of GUILDify [25]. The function conservation index, which makes use of this relevant data, determines how functionally comparable a protein is to the enriched GO terms subnetwork computed by GUILDify. In a nutshell, this value is the result of calculating the Hamming distance between two binary vectors indicating the presence or absence of a specific GO term, one for the single protein and one for the subnetwork. This information can be interpreted as the contribution of single proteins to the enriched network's function, and thus how much is responsible for the onset of the Adverse Event if perturbed. Mathematically the Hamming distance is expressed as a ratio, with 1 indicating total function overlap between the single protein function and the enriched network function.

The following example illustrates this index. Given proteins A, B, and C that are linked to a specific ADR. Protein A is enriched with four different Molecular Function GO

terms, according to Gene Ontology results. The underlying network, derived from the expansion of protein A, B, and C, has been enriched with two different molecular function GO terms, of which only one is present in protein A. These data can be represented as two binary vectors indicating the presence or absence of specific GO terms (Table 6).

Table 6. Vector representation of GO terms

	<i>GO term₁</i>	<i>GO term₂</i>	<i>GO term₃</i>	<i>GO term₄</i>	<i>GO term₅</i>
<i>Protein A</i>	1	1	1	1	0
<i>Protein B</i>	0	0	1	0	1
<i>Network enrichment</i>	0	0	1	0	1

$$\{ProteinA = [1, 1, 1, 1, 0] \text{ Network Enrichment} = [0, 0, 1, 0, 1]$$

$$\{ProteinB = [0, 0, 1, 0, 1] \text{ Network Enrichment} = [0, 0, 1, 0, 1]$$

We can now compute the Hamming distance between these two vectors, which is defined as their XOR operation, $ProteinA \oplus NetworkEnrich$. The final value is computed counting the total number of 1s in the resultant vector and then expressed as the proportion of values that are the same, in this case 0.2. As evident, the same operation performed with *ProteinB* yields result 1.

Programmatically, the Hamming distance has been computed with the SciPy function `spatial.distance.hamming`

4.2.2.5 - Shortest path to Very Important Targets

Safety panels include proteins and pathways that are well established as contributing factors to clinical ADRs representing the bare minimum of targets that qualify for early hazard detection, off-target risk assessment, and mitigation. These specific proteins can be accessed via the EuroFins Discovery Safety Screen Tier 1 panel, thanks to the work of

Whitebread and colleagues [94]. The panel is composed of 48 proteins renamed in this research as Very Important Targets (VITs). As one may expect, many of the VITs represent network hubs, or proteins that have particular relevance in critical biological functions. Indeed, the location of T-ARDIS' adverse reaction associated proteins in the human interactome in relation to VITs may provide important insight into the twos' relationships.

The VITs have been mapped in the interactome using the NetworkX utility. At the same time the shortest path distance between each of the proteins in our training set and any VIT has been computed [92]. As a representative distance between the T-ARDIS proteins and the VITs, the value of the first quartile was taken from the overall distribution of shortest path distances of any given protein.

4.3 - Machine Learning implementation

The method for predicting protein-adverse reaction relationships will be described in this subchapter. The strategy proposed, as previously stated, is a network-based application, which means it is focused on a topology-oriented collection of eight metrics generated for each protein and used as inputs to machine learning classifiers. The three types of classifiers used were Support Vector Machine (SVM), Random Forest (RF), and Neural Networks (NN). Specific models were trained and tested for each of the 84 ADR, as well as models at SOC, which group ADR belonging to the same SOC.

4.3.1 - Positive and negative sets definitions

The positive set, i.e., proteins associated with each of the 84 ADR considered, was obtained from the T-ARDIS database [11]. Since the number of positive examples per adverse reaction are typically low, the positive set was augmented using the concept of close connection. To that end, the DIAMOnD score [26] was calculated for the subnetworks associated with the positive set. The proteins forming the new obtained subnetwork (i.e., the closest to each other) were then sorted, and those with a DIAMOnD score greater than an arbitrary threshold were chosen to form the positive set. Multiple DIAMOnD threshold scores, namely at 0.6, 0.7, 0.8, and 0.9, were tested to obtain the best result during the training phase. The findings of this comparison can be found in the thesis supplementary materials [10].

Each of the ADR under consideration has its own negative set defined. The DIAMOnD criterion was used again, this time by randomly selecting proteins with scores lower than the chosen positive threshold. All of the negative sets produced in this fashion contain proteins with DIAMOnD scores close to zero, indicating that they are completely unrelated to the associated adverse reaction subnetwork (Figure 22).

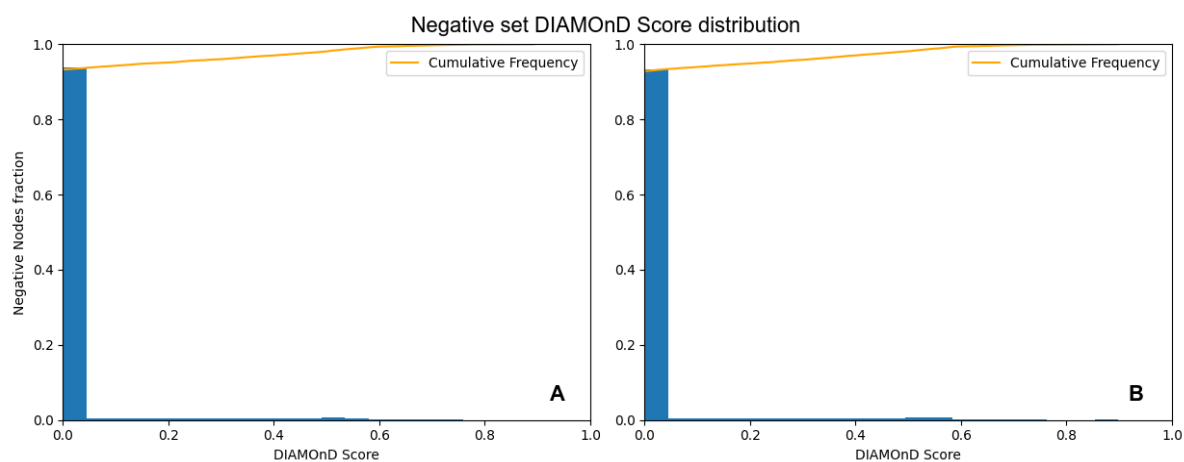


Figure 22. DIAMOnD Distribution for the Negative sets in the case of T-ARDIS self-reporting (A) and T-ARDIS controlled (B) datasets. The distribution is considered for all the accumulated ADR.

This is simply explained by the fact that the adverse reaction positive subset only accounts for a small portion of the overall human interactome. Also, for this reason, during the training and testing phases, various positive:negative case ratios were evaluated. Indeed, in addition to using a balanced training set, that is, an equal number of positive and negative instances, alternative ratios for training and testing the models such as 1:1.5, 1:3, and 1:5 (positives:negatives) were investigated. In parallel with the DIAMOnD negative distribution, a simple topological study was also carried out to assess the independence of the generated negative sets and their associated positives. The obtained distribution confirmed that the majority of negative proteins are at least three jumps apart from the positive subsets. (Figure 23)

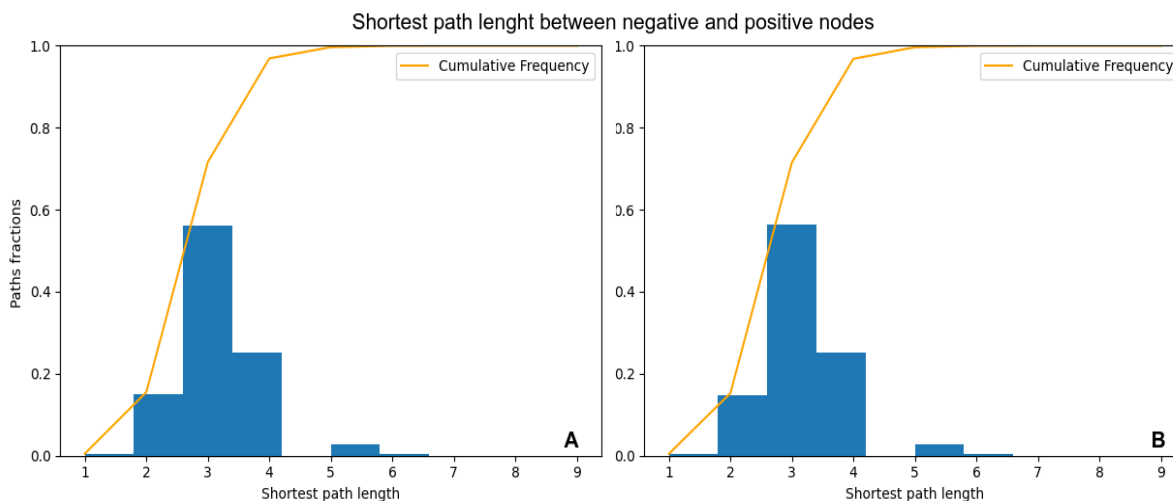


Figure 23. Distribution of negative node shortest path. The shortest path is computed taking in consideration as source the negative set nodes and as target the adverse reaction associated positives in the case of T-ARDIS self-reporting dataset (A) and T-ARDIS curated dataset (B)

4.3.2 - The predictive performance of individual features in the Self-reporting and curated datasets

As input features for the classifiers, eight different variables were considered as described in the previous sections. GUILDify scores, network topology (degree and betweenness centrality values), a function conservation score, module imputations, and distances to proteins in safety panels are among them. In Figure 24 it is shown the distribution of the different features for the positive and negative sets. As already mentioned, the positive cases (negative cases were selected randomly) were extracted from the T-ARDIS database [11] both for the self-reporting and curated sets. The data shown in Figure 24, in particular, derives from the self-reporting set of T-ARDIS.

Starting from the GUILDify rankings, positive and negative nodes present an evident overlap, but the positive sets have higher scores and a slightly skewed distribution toward high values (Figure 24A). The analysis of centrality-based features also shows a significant overlap between positive and negative sets, though positive sets have a more skewed distribution towards higher values, especially in the case of betweenness values (Figure 24B-C).

When quantifying function analysis as distance to enriched function(s) of the set (Figure 24D), the proteins in the negative set have larger distances, i.e., no shared functions with the GUILDify enriched GO terms, than those in the positive set. In fact, the majority of proteins with a value of 1.0 correspond to proteins in the positive set, while those with lower values, i.e., no shared GO terms, tend to be proteins in the negative set. However, it is fair to say that the overlap is substantial.

The tendency of functionally and disease-related proteins in the interactome to be close (i.e., shorter distances) was also considered as a feature for the prediction. This aspect was investigated, as described in the previous section, by using clustering algorithms to identify modules in the entire interactome where proteins associated with the same or similar ADR are grouped. If the number of modules needed to represent a given collection of proteins in an adverse reaction is small, the proteins are likely to share modules. Similarly, the presence of a large number of modules indicates that the proteins do not belong to the same cluster.

The K1 algorithm [29] identified 1170 different clusters, many of which were composed of three proteins, the smallest amount required to define a module (Figure 24E). As shown, proteins in the positive set have fewer clusters, implying that proteins associated with ADR tend to belong to a small number of clusters rather than being dispersed throughout the interactome. Similarly, the Louvain-Newman method [93], which grouped the entire interactome into only 95 distinct clusters, allowing for larger module analysis, demonstrated a similar distribution as K1, i.e., the positive set is drawn towards lower values (Figure 24F). Finally, in the Clustering Coefficient Analysis (Figure 24G), both negative and positive sets have the same value distribution. As a result, this feature does not appear to distinguish between positive and negative cases on the adverse reaction.

The distance of given proteins to so-called VITs was the final metric considered as an input variable (see previous section). The distance was calculated as the shortest path (i.e., the fewest number of links) to any given protein in the panel, using the first quartile value after computing all the distances all vs. all (protein in the given adverse reaction and proteins in the panel).

Once again, the distribution of values varies depending on whether the proteins are in the positive or negative sets (Figure 24H). While the most common distance is 2.0, only proteins in the positive set have values less than 2, indicating that proteins in the positive set are closer to proteins considered critical according to pharmacological profiling.

The analysis of the individual features already shows some promising behaviors with individual metrics able to differentiate between both the positive and negative case. While individual features are informative, the predictive power could be boosted by combining them with a machine-learning classifier. In the next section I will present how the different machine learning methods have been implemented and how, for each one of the 84 ADR, 12 different models have been obtained by the combination of positive threshold and Negative Ratio for a total of 1008 trained models.

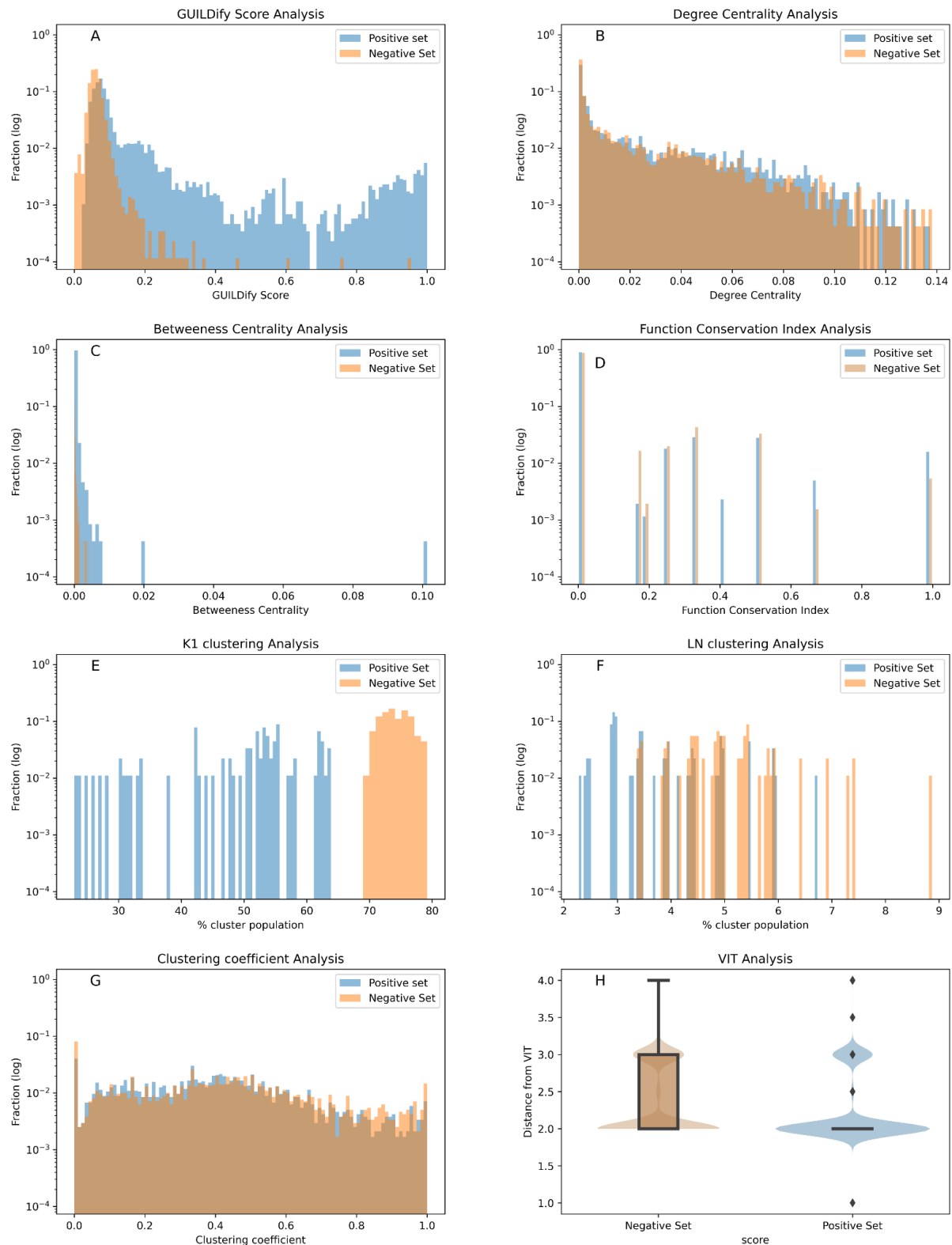


Figure 24. Distribution plots of 8 different input variables used by classifiers. The values of the positive and negative sets are shown in blue and red respectively in panels (A) to (G). Panels (A), (B), (C), (D), (E), (D) and (G) show the distribution of GUILDify scores, centrality values, betweenness values, function score, % of clusters K₁, % of clusters LN, and clustering coefficient values respectively. Panel (H) presents the

box-plots and a violin representation of the distribution of the shortest path values on the negative (orange) and positive (blue) sets.

4.3.3 - Features vectorization and model construction

Support Vector Machine (SVM) with nonlinear kernel (radial basis function - RBF), Random Forest (RF), and Neural Networks were the three machine-learning (ML) classifiers employed in this study (NN). All three approaches were trained using both the positive and negative sets described above. Particular ML libraries were used to implement the classifiers used in this project. The first is Scikit-learn [95], a free machine learning software library for the Python programming language. This package has been of particular use as it includes a variety of classification, regression, and clustering algorithms, such as support-vector machines and random forests in our case. The version used for this project is 1.1.0. The implementation of neural networks required the Keras [96] and TensorFlow [97] packages. Keras is a Python API for the development of artificial neural networks and acts as front end for the TensorFlow library. TensorFlow is a free and open-source machine learning and artificial intelligence software library. It can be used for a variety of tasks, but it is most commonly used for deep neural network training and inference. The versions used for this research are the 2.9.0 for both libraries.

The feature vectorization (i.e., the preparation of training data in a suitable format for the ML functions) has been performed using the Pandas python package [98]. Pandas is a software library for data manipulation and analysis, written in the Python programming language. This package has been especially useful for data structure operations such as manipulating and parsing numerical tables, making it indispensable for the analysis of large datasets such as this one. The Pandas version used is the 1.4.2. The feature vectorization process, as previously exposed, include the selection of adverse reaction positive set, filtering for different threshold of the diamond Score (*DiamondScore* \geq [0.6,0.7,0.8,0.9]) and the random selection of negative values from all the proteins with a diamond score \leq of the positive threshold. The negative has been

also sampled in different ratios with respect to the positive set to increase prediction difficulty (1:1,5, 1:3, 1:5). with the panda's *sample* function.

4.3.3.1 - Support Vector machine details

The support Vector machine was built using a combination of Pandas [98] and Scikit-learn [95]. The Pandas package was used to load and parse the training data into a suitable format as explained above, while the SVC function with an RBF kernel was used for the actual SVM implementation. For the hyperparameters optimization, a 5-fold grid-search cross validation was used. This has been obtained with the *StratifiedKFold()* and the *GridSearchCV()* functions. The *StratifiedKFold()* provides train/test indices to split data in train/test sets and has been applied during the cross-validation process on the training data. The *GridSearchCV()* function instead is used to train a machine learning model with various combinations of training hyperparameters, finding the best combination that optimizes a given evaluation metric.

The best approach to a SVM problem requires the optimization of two main parameters that will be explored in the *GRIDSearchCV()* process. The first one is the C hyperparameters. As mentioned in Chapter 1, the optimization problem that SVM training attempts to solve has two main terms: the first is a regularization term that benefits "simpler" weights, and the second is a loss term that ensures that the weights correctly classify the training data point.

The hyperparameter C represents the balance of importance between these two terms. If the C value is skewed toward high values, the SVM will prioritize the second term, whereas if the C value is low, the SVM will be optimized toward a more general model. In the applied optimization procedure, the C value has been explored in the log space between -4 and 4.

The Gamma Hyperparameter is the second hyperparameter that must be optimized. The Gamma value essentially controls the distance of influence of a single training point and its optimization can be applied only when working with multidimensional kernels

such as an RBF or Polynomial kernel. Low Gamma values indicate a large similarity radius, which results in more points being grouped together; on the other hand, high Gamma values require the points to be very close to each other in order to be considered in the same group. This parameter has a direct impact on the ability to discriminate between points and is crucial in avoiding overfitting. Gamma values have been explored with the SVC inner parameter “scale” and in the log space between -4 and 4.

As with any ML method developed, the use of cross validation in tandem with the defined grid search procedures allowed the best model, as well as the adverse reaction best combination of positive and negative thresholds, to be identified.

4.3.3.2 - Random Forest details

The `RandomForestClassifier()` function was used to implement the random forest models. As before this function is derived from the Scikit-Learn python package. The deployment is similar to the SVM, exploiting the panda’s library for the training set preparation and the `StratifiedKFold()` and `Gridsearchcv()` functions for the cross-validation and grid search processes. In this case the total number of estimators (trees) and the maximum number of features used for each estimator were the random forest hyper-parameters tuned. The total number of estimator tuning is still a point of contention in ML theory, and it is strongly related to the single problem under consideration [99]

Following the examples found in literature [99] [100] the total number of estimators is explored in the linear space between 64 and 128. The total number of features for each estimator are investigated with the “*auto*” inner validation offered by the `RandomForestClassifier()` function or the option without a maximum number (i.e., all the features were used in each estimator).

5.3.3.3 - Neural Network details

As already mentioned, the implementation of Neural Network relied on the Keras and TensorFlow architectures [96] [97]. Again, training data was prepared using the pandas' package, which allows for easy retrieval and differentiation of positive and negative sets. In contrast to the previous two machine learning implementations, which allowed for direct hyperparameter tuning, the Neural Network models necessitated the creation of a dummy function. This function creates an environment in which all NN hyperparameters that need to be tuned can be declared (Snippet 1).

```
def create_model(neurons=1, optimizer='adam', hidden_layers=1):

    model = Sequential() # initialize the model
    model.add(Dense(neurons, input_dim=len(training_df_X.columns),
                    activation='relu')) # add first layer

    for i in range(hidden_layers):
        # Add one hidden layer
        model.add(Dense(neurons, activation='relu'))

    model.add(Dense(1, activation='sigmoid'))
    model.compile(loss='binary_crossentropy', optimizer=optimizer,
                  metrics=['accuracy'])
    return model
```

Snippet 1. Code snippet for the dummy Neural Network function. Implemented in Python3.9 this code declares all the hyperparameters that need to be tuned such as the number of neurons, the optimizer and the number of hidden layers

All the parameters defined in the `create_model()` function will be investigated and enhanced during the grid search procedure. These include the number of hidden layers,

from 1 to 3 and the number of neurons for each layer, from a minimum of 4 to a maximum of 2048. Learning model parameters were also studied including the batch size, with the value of 32 or 64 and the number of epochs, 50 or 100. Due to computational time constraints, neither the optimizer nor the activation function were investigated, instead the implementation relied on the standard SGD optimizer function and the relu and sigmoid activation functions.

4.4 - Performance score implementation

The testing comes after the training phase in any ML approach. The independent testing dataset was obtained from the T-ARDIS benchmarking compendium, as previously stated. This dataset contains 188 proteins mined from external sources which are linked to the same ADR under investigation. Again, no overlap exists between the training and testing sets, implying that none of the proteins extracted externally are included in the training set.

The testing procedure is quite simple; in a nutshell, the proteins associated with each of the 84 ADR will be predicted using the appropriate model, and different evaluation metrics, as discussed in previous chapters, will be implemented based on the results to assess the prediction's quality. During the testing phase, different scores are used to validate the model, as described in sub-chapter 2.7.4. Accuracy, Precision, Recall, Receiver operator curve, and Matthew correlation coefficient are easily implemented using the Scikit-learn package's functions `accuracy_score()`, `precision_score()`, `recall_score()`, `roc_auc_score()`, and `matthews_corrcoef()`. The direct usage is also straightforward; in fact, these functions require only the labels of the testing set (in this case, 0 or 1 depending on whether the proteins are related to the adverse reaction or not) and the predicted label. Based on the evaluation scores obtained for each of the machine learning models implemented, only one model is chosen for each of the ADR under investigation.

4.5 - Single ML methods CV results

As previously stated, each of the models underwent cross-validation during the training phase to assess model reliability prior to the effective independent test. This procedure was also used for hyperparameter tuning, to see how the model scores changed as the metrics varied. The preliminary results of this phase can be used to have an idea of the model's performances. The Receiver Operator Curve AUC value distribution derived from single classifier cross-validation is shown in Figure 25. To determine the best hyperparameters combination, this score was used to rank the various models obtained during the cross-validation and GRID search procedures. The different classifiers appear to perform well, with the median value exceeding the 0.5 random threshold while the best ranking models easily outperform the 0.8 threshold. The top-ranking model for each adverse reaction will be chosen as representative and tested against an independent set. Figure 26 shows the evaluation results of the testing procedure for the single predictors as well as the evaluation scores for the meta-predictors.

In this case, the differences between the classifiers were minimal, performing similarly in terms of Accuracy, Precision, and AUC, though RF appeared to score higher, particularly in terms of sensitivity, with the highest value for the third quartile of the distribution. In terms of Matthew's Correlation Coefficient (MCC), the values are primarily distributed above zero, with the median value hovering around 0.25, indicating non-random predictions (Figure 26).

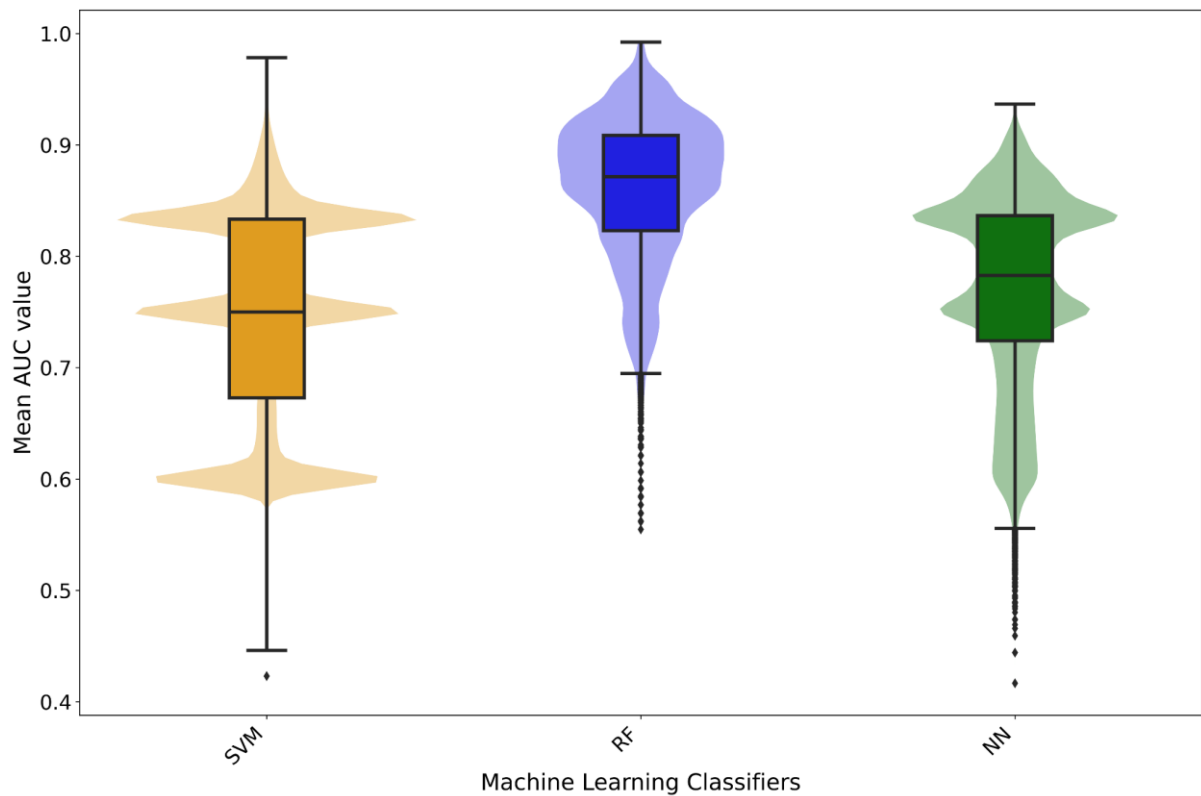


Figure 25. Box- and violin plots of the cross-validation AUC results for the three different classifiers. The different box-plots show the distribution of the mean AUC values for the best models developed for each adverse reaction using the three different classifiers: SVM (orange), Random Forest (Blue) and Neural Networks (green).

4.6 - Meta predictor implementation and details

To incorporate individual classifier predictions, three voting systems were proposed: a jury vote, a consensus score, and a red-flag schema. Since classifiers are binary, they can predict whether or not a certain protein is causing a given adverse reaction. Both the jury vote and the consensus aim to maximize comparable predictions, whereas the red-flag focuses on outliers.

The jury vote is simply a tally of the predicted outcomes and is one of the most basic voting systems. In a nutshell, given a protein X, if two ML methods predict the protein class as 1 - or "adverse reaction linked" and the other method as 0, the ensemble method output class will be 1. The consensus score c is more granular, as it uses the posterior probability p of each classifier instead of a yes/no answer. As a result, the consensus score can be used to rank proteins in the same class, such as those projected to be connected to a specific adverse reaction. In particular this method adds the single ML class probability multiplied by 1 or -1 depending on whether the prediction is adverse reaction-linked or not. If the sum of these values is positive, the ensemble method output is 1 or "adverse reaction-linked," otherwise it is 0 or "Not-adverse reaction-linked."

$$c = \sum_{i=1}^3 p_i * class(i); i = [SVM, RF, NN]; class \in [-1, +1] \quad \text{eq. 45}$$

Consider protein X once more for a further concrete example. If the SVM model predicts the protein class as 1 with a probability of 0.86, the RF predicts it as 0 with a probability of 0.63, and the NN predicts it as 1 with a probability of 0.53, the overall consensus prediction would be $(0.86 * +1) + (0.63 * -1) + (0.53 * +1) = 0.76$. Since this value is positive, the protein class can be accepted as 1.

Finally, the red flag schema simply accepts as final prediction the one which is not common among the different classifiers. In other words, as opposed to the jury vote system, the red flag approach will select as output for the ensemble method class with

least predictions. Considering again a protein X, if two ML methods predict the class as 1 - or “adverse reaction linked” and just one as 0, the ensemble method output class will be 0

4.7 - Single predictor vs Meta-Predictor

Since each adverse reaction was assigned three different classifiers, it is possible to combine the predictions using the scoring methods described above, resulting in improved prediction performance with respect to the single predictor. Indeed, Accuracy, Precision, Recall, and AUC increased when compared to individual predictors in the jury vote and consensus voting systems (figure 26). Generally speaking, there was not only an improvement, but also a gradual shift toward higher values as the distribution skewed toward better values. The red flag method, on the other hand, caused predictions to worsen. As previously stated, the red flag was designed to detect singular projections and to serve as a failsafe in the case of two classifiers that predict the same results but with low probability. A similar pattern can be seen in the case of MCC values (figure 26).

Indeed, the distribution of MCC values for jury vote and consensus voting systems was more skewed toward higher values when compared to individual predictors, directly indicating an improvement of the prediction quality. Lower MCC values ranging from 0 (random prediction) to negative (inverse) values are seen in the red flag consensus of Accuracy, Precision, and Recall. Accepting the most common prediction rather than any single predictor is thus a better strategy.

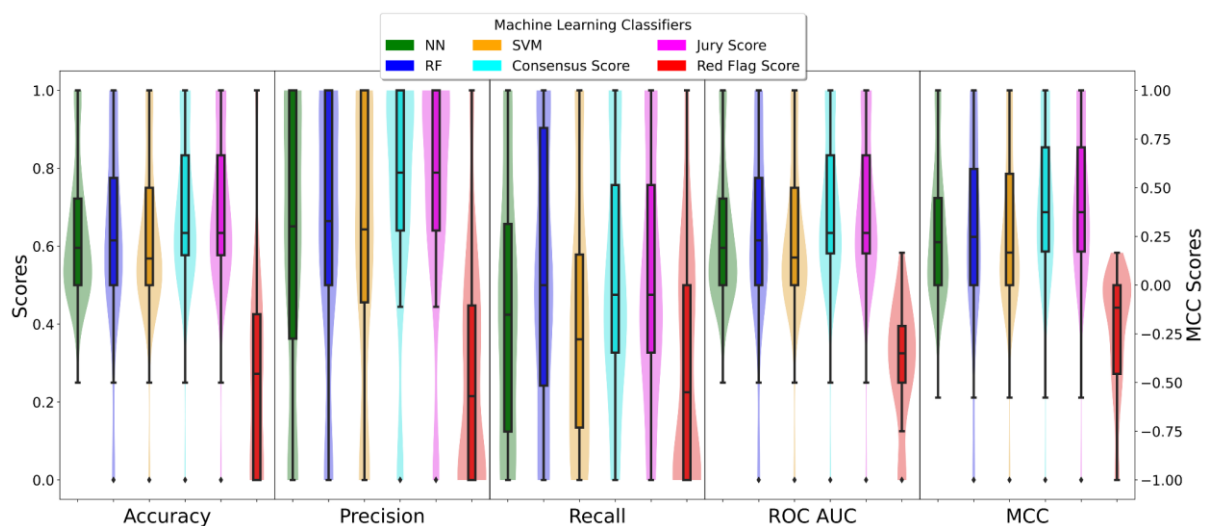


Figure 26. Box- and violin plots for accuracy (ACC), precision (PREC), recall (REC), Receiver Operating Area Under Curve (ROC AUC) and Matthew

Correlation Coefficient (MCC). Distribution of Accuracy, Precision, Recall and ROC AUC values for individual classifiers: NN (green), RF (blue) and SVM (orange) as well as meta-predictions: consensus (cyan), jury-vote (magenta) and red-flag (red)

4.7.1 - Predicting at SOC level

All of the models in the previous sections were adverse reaction-specific. However, in pure terms of applicability, there could be the need also of more generalist predictive models while maintaining biological and medicinal significance. This can be obtained exploiting the information contained in the MEDDRA classification system, to categorize the various ADR into distinct System Organ Classes (SOCs) [38]. Since certain MEDDRA reported ADR are very generic or not particular to body regions, tissues, or underlying human biology, not every SOC is contained in the database, as mentioned in the T-ARDIS publication [11].

