In silico tools to study diseases and polypharmacology through the lens of network medicine

Joaquim Aguirre Plans

TESI DOCTORAL UPF / 2021

THESIS SUPERVISORS

Dr. Baldo Oliva Miguel

Dr. Narcis Fernandez Fuentes (Universitat de Vic)

DEPARTMENT OF EXPERIMENTAL AND HEALTH SCIENCES

Als meus pares,
per fer-ho tot més fàcil.

Acknowledgments

M'agradaria començar la tesi donant les gràcies als meus pares, perquè tot això és gràcies a ells. A la meva **mare**, per ser la millor mare del món. Gràcies per donar-me sempre el teu suport incondicional, preocupar-te per mi, i regalar-me moltes de les teves qualitats. Al meu **pare**. Et trobo tant a faltar! Tan de bo estiguessis aquí per veure com acabo la tesi. Però sé perfectament com d'orgullós estaries, perquè sempre m'ho vas fer saber. Això és gràcies a tu, i va per tu.

M'agradaria donar les gràcies als meus supervisors, per ser els millors que un estudiant de doctorat pugui demanar. Primer de tot, a en **Baldo**. Per confiar en mi des del principi, quan només era un estudiant de màster que volia fer unes pràctiques a un laboratori del PRBB. Gràcies per dedicar-me tot el temps necessari en fer-me créixer com a investigador. Tot el que sé del món acadèmic ho he après de tu. I a en **Narcis**, gràcies per tots els consells i bones paraules que m'has donat durant el doctorat. Gràcies a tots dos per ser tan bons supervisors, heu posat el llistó als núvols. Tan de bo tots els caps que tingui en un futur siguin així.

Agrair també a tots els membres del Structural Bioinformatics Lab. Tan de bo coincidim, dins o fora del món científic, en un futur. En especial a l'**Alberto**, ha sigut genial estar literalment al teu costat durant tots aquests anys, primer com a companys de màster, i després de doctorat. M'emporto molts riures veient vídeos de YouTube random (d'entre els quals moltes perles!). Thank you **Emre**, for being the postdoc figure, network passionate and friend that I needed, all in one person. Thanks for the guidance, advice, and the nice times that you've shared with me. You made it way easier for me! Thank you **Patri**, for being a great friend and volleyball partner. For a lot of conversations and videos mainly about Oscar, but also Linking Park songs, frankfurts in the stadium bar, and many other happy moments! Gràcies **Ruben**, per ser tan bon amic i company de labo. Ha estat genial compartir tants bons moments amb tu! And to the rest of the members that I have met during this time: **Laura**, **Filip**, **Cristiano**, **Alexis**, **Altair, Gaurav**, **Ida**, **Alejandro**, **Guillem**, and many others!

No em vull oblidar de tota la gent del GRIB, que m'ha fet la vida una mica més fàcil durant aquest temps. Sobretot, a l'**Alfons** i en **Miguel**, per la paciència incansable que heu tingut amb mi, i per tot el suport i consells. A tots els amics i companys de departament, amb qui he compartit molts migdies a la cafeteria o a les terrasses: **Mariona, Judith, Andreu, Juanma, Janet** i molts altres! I a tots els companys de màster, que em van fer viure dos anys de la meva vida genials compartint passió per la bioinformàtica.

Als meus amics de Biotec, perquè sense ells no em puc imaginar haver arribat aquí. Aquesta tesi és en gran part gràcies a vosaltres. En especial a la **Natalia**, per ser una gran amiga i pilar fonamental des de que et vaig conèixer. Per aquells "birriernes" tan necessaris, que han fet més passables els moments més durs. Al **David**, per ser la meva constant des de l'institut, i per fer-me riure literalment sempre. A la **Miriam**, perquè ets una gran amiga, i es nota molt quan no estàs a prop. Al **Roger**, per tot el teu suport sense esperar res a canvi, i per avivar la meva passió per la ciència cada cop que ens veiem. Al **Guillermo**, el meu riojano preferit, per fer els meus anys de grau més fàcils. A la **Laura**, la meva pianista youtuber preferida! I a la **Silvia**, per ser l'alegria personificada.

Als meus amics del cole i institut, sobretot al **Guifré** i al **Jona**, per haver-me acompanyat durant les diferents etapes de la meva vida. I també a l'**Uri**, **Miguel**, **Carles**, **Ángel**, **Berni** i **Marc**.

Als meus companys de hockey, gràcies per acceptar-me de bon grat a l'equip i fer la meva vida extra-acadèmica molt més amena.

A la meva família, en especial als meus germans **Pau**, **Mar**, **Cesc** i **Andreu** per ser un altre pilar fonamental en la meva vida. Als meus **avis**, per estimar-me incondicionalment. Gràcies per fer-me sentir sempre tan estimat. I a tots els meus tiets i cosins, que sempre que us veig m'alegreu el dia.

I finalment, i no podia acabar d'una altra manera, a la **Maria**. Gràcies per ser la meva companya d'aventura, per estar en els millors i els pitjors moments. Sense tu, això hagués estat impossible. T'ho dec tot. Ets la millor.

ABSTRACT

The inner workings of cells can be understood as an interplay of interactions between biomolecules, forming a network known as the interactome. Drugs and diseases can be considered as perturbations in this network, modulating directly specific molecules, but indirectly communities of molecules whose interactions are affected by the perturbation. Network medicine seeks to accurately represent and analyze biological networks to understand diseases and find safer and more effective treatments. In this thesis, I present several *in silico* tools for network medicine, addressed to study the molecular mechanisms of diseases and drugs. These tools are used in a wide range of novel and diverse applications of network medicine, such as the study of comorbidities, endophenotypes, side effects, drug combinations and drug repurposing.

RESUM

El funcionament intern de les cèl·lules pot entendre's com un conjunt d'interaccions entre biomolècules, formant una xarxa que coneixem amb el nom d'interactoma. Els fàrmacs i malalties poden considerar-se pertorbacions d'aquesta xarxa, modulant directament molècules específiques, però indirectament comunitats de molècules les interaccions de les quals es veuen afectades per la pertorbació. La medicina de xarxes busca representar i analitzar amb precisió les xarxes biològiques, per tal d'entendre millor les malalties i aconseguir tractaments més segurs i eficaços. En aquesta tesi, presento diverses eines *in silico* basades en la medicina de xarxes, pensades per estudiar els mecanismes moleculars de malalties i fàrmacs. Aquestes eines s'utilitzen en un ampli ventall d'aplicacions de la medicina de xarxes, com per exemple l'estudi de comorbiditats, endofenotips, efectes secundaris, reutilització i combinació de fàrmacs.

PREFACE

It was the 1st of October of 2014 when I first heard about the words "network medicine". Professor Enrique Querol, our teacher of omics-sciences during my Bachelor's in Biotechnology, recommended us to go to the talk that was opening the course 2014-15, which was given by one of his former students. As Enrique was one of my favorite teachers, I decided to follow his advice and attend to the talk. At this moment, I couldn't imagine it, but this decision influenced a lot the shape that my career took until now.

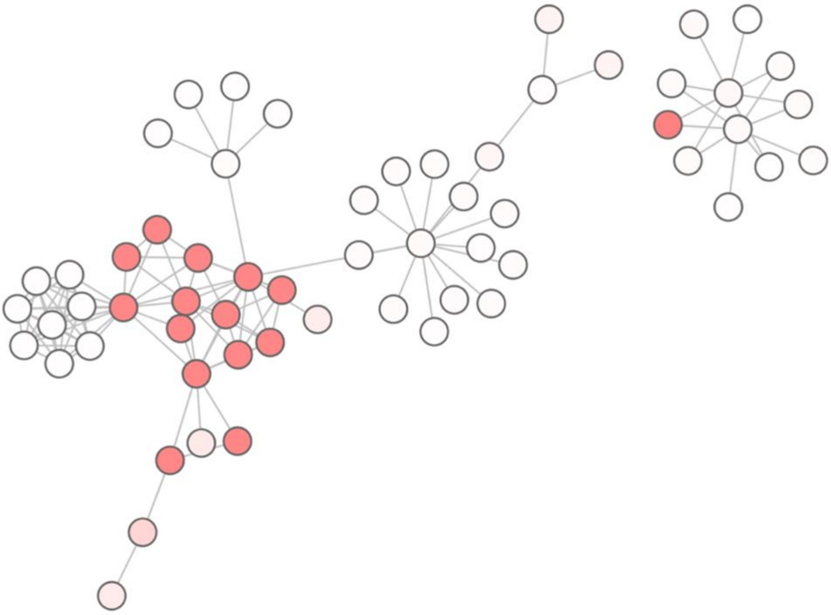
The talk, given by Dr. Patrick Aloy, introduced us the concept of network medicine. How this discipline allows us to organize and analyze the thousands of records of experimental data about molecules and interactions using networks. He also stressed the close relationship of this discipline with pharmacological and clinical data, aiding the understanding of diseases and guiding the drug development process. I suspect that it was many factors (maybe how organized are networks, or maybe how analyzing data can have an impact in medicine), but this talk clarified a lot my scientific itinerary.

Two years later, I found in Professor Baldo Oliva's lab, who is precisely one of Aloy's mentors, the perfect environment to start my personal quest in network medicine. Baldo gave me the freedom to explore different projects which at the beginning seemed disconnected, but after some time working in them, I could appreciate how network medicine was linking them. This thesis is the result of my initial years exploring this young discipline, learning bit by bit from it.

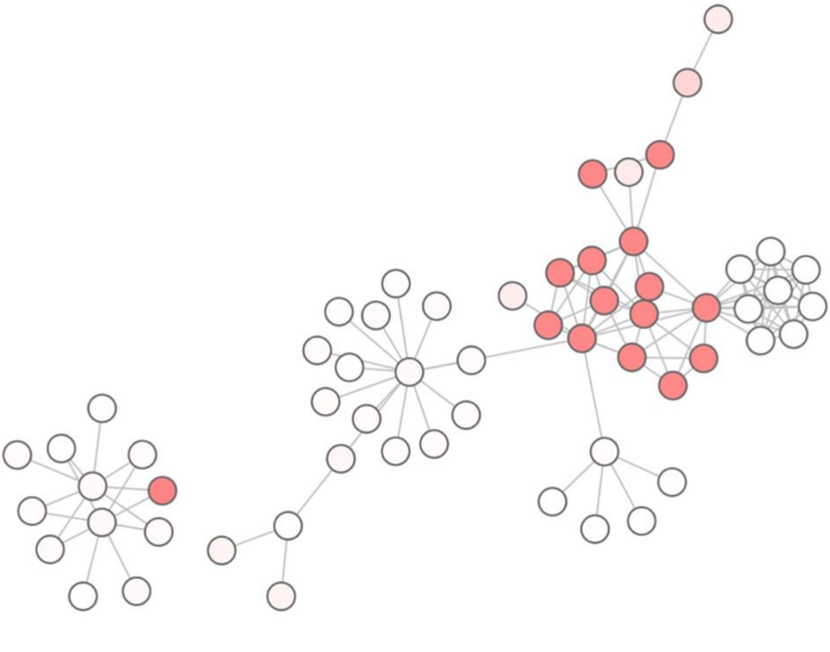
TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES



1. INTRODUCTION



1.1. The human interactome

1.1.1. Molecules conforming the human interactome

The human organism is made of cells, which are the structural and functional units of life. Each cell type is different and carefully placed within organs and tissues to undergo a specific function. At molecular level, cells are made of a wide collection of different molecules that interact between themselves: deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) molecules, genes, proteins, and metabolites, to name just a few.

The instructions for each organism are encoded within the sequence of the DNA. The DNA is a double helical structure composed of 4 types of nucleotides: adenine, guanine, cytosine and thymine (in eukaryotic organisms). These bases are organized precisely in a sequence that in human can be comprised of about 3 billion nucleotide bases (1). DNA molecules are organized in chromosomes, placed in the nuclei of cells, and are inherited from the parents after a molecular recombination of the parental chromosomes during the process of meiosis. A small fraction of the DNA is also located in the mitochondria, a small cellular organelle, and is inherited exclusively from the mother. The complete set of DNAs of an organism is called genome.

About 1% of the genome is made of genes, the coding regions of the DNA. Genes are sequences of nucleotides with the capacity to encode the synthesis of proteins. This process is described by the central dogma of molecular biology: DNA is transcribed by another type of sequence called messenger ribonucleic acid (mRNA) that in tun is translated into proteins.

But what are proteins? Proteins are the brick and mortar of the cell, the structural molecules that are used to build the scaffold of cells. But they are also functional units, being involved in virtually all the processes occurring in and out of cells that ensure the correct functioning of our body. In order to carry out their functions, proteins interact with each other and with other biomolecules. At a global view, each cell in the human body can be seen as a complex network of proteins specifically interacting with other proteins to execute the functions of the cell. This is often referred as **protein-protein interaction (PPI) network** or as the **human interactome**.

1.1.2. Omics sciences to study the human molecules

The last couple of decades has witnessed a notable advance on the technological development that has ushered the so-called ‘omics technologies. Indeed, several branches of science have emerged with the purpose to identify and characterize all the biomolecules that explain the functioning of the human organism. These disciplines are known as “omics” because their names end with the suffix -omics, which is used to refer to high-throughput technologies (2). They usually employ experimental and computational techniques to compile largescale datasets of biomolecules and understand their role. Therefore, they are closely related with bioinformatics. Here, we will focus on three omics: genomics, proteomics and interactomics:

- **Genomics:** Genomics is a discipline that studies the structure, content, and evolution of the genome. It has the objective of elaborate genetic maps, detailed genomes, annotation of genes, identify the genome variability between different individuals, analyze the expression and function of genes.

- **Proteomics:** Proteomics studies the proteome, which is the collection of proteins encoded by the genome. It has several sub-branches: (i) proteomics of expression, identifies and quantifies proteins and identifies their cellular location; (ii) functional proteomics, determines the function of proteins; (iii) structural proteomics, studies the tridimensional structure of proteins.

- **Interactomics:** Interactomics is a sub-branch of proteomics that studies the interactions between proteins. As mentioned, proteins do not act alone, carry out their function through complex network of interactions known as the interactome. Interactomics identifies the interactions between proteins and characterizes their role and mechanism.

1.1.3. The underlying PPI network of a cell

Recent advances in molecular biology offer us information on a wide range of cellular components as individuals. However, as we observed in the previous sections, to understand the complexity of the human organism, we need to represent the cell as a network of interacting components. These components (proteins, genes, metabolites…) interact with each other to exert the cellular functions (3).

The underlying network of a cell, also known as interactome, is made by the interactions of all these cellular components. The interactome can be represented either as a multilayered network of different molecules interacting with each other, or as separated networks of the distinct types of interaction. Proteins are the key molecules of the interactome, as they act as both structural and functional elements of the cell. For this reason, among the different types of biological networks, the PPI network is especially important to understand the molecular mechanisms. The question is, how to detect reliable PPIs? How can we unveil the molecular details of the PPIs? And how can we represent them as a complete network?

1.1.4. Types of PPI detection methods

The first step in the process of building a complete PPI network is to detect the PPIs, which are estimated to be approximately 650,000 (4) from 10,000 different types (5). There is a wide diversity of methods to find and study PPIs (**Table 1**). These methods can be classified in different ways depending on their nature, the scale of the experiment or the level of detail of the results (6).

Table 1. Experimental PPI detection methods classified by scale and result resolution.

Method	Yield	Result resolution				
		Co-expression / localization	**Complex**	**Binary interaction**	**Interface**	**Structure**
Co-localization	High	✓	✗	✗	✗	✗
Protein microarrays	High	✓	✗	✗	✗	✗
RNA-seq	High	✓	✗	✗	✗	✗
Tandem Affinity Purification	High	✓	✓	✗	✗	✗
Anti-tag coimmunoprecipitation	High	✓	✓	✗	✗	✗
Yeast Two-Hybrid	High	✓	✗	✓	✗	✗
Nuclear Magnetic Resonance	Low	✓	✓	✓	✓	✓
X-ray crystallography	Low	✓	✓	✓	✓	✓
Cryo-electron microscopy	Low	✓	✓	✓	✓	✓
Micro-electron diffraction	Low	✓	✓	✓	✓	✓

Depending on the nature of the techniques used, PPI detection methods can be considered experimental or computational. **Experimental methods** are timely and economically expensive. They can either be *in vivo* (using a living organism) or *in vitro* (outside the normal biological context). In contrast, **computational methods** use as basis the knowledge generated by experimental methods to infer new predictions. Therefore, computational methods have smaller costs and can be used as a complement of the experimental methods. The predictions of computational methods must be validated with experimental methods. Thus, the quality of their predictions will rely on the quality of the experimental data used to infer and validate the predictions.

Depending on the scale of the experiment, PPI detection methods can be classified as **low-throughput** when a small set of proteins is studied, or **high-throughput** when a large set of proteins is systematically studied. High-throughput methods can detect large datasets of PPIs, but they also tend to detect false positive interactions (7,8). Yet, there are also high-throughput methods such as yeast two-hybrid which, on the contrary, do not detect many false positives but tend to miss true positives.

Finally, depending on the level of detail of the results, PPI detection methods can be classified in five groups (see **Figure 1** for a schematic overview):

(1) Methods to detect co-expressed and co-localized proteins

A clear sign of an interaction between two or more proteins is that they are found in the same place (co-localization) at the same time (co-expression). Although co-expression and co-localization are not sufficient to ensure that an interaction will exist, they have been used to predict functional relationships between proteins, validate experimental results or remove false interactions (9). Co-expression methods include techniques such as microarray or RNA-seq. Co-localization methods are usually based on the fluorescent labelling of proteins (10).

(2) Methods to detect proteins belonging to the same cellular complex

Proteins belonging to the cellular complex are more likely to interact with each other. The methods that detect interactions between groups of proteins without pairwise determination are called *co-complex* methods. The most common co-complex method is tandem affinity purification. Tandem affinity purification consists in tagging an individual protein (bait) and using it to catch a group of proteins (preys), which later are separated and identified. Co-complex methods are usually high-throughput techniques, which can make PPI detections at large scale. However, they do not necessarily detect physical interactions, as not all the proteins of the complex have to interact with each other, giving rise to a high number of false positives in comparison with more specific methods (8,11).

(3) Methods to detect binary PPIs

These methods uncover physical, i.e., direct, interactions between pairs of proteins. They can either be low or high-throughput methods, but generally they are designed to study smaller sets of proteins than co-complex methods, therefore being more precise in detecting physical interactions. The most extended group of methods to predict binary PPIs are Protein Complementation Assays. In this group of methods, the two proteins of interest (bait and prey) are covalently linked to incomplete fragments of a third protein (reporter). If the bait and prey proteins interact, the reporter proteins get close enough to become functional and detectable. Among Protein Complementation Assays, we find the Yeast Two-Hybrid method, which is one of the most popular PPI detection methods (6,8).

There are also computational methods that have been developed to complement these experimental methods: genomic-based methods such as gene fusion, conservation of gene neighborhood or phylogenetic profiles; experimental knowledge-based methods such as interologs, domain profiles or sequence signatures; evolution-based methods such as correlated mutations or phylogenetic mirror trees (12).

(4) Methods to study the interface of PPIs

In a PPI, the interface is the part of the proteins that is interacting. Defining the interface is essential to understand the molecular mechanism of the interaction (13). The most straightforward methods to define the interface of interaction between two proteins are the methods that determine the structure of the interaction. Apart of these methods (which are explained in more detail in the following point), there are experimental methods that determine the interface without providing atomic details. For example, detecting the domains involved in the interaction by removing domains (14) or looking for specific mutations that disrupt the interaction. Also, Yeast Two-Hybrid variations applied to identify interacting domains (15,16).

There are also several computational methods to predict the interface of PPIs. These methods can either be focused on identifying binding sites of individual proteins, or on identifying pairs of interacting residues (6).

(5) Methods to obtain atomic details of PPIs

These methods focus on obtaining the structural details of the residues involved in the interaction. The traditional experimental methods to obtain precise structural information are Nuclear Magnetic Resonance (NMR) spectroscopy and X-ray crystallography. NMR consists in applying a magnetic field to the protein sample, producing an energy transfer emitting a signal that can be processed to generate a spectrum. However, the signal generated by this method is small in comparison with other techniques (17). X-ray crystallography is the most common method. It consists in making the protein sample form crystals that need to be suitable for X-ray radiation. The diffraction pattern of the X-rays is collected and used to calculate the electron density of the sample in three dimensions. The main disadvantage in X-ray crystallography is to obtain crystals of sufficient quality for diffraction. In the recent years there has been a revolution due to the advancements in cryo-electron microscopy (cryoEM) and micro-electron diffraction (microED). CryoEM consists in freezing protein samples in a cryogen and observe them in the transmission electron microscope. Freezing the samples allows to examine them without additional staining that reduce the resolution. Also, it permits the samples to tolerate higher electron beam doses (18). CryoEM is generally easier use in large complexes, which is ideal to solve PPI structures. Another technique gaining more and more recognition recently is microED. In microED, electrons are accelerated to interact with protein nanocrystals and generate a diffraction pattern. The clear advantage over X-ray crystallography is that it is easier to generate crystals that are suitable for microED.

There are also computational methods to predict the structure details of PPIs. The methods can either be based on comparative modelling (using as a modelling template the structures of two interacting homologs) or docking (sampling the possible orientations of the two unbound structures of the interacting proteins).

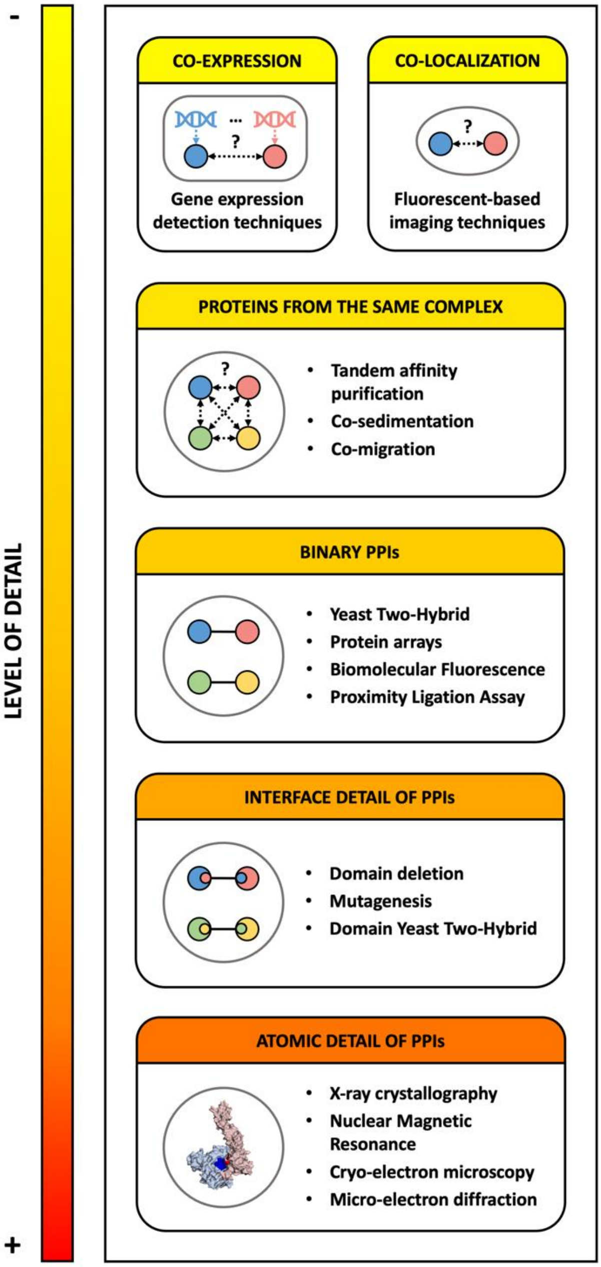

-
LEVEL OF DETAIL
+
CO-EXPRESSION
?
Gene expression detection techniques
CO-LOCALIZATION
?
Fluorescent-based imaging techniques
PROTEINS FROM THE SAME COMPLEX
?
• Tandem affinity purification
• Co-sedimentation
• Co-migration
BINARY PPIs
• Yeast Two-Hybrid
• Protein arrays
• Biomolecular Fluorescence
• Proximity Ligation Assay
INTERFACE DETAIL OF PPIs
• Domain deletion
• Mutagenesis
• Domain Yeast Two-Hybrid
ATOMIC DETAIL OF PPIs
• X-ray crystallography
• Nuclear Magnetic Resonance
• Cryo-electron microscopy
• Micro-electron diffraction

Figure 1. Scheme of the different types of experimental PPI detection methods classified by the level of detail of the results (from lower to higher). In each box, it is shown a schematic figure of the type of PPI detection method and a list of the most representative methods.

1.1.5. Structural characterization of protein folds and PPIs by computational methods

The determination of both individual proteins and PPIs in atomic detail is key to understand the molecular mechanism of interactions. As seen previously, there are several experimental methods to unveil the structural details of PPIs (X-ray crystallography, nuclear magnetic resonance, cryoEM, microED). We also dispose of important repositories of protein and PPI structures such as the Protein Data Bank (PDB) (19) or 3did (20). Yet, the amount of experimentally determined 3D structures is limited. Structural models derived by computational methods can be used to close the gap between the number of known interactions and their structures. The process of computationally predict the 3D structure of a proteins or PPIs is comprised of two steps: modelling and scoring. In the following subsections we review the main strategies to computationally model protein folds, PPIs and score them (see **Figure 2** for a schematic view).

1.1.5.1. Modelling protein folds

There are two types of methods to predict the 3D structure of protein folds (**Figure 2**) (21):

(1) Template-based modelling

This type of method uses a previously determined structure of a related protein as a template to model the unknown structure of the target protein. The basic steps in template-based modelling are: (i) select a closely-related structural template by using single-sequence search methods such as BLAST (22) to scan sequences from PDB database (19); (ii) align the sequence of the protein target with the sequence of the template; and (iii) use modelling tools such as Modeller (23) or SWISS-MODEL (24) to build models of the target protein by performing side-chain optimization of the residues that differ from the original template and rebuilding the backbone around the insertions and deletions (21).

(2) Template-free modelling

These methods are usually applied when the target protein is not similar to any of the known structures in PDB. The methods usually implement a conformational sampling strategy that generates multiple candidate models, and a scoring function that ranks the models by their quality. Briefly, the process usually starts with a multiple-sequence alignment of the target protein and related homolog sequences. The alignment is used to predict local structural features from the secondary structure, and non-local features such as residue-residue contacts or inter-residue distances. These features guide the process of building 3D models that are refined and ranked (21).

Many new methods are starting to use approaches from both categories: there are template-based methods that employ energy-guided model refinement, and template-free methods that exploit information from previously known structures applying machine learning approaches (21).

1.1.5.2. Modelling PPIs

There are two main strategies to predict the 3D structure of PPIs (**Figure 2**):

(1) Comparative modelling

This strategy is used if the sequences of the two interacting proteins are known and the structure of the interaction between the two homologs in another organism (interologs) is available. In this case, we can use the structure of the interologs as a template to model the new structure (25). This method is based on the principle that the structure tends to be more conserved across species than the sequence. Therefore, the sequences of proteins that are enough similar (homologs) acquire a very similar structural conformation (26). In **Appendix 6.5**, we present MODPIN, a method that automatizes the whole comparative modelling process to obtain an ensemble of structural models of the PPI of interest. These models are clustered according to common structural elements in their interfaces and evaluated using scoring functions (27).

(2) Docking

This strategy is used if the structures of the two unbound interacting proteins are known. Docking methods use these structures to sample the possible orientations of the proteins, produce several predictions and rank them according to a scoring function. Docking methods can be classified in two categories depending on their consideration of the conformational changes upon binding: (i) rigid-body docking algorithms ignore any conformational change occurring after the binding; (ii) flexible-body docking algorithms take into account this conformational change in several levels, for example by smoothing the protein surfaces, or by allowing sidechain and/or backbone flexibility, either during docking, or afterwards during a refinement step (28). There is a wide range of computational methods to predict PPIs through docking (29,30), among which we can find ZDOCK (31), V-D^{2}OCK (32), HADDOCK (33) or AutoDock (34).