However, as previously stated, the 84 ADR included in this study were chosen specifically to cover the entire spectrum of available SOC, being able to be divided into 18 separate SOCs, with each SOC containing an average of 5 ADR. There is a lot of variation in predictions for the accuracy, precision, sensitivity, and MCC scores at the single classifier level (Figure 27). When it came to "Respiratory, thoracic and mediastinal disorders" predictions were far more accurate than when it came to immunological or nervous disorders.

With the exception of red-flag voting, combining predictors enhanced forecasts in general, especially in terms of Recall. However, when compared to predictors acting at the adverse reaction level, sensitivity values were often low (figure 27). This fact emphasizes how difficult it is to predict at a higher level of abstraction rather than at the level of individual ADR. A similar issue can be found in terms of MCC values (figure 25). When individual predictions were combined in a jury vote or consensus voting, predictions improved, such as in the case of respiratory, thoracic, and mediastinal illnesses, which went from an MCC of 0.75 of the best predictors to 0.81 when combined.

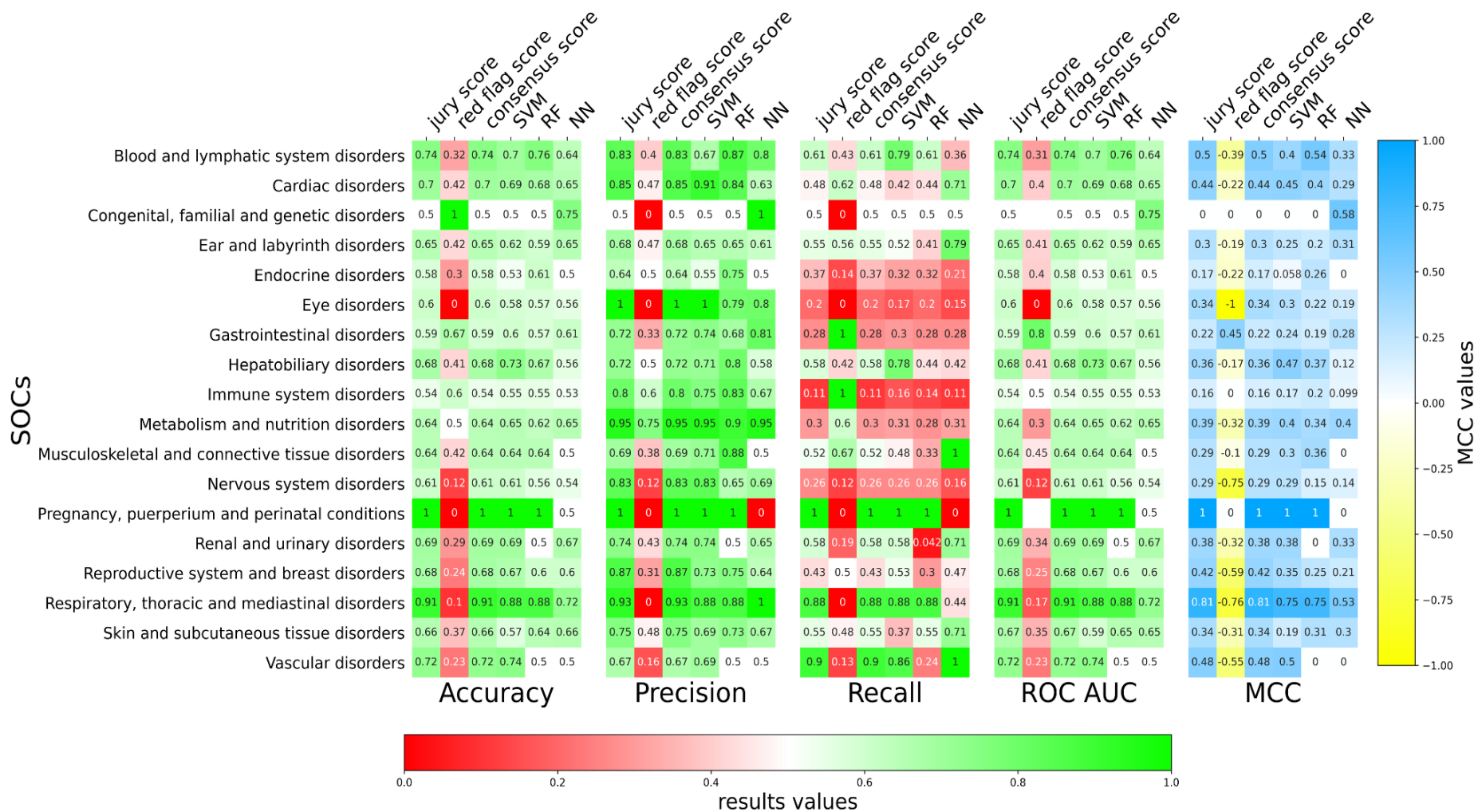


Figure 27. Evaluation of adverse reaction-protein association predictions of the different classifiers at SOCs level. Accuracy, Precision, Recall and ROC AUC values for predictions at SOCs for both individual classifiers (SVM, RF, NN) and voting (*jury vote*, consensus and *red flag*).

4.8 - Discussion

The goal of this study was to develop a method for predicting the potential liability of proteins in the context of ADR when they are targeted for therapeutic purposes. By analyzing the human interactome, a number of network-based metrics were developed to characterize the proteins under investigation. This wide range set of measures was then fed into three machine-learning classifiers, which were then combined using three different voting methods. Both the individual adverse reaction and SOC prediction models performed well, indicating that they can be used to forecast potential protein liabilities.

4.8.1 - Classifiers performances

There were eight variables used in the predictions, each representing a different characteristic of the proteins under investigation. As illustrated in Figure 24, the level of discrimination between positive and negative cases varies with GUILDify scores and K1 clustering analyses among top performers and degree centrality and clustering coefficient analyses as fewer discriminating factors. This reflects the small-world nature of the human interactome.

There were also distinctions among the ensemble classifiers (figure 26 - 27). Under training conditions, RF appeared to perform best, but the performance of the different classifiers was inferior in some cases for specific ADR, demonstrating the complexity and heterogeneity of this biological problem. As a result of the previous discovery, I devised a voting method to combine the individual predictors into a meta-predictor. With the exception of the red-flag vote, combining the methods produced better predictions, as illustrated in Figures 26-27. The jury vote and consensus voting systems both worked on the same premise: to improve classifiers that make similar predictions. In fact, the jury vote and consensus voting methods perform similarly (figures 25 - 26),

but the consensus voting system adds more specificity to the predictions, allowing for a more accurate ranking. Indeed, whereas a jury vote will assign a specific protein to a class, such as +2; both methods agree that the given protein is associated with a specific adverse reaction, the consensus scoring function will provide a quantitative metric that can be used to rank proteins within the same class. This feature is critical for establishing trust in the DocTOR application's predictions (see below). Finally, the red-flag voting mechanism, as previously stated, resulted in overall worse predictions. However, in certain cases, such as *nocturia*, *neutropenia*, or *ischaemia* adverse reaction, this method has been shown to be effective.

Another aspect of this study that was looked into was the nature of the predictions. In theory, one of the key achievements of protein-adverse reaction predictions would be determining whether targeting a protein will result in an undesirable adverse event, i.e., obtain a unique model to predict every adverse reaction. Given however the presence and possible concurrence of many different types of ADR, this is quite a difficult subject to convert into a predictive model which could lead directly to consider every protein linked to any adverse reaction.

This is why the predictive models were adverse reaction-specific; the prediction is not whether a protein will cause an undesirable event, but what type of reaction it will cause. Nonetheless, ADR can be classified into common SOCs.

Individual ADR are abstracted into a higher entity in this way, allowing for the development of more generalist prediction models, such as one that predicts whether the targeting of a specific protein is linked to a specific SOC perturbation. Predicting at this level resulted in some SOCs outperforming others, as illustrated in Figures 27. Moreover, it appears that SOCs with more clearly defined impacted tissues/organs had higher systemic representations in their predictions.

4.8.2 - Self-reporting vs curated Dataset results

The ML techniques' distribution of scores reflects the varying number of proteins linked with ADR in the self-reporting and curated T-ARDIS dataset. Due to the nature of the origin databases, the curated set presents a more specific and trustworthy direct correlation between targets and ADR, which is useful in the prediction of tissue-specific ADR such as Atrial Fibrillation (curated jury score ACC 0.88, PREC 0.88, RECALL 0.88, MCC 0.77). However, this has an effect on the dimension of the GUILDify subnetwork and, as a result, on the definition of positive set during the training phase, resulting in somewhat poorer accuracy and precision for some difficult-to-predict associations such as *Respiratory Failure* (self-reporting jury score ACC 0.722222, PREC 0.7, RECALL 0.77, MCC 0.44; curated jury score ACC 0.66, PREC 0.66, RECALL 0.66, MCC 0.33). Overall, however, the implemented method proved successful both in the case of single ADR and SOC for the controlled dataset as is shown in figure 28 -29.

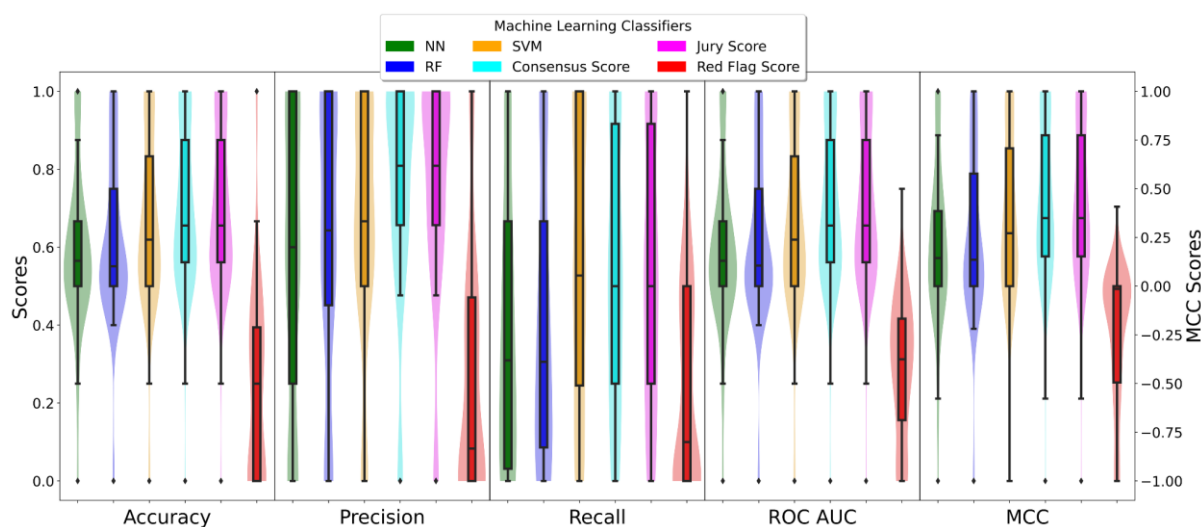


Figure 28. Box- and violin plots for accuracy (ACC), precision (PREC), recall (REC), Receiver Operating Area Under Curve (ROC AUC) and MCC for the curated dataset. Distribution of ACC, PREC, REC and ROC AUC values for individual classifiers: NN (green), RF (blue) and SVM (orange) as well as meta-predictions: consensus (cyan), jury-vote (magenta) and red-flag (red)

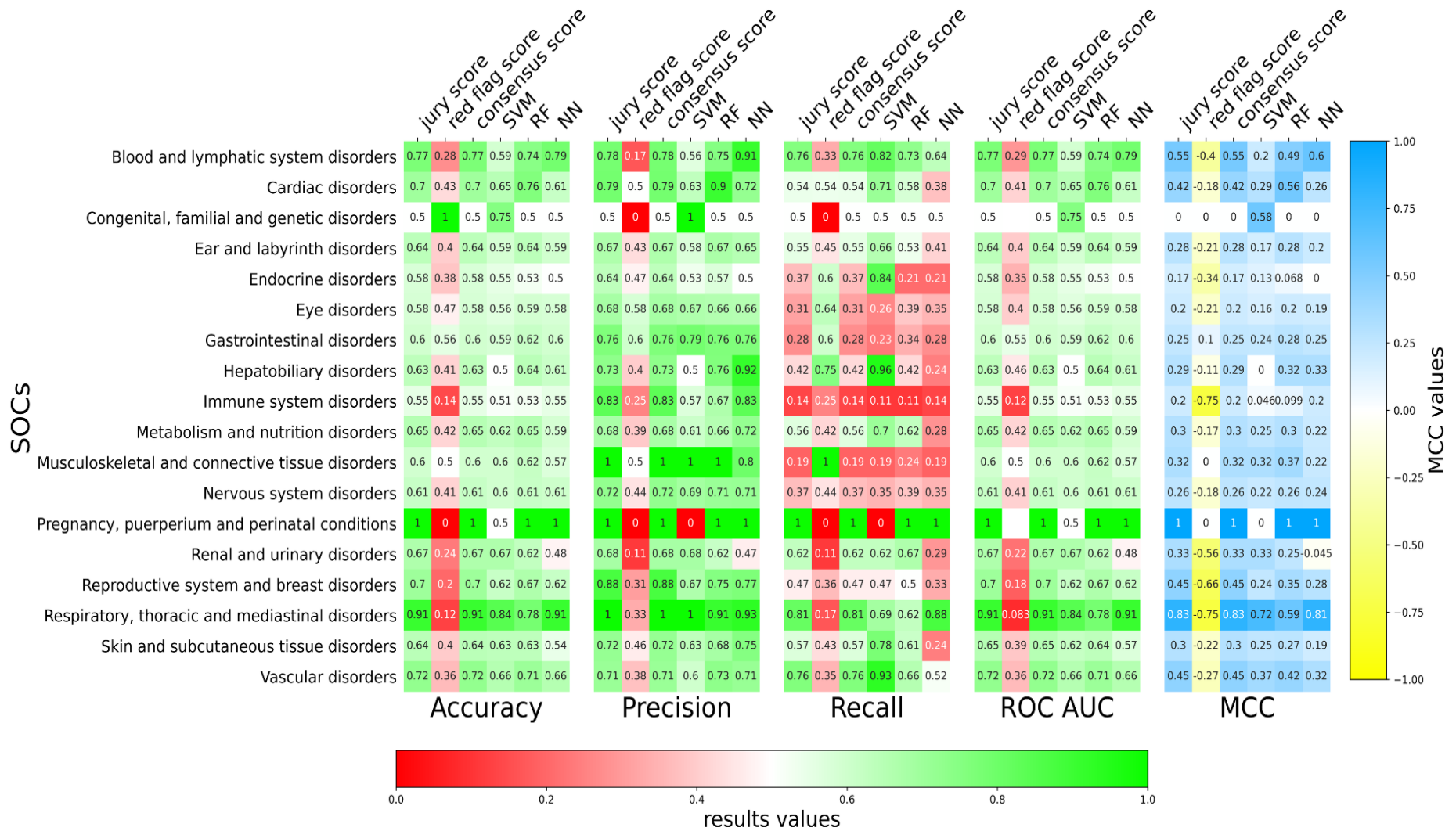


Figure 29. Heatmap of predictions at SOCs for curated dataset. ACC, PREC, REC, ROC AUC and MCC values for predictions at SOCs for both individual classifiers (SVM, RF, NN) and voting (*jury vote*, consensus and *red flag* for the controlled dataset).

4.8.3 - The DocTOR utility

All of the datasets used in this work, as well as the predictive models and auxiliary scripts to carry out the predictions, are available at the Direct fOreCast Target On Reaction (DocTOR) application, which can be found at <https://github.com/cristian931/DocTOR>. To explore the potential association between the two, users can submit a list of proteins in the form of UNIPROT identification codes and a list of ADR of interest (from the available models). For all three distinct classifiers (SVM, NN, and RF) and voting methods, the computer will assign a positive or negative class to the protein, as well as a probability associated with the chosen class (jury vote, consensus and red-flag). As a result, when analyzing the forecast results, users can take into account all of this information. When novel protein targets are identified to be related with certain ADR and/or new releases of the T-ARDIS database, the application lends itself to being quickly updated, allowing for the addition of new models for new ADR on demand or the retraining of current models.

4.9 - Summary

In this chapter, I looked at protein liabilities in the context of medicinal development from an interactome-centric standpoint gathering information on protein topology in the human interactome, insights in relation to certain in vitro verified adverse reaction-related hotspots and finally function connections. With the obtained features I trained three separate machine-learning models using the various variables to predict 84 different ADR, including a DILI-related subset and 20 different System Organ Classes. The models were optimized using grid-search and 5-fold cross-validation, and the results were tested on a separate dataset. The effectiveness of the models in both training and independent testing validates their use as a future computational tool for assessing protein liability at the level of specific adverse reaction type and SOC. Finally, I made the data, models, and prediction tool available to the scientific community through a GitHub repository.

**5 - LINKING AND IDENTIFYING THE
MOLECULAR BASES OF ADRS
THROUGH SHARED TARGETS:
SONG**

5.1 - Abstract

T-ARDIS opened the door to this project in Chapter 4, allowing statistical correlation of ADRs with Protein Targets retrieved from publicly available data-sets. T-ARDIS, unlike other resources, enables the use of the largest archives available, yielding a substantial amount of data. DocTOR used this information in chapter 5, to train and develop machine-learning based tools to predict the likelihood of a protein to elicit an ADR. While T-ARDIS deals with known associations (assessed statistically), DocTOR could be applied to proteins for which no information is available, i.e., *de novo* predictions. Despite the promising results obtained thus far, this research has not yet investigated the actual molecular basis of ADR onset.

I now propose an alternative viewpoint on the T-ARDIS discoveries, which will allow us to explore the proteins shared by different ADRs from a different network standpoint. In the following sections, I will examine and expand an "Adverse-Reactome," a different type of network formed by plotting the various ADRs extracted from T-ARDIS curated datasets as nodes and using the shared proteins between them as edges. The resulting network will go through a clustering procedure that will aid in identifying specific subsets and modules of ADRs linked by peculiar proteins. Investigating the cluster's protein-enriched functions and mining the literature for information, I hope to uncover the possible relationship between ADRs and associated protein roles, expanding our knowledge, and eventually identifying the molecular perturbation that causes ADRs to occur as well as the relationship between ADRs.

5.2 - The SONG Network

As exposed in chapter 4, the T-ARDIS methodology was able to mine a wide range of adverse reactions-protein relationships. This data contains the information mined from two distinct sources. one from self-reported databases and the other, defined "controlled" data-set, from more curated databases. Still, even if the proteins-ADRs relationships have been uncovered, the precise molecular mechanism of protein modulation and how this event is related to the onset of an ADR remains unknown. Examining the function of the proteins involved in the various ADRs, on the other hand, can help shed light on this problem. However, throughout this research, protein-ADR interactions have been regarded as stand-alone relationships, despite the fact that functions, and thus ADRs, are carried out finely through the synergy of multiple proteins. As a result, the retrieved ADR-protein information must be integrated in a more interconnected environment. This could lead to the discovery of seemingly unrelated ADRs that are actually linked by a shared biological process.

This idea resulted in the development of the SONG (Side-effect ON Graph) approach. Given computational time and demand, this method has been however developed limited to the more reliable curated set of T-ARDIS databases, which contains 4k statistically significant protein-ADR relationships shared between 537 and 194 ADRs and proteins, respectively [11]. Given the SONG method's multi-step nature, the first phase revolves around the creation of a network that represents the relationships between multiple extracted ADRs, with shared proteins acting as connecting elements. The resulting network will be similar to a protein-protein interaction network integrating the ADRs as nodes and the proteins they share as edges. The total number of shared proteins between two nodes has also been integrated in the network as edge score value. As a result, a densely connected network with 537 nodes and 20K edges was generated, along with 12 isolated nodes (ADRs that do not share proteins with other nodes) (Figure 30), which was then visualized and analyzed using the Cytoscape [101] utility.

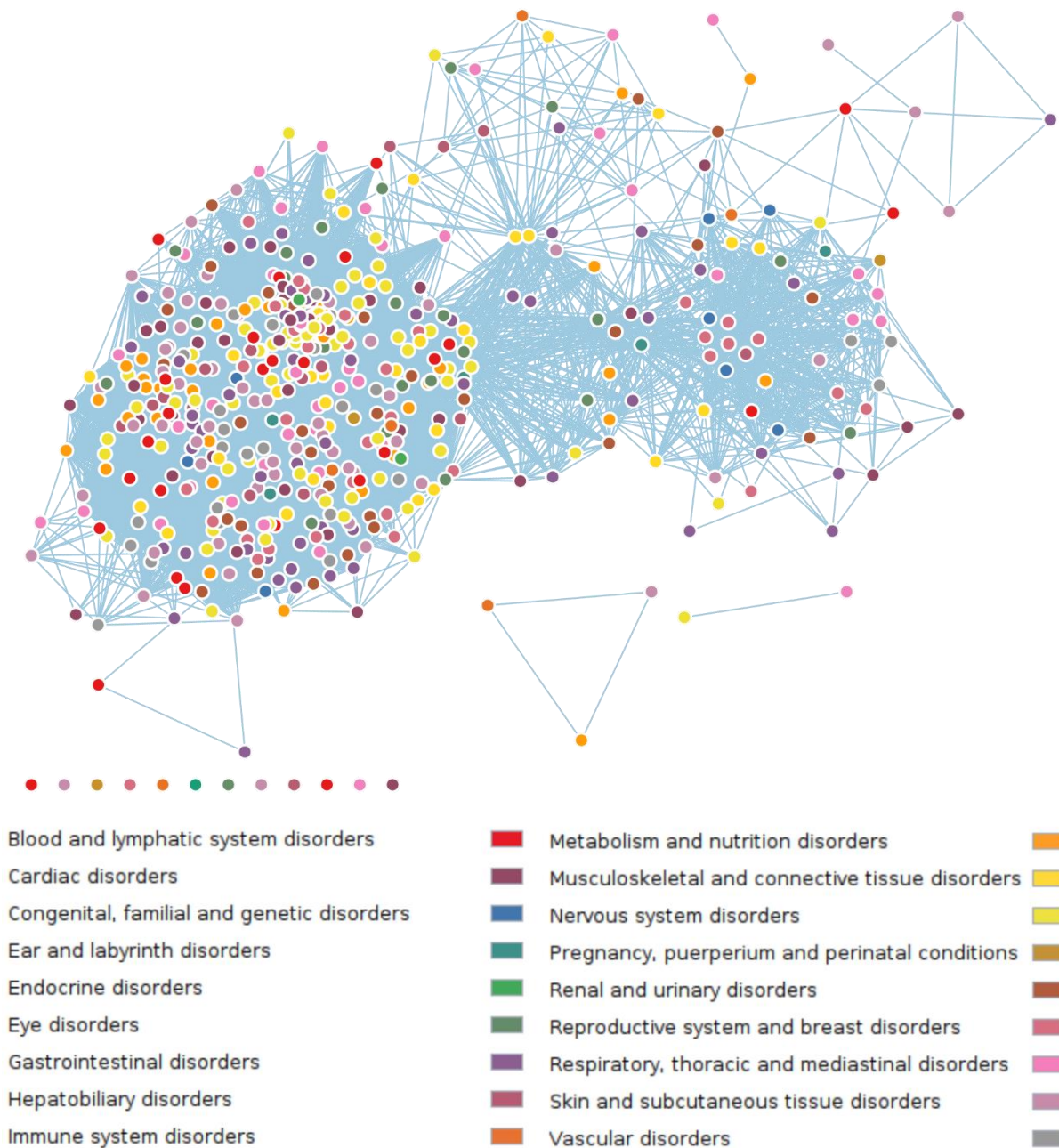


Figure 30. Representation of the Adverse Reactome - The network resulting from the integration of T-ARDIS curated dataset, the ADRs are represented as nodes while the edges contain the shared proteins. The edge weight is represented by the total number of proteins shared between two ADRs. Each different color represents the ADR associated SOC.

The high number of edges in comparison to the small number of nodes demonstrates how proteins are commonly shared among the various ADRs and SOC, explaining the

difficulties of predictions during the DocTOR method's development. This, however, suggests also the presence of a somewhat shared biological pathway even between seemingly unrelated ADRs explaining the possible onset of comorbidities. The "Adverse Reactome" also highlights a group of ADRs that share no protein with any other node. *Thrombotic thrombocytopenic purpura, Pruritus, Premature baby, Ovarian disorder, Hypersensitivity, Hypervolaemia, Dermatitis, Cholecystitis, Anaemia megaloblastic, Acute pulmonary oedema, and Acute myocardial infarction* are all examples. As can be seen, the majority of these ADRs are associated with various SOCs and are distinguished by a small number of T-ARDIS proteins [11]. Nevertheless, given their unique associated proteins, they are ideal candidates for DocTOR future models (to be explored in a later section devoted to future perspectives.)

5.2.1 - Clustering application

The more likely method of extracting functional information from the generated Adverse Reactome is to use a clustering procedure to identify specific subsets of ADRs with similar properties and characterized by the given proteins. The Affinity Propagation Clustering Algorithm [102] represent an appropriate choice for the problem at hand as it is based on the concept of "message passing" between data points and does not require the number of clusters to be a priori specified. Specifically, the algorithm is based on a repeated exchange of information between all nodes.

The base algorithm can be explained easily as follows: each data point sends signals to all other nodes, passing the information of the relative attraction of each target to the sender as a score. Given the attractiveness of the messages received, each target communicates to all senders its availability to associate with the sender. The sender node signals again to targets modifying the score based on the availability signals received before. The message-passing method is repeated until agreement is attained. When the sender is paired with one of its targets, that target becomes the exemplar of the point. Finally, all points with the same exemplar are clustered together.

Mathematically the algorithm proceeds by alternating between two message-passing steps, which update three matrices: a similarity (s) matrix, a responsibility (r) matrix, and an availability (a) matrix. Results are contained in a criterion matrix (c). These matrices are updated repeatedly using four equations, where i and k correspond to the rows and columns of the corresponding matrix.

$$r(i, k) \leftarrow s(i, k) - \max_{k' \text{ such that } k' \neq k} \{a(i, k') + s(i, k')\} \quad \text{Eq. 46}$$

$$a(k, k) \leftarrow \sum_{i, i \neq k} \max\{0, r(i', k)\} \quad \text{Eq. 47}$$

$$a(i, k) \leftarrow \min\{0, r(k, k) + a(k, k)\} \quad \text{Eq. 48}$$

$$c(i, k) \leftarrow r(i, k) + a(i, k) \quad \text{Eq. 49}$$

In a nutshell, the distances between elements are subtracted to create the similarity matrix. The sum of the squares of the differences between variables that make up the items are typically used to determine these distances. The algorithm's next step is to create an availability matrix with all of its entries set to zero. The responsibility matrix is then computed using Equation 46. Equations 47 and 48 are then used to update the availability matrix's diagonal and off-diagonal entries, respectively. Finally applying Equation 49 gives rise to the criterion matrix. For each row, the column with the highest criterion value specifies the exemplar for that row's item. A cluster is composed of rows that have the same exemplar.

The application of this algorithm to the Adverse Reactome identifies 24 clusters and 67 singletons. Excluding the singletons and clusters with ≤ 3 nodes we remain with 16 highly connected clusters (Figure 31)

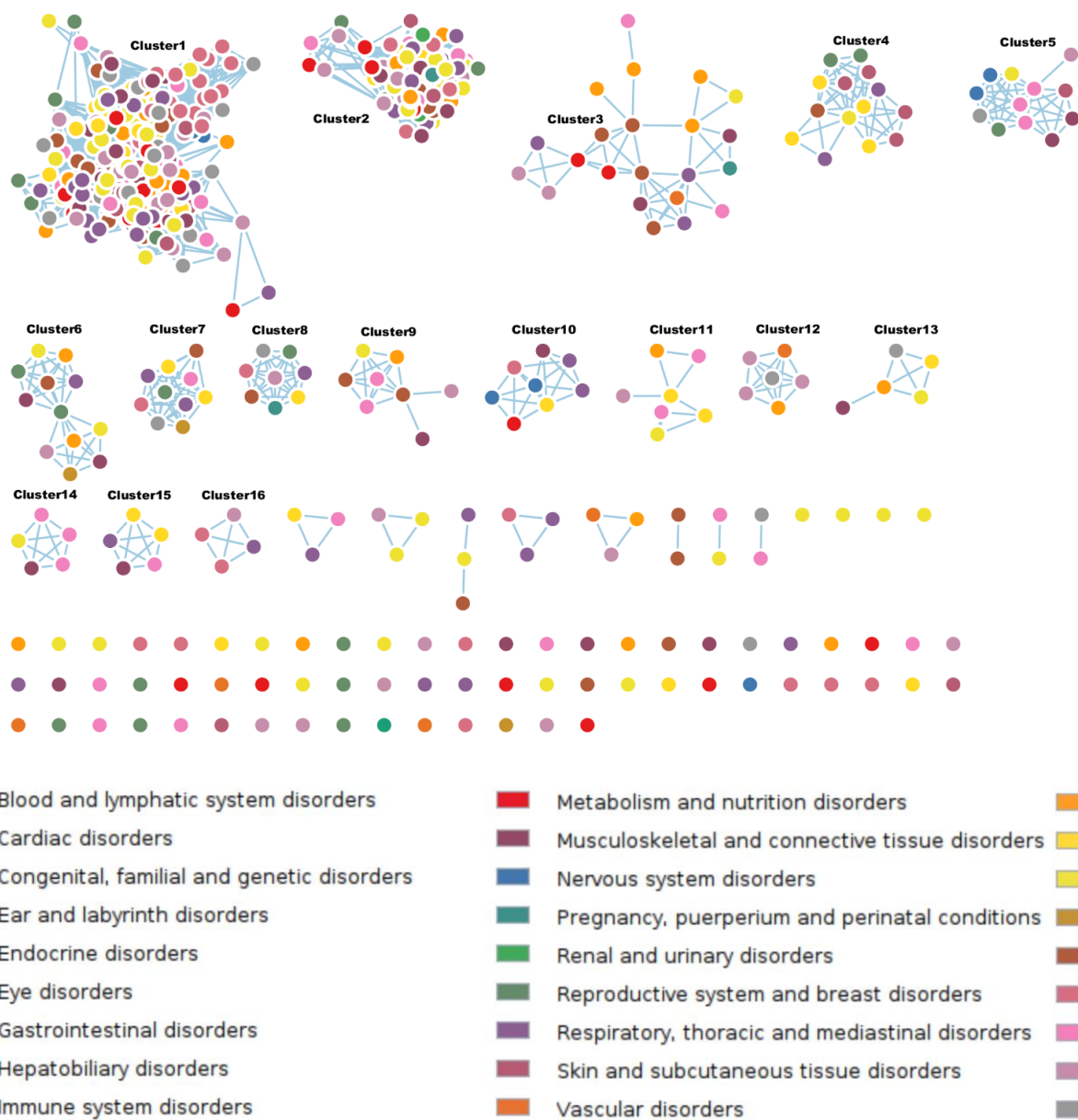


Figure 31. Results of the clustering procedure on the Adverse Reactome. The clusters are ordered from left to right in a crescent number of nodes, with the largest in the leftmost upper denominated as Cluster1. The nodes are classified by colors denoting the belonging of a distinct SOC.