1.1.5.3. Scoring models of proteins and PPIs

In the recent CASP and CAPRI competitions, we have observed a dramatic progress in the quality of the template-free models made by novel computational methods involving deep learning techniques (30,35,36). However, these methods need to be complemented by evaluation methods to know the margins of accuracy when we study the role of structural models in a biological system. The evaluation methods score and rank the protein fold and PPI models obtained so that the best ones are selected.

The evaluation methods applied to score protein folds and the ones applied to score PPIs are usually based on the same principles. The only difference is that the ones applied to protein folds evaluate the whole protein structure, whereas the ones applied to PPIs focus on the area of the interface (37).

Evaluation methods can be classified into two categories: single- and multiple-model methods. Single-model methods only require one model as input, whereas multiple-model methods require several. The latter ones take advantage of the similarity between the distinct models to evaluate them, but they are not based on the properties of the model itself. In contrast, single-model methods are often based on the geometric and energetic analysis of the model coordinates, although some of them may also use additional information (e.g. for evolutionary related proteins) (38,39).

For single-model methods, the most common approach is to use knowledge-based potentials, i.e. scoring functions derived from the analysis of empirical data (40). The global minimum of these scoring functions corresponds to the native structure (41). Several computational methods have been implemented from knowledge-based potentials (42–44).

Many scoring functions have been proposed to assess the quality of protein fold models (42–49). However, very few can be easily accessed as web servers by the non-specialized user. In most cases, the web servers have a reduced input flexibility (i.e. only accept models in PDB format, require chain identifiers and protein sequences, or do not accept multiple structures) and a complicated visualization of the results (i.e. do not permit to download results or do not have 3D visualization capabilities). There is a need of accessible web servers that facilitate this type of analysis (**Appendix 6.7**), or integrative platforms such as InteractoMIX (**Appendix 6.4**), which enable a combined, easy-use of different bioinformatics tools through the interface of Galaxy (50).

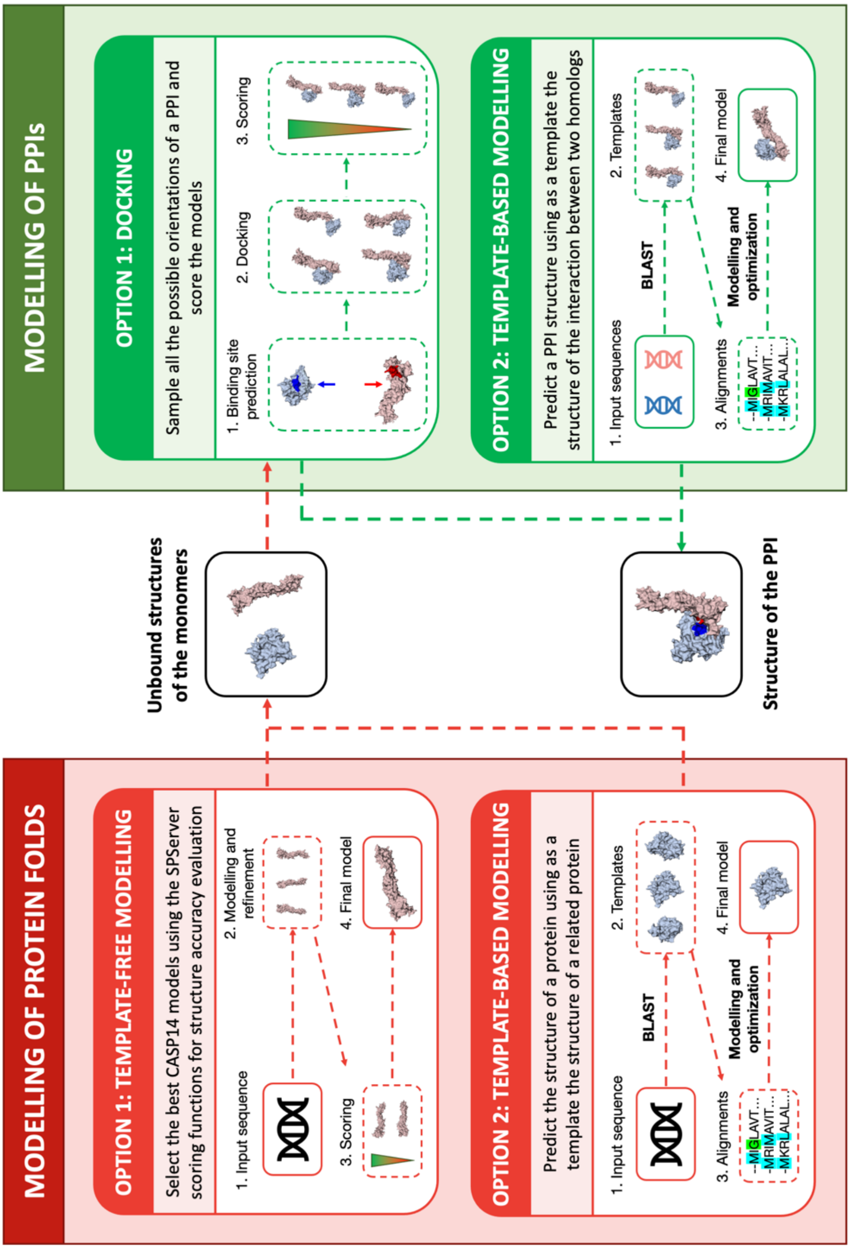

MODELLING OF PROTEIN FOLDS
OPTION 1: TEMPLATE-FREE MODELLING
Select the best CASP14 models using the SPServer scoring functions for structure accuracy evaluation
1. Input sequence
2. Modelling and refinement
3. Scoring
4. Final model
OPTION 2: TEMPLATE-BASED MODELLING
Predict the structure of a protein using as a template the structure of a related protein
1. Input sequence
BLAST
2. Templates
3. Alignments
--MIGLAVT…
-MRIMAVIT…
-MKRLALAL…
Modelling and optimization
4. Final model
Unbound structures of the monomers
Structure of the PPI
MODELLING OF PPIs
OPTION 1: DOCKING
Sample all the possible orientations of a PPI and score the models
1. Binding site prediction
2. Docking
3. Scoring
OPTION 2: TEMPLATE-BASED MODELLING
Predict a PPI structure using as a template the structure of the interaction between two homologs
1. Input sequences
BLAST
2. Templates
3. Alignments
--MIGLAVT…
-MRIMAVIT…
-MKRLALAL…
Modelling and optimization
4. Final model

Figure 2. Scheme of the computational modelling of protein folds and PPIs. On the left box, the two traditional options to model protein folds: (1) template-free modelling, and (2) template-based modelling. On the right box, the traditional options to model PPIs: (1) docking, and (2) template-based modelling.

1.1.6. Integration of PPIs

The amount of PPI data has importantly increased over the past few years. However, we are still far from having a complete, reliable interactome made of physical PPIs. The main problems to overcome to achieve a complete interactome are the following:

(1) **PPI data is spread across multiple databases and publications:** There are three different types of databases where PPI data can be collected. First, there are primary interaction databases, which annotate experimental interactions directly from the source publications (51–53). They provide services such as curating metadata or creating standards and ontologies. These databases are coordinating their efforts through the IMEx consortium (54). Second, there are databases of predicted interactions, where the data is provided by computational methods (55,56). Third, there are databases that integrate primary and/or predicted interactions into a unique database containing physical and/or functional interactions (57).

(2) **The nomenclature of the proteins is different:** The protein identifiers of the different PPI databases are not unique. They are usually different from database to database. This complicates to have a uniform system to identify proteins.

(3) **There are different formats to store PPIs:** There are still different types of formats to store PPIs, although standard file

formats such as PSI-MI and BIOPAX are improving the access to this type of data.

(4) **The reliability of the PPI data varies on each experiment:** All the PPI detection methods are subjected to some degree of error. Some experiments will be more prone to induce errors than others. For example, high-throughput methods will be more likely to have false positives, whereas low-throughput methods will have more false negatives. When integrating PPI data to create an interactome, the reliability of these methods has to be taken into account.

(5) **Tissue-specific interactions are not generally considered:** Some PPIs are tissue-specific, because the proteins are expressed in specific tissues in given circumstances. Therefore, some PPIs that are reflected in generals PPI networks might not happen depending on the tissue.

Several resources and databases have been developed during the recent years to integrate PPI data and other types of omics data (**Table 2**). These resources gather PPIs from different species by parsing multiple sources of data. The unification criteria of the proteins differs depending on the database: most of them unify proteins by Uniprot Accession ID (APID (58), ConsensusPathDB (59), InBioMap (60), IID (61)), whereas some others opt to use more complex unifications, such as BIANA (62), which gives the user flexibility to decide the criteria of unification.

The integration resources usually employ their own methods to assess the reliability of the interactions. For example, ConsensusPathDB (59) uses an algorithm called IntScore (63) based on the topology of the network. InBioMap (60) uses a score that combines network topology with the number of publications in which the PPI has been reported. In contrast, HIPPIE (64) assesses the interactions by combining multiple criteria such as number of publications, number and quality of experiments and number of interacting orthologs.

Some of these resources have started to incorporate tissue-specific interactions by integrating data from RNA-seq datasets. It is the case HIPPIE (64), which was pioneer in incorporating tissue-specific information from 53 human tissues from GTEx (65). They considered that a gene is expressing a protein in a given tissue if the median expression over samples exceeds the RPKM (Reads Per Kilobase Million) threshold of 1 (66).

Table 2. Resources and databases of PPI integration.

Integration resource	Sources	Species	Unification	Assessment of reliability	Tissue-specificity
APID (58)	BioGRID, DIP, HPRD, IntAct, MINT, BioPlex	All available	Uniprot accession ID	Based on the detection method, source, and publications	No
BIANA (46)	Defined by the user. By default: BioGRID, InnateDB, IntAct	All available	Defined by the user. By default: Gene ID, or same sequence from same species	Defined by the user	Yes, based on GTEx RNA-seq data

Consensus PathDB (59)	BIND, InnateDB, MatrixDB, PDZBase, PhosphoPOINT, PhosphoSitePlus, PINdb	Human, yeast, mouse	Uniprot accession ID	IntScore: score based on network topology	No
HIPPIE (64)	BIND, BioGRID, DIP, HPRD, IntAct, MIPS	Human	Gene symbol, or Entrez Gene ID, or Uniprot accession ID	Score based on the number of publications, number and quality of experiments, and number of interacting orthologs	Yes, based on GTEx RNA-seq data
InBioMap (60)	BIND, BioGRID, DIP, IntAct, MatrixDB	Human	Uniprot accession ID	Score based on the number of publications and network topology	No
iRefIndex (67)	BIND, BioGRID, CORUM, DIP, HPRD, InnateDB, IntAct, MatrixDB, MPact, MPIDB, MPPI, VirHostnet	All available	SEquence Global Unique IDentifiers	Score based on the number of publications	No
IID (61)	BioGRID, DIP, HPRD, I2D, InnateDB, IntAct, MINT	Human, mouse, pig, rabbit, rat, sheep, turkey, worm, yeast	Uniprot accession ID	Filter interactions based on types of detection methods	Yes, based on NCBI GEO
STRING (57)	BIND, BioGRID, DIP, HPRD, IntAct, MINT, PID	All available	Protein sequence	Confidence score based on pathway knowledge and orthologs data	No

1.2. Network biology: Apply network science to study the interactome

To fully understand the functioning of the human organism it is required a holistic and inclusive view that relies on a system-based approach studying the relationships between all the biomolecules. Given the interrelated nature of cellular processes, network analysis is particularly suited to study their molecular mechanisms. Applied to biological systems, the ***nodes*** of the network can represent proteins, genes or even diseases, while the ***edges*** are the relationships between these biological entities. There are different types of network representations that are used to study the human organism (**Table 3**): (i) ***protein interaction networks***: where nodes are proteins and edges are physical interactions; (ii) ***metabolic networks***: where nodes are metabolites and proteins, while edges are metabolic reactions; (iii) ***gene regulatory networks***: where nodes are transcription factors and genes, while edges are regulatory interactions; and (iv) ***disease networks***: where nodes are diseases, while edges represent different types of relationships such as shared genes (68). ***Network biology*** is the discipline that seeks to accurately represent biological networks and analyze them to understand the behavior of a biological system. In the following subsections, I will review the properties of networks and how they help us to better understand different biological systems.

Table 3. Properties of different types of biological networks.

Network	Type of nodes	Type of edges	Direction
Protein interaction network	Proteins	Protein-protein interactions	No
Metabolic network	Metabolites, proteins	Metabolic reactions	Yes
Gene regulatory network	Transcription factors, genes	Regulatory interactions	Yes
Disease network	Diseases	Disease-disease pairs sharing a property	No

1.2.1. Definition of network

Mathematically, a **network** or graph (G) can be defined as a pair $G = (V, E)$, where V is a set of elements called **nodes** (or vertices), and E is a set of paired nodes, whose elements are called **edges** (or links). More informally, a network can be defined as a structure containing a set of elements in which some of them are related. The elements composing the network are the nodes, while the connections between related nodes are the edges. Applied to biological systems, the nodes can represent proteins, genes or even diseases, while edges are the relationships between these biological entities.

1.2.2. Types of networks depending on the properties of the edges

Depending on the direction of the relationship that the edges represent, networks can be directed or undirected (see **Figure 3**). A network is **directed** when the interactions have a specific direction that goes from a source to a target and that is represented by an arrow. In contrast, a network is **undirected** when the interactions do not have a specific direction and are represented by lines. For example, PPI networks are mostly undirected, because their edges represent interactions between proteins, which generally do not follow a specific direction. In contrast, metabolic networks are directed, because the edges represent metabolic reactions that start with substrates and end with products. Gene regulatory networks are also directed, because they represent how the expression of genes regulates the expression of other genes.

Networks can also be weighted or unweighted depending on whether the edges carry an additional weigh or not. **Unweighted** networks have edges that are placed if a certain threshold of evidence for the relationship is reached, whereas **weighted** networks have edges in which a specific property of the relationship is indicated by the weight. For example, gene co-expression networks are networks where the nodes are genes that are connected by the correlation of their expression. These networks can be weighted, thus showing the correlation between the expression of two genes in the edge weight; or unweighted, only showing the edges that meet a certain threshold of correlation.

1.2.3. Types of nodes depending on the degree

The main property of a node is the **degree** (k), which is the number of edges directly linked to that node. When the degree of a node exceeds the average, the node is considered a **hub**. In contrast, a node of degree 1 (with only one edge) is considered a **leaf**, and a node without edges is an **isolated node**. In directed networks, we can distinguish between **in-degree** (number of incoming edges to a node) and **out-degree** (number of outcoming edges from a node). Also in directed networks, a **source** node has an in-degree of zero (relations only come out from this node and not come in), whereas a **sink** node has an out-degree of zero (relations only come into this node and not come out). Examples of the different types of nodes are represented in **Figure 3**.

The degree of a node in a biological network gives us highly valuable information about how this node behaves in the network. For example, in a PPI network context, the hubs tend to be involved in crucial functions for the survival of the cell. In the same way, less connected proteins tend to be less essential, and therefore more susceptible to changes through the evolution of the species (69,70).

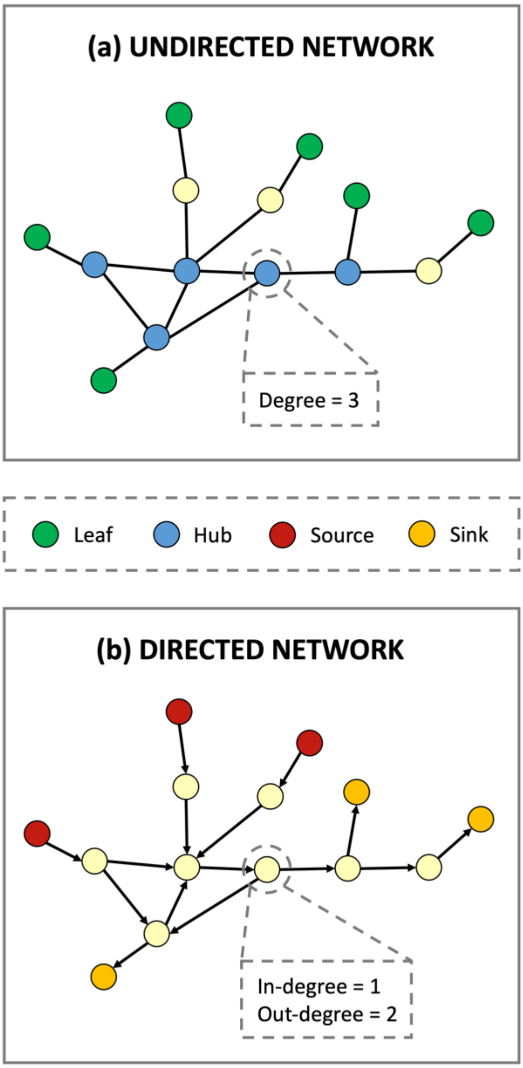


Figure 3. Types of nodes depending on the degree associated to undirected (a) and directed (b) networks. (a) Undirected network, containing edges without specific direction represented by lines. I highlight 5 hub nodes (in blue) which exceed the average degree, and 6 leave nodes (in green) with only one edge. As an example, I highlight one node of degree 3. (b) Directed network, containing edges with specific direction represented by arrows. I highlight 3 source nodes (in red) with in-degree of 0, and 3 sink nodes (in orange) with out-degree of 0. As an example, I highlight one node of in-degree 2 and out-degree 1.

1.2.4. Types of networks depending on the degree distribution

The **degree distribution** is the frequencies of degree values of the nodes of a network. In other words, it is the probability that a randomly selected node has a specific degree value. The degree distribution is a very characteristic metric, which allows to distinguish different types of networks (71). For example, in random networks, the degrees of the nodes follow a Poisson distribution, meaning that most nodes have a similar number of edges, approximately the same as the network's average degree (see **Figure 4-a**). In contrast, many biological networks have a degree distribution that follows a **power law distribution**:

$$P(k) \sim k^{-\gamma} \qquad \textit{Eq. 1}$$

where k is the degree, γ is the degree exponent and ~ indicates "proportional to". The value of γ is represented by many properties of the system. These types of networks are called **scale-free networks** and are characterized by not being uniform: most of their nodes have only a few edges, and very few nodes have a very large number of edges (hubs) (see **Figure 4-b**). Among the examples of scale-free biological networks, we find PPI networks, cellular networks, and genetic regulatory networks.

From the degree distribution of the network, we can also obtain the **average degree** ($\langle k \rangle$), which is calculated in undirected networks as:

$$\langle k \rangle = \frac{2E}{V} \qquad \textit{Eq. 2}$$

where V is the number of nodes and E is the number of edges of the network. In directed networks, it is calculated using the same formula (*Eq. 2*) but without multiplying by 2.

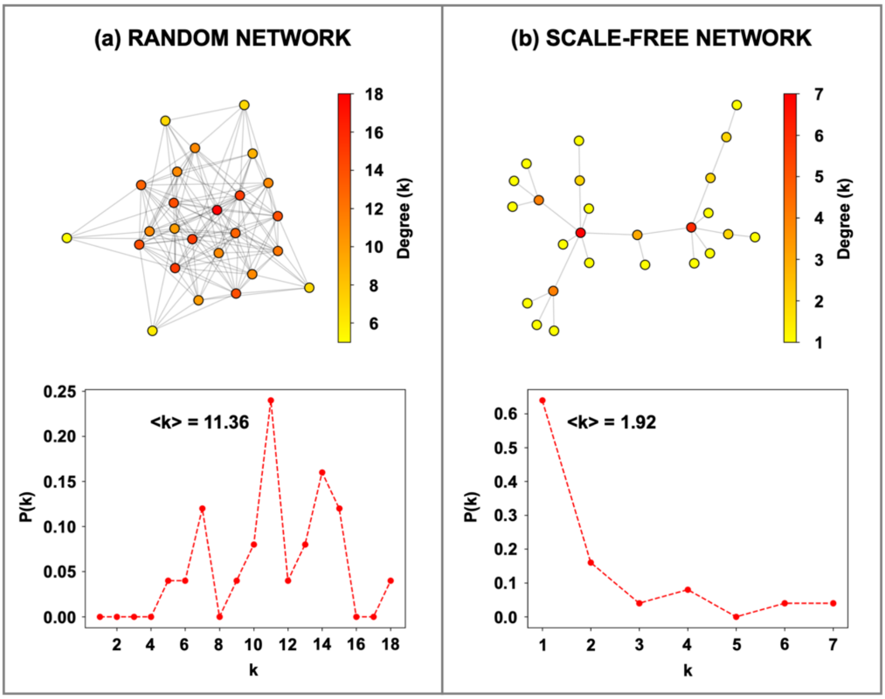


Figure 4. Examples of a random network (a) and a scale-free network (b) and their corresponding degree distributions. (a) Random network following a Poisson distribution, where the majority of nodes have a similar number of edges. The network has 25 nodes, 142 edges and an average degree of 11.36. (b) Scale-free network following a pawer law distribution, where the majority of nodes have few edges and very few nodes have a large number of edges (hubs). The network has 25 nodes, 24 edges and an average degree of 1.92.

1.2.5. Paths: distance between nodes

A **path** within a network is a connection between two nodes that follows a certain number of edges. The length of the path is quantified by the number of edges involved in the path. Mathematically, we can define a path (P) in an undirected graph as a sequence of nodes (v):

$$P = (v_1, \ldots, v_n) \qquad \textit{Eq. 3}$$

such that v_i is adjacent to v_{i+1} for $1 \leq i < n$. The path P is called a path of length $n - 1$ from v_1 to v_n.

The **shortest path** between two nodes is the path that contains the minimum number of edges to connect them (see **Figure 5** for an example). The mean shortest path length among all the nodes of the network is the **characteristic path length**. It can be calculated as:

$$a = \sum_{s,t \in V} \frac{d(s,t)}{n(n-1)} \qquad \textit{Eq. 4}$$

Where V is the set of nodes in the whole network of number n, and $d(s,t)$ is the shortest path from nodes s and t. The **network diameter** (d_{max}) is the longest shortest path between any pair of nodes of the network. If there is a path between every node in the network, the network is called **connected**. If not, we call the largest connected set of nodes as **largest connected component**.

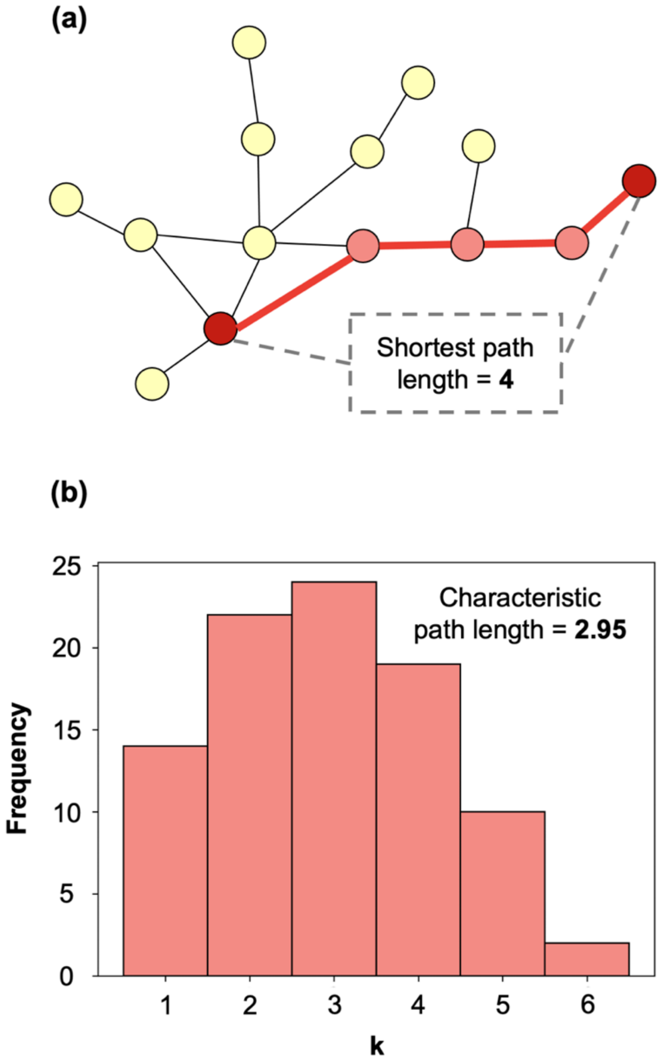


Figure 5. Examples of shortest path and shortest path distribution. (a) Example of network, where the shortest path between the dark red nodes involves the light red nodes and has a length of 4. (b) Shortest path distribution of the previous network, with a characteristic path length of 2.95.

The path is a key metric in network science, because it provides valuable information about the relationship between a given node and the rest of the nodes of the network. For this reason, there are many methods applied to the study of biological networks that are based on the calculation of shortest paths (72). In PPI networks, shortest path methods are used to assign directions to the edges (73,74). These methods focus on finding solutions that maximize the number of source-target pairs that admit a shortest path between them. Another common approach in PPI networks is to use methods based on the shortest path to find disease-associated proteins. This is because disease-associated proteins tend to cluster in the same neighborhoods of the network forming the so-called disease modules (75–77). Thus, the lower the shortest path is with known disease-associated proteins, the more likely it is of being associated with the disease. For example, some of the algorithms developed in GUILD software (78) employ shortest path methods to prioritize new disease-associated proteins. NetShort algorithm assigns higher scores to the proteins of the network if the shortest path between the protein and a disease-associated protein includes other disease-associated proteins. NetScore algorithm scores proteins by calculating how fast a message travels from the protein to the disease-associated proteins through the multiple shortest paths. The shortest path is also used Menche et al. (77) as part of a network-based metric to calculate the separation between the proteins associated with different diseases. The same measure was slightly modified in Guney et al. (79) to evaluate the proximity between the protein targets of a drug and the proteins associated with a disease.

1.2.6. Centrality measures

Centrality measures give information about the importance of the nodes and edges by assigning them scores. In biological networks, centrality measures are used to identify the nodes with crucial roles in biological functions. There are different types of measures accounting for network centralities that give different types of importance to the highest scored nodes. However, different centrality measures tend to be correlated with each other, and hubs tend to have high centrality. The most important centrality measures are the following (see **Figure 6** for an example):

(1) **Degree centrality:** Refers to the number of connections of a node. It is defined as:

$$C_D(v) = \deg(v) \qquad \textit{Eq. 5}$$

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 $n - 1$ where n is the number of nodes in the network of interest. By definition, the nodes with higher degree centrality tend to be hubs, nodes with a number of edges that exceeds the average. As said before, hubs tend to be related with essential functions, which are less susceptible to changes through the evolution of the species (69,70).

(2) Closeness centrality: Refers to how close a node is to the rest of the nodes of the network by measuring the shortest-path distance between them. It can be defined as:

$$C(u) = \frac{n-1}{\sum_{v=1}^{n-1} d(v,u)} \qquad \textit{Eq. 6}$$

Where $d(v,u)$ is the shortest-path distance between the nodes v and u, and n is the total number of nodes in the network. The nodes with larger closeness centrality have shorter average propagation length of information to the others. Therefore, closeness centrality shows how efficiently the nodes of the network transfer information with each other (72,80). For example, a study in an *Escherichia coli* metabolic network showed how 8 of the top 10 metabolites with higher closeness centrality are part of the glycolysis and citrate acid cycle pathway (the most central metabolic pathway) (81).

(3) Betweenness centrality: Refers to how often the node is present within the group of shortest paths of the network. The betweenness centrality of a node v can be calculated as:

$$c_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \qquad \textit{Eq. 7}$$

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 measures the number of times that the node of interest appears among the shortest paths between all pairs of nodes. Betweenness centrality can be very effective to predict “bridge” or “link” nodes that connect different neighborhoods or modules of a network (72,80). For example, a study in a *Saccharomyces cerevisiae* PPI network reported that proteins with high betweenness centrality and low degree are key for the modularization of the network (82). Also, in the study of Piñero et al. (83) it is reported how disease-associated proteins (especially if they are cancer related) have a higher betweenness centrality than the rest in PPI networks.

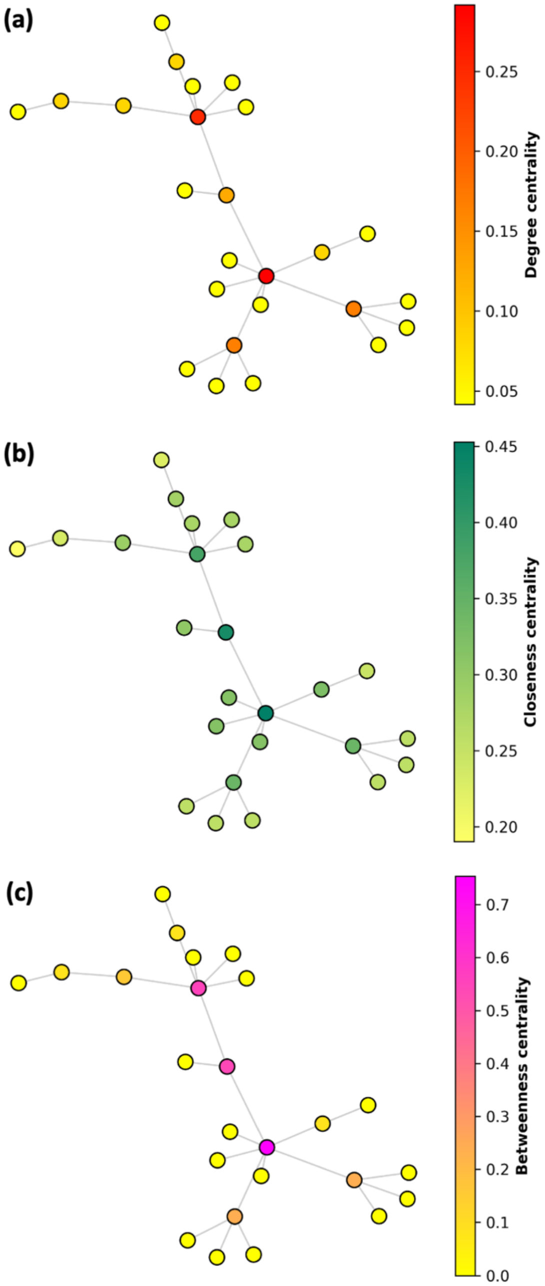


Figure 6. (a) Degree, (b) closeness and (c) betweenness centrality of the nodes of a scale-free network. The three centrality measures have been normalized to a scale from 1 (the highest) to 0 (the lowest).

1.2.7. Network modules and clustering measures

In many types of biological networks, it is common that the nodes that have similar roles or functions interact with each other, forming clusters known as modules or communities. For example, in PPI networks, it is possible to identify modules of:

(1) Proteins executing the same or similar functions (84,85).
(2) Proteins forming complexes (85,86).
(3) Proteins associated to the development of a disease (75–77).

Therefore, the identification modules in a network can be key to understand better the biological implications of the module members.

There are several metrics and algorithms that tell us about the level of clustering of nodes. A clear example is the **clustering coefficient**, which describes the probability that two nodes that are connected to another node are directly connected between themselves (forming a triangle). The **local clustering coefficient** of a node measures the number of possible triangles present in its neighborhood (**Figure 7**), and can be calculated as:

$$C_i(v) = \frac{2L_i}{k_i(k_i - 1)} \qquad \textit{Eq. 8}$$

Where i is the node of degree k_i, and L_i denotes the number of connections between the neighbors of node i. By calculating the average local clustering coefficient of all the nodes of the network we obtain the **average clustering coefficient** of the network, which describes the degree of clustering of the network.

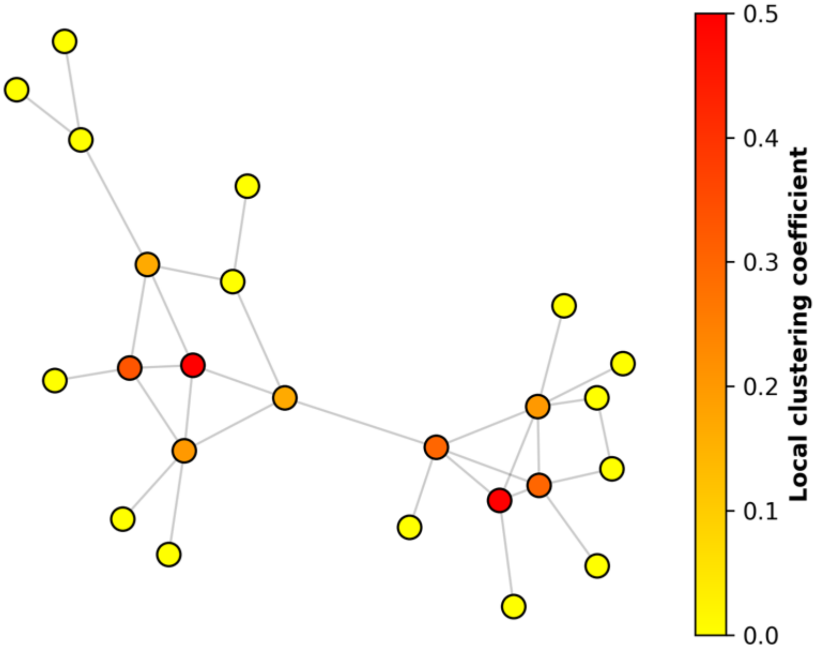


Figure 7. Local clustering coefficient of the nodes of a scale-free network. The nodes that form a higher number of triangles with other nodes obtain a higher local clustering coefficient. The average clustering coefficient of the network is 0.11.

There are also many different algorithms to directly identify modules in the network. These algorithms are usually classified into two different categories. The first category includes algorithms that use prior knowledge on nodes that have some kind of relationship which is common in certain modules of the network. These nodes are known as **seeds**. These algorithms identify the neighborhood of nodes that are topologically closer to the seeds. The second category includes methods that identify the modules “ab initio”, using community structure detecting algorithms. These methods analyze the topology of the network and identify the neighborhoods with properties that are common in modules, such as high within-edge density of connections. Module identification algorithms are extensively used to identify disease modules, which are modules composed by proteins whose function is perturbated by the same disease and which tend to be in the same neighborhood of the interactome. These algorithms are explained in more detail in **Chapter 1.3.1.2 “Identification of disease modules”**.

1.2.8. Network robustness

Robustness (or resilience) is the ability of the network to respond to internal mechanistic failures or external conditions while maintaining a relatively normal behavior (71). Robustness depends on several factors such as the topology of the network and the functional and dynamic changes that the perturbations involve.

The topology of the network is key to define the robustness that the network will have (71). Depending on the type of topology, the network will be more or less resilient to changes. The simplest definition of network robustness is to test the network topological changes by removing a proportion of nodes or edges. When the removal proportion exceeds a critical value, the network is disintegrated into smaller and disconnected components (72). For example, a random network is less robust, because if a critical number of nodes is removed, the network becomes a group of tiny, isolated nodes. In contrast, scale-free networks are extremely resilient to changes: even if 80% of randomly selected nodes are removed, the remaining 20% still form a fully connected network with paths connecting any two nodes. This is explained by the degree distribution of these networks. Scale-free networks have a high number of nodes with a small degree, and a small number of nodes with high degree. Therefore, random removal of nodes will affect mainly nodes with a small degree, the absence which will not affect the integrity of the network. In contrast, in random networks, the majority of nodes have a similar number of edges. This property makes them much more vulnerable to the removal of nodes (71).

In general, there are two types of network failures: random failures (e.g., random removal of nodes) and intentional attacks (e.g., removal of network hubs). Random networks are very vulnerable to both type of failures. In contrast, scale-free networks are robust against random failures but especially vulnerable to intentional attacks (87).

Robustness is a common property in many complex systems, and biological systems are not the exception. This robustness is achieved through different mechanisms such as feedback (88), redundancy (89), or functional modularity (90). Robustness protects the biological system from failures such as gene mutations that affect protein–protein interactions (91,92) or removal of enzymes in metabolic networks (93). In the context of PPI networks, Guney et al. (94) studied the robustness associated to the disease modules of complex diseases. They showed that even when randomly removing interactions, some diseases conserved the interconnectivity of the core proteins associated with the diseases (see **Figure 8** for a graphic explanation). This is related with the high betweenness centrality of disease-associated genes of some disease classes (83,95).

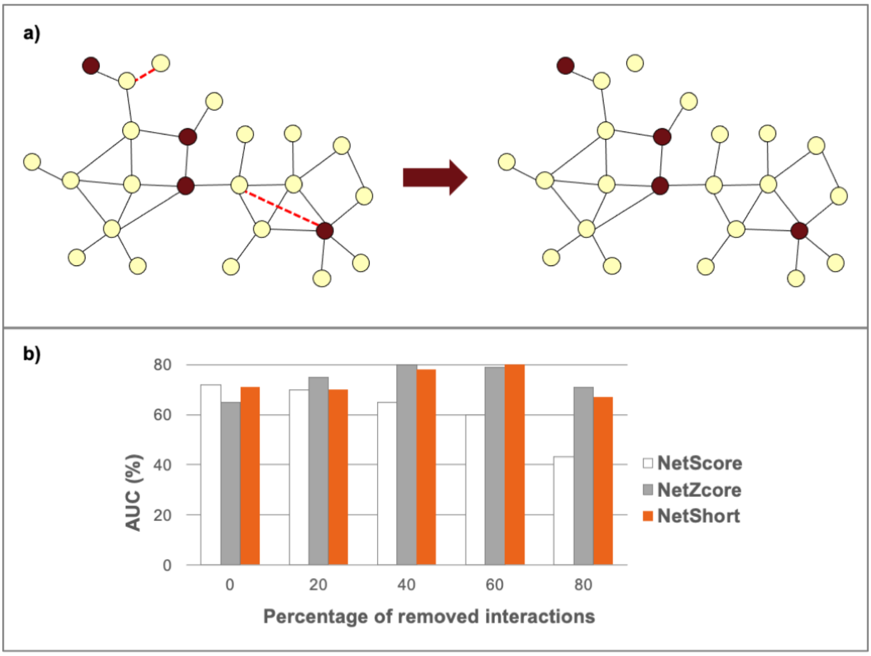


Figure 8. Impact of randomly removing edges of the interactome on the prediction of disease-associated genes. (a) Example of the removal of an edge in the protein interaction network. (b) Effect of the edge removal on the prediction of disease-associated genes (by the algorithms NetScore, NetZcore and NetShort) in the form of the Area Under the Receiver Operating Curve (AUROC). Surprisingly, the AUROC increases as the number of edges removed increases, which shows the ability of disease-associated genes to preserve their connectivity despite losing interactions. This is aligned with the high betweenness centrality of disease-associated genes of some diseases (83,95). The figure is based on the study of Guney et al. (94) and Figure 1 from Aguirre-Plans et al. (96).

1.3. Network medicine: towards a better understanding of human disease complexity

At molecular level, the inner workings of cells can be understood as an interplay of interactions between biomolecules underpinned by interconnected networks, forming the interactome. Accordingly, diseases can be considered as perturbations in these networks, influenced by genetic and/or environmental factors that affect the normal functioning of the organism. Although our knowledge of human biological networks is still far from being complete, network-based methods are valid approaches to explore the molecular mechanisms of diseases and comorbidities (i.e. two diseases or more that are more likely to co-occur in the same patient). ***Network medicine*** is the discipline of network biology that seeks to accurately represent biological networks and analyze them to understand diseases and find the right treatments. In the following subsections, I will review the advances of network medicine towards a better understanding of human disease complexity.

1.3.1. Disease-gene associations

1.3.1.1. The genetic basis of diseases

The **genome** is the genetic material within the organism. It is comprised by genes, DNA regions that synthetize and regulate the proteins of the interactome. The genome is one of the main factors that gives rise to the **phenome**, the observable structure, function, and behavior of an organism.

The genetic information encoded in the genome does not remain unchanged in an organism. On the contrary, it is a constant target for new or inherited alterations in the sequence of the DNA that we call **genetic variants**. Depending on the frequency of occurrence of the genetic variants in the population, we can classify them as common, when the frequency is equal or higher than 1% in the population; or as rare, when the frequency is lower than 1%. Also, depending on the number of nucleotides involved in the alteration, they can be classified as single-nucleotide variants or polymorphisms, when only a single nucleotide is involved; or as structural variants, when involve a large number of nucleotides; there are also small-scale variants such as insertions or deletions, involving a few nucleotides (97,98). If the genetic variants are detectable within germ cells (and therefore they can be inherited), they are called germline variants, while if they occur in any cell other than a germ cell (therefore not being inherited), they are called somatic variants. Although some of these alterations might not produce an observable effect, some others might affect molecular functions which alter the phenotype and end up being beneficial or counterproductive for the organism. The variants that are counterproductive or deleterious to the organism are called pathogenic variants and they might be one of the factors or the main cause of a disease (68).

Diseases are perturbations in the structural or functional parts of the body that alter homeostatic processes. They are characterized by specific signs, symptoms, and biochemical patterns, i.e., a specific phenotype. The diseases that are caused or influenced by genetic variants and are known as **genetic diseases**. Depending on the genetic architecture that is causing the phenotype of the disease, genetic diseases are classified in the following categories (68):

- **Single-gene disorders:** Also called monogenic or Mendelian diseases, they are diseases caused by germline variants that only affect an individual gene. These diseases tend to be strongly linked to a genetic component rather than to other factors such as the environment.

- **Oligogenic disorders:** They are diseases caused or influenced by variants in only a few genes.

- **Mitochondrial disorders:** They are caused by alterations in the mitochondrial DNA or the nuclear DNA encoding mitochondrial proteins, provoking a dysfunction in the mitochondrial respiratory chain. As they affect the mitochondrial DNA, they have different patterns of inheritance to the rest of germline variants.

- **Chromosomal abnormalities:** They are diseases caused by chromosomal abnormalities involving changes in chromosome number or large physical changes in chromosome structure.

- **Complex diseases:** They are diseases caused by the interplay of multiple genetic variants in different genes with small additive effects, apart from the influence of other factors such as the environment or epigenetics. The results of genomic studies and predictions from statistical models (99,100) indicate that the genetic architecture of complex diseases is composed by two types of genetic variants: (i) a large number of common variants of small effect distributed throughout the genome and (ii) a smaller contribution from rare variants with moderate effect in genes known to cause the familial form of the disease.

- **Cancer:** Cancer is characterized by the interplay of multiple genetic variants that causes an uncontrolled growth of cells and tissues. While cancer usually arises due to the accumulation of variants in somatic cells, it is also influenced by germline variants that confer susceptibility to certain types of cancer.

1.3.1.2. Disease-gene associations: Identification methods and databases

The genes which, due to a genetic variant, are involved in a disease are known as **disease-gene associations**. There are several methods to identify disease-gene associations. Traditionally, genetic diseases were studied with a direct analysis of a candidate gene. In the 1980's, genetic linkage maps, based on analyzing the frequencies of recombination between specific gene markers, were used to check if certain variants showed similar segregation to a particular disease (101). These methods were successful in identifying genetic variants associated with single-gene disorders, but insufficient in the case of complex diseases and cancer due to their polygenic nature. In this context, the first Genome-Wide Association Studies (GWAS) emerged, based on performing tests for the association of single-nucleotide variants with diseases in large populations of individuals (102). These large-scale studies are possible thanks to the rapid advances in Next Generation Sequencing (NGS) technologies, which allow to sequence the exomes of hundreds and thousands of patients across the world, searching for disease-gene associations.

Different resources and databases compile and curate the information on disease-gene associations that is spread in the scientific literature (**Table 4**). The most known and widely used resource is the Online Mendelian Inheritance in Man (OMIM) database (103), which was developed in 1966, and compiles associations between human genes and Mendelian diseases. Apart from OMIM, there are many other databases that compile and manually curate disease-gene associations from the scientific literature (104–109). Still, it is a tedious task to gather all the information spread over these resources, and this is why a database such as DisGeNET (110) emerges, integrating them in a unique repository.

Table 4. List of disease-gene association databases.

Database	Description	Drug-target association sources	URL
Cancer Genome Interpreter (CGI) (104)	Repository specialized on predicting and compiling cancer driver genes potentially causing oncogenic alterations	Scientific literature (compilation of different datasets by manual curation)	cancergenome interpreter.org
The Clinical Genome Resource (ClinGen) (105)	Repository that defines the clinical relevance of genes and variants for use in precision medicine and research	Scientific literature (by manual curation)	clinicalgenome.org
Comparative Toxigenomics Database (CTD) (106)	Database of associations between chemicals, genes and diseases across different species	Scientific literature (by manual curation)	ctdbase.org
DisGeNET (110)	Database that integrates disease-gene associations from expert curated repositories and text mining resources	Other databases (CGI, ClinGen, CTD, Genomics England, Orphanet, PsyGeNET, UniProt); text mining (LHGDN, BeFree)	disgenet.org
Online Mendelian Inheritance in Man (OMIM) (103)	Widely used database of disease-gene associations, reporting links between human genes and Mendelian disorders	Scientific literature (by manual curation)	omim.org
Orphanet (107)	The reference portal for information on rare diseases and orphan drugs	Scientific literature (by manual curation)	orpha.net
PsyGeNET (108)	Resource containing genes associated to psychiatric diseases	Scientific literature (by text mining and manual curation)	psygenet.org
UniProt (109)	Database of protein information from different species. It also includes knowledge on disease-gene associations	Scientific literature; other databases (OMIM)	uniprot.org

1.3.2. Identification of disease modules to explore disease mechanisms

1.3.2.1. Definition of disease modules

The distribution of nodes and edges in the human interactome is not homogeneous. There are regions where nodes are more densely connected, forming modular structures (111). As early as 1999, Hartwell already proposed that cellular functions could be accomplished by “modules” of different types of molecules (90). The modular organization of biological networks offers multiples advantages to the system in terms of adaptability: it can increase the robustness of the network, as it limits the number of components of the system affected by a perturbation, and it can be easily rewired to adapt to new conditions (112,113).

The interest in network modules increased after several studies which showed that proteins associated with similar diseases tend to interact directly with each other (114,115) and cluster in the same neighborhoods (regions) of the interactome (75). In a pioneer work in 2007 (75), Goh and coworkers found that proteins encoded by genes associated with similar diseases are more likely to interact with each other, indicating the existence of specific functional modules within the interactome. Building on this observation, they created the first human disease network, called **human diseasome**, by connecting the diseases with shared genetic component, extracting the disease-gene associations from the Online Mendelian Inheritance in Man (OMIM) database (103). This concept was further extended to complex and environmental diseases (116) and by Park et al. (117) with the objective of finding comorbidity patterns within the overlapping functional modules. They analyzed the US Medicare database to find co-occurring diseases. They found a high correlation between comorbidity and the number of shared genes from the diseasome.

These studies led to the hypothesis of the existence of **disease modules**: genes involved in the same disease that cluster together in the interactome. This hypothesis was explained in detail in the review of Barabási et al. (76). They defined three types of modules within the interactome (**Figure 9**): (i) the ***topological module***, a locally dense neighborhood which can be identified by clustering algorithms; (ii) the ***functional module***, the neighborhood of nodes with similar or related function; and (iii) the ***disease module***, the neighborhood of nodes that contribute to cellular functions whose disruption results in a particular disease. Disease modules are not expected to be identical to functional or topological modules, but rather to overlap to some extent with them. The identification of the so-called disease modules is key to achieve a comprehensive molecular understanding of diseases.

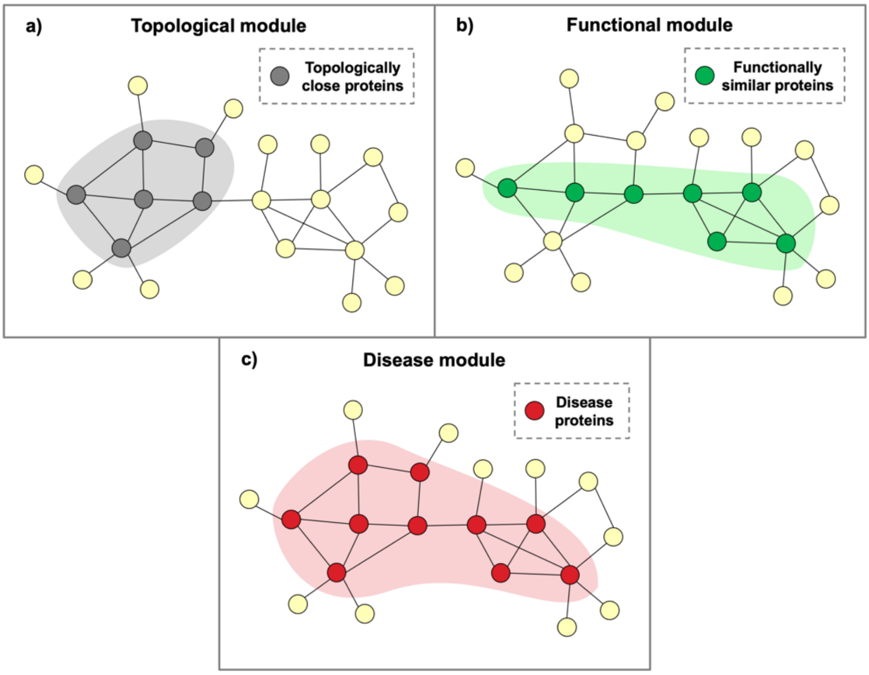


Figure 9. Schematic description of three types of modules that can be found in the interactome, which were defined in the review of Barabási et al. (76). (a) Topological module, containing proteins with a higher tendency to interact with each other rather than with proteins outside the module. (b) Functional module, containing a group of proteins with a higher tendency to develop similar or related biological functions. (c) Disease module, containing a group of proteins whose perturbation or disruption results in the development of a particular disease. Figure adapted from Figure 2 of Barabási et al. (76).

1.3.2.2. Identification of disease modules

There is a wide range of approaches proposed to predict or identify disease modules that can be classified into two categories (118):

(1) **Identification based on prior knowledge:** This category includes methods that use prior knowledge on disease-associated genes (also known as ***seed genes***). These methods identify the neighborhood of proteins that are topologically closer to the proteins encoded by the seed genes (78,119). We can classify these methods into three subcategories (**Figure 10**) (3,76):

(a) **Network neighbor methods:** These methods (also called linkage methods) assume that the proteins that directly interact with other proteins associated with a certain disease are more prone to be associated with the same disease. For example, in Oti et al. (120), the authors find new disease-gene associations by identifying the neighbor proteins of disease-associated gene products and checking that the chromosomal location of their corresponding genes was falling within the loci of the same disease. This method was 10 times more likely to predict true disease-gene associations than using only genomic information.

(b) **Topological community finding methods:** These methods (also known as graph partitioning methods) first identify if the seed proteins associated to the disease cluster together forming a subnetwork, and if the subnetwork is statistically significant (i.e., it is less likely than random to find this given number of connected nodes). Then, topological community finding tools (121–123) are applied to identify additional protein candidates that could be added in the subnetwork and that are topologically and functionally related with the rest of the module proteins. For example, DIAMOnD (119) employs an iterative algorithm to calculate the significance of the interactions of the proteins in the neighborhood of the seeds (i.e. if the number of interactions is higher than a random expectation). Another example is the Seed Connector Algorithm (124), which iteratively includes proteins to a pool of seeds associated with the disease if the size of their largest connected component is increased. This algorithm goes on until none of the neighbor proteins, when added, increases the coverage of seed proteins in the largest connected component of proteins associated with the disease.

(c) **Diffusion-based methods:** These methods consist in releasing signals (known as “random walkers”) from the seed nodes to the rest of the nodes of the network. The nodes that are closer to the seeds are more visited by the signals and therefore are scored higher by these methods. For example, the algorithms from GUILD software (78) are message-passing algorithms that transmit a signal from the 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.

(2) **“Ab-initio” identification:** This category includes methods that identify the modules “*ab initio*”, using community structure detecting algorithms. These methods, based on the topology of the network, identify neighborhoods of proteins with high within-edge density of connections such as algorithms based on the maximal clique enumeration problem (118,125).

Even though this is an active area of research, the identification of disease modules with high accuracy remains problematic. In a community effort to advance in this area, the Synapse platform launched in 2018 a DREAM challenge focused on the blind prediction of disease modules from different types of networks (126). In this challenge, different types of approaches were among the top performers (diffusion state distance, kernel clustering, modularity optimization, random-walk-based and local methods), suggesting that not a single approach was superior to the rest, including methods from both categories as described above. One of the top performers was diffusion state distance method, which is an improved measure of the network proximity between pairs of nodes and shows the importance of considering the full topology of the network instead of only local neighborhood (127,128).

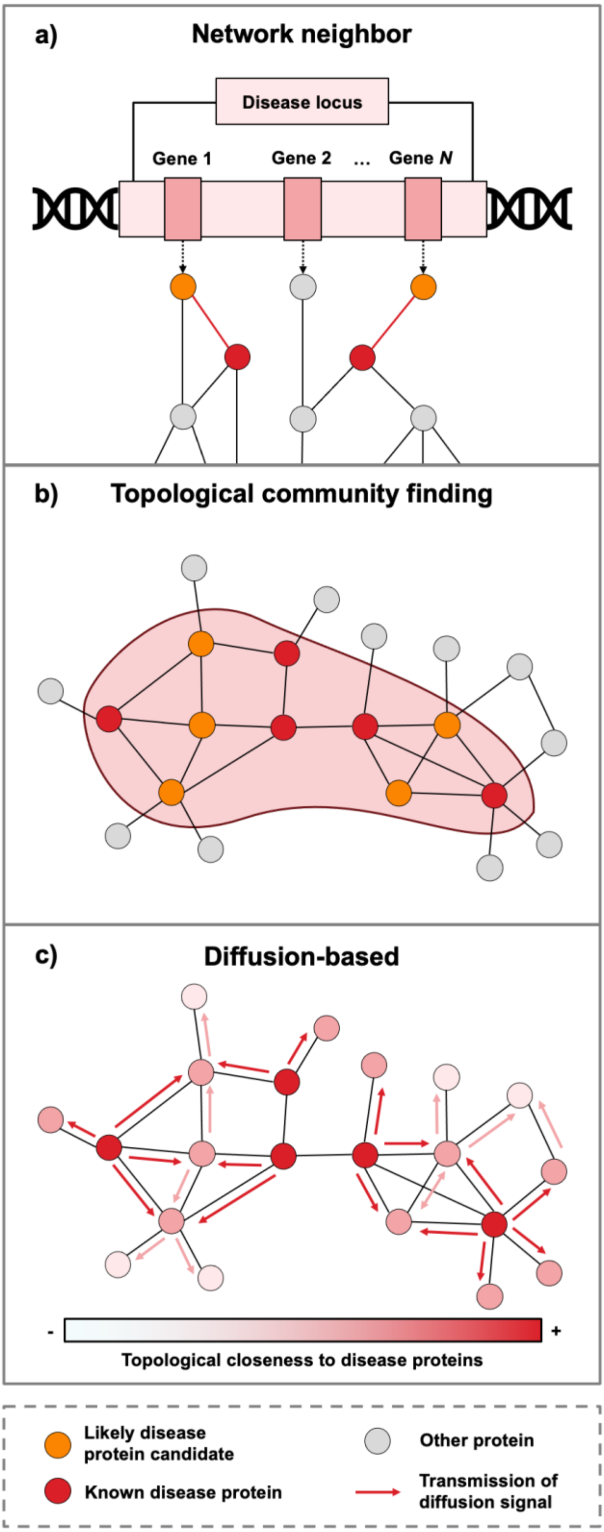

a)
Network neighbor
Disease locus
Gene 1
Gene 2
...
Gene N
b)
Topological community finding
c)
Diffusion-based
-
+
Topological closeness to disease proteins
Likely disease protein candidate
Other protein
Known disease protein
Transmission of diffusion signal

Figure 10. Schematic description of three types of disease module identification methods based on prior knowledge. (a) Network neighbor methods, based on finding genes in the locus of the disease whose products interact with known disease-associated proteins. (b) Topological community finding methods, based on finding a subnetwork of disease-associated proteins, and applying network-based algorithms to include new disease-associated proteins which are likely to be forming a disease module. (c) Diffusion-based methods, which consist in applying message-passing algorithms from the disease-associated proteins to the rest of the proteins of the network, simulating the perturbation caused by the disease. The proteins of the network are given a score based on how fast the signal reaches them. Figure adapted from Figure 4 of Barabási et al. (76).