Preliminary analysis of the clustered network reveals no distinct division between SOCs, as already suggested by the original network's high connectivity. The clusters have been numbered from left to right, beginning with the largest as Cluster1 to Cluster16 without considering, as previously stated, singletons and cluster with less or equal than 3 nodes.

5.3 - Functional data enrichment

The discovered modules, as previously stated, are composed of ADRs that share distinctive proteins or characteristics. The analysis of the underlying protein function may reveal the hidden relationship that characterizes these ADRs. To accomplish this, I used g:profiler [103], an online tool for gene functional enrichment.

5.3.1 - g:profiler

g:Profiler [42] is a set of tools that are used in biological entity (gene/protein)-centered computational analysis pipelines. It is constituted by several applications, specialized for the different aspects of gene functional enrichment analysis: g:GOSSt analyzes the functional enrichment of single or multiple gene lists, g:Convert converts gene/protein IDs between different name-spaces, and g:Orth allows orthologous genes to be mapped across species. g:SNPense is a program that connects human SNP identifiers to genes.

The cluster's associated proteins study has been performed using the g:GOSSt utility, the primary software specialized in functional enrichment analysis on a user-defined gene list input. The utility mines different databases to retrieve functional information and finds biological processes, pathways, regulatory motifs, and protein complexes that are statistically significantly enriched. The Ensembl database is the primary source of information on genes together with the Gene Ontology resources, already introduced in this thesis.

On the statistical evaluation point of view, the well-proven cumulative hyper-geometric test is used to assess the functional enrichment of the input gene list, analyzing large numbers of functional terms that are evaluated at once, together with multiple Bonferroni testing corrections. [103]. This is necessary to reduce the amount of false positive findings given the large number of terms evaluated at the same time, e.g., around 16 000 GO biological process keywords are taken into account only for the

human gene list. On completion, the g:GOST method provides a result table containing information about the enriched terms, overlap sizes and corresponding *P*-values (Figure 32). The enriched terms with lowest *p*-value at level of molecular function and biological process are selected as representative of the cluster function. Clusters presenting more interactions have been associated with more statistically meaningful functions as will be presented in the next subchapters.



Figure 32. Example of g:GOST method output. From this interface it is easy to extrapolate the relevant information regarding the enriched terms with the lower *p*-value. Moreover, the procedure also highlights the proteins whose enriched function has been validated in literature and in vitro (right column). The color mapping is available at the g:profile website. [103]

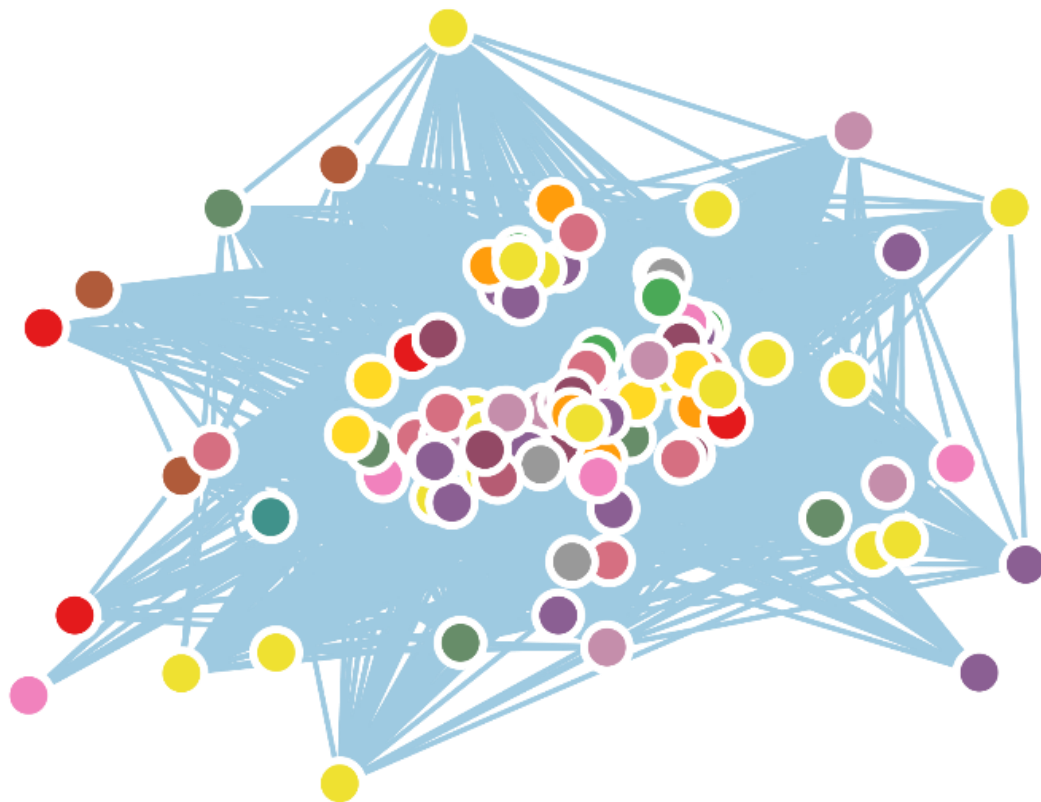
5.4 - Uncovered Associations and examples

The proposed approach was successful in linking fundamental functions to the development of a diverse set of ADRs. Various clusters have provided enriched functions in accordance with the belonging SOC, and the latter's perturbation may be directly correlated with the onset of ADRs. The majority of proteins and ADRs are found in Cluster 1, which contains 219 of the original network's 537 nodes. Cluster 2 is the second largest cluster, with 104 nodes. The other 14 clusters present from a maximum of 22 to a minimum of 4 nodes.

Given Cluster₁ contained more than half of all possible nodes, another clustering run was performed using again the Affinity Propagation Clustering Algorithm. As a result, nine new clusters and 23 singletons were discovered. Cluster₂ has also been subjected to a re-clustering procedure, but with no success. This is almost certainly due to the latter's high connectivity. (See Figure 31). In the next sub-section, I'll present some of the investigated clusters and their enriched functions, as well as a literature review, in order to link finally function perturbation to possible ADR manifestations.

5.4.1 - G-coupled serotonin receptor signaling pathway disruption causes multi-organ failure

Cluster 2 is, as previously stated, the cluster with the greatest connectivity presenting 104 nodes, and 35 distinct proteins. (Figure 33)



Blood and lymphatic system disorders	Metabolism and nutrition disorders	
Cardiac disorders	Musculoskeletal and connective tissue disorders	
Congenital, familial and genetic disorders	Nervous system disorders	
Ear and labyrinth disorders	Pregnancy, puerperium and perinatal conditions	
Endocrine disorders	Renal and urinary disorders	
Eye disorders	Reproductive system and breast disorders	
Gastrointestinal disorders	Respiratory, thoracic and mediastinal disorders	
Hepatobiliary disorders	Skin and subcutaneous tissue disorders	
Immune system disorders	Vascular disorders	

Figure 33. Cluster 2 resulting network. The nodes are presented without labels for visualization issues. The different ADRs are part of SOCs: Blood and lymphatic system disorders (red), Cardiac disorders (brown), Ear and labyrinth disorders (green), Endocrine disorders (light green), Eye disorders (dark green), Gastrointestinal disorders (dark purple), Metabolism and nutrition disorders (light orange), Musculoskeletal and connective tissue disorders (light orange), Nervous system disorders (yellow), Renal and urinary disorders (Dark pink), Reproductive system and breast disorders (pink-red), Respiratory, thoracic and mediastinal disorders (shocking pink), Skin and subcutaneous tissue disorders (violet), Vascular disorders (grey).

Many of the proteins found belong to the ADRA, CALM, CHRM, and HTR families, which have already been linked to the most ADRs in T-ARDIS [11]. These proteins serve

an important role as receptors in a variety of functions. So, it's no surprise that the enriched function indicated by g:profile is the most generic or related to the activity of G-coupled receptors. This also explains the high number of ADRs belonging to the most disparate SOCs. Given the high number of ADRs, just one for each of the SOC identified in this cluster will be brought as example: Thrombocytosis (Blood and lymphatic system disorders), Cardiomegaly (Cardiac disorders), Hyperacusis (Ear and labyrinth disorders), Hypothyroidism (Endocrine disorders), Glaucoma (Eye disorders), Parotid gland enlargement (Gastrointestinal disorders), Diabetes mellitus (Metabolism and nutrition disorders), Torticollis (Musculoskeletal and connective tissue disorders), Tardive dyskinesia (Nervous system disorders), Glycosuria (Renal and urinary disorders, but also related to Diabetes onset), Prostatitis (Reproductive system and breast disorders), Pleural fibrosis (Respiratory, thoracic and mediastinal disorders), Psoriasis (Skin and subcutaneous tissue disorders) and Hypertension (Vascular disorders). All of the presented ADR act on a diverse system with different gravity and incidence, however in vitro studies have proven their correlation to GPCR modulation (Table 7).

Table 7. List of publications for the Cluster2 ADRs

Adverse Reaction	Publication
<i>Thrombocytosis</i>	[104]
<i>Cardiomegaly</i>	[104]
<i>Hyperacusis</i>	[105]
<i>Hypothyroidism</i>	[106]
<i>Glaucoma</i>	[107]
<i>Parotid gland enlargement</i>	[108]
<i>Diabetes mellitus</i>	[109]
<i>Torticollis</i>	[110]
<i>Tardive dyskinesia</i>	[111]
<i>Glycosuria</i>	[112]
<i>Prostatitis</i>	[113]
<i>Pleural fibrosis</i>	[114]
<i>Psoriasis</i>	[115]
<i>Hypertension</i>	[104]

5.4.2 - Perturbation of Smooth muscle Adaptation and NADPH binding inficiate multi-level biological functions

One of the studied clusters included a wide range of adverse events that were quite difficult to link in an SOC point of view. Cluster 4 (Figure 31) is a small sized cluster composed of 15 different Adverse reactions. Between them we can find *Abnormal faeces*, *Arthritis*, *Jaundice cholestatic*, *Foetor hepaticus*, *Muscle atrophy*, *Lenticular opacities*, *Cognitive disorder*, *Cataract*, *Myoglobinuria*, *Dyspepsia*, *Hepatic necrosis*, *Lupus-like syndrome*, *Myopathy* and *Liver disorder*.

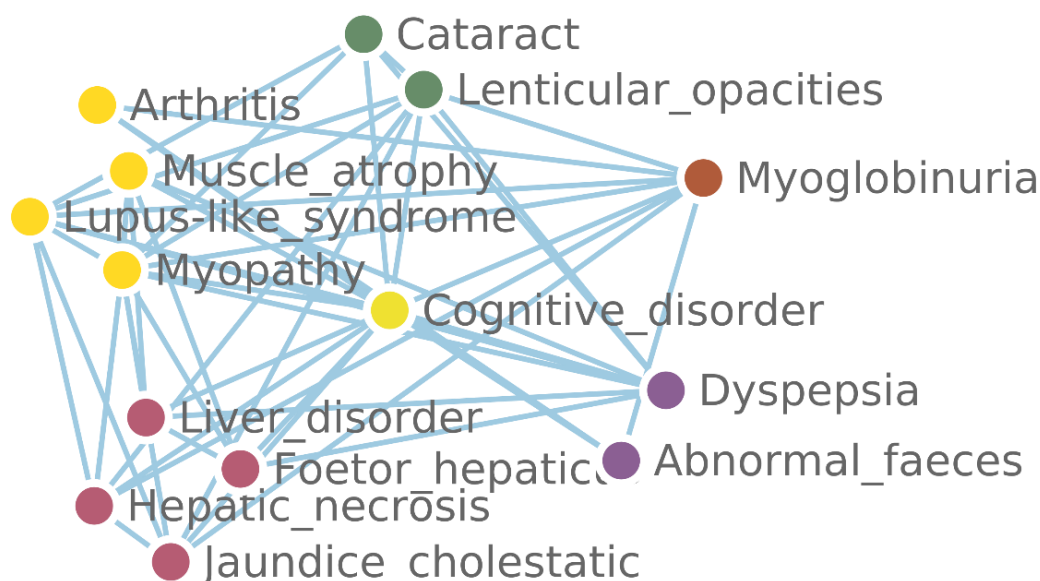


Figure 34. Cluster 4 resulting network. The nodes have been manually grouped by SOC to aid visualization. The different ADRs are part of 5 SOCs: Eye disorders (dark green), Gastrointestinal disorders (dark purple), Hepatobiliary disorders (radish), Musculoskeletal and connective tissue disorders (light orange), Nervous system disorders (yellow - central node), Renal and urinary disorders (dark red).

Finding a clear relationship between this subset of ADRs is not an easy task, as many are part of different SOC, despite the fact that the majority are part of the hepatobiliary system or the musculoskeletal system. (Figure 34). Moreover, the range of different and diverse ADRs (e.g., *Abnormal faeces*, *myopathy*) further complicate this task.

Nonetheless, the proposed study identified six underlying genes shared by these ADRs: CRP, HMDH, COG₂, APOB, HMOX₁, and NOS₃. The g:profiler analysis yielded 5 different statistically significant GO molecular functions and 4 biological processes (figure 35).

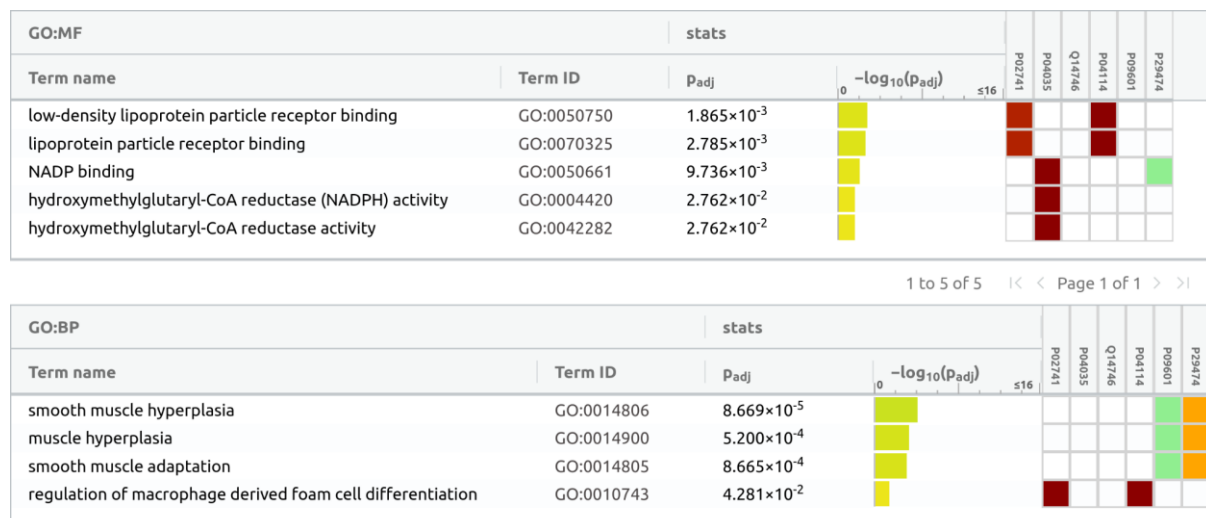


Figure 35. g:profiler results for cluster 4 analysis. The lowest p-value were obtained for the GO terms GO:0050750 (low density lipoprotein particle receptor binding) and GO:0014806 (smooth muscle hyperplasia)

Each one of the ADR has been related with the perturbation for these functions in literature or at least with the smooth muscle growth biological process perturbation. For example, changing in the bowel smooth muscle has been associated with the event of abnormal faeces and muscle hyperplasia has been also related to Crohn's disease [116]. On the other hand, a high level of α -smooth muscle actin has been also identified in patients with Arthritis [117]. Given the liver's smooth muscle structure and participation in the NADPH binding system, liver degeneration is not surprising when smooth muscle adaptation activities are disrupted, as in the cases of Foetor hepaticus, Jaundice cholestatic, and Hepatic necrosis. [118] [119] [120]. NADPH binding perturbation, on the other hand, may affect the brain and has been linked to cognitive impairment in rats [121]. Thus, starting from unrelated ADRs and through the analysis of SONG we can derive actionable hypothesis to provide an explanation to the molecular basis.

5.4.3 - Cyclooxygenase inhibition presents a multi-system impact

Cluster 12 is one of the smallest clusters identified (Figure 36). It presents 6 Adverse reactions associated between them by just three proteins. The Adverse reactions presented appear to be multi-systemic: Anaphylactoid reaction (being part of the Immune system disorders SOC), Hyperkalaemia (part of the Metabolism and nutrition disorders SOC), Angioedema, Erythema multiforme, Dermatitis exfoliative (all three parts of the Skin and subcutaneous tissue disorders SOC) and finally Vasculitis (part of the Vascular disorders SOC).

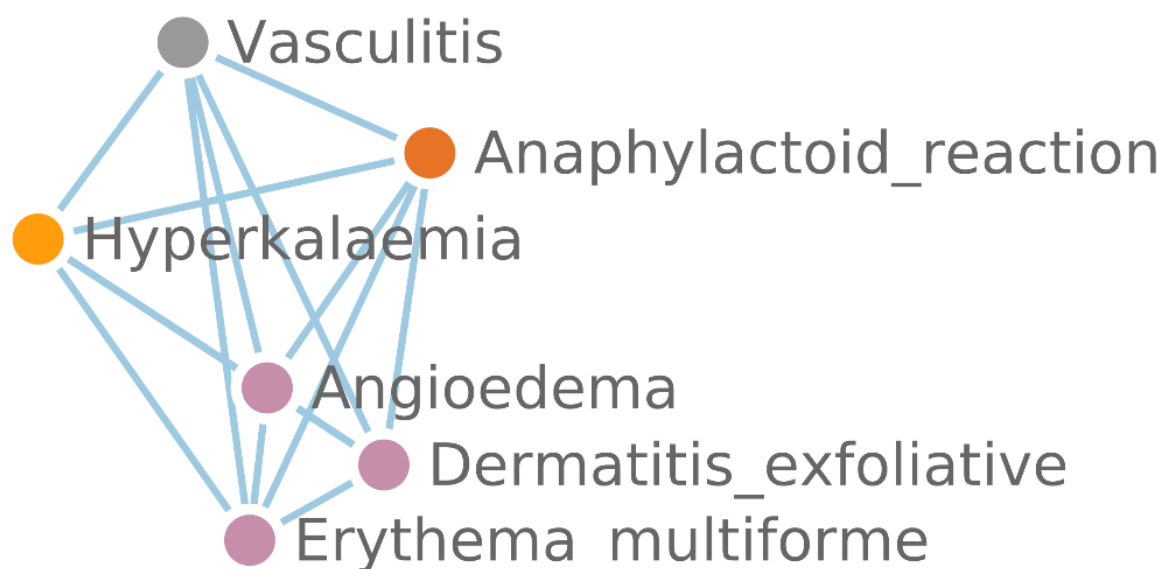


Figure 36. Cluster 12 resulting network. The nodes have been manually grouped by SOC to aid visualization. The different ADRs are part of 4 SOCs: Immune system disorders (dark orange), Metabolism and nutrition disorders (orange), Skin and subcutaneous tissue disorders (purple), Vascular disorders (grey).

The three proteins have been identified as ACE, PGH₁ and PGH₂. Following the g:profile analysis this sub-network has been enriched with various functions, mainly related to *prostaglandin-endoperoxide synthase* and *peroxidase activity* molecular functions and *cyclooxygenase / prostaglandin regulation* biological process (figure 36).

GO:MF		stats						
Term name	Term ID	P _{adj}	$-\log_{10}(P_{adj})$	≤ 16	P12821	P23219	P35354	
prostaglandin-endoperoxide synthase activity	GO:0004666	8.943×10^{-7}						
peroxidase activity	GO:0004601	1.230×10^{-3}						
oxidoreductase activity, acting on peroxide as acceptor	GO:0016684	1.424×10^{-3}						
antioxidant activity	GO:0016209	3.335×10^{-3}						
dioxygenase activity	GO:0051213	3.980×10^{-3}						
heme binding	GO:0020037	9.161×10^{-3}						
tetrapyrrole binding	GO:0046906	1.048×10^{-2}						
oxidoreductase activity, acting on paired donors, with inc...	GO:0016705	1.203×10^{-2}						
bradykinin receptor binding	GO:0031711	1.670×10^{-2}						
peptidyl-dipeptidase activity	GO:0008241	2.505×10^{-2}						
tripeptidyl-peptidase activity	GO:0008240	2.505×10^{-2}						

1 to 11 of 11 | < < Page 1 of 1 > >

GO:BP		stats						
Term name	Term ID	P _{adj}	$-\log_{10}(P_{adj})$	≤ 16	P12821	P23219	P35354	
cyclooxygenase pathway	GO:0019371	6.760×10^{-4}						
regulation of blood pressure	GO:0008217	7.879×10^{-4}						
prostaglandin biosynthetic process	GO:0001516	6.484×10^{-3}						
prostanoid biosynthetic process	GO:0046457	6.484×10^{-3}						
unsaturated fatty acid biosynthetic process	GO:0006636	1.627×10^{-2}						
prostanoid metabolic process	GO:0006692	1.627×10^{-2}						
prostaglandin metabolic process	GO:0006693	1.627×10^{-2}						
blood circulation	GO:0008015	1.851×10^{-2}						
icosanoid biosynthetic process	GO:0046456	1.958×10^{-2}						
arachidonic acid metabolic process	GO:0019369	2.245×10^{-2}						
regulation of vasoconstriction	GO:0019229	2.473×10^{-2}						
circulatory system process	GO:0003013	2.841×10^{-2}						
vasoconstriction	GO:0042310	4.071×10^{-2}						

Figure 37. g:profiler results for cluster 12 analysis. The lowest p-value were obtained for the GO terms GO:0004666 (prostaglandin-endoperoxide synthase activity) and GO:0014806 (cyclooxygenase pathway).

The literature review reveals a strong link between the disruption of these functions and the development of such ADRs, which is mostly due to an inhibitory mechanism. The most well-known anaphylactic and anaphylactoid reactions to aspirin, the cyclooxygenase inhibitor for excellence, has been extensively investigated [122]. Selective COX-2 inhibitors and the onset of Hyperkalemia have also been proven [123]. Angioedema, Erythema multiforme and Dermatitis exfoliative, while being quite different adverse Events have been all three again related to a possible COX-2 inhibition [124] [125] [126]. Finally, even the event of vasculitis has been related to a COX-2 inhibition process [127].

5.4.4 - The Cluster-1 Analysis

As previously stated, Cluster₁ contains more than half of the network total nodes linked by 74 unique Uniprot IDs. The functional enrichment procedure performed on the total of Cluster₁ yielded significant results for very general molecular function like G-coupled receptors activity and serotonin receptor activity (Figure 38), such as in the case of Cluster₂.

For this reason, it was decided to perform again the clustering procedure on Cluster-1 which yielded successfully a number of modules or sub-clusters.

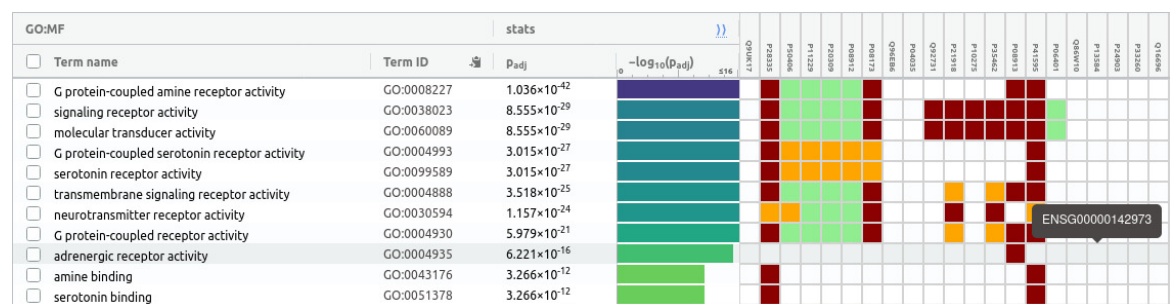


Figure 38. Significant molecular functions extracted from the functional enrichment procedure of the entire Cluster₁'s related proteins.

To avoid confusion with the findings of the first clustering round, the new clusters will be referred through Cluster 1-1 to 1-8. Even in this situation, the clusters are composed of various nodes, ranging from 75 ADRs in Cluster 1-1 to only 3 in Cluster 1-9. (Figure 39) Again, no specific distribution of SOCs can be identified.

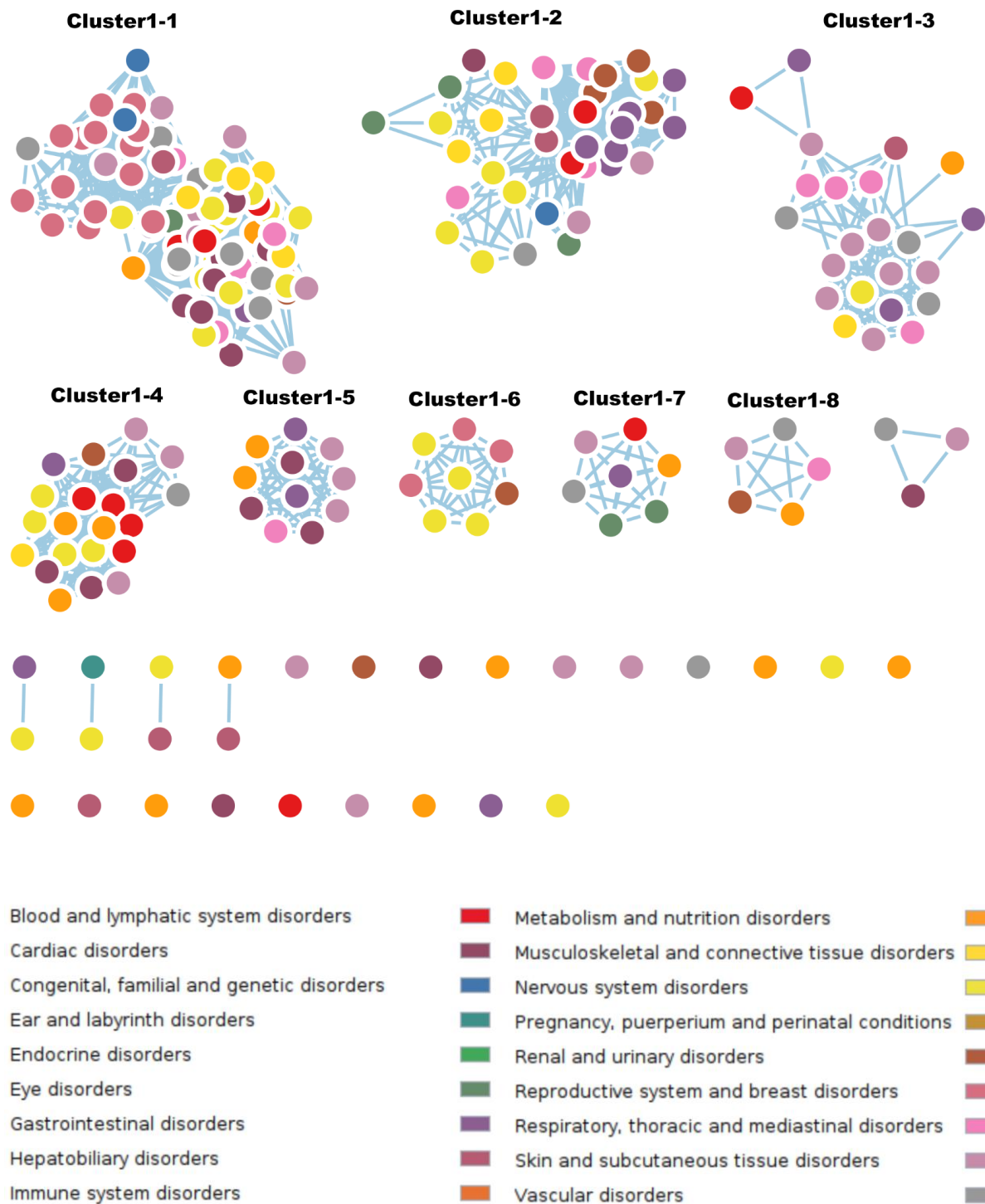


Figure 39. Results of the clustering procedure on Cluster₁ - The clusters are numbered from left to right in crescent order, with the largest in the leftmost upper denominated as Cluster₁₋₁. The nodes are classified by colors denoting the belonging of a distinct SOC.

The g:profiler analysis proved to be effective even in the case of sub-cluster. However, the proteins analyzed are majorly part of the tyrosine-kinase and G-coupled receptor family giving the obtained result a more generalist standpoint. Following are the most interesting cases that are not part of the mentioned super-families.

5.4.4.1 - Aromatase binding influences multisystemic ADRs onset.

Cluster 1-3 is one of the largest sub-clusters containing 24 ADRs that share 14 proteins mapped to the EGFR, ERBB, IGF1R, MTOR, P53, PGFRB, PGH2, THRB and VGFR2 families. (Figure 40)

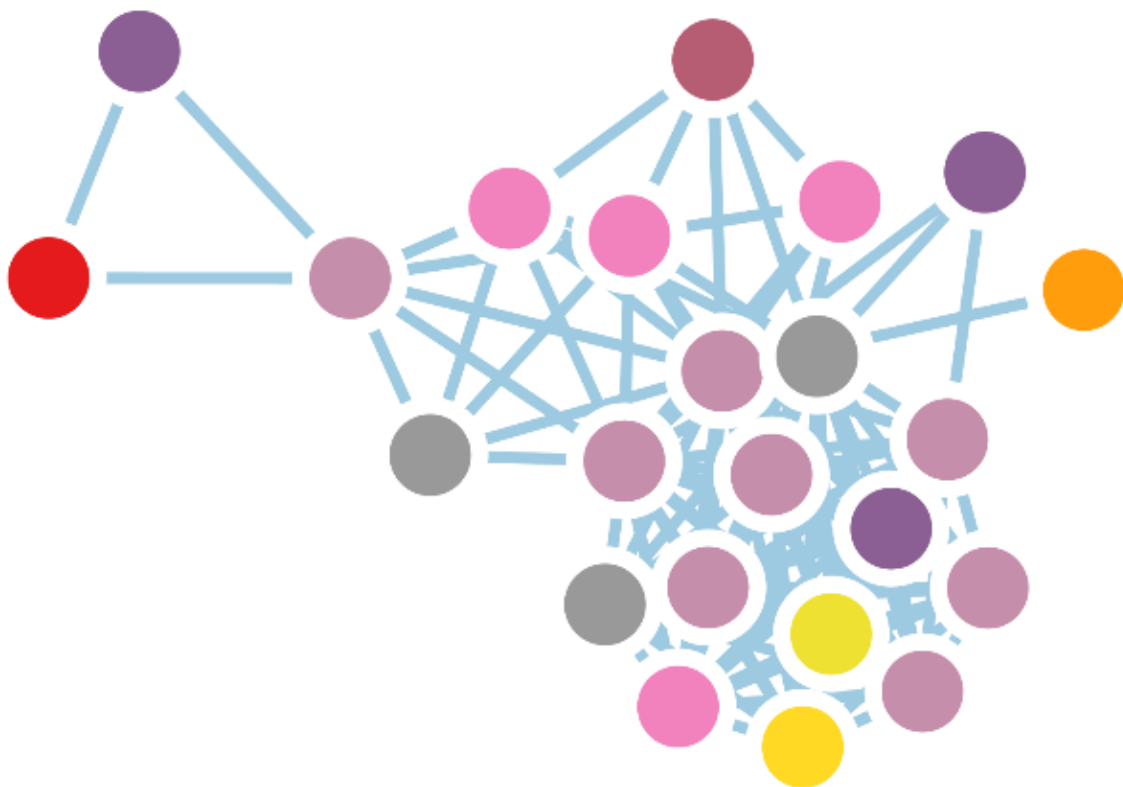


Figure 40. Cluster 1-3 resulting network. The nodes are presented without labels for visualization issues. The different ADRs are part of SOCs: Blood and lymphatic system disorders (red), Gastrointestinal disorders (dark purple), Hepatobiliary disorders (radish), Metabolism and nutrition disorders (light orange), Musculoskeletal and connective tissue disorders (light orange), Nervous system disorders (yellow), Respiratory, thoracic and mediastinal disorders (shocking pink), Skin and subcutaneous tissue disorders (violet), Vascular disorders (grey).