1.3.2.3. Cases of disease module identification that explain disease mechanisms

The recent advances in network medicine are starting to offer enough coverage and accuracy to permit an exhaustive identification of disease modules for some complex diseases. Here, I review some case studies of diseases where the identification of disease modules helped to uncover the molecular mechanisms of disease causation and identify new disease genes, pathways and potential drug targets.

- **Cancer:** Disease module identification methods have been extensively applied to study the molecular mechanisms of several types of cancer, including breast cancer (129–132), colon cancer (133), gastric cancer (134), prostate cancer (135) or acute myeloid leukemia (136). For example, Chang et al. (134) searched for genes whose expression was different in metastatic types of breast cancer, and identified the subnetworks of their gene products in the human interactome. By analyzing the resulting subnetworks, they found genes whose products had a central role interconnecting other disease proteins that remained undetected by differential expression methods. Additionally, the accuracy of classifying different types of breast cancer as metastatic or non-metastatic increased when using the markers obtained from the subnetworks. In a different case study, Taylor et al. (132) identified disease modules associated to breast cancer, and investigated the role of the different hubs within these modules. The analysis of expression of the genes encoding these hubs was useful to find markers for predicting breast cancer outcome.

- **Asthma:** The inflammatory disease of asthma has also been thoroughly characterized using network-based methodologies. Sharma et al. (137) defined the asthma disease module by applying clustering algorithms to asthma-associated genes. The authors compiled 129 asthma-associated genes that were represented in a human PPI network. 37 of these genes were highly interconnected, forming a cluster (called “proto-module”). The rest were spread through the interactome, either forming small clusters or disconnected from other asthma-associated genes. They expanded the proto-module by applying the DIAMOnD (119) topological community finding algorithm. DIAMOnD follows an iterative process to select the proteins that have a significant fraction of their interactions with asthma-associated genes, thus being more likely involved with the mechanism of the disease. The union of the proteins selected by DIAMOnD and the initial asthma-associated proteins conformed the “asthma disease module”. Within the asthma disease module, the authors found an inflammatory response signature that is shared with other auto-immune diseases (Crohn’s disease, multiple sclerosis, rheumatoid arthritis…). In contrast, another part of the disease module is specific for asthma. They investigated the asthma-specific region, identifying an enrichment in GAB1 expression. Their study emphasizes the importance of abnormal steroid response to asthma development and adds options for novel treatments. The study of Sharma et al. (137) was expanded recently by Maiorino et al. (138). In the new study, the authors applied a similar procedure to obtain the disease modules of Chronic Obstructive Pulmonary Disease (COPD) and asthma, and investigated the genes mediating the relationship between the two diseases.

- **Cardiovascular diseases:** They are another clear example of complex diseases, where their development is influenced by the modulation of many genes. Wang et al. (124) applied a network-based algorithm called Seed Connector Algorithm to identify the disease module of coronary artery disease. The Seed Connector Algorithm is a topological community finding method which iteratively includes neighbor proteins to an initial pool of known disease-associated proteins if the size of the largest connected component made by disease-associated proteins is increased. Without applying the algorithm, the 65 proteins associated with the disease were already forming a subnetwork of 18 proteins and 15 interactions, which is a significantly higher number than by chance. However, after applying the Seed Connector Algorithm, the authors identified a disease module of 88 proteins and 111 interactions. They also identified novel drug targets such as neuropilin-1 protein.

- **COVID-19:** The previously mentioned case studies were applied to complex diseases with a strong background genetic component. However, methods to identify disease modules can also been applied to study viral infectious diseases with a much lower genetic component involved. This is explained because diseases associated with viral infections can be studied by looking at the interaction between viral and host proteins. As in other types of diseases, the infection of the virus can be understood as a perturbation in a specific neighborhood of the interactome, which is also a disease module. For example, this type of methodology has been recently employed to understand the infection mechanisms of SARS-CoV-2, the virus causing the pandemic disease COVID-19. Gysi et al. (139), in a recent publication, compiled a list of 332 human target proteins of SARS-CoV-2 (140). 208 of the 332 SARS-CoV-2 targets were connected in the PPI network forming a module. By analyzing this module, they found that most of their proteins were expressed in the lung, directly linking the infection with lung affection. They also proposed potential drug treatments based on their effect to the proteins of the disease module. In a similar case study that can be found in this thesis (**Appendix 6.6**), a network diffusion-based method was used to identify the network modules associated to the infection of SARS-CoV-2, and to the severe effects produced by the infection (acute respiratory distress). The molecular information provided by this study helped to understand the effect of SARS-CoV-2 infection in severe cases and propose potential drug candidates.

1.3.3. Identification of relationships between disease modules to explore comorbidity

1.3.3.1. Definition of comorbidity

Because of the interconnected nature of the biological systems, it has been proposed that many human diseases are not independent of each other. Thus, the perturbation of specific shared components on the system might be the underlying cause of certain comorbidities or multimorbidities. The term "**comorbidity**" was already defined in 1970 by Feinstein as "any distinct clinical entity that has co-existed or that may occur during the clinical course of a patient who has the index disease under study" (141). Since then, the terms comorbidity and multimorbidity have been used in the relevant literature to refer to the presence of several disease conditions in a patient (142,143). While the term comorbidity is frequently used when there is a focus on a main, or index, disease and their associated conditions, the term multimorbidity is used when none of the diseases have higher importance over the rest. Although the meaning is very similar, in the case of comorbidity the therapeutic treatments or strategy are centered in a primary disease, whereas in multimorbidity no disease is given a higher priority (144).

1.3.3.2. *Dynamic properties of the interactome that explain comorbidities*

Traditionally, the network models based on protein and genetic interactions were static visions of the system, created under a single condition. In reality, complex systems, including complex diseases and comorbidities, are dynamic. The systems change and evolve depending on the context and time; and thus, properties such as pleiotropy, robustness and rewiring must be considered when studying complex diseases and comorbidities (145):

- **Pleiotropy:** Pleiotropy is the property of a given genetic locus to affect two or more phenotypic traits. It follows from this definition that genes are necessary multifunctional, encoding for proteins that can play multiple roles depending on the context. Pleiotropy and multifunctionality are closely related with comorbidity, as they explain why two genetically related phenotypes are more likely to co-occur. However, there are many ways in which a gene can lead to multiple functions. Hu et al. (145) unified the different theories into 5 models (**Figure 11**):

    - **Model 1:** Genetic changes affect multiple genes. This is the case of large deletions or insertions affecting multiple genes.

- **Model 2:** Alternative splicing as the source of functional diversity.
- **Model 3:** Multidomain proteins including domains with different functions; or proteins having different functions depending on the tissue.
- **Model 4:** Protein with a single function that affects multiple phenotypes, for example, by being present in multiple tissues.
- **Model 5:** Physiological changes caused by one phenotype that lead to the appearance of another phenotype.

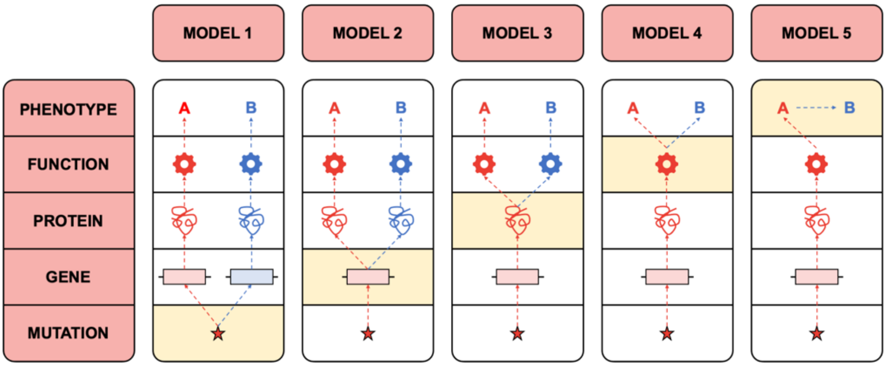


Figure 11. Five models of pleiotropy and multifunctionality. Scheme of the 5 models of pleiotropy and multifunctionality, which describe how a given genotype can lead to multiple phenotypes. Model 1: Genetic mutations affect multiple genes, possibly due to insertions or deletions. Model 2: Alternative splicing causes a gene to have multiple functions and encode different gene products. Model 3: A protein carries out distinct functions depending on the domain or the tissue where it is located. Model 4: The function of a protein is involved in multiple phenotypes. Model 5: The physiological changes caused by one phenotype lead to another phenotype. Figure adapted from Figure 3 of Hu et al. (145).

These models are useful to find etiological relationships in comorbidities. However, most comorbidities tend to be combinations of these models.

- **Robustness:** Robustness is the property that allows a system to maintain its functions against internal and external perturbations (113). Depending on the robustness of a biological system, the system maintains its functionality, or it changes when exposed to perturbations. Hence, diseases can be considered perturbations in the interconnected networks of molecular interactions. If the biological system is not robust enough, the perturbation of the system may lead to the development of multiple diseases. But even if the system is sufficiently robust, the perturbation may cause a change in the internal state of the system (molecular interactions) that makes it more fragile towards other new perturbations (comorbidities) (145). Robustness is not a general feature but depends on the system of the individual. Therefore, among individuals there will be several degrees of robustness, where individuals with the less robust systems are more prone to develop comorbidities.

- **Rewiring:** Rewiring is the restructuring of interactions between biological components due to conditional changes. In the context of graph theory, it would imply the destruction and creation of new interactions between the elements of the system. Rewiring has been measured with genetic interactions and protein-protein interactions, by mapping the changes across different conditions in differential networks

(146). However, differential networks are just the representation of the edge changes between static networks at different measures, and they do not necessarily reflect all the changes of a system. Rewiring is closely related to robustness: in response to a perturbation, the system rewires the interactions in order to maintain the equilibrium. But the rewiring of the system can also be caused by disease-causing mutations, leading to a disease state. Therefore, analyzing the rewiring of the system is key to find disease-associated genes and to understand the mechanism of complex diseases. The impact of rewiring depends on the position of the proteins in the network. Proteins with multiple interactions are likely to have multiple functions and change function and rewiring depending on the conditions (145,147). The relationship between robustness and rewiring of complex diseases was studied by Guney et al. (94), showing that even when randomly removing interactions, some diseases conserve the interconnectivity of the core proteins associated with the diseases (see **Figure 8** in the previous chapter).

1.3.3.3. Network separation between disease modules explains some comorbidities

Network medicine approaches have been widely used to reveal genetic connections between diseases and predict comorbidities. In some cases, comorbidities can be explained by shared genetic components such as disease-associated genes or biological pathways (117,148,149). However, a disease pair sharing genes does not necessarily mean that comorbidity exists (145,150,151), because the pleiotropy of genes can associate them with multiple pathophenotypes without necessarily give rise to a comorbidity. Likewise, comorbid diseases may not have a common genetic component; they can also arise due to shared pathways or the adverse effect of a clinical treatment (145).

Menche et al. (77) went a step further by identifying disease modules and analyzing the distances between them, finding comorbid relationships between diseases. They first created a protein interaction network compiling 141,296 physical interactions between 13,460 proteins. Then, they retrieved 299 diseases that were associated with at least 20 genes in OMIM and GWAS databases (103,152), obtaining 2,436 disease-gene associations. From each group of genes associated to a given disease, only the ones that were connected forming a subgraph were considered, ending up with 226 significant subgraphs that they considered disease modules. The underlying hypothesis was that the overlapping modules were more likely to disrupt pathways involved in the other disease module, resulting in shared clinical features. To answer this, they created a measure of topological distance between the disease modules. The measure, called "network-based separation" or s_{AB} (*Eq. 9*), was defined as the difference between the mean shortest distance between the proteins within each disease module (d_{AA} and d_{BB}) and the mean shortest distance between the proteins of the disease pair (d_{AB}).

$$s_{AB} \equiv \langle d_{AB} \rangle - \frac{\langle d_{AA} \rangle - \langle d_{BB} \rangle}{2} \qquad \textit{Eq. 9}$$

According to the authors, the disease modules topologically overlap when $s_{AB} < 0$, and they are topologically separated when $s_{AB} > 0$ (**Figure 12**). They compared this measure with 4 measures of pathobiological similarity: (i) biological similarity; (ii) co-expression; (iii) disease symptoms similarity; and (iv) comorbidity, measuring the relative risk (i.e., the fraction of patients in the population affected by both diseases divided by the product of the prevalence of the diseases) from US Medicare data. The results showed concordance between s_{AB} and the four measures, finding higher co-expression, comorbidity and biological and symptomatic similarity when the disease modules topologically overlap. This study showed that, despite the incompleteness of the interactome, it had reached enough coverage to allow a systematic investigation of disease mechanisms and helped to uncover some of the pathobiological relationships between diseases.

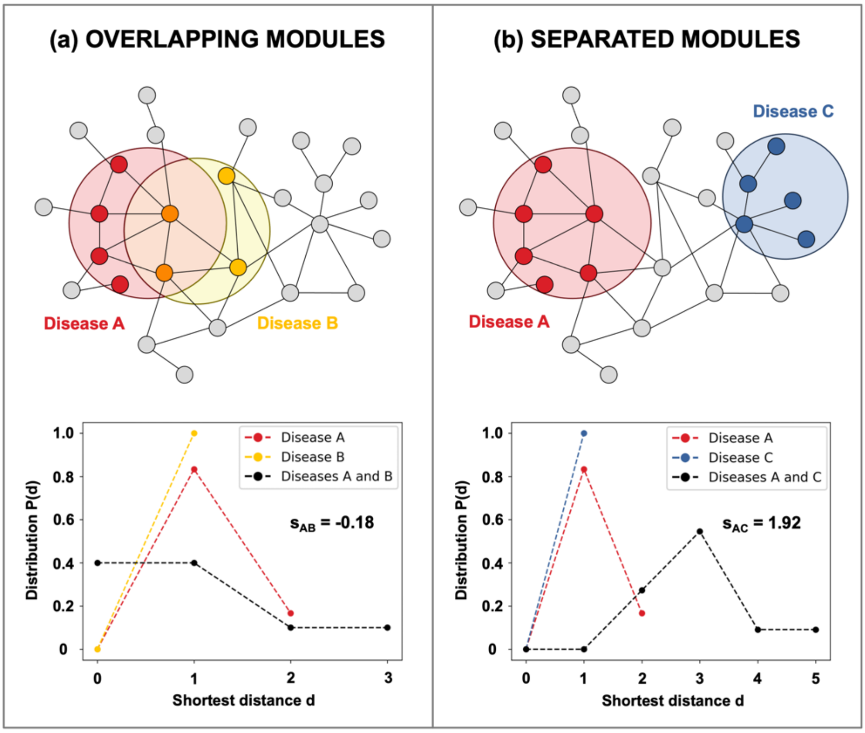


Figure 12. Network separation between disease modules. Examples of separation between the overlapping modules of disease A and B in (a), and separated modules of disease A and C in (b). The plots below show the distribution of shortest distance values of the proteins of the disease modules alone (in red, yellow and blue), and the proteins of the different disease modules (in black). The distributions of distances between proteins of the same disease module is similar to the distribution of distances between proteins of different modules when the modules are overlapped (s_{AB} < 0). In contrast, the distributions are different when the disease modules are separated (s_{AB} > 0). Figure adapted from Figures 3B and 3C of Menche et al. (77).

1.3.3.4. Including functional information to network-based methods improves the study of comorbidities

As discussed in the previous section, some comorbidities can be explained by a shared genetic component. However, the existence of shared genes does not necessarily involve comorbidity (150,151) and vice versa, comorbid diseases do not always show a shared genetic component (145). As gene products rarely act in isolation, it is necessary to consider the interactions between the disease-associated gene products to fully understand comorbidities. The work of Menche et al. (77) was especially relevant at this point, demonstrating that although the human interactome remains incomplete, the current view provides enough information to uncover the molecular mechanisms between disease relationships, including comorbidities.

Nevertheless, a close view of the disease relationships of Menche et al. (77) shows that 59% of disease pairs do not share any genes, suggesting that their relationship cannot be uncovered solely based on the shared gene hypothesis. This fact raises some important questions: Is it possible to complete the knowledge on disease-associated genes using the human interactome? Can we find molecular relationships for those comorbidities that are not apparently related by a genetic component, potentially due to such incompleteness?

In a recent work by Rubio-Perez et al. (92), these problems were tackled by considering disease pathways as an interactome-based extension of the known disease-genes and introducing measures of functional overlap. The approach consisted of: (i) integrating protein interaction data from several public resources using BIANA software (62) to derive one of the most up-to-date and comprehensive human protein interaction network; (ii) extending the incomplete current knowledge of disease-associated genes using the GUILD software (78), based on the topological closeness in the interactome to the initial disease-gene associations; and (iii) incorporating functional information that reveals comorbid links from shared disease-pathways. This process is explained schematically in **Figure 13**. The functional information was incorporated by considering the Gene Ontology (GO) biological processes, GO molecular functions and Reactome pathways enriched in the extended number of genes of each disease. Finally, Rubio-Perez et al. defined a genetic measure based on the overlap of (extended) genes between two diseases, and two functional measures based on the overlap of functions between two diseases and the enriched biological functions of the common genes (see **Figure 14** for a schematic representation).

The approach revealed 206 significant links among 94 diseases on a highly clustered disease association network. Moreover, around 95% of the links in the disease network, though not identified by genetic overlap, were discovered by functional overlap. Among the discovered connections, multiple sclerosis and rheumatoid arthritis were functionally linked through GO terms related to inflammation, consistent with previous findings (77,153). Similarly, the association between asthma and celiac disease was also identified using functional measures in agreement with the same previous studies (77,153). Some of the comorbidities uncovered by these studies are listed in **Table 5**.

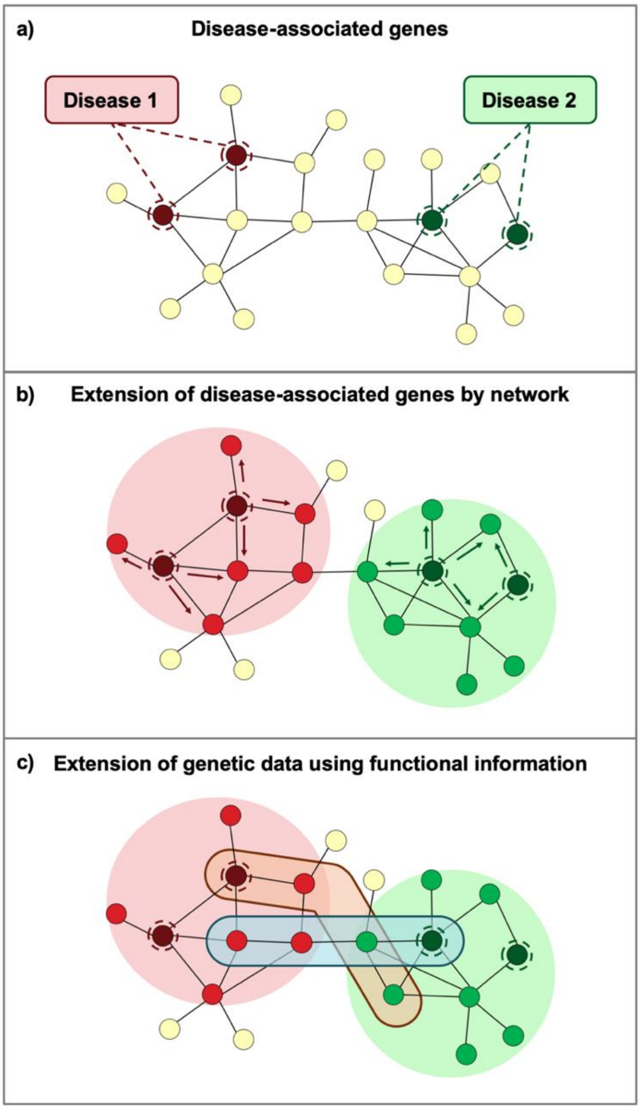

a) Disease-associated genes
Disease 1
Disease 2
b) Extension of disease-associated genes by network
c) Extension of genetic data using functional information

Figure 13. Using functional information to uncover disease comorbidities. (a) Network containing the disease-associated genes for a potential comorbidity of two diseases (shown by nodes with different hatching). (b) Extending the number of disease-gene associations using network-based approaches. Still, the incompleteness of the PPI network may hamper to uncover some of the relationships; this (c) can be unveiled by calculating the common enriched biological functions. Figure retrieved from Figure 2 of Aguirre-Plans et al. (96).

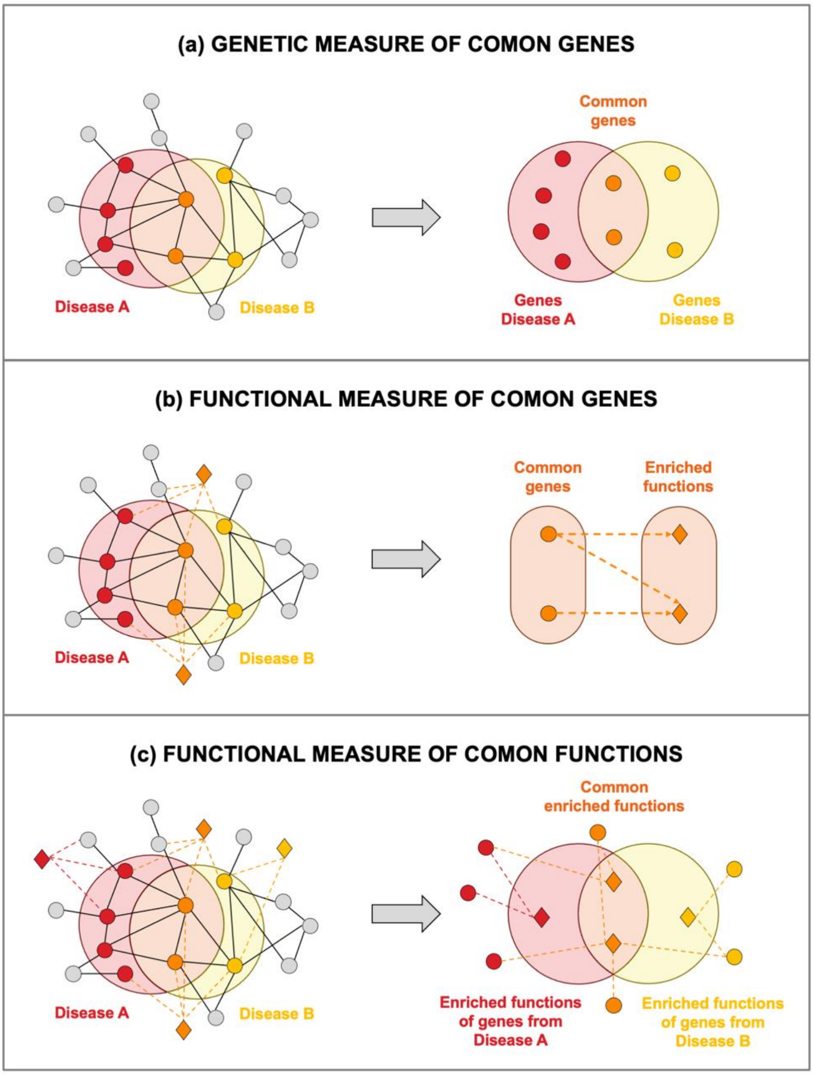

(a) GENETIC MEASURE OF COMON GENES
Disease A
Disease B
Common genes
Genes Disease A
Genes Disease B
(b) FUNCTIONAL MEASURE OF COMON GENES
Disease A
Disease B
Common genes
Enriched functions
(c) FUNCTIONAL MEASURE OF COMON FUNCTIONS
Disease A
Disease B
Common enriched functions
Enriched functions of genes from Disease A
Enriched functions of genes from Disease B

Figure 14. Schematic representation of the genetic and functional measures of disease overlap defined by Rubio-Perez et al. (92). (a) Genetic measure of common genes, containing the shared genes between the disease modules of two diseases. (b) Functional measure of common genes, containing the significantly enriched biological functions associated to the common genes of the two diseases. (c) Functional measure of common functions, containing the significantly enriched biological functions associated to each disease that are shared by both diseases. Figure adapted from Figures 1C, 1D and 1E of Rubio-Perez et al. (92).

Table 5. List of comorbidities predicted by Menche et al. (77) with a negative s_{AB} that are also uncovered by at least one of the measures of Rubio-Perez et al (92). Table retrieved from Aguirre-Plans et al. (96).

Disease 1	Disease 2	Menche et al. (s_{AB})	Rubio-Perez et al. (measures)
Asthma	Respiratory Hypersensitivity	-1.95	FCG-GObp
Hypersensitivity, Immediate	Respiratory Hypersensitivity	-1.80	FCG-GObp FCG-RP
Asthma	Hypersensitivity, Immediate	-1.75	FCF-GObp
Retinal Degeneration	Retinitis Pigmentosa	-1.12	FCF-GObp FCF-GOmf FCG-GObp FCG-GOmf
Cardiomyopathy, Dilated	Hypertrophic Cardiomyopathy	-0.55	FCG-GObp FCG-GOmf FCG-RP
Colitis, Ulcerative	Crohn Disease	-0.52	FCF-GObp
Leukemia	Neoplasms	-0.37	FCG-GObp
Arthritis, Rheumatoid	Systemic lupus erythematosus	-0.26	FCF-GObp
Celiac Disease	Multiple Sclerosis	-0.12	CG FCG-GObp FCG-RP
Lung Diseases	Coronary Arteriosclerosis	-0.08	FCG-GObp FCG-GOmf FCG-RP
Carcinoma	Neoplasms	-0.08	FCG-GObp
Arthritis, Rheumatoid	Celiac Disease	-0.05	FCF-GObp
Leukemia, B-cell, Chronic	Leukemia, Myelocytic, Acute	-0.04	FCF-GObp
Arthritis, Rheumatoid	Asthma	-0.03	FCF-GObp FCF-GOmf FCF-RP FCG-GObp FCG-GOmf FCG-RP

Arthritis, Rheumatoid	Respiratory Hypersensitivity	-0.03	CG FCG-GObp FCG-GOmf FCG-RP

A negative s_{AB} means that genes of two diseases are significantly closer than genes associated with the same disease, which could indicate that the disease modules of both diseases are topologically overlapped. There are three types of measures from *Rubio-Perez et al*: Common Genes (CG), Functional measure of Common Genes (FCG), and Functional measure of Common Functions (FCF). For the two functional measures, the functional terms are defined using Gene Ontology (GO) biological processes (named FCG-GObp and FCF-GObp measures), GO molecular functions (named FCG-GOmf and FCF-GOmf measures) and Reactome pathways (named FCG-RP and FCF-RP measures).

An important aspect of the work by Rubio-Perez et al. (92) was that, by taking into account functional information, the effect of pleiotropy and the multifunctionality of genes was implicitly considered. This is important because pleiotropy is one of the causes that gives rise to different pathophenotypes. By combining topological information (using the network-based algorithm of GUILD), functional information (incorporating data from GO and Reactome) and disease information (extending disease-gene associations), the authors provide insights on the potential causes of comorbidities. Apart from its contribution in the study of comorbidities, the same approach has been recently applied to study novel disease treatments (53,54).

1.3.3.5. Examples of comorbidities studied using network medicine

Here I review several examples of comorbidities studied using different network medicine methods:

- **Asthma and rheumatoid arthritis**

Rheumatoid arthritis and asthma have been related in multiple studies as a potential comorbidity. The link was already unveiled both in Hidalgo et al. (153) and Menche et al. (77). Several strategies proposed Tumor Necrosis Factor (TNF) as a therapeutic target for asthma (i.e., adalimumab, etanercep, infliximab) (155,156) and rheumatoid arthritis (i.e., in the DREAM Challenge 8.5 (157), in several cohorts with similar surveys (158–161)). However, the mechanisms underlying this association are still unclear (162). Rubio-Perez et al. (92) reported asthma as the disease with a highest degree of connectivity in the diseasome (20 links), followed by rheumatoid arthritis (16 links). Both diseases were in the largest cluster of the diseasome, together with other diseases such as Crohn's disease, chronic obstructive pulmonary disease, and respiratory hypersensitivity among others. Thus, according to that study, both diseases were prone to be associated with other diseases. The authors specifically explored the links underlying asthma and rheumatoid arthritis and found that both diseases were related with inflammatory processes. To further study the molecular mechanisms of the comorbidity, they focused on the link between the TNF, associated with rheumatoid arthritis, and its receptor in the superfamily 1B (TNFR1B), associated with both diseases. They highlighted one mutation in the interface of the interaction directly associated with rheumatoid arthritis.

In **Article 3.1** of this thesis (154), we decided to further investigate the genetic and functional relationship between these two diseases. We used GUILDify v2.0 to identify two disease modules of 290 and 181 proteins in the protein interaction network related with rheumatoid arthritis and asthma respectively. There were 55 proteins in common in both neighborhoods, and 31 shared enriched functions, which show a significant overlap based on both genetic and functional relationships. Among the shared top-ranking genes, we found TNF, previously highlighted as a potential precursor of the comorbidity. We also found HLA-DRB1 and several interleukins (IL18, IL1B, IL3), taking part of the immune response, which is potentially involved in both diseases. Moreover, most of the shared enriched functions between both diseases are related with inflammatory processes such as: “inflammatory response”, “positive regulation of interferon-gamma production” and “positive regulation of T-helper 1 cell cytokine production”.

- **Asthma and Chronic Obstructive Pulmonary Disease (COPD)**

Asthma and COPD are two of the most common respiratory diseases, which cause approximately 3 million deaths worldwide (163). Both diseases share many similar phenotypes (such as airflow obstruction, inflammation, and shortness of breath), but still very little is known about the shared molecular mechanisms between these diseases. Maiorino et al. (138) constructed the disease modules of asthma and COPD and investigated the commonalities between modules by applying a betweenness centrality-based measure. The authors first compiled 35 asthma-associated genes and 30 COPD-associated genes. Both sets of genes can be mapped in a PPI network of 16,656 proteins and 243,592 interactions. They extended the incomplete knowledge on disease-gene associations by applying DIAMOnD (119), an algorithm that identifies the proteins that significantly interact with the disease genes, thus being more likely involved with the mechanism of the disease. Using DIAMOnD, the authors defined the disease modules of asthma (373 genes) and COPD (228 genes). They found 14 overlapping genes, mainly involved in the regulation of apoptosis, proliferation, inflammation, cellular remodeling, and differentiation. Although the biological processes identified play a main role in both diseases, they are very common in many other diseases, and the number of overlapped genes was not significant.

The authors hypothesized that the perturbation in the interactome that leads asthma patients to develop COPD symptoms may not be carried exclusively by direct interactors of disease genes, but by mediating genes that are not specifically linked to a single disease. They defined these connecting genes as genes that participate in the majority of interactions between the two modules, becoming a “bottleneck” in the communication between them. To identify mediating genes, they introduced a measure called flow centrality, based on betweenness centrality. Flow centrality is a betweenness centrality measure with reduced coverage, which spans exclusively the shortest paths connecting the two modules, instead of the whole network. A high flow centrality score indicates that a node is highly central with respect to the genes of the two modules. Unlike the genes identified by the overlap of disease modules, flow central genes showed high specificity and were not directly interacting with disease genes. Flow central genes also showed functional similarity and co-expression with the genes of the disease modules.

- **Asthma, rhinitis and atopic dermatitis**

In the last years, the multimorbidity between asthma, rhinitis and atopic dermatitis (so-called allergic diseases) has gained an increasing attention. The process known as atopic march (or, more recently, atopic multimorbidity) (164) recognizes the increased occurrence of asthma, allergic rhinitis, or both, after atopic dermatitis onset (165,166).

The knowledge about the common mechanisms of allergic multimorbidity relies on a few candidate mechanisms, some are common to all allergic diseases (e.g., Type 2 immune-related response) and some are more specific (167–169). GWAS studies have identified many genes that contribute to Type 2 immune responses as well as to asthma, eczema, rhinitis as individual conditions through a variety of mechanisms (168–170). However, despite these evidences, how these different mechanisms jointly contribute to the allergic multimorbidity is still unclear (164,170).

Recently, several computational analyses of the diseasome of asthma, dermatitis and rhinitis helped to better characterize the mechanisms of atopic multimorbidity (171,172). The authors identified a core mechanism linking the three diseases and allowing to weight the involvement of different cellular mechanisms in the multimorbidity. Despite the limitations of the analysis (e.g., networks represented a diseasome static in time, while atopic multimorbidity is a progressive condition), network analysis provided a framework to integrate the current knowledge of the diseases' molecular mechanisms into a full picture of how the three diseases might be interconnected.

- **Alzheimer's disease and cancer**

The comorbidity between Alzheimer's disease and cancer has recently attracted the interest of the scientific community, although the mechanism is still poorly understood. Several studies suggested that for some cancer types, there is an inverse comorbidity (173,174), that is, a lower-than-expected probability of developing one of the diseases in individuals diagnosed with the other. In this case, patients with Alzheimer's disease are less prone to develop cancer and *vice versa*.

In Rubio-Perez et al. (92), Alzheimer's disease was observed to have significant links with different types of neoplasms, giving rise to a disease cluster containing Alzheimer and 7 neoplasms. Interestingly, the association between the two types of diseases was driven by common enriched biological functions. The main functions pointed towards the induction of apoptosis through caspase activation by mitochondrial cytochrome C. This suggests that apoptosis triggered by neurodegeneration in Alzheimer's disease may play a protective role in various cancer types by promoting programmed death of cancer cells. The authors further explored the role of the protein interactions in the comorbidity, finding 5 mutations potentially disrupting the interaction between the apoptosis regulator BAX and the BH3-interacting domain death agonist BID. The energy analysis and the predictions of hot spots of the interaction show that these mutations have a high potential to disrupt the BAX-BID interaction. The loss of the BAX-BID interaction could reduce apoptosis and its p53-dependent induction. Hypothetically, this could produce predisposition for cancer and inhibition of neurodegeneration, which may explain the inverse comorbidity between the two diseases.

1.3.4. Identification of endophenotypes and their relevance in the mechanism of diseases

Biological pathways tend to crosstalk. Their proteins are interconnected through the proteins of the interactome, influencing each other. Due to crosstalk, the pathways affected by a disease might as well influence other pathways, thus being relevant in modulating the pathophysiology of the disease (175). Crosstalk plays also an important role in some comorbid diseases, as they might be caused by the modulation of proteins belonging to common pathways (77,92,176). These crosstalking intermediate pathways shared among diseases are called **endophenotypes** (177), and they are key in the mechanism of many diseases and comorbidities.

Endophenotypes are collections of biological pathways interconnected with each other that play an important role in the development of many diseases (**Figure 15**) (3). The best known endophenotypes, which are present in most complex diseases, are inflammation, thrombosis and fibrosis (178). These endophenotypes facilitate the organism's adaptation to injury, with the goal of restoring the normal functioning of the organism. Ghiassian et al. (179) recently applied network medicine methods to study these endophenotypes, demonstrating their essential role in the progression of cardiovascular diseases. Specifically, they applied the topological community finding algorithm DIAMOnD (119) to identify the modules associated to the respective endophenotypes. They observed that the three modules have a large common core of proteins, meaning that they are endophenotypes closely related with each other. They analyzed the genes of the endophenotype modules, finding a high number of genes associated with cardiovascular risk factors. Additionally, they explored the topological properties of the endophenotype modules, showing high robustness, degree and betweenness centrality in proteins associated to inflammation and fibrosis.

Ghiassian et al. (179) showed that the study of endophenotypes through the identification of their network module can be key to understand the mechanisms and commonalities of complex diseases. After this landmark, endophenotypes are emerging as an attractive pharmacological target, as they are the perfect targets to treat comorbidities (180,181). In **Article 3.2**, we introduce a novel network medicine method to find drug candidates targeting endophenotypes, paving the way towards endopharmacology.

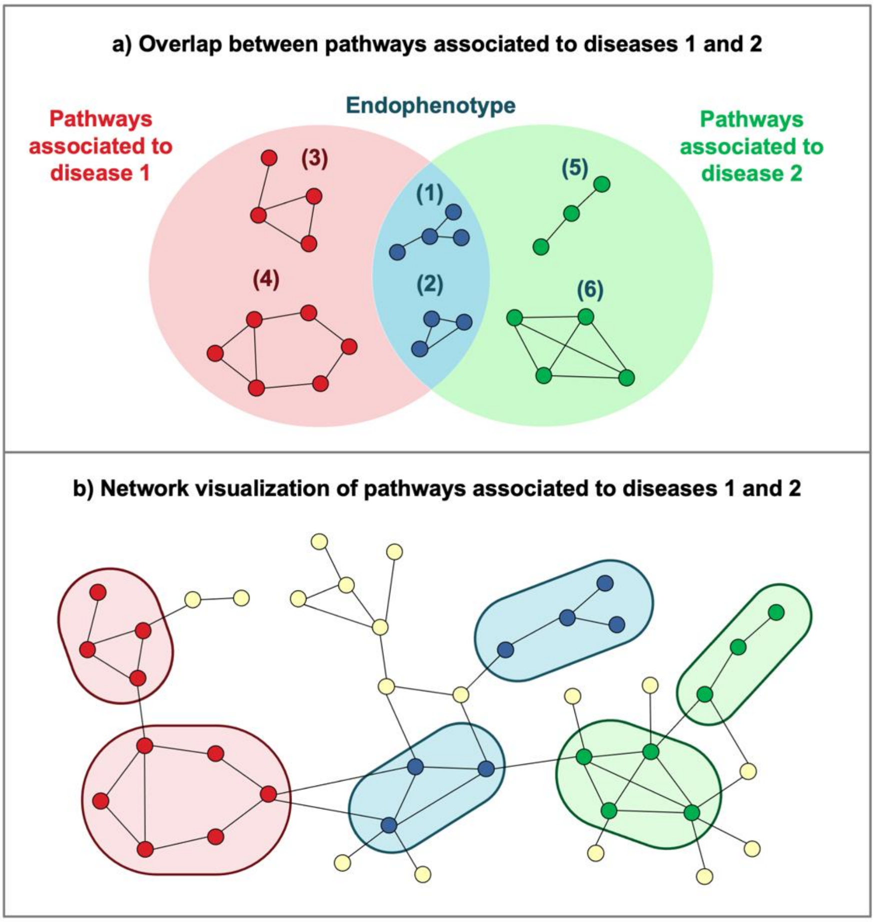

a) Overlap between pathways associated to diseases 1 and 2
Pathways associated to disease 1
Endophenotype
Pathways associated to disease 2
(3)
(4)
(1)
(2)
(5)
(6)
b) Network visualization of pathways associated to diseases 1 and 2

Figure 15. Schematic representation of endophenotype as a set of pathways shared by two diseases. (a) Venn plot between the pathways of disease 1 and 2, where the overlapped pathways represent an endophenotype. (b) Network visualization of the pathways associated to diseases 1 and 2.

1.4. Network medicine: towards finding better treatments

Pharmacology is the branch of medicine that studies the mechanism of action of drugs. From a network medicine perspective, pharmacology can be understood as the study of the effect of drugs at the interactome level, resulting on therapeutic and/or side effects. For this purpose, it is critical to identify the target biomolecules affected by the drug and understand the effect of their perturbation to the rest of the network. In the following subsections, I will introduce the strategies of network medicine towards a better understanding of drugs, their targets, and finding more effective, safer and personalized treatments.

1.4.1. From “one drug, one target, one disease” to network medicine

Traditionally, the pipeline followed by pharmaceutical industry to discover new drugs was based on Ehrlich’s paradigm of ‘magic bullets’: a drug interacts with an individual protein (182). This approach has produced many successful drugs, as the industry focused on: first, finding unique genes or proteins responsible of producing a disease; then, employing high-throughput screening techniques to find drug candidates that specifically modulate the selected target. However, there are two important assumptions of this paradigm that lead to two important problems (183):

(1) **“One disease, one gene” assumption:** Many diseases are not caused by a single gene or protein but by multiple ones.

If we focus on a unique target, there could be alternative signaling pathways or interactions, due to the robustness associated to the disease module of the network, that reduce the efficacy of the drug.

(2) **"One drug, one target" assumption:** Most of the drugs, either target multiple proteins, or affect the neighborhood of proteins close to their targets. If the drug modulates multiple proteins, this could produce more side effects than the ones expected for only one target.

Therefore, there are two important problems on assuming the 'one drug, one target, one disease' paradigm: a potential decrease of the efficacy and an increase in the toxicology of the drug candidates (182). These are the most important reasons why the attrition rates (compound failure) of the novel drugs are so high, which is one of the main concerns of pharmaceutical industry. From data of 2014, only the 10.4% of the novel compounds that started at phase 1 were launched to the market (184). The two main reasons of this failure were the lack of efficacy (in phase 2) and safety (in phase 3) of the new compounds (185).

Network medicine, which understands the organism as a group of interconnected networks, could be the solution to the late-stage high attrition rates of novel compounds. Network medicine suggests that, instead of focusing on searching disease-causing genes, we identify the subnetwork of molecules in the interactome that is perturbated by the drug. By understanding the **polypharmacology** of drug, how it perturbates the disease module of the interactome on multiple targets and pathways, it is possible to understand better the molecular effect of the drug in the organism and avoid novel drug failures.

Moreover, network medicine permits to explore alternative options that are faster and cheaper than the traditional drug discovery pipeline. The first alternative is **drug repurposing**, which consists in finding new indications for approved or investigational drugs that are outside the scope of the original indication. The second alternative are **drug combinations**, which are combinations of more than one drug that can be more effective than the single drugs. In the following subsections, the use of network medicine to study drug-target associations and identify more effective and safer drugs is explained in detail.

1.4.2. Drug-target associations: Identification methods and databases

A **drug target** is a biomolecule whose activity is modified by the interaction with a drug, resulting in a specific effect that might be positive (**therapeutic effect**) or negative (**side effect**) for the patient (186). Drug targets are mainly proteins, but they also can be nucleic acids. Depending on the type of interaction of the drug with the molecules in our body, there are two main concepts that need to be introduced: (i) **pharmacodynamics**, which is the study of how the drugs affects the molecules in the organism; and (ii) **pharmacokinetics**, which is the study of how the organism metabolizes the drugs (187). During this thesis, when discussing about drug-target associations, I will refer specifically to

pharmacodynamical interactions, involving the biochemical and physiological effect of drugs to targets (mainly human proteins).

1.4.2.1. Drug-target association identification methods

The identification of drug targets is key to understand the mechanism of action of the drug: the effect of the drug in the organism. The traditional way to identify drug-target associations is via biological experiments. Experimental methods can either be genetic interaction methods, based on monitoring the gene expression after the application of the drugs, or direct biochemical methods, based on detecting the binding affinity between the target and the drug (188). Binding affinity provides information on the strength of the interaction between a drug-target pair and is measured with different metrics: inhibition constant (K_i), dissociation constant (K_d), half-maximal inhibitory concentration (IC_{50}) and half-maximal effective concentration (EC_{50}) (189).

Although experimental methods are essential to have reliable information about drug-target associations, they are also expensive and time-consuming. For this reason, computational inference methods have risen as a fast, low-cost alternative to predict drug-target associations. According to the type of prediction, these methods can be divided into two categories: qualitative methods (which classify the drug-target pairs into association or non-association) and quantitative methods (which determine a value that indicates the strength of the association). Depending on the computational approach followed, these methods can be divided in several categories (190):

(1) **Molecular docking-based methods:** These methods can be used when the unbound three-dimensional structures of the target and drug are known. Docking methods use these structures to sample possible orientations of the bound drug-target interaction structure and rank them according to scoring functions that can be correlated with binding affinities (190,191).

(2) **Pharmacophore-based methods:** These methods are based on finding a pharmacophore model, which is the spatial arrangement of features that are necessary for a drug to interact with a specific target. The drug is screened for matching predefined protein-ligand structure-based pharmacophoric features (191,192).

(3) **Similarity-based methods:** These methods are based on the hypothesis that similar drugs are associated with similar targets and vice versa (190). There are different types of similarity-based methods, such as two-dimensional (193,194) and three-dimensional (195) structure similarity, or side-effect similarity (196).

(4) **Machine learning-based:** These methods apply machine learning algorithms to different types of biological data (e.g. drug chemical features, target protein sequences, known drug-target interactions, etc.) to predict drug-target associations (197).

(5) **Network-based methods:** These methods apply network-based algorithms usually based on topology similarity in

different types of biological networks (e.g. PPI networks or drug-target interaction networks) to predict new drug-target associations (190).

1.4.2.2. Drug-target association databases

The information on drug-target associations is spread over many different databases and repositories (197,198). This information can come from different types of sources: (i) from scientific literature, extracted via text-mining and/or manual curation; (ii) from other drug-target association databases; and (iii) from computational methods, thus being predictions that are not as reliable as other types of associations. **Table 6** provides information about the most important current drug-target association databases.

One of the main problems when working with drug-target associations is that, although there are many resources, the overlap between them is very poor. One of the main causes is that the determination of drug-target associations can be based on multiple qualitative parameters (e.g. multiple binding affinity metrics from different biological experiments, gene expression values, etc.). Therefore, depending on the criteria, the associations determined by each resource may be different. For this reason, integrating drug-target associations for an experiment is not a trivial problem. Here there are some examples of different efforts to integrate drug-target associations on specific experiments:

- Piñero et al. (95) compiled drug-target associations to characterize their transcriptomics, genomics and network features. The authors integrated data from DrugBank (199), DrugCentral (200), DGIdb (201) and ChEMBL (202). The associations from each database were filtered to select the associations from the most reliable sources: DrugCentral targets where only kept when belonging to the “Tclin” category (targets with comprehensive knowledge on their mechanism); DGIdb associations were selected when coming from ChEMBL, GuideToPharmacology, Tdg Clinical Trial, FDA, TEND and TTD; ChEMBL associations were only kept if the drugs could be mapped to DrugBank identifiers; and all the proteins reported as drug transporters, drug carriers or enzymes in DrugBank were excluded, therefore avoiding potential pharmacokinetic associations.

- Cheng et al. (203,204) used drug-target associations to measure the network distance between drug targets and disease genes and find novel drug repurposing candidates and drug combinations. The authors compiled drug-target associations from DrugBank (199), TTD (205) and PharmGKB (206). They also collected binding affinity data from ChEMBL (202), BindingDB (207) and Guide to PHARMACOLOGY (208). They only kept the human protein targets that could be represented by a unique Uniprot accession number and marked as reviewed in the Uniprot database (109). Finally, they selected drug-target associations with binding affinities including Ki, Kd, IC50 or EC50 each equal or below 10 µM.

- Yamanishi et al. (209) compiled a drug-target associations dataset as a gold standard to evaluate their performance predicting drug-target association networks. Their dataset integrates drug-target associations from KEGG BRITE (210), BRENDA (211), SuperTarget (212) and DrugBank (199) databases. This dataset had the particularity that it could be divided into four groups according to the types of protein targets (enzymes, G-protein coupled receptors, ion channels and nuclear receptors). Their gold standard has been used widely in many other studies of drug-target association prediction.

Table 6. List of drug-target association databases.

Database	Description	Drug-target association sources	URL
BindingDB (207)	Database of protein-ligand binding affinities	Scientific literature (by manual curation); other databases (PubChem, ChEMBL, PDSP Ki, CSAR)	bindingdb.org
ChEMBL (202)	Database of binding, functional and pharmacokinetic information for drugs	Scientific literature (by manual curation); other databases (PubChem, BindingDB)	ebi.ac.uk/chembl
DGIdb (201)	Database of drug-gene interactions from publications and other databases	Scientific literature (by text mining and manual curation); other databases (ChEMBL, DrugBank, Drug Target Commons, PharmGKB, TTD…)	dgidb.org
DrugBank (199)	Database of molecular information about drugs, their mechanisms, interactions and targets	Scientific literature (by text mining and manual curation)	go.drugbank.com

DrugCentral (200)	Database of drug information, including structure, bioactivity, regulatory and pharmacologic actions	Scientific literature (by manual curation), other databases (ChEMBL, Guide to Pharmacology, DrugMatrix, WOMBAT-PK, PDSP)	drugcentral.org
Drug Target Commons (213)	Database for community-driven drug bioactivity data integration	Scientific literature (by manual curation), users (by manual curation), other databases (ChEMBL)	drugtargetcommons.fimm.fi
Guide to Pharmacology (208)	Database of expert-curated ligand-activity-target relationships	Scientific literature (by manual curation)	guidetopharmacology.org
PharmGKB (206)	Database of information about human genetic variation on drug responses	Scientific literature (by manual curation)	pharmgkb.org
PubChem (214)	Database of chemical information (biological activities, safety, structure)	Other databases (BindingDB, ChEMBL, DGIdb, DrugBank, CTD, PDB)	pubchem.ncbi.nlm.nih.gov
STITCH (215)	Database of known and predicted chemical-protein interactions	Scientific literature (by text mining); other databases (DrugBank, Matador, TTD, CTD, ChEMBL, PDSP, PDB…); structure-based prediction	stitch.embl.de
SuperTarget (212)	Database of drug information about indications, side effects, metabolism and pathways for target proteins	Scientific literature (by text mining and manual curation); other databases (DrugBank, KEGG, PDB, SuperLigands, TTD)	bioinformatics.charite.de/supertarget
TTD (205)	Database of therapeutic target information, targeted diseases, pathways and drugs associated	Scientific literature (by text mining and manual curation)	db.idrblab.net/ttd

1.4.3. Network medicine for drug repurposing

1.4.3.1. Introduction to drug repurposing

Drug repurposing (also known as drug repositioning, reprofiling or re-tasking) consists in finding new uses to approved or investigational drugs that already have a given indication (216). This strategy offers important advantages over the traditional procedure of developing a new drug (217):

(1) **The safety of the compound has already been tested:** The drug has already been tested in preclinical models and humans (if early-stage trials have been completed). Therefore, it is less likely to fail due to safety reasons for the new indication.

(2) **The approval process is faster:** As most of the preclinical and clinical testing have already been completed, the approval procedure of a drug repurposing candidate is much faster than for a new drug.

(3) **The investment needed is lower:** As the approval process is shorter because there are phases that were previously completed for the original indication, the investment needed is lower. However, this can vary depending on the stage and process of development of the drug (218). The regulatory and phase III costs may remain similar as the ones of a new drug in the same indication. The most important savings would be in the costs of preclinical, phase I and II. Indeed, it was

estimated that taking a new drug to the market costs US$2-3 billion on average, whereas a drug repurposing candidate costs $300 million on average (219).

Historically, the first drug repurposing candidates were usually found by serendipity: if the drug showed an interesting off-target effect or a new on-target effect, it was considered for commercial exploitation. The most famous example is the case of sildenafil citrate (marketed as Viagra), originally indicated for hypertension, but repurposed by Pfizer for erectile dysfunction after clinical experience (216).

The multiple successes of drug repurposing have encouraged the development of systematic approaches to identify drug repurposing candidates. There are many systematic methods to identify new indications in drugs, most of them computational-based (217). Some of them are based on the binding between the structures of the drug and the targets (molecular docking). Others are based on analyzing how the drug modulates the expression of some genes. There are methods based on the analysis of clinical data reports. However, the only type of methods that put the drug in the molecular context of the disease are the network-based methods.

1.4.3.2. Network-based drug repurposing methods

Network-based drug repurposing methods involve the analysis of biological networks to predict new indications for drugs. These methods can be classified in different categories depending on the type of biological networks employed, including PPI networks, gene-regulatory networks or drug-target interaction networks (220):

(1) **Based on PPI networks:** Drug repurposing methods using PPI networks are based on the idea of identifying drugs targeting the disease modules within the PPI network. Guney et al. (79) hypothesized that a drug was effective against a disease by targeting proteins within or in the immediate neighborhood of the corresponding disease module. Based on this assumption, the authors proposed a drug-disease proximity measure where they quantified the distance between the proteins in the disease module and the targets of the drug of interest (see **Figure 16** for a schematic example). The measure, called "drug-disease proximity" or $d(S,T)$ (*Eq. 10*), was defined as the shortest path lengths ($d(s,t)$) between targets (t) of a drug (T) and proteins (s) associated with the disease module (S).

$$d\ (S,T) = \frac{1}{\|T\|}\sum_{t\in T} min_{s\in S} d(s,t) \qquad \textit{Eq. 10}$$

The implicit bias towards nodes with high degree when calculating the shortest path was addressed by calculating a z-score $\left(z = \frac{d-\mu}{\sigma}\right)$. They used a reference distance distribution corresponding to the expected distance between two randomly selected groups of proteins of the same size and degree as the original disease proteins and drug targets. They repeated the procedure 1000 times, and the mean (μ) and standard distribution (σ) of the distance $d(S,T)$ were used to calculate the z-score. This proximity measure was validated using pharmacoepidemiologic records from 220 million cardiovascular disease patients in a posterior study (203). Additionally, it was further used to propose drug repurposing candidates against the SARS-CoV-2 virus during COVID-19 pandemic (139,221).

(2) **Based on drug-target interaction networks:** These methods are based on the idea that proteins targeted by similar drugs are functionally related and close in the drug-target interaction network. For example, Cheng et al. (222) proposed a network-based inference method that, using a drug as seed, scores the rest of elements of the drug-target interaction network employing a process analogous to mass diffusion. Other methods apply machine learning algorithms to predict new drug-target interactions based on the known interactions of the network (220).

(3) **Based on gene-regulatory networks:** Gene expression patterns are known to change systematically in response to a drug or a disease. Thanks to gene expression detection technologies such as microarrays and RNA-seq, it is possible to monitor the gene expression changes and understand better the mechanism of action of drugs. For example, the Library of Integrated Network-Based Cellular Signatures (LINCS) (223,224) provides a large-scale gene expression catalog of perturbation-response signatures. The Drug Repurposing Hub (225) provides applications to analyze this type of data and identify drugs that produce a therapeutic response in cellular models and could be suitable repurposing candidates. Alternative methods, such as Greene et al. (226), construct a gene-regulatory network and identify disease modules from the analysis of genome-wide association studies to find potential new therapies.