The subnetwork has been enriched with Aromatase and oxidoreductase activity with the g:profiler method. (Figure 41)

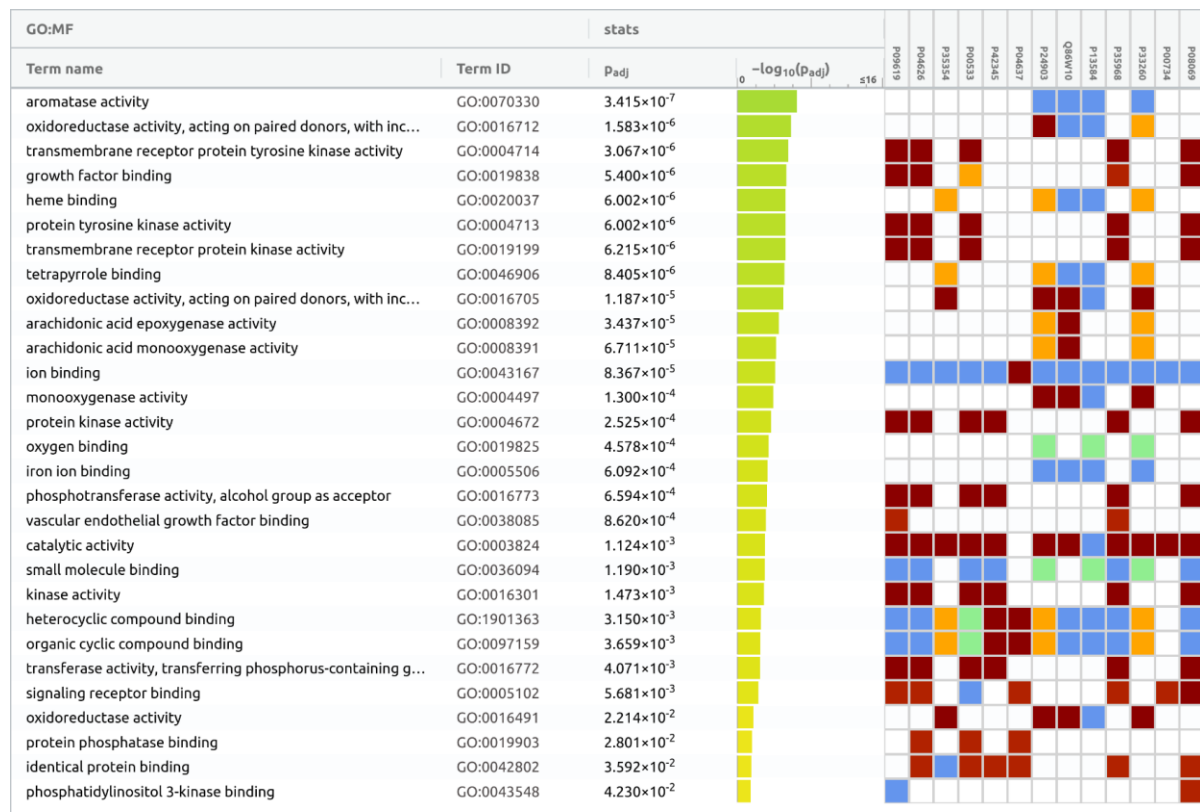


Figure 41. g:profiler results for Cluster1-3 analysis - The lowest p-value were obtained for the GO terms GO:0070330 (Aromatase activity) and GO:0016712 (Oxidoreductase activity). Biological processes results are not shown for visualization issues. The lowest p-values for the BP are obtained for the GO:0014065 (phosphatidylinositol 3-kinase signaling) and GO:0042327 (positive regulation of phosphorylation).

The ADR contained in this cluster ranges from different SOC: Coagulopathy (Blood and lymphatic system disorders), Gastrointestinal disorder, Retroperitoneal haemorrhage, Intestinal perforation (Gastrointestinal disorders), Cholangitis (Hepatobiliary disorders), Hypernatraemia (Metabolism and nutrition disorders), Musculoskeletal pain (Musculoskeletal and connective tissue disorders), Dysgeusia (Nervous system disorders), Interstitial lung disease, Epistaxis, Pulmonary toxicity, Respiratory distress (Respiratory, thoracic and mediastinal disorders), Alopecia, Skin necrosis, Erythema, Rash papular, Nail disorder, Blister, Rash vesicular, Skin disorder (Skin and

subcutaneous tissue disorders), Deep vein thrombosis, Angiopathy, Venous thrombosis (Vascular disorders).

Again, the literature review manages to link all of the reported ADRs with a perturbation of Aromatase or, as present in other clusters, Oxidoreductase Activity.

Table 8. List of publications for Cluster1-3 ADRs

Adverse Reaction	Publication
<i>Coagulopathy</i>	[128]
<i>Gastrointestinal disorder</i>	[129]
<i>Retroperitoneal haemorrhage</i>	[130]
<i>Intestinal perforation</i>	[131]
<i>Cholangitis</i>	[132]
<i>Hypernatraemia</i>	[133]
<i>Musculoskeletal pain</i>	[134]
<i>Dysgeusia</i>	[135]
<i>Interstitial lung disease</i>	[136]
<i>Epistaxis</i>	[137]
<i>Pulmonary toxicity</i>	[138]
<i>Respiratory distress</i>	[139]
<i>Alopecia</i>	[140]
<i>Skin necrosis</i>	[141]
<i>Erythema</i>	[142]
<i>Rash papular</i>	[143]
<i>Nail disorder</i>	[144]
<i>Blister</i>	[145]
<i>Rash vesicular</i>	[141]
<i>Skin disorder</i>	[141]
<i>Deep vein thrombosis</i>	[128]
<i>Angiopathy</i>	[146]
<i>Venous thrombosis</i>	[128]

5.4.4.2 - Bradykinin binding proved to be linked to inflammatory – related ADRs

Cluster 1-8 is one of the smallest sub-clusters constituted by only 5 ADRs and 2 proteins, being mapped to the ACE and REN1 gene family. The ADRs includes Chronic kidney disease, Gout, pemphigus, Flushing and Pulmonary eosinophilia. Even if these ADRs belong to different SOCs, it is evident that it can be directly correlated with a perturbation of the Renin-Angiotensin pathway, in particular as highlighted by the g:profiles results, to the Bradykinin binding. (Figure 42)



Figure 42. g:profiler results for Cluster1-8 analysis. The lowest p-value were obtained for the GO terms GO:0031711 (bradykinin receptor binding) and GO:0002016 (regulation of blood volume by renin - angiotensin).

In the case of chronic kidney disease in fact it has been proved that chronic overexpression of Bradykinin may lead to Kidney injury [147]. Same applies in the case of Gout, where an overexpression of Bradykinin receptors may extend the gout inflammatory process [148]. The bradykinin mediated inflammatory process appear to be at the basis of also the flushing ADR [149] and Pulmonary eosinophilia [150].

Even if it was not conclusive, this approach was successful in mining literature and in vivo research for clues on the logical sequence drug → protein → perturbation of function → ADR.

5.5 - Cluster's related drugs exploration

As we saw in the previous subsections, even seemingly disparate adverse events share proteins with common biological functions. At this point, with a new understanding of the potential molecular causes of ADR onset, it is reasonable to investigate the physical agents that cause this function disruption, i.e., the drugs. Many branded pharmaceutical agents are known to act on the same targets through different biochemical pathways or mechanical functions (dosage, route of prescription, biological half-life), achieving the same results. This is due to the fact that, while the manufacturing methods differ, the drug's effective stereochemical components are structurally similar.

This type of redundancy has already been addressed in T-ARDIS, where a chemical similarity screening was developed to reduce the number of false positives and duplicated drugs. The newly implemented data exploration step will investigate whether the drugs linked to the proteins found in the clusters, which also present the associated module's ADR, share any physical or chemical properties by utilizing the drug information and structure contained in STITCH 5.0. [36]

The drug's structure similarity has been assessed with the Rdkit python package [151], taking in consideration the drug's SMILE codes. In particular a pairwise Tanimoto scoring index has been performed distinguishing between "intra" and "outra" drugs. On the one hand, the 'outra' Tanimoto index is determined by calculating the Tanimoto score for drugs that act on different proteins that are still in the same cluster. (Figure 43 - A). On the other hand, the "intra" Tanimoto index is computed using the Tanimoto scores of different drugs that act on the same proteins in the cluster (Figure 43 - B).

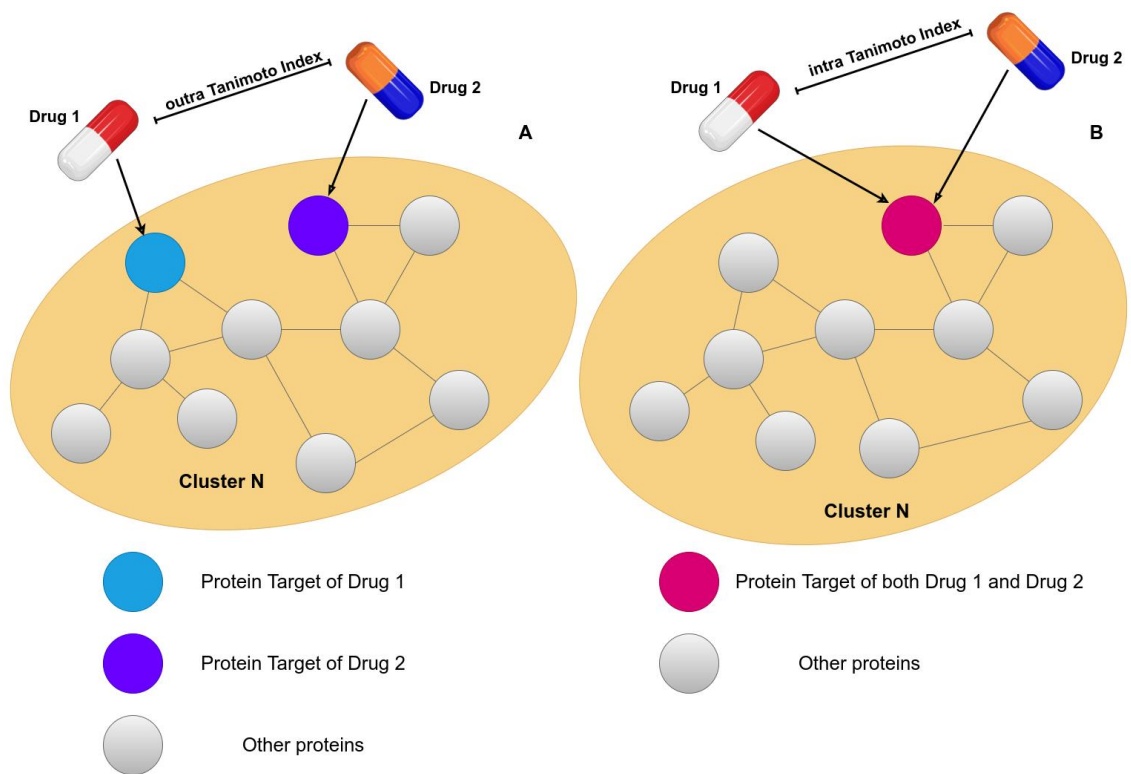


Figure 43. Definition of “outra” and “intra” Tanimoto Scores. The 'outra' Tanimoto index is determined by calculating the Tanimoto score for drugs that act on different proteins that are still in the same cluster. (A). The “intra” Tanimoto index is computed using the Tanimoto scores of different drugs that act on the same proteins in the cluster (B).

Following, the results of this analysis for some of the clusters already mentioned, Cluster₂, Cluster₄ and Cluster₁₂ (Figure 44).

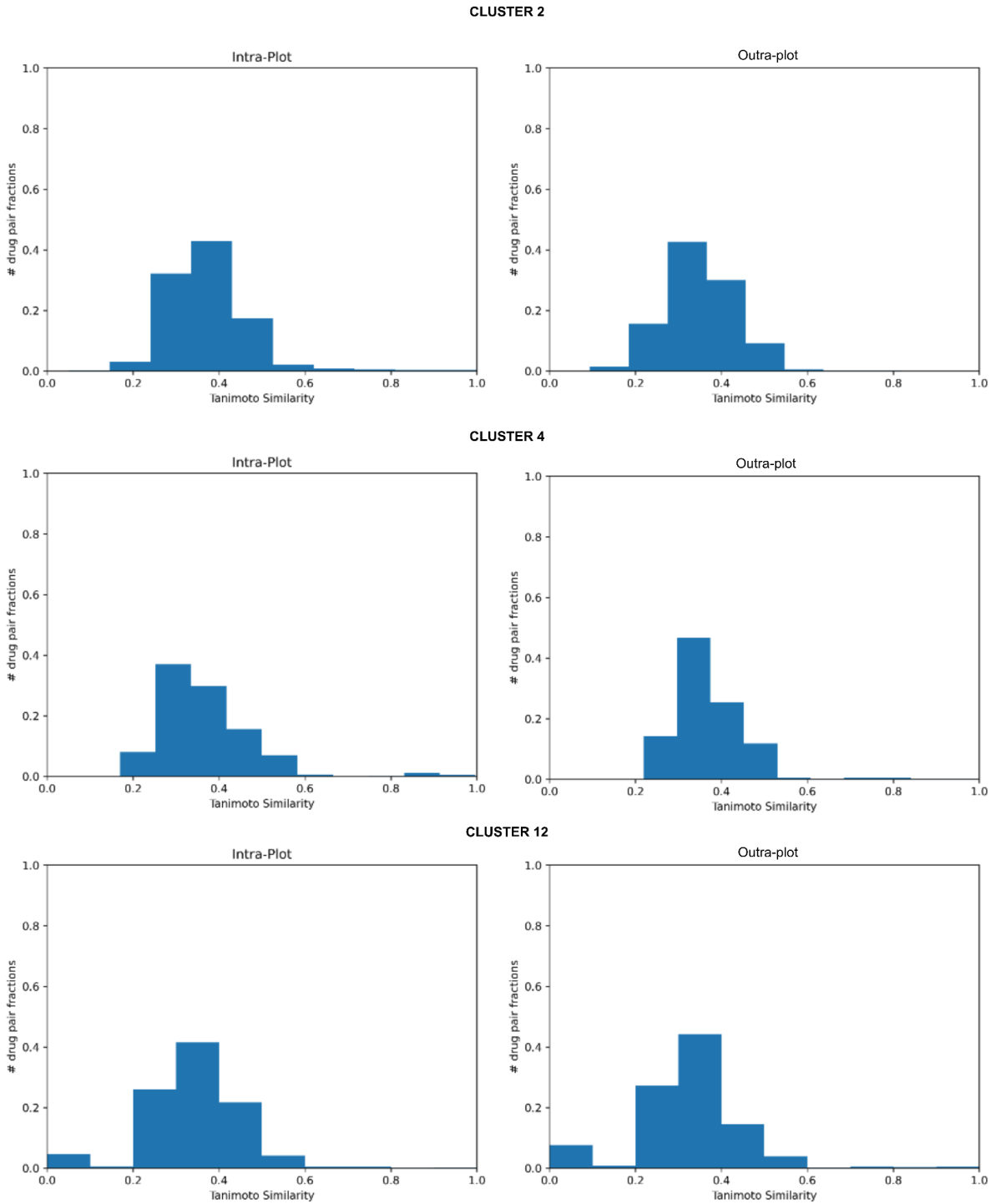


Figure 44. Distribution of “intra” ed “outra” Tanimoto scores in the different cluster.

The results obtained, as shown in figure 43, are not definitive; both the intra and outra approaches revealed a value distribution around the Tanimoto index of 0,3-0,5. Given that two drugs can be classified as comparable if their Tanimoto index is greater than

0,7, the enriched medications appear to lack structural similarity. This is probably due to the fact that structurally similar drugs are already being filtered in T-ARDIS. [11] and because, while the catalytic region is shared by all drugs in order for them to bind accurately to the target, the rest of the patented drug differ structurally. [2]

5.6 - Summary

The proposed SONG analysis was developed to try to answer the question of how ADRs occur. While much remains unknown about this subject, it appears that the answer must be sought at the molecular level of the drug's protein targets. SONG contributed to the effort by providing new perspectives on the mined T-ARDIS ADRs-protein relationships with its Adverse Reactome. This new network makes use of graph theory to identify specific groups of ADRs that share proteins and topological characteristics. The functional enrichment of the identified proteins allowed for a better molecular understanding of the potential biological pathway that, if disrupted, could result in the module's associated ADRs. While the scientific literature confirmed multiple examples of this theory, a study of the structural similarity of drugs targeting protein clusters revealed no direct relationships between the analyzed drugs. The reason for this could be reconducted directly to T-ARDIS, where similar drugs were removed to increase the statistical significance of the findings, and possibly to the various legislation of trademarked drugs. Despite this, SONG proved useful in investigating the potential causes of ADRs, and while many aspects remain unknown due to a lack of biological sources, it may prove to be an excellent tool for other researchers preparing to delve into this topic.

6 - GENERAL DISCUSSION

The relationship between Adverse Reactions and drug targets remains a difficult topic. There are several databases that relate ADRs-drugs and drug-protein targets, but the information contained still remains dispersed, making it difficult to interpret. Due to the obvious cost and research time savings, the ability to link adverse events to specific protein targets could be the first step toward more reliable and safer drugs. Statistical techniques and machine learning approaches arise in this context, giving tools to organize all of the data as well as algorithms to evaluate it and better understand the subtle link that connects medicines, the disruption of functions performed by the targeted proteins involved and the rise of Adverse Reactions. My thesis proposal fits right in this context, focusing on the development of tools and methodologies to better understand the Drug - Protein-ADR associations as well as the possible molecular mechanisms that characterize these relationships. From here on out, I'll continue to explore my thesis by describing the importance of my study to the area, its limits, and possible future advances.

6.1 - The T-ARDIS database: “*Allons-y*” towards the identification of ADR-Target relationships

Filling the knowledge gap between drug’s target and Adverse reaction emergence is one of the main objectives of this thesis. If correctly applied this information may give rise to a trickle-down effect that can directly influence the quality and cut the times of pharmaceutical research. For this purpose, as exposed in chapter 4, I developed the T-ARDIS database [11], a method to statistically identify the relationship between the modulation of proteins and the onset of Adverse reactions. The estimation of these relationships has been based on the mining, cleaning and filtering of various Drug – ADR and Drug – Target databases in tandem with statistical approaches that allowed to retrieve only meaningful associations removing misinformation and redundancy. The data contained in T-ARDIS has been validated by the existing literature, which supports the highly significant connections, i.e., low q-values, identified. This highlights the convenient mapping role of T-ARDIS and its discovered relationships can be beneficial as guiding evidence for drug repurposing or discovery.

Nonetheless the data contained in T-ARDIS is far from complete, unknown information in self-reporting (FAERS, MEDEFFECT) or curated databases (OFFSIDES, SIDER), lead directly to no correlations in T-ARDIS. It’s also possible that neither of the two drug-target databases employed in this investigation, DTC and STITCH, finds a relationship between the given drug and the target thus reducing the data coverage. This can be solved by constantly updating T-ARDIS in line with new database releases or by integrating information from other available resources. The database mining, statistical inference, and database updating are all entirely automated, ensuring that data is merged as it becomes available, further aiding our knowledge of ADR processes. Apart from the basic usage as a data repository, T-ARDIS has been used as a foundation for the development of DocTOR and SONG methods proving its value also as launching pad for other applications.

6.2 - The DocTOR approach: Precise predictions from blue black-box methods

While T-ARDIS is a valuable tool, as one of the few repositories investigating the relationship between ADRs and proteins, its structure is based solely on information found in publicly accessible databases, limiting the amount of data that can be retrieved. This can be directly translated to an increase in the amount of missing information and associations. DocTOR proposes a machine-learning-based method for predicting the association between ADRs and protein targets in order to integrate unknown information or to discover novel associations.

The DocTOR utility, in particular, presents itself as a tool for extrapolating relevant information from a network-based perspective, employing a combination of network and function-based measures to distinguish proteins that are associated with or completely unrelated to the various Adverse reactions. Unknown proteins can thus be framed as responsible or not for the onset of an ADR without having to mine through countless databases or run expensive experiments. DocTOR's power is based on a trickle-down strategy starting with the topological and functional information extracted from T-ARDIS, which are used as input for three machine-learning classifiers, which are in turn integrated by three distinct voting methods.

The model's evaluation methodology endorsed DocTOR performance at both the individual ADR and SOC levels, justifying its use as a general tool to predict potential protein vulnerabilities. Unfortunately, every approach has its own limitation, and DocTOR is no exception, for instance the massive computational power requirements for training and testing the models. As explained in chapter 5, this is also the reason why this study only managed to create models for 84 different ADR out of the thousands available. Limitations came also from the biological point of view, proving how complex are the role of particular proteins in the human interactome. In particular, the worst results have been obtained in 17 different ADRs which obtained a negative or equal to 0 MCC (random predictions). These includes *Hyper-coagulation*, *Ichthyosis*, *Coordination*

abnormal, Biliary cirrhosis, Acute hepatic failure, Hyper-ammonaemia, Azoospermia, Diplegia, Glucose tolerance impaired, Haemorrhagic diathesis, Hypoacusis, Ophthalmoplegia, Renal tubular acidosis, Hepatic failure, Coagulopathy and Ischaemia. These ADRs have been linked to 40 of the most highly connected genes in the human interactome, including TP53, 5HT1A, ACE, CALM family members, LEP, and IL8 [10]. These genes are associated with the majority of basic biological processes and serve as the foundation for many functions, making them directly related to the onset of various ADRs.

In spite of this, as shown in chapter 5 reliable results have been obtained with the different classifier, as in the case of ADR *malnutrition* where the random forest had the greatest results, with 0.95, 0.92, 1.00, and 0.91 for Accuracy, Precision, Recall, and MCC, respectively. In the case of ADR *febrile neutropenia*, however, NN was far and away the strongest predictor, with Accuracy, Precision, Recall, and MCC values of 0.80, 0.87, 0.70, and 0.77, respectively, compared to a virtually random prediction by SVM and RF (MCC 0.0). Finally, SVM surpassed the other two ML techniques in additional circumstances, such as *Nasal Congestion*, with an Accuracy of 0.90, Precision of 0.83, Recall of 1 and MCC of 0.81, whereas RF and NN barely reached 0.70.

The meta-predictor approach also proved its efficacy, strengthening the obtained results from the single ML methods in the case of single ADRs and SOCs as described in chapter 5, or at least for the *jury vote* and consensus systems. Indeed, the *red flag* method, which performs the worst as evidenced by the various scoring criteria, remains one of the main flaws of the implemented procedure. As counterintuitive as it may seem, the red flag method represents an attempt to develop a score system that would aid in detecting instances where the overall consensus system would fail. As mentioned in Chapter 5, this method was useful when two ML techniques implemented agreed but with low probability estimates. In addition, the red-flag approach acts as a failsafe in the event of an unknown prediction, such as when using the DocTOR utility.

From a tool usage standpoint, DocTOR lends itself to being rather easily updated, allowing the user to add new models for new ADRs on demand or retrain existing

models when new protein targets are identified to be related with certain ADRs and/or given new T-ARDIS database releases. The tool is nicely packed in a git-hub repository, so the maintenance and update, as well as the deployment is highly simplified. Unfortunately, as already mentioned, updates come with increased computation power costs, having to retrain the different features and models. Despite this, researchers will be able to use the DocTOR tool in conjunction with in vitro studies to evaluate the possible link between protein target modification and the development of ADR, cutting down on research time.

6.3 – SONG: Echoes from the ADRs' choirs

As described in Chapter 6, the SONG analysis is a procedure for determining the relationships between molecular functions and ADRs. The novel aspect of this approach is the unusual way in which T-ARDIS data is exploited, introducing a topological perspective on the ADR-protein relationships mined. This resulted in the development of a special network known as the "adverse reactome," which enabled the use of a variety of clustering and network-based measurements to identify groups of ADRs and proteins with similar properties. The proposed analysis identified 16 highly connected clusters, which have been enriched with meaningful functional annotations using the g:profile utility [103]. The identified molecular functions proved the direct onset of the ADRs in case of perturbation as confirmed by an extensive literature review.

SONG presents the same already discussed limitation of T-ARDIS, since it relies solely on the latter's data; missing links between ADRs and proteins, combined with constantly expanding knowledge of the human interactome, results in a limited view of the possible entire "adverse reactome" thus reducing modules identification.

As the analysis of modules revealed, for instance in the case of Cluster₁, another relevant limitation of the SONG approach resides in the many proteins that share common general functions, such as the previously mentioned G-coupled receptors. Since these functions play a critical biological role, their disruption results in an overabundance of ADRs, drastically reducing the significance of the associations.

SONG analyses also provide relationships between drugs that target the identified protein's modules and elicit the ADRs associated with the clusters. However, the computation and analysis of the pairwise Tanimoto index produced inconclusive results, highlighting different drug chemical structures. This is primarily because drugs with similar structures have already been filtered in T-ARDIS to reduce false negatives and redundancy. Another possibility is that patented drug structures differ from one

medication to the next due to copyright legislation while retaining the same catalytic pocket.

6.4 - Future perspectives and implication in the field of Protein-ADRs relationships

The identification and prediction of protein-ADR relationships has emerged as a powerful approach for limiting the cost and speed of drug discovery research. The information retrieved could also be broadened with the goal of better understanding the molecular complexity of ADRs and discovering better ways to prevent or exploit them, as in drug repurposing [152]. The notion to investigate the link between ADRs and protein is not new [7], but the amount of data now available has grown dramatically, allowing for a more in-depth understanding of the problem. Nonetheless, there are numerous restrictions and difficulties that must be addressed.

The first is (1) how to integrate different types of data to represent an interactome as completely as possible; This research highlights how ADRs are a multilayered problem that is dependent on the functions and interactions of various genes; a more in-depth understanding of the human interactome is the first step in extrapolating reliable information for the development of safer drugs.

(2) How to represent the signal of drug-induced network perturbations and the resulting ADRs. This thesis does not address the drug's effective activity (i.e., inhibitory or activation), including such information will help to better understand how the drugs selectively act on the function disruption that causes ADR emergence and how to prevent them.

(3) Testing different data architecture approaches; data dimensionality, computing time and memory processes are all issues that should not be overlooked. This thesis is a clear example of how, no matter how abundant the data, for computational time and requirements, only a small subset of ADRs could be analyzed. The future evolution of computational approaches, as well as data architectures in general, will undoubtedly make analysis faster and more reliable.

7 - CONCLUSIONS

In this thesis, I developed a set of networks and statistical-based in silico tools and studies to better understand the relationships between Drug's Adverse reactions and Drug's target. The knowledge's expansion in the protein-ADR interactions landscape has also aided in the investigation of the molecular basis of ADR and ADRs relationships.

The following conclusions can be drawn from the various methodologies used and results obtained:

- Pharmacovigilance and publicly accessible databases are critical in the study and control of the emergence of ADRs; however, the information contained must be cautiously used and validated before drawing conclusions
- The T-ARDIS development evidenced that the statistical inference of drug-protein and ADRs relationships can yield useful information. These associations are invaluable for drug discovery and can be further mined to extrapolate unique properties useful in a wide range of applications:
 - The T-ARDIS database, a repository of the identified ADR-protein relationship, is easily accessible and customizable for the advancements of drug discovery.
 - The DocTOR utility, a machine learning compendium which proved that ADR linked proteins are rich in topological and functional information, allowing them to characterize and mediate the ADR's emergence. These features have been used to predict whether or not unknown proteins are linked to specifically selected ADRs.
 - The SONG analysis, a network-based that considers the protein-ADR relationship in a more interconnected framework, improving understanding of the protein's functional properties and leading directly to the molecular basis of ADR and ADRs associations. The information

retrieved can be used to improve drug safety and avoid the perturbation of particular biological processes.

8 - BIBLIOGRAPHY

1. Waring, M.J., et al., *An analysis of the attrition of drug candidates from four major pharmaceutical companies*. *Nature Reviews Drug Discovery*, 2015. **14**(7): p. 475-486.
2. Talevi, A., *Multi-target pharmacology: possibilities and limitations of the “skeleton key approach” from a medicinal chemist perspective*. *Frontiers in Pharmacology*, 2015. **6**.
3. Bailey, J., M. Thew, and M. Balls, *An Analysis of the Use of Animal Models in Predicting Human Toxicology and Drug Safety*. *Alternatives to Laboratory Animals*, 2014. **42**(3): p. 181-199.
4. Singh, V.K. and T.M. Seed, *How necessary are animal models for modern drug discovery?* *Expert Opinion on Drug Discovery*, 2021. **16**(12): p. 1391-1397.
5. Zhang, J.-X., et al., *DITOP: drug-induced toxicity related protein database*. *Bioinformatics*, 2007. **23**(13): p. 1710-1712.
6. Huang, L.-H., et al., *ADReCS-Target: target profiles for aiding drug safety research and application*. *Nucleic acids research*, 2018. **46**(D1): p. D911-D917.
7. Ji, Z.L., et al., *Drug Adverse Reaction Target Database (DART)*. *Drug Safety*, 2003. **26**(10): p. 685-690.
8. Portanova, J., et al., *aer2vec: Distributed Representations of Adverse Event Reporting System Data as a Means to Identify Drug/Side-Effect Associations*. *AMIA Annu Symp Proc*, 2019. **2019**: p. 717-726.
9. Kuhn, M., et al., *Systematic identification of proteins that elicit drug side effects*. *Molecular Systems Biology*, 2013. **9**(1): p. 663.
10. Galletti, C., *Exploring the Drug - Adverse Reaction and Drug - Target Landscape through Networks, Statistics and Machine Learning Approaches - Supplementary materials*. 2022, Universitat central de Catalunya - Universitat de Vic.
11. Galletti, C., et al., *Mining drug-target and drug-adverse drug reaction databases to identify target-adverse drug reaction relationships*. *Database*, 2021. **2021**.
12. Hughes, J.P., et al., *Principles of early drug discovery*. *British Journal of Pharmacology*, 2011. **162**(6): p. 1239-1249.
13. Augen, J., *The evolving role of information technology in the drug discovery process*. *Drug Discovery Today*, 2002. **7**(5): p. 315-323.
14. Ratti, E. and D. Trist, *Continuing evolution of the drug discovery process in the pharmaceutical industry*. *Pure and Applied Chemistry*, 2001. **73**(1): p. 67-75.
15. Xie, L., L. Xie, and P.E. Bourne, *Structure-based systems biology for analyzing off-target binding*. *Current Opinion in Structural Biology*, 2011. **21**(2): p. 189-199.
16. Lee, H., *Genetically Engineered Mouse Models for Drug Development and Preclinical Trials*. *Biomolecules & Therapeutics*, 2014. **22**(4): p. 267-274.
17. Holbein, M.E., *Understanding FDA regulatory requirements for investigational new drug applications for sponsor-investigators*. *J Investig Med*, 2009. **57**(6): p. 688-94.
18. McBride, W.G., *THALIDOMIDE AND CONGENITAL ABNORMALITIES*. *The Lancet*, 1961. **278**(7216): p. 1358.
19. Edwards, I.R. and J.K. Aronson, *Adverse drug reactions: definitions, diagnosis, and management*. *Lancet*, 2000. **356**(9237): p. 1255-9.
20. Goh, K.-I., et al., *The human disease network*. *Proceedings of the National Academy of Sciences*, 2007. **104**(21): p. 8685-8690.

21. Barabási, A.-L., N. Gulbahce, and J. Loscalzo, *Network medicine: a network-based approach to human disease*. *Nature Reviews Genetics*, 2011. **12**(1): p. 56-68.
22. Somekh, J., et al., *A model-driven methodology for exploring complex disease comorbidities applied to autism spectrum disorder and inflammatory bowel disease*. *Journal of biomedical informatics*, 2016. **63**: p. 366-378.
23. Goh, K.-I. and I.-G. Choi, *Exploring the human diseasome: the human disease network*. *Briefings in Functional Genomics*, 2012. **11**(6): p. 533-542.
24. Menche, J., et al., *Uncovering disease-disease relationships through the incomplete interactome*. *Science*, 2015. **347**(6224): p. 1257601.
25. Aguirre-Plans, J., et al., *GUILDify v2.0: A Tool to Identify Molecular Networks Underlying Human Diseases, Their Comorbidities and Their Druggable Targets*. *Journal of Molecular Biology*, 2019. **431**(13): p. 2477-2484.
26. Ghiassian, S.D., J. Menche, and A.-L. Barabási, *A Disease Module Detection (DIAMOND) Algorithm Derived from a Systematic Analysis of Connectivity Patterns of Disease Proteins in the Human Interactome*. *PLOS Computational Biology*, 2015. **11**(4): p. e1004120.
27. Barrenäs, F., et al., *Highly interconnected genes in disease-specific networks are enriched for disease-associated polymorphisms*. *Genome Biology*, 2012. **13**(6): p. R46.
28. Choobdar, S., et al., *Assessment of network module identification across complex diseases*. *Nature Methods*, 2019. **16**(9): p. 843-852.
29. Cao, M., et al., *New directions for diffusion-based network prediction of protein function: incorporating pathways with confidence*. *Bioinformatics*, 2014. **30**(12): p. i219-i227.
30. Moore, N., et al., *Pharmacovigilance – The next chapter*. *Therapies*, 2019. **74**(6): p. 557-567.
31. Kumar, A., *The Newly Available FAERS Public Dashboard: Implications for Health Care Professionals*. *Hospital Pharmacy*, 2018. **54**(2): p. 75-77.
32. Re3data.Org, *MedEffect Canada - Adverse Reaction Database*. 2014.
33. Re3data.Org, *EU Clinical Trial Register*. 2020: p. 37.606-clinical trials; 18.700 paediatric trials.
34. Kuhn, M., et al., *The SIDER database of drugs and side effects*. *Nucleic Acids Research*, 2015. **44**(D1): p. D1075-D1079.
35. Tatonetti Nicholas, P., et al., *Data-Driven Prediction of Drug Effects and Interactions*. *Science Translational Medicine*, 2012. **4**(125): p. 125ra31-125ra31.
36. Szklarczyk, D., et al., *STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data*. *Nucleic acids research*, 2016. **44**(D1): p. D380-D384.
37. Tanoli, Z., et al., *Drug Target Commons 2.0: a community platform for systematic analysis of drug-target interaction profiles*. *Database*, 2018. **2018**.
38. Mozzicato, P., *MedDRA*. *Pharmaceutical Medicine*, 2009. **23**(2): p. 65-75.
39. Kass-Hout, T.A., et al., *OpenFDA: an innovative platform providing access to a wealth of FDA's publicly available data*. *Journal of the American Medical Informatics Association*, 2016. **23**(3): p. 596-600.
40. Wishart, D.S., et al., *DrugBank 5.0: a major update to the DrugBank database for 2018*. *Nucleic Acids Res*, 2018. **46**(D1): p. D1074-d1082.

41. Günther, S., et al., *SuperTarget and Matador: resources for exploring drug-target relationships*. Nucleic Acids Research, 2008. **36**(suppl_1): p. D919-D922.
42. Zhou, Y., et al., *Therapeutic target database update 2022: facilitating drug discovery with enriched comparative data of targeted agents*. Nucleic Acids Research, 2022. **50**(D1): p. D1398-D1407.
43. Kooistra, A.J., et al., *GPCRdb in 2021: integrating GPCR sequence, structure and function*. Nucleic Acids Research, 2021. **49**(D1): p. D335-D343.
44. Davis, A.P., et al., *Comparative Toxicogenomics Database (CTD): update 2021*. Nucleic Acids Research, 2021. **49**(D1): p. D1138-D1143.
45. Kanehisa, M. and S. Goto, *KEGG: Kyoto Encyclopedia of Genes and Genomes*. Nucleic Acids Research, 2000. **28**(1): p. 27-30.
46. Schaefer, C.F., et al., *PID: the Pathway Interaction Database*. Nucleic Acids Research, 2009. **37**(suppl_1): p. D674-D679.
47. Gillespie, M., et al., *The reactome pathway knowledgebase 2022*. Nucleic Acids Research, 2022. **50**(D1): p. D687-D692.
48. Karp, P.D., et al., *The BioCyc collection of microbial genomes and metabolic pathways*. Briefings in Bioinformatics, 2019. **20**(4): p. 1085-1093.
49. Gaulton, A., et al., *The ChEMBL database in 2017*. Nucleic acids research, 2017. **45**(D1): p. D945-D954.
50. Roth, B.L., et al., *The Multiplicity of Serotonin Receptors: Uselessly Diverse Molecules or an Embarrassment of Riches?* The Neuroscientist, 2000. **6**(4): p. 252-262.
51. Berman, H.M., *The Protein Data Bank: a historical perspective*. Acta Crystallogr A, 2008. **64**(Pt 1): p. 88-95.
52. *The Gene Ontology resource: enriching a GOLD mine*. Nucleic Acids Res, 2021. **49**(D1): p. D325-d334.
53. The UniProt, C., *UniProt: a worldwide hub of protein knowledge*. Nucleic Acids Research, 2019. **47**(D1): p. D506-D515.
54. Boeckmann, B., et al., *The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003*. Nucleic Acids Research, 2003. **31**(1): p. 365-370.
55. Huang, L., J. Zalkikar, and R. Tiwari, *Likelihood-Ratio-Test Methods for Drug Safety Signal Detection from Multiple Clinical Datasets*. Computational and Mathematical Methods in Medicine, 2019. **2019**: p. 1526290.
56. Huang, L., J. Zalkikar, and R.C. Tiwari, *A likelihood ratio test based method for signal detection with application to FDA's drug safety data*. Journal of the American Statistical Association, 2011. **106**(496): p. 1230-1241.
57. Yang, Y., et al., *Sixty-five years of the long march in protein secondary structure prediction: the final stretch?* Briefings in Bioinformatics, 2018. **19**(3): p. 482-494.
58. Dara, S., et al., *Machine Learning in Drug Discovery: A Review*. Artificial Intelligence Review, 2022. **55**(3): p. 1947-1999.
59. Kourou, K., et al., *Machine learning applications in cancer prognosis and prediction*. Computational and Structural Biotechnology Journal, 2015. **13**: p. 8-17.
60. Mirtskhulava, L., et al. *Artificial Neural Network Model in Stroke Diagnosis*. in *2015 17th UKSim-AMSS International Conference on Modelling and Simulation (UKSim)*. 2015.
61. Breiman, L., *Random Forests*. Machine Learning, 2001. **45**(1): p. 5-32.

62. Rosenblatt, F., *The perceptron: a probabilistic model for information storage and organization in the brain*. Psychol Rev, 1958. **65**(6): p. 386-408.
63. Qian, N. and T.J. Sejnowski, *Predicting the secondary structure of globular proteins using neural network models*. Journal of Molecular Biology, 1988. **202**(4): p. 865-884.
64. Silver, D., et al., *Mastering the game of Go with deep neural networks and tree search*. Nature, 2016. **529**(7587): p. 484-489.
65. Ciregan, D., U. Meier, and J. Schmidhuber. *Multi-column deep neural networks for image classification*. in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. 2012.
66. Fawcett, T., *An introduction to ROC analysis*. Pattern Recognition Letters, 2006. **27**(8): p. 861-874.
67. Banda, J.M., et al., *A curated and standardized adverse drug event resource to accelerate drug safety research*. Scientific data, 2016. **3**(1): p. 1-11.
68. Hripcsak, G., et al., *Observational Health Data Sciences and Informatics (OHDSI)*, in *15th World Congress on Health and Biomedical Informatics, MEDINFO 2015*. 2015, IOS Press. p. 574-578.
69. Nelson, S.J., et al., *Normalized names for clinical drugs: RxNorm at 6 years*. Journal of the American Medical Informatics Association, 2011. **18**(4): p. 441-448.
70. Ietswaart, R., et al., *Machine learning guided association of adverse drug reactions with in vitro target-based pharmacology*. EBioMedicine, 2020. **57**: p. 102837.
71. Nakamura, H., *[Cyclooxygenase (COX)-2 selective inhibitors: aspirin, a dual COX-1/COX-2 inhibitor, to COX-2 selective inhibitors]*. Nihon Yakurigaku Zasshi, 2001. **118**(3): p. 219-30.
72. Flower, R., *What are all the things that aspirin does?*, in *Bmj*. 2003. p. 572-3.
73. Shim, Y.K. and N. Kim, *Nonsteroidal anti-inflammatory drug and aspirin-induced peptic ulcer disease*. The Korean Journal of Gastroenterology, 2016. **67**(6): p. 300-312.
74. Muir, A. and I. Cossar, *Aspirin and ulcer*. British medical journal, 1955. **2**(4930): p. 7.
75. Gutierrez, M.A., G.L. Stimmel, and J.Y. Aiso, *Venlafaxine: a 2003 update*. Clinical therapeutics, 2003. **25**(8): p. 2138-2154.
76. Pauwels, R.A., et al., *Effect of inhaled formoterol and budesonide on exacerbations of asthma*. New England Journal of Medicine, 1997. **337**(20): p. 1405-1411.
77. Arntzenius, A. and L. van Galen, *Budesonide-related adrenal insufficiency*. Case Reports, 2015. **2015**: p. bcr2015212216.
78. Laugesen, K., et al., *Management of endocrine disease: Glucocorticoid-induced adrenal insufficiency: Replace while we wait for evidence?* European Journal of Endocrinology, 2021. **1**(aop).
79. Spencer, C.M., N.S. Gunasekara, and C. Hills, *Zolmitriptan*. Drugs, 1999. **58**(2): p. 347-374.
80. Salwinski, L., et al., *The Database of Interacting Proteins: 2004 update*. Nucleic Acids Research, 2004. **32**(suppl_1): p. D449-D451.
81. Smit, I.A., et al., *Systematic Analysis of Protein Targets Associated with Adverse Events of Drugs from Clinical Trials and Postmarketing Reports*. Chemical Research in Toxicology, 2020.

82. Sheffield, J.S., et al., *Designing Drug Trials: Considerations for Pregnant Women*. *Clinical Infectious Diseases*, 2014. **59**(suppl_7): p. S437-S444.
83. Mizutani, S., et al., *Relating drug-protein interaction network with drug side effects*. *Bioinformatics*, 2012. **28**(18): p. i522-i528.
84. Garcia-Garcia, J., et al., *Biana: a software framework for compiling biological interactions and analyzing networks*. *BMC Bioinformatics*, 2010. **11**(1): p. 56.
85. Orchard, S., et al., *The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases*. *Nucleic acids research*, 2014. **42**(Database issue): p. D358-63.
86. Keshava Prasad, T.S., et al., *Human Protein Reference Database—2009 update*. *Nucleic Acids Research*, 2008. **37**(suppl_1): p. D767-D772.
87. Stark, C., et al., *BioGRID: a general repository for interaction datasets*. *Nucleic Acids Research*, 2006. **34**(suppl_1): p. D535-D539.
88. Güldener, U., et al., *MPact: the MIPS protein interaction resource on yeast*. *Nucleic acids research*, 2006. **34**(Database issue): p. D436-D441.
89. Ceol, A., et al., *MINT, the molecular interaction database: 2009 update*. *Nucleic Acids Research*, 2009. **38**(suppl_1): p. D532-D539.
90. Guney, E. and B. Oliva, *Exploiting Protein-Protein Interaction Networks for Genome-Wide Disease-Gene Prioritization*. *PLOS ONE*, 2012. **7**(9): p. e43557.
91. Aguirre-Plans, J., et al., *An ensemble learning approach for modeling the systems biology of drug-induced injury*. *Biology Direct*, 2021. **16**(1): p. 5.
92. Hagberg, A.A., D.A. Schult, and P.J. Swart, *Exploring Network Structure, Dynamics, and Function using NetworkX*, in *Proceedings of the 7th Python in Science Conference*, G. Varoquaux, T. Vaught, and J. Millman, Editors. 2008. p. 11-15.
93. Blondel, V.D., et al., *Fast unfolding of communities in large networks*. *Journal of Statistical Mechanics: Theory and Experiment*, 2008. **2008**(10): p. P10008.
94. Bender, A., et al., *Analysis of Pharmacology Data and the Prediction of Adverse Drug Reactions and Off-Target Effects from Chemical Structure*. *ChemMedChem*, 2007. **2**(6): p. 861-873.
95. Pedregosa, F., et al., *Scikit-learn: Machine Learning in Python*. *Journal of Machine Learning Research*, 2011. **12**(85): p. 2825-2830.
96. Chollet, F. and Others, *Keras*. 2015.
97. Developers, T., *TensorFlow*. 2022, Zenodo.
98. McKinney, W., *Data Structures for Statistical Computing in Python*, in *Proceedings of the 9th Python in Science Conference*, S. van der Walt and J. Millman, Editors. 2010. p. 56-61.
99. Shaoqing, R., et al. *Global refinement of random forest*. in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015.
100. Oshiro, T.M., P.S. Perez, and J.A. Baranauskas. *How Many Trees in a Random Forest?* in *Machine Learning and Data Mining in Pattern Recognition*. 2012. Berlin, Heidelberg: Springer Berlin Heidelberg.
101. Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks*. *Genome Res*, 2003. **13**(11): p. 2498-504.
102. Frey Brendan, J. and D. Dueck, *Clustering by Passing Messages Between Data Points*. *Science*, 2007. **315**(5814): p. 972-976.

103. Raudvere, U., et al., *g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update)*. Nucleic Acids Research, 2019. **47**(W1): p. W191-W198.
104. Brinks, H.L. and A.D. Eckhart, *Regulation of GPCR signaling in Hypertension*. Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease, 2010. **1802**(12): p. 1268-1275.
105. Tang, Z.-Q. and Y. Lu, *Two GABAA responses with distinct kinetics in a sound localization circuit*. The Journal of Physiology, 2012. **590**(16): p. 3787-3805.
106. Penela, P., et al., *Effect of Hypothyroidism on G Protein-Coupled Receptor Kinase 2 Expression Levels in Rat Liver, Lung, and Heart**. Endocrinology, 2001. **142**(3): p. 987-991.
107. Itakura, T., et al., *GPR158 in the Visual System: Homeostatic Role in Regulation of Intraocular Pressure*. J Ocul Pharmacol Ther, 2019. **35**(4): p. 203-215.
108. Morshed, S.A., et al., *Biased signaling by thyroid-stimulating hormone receptor-specific antibodies determines thyrocyte survival in autoimmunity*. Science signaling, 2018. **11**(514): p. eaah4120.
109. Riddy, D.M., et al., *G Protein-Coupled Receptors Targeting Insulin Resistance, Obesity, and Type 2 Diabetes Mellitus*. Pharmacological Reviews, 2018. **70**(1): p. 39-67.
110. Vemula, S.R., et al., *Role of Gα(olf) in familial and sporadic adult-onset primary dystonia*. Human Molecular Genetics, 2013. **22**(12): p. 2510-2519.
111. Hernandez, G., et al., *Tardive dyskinesia is associated with altered putamen Akt/GSK-3β signaling in nonhuman primates*. Movement Disorders, 2019. **34**(5): p. 717-726.
112. Tao, Y.-X. and X.-F. Liang, *Chapter One - G Protein-Coupled Receptors as Regulators of Glucose Homeostasis and Therapeutic Targets for Diabetes Mellitus*, in *Progress in Molecular Biology and Translational Science*, Y.-X. Tao, Editor. 2014, Academic Press. p. 1-21.
113. Cao, W., et al., *Prostate specific G protein coupled receptor is associated with prostate cancer prognosis and affects cancer cell proliferation and invasion*. BMC Cancer, 2015. **15**(1): p. 915.
114. Haak, A.J., et al., *Targeting GPCR Signaling for Idiopathic Pulmonary Fibrosis Therapies*. Trends Pharmacol Sci, 2020. **41**(3): p. 172-182.
115. Krejner, A., et al., *Decreased expression of G-protein-coupled receptors GPR43 and GPR109a in psoriatic skin can be restored by topical application of sodium butyrate*. Archives of Dermatological Research, 2018. **310**(9): p. 751-758.
116. Chen, W., et al., *Smooth Muscle Hyperplasia/Hypertrophy is the Most Prominent Histological Change in Crohn's Fibrostenosing Bowel Strictures: A Semiquantitative Analysis by Using a Novel Histological Grading Scheme*. Journal of Crohn's and Colitis, 2017. **11**(1): p. 92-104.
117. Song, H.Y., et al., *Synovial fluid of patients with rheumatoid arthritis induces α-smooth muscle actin in human adipose tissue-derived mesenchymal stem cells through a TGF-β1-dependent mechanism*. Experimental & Molecular Medicine, 2010. **42**(8): p. 565-573.
118. Paik, Y.-H., et al., *Role of NADPH Oxidases in Liver Fibrosis*. Antioxidants & Redox Signaling, 2013. **20**(17): p. 2854-2872.

119. Nakanuma, Y., et al., *Pathology and Pathogenesis of Idiopathic Portal Hypertension with an Emphasis on the Liver*. Pathology - Research and Practice, 2001. **197**(2): p. 65-76.
120. Jiang, J.X. and N.J. Török, *NADPH Oxidases in Chronic Liver Diseases*. Advances in Hepatology, 2014. **2014**: p. 742931.
121. Raz, L., et al., *Role of Rac1 GTPase in NADPH Oxidase Activation and Cognitive Impairment Following Cerebral Ischemia in the Rat*. PLOS ONE, 2010. **5**(9): p. e12606.
122. Berkes, E.A., *Anaphylactic and anaphylactoid reactions to aspirin and other NSAIDs*. Clinical Reviews in Allergy & Immunology, 2003. **24**(2): p. 137-147.
123. Aljadhey, H., et al., *Risk of hyperkalemia associated with selective COX-2 inhibitors*. Pharmacoepidemiology and Drug Safety, 2010. **19**(11): p. 1194-1198.
124. Laouini, D., et al., *COX-2 inhibition enhances the TH2 immune response to epicutaneous sensitization*. J Allergy Clin Immunol, 2005. **116**(2): p. 390-6.
125. Thirion, L., A.F. Nikkels, and G.E. Piérard, *Etoricoxib-Induced Erythema-Multiforme-Like Eruption*. Dermatology, 2008. **216**(3): p. 227-228.
126. Kelkar, P.S., J.H. Butterfield, and H.G. Teaford, *Urticaria and angioedema from cyclooxygenase-2 inhibitors*. The Journal of Rheumatology, 2001. **28**(11): p. 2553.
127. Drago, F., et al., *Cutaneous vasculitis induced by cyclo-oxygenase-2 selective inhibitors*. Journal of the American Academy of Dermatology, 2004. **51**(6): p. 1029-1030.
128. Xu, X., et al., *Aromatase inhibitor and tamoxifen use and the risk of venous thromboembolism in breast cancer survivors*. Breast Cancer Research and Treatment, 2019. **174**(3): p. 785-794.
129. Khosrow-Khavar, F., et al., *Aromatase inhibitors and the risk of colorectal cancer in postmenopausal women with breast cancer*. Annals of Oncology, 2018. **29**(3): p. 744-748.
130. Sajnani, N. and D.B. Bogart, *Retroperitoneal hemorrhage as a complication of percutaneous intervention: report of 2 cases and review of the literature*. The open cardiovascular medicine journal, 2013. **7**: p. 16-22.
131. Fujii, Y., et al., *Bevacizumab-induced intestinal perforation in a patient with inoperable breast cancer: a case report and review of the literature*. Journal of Medical Case Reports, 2018. **12**(1): p. 84.
132. Murakami, K., et al., *Aromatase in normal and diseased liver*. Hormone Molecular Biology and Clinical Investigation, 2020. **41**(1).
133. Leschek, E.W., et al., *Effect of Antiandrogen, Aromatase Inhibitor, and Gonadotropin-releasing Hormone Analog on Adult Height in Familial Male Precocious Puberty*. J Pediatr, 2017. **190**: p. 229-235.
134. Tenti, S., et al., *Aromatase Inhibitors-Induced Musculoskeletal Disorders: Current Knowledge on Clinical and Molecular Aspects*. Int J Mol Sci, 2020. **21**(16).
135. Kan, Y., J. Nagai, and Y. Uesawa, *Evaluation of antibiotic-induced taste and smell disorders using the FDA adverse event reporting system database*. Scientific reports, 2021. **11**(1): p. 9625-9625.
136. Mascella, F., et al., *Aromatase inhibitors and anti-synthetase syndrome*. International journal of immunopathology and pharmacology, 2016. **29**(3): p. 494-497.

137. Lustberg, M.B., et al., *Randomized placebo-controlled pilot trial of omega 3 fatty acids for prevention of aromatase inhibitor-induced musculoskeletal pain*. Breast Cancer Research and Treatment, 2018. **167**(3): p. 709-718.
138. Yavas, G., et al., *Comparison of the effects of aromatase inhibitors and tamoxifen on radiation-induced lung toxicity: results of an experimental study*. Support Care Cancer, 2013. **21**(3): p. 811-7.
139. Stabile, L.P., et al., *Preclinical Evidence for Combined Use of Aromatase Inhibitors and NSAIDs as Preventive Agents of Tobacco-Induced Lung Cancer*. Journal of Thoracic Oncology, 2018. **13**(3): p. 399-412.
140. Freites-Martinez, A., et al., *Endocrine Therapy-Induced Alopecia in Patients With Breast Cancer*. JAMA Dermatol, 2018. **154**(6): p. 670-675.
141. Santoro, S., et al., *Aromatase inhibitor-induced skin adverse reactions: exemestane-related cutaneous vasculitis*. Journal of the European Academy of Dermatology and Venereology, 2011. **25**(5): p. 596-598.
142. Jhaveri, K., et al., *Erythema nodosum secondary to aromatase inhibitor use in breast cancer patients: case reports and review of the literature*. Breast Cancer Research and Treatment, 2007. **106**(3): p. 315-318.
143. Zarkavelis, G., et al., *Aromatase inhibitors induced autoimmune disorders in patients with breast cancer: A review*. Journal of Advanced Research, 2016. **7**(5): p. 719-726.
144. Lacouture, M. and V. Sibaud, *Toxic Side Effects of Targeted Therapies and Immunotherapies Affecting the Skin, Oral Mucosa, Hair, and Nails*. American journal of clinical dermatology, 2018. **19**(Suppl 1): p. 31-39.
145. Kim, Y.J. and P.R. Cohen, *Anastrozole-Induced Dermatitis: Report of a Woman with an Anastrozole-Associated Dermatitis and a Review of Aromatase Inhibitor-Related Cutaneous Adverse Events*. Dermatology and Therapy, 2020. **10**(1): p. 221-229.
146. Gara, E., et al., *Anti-cancer drugs-induced arterial injury: risk stratification, prevention, and treatment*. Medical Oncology, 2019. **36**(8): p. 72.
147. Barros, C.C., et al., *Chronic Overexpression of Bradykinin in Kidney Causes Polyuria and Cardiac Hypertrophy*. Frontiers in Medicine, 2018. **5**.
148. Shaw, O.M. and J.L. Harper, *Bradykinin receptor 2 extends inflammatory cell recruitment in a model of acute gouty arthritis*. Biochemical and Biophysical Research Communications, 2011. **416**(3): p. 266-269.
149. Luo, M.-C., et al., *Spinal Dynorphin and Bradykinin Receptors Maintain Inflammatory Hyperalgesia*. The Journal of Pain, 2008. **9**(12): p. 1096-1105.
150. Magalhães, G.S., et al., *Oral Formulation of Angiotensin-(1-7) Promotes Therapeutic Actions in a Model of Eosinophilic and Neutrophilic Asthma*. Frontiers in Pharmacology, 2021. **12**.
151. Landrum, G., *Rdkit documentation*. Release, 2013. **1**(1-79): p. 4.
152. Pushpakom, S., et al., *Drug repurposing: progress, challenges and recommendations*. Nature Reviews Drug Discovery, 2019. **18**(1): p. 41-58.

9 – PAPERs ANNEX

Mining drug–target and drug–adverse drug reaction databases to identify target–adverse drug reaction relationships

Cristiano Galletti¹, Patricia Mirela Bota^{1,2}, Baldo Oliva² and Narcis Fernandez-Fuentes^{1,*}

¹Department of Biosciences, U Science Tech, Universitat de Vic-Universitat Central de Catalunya, Carrer Laura 13, Vic, Catalonia 08500, Spain

²Department of Experimental and Health Sciences, Structural Bioinformatics Group, Research Programme on Biomedical Informatics, Universitat Pompeu Fabra, Barcelona, Catalonia 08003, Spain

*Corresponding author: Tel: +34 938885519; Fax: +34 938861213; Email: narcis@bioinsilico.org

Citation details: Galletti, C., Mirela Bota, P., Oliva, B. *et al.* Mining drug–target and drug–adverse drug reaction databases to identify target–adverse drug reaction relationships. *Database* (2021) Vol. 2021: article ID baab068; DOI: <https://doi.org/10.1093/database/baab068>

Abstract

The level of attrition on drug discovery, particularly at advanced stages, is very high due to unexpected adverse drug reactions (ADRs) caused by drug candidates, and thus, being able to predict undesirable responses when modulating certain protein targets would contribute to the development of safer drugs and have important economic implications. On the one hand, there are a number of databases that compile information of drug–target interactions. On the other hand, there are a number of public resources that compile information on drugs and ADR. It is therefore possible to link target and ADRs using drug entities as connecting elements. Here, we present T-ARDIS (Target–Adverse Reaction Database Integrated Search) database, a resource that provides comprehensive information on proteins and associated ADRs. By combining the information from drug–protein and drug–ADR databases, we statistically identify significant associations between proteins and ADRs. Besides describing the relationship between proteins and ADRs, T-ARDIS provides detailed description about proteins along with the drug and adverse reaction information. Currently T-ARDIS contains over 3000 ADR and 248 targets for a total of more 17 000 pairwise interactions. Each entry can be retrieved through multiple search terms including target Uniprot ID, gene name, adverse effect and drug name. Ultimately, the T-ARDIS database has been created in response to the increasing interest in identifying early in the drug development pipeline potentially problematic protein targets whose modulation could result in ADRs.

Database URL: <http://www.bioinsilico.org/T-ARDIS>

Introduction

One of the main major problems faced in drug development is the lack of toxicology or safety information for targets (1). This fact results in a high level of attrition of drugs entering clinical trials due to the severity of adverse drug reactions (ADRs) associated with toxicity, significantly increasing the costs and therefore limiting the development of novel drugs for emerging targets (2). One of the most conventional methods in past years relied on the use of animal models. However, animal models imply high maintenance cost and ethical drawbacks and not always transferable to human biology (3), and thus computational approaches can provide useful predictions.

There are a number of approaches that can be used to decrease the risk associated with the development of novel drugs from a drug-centric point of view. *In-silico* approaches have demonstrated their utility in estimating the toxicity of drug candidates, exploiting features such as composition, structure and binding affinity. These methods include various examples of machine learning and deep learning (4). Other studies are based on target-based predictions, analy-

ses of the underlying protein network and interactions and quantitative structure–activity relationships. The latter have been used to model numerous drug safety endpoints including drug lethal dose of 50%, the so-called LD50 values, skin/eye irritation and tissue-specific toxicity, making it one of the most used parameters for estimating the toxicity of a drug (5). The use of curated protein target sets, conforming so-called safety panels, are also used to assess the potential liability of novel drugs during pre-clinical stages (6). Finally, information about potential liability of drugs can be also obtained post-development in the context of pharmacovigilance including a number of approaches that mine information for a range of databases such the Food and Drug Administration (FDA) spontaneous reporting systems database (5, 7, 8).

All the methods presented above are drug-centric, i.e. the prediction of potential ADR is based solely on the properties of the drug but not on the putative or known protein targets. In fact, while there are well-established methodologies and resources, as shown above, to associate drugs to ADR, it is less so to associate ADR to protein targets. Examples of the latter include the ADReCS-Target database (9) and a recent

study on ADRs compiled from clinical trials and post marketing reports (10). A different take on the issue would be to identify the link between ADR and proteins, using drugs as a connecting element. In principle, the idea is very straightforward: if drug X causes ADR Y and drug X binds to protein Z, then protein Z is related to ADR Y. This simple statement is, however, incorrect. As pointed out by Kuhn and colleagues (11), most drugs bind to sets of pharmacologically similar proteins, for example, members of the same protein family. While it is likely that only one of the targets is responsible for a given ADR, a direct Target–ADR association, as in this simple approach, would relate each target to each possible ADR of the same drug, creating erroneous or non-existent relationships, i.e. false positives. This association needs to be validated statistically, and the method described by Kuhn *et al.* (11) provides a defined path to identify statistically significant associations between ADR and proteins using drugs as the connecting elements.

T-ARDIS (Target—Adverse Reaction Database Integrated Search), the database presented here, contains statistically validated associations between protein targets and potential ADR derived from the association drug–ADR and drug–protein. In the first stage, drug–ADR and drug–protein associations were mined from different databases. In the case of drug–protein, the databases included the Drug–Target Commons (12) and STITCH (13) databases. Drug–ADR associations were mined from FDA Adverse Event Reporting System (FAERS) (14), MEDEFECT (15), SIDER (16) and OFFSIDES (17). Upon mining, by parsing and filtering these databases, the associations between proteins and ADRs were established using the method described by Kuhn *et al.* (11) as described above. The results are therefore a number of protein–ADR associations that are statistically significant and that can be of use as complement to other approaches to identify potential liabilities associated with protein targets.

Currently, T-ARDIS compiles over 3000 ADRs associated with over 200 proteins. Users can easily access the data searching by the drug name (common name), type of ADR as defined in MedDRA dictionary (18) or the protein UNIPROT (19) identification code or gene name. The results are returned in a tabular form listing the principal descriptor for each entry such as the drug name, the target UniProt ID, gene name, the MedDRA classification for ADR, together with the results of the statistical validation (*P*-value of association and its correction for multiple testing, *q*-value, including the contingency table used). Moreover, it will be possible to access external links to the native drug target or drug–ADR database, together with related repositories.

Material and methods

Databases containing drug–ADR information

Four different databases were parsed and mined to identify drug–ADR associations: OFFSIDES (17), SIDER4.1 (16), MEDEFECT (15) and FAERS (14). OFFSIDES is a manually curated database available at <http://tatonnettilab.org/resources/nsides/>. SIDER4.1 is a database of drugs, ADR and indications mined from the FDA drug labels. The version used in this study is SIDER4.1 released 21 October 2015 available at <http://sideeffects.embl.de/>. The FAERS or AERS is a centralized pharmacovigilance database developed

to integrate the U.S. FDA's post marketing safety surveillance program. The data stored in this database represent one of the major repositories regarding drug–ADR relationships, although it requires a curation before that can be used (see below 'Curation of FAERS database'). The version included in T-ARDIS was last updated in March 2020 and is available at: <https://fis.fda.gov/extensions/FPD-QDE-FAERS/FPD-QDE-FAERS.html>. Finally, the MEDEFECT, Canada's sister database of the FAERS. Adverse reaction reports are submitted by consumers and health professionals, who submit reports voluntarily, and manufacturers and distributors (also known as market authorization holders), who are required to submit reports according to the Canadian Food and Drugs Act. The version of MEDEFECT included in T-ARDIS was updated in May 2020 and is accessible at <https://www.canada.ca/en/health-canada/services/drugs-health-products/medeffect-canada/adverse-reaction-database/canada-vigilance-online-database-data-extract.html>.

The adverse event report descriptions are coded as medical terms as defined in the MedDRA vocabulary and ontology (18). The entries in MedDRA are reported using five hierarchical levels of medical terminology, ranging from a very general System Organ Class (SOC—e.g. gastrointestinal disorders) term to a very specific Lowest Level Term (e.g. feeling queasy). Each term is linked to only one term on a higher level. For each drug–ADR database, we manually checked that all adverse reactions were registered as MedDRA Reaction terms at Preferred Term (PT) level that describes a single medical concept. We also used the SOC definition of MedDRA to filter unspecific ADR (see the 'Filtering of ADR based on SOCs' section).

Curation of FAERS and MEDEFECT databases

Prior to using the data present on the FAERS database, a curation of the records was performed. This step is required due to the heterogeneity in the reports as these are uploaded directly by health-care professionals (physicians, pharmacists, nurses and others) and other actors (patients, family members, lawyers and others.) Thus, the quality of the reports varies substantially and there are often typos (e.g. misspelled drug names), missing information and other errors. To obtain a curated and standardized version of FAERS and MEDEFECT, we relied on a modified pipeline specially developed for the standardization of FAERS records (20) and adapted to MEDEFECT. In particular, this pipeline uses standardized vocabularies with drug names mapped to RxNorm concepts (21) and exploits the demographic information on the patients in order to remove duplicates. To identify statistically significant associations between drugs and ADRs, the method proposed by Huang *et al.* (22). was applied to the resulting databases originating from the standardization pipeline described above. Finally, only those drug–ADR associations that are statistically significant, i.e. the likelihood ratio value is above the 5th percentile of the multinomial distribution, and present both in FAERS and MEDEFECT were kept.

Filtering of ADR based on SOCs

Some of the ADRs reported are very general or not specific to body parts, tissues or underlying human biology. For this

reason and as described in (23), any ADR belonging to the following SOCs were discarded.

General disorders and administration site conditions

As the name suggests, this SOC contains terms that do not readily fit into the hierarchy of any one SOC or are non-specific disorders that impact several body systems or sites. To be noted that representing PTs in this SOC in each potential secondary SOC would create an inordinately large number of redundancies. Therefore, most of the PTs in this SOC are primarily linked to SOC General disorders and administration site conditions and have limited representation in secondary SOCs (e.g. PT Injection site atrophy is primarily to SOC General disorders and administration site conditions and secondarily only to SOC injury, poisoning and procedural complications).