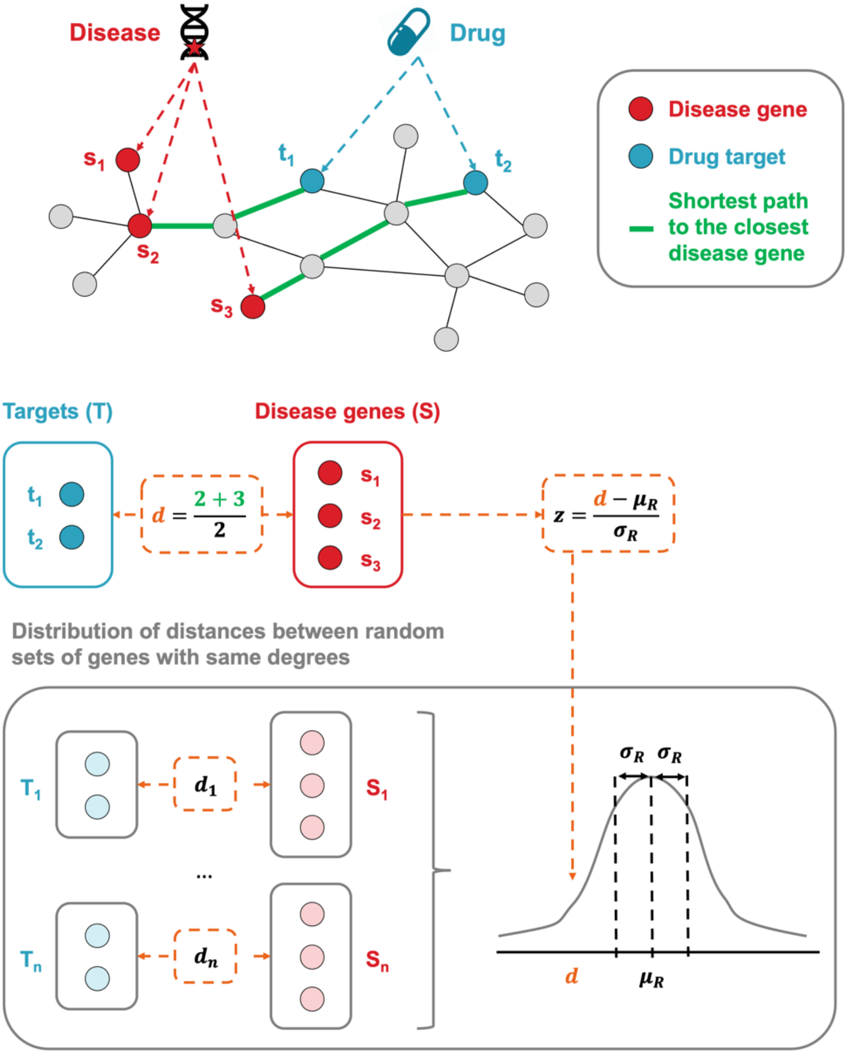


Figure 16. Schematic description of the network-based drug-disease proximity defined in Guney et al. (79). Calculation of the distance (d) between the targets of the drug (T) and the disease genes (G), by averaging the shortest paths from the targets to the closest disease genes. To measure the relative proximity (z), the distance is compared with a distribution of distances between random sets of genes with same degrees. The relative proximity (z) indicates if the network-based distance (d) is smaller than what is expected by chance. Figure adapted from Figure 2 of Guney et al. (79).

1.4.4. Network medicine for drug combination discovery

1.4.4.1. Introduction to drug combinations

A drug combination is a medicine that includes two or more active ingredients combined in a single dosage form. Drug combinations often show important advantages when compared to individual drugs (227):

(1) Drug combinations are able to simultaneously target multiple disease-related pathways in order to alleviate factors such as network robustness, redundancy or crosstalk, therefore increasing the efficacy of the treatments (228).

(2) The interaction between two drugs can sometimes lead to synergistic effects where the combined effect is better than individual ones. This is particularly relevant in cases where side effects are important as synergistic drugs can be used at lower dosage (229).

Drug combinations can be divided into two categories depending on the type of interaction between the drugs (227,230): **pharmacodynamic**, when the interaction between drugs directly influences their mechanism of action; or **pharmacokinetic**, when the interaction between drugs produces changes in the levels of absorption, distribution, metabolism or excretion of the drugs.

On the one side, pharmacodynamic drug combinations can be classified in three different groups depending on the effect that they produce:

(1) **Synergistic:** When the effect of the combination is greater than the summed effects of the individual drugs. The mechanisms of action that provoke the synergy can be divided in three categories (227):

   (a) **Anti-counteractive actions:** The synergy arises because the anti-counteractive actions of the drugs reduce the counteractive activities in the network against the therapeutic effect of the drugs.

   (b) **Complementary actions:** The synergy is produced because the drugs positively regulate a target or multiple points of a pathway, or collectively modulate the expression and activity of a target.

   (c) **Facilitating actions:** The synergy is caused because the drug-target interactions of one of the drugs produces an effect that facilitates the action of the other drug/s.

(2) **Additive:** When the effect of the combination is equal to the summed effects of the individual drugs. The mechanisms of action that provoke the additive effect can be divided in two categories (227):

   (a) **Overlapping or equivalent actions:** The additive effect is caused because the drugs interact with the same targets, or with different targets of the same pathway that equivalently regulate the same target.

(b) **Independent actions:** The additive effect is caused because the drugs interact with different targets of unrelated pathways, or with different sites of the same target.

(3) **Antagonistic:** When the effect of the combination is lower than the summed effects of the individual drugs. The antagonistic effect is always caused by counteractive or interfering actions between the drugs, but they can either be at the same target, or at different targets of related pathways.

On the other side, pharmacokinetic drug combinations can be classified in two different groups depending on their effect (227):

(1) **Potentiative:** When the therapeutic activity of one of the drugs is increased by the other drug/s due to a positive modulation of absorption, distribution, metabolism or excretion.

(2) **Reductive:** When the therapeutic activity of one of the drugs is decreased by the other drug/s due to a negative modulation of absorption, distribution, metabolism or excretion.

Finally, there is also another type of drug combination known as coalistic, in which the ingredients of the combination are individually inactive but active in combination (227).

1.4.4.2. *Drug combination discovery methods based on network medicine*

Most drug combinations used in clinic have been found empirically, meaning that only a small fraction of the full drug combination spectrum has been effectively exploited. In consequence, there is a large therapeutic space of drug combinations to uncover. In this aspect, computational approaches can provide an avenue to explore and predict drug combinations for diseases of interest through network medicine.

The field of computational prediction of drug combination is a fertile ground as proved by the number of methods available (231). Most of these methods (232–237) are focused on predicting the synergy of the compounds by using high-throughput screening assays from cancer cell lines (238,239). Although valuable, these type of prediction methods are very limited to very specific diseases and rarely delve into the actual molecular mechanism of the drug combination.

In a recent work, Cheng et al. (204) approached the challenge of predicting combinations of drug pairs by measuring the distance between the proteins targeted by the two drugs (i.e. drug-target modules) and the neighborhood of the interactome affected by the disease (i.e. disease module). To measure the distance between two drug-target modules, the authors used the network-based separation metric defined by Menche et al. (77) (*Eq. 9*), whereas to measure the distance between a drug-target module and a disease module, they used the drug-disease proximity metric defined by Guney et al. (79) (*Eq. 10*). The authors defined 6 classes of drug combination pairs depending on how their drug-target modules were overlapping the disease module (**Figure 17**):

(1) **Overlapping exposure:** Two overlapping drug-target modules are overlapping the disease module.

(2) **Complementary exposure:** Two separated drug-target modules are overlapping individually with the disease module (in different areas).

(3) **Indirect exposure:** One of the two overlapping drug-target modules overlaps with the disease module.

(4) **Single exposure:** One drug-target module separated from the other drug-target module overlaps with the disease module.

(5) **Non-exposure:** Two overlapping drug-target modules are topologically separated from the disease module.

(6) **Independent action:** Each of the drug-target modules and the disease module are topologically separated.

This study found that when looking for therapeutically effective drug combinations, the two drug-target modules, although separated in the interactome, were overlapping in the disease module. These findings were only applied for two complex diseases (hypertension and cancer) but can be generalized to other diseases, as they are proof that the understanding of the human interactome is key to predict drug combinations.

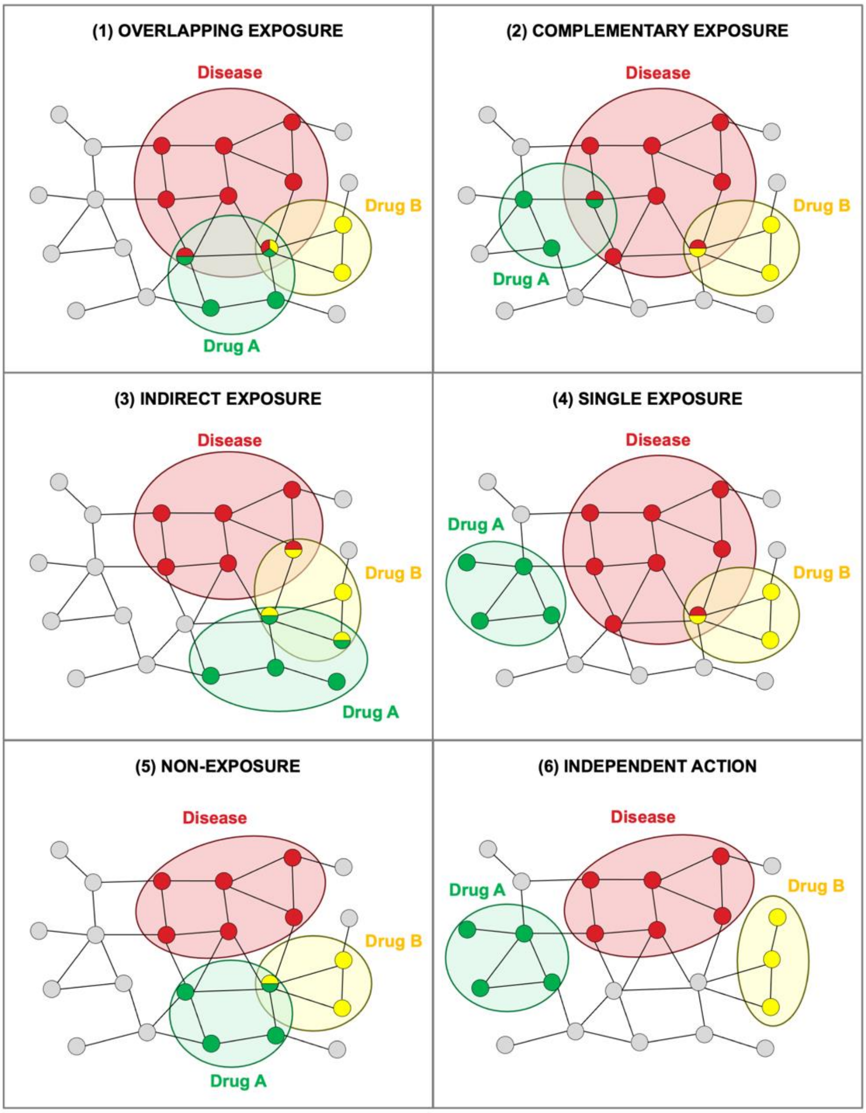


Figure 17. Schematic description of the six classes capturing the network-based relationship between two drug-target modules of a drug combination and one disease module, as defined by Cheng et al. (204). Figure adapted from Figure 2 of Cheng et al. (204).

1.4.5. Network medicine towards a more precise and personalized medicine

The first step towards a rational drug design is to understand the perturbation in the system caused by the disease. Network medicine addresses this problem by providing tools to identify the disease module, the neighborhood of the interactome perturbated by the disease (76). Thus, the next step is to understand the mechanism of action of drugs. Getting to know how a drug perturbates the interactome, and more specifically the module of the disease of interest, is the key to find the right treatment for the disease.

The most straightforward way to know how the drugs perturbate the interactome is to identify their drug targets. As seen in the previous section, there are plenty of varied computational methods to predict drug-target associations. Network medicine provides multiple approaches to identify unknown drug targets. They are usually based on the construction of either a drug-target network or a PPI network from the integration of public databases, and the exploitation of network-based algorithms (190).

However, thanks to the rapid growth of omics datasets and detailed phenotyping, network medicine is moving from a drug-centered view (drug discovery) to a patient-centered view (personalized medicine) with the following objectives: discover new disease subgroups, enhance patient risk stratification and develop individualized treatment strategies (72,181,240). To achieve this, several approaches to integrate and exploit multi-omics datasets have been developed, which can be divided in 5 groups (241) (**Figure 18**):

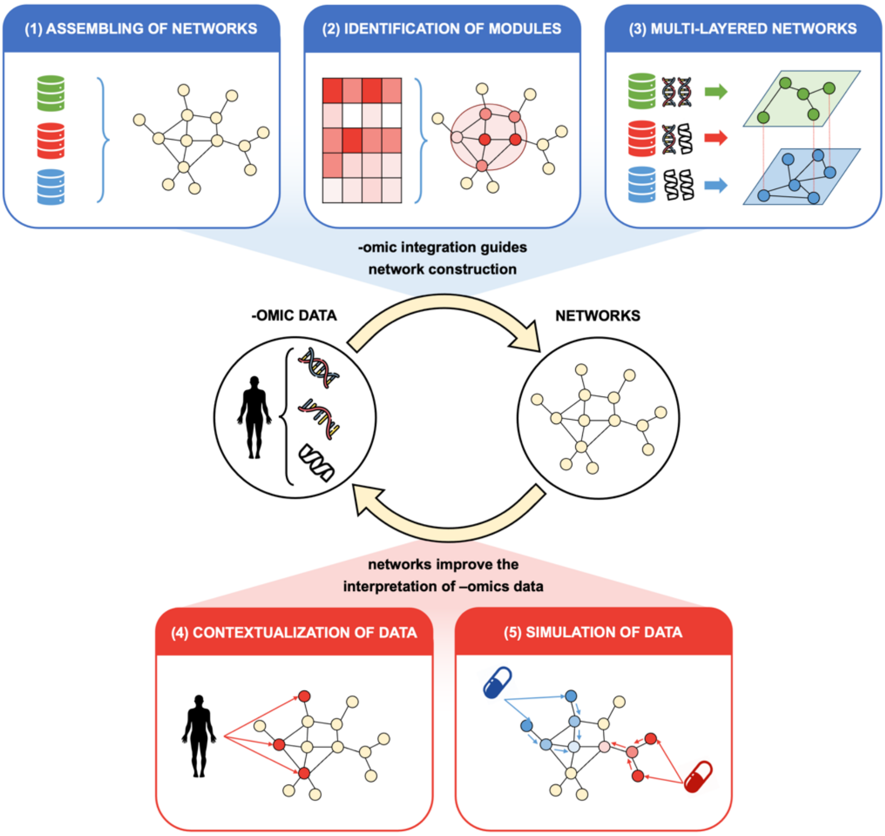

(1) ASSEMBLING OF NETWORKS
(2) IDENTIFICATION OF MODULES
(3) MULTI-LAYERED NETWORKS
-omic integration guides network construction
-OMIC DATA
NETWORKS
networks improve the interpretation of –omics data
(4) CONTEXTUALIZATION OF DATA
(5) SIMULATION OF DATA

Figure 18. Schematic representation of the feedback between -omics data and networks. -Omics data guides the construction of simple and multi-layered networks and the identification of modules inside the network. Networks provide the ideal context to analyze -omics data and make simulations and models. Figure adapted from Figure 12-1 of Ma et al. (241).

(1) **Multi-omics data-driven assembling of networks:** It consists in the reconstruction of knowledge-based networks at the genome or cell scale by compiling information from different types of experiments and research articles. Typical examples of network reconstruction based on multi-omics data are genome-scale metabolic networks. Their

reconstruction requires compilation of chemical reactions, genome annotations and literature findings (242).

(2) **Multi-omics data-driven identification of network modules:** It consists in the application of machine learning and statistical inference methods on multi-omics data to predict previously unknown network structures or identify functional network modules. For example, the application of biclustering (i.e. two-dimensional clustering of biomolecules and conditions) to infer functional modules (243). Another example would be the study of Oldham et al. (244), which applied network-based methods to analyze clinical data from patients with exercise intolerance and stratify the clinical risk. They assembled a network based on correlations between invasive cardiopulmonary exercise testing variables of patients. Using K-means analysis, they identified 4 distinct patient clusters.

(3) **Integration of multi-layered networks:** It consists in the integration and exploitation of different types of omics data by organizing them in multi-layered (or multiplex) networks, where each type of network (e.g. PPI network, metabolic network, drug-target network, etc.) is a different layer (245). Each interaction is not a pair of nodes, but a tuple of node-layers (246). The elements of the network can be analyzed at an intra-layer level (focusing on a single homogeneous network) or at an inter-layer level. By studying the connectivity of multi-layered networks, it is possible to identify certain areas that could provide biological insights. Still, such networks can also be more difficult to visualize or analyze,

requiring the use of fast-filtering or pattern-matching approaches based on machine learning (245). Multi-layer networks can be used in multiple applications, such as studying human diseases (247) or veterinary epidemiology (248).

(4) **Contextualization of multi-omics data using networks:** Using established networks as contextual basis, there are methods that apply statistical analyses to high-throughput data types to reduce their dimensionality and provide a clearer interpretation of the results. For example, genome-scale metabolic networks can be a useful framework for contextualizing disease-associated genes. It is the case of Lee et al. (249) study, where the authors mapped into a human metabolic network the disease-gene associations from OMIM (103). Using this network as basis, they created a new disease-disease network linking metabolic diseases if their associated genes catalyzed neighboring reactions in the metabolic network. The resulting network unveiled the metabolic mechanism of some comorbidities.

Another example would be the study of Cheng et al. (250) where the authors integrated patients' DNA and RNA sequencing profiles into the human PPI network. Specifically, they collected significantly mutated genes from large-scale genome datasets across 15 cancer types and mapped them to the PPI network. They found that the significantly mutated genes formed disease modules in the PPI network. Each disease module was expanded by applying a random walk network algorithm. Finally, they applied an arsenal of different drug repurposing algorithms to prioritize several drugs as potential treatments of these diseases.

(5) **In-silico network simulations of multi-omics data:** It consists in the integration of multi-omics data in established networks to predict the mechanism of cellular behavior. An example of this approach would be the Therapeutic Performance Mapping System (TPMS), a software to create models of all the possible mechanisms of actions that could exist between a drug and a disease or side effect (251,252). TPMS uses as basis a PPI network created from the integration of public repositories of PPIs. TPMS simulates the transmission of the perturbation of the drug through the PPI network from the stimulus (the drug targets) until the response (the disease-associated genes). The simulation of the perturbation is carried out mimicking the transmission of signal of a Multilayer Perceptron algorithm in a network. It takes as input signals the activation (+1) or inactivation (-1) of the drug targets, and as output the protein states of the disease. The models are trained by using restrictions (known activated or inactivated genes) retrieved from gene expression datasets for the particular disease. The models that best fit the conditions are considered the potential responses of the drug on patients with the studied disease (see **Article 3.3** for more details).

1.4.6. Network medicine towards safer treatments

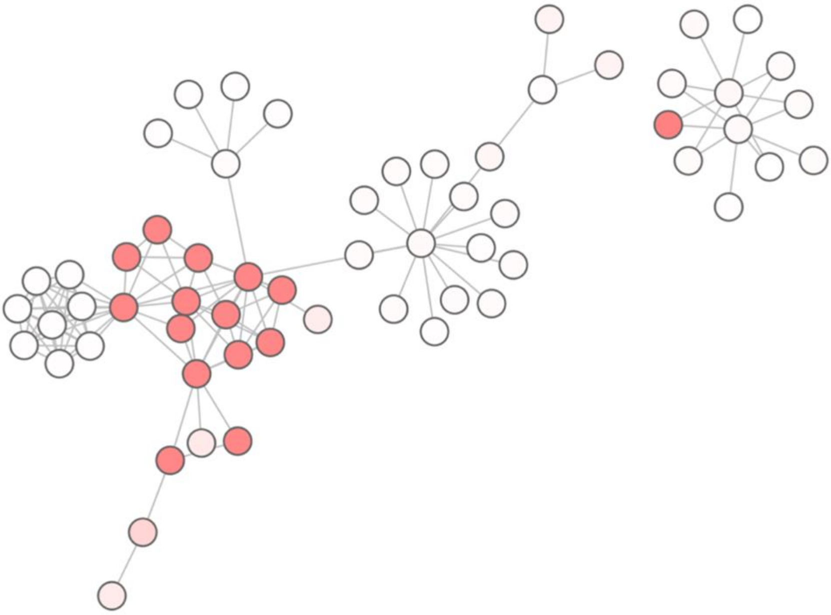
One of the main reasons of failure when launching new drug candidates is the safety of the compounds (185). Such problems are caused by undesired **side effects**, which are toxic reactions caused by the interaction of the drug with other molecules in the organism. The prediction of side effects remains as one of the main challenges of the pharmaceutic industry in the last years. In this aspect, computational methods are key to systematically exploit the biological information known about the drug to predict potential side effects.

Many computational methods focus on the prediction of drug-side effect associations, using biological features related with the drug. Among them, many methods are based on analyzing the chemical structure of the drug (253,254). Specifically, they identify commonalities between drug structures or specific chemical substructures that may arise such side effects. Others combine it with information such as the Anatomic Therapeutic Chemical (ATC) classification, literature mining or drug-target associations, using machine learning algorithms (255).

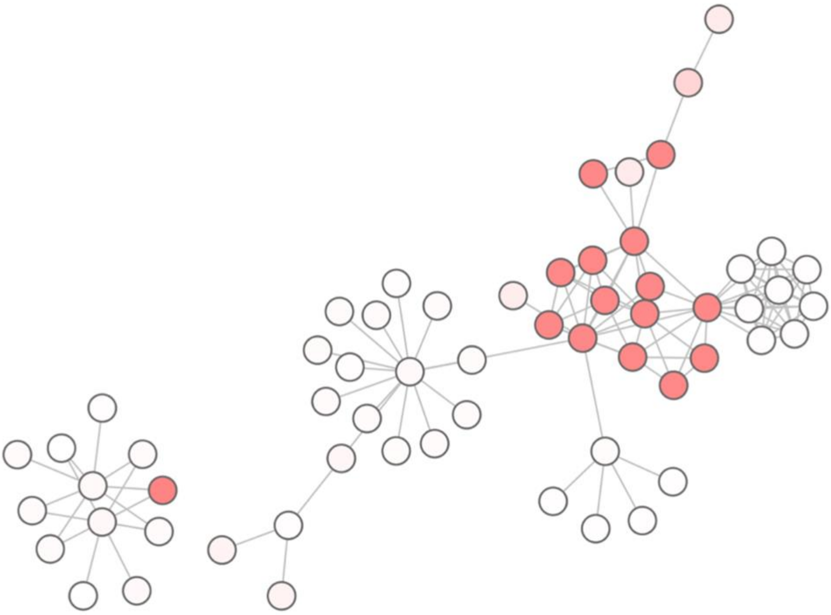
In a pioneer study, Kuhn et al. (256) decided to focus on finding proteins elucidating side effects when perturbated by the drug. Their idea was to combine drug-side effect relationships obtained from the SIDER database (257) with drug-target associations from the STITCH database (215), and identify those target-side effect relationships that were significantly more common. To do so, they applied a Fisher’s exact test to determine the significance of these relationships.

Although many side effects can be explained by the proteins targeted by the drug, many of them originate from the perturbation of proteins and biological functions in the neighborhood of the targets (183). Network medicine can be the key to explore how the drugs perturbate the interactome and the side effects that arise from such perturbations. Guney (258) started the way, by proposing a network medicine approach to identify side effect modules in the interactome. The author applied the approach of Kuhn et al. (256) to identify protein targets significantly associated to side effects. The groups of proteins associated to a common side effect constituted the module of such side effect. Then, the author applied a battery of network topology-based methods to quantify the closeness of the side effect modules to the targets of drugs. Still, there is a lack of methods applying network medicine to elucidate side effects.

More recently, the identification of gene expression patterns in genomics datasets using machine learning is becoming key for the prediction of toxic reactions in patients (259). This is possible due to the appearance of datasets such as LINCS (223,224), which provides a large-scale gene expression catalog of perturbation-response signatures. Additionally, initiatives such as the International Conference on Critical Assessment of Massive Data Analysis (CAMDA) are facilitating to assess the value of genomics and the state-of-the-art machine learning methods on predicting drug adverse reactions (260–262). The combination of the fields of toxicogenomics, pharmacogenomics and network medicine is the future for an *in silico*, personalized and precise prediction of side effects (245).



2. OBJECTIVES



The **focus of the thesis** is to develop *in silico* tools for network medicine that contribute to the research on diseases and the development of safer and more effective drugs.

Network medicine is improving our knowledge on the molecular mechanisms of human diseases by providing resources to organize and analyze biological systems. The **first objective** of the thesis is:

1. Develop an easy-to-use application to identify disease modules and analyze the protein interactions and biological functions perturbated, the relationship with other diseases and propose drug repurposing candidates.

This objective involves the following sub-objectives:

- Compile and integrate information on genes, proteins, drugs, diseases, PPIs, disease-gene associations and drug-target associations. Store this information in a database.
- Derive PPI networks from the previously integrated information on PPIs.
- Update the web server GUILDify, adding the following functionalities:
    - Calculate the biological functions enriched among the proteins of a disease module.
    - Explore the molecular mechanisms of comorbid disease pairs based on the previous work of Rubio-Perez et al. (92).
    - Detect drug-target modules associated with drugs.

  - Propose drug repurposing candidates based on the molecular and functional overlap of disease modules and drug-target modules.

Network medicine is especially suited to find drug repurposing candidates by placing the drug information in the molecular context of the interactome. This permits to explore specifically which pathophenotypes are targeted by the drug. For instance, endophenotypes, intermediate pathophenotypes shared by different diseases. The **second objective** of my thesis is:

2. Develop a method to repurpose drugs targeting endophenotypes.

This objective involves the following sub-objectives:

- Develop a method to rank drugs, based on the network-based drug-disease proximity (79) of their targets to a list of pathways of interest (i.e., pathways belonging to an endophenotype).
- Identify pathways proximal to disease genes across various autoimmune disorders using the previously developed method.
- Investigate whether the drugs promiscuously used in autoimmune disorders target specifically the pathways associated with one disease or the pathways shared across the diseases.
- Explore the potential endophenotypes shared by Type 2 Diabetes and Alzheimer’s Disease using the previously developed method.

Network medicine is moving towards a more personalized medicine where the center is the patient. This means to understand the effect of diseases on different types of patients and find drugs that are more effective and produce less side effects. This includes the consideration of novel types of treatments, such as drug repurposing candidates or drug combinations. The **third objective** of my thesis is:

3. Propose a network medicine method to identify the potential mechanisms of action of a drug for the treatment of a disease, stratifying different types of patients.

This objective involves the following sub-objectives:

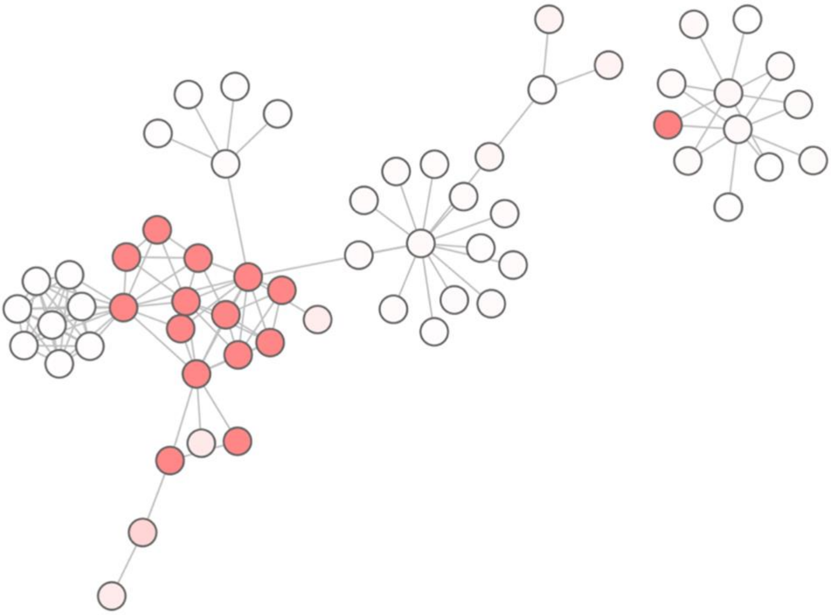
- Use the program TPMS to investigate the potential mechanisms of action of the drug combination sacubitril/valsartan in heart failure, associating the different mechanisms of action to different prototype-patients.
- Use the program TPMS to assess the potential mechanisms of action that could lead the drug combination sacubitril/valsartan to produce macular degeneration, associating these mechanisms of action to different prototype-patients.
- Identify biomarker proteins whose modulation is key to differentiate between classes of prototype-patients.
- Use GUILDify web server to explore the protein interactions and biological functions modulated by the drug combination sacubitril/valsartan in the disease modules of heart failure and macular degeneration.