Injury, poisoning and procedural complications

This SOC provides a grouping for those medical concepts where an injury, poisoning, procedural or device complication factor is significant in the medical event being reported. As a general rule, in this SOC all the events appear directly attributed to trauma, poisoning and procedural complications, in other words, all the events due to an external cause.

Investigations

For MedDRA, an 'investigation' is a clinical laboratory test concept (including biopsies), radiologic test concept, physical examination parameter and physiologic test concept (e.g. pulmonary function test). Only PTs representing investigation procedures and qualitative results (e.g. PT blood sodium decreased, PT blood glucose normal) appeared in this SOC. Terms representing conditions (e.g. hyperglycemia) or mixed concepts of conditions with an investigation are excluded from this SOC and can be found in the respective 'disorder' SOCs (e.g. PT hyperosmolar state, PT haemosiderosis, PT orthostatic proteinuria and PT renal glycosuria).

Neoplasms benign, malignant and unspecified (incl.cysts and polyps)

This SOC is classified anatomically, with pathologic subclassifications for staging of both benign and malignant neoplasms.

Product issues

This SOC includes terms relevant for issues with product quality, devices, manufacturing quality systems, product supply and distribution and counterfeit products.

Social circumstances

The purpose of this SOC is to provide a grouping for those factors that may give insight into personal issues that could have an effect on the event being reported. Essentially, this SOC contains information about the person, not the adverse event. As an example, terms such as PT drug abuser and PT death of relative are found in this SOC.

Surgical and medical procedures

This SOC contains only those terms that are surgical or medical procedures. The nature of this SOC makes it more

of a 'support' SOC for recording case information and for developing queries.

Infections and infestations

This SOC just provides information on location linked to infectious disorders but not to specific targets.

Psychiatric disorders

The following high-level general terms and high-level terms were excluded from this specific SOC due to being too general and/or broad. These included the terms: depressed mood disorders and disturbances; eating disorders and disturbances; impulse control disorders not elsewhere classified (NEC); manic and bipolar mood disorders and disturbances; personality disorders and disturbances in behaviour; psychiatric disorders NEC; suicidal and self-injurious behaviours NEC; paraphilias and paraphilic disorders and sexual and gender identity disorders NEC.

Databases containing drug-protein information

Two different databases were used to extract drug-protein associations. These include Drug-Target Commons (DTC) database (<https://drugtargetcommons.fimm.fi>) (12). The DTC aims at providing an open-data platform for a community-driven crowd-sourcing effort to annotate drug-target associations and provides information on drugs' bioactivity such IC50, EC50 and potency values. The version included in T-ARDIS was downloaded in April 2021 from <https://drugtargetcommons.fimm.fi>. The second database considered was STITCH (13). STITCH provides a complementary view on drug-target associations as it relies on different sources of information combined into a composite scoring function (24). The version included in T-ARDIS is 5.0 and is accessible at <http://stitch.embl.de>.

The starting databases were subjected to two filter steps to ensure that biologically/therapeutically relevant associations are captured and that redundant entries originating from the same drug been named differently. The Uniprot ID was used to ensure that the target was the same in both databases. DTC provide already this information for each pair drug-target but in the case of STITCH the Uniprot ID was retrieved programmatically from the Uniprot database (19) using the STRING (25) identification code. In the case of DTC, only drug-protein association with a reported IC50 (or EC50) of 100 nM or better was considered. In the case of the STITCH database, a cut-off of 0.8 was applied, thus only association with a better score was considered. To avoid redundancy, the drug entries were unified using the InChIKey hash descriptors and the drug's standard name ensuring that not redundant entries appear in the consolidated dataset.

Statistical association protein-ADR using drug-protein and drug-ADR relationships

The statistical significance of ADR-protein associations was calculated following the method proposed by Kuhn *et al.* (11). In a nutshell, the method computes a contingency matrix for each ADR-protein pair and calculates the *P*-value using Fisher's exact test. The elements of the contingency matrix are as follows: (i) the number of drugs that present the given ADR; (ii) the number of drugs that binds to the given protein; (iii) the number of drugs that both present the given ADR and

bind to the given protein and (iv) how many drugs neither present the ADR nor bind to the given target. Given the high number of relationships, P -values were corrected for multiple testing using the ‘ q -value’ module contained in the python

package ‘MultyPy’ (26). An ADR–protein relationship was accepted if the computed q -value is equal or smaller than 0.05. Figure 1 shows an outline of this annotation approach, from the mining of individual databases to statistical association.

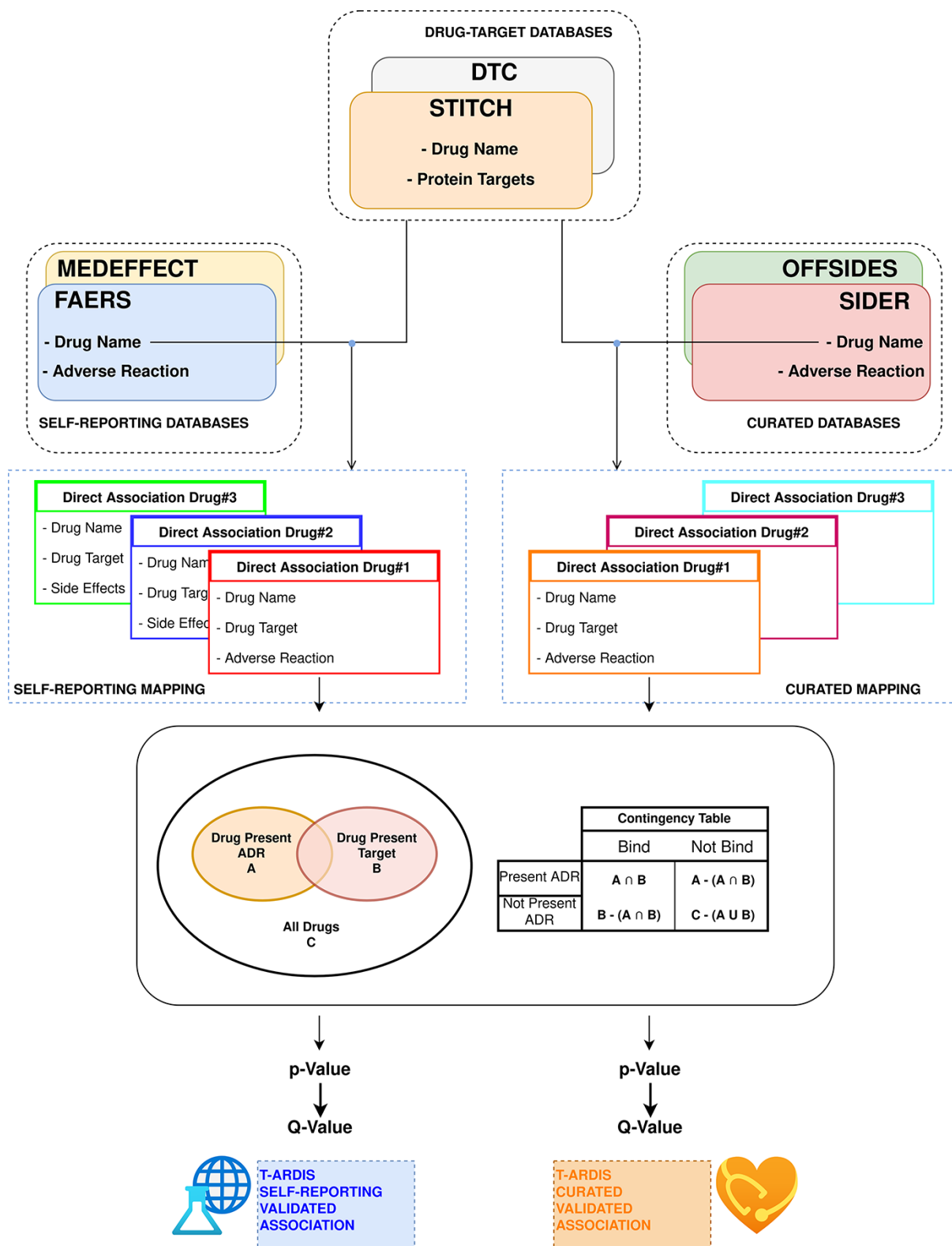


Figure 1. Workflow followed to combine and derive statistical associations between proteins and ADR. Drug–ADR and drug–target associations are retrieved from relevant databases. Subsequently, statistical association between proteins and ADRs is computed as described by Kuhn *et al.* (10).

Prior to the calculation of protein–ADR statistical associations, the drug–ADR databases were divided in two different sets: curated and self-reporting drug–ADR association. The curated included drug–ADR associations extracted from SIDER and OFFSIDES, while the self-reporting set included drug–ADR association from FAERS and MEDEFECT. The logic follows on distinguish between these two groups as the origin of information is very different as mentioned above. Therefore, the statistical associations between protein–ADR present in T-ARDIS originate from any of these two sets as the drug–target associations are common to both, i.e. DTC and STITCH databases. The unifying entity between drug–protein and drug–ADR is of course the drug entity, and the unification between both groups was done the using the drug’s standard name. To make sure an unequivocal association, a Tanimoto 2D chemical similarity score was computed with a cut-off of 0.7 using the Rdkit Conda package (27). Finally, drugs presenting less than 10 ADRs were also discarded.

In the case of the drug–target databases, a filtering procedure was implemented as described in Kuhn *et al.* (11). First, proteins related to drug metabolism were discarded. These were selected using the Gene Ontology annotation (28), and thus proteins belonging to GO terms: GO:0042737 (drug catabolic process) and GO:0017144 (drug metabolic processes) were discarded. Second, a sequence similarity filter was implemented to remove highly redundant proteins using CD-HIT (29) at 90% sequence identity cut-off. A subsequent clustering step was devised to group proteins into families using a sequence identity cut-off of 70% and families with more than 10 members for same drug were excluded preserving just the association with the centroid of the cluster. Finally, as discussed in Kuhn *et al.* (11), for each of the protein–ADR groups, the main target was identified as reported (30) and the rest of the members of the group were kept if sharing at least 50% of the drugs binding to the main target.

Benchmarking datasets

Four different datasets were used to compare the associations uncovered by T-ARDIS. The first set was extracted from the ADRCS-Target database (9) from which 1710 protein–ADR top scoring associations were compiled. The second set derives from the recent work by Smit *et al.* (10) that albeit containing an older release of SIDER (ver.3) was used to extract circa 2000 protein–ADR associations. The third set relates to a set of 225 pairwise interactions validated in the work of Kuhn *et al.* (11). Finally, the fourth set is a manually curated set mined for scientific publications presented in the work by Kuhn *et al.* (11), which includes 816 protein–ADR associations (Table 1).

Results

Combining different databases increases the coverage of associations

We first consider the databases with drug–ADR associations. As described in the ‘Materials and methods’ section, the nature and purpose as well as the level of curation of these databases vary. There is a core of drug–ADR associations, which are common to all databases (Figure 2). The overlap between OFFSIDES and FAERS databases is relatively high and expected as drug–ADR associations annotated in

Table 1. Comparison of different datasets and T-ARDIS

SET	# Associations	Self-reporting ^a	Curated ^b
Associations mined from the literature in Kuhn <i>et al.</i> (11)	224	27 (4)	17 (6)
Associations validated in vivo in Kuhn <i>et al.</i> (11)	2170	115 (69)	113 (85)
Associations described in Smit <i>et al.</i> (10)	2153	340 (48)	297 (167)
Associations from ADRCD-Target database (9)	816	171 (14)	87 (11)

^aAssociations present in the self-reporting set of T-ARDIS; significant associations shown within parentheses (q -values < 0.05).

^bAssociations present in the curated set of T-ARDIS; significant associations shown within parentheses (q -values < 0.05).

OFFSIDES are subsequently added to FAERS on new releases. FAERS and MEDEFECT rely on multiple sources and spontaneous reporting systems and contain the largest number of drugs–ADRs associations as well as the largest percentage of unique entries. Following the curation approach, over 4 million pairwise interactions originating from over 9000 compounds and around 17 000 unique ADR were obtained from FAERS. In the case of MEDEFECT, 1.5 M drug–ADR associations were uncovered from a total of over 4000 and 12 000 drugs and ADR events annotated in the database, respectively.

Unlike FAERS and MEDEFECT, SIDER and OFFSIDES contain manually curated associations of drugs and ADRs. These databases have a lower number of associations when compared to spontaneous reporting databases FAERS and MEDEFECT (between 1 and 2 orders of magnitude less). In the case of SIDER, over 108 000 pairwise interactions were mined for a total of 1344 unique drugs and 2303 ADRs. OFFSIDES yielded a large number of pairwise drugs–ADR associations: 1.5 M associations from a total of 2708 and 4368 unique drugs and ADRs. In terms of uniqueness of information, FAERS and MEDEFECT show a larger percentage of shared drugs between the different databases (Figure 2).

The second group of databases considered were those describing drug–protein target associations including DTC (12) and STITCH (13). The nature of both databases is rather different and so it is reflected in the number of associations extracted from each individual database. In the case of STITCH, over 10 000 drug–target associations were retrieved after applying the filter described in the ‘Materials and methods’ section accounting for 5007 and 1075 different drug and chemical compounds and proteins (as per Uniprot IDs), respectively. In the case of STITCH, the number of associations was much larger: over 6 M from over 42 000 chemical compounds (including approved drugs) and 7264 different proteins. The overlap between both databases in terms of shared drugs was around 1600.

Proteins–ADR relationship from mined drug–ADR and drug–protein associations

After curation of drug–target and drug–ADR database and filtering, the associations between proteins and ADRs were obtained. The association was based on the drug entities shared among the databases. It is important to stress

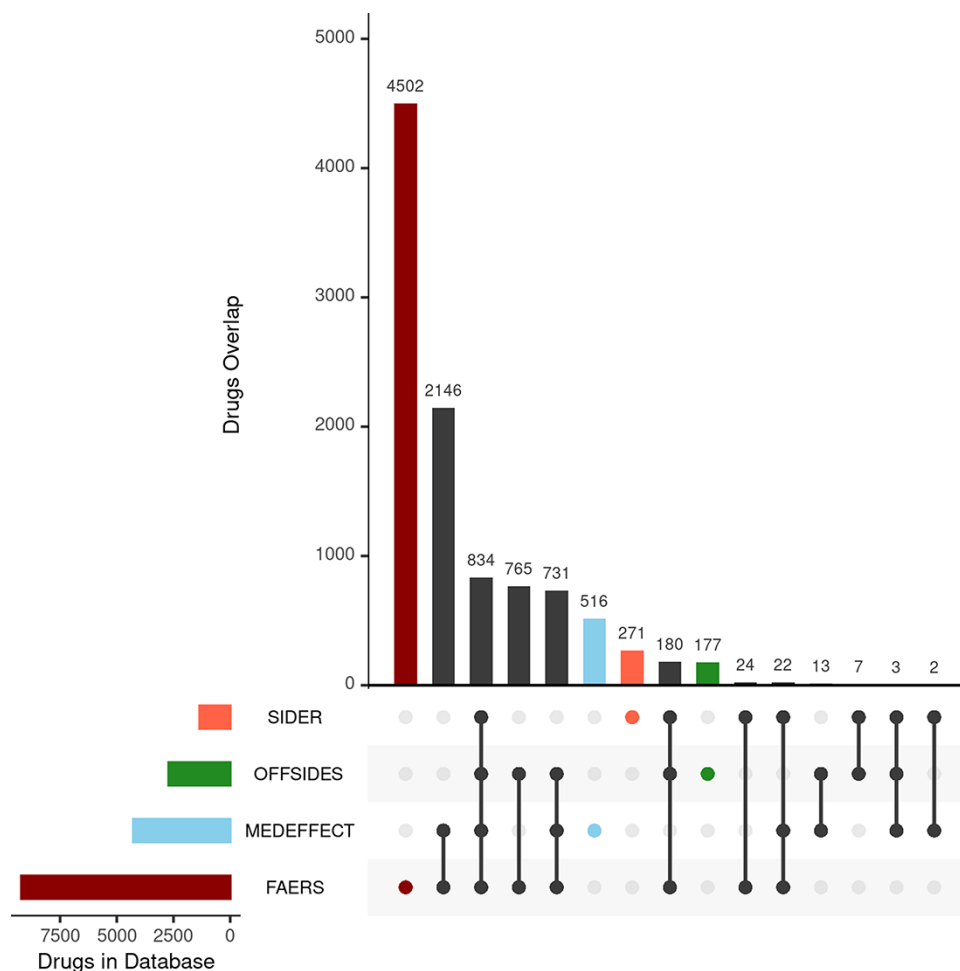


Figure 2. Upset plot showing the overlap between the different databases compiling drug-ADR associations. FAERS, MEDEFECT, OFFSIDES and SIDER represented as dark red, light blue, green and orange, respectively.

that self-reporting (FAERS and MEDEFECT) and curated (OFFSIDES and SIDERS) drug-ADR sources of information were not combined but treated independently. In the case of protein-ADR associations uncovered from combining drug-target and drug-ADR (self-reporting), a total of 998 drugs were mapped unequivocally on both sets (i.e. drug-target, drug-ADR) yielding over 100k statistically significant (i.e. q -value ≤ 0.05) protein-ADR associations accounting for around 3k and 211 different ADRs and proteins, respectively. In the case of the second group of drug-ADR databases, the curated set (or not self-reporting), i.e. SIDER and OFFSIDES, a total of 1135 common drug entities were identified between drug-target, yielding circa 40k statistically significant associations protein-ADR including 537 and 194 ADRs and proteins, respectively.

The number of ADR associated with a given protein target varies but in most cases the number of associated ADR to proteins is low both in the case of data extracted from the self-reporting and curated dataset (Figure 3). As expected, the number of associated ADRs to a given target relates to the number of drugs identified to target the given protein; as the number increases, the number of ADRs

also increases, albeit with a clearer trend in the case of the curated dataset (Figure 3B). Nonetheless there are a number of proteins associated with a large number of ADRs. In the case of the protein-ADR associations uncovered from the self-reporting dataset proteins, interleukin-8 (Uniprot ID: P10145), endothelin-1 (Uniprot ID: P05305) and leptin (Uniprot ID: P41159) were associated with 1532, 933 and 717 ADRs, respectively. In the case of the curated dataset, the figures are smaller and among the top three proteins are the 5-hydroxytryptamine receptor 2C (Uniprot ID P28335), the 5-hydroxytryptamine receptor 1A (Uniprot ID: P08908) and the alpha-2A adrenergic receptor (Uniprot ID: P08913) with 119, 104 and 98 associated ADRs, respectively. The explanation to this high number relates to the biological role played by these proteins. For instance, leptin is associated with over 150 biological processes (as per GO classification) ranging from signal transduction (GO:0007165) to autophagy regulation (GO:0010507). Moreover, the distribution of the number of ADR per target is in line with the work presented by Kuhn *et al.* (11) where the statistical association approach was described and that is the basis of T-ARDIS.

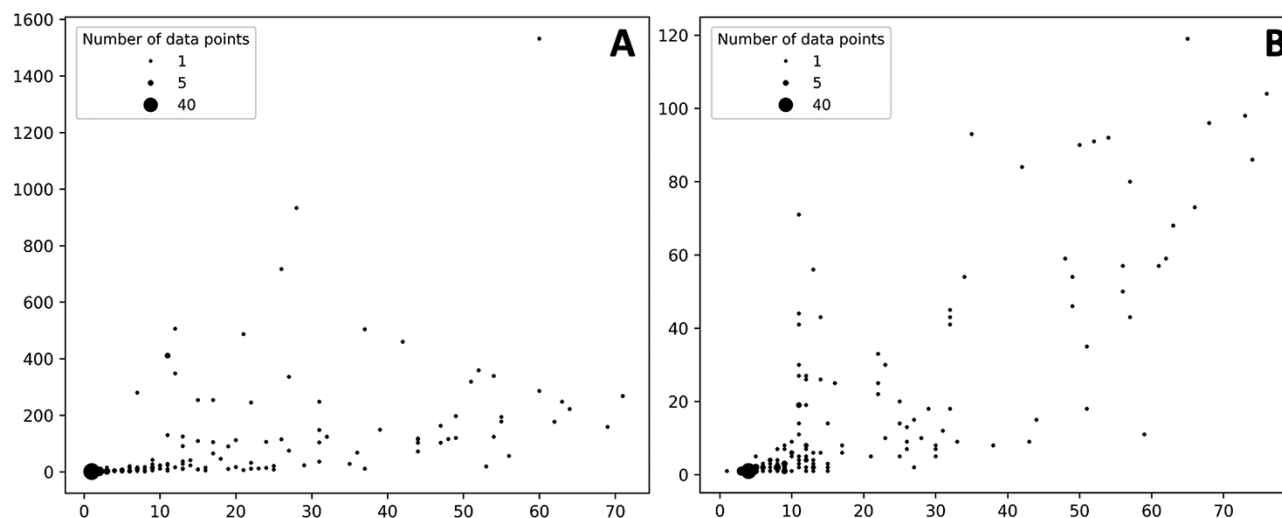


Figure 3. Bubble plots showing the number of drugs per protein (X axis) vs number of statistically significant ADR per protein (Y axis). (A) Distribution of the self-reporting set; (B) distribution of the curate set. Refer to the 'Material and methods' section for the description of self-reporting and curated sets.

T-ARDIS associations complement those of other resources

Association between ADRs and proteins uncovered in T-ARDIS were compared to previous works to assess the level of agreement and complementarity. The overall representation of target-ADR associations described in these four datasets, i.e. regardless of whether significant or non-significant, is low (Table 1). For instance, in the case of the set A (target-ADR associations mined from the literature), only 12% and 8% are presented in the self-reporting and curated sets of T-ARDIS, respectively. Overall, the values range from 20% to 5% in the case of self-reporting set and from 8% to 5% in the case of the curated set. These relatively low values can be due to two different causes. On the one hand, the lack of target-ADR associations in T-ARDIS can be due the fact that no safety issues have been reported either in self-reporting (FAERS, MEDEFECT) or curated databases (OFFSIDES, SIDER). It could also be that association between the given drug and target is not present in any of the following two databases used in this study: DTC and STITCH. On the other hand, and as described in the 'Methods' section, a robust and stringent procedure is followed when compiling and integrating the databases used to derive T-ARDIS. Thus, the given drug-ADR and/or drug-target association can be present but do not succeed to pass the filtering steps. In any case, these results come to illustrate the complementary nature of T-ARDIS to that of other resources available in the field and thus achieving a more comprehensive and complete view of target-ADR associations.

Examples of uncovered associations

Examples of protein-ADR associations uncovered by the approach presented here have been confirmed in the literature. For example, the cyclo-oxygenase 2 enzyme found in the gastric mucosa (COX-2 or PTGS2; Uniprot ID: P35354) is inhibited by the anti-inflammatory drug aspirin (acetylsalicylic acid). The aspirin also acts against the prostaglandin G/H synthase 1 (COX-1 or PTSG1; Uniprot ID: P23219) (31, 32). These secondary interactions may be the concomitant

cause for gastritis and bleeding ulcer as mentioned in various publications even since 1955 (33, 34). In our analyses, both PTGS1 and PTGS2 proteins are linked to Peptic ulcer and Peptic ulcer haemorrhage ADRs with significant q -values.

The sodium-dependent serotonin transporter (SLC6A4; Uniprot ID P31645) is inhibited by the serotonin norepinephrine reuptake inhibitor Venlafaxine, which in turn has been associated with sexual-dysfunction (35). In our analyses, SLC6A4 appears highly significantly associated (i.e. q -value $\ll 0.05$) with a range of different sexual dysfunctions (e.g. ejaculation failure and female sexual dysfunction).

Another example is illustrated by Budesonide and the glucocorticoid receptor (Uniprot ID: P04150). Identified ADRs to budesonide treatment include respiratory infections, coughs and headaches in the case of the inhaled form and tiredness, vomiting and joint pains in the oral form. A much rarer condition, adrenal insufficiency, has been identified in the case of the long-term use of the oral form of budesonide (36), which in T-ARDIS appears as a potential ADRs associated with the glucocorticoid receptor with a highly significant q -value. Furthermore, the association between glucocorticoids and adrenal insufficiency is an active topic of discussion in the current literature (37).

The activation of the 5-hydroxytryptamine receptor family (HTR1A, HTR1B and HTR1E; Uniprot IDs: P08909, P28222, and P28566, respectively) by zolmitriptan is reported to cause hyperaesthesia. In our analysis, the association between these proteins and hyperaesthesia were all significant, with q -values of 0.0001, 0.006 and 0.02 for HTR1A, HTR1B and HTR1E, respectively. It is worth mentioning that this association was identified and validated in vitro by Kuhn *et al.* (11). Overall, these examples, by no means a representative sample, show the usefulness of the data presented here that can be of use to identify potential liabilities associated with the targeting of proteins.

Accessing and querying T-ARDIS

All the association between drugs-proteins including the original sources, i.e. drug-protein and drug-ADR, has

Tardis About Drug-Target DB Drug-ADR DB [Download](#) [Github](#)

Show 10 entries Search:

Drug name	Uniprot ID	Gene	Adverse Reaction	q-value	Table	Drug-Side Effect DB	Drug-Target DB
<input type="text" value="Search Drug"/>	<input type="text" value="Search Unipr"/>	<input type="text" value="Search Gen"/>	<input type="text" value="Search Advers"/>	<input type="text" value="Search q-value"/>	<input type="text" value="Search Tabl"/>	<input type="text" value="Search Drug-"/>	<input type="text" value="Search Drug"/>
ASPIRIN	P23219	PTGS1	Peptic ulcer	8.89932e-8	807, 20, 53, 17	SIDER	STITCH
ASPIRIN	P35354	PTGS2	Peptic ulcer	0.00000572206	781, 46, 50, 20	SIDER	STITCH
ASPIRIN	P02768	ALB	Acidosis	0.0000968287	824, 42, 20, 11	SIDER	STITCH
ASPIRIN	P35354	PTGS2	Renal impairment	0.000337175	721, 41, 110, 25	SIDER	STITCH
ASPIRIN	P23219	PTGS1	Deafness	0.000347234	750, 20, 110, 17	SIDER	STITCH
ASPIRIN	P35354	PTGS2	Deafness	0.000390525	728, 42, 103, 24	SIDER	STITCH
ASPIRIN	P35354	PTGS2	Melaena	0.000436835	743, 44, 88, 22	SIDER	STITCH
ASPIRIN	P23219	PTGS1	Tinnitus	0.000551253	627, 13, 233, 24	SIDER	STITCH
ASPIRIN	P23219	PTGS1	Vascular purpura	0.000661393	743, 20, 117, 17	SIDER	STITCH
ASPIRIN	P35354	PTGS2	Vascular purpura	0.000898463	721, 42, 110, 24	SIDER	STITCH

Showing 1 to 10 of 47 entries

[Copy](#) [CSV](#) [PDF](#)

Previous 1 2 3 4 5 Next

Figure 4. Snapshot of the result page example upon querying by drug 'Aspirin'.

been deposited and compiled in a biological database: T-ARDIS. T-ARDIS is available at: <http://bioinsilico.org/T-ARDIS>. T-ARDIS provides a convenient and easy access to the information including the option of searching and filtering associations based on tailored queries. The database is searchable by protein (Uniprot ID or gene name), drug or ADR name. The resulting tables provide information on the association between protein-ADR as well as the q -value of the association and parent databases, both drug-protein and drug-ADR (Figure 4). External links to native drug-target or drug-ADR databases, together with protein-related repositories, are also provided. Users also have the option to further filter the resulting table by querying by specific drug, ADR or parent databases (e.g. filtering those associations resulting from FAERS). The table can be also sorted by q -values, so most significant associations could be shown first. The tables can be downloaded in the different formats (simple copy, CSV or PDF). Finally, bulk downloads of the database and associated scripts to recreate the database are also available from the home page links.

Discussion

Predicting associations between protein targets and ADR is desirable particularly in pre-clinical drug development in order to identify early in the process potential liabilities and toxicity-related aspects linked to proteins. Here, we present a fully automatic, large-scale, analysis to identify potential links between proteins and ADRs. By integrating public databases on drug-protein and drug-ADR associations, we have statistically identified significant relationships between protein and ADR using drugs as connecting elements. Highly significant associations, i.e. low q -values, are supported in the current literature and thus proving that uncovered associations could

be useful as guiding evidence. The data compiled in this work have been deposited in a freely accessible database, T-ARDIS, which allows a convenient and easy access to the information. The mining of the databases, statistical inference and database updating is fully automatic and thus ensuring that data will be integrated as become available further facilitating our understanding of the mechanisms behind ADRs. We envisage that T-ARDIS represents a resource that will be useful to both academic and industry researchers working on drug development.

Funding

Authors acknowledge support from MINECO grant numbers RYC2015-17519 and BIO2017-85329-R.

Conflict of interest

None declared.

Data availability

All the data and scripts required to recreate T-ARDIS Database are available on GitHub at <https://github.com/cristian931/Target-Adverse-Reaction-Database-Integrated-Search>. The database is also available at <http://bioinsilico.org/T-ARDIS>.

References

1. Seyhan, A.A. (2019) Lost in translation: the valley of death across preclinical and clinical divide – identification of problems and overcoming obstacles. *Trans. Med. Commun.*, 4, 18.
2. Waring, M.J., Arrowsmith, J., Leach, A.R. *et al.* (2015) An analysis of the attrition of drug candidates from four major pharmaceutical companies. *Nat. Rev. Drug. Discov.*, 14, 475–486.

3. Ferreira,G.S., Veening-Griffioen,D.H., Boon,W.P.C. *et al.* (2020) Levelling the translational gap for animal to human efficacy data. *Animals (Basel)*, **10**, 1199–1211.
4. Lo,Y.-C., Rensi,S.E., Torng,W. *et al.* (2018) Machine learning in cheminformatics and drug discovery. *Drug Discov. Today*, **23**, 1538–1546.
5. Basile,A.O., Yahi,A. and Tatonetti,N.P. (2019) Artificial intelligence for drug toxicity and safety. *Trends Pharmacol. Sci.*, **40**, 624–635.
6. Hamon,J., Whitebread,S., Techer-Etienne,V. *et al.* (2009) In vitro safety pharmacology profiling: what else beyond hERG? *Future Med. Chem.*, **1**, 645–665.
7. Portanova,J., Murray,N., Mower,J. *et al.* (2019) aer2vec: distributed representations of adverse event reporting system data as a means to identify drug/side-effect associations. *AMIA Annu. Symp. Proc.*, **2019**, 717–726.
8. Michel,C., Scosyrev,E., Petrin,M. *et al.* (2017) Can disproportionality analysis of post-marketing case reports be used for comparison of drug safety profiles? *Clin. Drug Investig.*, **37**, 415–422.
9. Huang,L.H., He,Q.S., Liu,K. *et al.* (2018) ADReCS-Target: target profiles for aiding drug safety research and application. *Nucleic Acids Res.*, **46**, D911–D917.
10. Smit,I.A., Afzal,A.M., Allen,C.H.G. *et al.* (2021) Systematic analysis of protein targets associated with adverse events of drugs from clinical trials and postmarketing reports. *Chem. Res. Toxicol.*, **34**, 365–384.
11. Kuhn,M., Al Banchaabouchi,M., Campillos,M. *et al.* (2013) Systematic identification of proteins that elicit drug side effects. *Mol. Syst. Biol.*, **9**, 663.
12. Tanoli,Z., Alam,Z., Vaha-Koskela,M. *et al.* (2018) Drug Target Commons 2.0: a community platform for systematic analysis of drug-target interaction profiles. *Database (Oxford)*, **2018**, 1–13.
13. Szklarczyk,D., Santos,A., von Mering,C. *et al.* (2016) STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data. *Nucleic Acids Res.*, **44**, D380–384.
14. Wong,C.K., Ho,S.S., Saini,B. *et al.* (2015) Standardisation of the FAERS database: a systematic approach to manually recoding drug name variants. *Pharmacoepidemiol. Drug Saf.*, **24**, 731–737.
15. Canada H. MedEffect Canada - Adverse Reaction Database; editing status 2019-01-15; re3data.org - Registry of Research Data Repositories. [10.17616/R3J03W](https://doi.org/10.17616/R3J03W) (18 October 2021, date last accessed).
16. Kuhn,M., Letunic,I., Jensen,L.J. *et al.* (2016) The SIDER database of drugs and side effects. *Nucleic Acids Res.*, **44**, D1075–D1079.
17. Tatonetti,N.P., Ye,P.P., Daneshjou,R. *et al.* (2012) Data-driven prediction of drug effects and interactions. *Sci. Transl. Med.*, **4**, 125ra31.
18. Chang,L.-C., Mahmood,R., Qureshi,S. *et al.* (2017) Patterns of use and impact of standardised MedDRA query analyses on the safety evaluation and review of new drug and biologics license applications. *PLoS One*, **12**, e0178104.
19. The UniProt C (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, **47**, D506–D515.
20. Banda,J.M., Evans,L., Vanguri,R.S. *et al.* (2016) A curated and standardized adverse drug event resource to accelerate drug safety research. *Sci. Data*, **3**, 1–11.
21. Nelson,S.J., Zeng,K., Kilbourne,J. *et al.* (2011) Normalized names for clinical drugs: RxNorm at 6 years. *J. Am. Med. Inf. Assoc.*, **18**, 441–448.
22. Huang,L., Zalkikar,J. and Tiwari,R.C. (2013) Likelihood ratio test-based method for signal detection in drug classes using FDA's AERS database. *J. Biopharm. Stat.*, **23**, 178–200.
23. Ietswaart,R., Arat,S., Chen,A.X. *et al.* (2020) Machine learning guided association of adverse drug reactions with in vitro target-based pharmacology. *EBioMedicine*, **57**, 102837.
24. von Mering,C., Jensen,L.J., Snel,B. *et al.* (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res.*, **33**, D433–437.
25. Szklarczyk,D., Gable,A.L., Lyon,D. *et al.* (2019) STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.*, **47**, D607–D613.
26. Puoliväli,T., Palva,S. and Palva,J.M. (2020) Influence of multiple hypothesis testing on reproducibility in neuroimaging research: a simulation study and Python-based software. *J. Neurosci. Methods*, **337**, 108654.
27. Bento,A.P., Hersey,A., Felix,E. *et al.* (2020) An open source chemical structure curation pipeline using RDKit. *J. Cheminform.*, **12**, 51.
28. The Gene Ontology C (2019) The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res.*, **47**, D330–D338.
29. Li,W., Jaroszewski,L. and Godzik,A. (2002) Tolerating some redundancy significantly speeds up clustering of large protein databases. *Bioinformatics*, **18**, 77–82.
30. Imming,P., Sinning,C. and Meyer,A. (2006) Drugs, their targets and the nature and number of drug targets. *Nat. Rev. Drug Discov.*, **5**, 821–834.
31. Boutaud,O., Sosa,I.R., Amin,T. *et al.* (2016) Inhibition of the biosynthesis of prostaglandin E2 by low-dose aspirin: implications for adenocarcinoma metastasis. *Cancer Prev. Res. (Phila)*, **9**, 855–865.
32. Flower,R. (2003) What are all the things that aspirin does? *BMJ*, **327**, 572–573.
33. Muir,A. and Cossar,I.A. (1955) Aspirin and ulcer. *Br. Med. J.*, **2**, 7.
34. Shim,Y.K. and Kim,N. (2016) Nonsteroidal anti-inflammatory drug and aspirin-induced peptic ulcer disease. *Korean J. Gastroenterol.*, **67**, 300–312.
35. Higgins,A., Nash,M. and Lynch,A.M. (2010) Antidepressant-associated sexual dysfunction: impact, effects, and treatment. *Drug Healthcare Patient Saf.*, **2**, 141–150.
36. Arntzenius,A. and van Galen,L. (2015) Budesonide-related adrenal insufficiency. *BMJ Case Rep.*, **2015**.
37. Laugesen,K., Broersen,L.H.A., Hansen,S.B. *et al.* (2021) Management of endocrine disease: glucocorticoid-induced adrenal insufficiency: replace while we wait for evidence? *Eur. J. Endocrinol.*, **184**, R111–R122.



Prediction of Adverse Drug Reaction Linked to Protein Targets Using Network-Based Information and Machine Learning

Cristiano Galletti¹, Joaquim Aguirre-Plans², Baldo Oliva³ and Narcis Fernandez-Fuentes^{1*}

¹Department of Biosciences, U Science Tech, Universitat de Vic-Universitat Central de Catalunya, Barcelona, Spain, ²Department of Physics, Network Science Institute, Northeastern University, Boston, MA, United States, ³Department of Experimental and Health Sciences, Structural Bioinformatics Group, Research Programme on Biomedical Informatics, Universitat Pompeu Fabra, Barcelona, Spain

OPEN ACCESS

Edited by:

Tatsuya Akutsu,
Kyoto University, Japan

Reviewed by:

Surabhi Naik,
University of Tennessee Health
Science Center (UTHSC),
United States
Olga Kalinina,
Helmholtz-Institute for Pharmaceutical
Research Saarland (HIPS), Germany

*Correspondence:

Narcis Fernandez-Fuentes
narcis@bioinsilico.org

Specialty section:

This article was submitted to
Network Bioinformatics,
a section of the journal
Frontiers in Bioinformatics

Received: 28 March 2022

Accepted: 02 June 2022

Published: 14 July 2022

Citation:

Galletti C, Aguirre-Plans J, Oliva B and
Fernandez-Fuentes N (2022)
Prediction of Adverse Drug Reaction
Linked to Protein Targets Using
Network-Based Information and
Machine Learning.
Front. Bioinform. 2:906644.
doi: 10.3389/fbinf.2022.906644

Drug discovery attrition rates, particularly at advanced clinical trial stages, are high because of unexpected adverse drug reactions (ADR) elicited by novel drug candidates. Predicting undesirable ADRs produced by the modulation of certain protein targets would contribute to developing safer drugs, thereby reducing economic losses associated with high attrition rates. As opposed to the more traditional drug-centric approach, we propose a target-centric approach to predict associations between protein targets and ADRs. The implementation of the predictor is based on a machine learning classifier that integrates a set of eight independent network-based features. These include a network diffusion-based score, identification of protein modules based on network clustering algorithms, functional similarity among proteins, network distance to proteins that are part of safety panels used in preclinical drug development, set of network descriptors in the form of degree and betweenness centrality measurements, and conservation. This diverse set of descriptors were used to generate predictors based on different machine learning classifiers ranging from specific models for individual ADR to higher levels of abstraction as per MEDDRA hierarchy such as *system organ class*. The results obtained from the different machine-learning classifiers, namely, support vector machine, random forest, and neural network were further analyzed as a meta-predictor exploiting three different voting systems, namely, *jury vote*, *consensus vote*, and *red flag*, obtaining different models for each of the ADRs in analysis. The level of accuracy of the predictors justifies the identification of problematic protein targets both at the level of individual ADR as well as a set of related ADRs grouped in common system organ classes. As an example, the prediction of ventricular tachycardia achieved an accuracy and precision of 0.83 and 0.90, respectively, and a Matthew correlation coefficient of 0.70. We believe that this approach is a good complement to the existing methodologies devised to foresee potential liabilities in preclinical drug discovery. The method is available through the DocTOR utility at GitHub (<https://github.com/cristian931/DocTOR>).

Keywords: network biology, drug adverse reaction, drug target, machine learning, protein-adverse reaction association

1 INTRODUCTION

Protein–protein interactions are central to all aspects of cell biology, including processes linked to diseases. The phenomenal technological development in recent years allowed the comprehensive charting of the protein–protein interactions that take place in human cells, the interactome [(Gavin et al., 2011; Xing et al., 2016; Xiang et al., 2021)]. Indeed, high-quality and high-coverage protein interaction maps are now available for a number of model organisms, including humans (Kotlyar et al., 2022). Such resources present a number of opportunities to the pharmaceutical industry, which can exploit this information to, for instance, identify plausible therapeutic targets from which to develop or repurpose drugs [as in the most recent case of COVID-19 drug race (Sahoo et al., 2021; Gysi et al., 2021)]. At the same time, these recent advances have also led to increased efforts to fill the gap of toxicology or safety information for drug's targets. This problem has always crippled the development of novel drugs, increasing the attrition of the latter entering clinical trials due to the severity of adverse drug reactions (ADRs) associated with unforeseen toxicity, directly increasing the cost of research (Seyhan, 2019).

Currently, several drug-centered approaches exist that can be used to reduce the risk of ADRs associated with novel drugs (Basile et al., 2019), such as the use of animal models (Bailey et al., 2014) and *in vitro* toxicology research (Madorran et al., 2020). However, these approaches involve high maintenance costs and ethical limitations and are not always transferable to human biology (Singh and Seed, 2021). Many *in silico* approaches have also proved to be useful in estimating the toxicity of drug candidates, exploiting features such as composition, structure, and binding affinity [(Lo et al., 2018), (Bender et al., 2007)]. These methods include various examples of machine learning (ML) and deep learning (Dara et al., 2022). Contributing to these efforts, we recently described the T-ARDIS database (Galletti et al., 2021). T-ARDIS is a curated collection of relationships between proteins and ADRs. The associations are statistically assessed and derive from existing resources of drug-target and drug-ADR association (Galletti et al., 2021). Since T-ARDIS provides a direct link between proteins and ADRs, the question arose of whether this information can be exploited to predict potential ADR linked to proteins. Therefore, the major driver of this project was to develop a target-centric approach to predict whether the targeting of a given protein target is likely to result in ADR using the curated information to train machine-learning classifiers.

To that end, different machine-learning classifiers were assessed including support vector machine (SVM), random forest (RF), and neural networks (NN). Highly significant associations between proteins and ADRs were extracted from T-ARDIS and characterized using 8 different features. These include the following: 1) the network diffusion-based score from GUILDify (Aguirre-Plans et al., 2019); 2) several network-based clustering algorithms [(Cao et al., 2014), (Blondel et al., 2008)]; 3) a functional similarity index; 4) network distance to proteins that are part of safety panels used in preclinical drug development; and 5) network descriptors in the form of degree and betweenness centrality

measurements and conservation. All of the measurements use network-based information in some way and hence incorporate aspects that are intrinsic not only to the protein but also to the network. As a result, the proteins are framed within the interactome, and the potential impact of changes on neighboring proteins is assessed.

According to the MEDDRA nomenclature (Chang et al., 2017), specific models were built for each individual ADR, as well as clusters of ADRs within the same system organ class (SOC), allowing the analysis to be extended to a more general anatomical or physiological system. Besides the datasets derived from T-ARDIS to train and test the models, we also benchmarked our prediction in independent datasets including manually curated dataset compiled from literature [(Huang et al., 2018), (Mizutani et al., 2012), (Smit et al., 2021), (Kuhn et al., 2013)]—**Supplementary Table S2**, including a dataset submitted to the critical assessment of massive data analysis competition (Aguirre-Plans et al., 2021). Finally, as three different machine-learning predictions were developed, we also explored the accuracy of a meta-predictor that combines the predictions of each individual classifier. Three different meta-predictors were assessed based on the way the predictions were combined: 1) *jury vote*, 2) *consensus*, and 3) *red flag*. While *jury vote* and *consensus* scoring function are similar and seek to promote associations with high scores, *red flag* takes into account the divergent opinion.

The proposed method achieves a high level of reliability. For example, taking into account the undesirable effect of atrial fibrillation, the resulting model scored high in accuracy (0.88), precision (0.87), recall (0.85), and Matthew correlation coefficient (MCC) (0.77) for both the SVM and RF approaches. The neural network gives slightly lower results with 0.66 accuracy, 0.71 precision, and an MCC of 0.34. The obtained meta-predictors achieved similar results in jury voting and consensus methods with accuracy 0.89, precision 0.89, recall 0.88, and MCC 0.78. To be noted, the reliability of the model is closely related to the biological complexity and tissue specificity of various ADRs. The dataset employed in this study as well as the models, meta-predictors, and accessory scripts are available at <https://github.com/cristian931/DocTOR>. Upon installing the application, users will be able to upload a list of proteins in order to assess their relationship with the studied ADR.

2 MATERIALS AND METHODS

2.1 Datasets

2.1.1 Training Set

The set used to train and cross-validate the models was derived from T-ARDIS (Galletti et al., 2021). T-ARDIS is a database that compiles statistically significant relationships between proteins and ADRs. As described in original publication, T-ARDIS undergoes a series of filtering and quality control steps to ensure a reliable and significant relationship between the ADR and the protein targets. Depending on the source of ADRs associations used to derive target ADRs relationships, two groups were defined: relationships derived from self-reporting databases FAERS (Kumar, 2018) and MEDEFECT

(Re3data.Org, 2014); and relationships derived from curated databases SIDER (Kuhn et al., 2015) and OFFSIDES (Tatonetti et al., 2012). Both groups have been used to obtain the training set used in this work. For the self-reporting dataset, T-ARDIS currently contains about 17k paired protein–ADR interactions, including 3k adverse reactions and 300 Uniprot ids. The smaller curated dataset contains approximately 3,000 pairwise associations for 537 adverse events and 200 proteins. From the initial list of approximately 500 ADRs, only the 84 that were best characterized in terms of number of proteins associated and that covered the entire range of SOC classes, as defined by MEDDRA (Chang et al., 2017), were considered, i.e., included at least 5 numbers of ADR per SOC.

2.1.2 Independent Test Datasets

For external validation, we employed five different independent datasets sourced from literature containing protein–ADR relationships from Kuhn et al. (2013)—**Supplementary Table S2**, Smit et al. (2021), Mizutani et al. (2012) the ADRcCs-Target database (Huang et al., 2018), and the DisGeNet Drug-induced Liver Injury dataset (Piñero et al., 2019). In particular, the latter contains a specific subset of liver injuries caused by drugs composed by 12 different MEDDRA-defined events ranging from “Acute hepatic failure” to “Non-Alcoholic Steatohepatitis.”

More than 600 distinct adverse events and 428 proteins were retrieved, resulting in a total of 15 k interactions. Then, the 84 selected ADR were extracted, resulting in 188 associated proteins. The independent and the training dataset are totally independent in the sense that they do not share proteins between them on each particular ADR.

2.2 Protein Network

The protein network, or interactome, used in this study, was integrated using BIANA (Garcia-Garcia et al., 2010) and GUILDIfyv2 (Aguirre-Plans et al., 2019). The original BIANA network includes interactomic information from IntAct (Kerrien et al., 2006), DIP (Wong et al., 2015), HPRD (Keshava Prasad et al., 2008), BioGrid (Stark et al., 2006), MPACT (Güldener et al., 2006), and MINT (Ceol et al., 2009) databases. The most recent version composed of 13,090 proteins (or nodes) and 320,337 interactions (or edges) has been used in this work.

2.3 Features

2.3.1 GUILDIfy Score

GUILDIfy is a web server of network diffusion-based algorithms used for a wide range of network medicine applications (Aguirre-Plans et al., 2019). The message-passing algorithms of GUILDIfy (Guney and Oliva, 2012) transmit a signal from a group of proteins associated with a phenotype or drug (known as seeds) to the rest of the network nodes and score them depending on how fast the message reaches them, taking into account several network properties. Originally, GUILDIfy had been developed to prioritize gene–disease relationships and identify disease modules (Aguirre-Plans et al., 2019), but it was recently used to identify disease co-morbidities and drug repurposing options (Aguirre-Plans et al., 2019; Artigas et al., 2020). In this study, GUILDIfy was used as a feature to predict protein–ADR associations. Upon

expansion, a GUILD score was assigned to each protein in the interactome based on the ADR’s linked protein used as the seed. The higher the score, the more likely that an association exists between the protein and the set of seeds used to expand.

2.3.2 Degree and Betweenness Centrality

Degree and betweenness centrality are two network analysis measures. Degree centrality is the number of edges connected to a node, while betweenness centrality is the number of times a node acts as a bridge along the shortest path between two other nodes. Both measures define how relevant a given node is inside a network and, in terms of the interactome, how much a protein tends to be part of a cascade of signals and participate in the same biological process. Degree and betweenness centrality values were computed using NetworkX (Ceol et al., 2009).

2.3.3 Clustering-Based Algorithms

Another interpretation of the “guilt-by-association” principle is the definition of “disease module,” i.e., a neighborhood of a molecular network whose components are jointly associated with one or several diseases or risk factors (Choobdar et al., 2019). As shown, disease modules can be used to identify protein/genes associated with given diseases (Goh and Choi, 2012). In the context of ADRs, the assumption is that proteins linked to the same ADRs would cluster in local regions of the interactome, forming ADR modules (Guney, 2017).

To identify these modules, two different clustering algorithms were used. First, the K1 clustering algorithm is based on the so-called diffusion state distance (DSD) metric (Cao et al., 2014). The DSD metric is used to define a pairwise distance matrix between all nodes, on which a spectral clustering algorithm is applied. In parallel, dense bipartite subgraphs are identified using standard graph techniques. Finally, results are merged into a single set of non-overlapping 858 clusters. The second clustering method is based on the work by Lefebvre and col ((Blondel et al., 2008)), which is based on modularity optimization, assigning, and removing recursively the nodes to the modules found, each time evaluating the loss or gain of modularity. We applied this method to the interactome, retrieving 46 modules. Together with clustering approaches mentioned above, we compute for each node the “clustering coefficient” using the NetworkX utility (Ceol et al., 2009).

2.3.4 Function Conservation Index

A new feature included in the newer version of GUILDIfy is the identification of enriched Gene Ontology (GO) functions among top ranking proteins using Fisher’s exact test (Aguirre-Plans et al., 2019). The function conservation index, which takes advantage of this resource, considers the functional similarity between a protein and GUILDIfy’s enriched GO terms. In a nutshell, this value is the result of a Hamming distance between two binary vectors that represent the presence or absence of a specific GO term. The shorter the distance, the higher the similarity between the given protein and the enriched functions identified from a set of protein–ADRs. The scale represents the ratio where a 1 would indicate full overlap of functions.

2.3.5 Shortest Path to Very Important Targets

Targets and pathways that are now well established as contributors to clinical ADRs are included in safety panels, which constitute the minimal lists of targets that qualify for early hazard detection, off-target risk assessment, and mitigation. (Bowes et al., 2012a). Here, we considered the Safety Screen Tier 1 panel of EuroFins Discovery based on the work by Whitebread and co (Bowes et al., 2012b). This panel is composed of 48 proteins that we call Very Important Targets (VITs). We positioned the VITs in the interactome and calculated the shortest path distance of each one of the proteins considered in our training set to any VITs using NetworkX (Ceol et al., 2009). Of the overall distribution of shortest path distances to VITs of any given protein, the value of the first quartile was considered. This value represents the relative position of the given protein with respect to the VITs panel.

2.4 Model Construction

2.4.1 Positive and Negative Sets

The positive set, i.e., proteins related to a given ADR, for each of the 84 ADRs considered were extracted from the T-ARDIS database (Galletti et al., 2021). For the purpose of training and since the number of positive cases per ADR was generally low, the positive set was augmented using the definition of close connectivity as follows. The DIAMOnD score (Drozdetzkiy et al., 2015) was computed for the subnetworks associated with the ADR's associated proteins extracted from T-ARDIS. In doing so, we ranked the most immediate neighboring proteins and selected those with a DIAMOnD score over a certain threshold to conform to the positive set. Also, multiple DIAMOnD threshold scores have been tested to obtain the best result during the training phase, namely, at 0.6, 0.7, 0.8, and 0.9. Likely, the negative sets were specific to each of the ADRs under consideration by randomly selecting proteins with a DIAMOnD score below the given positive threshold. During the training and testing phase, different ratios of positive and negative cases were tested to account for class imbalance. Indeed, besides using a balanced training set, i.e., equal number of positive and negative cases, to train and test the models, different ratios including 1:1.5, 1:3, and 1:5 (positives:negatives) were also considered. Thus, in the end, for each one of the 84 ADRs, 12 different models have been obtained by the combination of positive and negative thresholds as well as imbalance ratios resulting in 1,008 trained models.

2.4.2 Features Vectorization and Model Construction and Training

The approach to predict protein-ADR associations is described below. In a nutshell, the approach is network-based, i.e., relies on a network-based set of 8 metrics computed for each protein that were used as inputs to machine-learning classifiers. Three different types of classifiers were used: SVM with nonlinear kernel (radial basis function—RBF), RF, and NN. The different ML classifiers were implemented in python3.9 using the following libraries. SVM and RF classifiers were implemented using the *Scikit-learn* package (Pedregosa et al., 2011), while NN

made use of the Keras and Tensorflow packages (Abadi et al., 2015; Gaulton et al., 2017). Specific models were trained and tested for each of the 84 ADR as well as models at SOC, i.e., grouping ADRs belonging to the same SOC. A schematic representation of the overall process is depicted in **Figure 1**.

Each protein in a given ADR is represented by an 8-dimensional vector composed by the features described above (or see **Figure 1**) that is used as an input to the classifier together with the labels (positive/negative) in supervised learning. Note that balanced and unbalanced sets were used, and thus, 4 specific models were built for each ADR depending on the set used. The training involved the optimization of a set of parameters using a grid-search approach and validated with an internal stratified five-fold cross-validation approach using the Scikit-learn python package. In the case of SVM classifiers, the grid search included the *gamma* and *C* parameters; for the RF, the *maximum number of features* and the *depth* for each tree; lastly, for the basic model architecture of NN, an *SGD optimizer function* was combined with a *relu activation function* (for the first layer) and then with a simple *sigmoid activation function*. A grid search was used to optimize the *learning rate*, *number of epochs*, *number of hidden layers*, and *neurons*, the same as it was for the other ML algorithms. Finally, in the case of ML classifiers derived for SOC, i.e., groups of ADRs, the training and testing was done in the same way after merging all the elements in each individual ADR. The training dataset, including the ML classifiers for individual ADRs and SOCs, can be obtained from <https://github.com/cristian931/DocTOR> together with the relative parameters of the best model for each ADR (**Supplementary Material**—NN_parameters.tsv, RF_parameters.tsv, SVM_parameters.tsv).

2.5 Assessing Performance of Models

The performance of models was assessed using four widely used statistical descriptors, namely, the accuracy (ACC), precision (PREC), recall (REC), and MCC calculated using the Scikit-learn python package (Pedregosa et al., 2011). In addition, the scores of AUPRC have been computed and compared to the NPV and PPV values available in the **Supplementary Material S1**.

2.6 Combining Predictions: Voting Systems

Three different voting systems were envisaged to integrate the prediction of individual classifiers: a *jury vote*, a *consensus* score, and a *red-flag* schema. Both jury votes and consensus seek to maximize similar predictions, while the *red-flag* prioritizes outliers. Jury voting is simply the count of prediction outcomes. Classifiers are binary and thus will predict whether a given protein is or is not causing a given ADRs. Each method exhibits a vote, and the most voted option is selected. The consensus score *c* is more granular, namely instead of a yes/no the posterior probability *p* of each classifier is used. Therefore, the consensus score can rank proteins within the same class, e.g., predicted to be related to a given ADR. Finally, the *red-flag* schema simply accepts as a final prediction the one which is not common among the different classifiers.

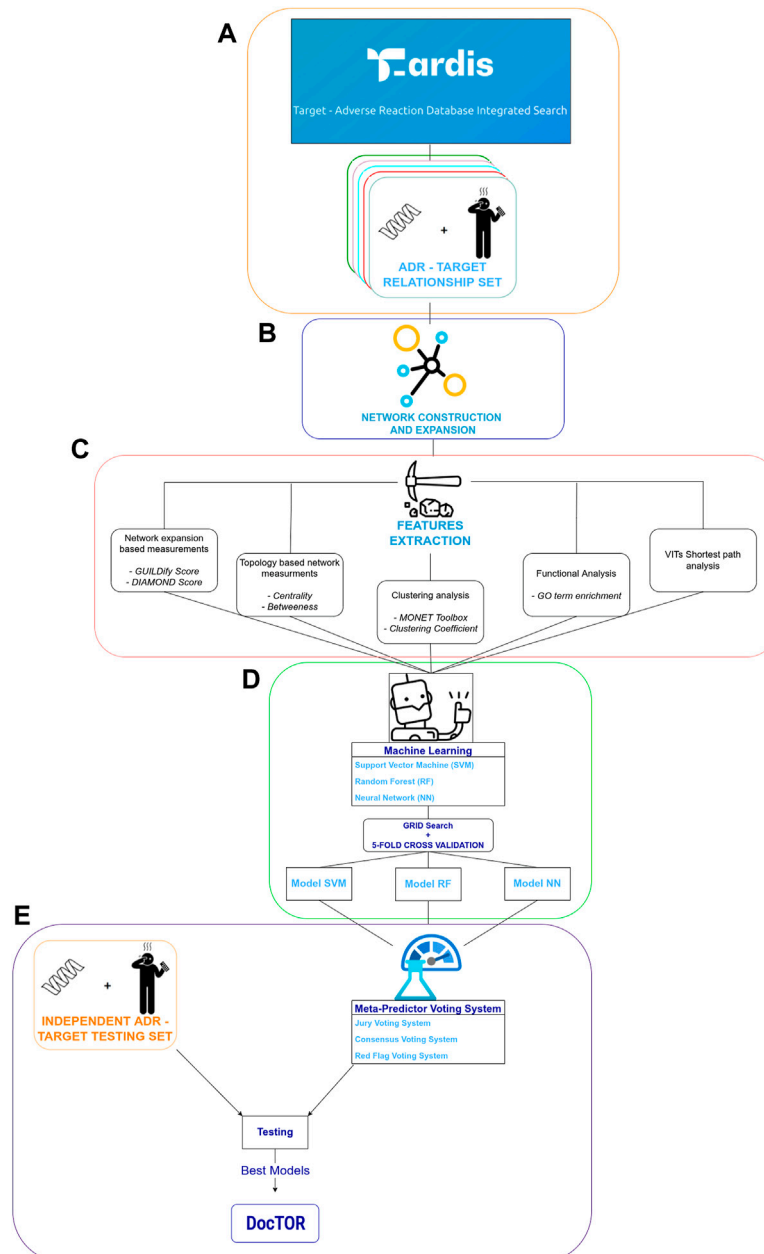


FIGURE 1 | Schematic depiction of feature extraction, training, and testing procedures. **(A)** indicates the process of extraction of training dataset from T-ARDIS (Galletti et al., 2021). **(B)** indicates the process of network expansion of targets extracted in **(A)** using GUILDify (Aguirre-Plans et al., 2019). **(C)** summarizes the process of computation of different input features. **(D)** Represents the development of machine-learning classifiers. Finally, **(E)** illustrates the development of the meta-predictors together with the testing of the classifiers and consensus functions on the independent dataset.

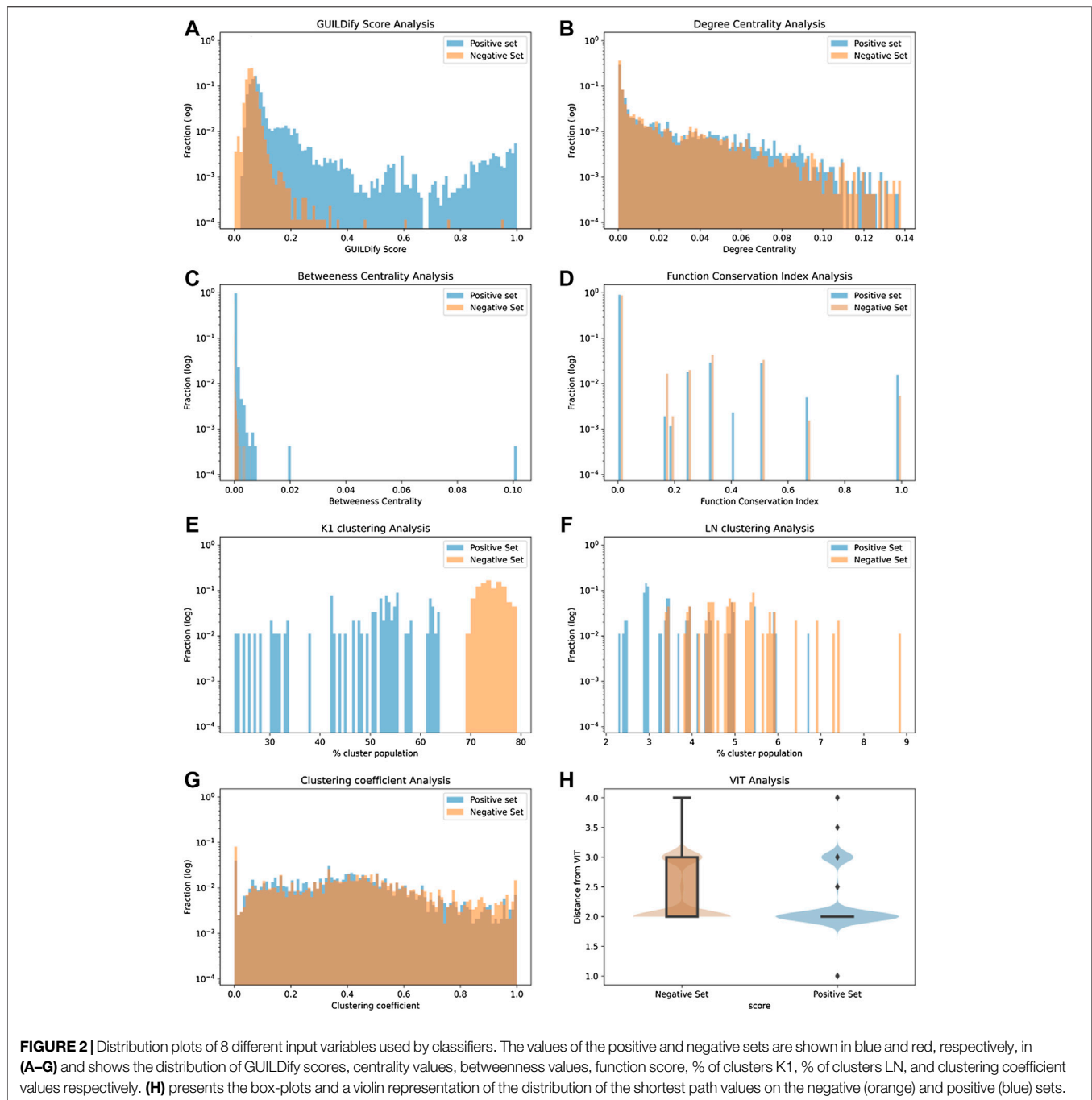
$$c = \sum_{i=1}^3 p_i * class(i); i = [SVM, RF, NN]; class \in [-1, +1] \quad 1$$

3 RESULTS

3.1 Individual Features

Eight different variables were considered as input features of the classifiers. These include the GUILDify scores, network topology

(degree and betweenness centrality values), a function conservation score, module imputations, and distances to proteins belonging to safety panels. In **Figure 2**, the distribution of the different features for the positive and negative sets is shown. As mentioned in the Methods section, the positive cases (negative cases were selected randomly) were extracted from the T-ARDIS database (Galletti et al., 2021), both for the self-reporting and curated sets. The data shown in **Figure 2** derives from the self-reporting set of T-ARDIS. The



equivalent information for the curated set is shown in **Supplementary Figure S1**; **Supplementary Material S1**. Likewise, equivalent information, as in **Figures 3, 4**, is presented in the **Supplementary Material S1**.

In the case of GUILDiFy scores, a high overlap is found, but nonetheless, the positive sets demonstrate higher scores and a distribution slightly skewed toward high values (**Figure 2A**). The analysis of centrality-based features also indicates a substantial overlap between positive and negative sets, although positive sets present a more skewed distribution toward higher values particularly

in the case of betweenness values (**Figures 2B,C**). A similar situation is presented when a quantifying function analysis as distance to enriched function(s) of the set (**Figure 2D**); the proteins in the negative set tend to demonstrate larger distances, i.e., no shared functions with the GUILDiFy enriched GO terms, respect to those on the positive set. In fact, the largest number of proteins with a value of 1.0 correspond to the proteins in the positive set and, conversely, those with lower values, i.e., no shared GO terms, tend to be proteins in the negative set. However, it is fair to say that the overlap is very high.

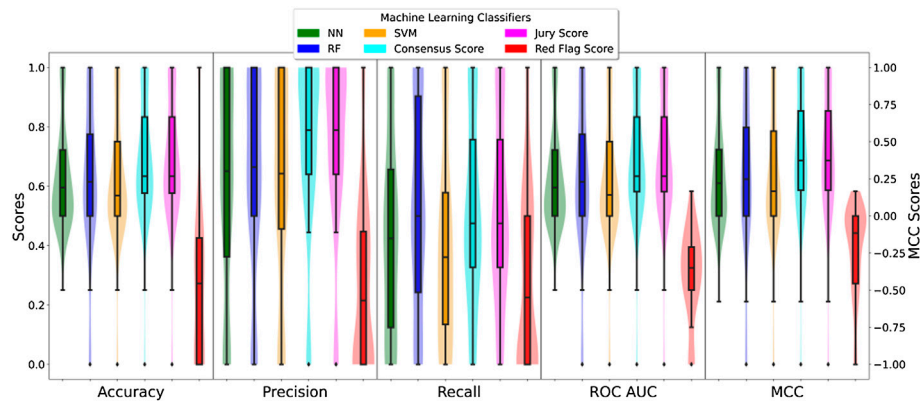


FIGURE 3 | Box- and violin plots of the cross-validation AUC results for the three different classifiers. The different box-plots show the distribution of the mean AUC values for the best models developed for each ADR using the three different classifiers: SVM (orange), random forest (blue), and neural networks (green).