Drug safety is one of the main problems of pharmaceutical industry and one of the main reasons of drug attrition during drug development. Network medicine could be key to understand the mechanism of action of drugs and predict side effects by integrating multi-omics data and applying machine learning techniques. The **fourth objective** of my thesis is:

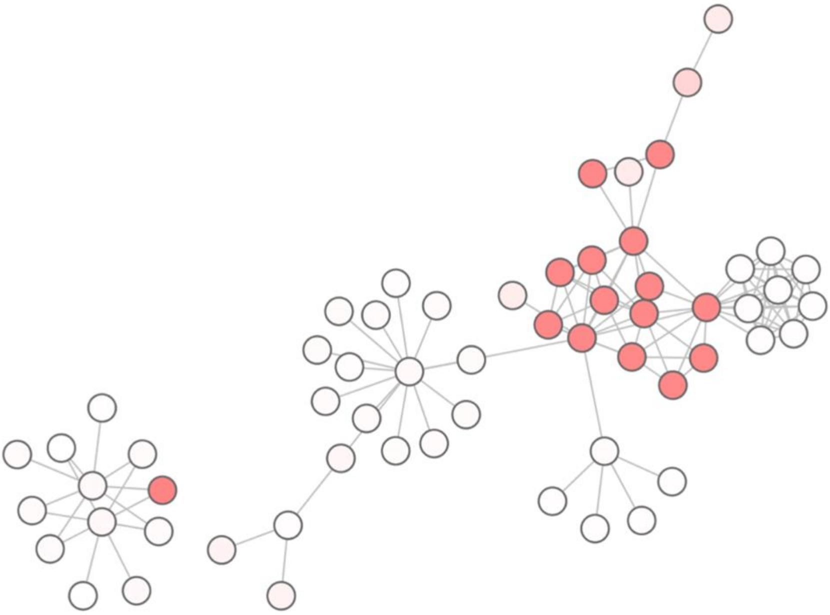
4. Implement a machine learning strategy to predict the drugs that cause drug-induced liver injury by analyzing multi-omics data.

This objective involves the following sub-objectives:

- Propose strategies to identify drug-induced liver injury gene signatures among L1000 datasets of CMap (224).
- Develop a machine learning ensemble to combine the predicted gene signatures with other types of omics data to predict drugs causing drug-induced liver injury.



3. RESULTS



3.1. GUILDify v2.0: A tool to identify disease modules and their relationships with other diseases and their druggable targets

In the first article of the thesis, I present GUILDify v2.0, a web server that applies the network diffusion method of GUILD (78) and the topological community finding method of DIAMOnD (119) to extend the information on disease-associated genes and identify disease modules. The work done to develop the update of the web server can be divided in 8 parts:

(1) Proportionate 7 species-specific PPI networks and 22 human tissue-specific PPI networks from the integration of multiple sources of biological interactions.

(2) Increase the quantity and quality of our dataset of disease-gene associations by incorporating DisGeNET (110).

(3) Integrate drug-target associations and incorporate the option to search by drug name, so that the user can extend the information on drug targets through the interactome and identify network modules affected by the drug.

(4) Improve the network visualization results by the incorporation of cytoscape.js (263) to the web server.

(5) Redefine the selection of proteins forming the disease module based on whether they have similar functional annotations as the disease-associated genes.

(6) Measure the genetic and functional overlap between the disease modules of two diseases, aiding the molecular understanding of disease-disease relationships and comorbidities.

(7) Implement a drug repurposing strategy based on the genetic and functional overlap between the modules associated to a disease and a drug.

(8) Develop an R package to facilitate the programmatic access to the web server.

As snapshot of the time of use, from 14th June of 2019 until 4th of February of 2021, GUILDify v2.0 has been accessed 1,238 times by 721 users.

Aguirre-Plans J, Piñero J, Sanz F, Furlong LI, Fernandez-Fuentes N, Oliva B, Guney E. GUILDify v2.0: A Tool to Identify Molecular Networks Underlying Human Diseases, Their Comorbidities and Their Druggable Targets. ***J Mol Biol*.** 2019; 431(13): 2477-2484. DOI: 10.1016/j.jmb.2019.02.027

GUILDify v2.0: A tool to identify molecular networks underlying human diseases, their comorbidities and their druggable targets

J Mol Biol. 2019 Jun 14;431(13):2477-2484

Joaquim Aguirre-Plans[1], Janet Piñero[2], Ferran Sanz[2], Laura I. Furlong[2], Narcis Fernandez-Fuentes[3,4], Baldo Oliva[1,*] and Emre Guney[2,5,*]

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

2- Integrative Biomedical Informatics Group, Research Programme on Biomedical Informatics, Hospital del Mar Medical Research Institute, Department of Experimental and Health Sciences, Universitat Pompeu Fabra, Barcelona, Catalonia 08003, Spain

3- Department of Biosciences, U Science Tech, Universitat de Vic-Universitat Central de Catalunya, Vic, Catalonia 08500, Spain

4- Institute of Biological, Environmental and Rural Sciences, Aberystwyth University, SY23 3EB Aberystwyth, United Kingdom

5- Department of Pharmacology and Personalised Medicine, CARIM, FHML, Maastricht University, Universiteitssingel 50, 6229 ER Maastricht, The Netherlands

***Correspondence to Emre Guney**: emre.guney@upf.edu

***Correspondence may also be addressed to Baldo Oliva**: baldo.oliva@upf.edu

Abstract

The genetic basis of complex diseases involves alterations on multiple genes. Unraveling the interplay between these genetic factors is key to the discovery of new biomarkers and treatments. In 2014, we introduced GUILDify, a web server that searches for genes associated to diseases, finds novel disease genes applying various network-based prioritization algorithms and proposes candidate drugs. Here, we present GUILDify v2.0, a major update and improvement of the original method, where we have included protein interaction data for seven species and 22 human tissues and incorporated the disease-gene associations from DisGeNET. To infer potential disease relationships associated with multi-morbidities, we introduced a novel feature for estimating the genetic and functional overlap of two diseases using the top-ranking genes and the associated enrichment of biological functions and pathways (as defined by GO and Reactome). The analysis of this overlap helps to identify the mechanistic role of genes and protein-protein interactions in comorbidities. Finally, we provided an R package, guildifyR, to facilitate programmatic access to GUILDify v2.0 (http://sbi.upf.edu/guildify2).

Keywords: disease comorbidity; drug repurposing; network analysis; systems medicine; target prioritization.

Introduction

Complex diseases such as cancer, diabetes, neurodegenerative disorders or cardiovascular diseases are rarely caused by a single genetic perturbation and usually involve polygenic modifications on the underlying interconnected cellular network. Understanding the genetic basis of diseases and the interactions of disease-associated proteins in the protein interaction network (PIN) is essential for the development of new rational therapeutic strategies. Despite recent large-scale genotyping efforts, information on disease-gene associations is still limited, often explaining a small percentage of the phenotypic variance observed among individuals (1). To address this limitation and infer novel disease-gene associations, various disease-gene prioritization methods have been suggested, exploiting the “guilt-by-association” principle over certain features of disease-genes such as similarity in sequence and functional annotations, clustering in the linkage interval, or proximity in the PIN (2). Indeed, albeit the PINs being incomplete (3), the proximity to disease-genes in the PIN has proven extremely useful in prioritizing disease-associated genes (4). Consequently, a number of tools and web servers has been developed to expand the number of disease-associated genes using the interactome (5–9).

Previously, we presented GUILDify, a web server that applies the prioritization algorithms developed in GUILD software to find novel disease-gene associations based on the connectedness of genes in the PIN (10,11). GUILDify searches for genes starting from user-provided keywords such as the names of diseases or gene symbols in the BIANA knowledge database. It uses the genes associated to the keywords as seeds and the PIN for the selected organism to apply graph theory algorithms to prioritize new disease genes. Recently, GUILDify has been applied to: (i) find comorbidities across genetic diseases (12); (ii) construct PINs specific to breast cancer metastasis to lung and brain (13); (iii) identify candidate genes for body size in sheep (14) and (iv) prioritize preeclampsia pathogenesis (15).

Here, we present a comprehensive upgrade, GUILDify v2.0, where we updated the underlying biological databases in BIANA knowledge database (protein and drug-target interactions, functional and disease annotations) and: (i) facilitated the use of seven species-specific PINs and 22 human tissue-specific PINs; (ii) increased the quality and number of disease-gene associations by incorporating DisGeNET to our datasets; (iii) incorporated the option to search by drug name, allowing the prioritization of genes based on known drug targets to uncover the neighborhood of the PIN affected by the drug; (iv) improved the visualization of the results using cytoscape.js; (v) refined the definition of top-ranking genes based on whether they had similar functional annotations as the seeds, thus providing the biologically most coherent subnetwork relevant to a given disease; (vi) introduced a feature to measure the genetic and functional overlap of the top-ranking genes of two different diseases, supporting the investigation of disease comorbidities; (vii) implemented a new drug repurposing functionality to propose novel indications for a given drug based on the genetic and functional overlap; and (viii) developed an R package to facilitate the programmatic access to the methods implemented in the web server.

Results and Discussion

Advances

Identifying genetic and functional similarities across diseases

In recent works, we have shown that the genetic and functional similarities of diseases in the PIN can be used to characterize co- and multi-morbidities across diseases (12) and also to repurpose existing drugs targeting these diseases (16). Motivated by these findings and to provide systematic insights on disease-disease relationships, GUILDify v2.0 now allows users to identify the overlap between two previously submitted results, i.e. sets of genes linked to two different diseases. Accordingly, given two job IDs corresponding to the prioritization results of two different diseases, GUILDify v2.0 provides: (i) the overlap between the top-ranking genes of the two diseases; (ii) the overlap between the enriched functions among the top-ranking genes of the two diseases; (iii) the enriched functions among the common top-ranking genes; and (iv) a network visualization of the interactions between common top-ranking genes. Moreover, GUILDify v2.0 also calculates the Fisher's exact test to quantify the significance of the overlap between genes and functions and report one-sided P-value (see details in **Supplementary Material**). GUILDify v2.0 is the first server that permits the use of gene prioritization results to explore disease-disease relationships with such simplicity and flexibility.

Prioritization of drug targets

GUILDify v2.0 now allows to search by a drug in addition to a phenotype and returns a list of drug-target associations integrated from DrugBank (17), DGIdb (18), DrugCentral (19) and ChEMBL (20) (see details in **Supplementary Material**). This new functionality allows the characterization of the neighborhood of the drug in the PIN, i.e. neighboring proteins to those targeted by the drug, and thus providing insights on the potential mechanism of action of the drug. Moreover, the novel feature of assessing the overlap between two network expansion runs (i.e. two job IDs) can also be applied in multiple scenarios to: (i) identify the similarity between the neighborhood of two drugs in the PIN, which can be useful to identify drug interactions; (ii) compare the neighborhood of a disease with the neighborhood of a drug in the PIN, which can be applied to drug repurposing. Such novel features make GUILDify v2.0 one of the most easy-to-use and flexible web servers to inspect the effect of drugs in the PIN.

Screening diseases to identify potential new indications of known drugs

Building upon new technical developments mentioned above, GUILDify v2.0 now offers a novel drug repurposing functionality. Given a job ID associated with a drug (or a list of drug targets), this feature automatically calculates the overlap of genes (or functions) between the given drug and a set of pre-calculated diseases. Details on the method and validation of drug repurposing are described in detail at **Supplementary Material**.

Tissue and species-specific PINs

The analysis of the protein interactions in a tissue-specific context is becoming increasingly relevant to understand genetic diseases and find improved treatments (21). We have included tissue-specific networks derived from 22 different human tissues (see **Supplementary Table S1**). To create these networks, we filtered the interactions in the global PIN using RNAseq data from GTEx (22), keeping only the interactions between proteins encoded by genes that are expressed in a given tissue (i.e. considering only transcripts with TPM (transcripts per kilobase million) expression values of 1 or higher (see details in **Supplementary Material**). We have also included 7 species-specific PINs derived from experimentally determined protein-protein interactions. Although the coverage of interactomic data for some species is low (e.g., 11,943 interactions in rat vs 320,337 interactions in human), these PINs provide a reliable backbone for interactome-based analyses (e.g., in preclinical research) as opposed to PINs generated by predicted interactions based on homology information.

Disease-gene information from DisGeNET

We incorporated DisGeNET, one of the largest repositories of genes and variants associated to human diseases (23). DisGeNET relies on data from UniProt (24), CTD (25), CLINVAR (26), ORPHANET (27), GWAS Catalog (28), PsyGeNET (29) and HPO (30) and is integrated in BIANA (31). To investigate the increase in the number of disease-gene associations between versions 1 and 2 of GUILDify, we checked the number of associations for the lowest-level non-obsolete diseases from Disease Ontology (32) that were available in our repositories (2,190 terms). GUILDify v1 contains gene associations for 1,505 diseases and 4,171 genes (2.8 genes per disease), while updated GUILDify v2.0 has gene associations for 2,064 diseases and 11,615 genes (5.6 genes per disease on average).

Functional-coherency based selection of top-ranking genes

One of the main issues when working with disease-gene prioritization is to select the most relevant (top ranked) genes associated with a given disease. The user can select top 1% or 2% highest scoring genes among all the proteins in the PIN as top ranked genes. In GUILDify v2.0, we also introduced a cutoff based on the functional validation approach described in Ghiassian *et al.* (5) and provided a new panel visualizing the significance of the functional enrichment (P-value) as a function of the number of top-ranking genes included in the validation (implemented in Plotly). In brief, the highest-scoring non-seed proteins are iteratively included in the top-ranking set, provided that they maintain the functional coherency of the existing top-ranking set (see details in **Supplementary Material**). Note that this approach might be too restrictive for some complex diseases in which the information on known disease-gene associations is limited, failing to represent the functional diversity involved in the disease.

Visualization of the top-ranking subnetwork

GUILDify v2.0 uses the JavaScript-based network visualization library, Cytoscape.js (33), to show the subnetwork of the top-ranking proteins and the drugs targeting these proteins. The user can decide the cutoff to define the top ranked proteins to be visualized (top 1%, top 2% or functionally-coherent as mentioned above). In addition to seeds (green hexagons), top-ranking proteins (yellow circles) and drugs (blue diamonds), the subnetwork includes the proteins that connect the seeds to the largest connected component induced by seeds (named “linkers” and shown as grey circles, see details in **Supplementary Material**).

R package

We have included an R package in order to provide programmatic access to GUILDify v2.0 through R statistical computing environment (https://www.r-project.org/). The package implements methods to query and retrieve results from the web server as an R data frame, allowing users to run multiple queries for more high-throughput and/or systematic analyses. The package and documentation are available online at: http://sbi.upf.edu/guildify2.

GUILDify v2.0 workflow

Input

The interface of GUILDify v2.0 is designed to be simple and intuitive. The input varies slightly depending on the desired task: (i) a new search; (ii) retrieving results from a previous run; and (iii) calculating genetic and functional overlap between two previous runs. For a new search, we require two steps: first the selection of seeds (genes associated with a phenotype or drug) and second the selection of parameters to run the prioritization algorithms. For the selection of seeds the user has to provide: (i) either keyword(s) describing the phenotype/drug of interest or a set of specific gene names separated by a semicolon; (ii) the species of interest (default value: *Homo sapiens*); (iii) the tissue of interest (default value: *All*); and (iv) the PIN source (default value: BIANA). If the user provides a keyword (or set of keywords) describing a phenotype or drug, the server searches genes containing the keyword in BIANA knowledge database (i.e. integrating information from many resources), otherwise it uses the list of provided gene names. The server shows the selected seeds, which can still be filtered and selected by the user. Then, for the prioritization parameters the user can select to run the “disease module detection algorithm” (DIAMOnD, downloaded from https://github.com/dinaghiassian/DIAMOnD) (5) or to use one of the several prioritization algorithms from the GUILD package (default value: NetScore with default parameters). Finally, to retrieve results, the required input is the job ID of a previous run, while for calculating genetic and functional overlap the inputs are two job IDs of previous runs.

Output

GUILDify v2.0 outputs the ranking of the nodes in the PIN and the visualization of the subnetwork involving the top-ranking genes in a cytoscape.js panel. In addition, the output page has: (i) a panel showing the P-values of functional enrichment of the ranked nodes; (ii) two panels with functions enriched among the top-ranking nodes and seeds, respectively; and (iii) one panel with the drugs that target the top-ranking proteins.

For the "*Overlap between two results*" option, the server provides: (i) the list of the common top-ranking genes and the significance of the overlap assessed by a Fisher's exact test (see details in **Supplementary Material**); (ii) the network visualization of the common top-ranking genes including the "linkers" (see above); (iii) the list of enriched functions of the common genes; iv) the list of common enriched functions of both results and the significance of the overlap; and v) the drugs targeting the proteins of the common PIN. Using this functionality, the users can identify the overlap between any two queries such as between two diseases, two drugs or a disease and a drug. Although we do not provide the overlap between interactions of top-ranking proteins in a separate table, these interactions can be investigated in the network visualization panel.

Case studies

Exploring the mechanistic links between rheumatoid arthritis and asthma

In multiple studies, rheumatoid arthritis and asthma are linked as a potential comorbidity, although the mechanisms underlying this association remain unclear (34). Using the new functionality of GUILDify v2.0, we can assess the overlap between diseases and thus propose a potential mechanism to explain the association between them. Querying for "rheumatoid arthritis" and "asthma" returns 156 and 96 seeds, respectively coming from DisGeNET, OMIM, and UniProt. There are already 12 seeds in common (Fisher's exact test, one-sided P-value = $1.4{\cdot}10^{-9}$) and 18 common functions out of the total enriched functions of the seeds (P-value = $9.3 \cdot 10^{-23}$). After running GUILDify v2.0, we select 290 and 181 top ranked genes using functional-coherency based cutoff for rheumatoid arthritis and asthma, respectively. We find that the number of common genes increases to 55 (yielding a P-value = $5.9 \cdot 10^{-48}$), while the number of common functions (biological processes) increases to 31 (P-value = $8.1 \cdot 10^{-46}$). The link between these diseases is significant even when the seeds are removed from the top-ranking genes (see **Supplementary Material**). Among the shared top-ranking genes, we find Tumor Necrosis Factor (TNF), which has been proposed as a potential drug target for asthma and rheumatoid arthritis, and highlighted as a potential precursor of the comorbidity (12). We also find HLA-DRB1 and several interleukins (IL18, IL1B, IL3), taking part of the immune response potentially involved in both diseases. Furthermore, the most common enriched functions relate to inflammatory processes such as “inflammatory response”, “positive regulation of interferon-gamma production” and “positive regulation of T-helper 1 cell cytokine production”. These functions appear again if we check the functions enriched by the common genes, along with other functions such as “T-helper 1 type immune response” or “negative regulation of type 2 immune response”, highlighting the involvement of type 1 immune response in both diseases. As negative controls, we repeated the analysis using other disease pairs that are not likely to be comorbid such as “rheumatoid arthritis” - “breast cancer” and “asthma” - “breast cancer”, finding drastically reduced number of genes in the overlap between these disease pairs (see **Supplementary Material**). The results can be further explored in **Figure 1** and in the pre-calculated examples section of the web. Additionally, we compared the functional relevance of the top-ranking genes identified by NetScore with DIAMOnD, based on the analysis in *Sharma et al.* (35) (see **Supplementary Material**)*.* We checked the enrichment of top-ranking genes among the pathways containing the seed genes of asthma and rheumatoid arthritis, showing that both methods significantly recover the pathways in each disease. Furthermore, NetScore identified more genes that belonged to the pathways shared between asthma and rheumatoid arthritis compared to DIAMOnD.

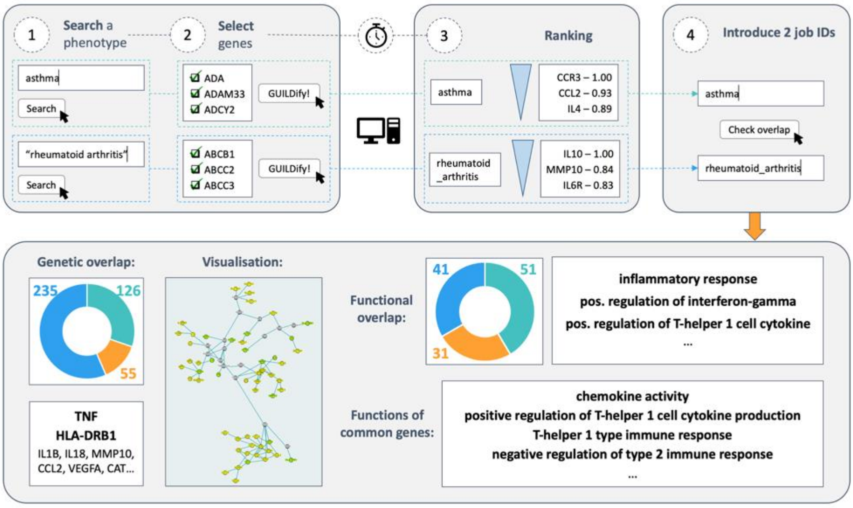


Figure 1. GUILDify v2.0 example study on the comorbidity between asthma and rheumatoid arthritis. First, we run the prioritizations of the two diseases by searching (1) and selecting (2) the genes. After obtaining the ranking of proteins from the prioritization (3), we use both job IDs to check their overlap (4) and inspect the genetic and functional relationships between them (see details at http://sbi.upf.edu/guildify2 in the pre-calculated examples section).

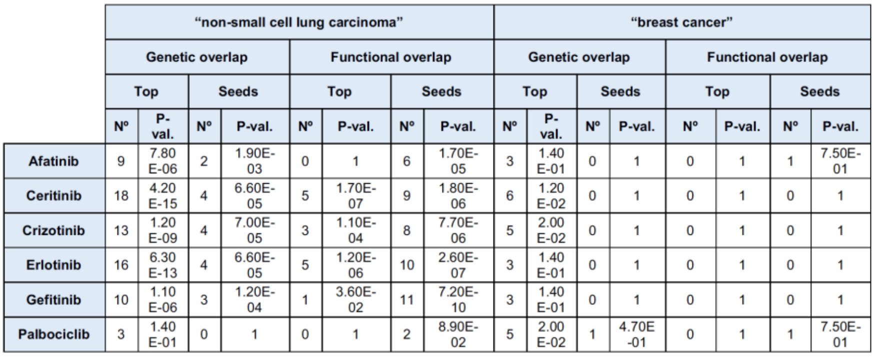
Study of the mechanism of non-small cell lung carcinoma drugs

Non-small cell lung cancer (NSCLC) is the most common type of lung cancer. Typically induced by exposure to toxic substances, the NSCLC pathology has been specially associated with a mutation in the Epidermal Growth Factor Receptor (EGFR) (36). In a recent study, 9 drugs were proposed to treat this disease (37), 6 of them having drug-target interactions reported: Afatinib, Ceritinib, Crizotinib, Erlotinib, Gefitinib and Palbociclib. Given that we can now identify potentially new relationships between drugs and diseases using drugs as queries, we investigate whether the neighborhood of the targets of these drugs in the PIN significantly overlaps with the neighborhood of the genes associated with NSCLC. We used GUILDify v2.0 to define this neighborhood. We observe that the genetic overlap is always significant, except for one of the drugs (Palbociclib, see **Table 1**).

We confirm the significance by applying the same approach to breast cancer, showing that Ceritinib, Crizotinib and Palbociclib produce a significant genetic overlap, although the number of common genes in each case is substantially lower than it is in NSCLC (see **Table 1**). These results are consistent with the fact that Palbociclib is primarily indicated for breast cancer and it has been recently repurposed for NSCLC (38).

The small but significant overlap of Ceritinib and Crizotinib suggests that these two drugs might also be considered as potential repurposing candidates. We note that using the top-ranking nodes increases the significance of the genetic overlap (with lower P-values) compared to the overlap using only seeds (genes associated with a pathophenotype and direct targets of drugs). The significant overlap between the top ranked genes identified using these drugs and the top ranked genes for NSCLC (but not for the top ranked genes for breast cancer) suggests that GUILDify v2.0 can help understanding how drugs exert their action on certain diseases. Indeed, the characterization of the neighborhood in the PIN that is affected by drugs opens a wide range of possibilities for drug repurposing research.

Table 1. Results of the genetic and functional overlap between the subnetwork of genes associated with “non small cell lung carcinoma” and “breast cancer” (top ranking genes and seeds) and the subnetwork of genes associated with the targets of drugs Afatinib, Ceritinib, Crizotinib, Erlotinib, Gefitinib and Palbociclib (drug targets and top-ranking genes obtained with GUILDify v2.0). P-values shown have been corrected using the Benjamini-Hochberg correction for multiple tests. Results with non-significant P-value are highlighted in red.



	“non-small cell lung carcinoma”								“breast cancer”							
	Genetic overlap				Functional overlap				Genetic overlap				Functional overlap			
	Top		Seeds		Top		Seeds		Top		Seeds		Top		Seeds	
	Nº	P-val.	Nº	P-val.	Nº	P-val.	Nº	P-val.	Nº	P-val.	Nº	P-val.	Nº	P-val.	Nº	P-val.
Afatinib	9	7.80E-06	2	1.90E-03	0	1	6	1.70E-05	3	1.40E-01	0	1	0	1	1	7.50E-01
Ceritinib	18	4.20E-15	4	6.60E-05	5	1.70E-07	9	1.80E-06	6	1.20E-02	0	1	0	1	0	1
Crizotinib	13	1.20E-09	4	7.00E-05	3	1.10E-04	8	7.70E-06	5	2.00E-02	0	1	0	1	0	1
Erlotinib	16	6.30E-13	4	6.60E-05	5	1.20E-06	10	2.60E-07	3	1.40E-01	0	1	0	1	0	1
Gefitinib	10	1.10E-06	3	1.20E-04	1	3.60E-02	11	7.20E-10	3	1.40E-01	0	1	0	1	0	1
Palbociclib	3	1.40E-01	0	1	0	1	2	8.90E-02	5	2.00E-02	1	4.70E-01	0	1	1	7.50E-01

Methods

Datasets

GUILDify v2.0 uses BIANA (31) for the integration of biological interaction databases with information on drugs, genes, proteins, functions, pathways and diseases. To create the tissue-specific PINs, we use the RNAseq data from GTEx V7 (22). Phenotype-gene associations are extracted from DisGeNET, OMIM, Uniprot, and Gene Ontology. Drug-target associations are taken from DrugBank (17), DGIdb (18), DrugCentral (19) and ChEMBL (20). See **Supplementary Material** for details on the datasets.

Prioritization algorithms

GUILDify v2.0 uses four different network-based prioritization algorithms: NetShort, NetZcore, NetScore and DIAMOnD. For details on these algorithms see references (5,10,11) and the **Supplementary Material**.