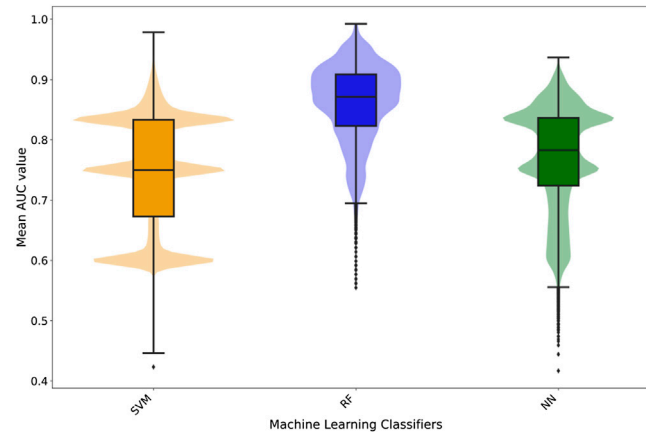
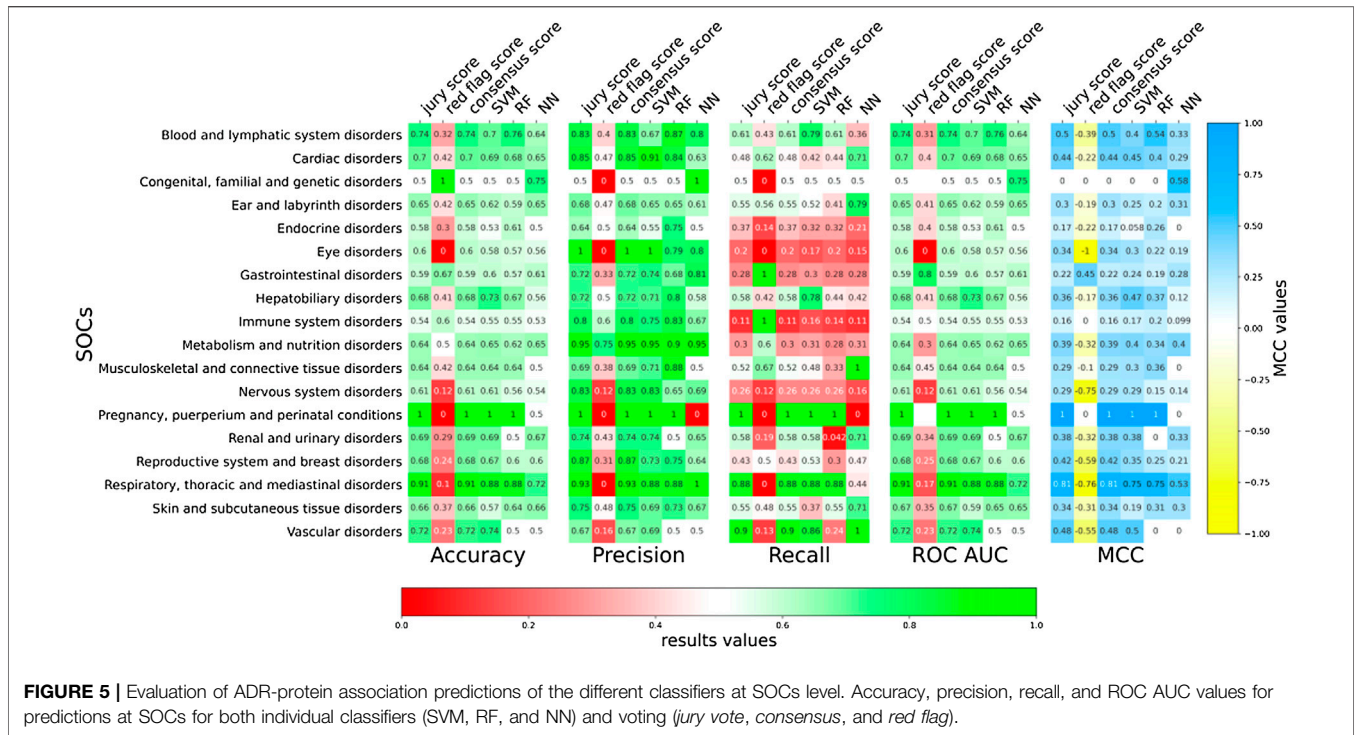


FIGURE 4 | Box- and violin plots for accuracy (ACC), precision (PREC), recall (REC), receiver operating area under curve (ROC AUC), and Matthew correlation coefficient (MCC). Distribution of accuracy, precision, recall, and ROC AUC values for individual classifiers: NN (green), RF (blue), and SVM (orange) as well as meta-predictions: *consensus* (cyan), *jury vote* (magenta), and *red flag* (red).

The tendency of functionally and disease-related proteins to be close (i.e., shorter distances) in the interactome was also considered as a feature for the prediction. As described in the Methods section, this aspect was studied by applying clustering algorithms to identify modules in the entire interactome where the proteins associated with the same or similar ADRs are grouped. Next, if the number of modules required to represent a given collection of proteins in an ADR is small, it is likely that the proteins will share modules. Similarly, a large number of modules indicate that the proteins do not share the same cluster. The K1 algorithm (Cao et al., 2014) identified 1,170 different clusters, many of them composed of 3 proteins, the least amount for defining a module (Figure 2E). As shown, proteins in the positive set present a lower number of clusters, meaning that proteins associated with ADRs tend to belong to a limited group of clusters, rather than being scattered through the interactome. Similarly, the Louvain-Newman method (Blondel et al., 2008),

which grouped the whole interactome into only 95 distinct clusters, allowing the analysis of bigger modules, demonstrated a similar distribution as K1, i.e., the positive set is drawn toward lower values (Figure 2F). Finally, in the case of the Clustering Coefficient Analysis (Figure 2G), in this case, both negative and positive sets share the same distribution of values. Therefore, this feature does not seem to provide a clear distinction between positive and negative cases on the ADR.

The final metric considered as an input variable was the distance of given proteins to the so-called VITs (see Methods). The distance was computed in the form of the shortest path (i.e., lowest number of links) to any given protein belonging to the panel, taking the value of the first quartile upon computing all the distances all vs. all (protein in the given ADR and proteins in the panel). Once again, the distribution of values is different depending if the proteins are part of the positive or negative sets (Figure 2H). While the most common distance is 2.0, only



the proteins in the positive set would demonstrate values smaller than 2, therefore showing that proteins in the positive set are closer to proteins considered critical as per pharmacological profiling.

3.2 Training and Cross-Validation

The input features described above represent the input variables to the different classifiers explored in this work. Three different machine-learning methods were used: NN, SVM, and RF. In order to define the best parameter values, each classifier was trained and validated on a 5-fold cross-validation and grid-search approach.

It is important to mention that specific classifiers were developed for each ADRs. The classifiers are not generic predictors of the likelihood of a protein to elicit an ADR, any, but to elicit a particular ADR, e.g., diarrhoea. Therefore, the predictions are tailored to the specific ADR (84 considered in this study) and, therefore, present unique characteristics. Next, **Figure 3** presents the distribution of mean area under the ROC curve (AUC) calculated for the training and testing as described (for details on individual classifiers and ADRs refer to the **Supplementary Material S1**—Supporting information 7 “cv scores. zip”). In general RF classifiers appear to demonstrate higher performance with mean AUC values around 0.85. Also, RF presents a more bell-shaped distribution of values when compared to SVM and NN. On the other hand, SVM and NN demonstrate a comparable performance, with a median AUC around 0.75, although the first quartile in SVM is slightly better than in NN (0.72 vs. 0.68).

Overall RF appeared to demonstrate the best performance under training conditions, but in some cases, the performance of the different classifiers was lower for particular ADRs, highlighting the complexity and heterogeneity of this biological problem. For instance, in the case of the ADR *malnutrition*, RF achieved the best performance with an accuracy, precision, recall, and MCC values of 0.95, 0.92, 1.00, and 0.91, respectively. However, in the case of the ADR *febrile neutropenia*, NN was by far the best predictor with an accuracy, precision, recall, and MCC values of 0.80, 0.87, 0.70, and 0.77, respectively, against an almost random prediction by SVM and RF (MCC ~0.0). Finally, SVM outperformed the other two ML approaches in other cases, such as *Nasal Congestion*, with an accuracy of 0.90, a precision of 0.83, a recall of 1, and a MCC of 0.81, while RF and NN barely reached values of 0.70 (see **Supplementary Material S1** for detailed information of individual performances across all ADR studied).

3.3 Testing on Independent Set

For independent testing purposes, we relied on proteins associated with the same ADRs retrieved from external sources, as described in the Methods section. This testing set is formed of 188 different proteins associated with 84 ADRs. Also, the training and the testing set do not overlap, meaning none of the 188 proteins present in the test set were present in the training set. The proteins associated with each one of the 84 ADRs are predicted using the respective model, and then, the performance score is computed based on the results (**Figure 4**).

Very large differences were not found between the different classifiers. They appear to perform at a

comparable level in terms of accuracy, precision, and AUC, although RF appeared to achieve a higher performance particularly in the case of sensitivity with the highest value for the 3rd quartile of the distribution. In terms of MCC, values are distributed mainly above 0 values with the median values around 0.25, thus indicating non-random predictions (Figure 4).

3.4 Combining Predictors

Since three different classifiers were developed for each ADR, the possibility exists of combining the predictions using consensus scoring functions. Three different approaches were used as described in Methods. In terms of accuracy, precision, recall, and AUC, the values increased when compared to individual predictors in the *jury vote* and *consensus* voting systems (Figure 4). There was not only an improvement but also a general shift toward higher values as distributions were skewed toward higher values. The exception was the *red-flag* consensus that resulted in a worsening of predictions. As described in the Methods section, the *red-flag* method was devised to identify singular predictions.

A similar pattern is observed in the case of MCC values (Figure 4). The distribution of MCC values for *jury vote* and *consensus* voting systems were skewed toward higher values when compared with individual predictors. Thus, the quality of the prediction improved when combining individual predictors. As shown in the of accuracy, precision, and recall, *red-flag* consensus decreased resulted in worse MCC values distributing between 0 (random prediction) and negative (inverse) values. Therefore, it is a better strategy to accept the most common prediction rather than any singular predictor.

3.5 Predicting at SOC Level

The models presented in the previous sections were ADR-specific. However, we also wanted to develop more generalist predictive models that at the same time preserve the biological and medical meaning. For this purpose, we grouped the different ADRs into specific SOCs as per MEDDRA classification (Chang et al., 2017). The MedDRA SOC is defined as the highest level of the MedDRA terminology, distinguished by anatomical or physiological system, aetiology (disease origin), or purpose. Also, most of these describe disorders of a specific part of the body. As explained in the T-ARDIS manuscript (Galletti et al., 2021), not every SOC is present in the database due the fact that some MEDDRA reported ADRs are very general or not specific to body parts, tissues, or underlying human biology (Ietswaart et al., 2020). Specifically, in this study, the 84 ADRs considered were grouped into 18 different SOCs with an average number of 5 ADRs per SOC. At a single classifier level, a large variability of predictions was found in terms of accuracy, precision, sensitivity, and MCC (Figure 5). Predictions were highly accurate in the cases of “*pregnancy, puerperium, and perinatal conditions*” compared to those in the case of *immune* or *nervous* disorders. In general, combining predictors resulted in improved predictions, with the exception of *red-flag* voting,

particularly in terms of recall. However, sensitivity values were generally low when compared to those achieved by predictors working at ADR level (Figure 4). This fact highlights the difficulty of predicting at a higher level of abstraction rather than at individual ADR level.

In terms of MCC values, a similar situation can be observed (Figure 5). There was an improvement of predictions when combining individual prediction in a *jury vote* or *consensus* voting, such in the case of *respiratory, thoracic, and mediastinal disorders* going from a MCC of 0.75 of the best predictor to 0.81 when combining.

4 DISCUSSION

In this work, we set to develop an approach to predict the potential liability of proteins in the context of adverse reactions when targeted for therapeutic purposes. By analyzing the human interactome, a range of network-based metrics were derived to characterize the proteins under study. This range of heterogeneous measurements was then fed into three machine-learning classifiers that were in turn combined using three different voting approaches. The prediction models both at individual ADRs and SOCs level provided a reasonable performance that justified its use as a tool to foresee potential liabilities of proteins. We looked at 84 different ADR in total, being able to create reliable models for each of them.

4.1 Classifiers Performances

The variables used in the predictions were of eight accounting for different aspects of the proteins under study. As shown in Figure 3, the level of discrimination among positive and negative cases varies with GUILDify scores and K1 clustering analyses among the top performers and degree centrality and clustering coefficient analyses as fewer discriminating features. This reflects the small world nature of the human interactome (Zhang and Zhang, 2009). As shown in the results, the performance of the different classifiers varied, with RF being the overall best performed predictor under training conditions, although in particular, ADRs, SVM, and NN were superior. This observation prompted us to develop a voting system to combine the individual predictors in a meta-predictor fashion. As shown in Figures 4, 5, combining the methods resulted in better predictions with the exception of the *red-flag* consensus. Both the *jury vote* and *consensus* voting systems followed the same principle, i.e., to boost coincident predictions among classifiers. In fact, the level of performance of *jury vote* and *consensus* voting systems are comparable (Figures 4, 5), but critically, the *consensus voting system* provides further granularity to the predictions that allows a finer ranking. Indeed; however, for instance, a *jury vote* will place a given protein in a class, e.g., +1; the two methods will agree that the given protein might be linked to a given ADR, and the *consensus* scoring function, however, will provide a quantitative measure that can allow the ranking of proteins within the same class. This aspect is pivotal in

order to establish a degree of confidence in the predictions of the DocTOR application (see below). Finally, as mentioned, the *red-flag* voting system resulted in worse predictions overall. The idea in itself seems counter-intuitive, i.e., promoting the marginal view. However, a few cases are found where this strategy was successful such in the cases of *nocturia*, *neutropenia*, or *ischaemia* ADR (see **Supplementary Figure S1**. tsv or **Supplementary Figure S2**. tsv). Furthermore, the *red-flag* approach serves as a failsafe in the event of an unknown prediction, such as in the instance of the DocTOR utility (explained below), or while two ML approaches, while agreeing, report low probabilities in their respective predictions.

The other aspect to consider in this work was the nature of the predictions. In theory, one of the major achievements of protein-ADR predictions would be determining if targeting a protein would result in an unwanted adverse response, i.e., ADR. However, this is a very difficult question to turn into a predictive model, as the types of ADR are very diverse, and we might end up considering any protein susceptible to causing an ADR to a certain extent. This is the reason why the predictive models were ADR-specific, so that the prediction is not whether a protein might cause an undesired reaction, but what type of adverse reaction. However, grouping ADRs into common SOCs is possible. In doing so, individual ADRs are abstracted into a higher entity, and, thus, more generalist prediction models can be developed, i.e., a model to predict whether the targeting of a given protein can be associated to a specific SOC perturbation. As shown in **Figures 5, 6**, predicting at this level resulted in some SOCs demonstrating better prediction performances than others. SOCs with more defined affected tissues/organs tended to demonstrate better predictions that include more systemic representations. For instance, comparing predictions on the *respiratory, thoracic, and mediastinal disorders vs. immune system disorders* resulted in the former achieving better performances (accuracy: 0.90 vs. 0.54; precision: 0.93 vs. 0.87; recall: 0.87 vs. 0.10; MCC: 0.81 vs. 0.16). Finally, researchers also found that better performance at SOCs related to cases with models already predicted successfully at the individual ADRs included in the particular SOC.

4.2 Difficult to Predict Adverse Drug Reactions

On the other hand, given the complexity of the biological problem, some ADR results are harder to predict. In particular, the worst results have been obtained in 17 different ADRs which obtained a negative or equal to 0 MCC (random predictions). These includes *Hypercoagulation, Ichthyosis, Coordination abnormal, Biliary cirrhosis, Acute hepatic failure, Hyper-ammonaemia, Azoospermia, Diplegia, Glucose tolerance impaired, Haemorrhagic diathesis, Hypoacusis, Ophthalmoplegia, Renal tubular acidosis, Hepatic failure, Coagulopathy, and Ischaemia*. Target on these ADRs included common genes

(**Supplementary Figure S6**. tsv), such as TP53, 5HT1A, ACE, members of the CALM family, LEP, and IL8. In particular, these genes have been already annotated in T-ARDIS as targets with the highest number of associated ADRs (Galletti et al., 2021), thus partially explaining prediction's inaccuracy.

4.3 The DocTOR Utility

The predictive models and accessory scripts to carry out the predictions as well as all the datasets employed in this study are available at the Direct fOreCast Target On Reaction (DocTOR) application available at <https://github.com/cristian931/DocTOR>. The application allows users to upload a list of proteins in the form of UNIPROT identification codes and a list of ADRs of interest (from the available models), in order to study the potential relationship between the two. The program will assign a positive or negative class to the protein output and a probability associated to the given class for all three different classifiers (SVM, NN, and RF) and voting systems (*jury vote*, *consensus*, and *red flag*). Users can, therefore, consider all this information when analyzing the prediction results. Also, the application lends itself to being easily updated, allowing the user to add new models for new ADR on request or retrain existing models when new protein targets are discovered to be associated with certain ADRs and/or given new releases of the T-ARDIS database.

5 CONCLUSION

Predicting associations between protein targets and ADR is desirable, particularly in preclinical drug development, in order to identify early in the process potential liabilities and toxicity-related aspects linked to proteins. In this study, we addressed this problem from an interactome-centric point of view. Next, we collected a range of protein features, including their topology characteristic in the human interactome, the spatial position related to specific *in vitro* validated ADR-related hotspots and their function associations. Also, we trained three different machine-learning approaches to construct models for 84 different ADRs, including a specific DILI related subset and 20 different SOCs using the various features. The models were optimized via grid-search and 5-fold cross-validations, and the results were tested in an independent dataset. The analysis of the performance of the models both under training and independent testing validated its use as a prospective computational tool, to assess the liability of proteins both at the level of specific ADR type and SOC. Finally, we provided access to the data, models, and predictive tools through a dedicated GitHub repository for the use of the scientific community. Researchers will be able to use the DocTOR utility in combination with *in vitro* investigations to assess the potential association between protein target modulation and the onset of ADR, reducing research time.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

NF-F and BO contributed to conception and design of the study. CG carried out the main bulk of the research including the development of methods and data acquisition with help from JA-P. NF-F and CG analyzed the data with help from BO and JA-P. CG wrote the first draft of the manuscript. All

authors contributed to manuscript revision, read, and approved the submitted version.

ACKNOWLEDGMENTS

Authors acknowledge support from MINECO, grant number RYC 2015-17519.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fbinf.2022.906644/full#supplementary-material>

REFERENCES

- Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. <https://www.tensorflow.org/>.
- Aguirre-Plans, J., Piñero, J., Sanz, F., Furlong, L. I., Fernandez-Fuentes, N., Oliva, B., et al. (2019). GUILDiFy v2.0: A Tool to Identify Molecular Networks Underlying Human Diseases, Their Comorbidities and Their Druggable Targets. *J. Mol. Biol.* 431 (13), 2477–2484. doi:10.1016/j.jmb.2019.02.027
- Aguirre-Plans, J., Piñero, J., Souza, T., Callegaro, G., Kunnen, S. J., Sanz, F., et al. (2021). An Ensemble Learning Approach for Modeling the Systems Biology of Drug-Induced Injury. *Biol. Direct* 16 (1), 5. doi:10.1186/s13062-020-00288-x
- Artigas, L., Coma, M., Matos-Filipe, P., Aguirre-Plans, J., Farrés, J., Valls, R., et al. (2020). In-silico Drug Repurposing Study Predicts the Combination of Pirfenidone and Melatonin as a Promising Candidate Therapy to Reduce SARS-CoV-2 Infection Progression and Respiratory Distress Caused by Cytokine Storm. *PLoS One* 15 (10), e0240149. doi:10.1371/journal.pone.0240149
- Bailey, J., Thew, M., and Balls, M. (2014). An Analysis of the Use of Animal Models in Predicting Human Toxicology and Drug Safety. *Altern. Lab. Anim.* 42 (3), 181–199. doi:10.1177/026119291404200306
- Basile, A. O., Yahi, A., and Tatonetti, N. P. (2019). Artificial Intelligence for Drug Toxicity and Safety. *Trends Pharmacol. Sci.* 40 (9), 624–635. doi:10.1016/j.tips.2019.07.005
- Bender, A., Scheiber, J., Glick, M., Davies, J. W., Azzaoui, K., Hamon, J., et al. (2007). Analysis of Pharmacology Data and the Prediction of Adverse Drug Reactions and Off-Target Effects from Chemical Structure. *ChemMedChem* 2 (6), 861–873. doi:10.1002/cmdc.200700026
- Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast Unfolding of Communities in Large Networks. *J. Stat. Mech.* 2008 (10), P10008. doi:10.1088/1742-5468/2008/10/p10008
- Bowes, J., Brown, A. J., Hamon, J., Jarolimek, W., Sridhar, A., Waldron, G., et al. (2012). Reducing Safety-Related Drug Attrition: the Use of *In Vitro* Pharmacological Profiling. *Nat. Rev. Drug Discov.* 11 (12), 909–922. doi:10.1038/nrd3845
- Bowes, J., Brown, A. J., Hamon, J., Jarolimek, W., Sridhar, A., Waldron, G., et al. (2012). Reducing Safety-Related Drug Attrition: the Use of *In Vitro* Pharmacological Profiling. *Nat. Rev. Drug Discov.* 11 (12), 909–922. doi:10.1038/nrd3845
- Cao, M., Pietras, C. M., Feng, X., Doroschak, K. J., Schaffner, T., Park, J., et al. (2014). New Directions for Diffusion-Based Network Prediction of Protein Function: Incorporating Pathways with Confidence. *Bioinformatics* 30 (12), i219–27. doi:10.1093/bioinformatics/btu263
- Ceol, A., Chatr Aryamontri, A., Licata, L., Peluso, D., Briganti, L., Perfetto, L., et al. (2009). MINT, the Molecular Interaction Database: 2009 Update. *Nucleic Acids Res.* 38 (Suppl. 1_1), D532–D539. doi:10.1093/nar/gkp983
- Chang, L. C., Mahmood, R., Qureshi, S., and Breder, C. D. (2017). Patterns of Use and Impact of Standardised MedDRA Query Analyses on the Safety Evaluation and Review of New Drug and Biologics License Applications. *PLOS ONE* 12 (6), e0178104. doi:10.1371/journal.pone.0178104
- Chooobar, S., Ahsen, M. E., Crawford, J., Tomasoni, M., Fang, T., Lamparter, D., et al. (2019). Assessment of Network Module Identification across Complex Diseases. *Nat. Methods* 16 (9), 843–852. doi:10.1038/s41592-019-0509-5
- Dara, S., Dhamecherla, S., Jadav, S. S., Babu, C. M., and Ahsan, M. J. (2022). Machine Learning in Drug Discovery: A Review. *Artif. Intell. Rev.* 55 (3), 1947–1999. doi:10.1007/s10462-021-10058-4
- Drozdzetskiy, A., Cole, C., Procter, J., and Barton, G. J. (2015). JPred4: a Protein Secondary Structure Prediction Server. *Nucleic Acids Res.* 43 (W1), W389–W394. doi:10.1093/nar/gkv332
- Galletti, C., Mirela Bota, P., Oliva, B., and Fernandez-Fuentes, N. (2021). Mining Drug–Target and Drug–Adverse Drug Reaction Databases to Identify Target–Adverse Drug Reaction Relationships. *Database (Oxford)*. 2021: baab068. doi:10.1093/database/baab068
- García-García, J., Guney, E., Aragües, R., Planas-Iglesias, J., and Oliva, B. (2010). Biana: a Software Framework for Compiling Biological Interactions and Analyzing Networks. *BMC Bioinforma.* 11 (1), 56. doi:10.1186/1471-2105-11-56
- Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., et al. (2017). The ChEMBL Database in 2017. *Nucleic Acids Res.* 45 (D1), D945–D954. doi:10.1093/nar/gkw1074
- Gavin, A. C., Maeda, K., and Kühner, S. (2011). Recent Advances in Charting Protein–Protein Interaction: Mass Spectrometry-Based Approaches. *Curr. Opin. Biotechnol.* 22 (1), 42–49. doi:10.1016/j.copbio.2010.09.007
- Goh, K. I., and Choi, I. G. (2012). Exploring the Human Diseaseome: the Human Disease Network. *Brief. Funct. Genomics* 11 (6), 533–542. doi:10.1093/bfpg/els032
- Güldener, U., Münsterkötter, M., Oesterheld, M., Pagel, P., Ruepp, A., Mewes, H. W., et al. (2006). MPact: the MIPS Protein Interaction Resource on Yeast. *Nucleic Acids Res.* 34. (Database issue), D436–D441. doi:10.1093/nar/gkj003
- Guney, E., and Oliva, B. (2012). Exploiting Protein–Protein Interaction Networks for Genome-wide Disease–Gene Prioritization. *PLOS ONE* 7 (9), e43557. doi:10.1371/journal.pone.0043557
- Guney, E. 2017. “Investigating Side Effect Modules in the Interactome and Their Use in Drug Adverse Effect Discovery,” in *Complex Networks VIII*. CompleNet 2017. Editors B. Gonçalves, R. Menezes, R. Sinatra, and V. Zlatić (Cham: Springer Proceedings in Complexity), 239–250. doi:10.1007/978-3-319-54241-6_21
- Gysi, D. M., Valle, I. D., Zitnik, M., Ameli, A., Gan, G., Varol, O., et al. (2021). Network Medicine Framework for Identifying Drug-Repurposing Opportunities for COVID-19. *Proc. Natl. Acad. Sci.* 118 (19), e2025581118. doi:10.1073/pnas.2025581118
- Huang, L. H., He, Q. S., Liu, K., Cheng, J., Zhong, M. D., Chen, L. S., et al. (2018). ADReCS-Target: Target Profiles for Aiding Drug Safety Research and Application. *Nucleic Acids Res.* 46 (D1), D911–D917. doi:10.1093/nar/gkx899

- Ietswaart, R., Arat, S., Chen, A. X., Farahmand, S., Kim, B., DuMouchel, W., et al. (2020). Machine Learning Guided Association of Adverse Drug Reactions with *In Vitro* Target-Based Pharmacology. *EBioMedicine* 57, 102837. doi:10.1016/j.ebiom.2020.102837
- Kerrien, S., Alam-Farouque, Y., Aranda, B., Bancarz, I., Bridge, A., Derow, C., et al. (2006). IntAct—open Source Resource for Molecular Interaction Data. *Nucleic Acids Res.* 35 (Suppl. 1_1), D561–D565. doi:10.1093/nar/gkl958
- Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., et al. (2008). Human Protein Reference Database--2009 Update. *Nucleic Acids Res.* 37 (Suppl. 1_1), D767–D772. doi:10.1093/nar/gkn892
- Kotlyar, M., Pastrello, C., Ahmed, Z., Chee, J., Varyova, Z., and Jurisica, I. (2022). IID 2021: towards Context-specific Protein Interaction Analyses by Increased Coverage, Enhanced Annotation and Enrichment Analysis. *Nucleic Acids Res.* 50 (D1), D640–D647. doi:10.1093/nar/gkab1034
- Kuhn, M., Al Banchaabouchi, M., Campillos, M., Jensen, L. J., Gross, C., Gavin, A. C., et al. (2013). Systematic Identification of Proteins that Elicit Drug Side Effects. *Mol. Syst. Biol.* 9 (1), 663. doi:10.1038/msb.2013.10
- Kuhn, M., Letunic, I., Jensen, L. J., and Bork, P. (2015). The SIDER Database of Drugs and Side Effects. *Nucleic Acids Res.* 44 (D1), D1075–D1079. doi:10.1093/nar/gkv1075
- Kumar, A. (2018). The Newly Available FAERS Public Dashboard: Implications for Health Care Professionals. *Hosp. Pharm.* 54 (2), 75–77. doi:10.1177/0018578718795271
- Lo, Y. C., Rensi, S. E., Torng, W., and Altman, R. B. (2018). Machine Learning in Chemoinformatics and Drug Discovery. *Drug Discov. Today* 23 (8), 1538–1546. doi:10.1016/j.drudis.2018.05.010
- Madorran, E., Stožer, A., Bevc, S., and Maver, U. (2020). *In Vitro* toxicity Model: Upgrades to Bridge the Gap between Preclinical and Clinical Research. *Bosn. J. Basic Med. Sci.* 20 (2), 157–168. doi:10.17305/bjbms.2019.4378
- Mizutani, S., Pauwels, E., Stoven, V., Goto, S., and Yamanishi, Y. (2012). Relating Drug-Protein Interaction Network with Drug Side Effects. *Bioinformatics* 28 (18), i522–i528. doi:10.1093/bioinformatics/bts383
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., and Thirion, B. (2011). Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* 12 (85), 2825–2830.
- Piñero, J., Ramírez-Anguita, J. M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., et al. (2019). The DisGeNET Knowledge Platform for Disease Genomics: 2019 Update. *Nucleic Acids Res.* 48 (D1), D845–D855. doi:10.1093/nar/gkz1021
- Re3data.Org (2014). *MedEffect Canada - Adverse Reaction Database*. re3data.org - Registry of Research Data Repositories. doi:10.17616/R3J03W
- Sahoo, B. M., Ravi Kumar, B. V. V., Sruti, J., Mahapatra, M. K., Banik, B. K., and Borah, P. (2021). Drug Repurposing Strategy (DRS): Emerging Approach to Identify Potential Therapeutics for Treatment of Novel Coronavirus Infection. *Front. Mol. Biosci.* 8, 628144. doi:10.3389/fmolb.2021.628144
- Seyhan, A. A. (2019). Lost in Translation: the Valley of Death across Preclinical and Clinical Divide - Identification of Problems and Overcoming Obstacles. *Transl. Med. Commun.* 4 (1), 18. doi:10.1186/s41231-019-0050-7
- Singh, V. K., and Seed, T. M. (2021). How Necessary Are Animal Models for Modern Drug Discovery? *Expert Opin. Drug Discov.* 16 (12), 1391–1397. doi:10.1080/17460441.2021.1972255
- Smit, I. A., Afzal, A. M., Allen, C. H. G., Svensson, F., Hanser, T., and Bender, A. (2021). Systematic Analysis of Protein Targets Associated with Adverse Events of Drugs from Clinical Trials and Postmarketing Reports. *Chem. Res. Toxicol.* 34 (2), 365–384. doi:10.1021/acs.chemrestox.0c00294
- Stark, C., Breitkreutz, B. J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). BioGRID: a General Repository for Interaction Datasets. *Nucleic Acids Res.* 34 (Suppl. 1_1), D535–D539. doi:10.1093/nar/gkj109
- Tatonetti, N. P., Ye, P. P., Daneshjoui, R., and Altman, R. B. (2012). Data-Driven Prediction of Drug Effects and Interactions. *Sci. Transl. Med.* 4 (125), 125ra31. doi:10.1126/scitranslmed.3003377
- Wong, C. K., Ho, S. S., Saini, B., Hibbs, D. E., and Fois, R. A. (2015). Standardisation of the FAERS Database: a Systematic Approach to Manually Recoding Drug Name Variants. *Pharmacoepidemiol Drug Saf.* 24 (7), 731–737. doi:10.1002/pds.3805
- Xiang, Z., Gong, W., Li, Z., Yang, X., Wang, J., and Wang, H. (2021). Predicting Protein-Protein Interactions via Gated Graph Attention Signed Network. *Biomolecules* 11 (6), 799. doi:10.3390/biom11060799
- Xing, S., Wallmeroth, N., Berendzen, K. W., and Grefen, C. (2016). Techniques for the Analysis of Protein-Protein Interactions *In Vivo*. *Plant Physiol.* 171 (2), 727–758. doi:10.1104/pp.16.00470
- Zhang, Z., and Zhang, J. (2009). A Big World inside Small-World Networks. *PLOS ONE* 4 (5), e5686. doi:10.1371/journal.pone.0005686

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Galletti, Aguirre-Plans, Oliva and Fernandez-Fuentes. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.