Overlap and functional enrichment analysis

We use one-sided Fisher's exact test to calculate the overlap between two sets of genes or functions and use Benjamini-Hochberg multiple hypothesis testing procedure (where applicable). The functions enriched among seeds and top-ranking nodes as well as common functions between two diseases are calculated as explained in a previous work (12) (see details in **Supplementary Material**).

Conflicts of Interest Statement

None declared.

Acknowledgements

J.A.P. and E.G. would like to acknowledge the technical support from GRIB IT team, in particular A. Gonzalez Pauner and M. A. Sánchez Gómez.

The authors received support from: ISCIII-FEDER (PI13/00082, CP10/00524, CPII16/00026); IMI-JU under grants agreements no. 116030 (TransQST) and no. 777365 (eTRANSAFE), resources of which are composed of financial contribution from the EU-FP7 (FP7/2007- 2013) and EFPIA companies in kind contribution; the EU H2020 Programme 2014-2020 under grant agreements no. 634143 (MedBioinformatics) and no. 676559 (Elixir-Excelerate); the Spanish Ministry of Economy (MINECO) [BIO2017-85329-R] [RYC-2015-17519]; "Unidad de Excelencia María de Maeztu", funded by the Spanish Ministry of Economy [ref: MDM-2014-0370]. The Research Programme on Biomedical Informatics (GRIB) is a member of the Spanish National Bioinformatics Institute (INB), PRB2-ISCIII and is supported by grant PT13/0001/0023, of the PE I+D+i 2013-2016, funded by ISCIII and FEDER.

References

1. Wangler MF, Yamamoto S, Chao H-T, Posey JE, Westerfield M, Postlethwait J, et al. Model Organisms Facilitate Rare Disease Diagnosis and Therapeutic Research. Genetics. 2017;207(1):9–27.
2. Bromberg Y. Chapter 15: Disease Gene Prioritization. PLoS Computational Biology. 2013;9(4):e1002902.
3. Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, Loscalzo J, et al. Uncovering disease-disease relationships through the incomplete interactome. Science. 2014;347(6224):1257601-1-1257601–8.
4. Wang X, Gulbahce N, Yu H. Network-based methods for human disease gene prediction. Briefings in Functional Genomics. 2011;10(5):280–93.
5. Ghiassian SD, Menche J, Barabási AL. 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):e1004120.
6. Nitsch D, Tranchevent LC, Gonalves JP, Vogt JK, Madeira SC, Moreau Y. PINTA: A web server for network-based gene prioritization from expression data. Nucleic Acids Research. 2011;39(SUPPL. 2):334–8.
7. Zuberi K, Franz M, Rodriguez H, Montojo J, Lopes CT, Bader GD, et al. GeneMANIA prediction server 2013 update. Nucleic acids research. 2013;41(Web Server issue):115–22.
8. Gottlieb A, Magger O, Berman I, Ruppin E, Sharan R. Principle: A tool for associating genes with diseases via network propagation. Bioinformatics. 2011;27(23):3325–6.
9. Kacprowski T, Doncheva NT, Albrecht M. NetworkPrioritizer: A versatile tool for network-based prioritization of candidate disease genes or other molecules. Bioinformatics. 2013;29(11):1471–3.

10. Guney E, García-garcía J, Oliva B. GUILDify : A web server for phenotypic characterization of genes through biological data integration and network-based prioritization algorithms. Bioinformatics. 2014;30(12):1789–90.

11. Guney E, Oliva B. Exploiting Protein-Protein Interaction Networks for Genome-Wide Disease-Gene Prioritization. PLoS ONE. 2012;7(9):e43557.

12. Rubio-Perez C, Guney E, Aguilar D, Piñero J, Garcia-Garcia J, Iadarola B, et al. Genetic and functional characterization of disease associations explains comorbidity. Scientific Reports. 2017;7(1):6207.

13. Halakou F, Kilic E Sen, Cukuroglu E, Keskin O, Gursoy A. Enriching Traditional Protein-protein Interaction Networks with Alternative Conformations of Proteins. Scientific Reports. 2017;7(1):7180.

14. Kominakis A, Hager-Theodorides AL, Zoidis E, Saridaki A, Antonakos G, Tsiamis G. Combined GWAS and ‘guilt by association’-based prioritization analysis identifies functional candidate genes for body size in sheep. Genetics Selection Evolution. 2017;49(1):41.

15. Tejera E, Cruz-Monteagudo M, Burgos G, Sánchez ME, Sánchez-Rodríguez A, Pérez-Castillo Y, et al. Consensus strategy in genes prioritization and combined bioinformatics analysis for preeclampsia pathogenesis. BMC Medical Genomics. 2017;10(1):50.

16. Aguirre-Plans J, Piñero J, Menche J, Sanz F, Furlong LI, Schmidt HHHW, et al. Proximal pathway enrichment analysis for targeting comorbid diseases via network endopharmacology. Pharmaceuticals. 2018;11(3):61.

17. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: A major update to the DrugBank database for 2018. Nucleic Acids Research. 2018;46(D1):D1074–82.

18. Cotto KC, Wagner AH, Feng Y, Kiwala S, Coffman C, Spies G, et al. DGIdb 3.0 : a redesign and expansion of the drug-gene interaction database. Nucleic Acids Research. 2018;46(November 2017):1068–73.

19. Ursu O, Holmes J, Knockel J, Bologa CG, Yang JJ, Mathias SL, et al. DrugCentral : online drug compendium. Nucleic Acids Research. 2017;45(October 2016):932–9.
20. Gaulton A, Hersey A, Patr A, Chambers J, Mendez D, Mutowo P, et al. The ChEMBL database in 2017. Nucleic Acids Research. 2017;45(November 2016):945–54.
21. Kitsak M, Sharma A, Menche J, Guney E, Ghiassian SD, Loscalzo J, et al. Tissue Specificity of Human Disease Module. Scientific Reports. 2016;6(October):35241.
22. Consortium G. Genetic effects on gene expression across human tissues. Nature. 2017;550(7675):204–13.
23. Piñero J, Bravo Á, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, Centeno E, et al. DisGeNET: A comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Research. 2017;45(D1):D833–9.
24. The UniProt Consortium. UniProt: The universal protein knowledgebase. Nucleic Acids Research. 2017;45(D1):D158–69.
25. Davis AP, Grondin CJ, Johnson RJ, Sciaky D, King BL, McMorran R, et al. The Comparative Toxicogenomics Database: Update 2017. Nucleic Acids Research. 2017;45(D1):D972–8.
26. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: Public archive of interpretations of clinically relevant variants. Nucleic Acids Research. 2016;44(D1):D862–8.
27. Rath A, Olry A, Dhombres F, Brandt MM, Urbero B, Ayme S. Representation of rare diseases in health information systems: The orphanet approach to serve a wide range of end users. Human Mutation. 2012;33(5):803–8.
28. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Research. 2017;45(D1):D896–901.
29. Gutiérrez-Sacristán A, Bravo À, Portero M, Valverde O, Armario A, Blanco-Gandía MC, et al. Text mining and expert curation to develop

a database on psychiatric diseases and their genes. Database. 2017;1650:48–55.

30. Köhler S, Vasilevsky NA, Engelstad M, Foster E, McMurry J, Aymé S, et al. The human phenotype ontology in 2017. Nucleic Acids Research. 2017;45(D1):D865–76.

31. Garcia-Garcia J, Guney E, Aragues R, Planas-Iglesias J, Oliva B. Biana: a software framework for compiling biological interactions and analyzing networks. BMC Bioinformatics. 2010;11(1):56.

32. Kibbe WA, Arze C, Felix V, Mitraka E, Bolton E, Fu G, et al. Disease Ontology 2015 update : an expanded and updated database of human diseases for linking biomedical knowledge through disease data. Nucleic Acids Research. 2015;43(October 2014):1071–8.

33. Franz M, Lopes CT, Huck G, Dong Y, Sumer O, Bader GD. Cytoscape.js: A graph theory library for visualisation and analysis. Bioinformatics. 2015;32(2):309–11.

34. Rolfes MC, Juhn YJ, Wi SI, Sheen YH. Asthma and the risk of rheumatoid arthritis: An insight into the heterogeneity and phenotypes of asthma. Tuberculosis and Respiratory Diseases. 2017;80(2):113–35.

35. Sharma A, Menche J, Chris Huang C, Ort T, Zhou X, Kitsak M, et al. A disease module in the interactome explains disease heterogeneity, drug response and captures novel pathways and genes in asthma. Human Molecular Genetics. 2014;24(11):3005–20.

36. Bethune G, Bethune D, Ridgway N, Xu Z. Epidermal growth factor receptor (EGFR) in lung cancer: an overview and update. Journal of Thoracic Disease. 2010;2(1):48–51.

37. Rubio-Perez C, Tamborero D, Schroeder MP, Antolín AA, Deu-Pons J, Perez-Llamas C, et al. In Silico Prescription of Anticancer Drugs to Cohorts of 28 Tumor Types Reveals Targeting Opportunities. Cancer Cell. 2015;27(3):382–96.

38. Zhou J, Zhang S, Chen X, Zheng X, Yao Y, Lu G, et al. Palbociclib, a selective CDK4/6 inhibitor, enhances the effect of selumetinib in

RAS-driven non-small cell lung cancer. Cancer Letters. 2017;408:130–7.

Supplementary material

1. Datasets

1.1. Protein-protein interaction networks

GUILDify v2.0 relies on a knowledge database called BIANA (1), which integrates biological interaction databases together with information on genes and proteins and its associated functions, diseases and phenotypes. Currently, we have up-to-date information of protein-protein interactions from IntAct (2), BioGRID (3) and DIP (4). As an additional option, we provide the user with 5 other PIN sources: HIPPIE (high confidence score threshold >= 0.7) (5), InBio_Map (score threshold >= 0.15) (6), ConsensusPathDB (7), I2D (8), and STRING (score >= 0.7) (9).

1.2. Tissue-specific protein-protein interaction networks

To create the tissue-specific PINs, we retrieved the RNA-sequencing gene TPMs from GTEx V7 (10). We use the samples from subjects for which the reason for death was traumatic injury (point 1 in Hardy Scale) and we discard the tissues with less than 5 samples (a total of 675 samples from 40 tissues). For each gene, we calculate the median expression of all samples of a tissue. We unify tissues into a unique “main” tissue (i.e. “Adipose – Subcutaneous” and “Adipose – Visceral Omentum” belong to the main tissue “Adipose”) by considering the highest median expression of all samples (11). We note that using this approach, the final tissue profiles could be biased towards the subtypes of tissue that have a higher number of samples. This is a limitation, as there may be subtypes of tissue that are more represented than others in the final expression of the tissue (Supplementary Table S1). GUILDify v2.0 includes a total of 22 tissues (see details in Supplementary Material).

1.3. Phenotype-gene associations

Phenotype-gene associations are extracted from DisGeNET, OMIM, Uniprot and Gene Ontology extracting information from relevant sections. In the case of DisGeNET, we parse disease-gene associations from curated sources: UniProt (12), CTD (13), ORPHANET (14), PsyGeNET (15) and HPO (16). In OMIM (17), we retrieve disease-gene associations from the OMIM's Synopsis of the Human Gene Map. For Uniprot [30], we collect the protein information of the categories "Description", "Function", "Keyword" and "Disease". Finally, in the case of the Gene Ontology (GO) (18,19), we parse the functional annotations of genes.

1.4. Drug-target integration

Drug-target interactions are retrieved from DrugBank (20), DGIdb (21), DrugCentral (22) and ChEMBL (23), and integrated following the procedure in *Piñero et al.* (24). In DrugBank, we only select the therapeutic targets (excluding enzymes, transporters and carriers). In the case of data from DrugCentral, we retrieve the targets in the "Tclin" category. From DGIdb we select the targets from "Chembl", "GuideToPharmacology", "Tdg Clinical Trial", "FDA", "TEND" and

“TTD”. Finally in ChEMBL, we collect targets with a DrugBank identifier cross-reference.

2. Prioritisation algorithms

GUILDify v2.0 uses four different network-based prioritisation algorithms: NetShort, NetZcore, NetScore and DIAMOnD. For details on these algorithms see references (25–27). In brief, **NetShort** (10-20 minutes of computation time) incorporates “phenotypic-relevance” of the path between a node and the nodes of a given phenotype by considering the number of edges to phenotype-associated nodes (seeds). **NetZcore** (5-10 minutes of computation time) iteratively assesses the relevance of a node for a given phenotype by averaging the normalised scores of the neighbours. **NetScore** (5-10 minutes of computation time) is based on the propagation of information through the nodes of the network by considering multiple shortest paths from the source of information to the target. **NetCombo** (10-20 minutes of computation time) combines NetScore, NetShort and NetZcore by calculating the mean of the normalised score of each prioritisation method. **DIAMOnD** (27) (5 minutes of computation time) determines the “connectivity significance” of all the proteins of the network, iteratively ranking and selecting the nodes with highest scores.

3. Functional enrichment analysis

We calculate the enriched functions of seeds and top-ranking nodes and calculate the significance of common functions between two diseases (or phenotypes) as in a previous work on comorbidities (28). Briefly, functions are defined by GO biological processes, GO molecular functions and Reactome pathways. In the case of GO, we only use high confident annotations (codes of evidence EXP, IDA, IMP, IGI, IEP, ISS, ISA, ISM or ISO). We calculate the significance of the enrichment using a one-sided Fisher's exact test (the alternative hypothesis is that the overlap would be greater than observed overlap). Then, we correct the P-value by either applying the Benjamini-Hochberg correction for multiple tests and keeping the functions for which the adjusted P-value < 0.05. We also offer the user the possibility to use Bonferroni correction at the results page.

4. Description of BIANA integration pipeline

We used BIANA (1) to compile different types of biological data in an integrated database and to create the protein-protein interaction networks (PIN). The information in BIANA is updated annually to keep the resources underlying the web server up-to-date. The details of data retrieval, integration, unification and network generation pipeline are as follows:

1. Download the data:

We use five sources of protein-protein interaction data:

- **IntAct**: retrieved from https://www.proteinatlas.org/download/normal_tissue.tsv.zip (Release of 22-Mar-2018).
- **BioGRID**: downloaded from https://downloads.thebiogrid.org/BioGRID (Version 3.4.159).
- **DIP**: downloaded from http://dip.doe-mbi.ucla.edu/dip/Download.cgi (Release of 05-Feb-2017).

- **iRefIndex**: downloaded from http://irefindex.org/download/irefindex/data/archive/release_14.0/psi_mitab/MITAB2.6/ (Version 15.0).
- **HIPPIE**: downloaded from http://cbdm-01.zdv.uni-mainz.de/~mschaefer/hippie/download.php (Version 2.1).

And we also incorporate additional databases to complement interactomics data:

- **UniProt swissprot**: retrieved from ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/ (Release of 28-Mar-2018).
- **Taxonomy**: downloaded from ftp://ftp.ncbi.nih.gov/pub/taxonomy (Release of 19-Apr-2018).
- **Gene Ontology**: downloaded from http://www.geneontology.org/ontology/ (Release of 19-Apr-2018).
- **NCBI Gene**: downloaded from ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2ensembl.gz (Release of 28-Nov-2017).
- **DisGeNET**: downloaded from http://www.disgenet.org/web/DisGeNET/menu/downloads (Version 5.0).
- **DrugBank**: downloaded from https://www.drugbank.ca/releases/latest (Version 5.1.0).
- **DrugCentral**: downloaded from http://drugcentral.org/download (Release of 29-Aug-2017).
- **DGIdb**: downloaded from http://dgidb.org/downloads (Version 3.0.2).
- **ChEMBL**: (Version ChEMBL_24) downloaded from:

  - https://www.ebi.ac.uk/chembl/drug/targets > Downloads > Download all txt
  - https://www.ebi.ac.uk/chembl/drugstore > Downloads > Download all txt
  - https://www.ebi.ac.uk/chembl/drug/indications > Downloads > Download all txt
  - https://www.ebi.ac.uk/chembl/target/browser > Select All > Fetch selected targets > Please select… > Download All (tab-delimited)

2. Parse and unify the data:

We parse all the external databases according to the manual of BIANA (available at http://sbi.imim.es/web/BIANA.php). Once we have all the databases incorporated in BIANA knowledge database, we find the equivalent entries across databases to unify the data. This means that two entities coming from different databases (or in some cases from the same database) can be unified in a unique entity provided that they satisfy certain equivalence criteria. If they do so, they will be given the same unique ID called **BIANA ID**. The rules to unify data (equivalence criteria) are the following:

- Same **Entrez Gene ID** (applied to all databases)
- Same **Taxonomy ID AND protein sequence** (applied to all databases)
- Same **UniProt entry** (only applied to ConsensusPathDB and Uniprot databases)
- Same **UniProt accession** (applied to InBio_Map, I2D, HitPredict and Uniprot)
- Same **UniProt accession** (applied to DrugBank, DrugCentral, ChEMBL and Uniprot databases, to unify the drug targets with the rest of proteins)

- Same **DrugBank ID** (applied to DrugBank, DCDB and DrugCentral to unify drugs)
- Same **PubChem Compound** (applied to DrugBank and DCDB to unify drugs)
- Same **ChEMBL ID** (applied to DrugBank, DGIdb and ChEMBL to unify drugs)

3. Generate protein-protein interaction networks (PIN):

To generate PINs, we retrieve the protein-protein interactions from BIANA knowledge database that have the same Taxonomy ID (reported in the same organism) for human, mouse, rat, yeast, worm, fly and plant.

Once we have the PIN, we filter the interactions depending on the detection method used to characterise each interaction. Recent studies highlighted that several protein interaction detection techniques tended to provide a higher number of interactions for more studied proteins in the interactome (29,30). Thus, for human, we only include interactions coming from the following detection methods that we consider to be less biased:

- Two hybrid (ID: 18).
- Cross-linking study (ID: 30).
- Protein array (ID: 89).
- Two hybrid array (ID: 397).
- Two hybrid pooling approach (ID: 398).
- Biochemical (ID: 401).
- Enzymatic study (ID: 415).
- Two hybrid prey pooling approach (ID: 1112).
- Proximity labelling technology (ID: 1313).
- Validated two hybrid (ID: 1356).

For mouse, we included all the methods listed above and the following ones:

- Affinity chromatography technology (ID: 4)
- Anti tag coimmunoprecipitation (ID: 7)
- Coimmunoprecipitation (ID: 19)
- Cosedimentation in solution (ID: 28)
- Pull down (ID: 96)
- X-ray crystallography (ID: 375)
- Chromatin immunoprecipitation assay (ID: 810)
- Tandem affinity purification (ID: 676)

5. Description of tissue-specific PIN generation pipeline

1. **Download the data:** We used the RNA-Seq data from GTEx Portal version 7, downloaded from https://gtexportal.org/home/datasets.
2. **Process the data:**
   - Process the subjects file (GTEx_v7_Annotations_SubjectPhenotypesDS.txt) and get the subjects with cause of death by traumatic injury (DTHHRDY=1). We end up with 29 subjects.
   - Process the samples file (GTEx_v7_Annotations_SampleAttributesDS.txt) and get the samples coming from the subjects filtered in the previous step. We end up with 699 samples.
   - Count the tissues present in the samples and the number of samples for each tissue. Remove the tissues that have less than 5 samples. We end up having 40 tissues and 675 samples.

- Read the TPM file (GTEx_Analysis_2016-01-15_v7_RNASeQCv1.1.8_gene_tpm.gct.gz) and get the TPM values of the 675 samples mentioned in the previous step.
- For each gene in the TPM file, calculate the median of TPM values across all the samples in each tissue.
- If several tissues belong to a more general tissue (i.e. “Adipose – Subcutaneous” and “Adipose – Visceral Omentum” belong to the main tissue “Adipose”), we unify them by considering the highest median expression value among these tissues. After unifying the tissues, we end up having 22 main tissues. The tissues unified are listed in Supplementary Table S1.

3. **Filter the PIN:** The last step is to filter the interactions of the PIN based on the tissue annotation of the proteins. For each pair of interacting proteins, first, we check if we have information of the proteins in the GTEx file that we have processed. If so, we get the TPM values for the tissue of interest and the two proteins, and if in both cases the TPM values are higher or equal than 1, we maintain the interaction in the network. If not, we remove the interaction.

Supplementary Table S1. Tissues considered for the creation of tissue-specific PIN. In the left we show the 40 initial tissues and in the right the 22 final tissues after the unification of some of the tissues into a broader one. We included the number of samples considered for each tissue.

GTEx tissues		**Unified tissues**	
Name	**Num. samples**	**Name**	**Num. samples**
Adipose – Subcutaneous	18	Adipose	28
Adipose – Visceral (Omentum)	10		
Adrenal Gland	5	Adrenal Gland	5
Artery – Aorta	16	Artery	52
Artery – Coronary	9		
Artery – Tibial	27		
Brain – Amygdala	10	Brain	144
Brain – Anterior cingulate cortex (BA24)	12		
Brain – Caudate (basal ganglia)	13		
Brain – Cerebellar Hemisphere	13		
Brain – Cerebellum	16		
Brain – Cortex	14		
Brain – Frontal Cortex (BA9)	8		
Brain – Hippocampus	13		
Brain – Hypothalamus	11		
Brain – Nucleus accumbens (basal ganglia)	12		
Brain – Putamen (basal ganglia)	12		
Brain – Spinal cord (cervical c-1)	10		
Breast – Mammary Tissue	16	Breast – Mammary Tissue	16

Cells – Transformed fibroblasts	20	Cells – Transformed fibroblasts	20
Colon – Sigmoid	8	Colon – Sigmoid	8
Esophagus – Gastroesophageal Junction	8	Esophagus	36
Esophagus – Mucosa	15		
Esophagus – Muscularis	13		
Heart – Atrial Appendage	14	Heart	39
Heart – Left Ventricle	25		
Liver	11	Liver	11
Lung	25	Lung	25
Muscle – Skeletal	31	Muscle – Skeletal	31
Nerve – Tibial	26	Nerve – Tibial	26
Ovary	5	Ovary	5
Pituitary	14	Pituitary	14
Prostate	10	Prostate	10
Skin – Not Sun Exposed (Suprapubic)	16	Skin	44
Skin – Sun Exposed (Lower leg)	28		
Testis	15	Testis	15
Thyroid	25	Thyroid	25
Uterus	5	Uterus	5
Vagina	6	Vagina	6
Whole Blood	110	Whole Blood	110

6. Functional-based selection of top-ranking genes

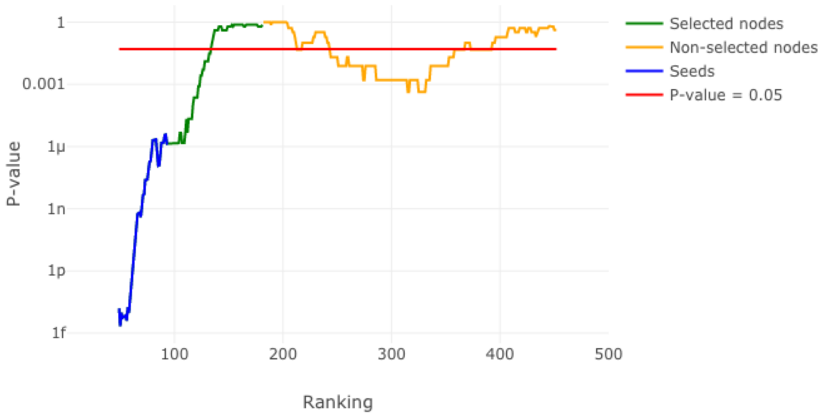
The functional-based selection of top-ranking genes is a procedure that we followed to identify if the set of ranking genes is functionally similar to the set of initial seed genes. The procedure was first implemented in *Ghiassian et al*. (27).

First, we find the enriched Gene Ontology (GO) terms in the seeds by calculating the functional enrichment following the procedure in Rubio-Perez, et al (28). We only use high confidence annotations from Biological Processes or Molecular Functions associated with the evidence codes EXP, IDA, IMP, IGI, IEP, ISS, ISA, ISM or ISO. From the enriched GO terms, we identify the ones that are significantly enriched using a one-sided Fisher's exact test of significance where the alternative hypothesis is that the overlap would be greater than observed overlap. We correct the significance applying either a Bonferroni or a Benjamini-Hochberg correction for multiple tests and selecting a P-value < 0.05 (the case studies use Benjamini-Hochberg).

For each candidate gene in the top-ranking genes, we search if it is annotated within any of the significant GO terms of the seeds. The genes annotated are considered true positives.

For each candidate gene in the ranking, we define a sliding window with a size corresponding to the number of seeds. For instance, if there are 66 seeds, the interval for top ranking node i will be [i-66/2, i+66/2]. We calculate the number of true positives among the proteins in the sliding window. We calculate the statistical significance in the sliding window by using a Fisher’s exact test.

In the end, we obtain a plot that goes from the first position of the sliding window (ranking = # of seeds / 2 + 1) to the final position (500 - # of seeds/2). In each position, we show the result of the Fisher’s test calculation for the positions of the sliding window. We consider as *enriched positions* all the positions until the last sliding window giving a P-value < 0.05.



Supplementary Figure S1. Functional-based selection plot in the case of the example GUILDify v2.0 run using 96 seeds for *asthma* keyword and the NetScore algorithm with default parameters.

In the figure S1, we observe an example of the functional-based selection plot for *asthma* using 96 seeds. Therefore, the plot starts at position 49 and ends at position 452. The last sliding window with a P-value < 0.05 is at position 133, and ranges from position 85 to 181. We consider as enriched positions all the top-ranking genes until the 181st position, including the 96 initial seeds and 85 additional non-seed genes.

7. Significance of the overlap between genes/functions

To calculate the significance of the overlap between top-ranking genes and their functions, we use a one-sided (the alternative hypothesis is that the odds ratio based on the overlap is greater than the observed odds ratio) Fisher's exact test with the contingency table given in Supplementary Table S2.

Supplementary Table S2. Contingency table used to calculate the significance of the overlap between top-ranking genes and their functions.

	Top 2	**Non-top 2**
Top 1	Nº common	Nº top 1 – Nº common
Non-top 1	Nº top 2 – Nº common	Nº total – Nº top 1 – Nº top 2 – Nº common

Supplementary Table S3 is an example of the contingency table for the genetic overlap between asthma and rheumatoid arthritis, where we obtain an Odds Ratio of 23.542 and a P-value of 5.9*10^{-48}.

Supplementary Table S3. Contingency table used in Fisher's exact text for the genetic overlap between asthma and rheumatoid arthritis.

	Top genes 2	**Non-top genes 2**
Top genes 1	55	290 – 55 = 235
Non-top genes 1	181 – 55 = 126	13,090 – 181 – 290 + 55 = 12,674

The contingency table for the functional overlap of the same example, where we obtain an Odds Ratio of 155.334 and a P-value of 1.3*10^{-58} is below (Supplementary Table S4).

Supplementary Table S4. Contingency table used in Fisher's exact text for the functional overlap between asthma and rheumatoid arthritis

	Top functions 2	**Non-top functions 2**
Top functions 1	38	84 – 38 = 46
Non-top functions 1	94 – 38 = 56	10,670 – 94 – 84 + 38 = 10,530

Supplementary Table S5. Job IDs to access to the case studies in the web server and parameters used.

Keyword	Parameters	Job ID
asthma	Seeds: all genes Organism: Homo sapiens Tissue: All Network: BIANA Method: NetScore (nRepetition=3, nIteration=2)	**asthma**
"rheumatoid arthritis"	Seeds: all genes Organism: Homo sapiens Tissue: All Network: BIANA Method: NetScore (nRepetition=3, nIteration=2)	**rheumatoid_arthritis**
"non small cell lung carcinoma"	Seeds: all genes Organism: Homo sapiens Tissue: All Network: BIANA Method: NetScore (nRepetition=3, nIteration=2)	**non_small_cell_lung_carcinoma**
"breast cancer"	Seeds: all genes Organism: Homo sapiens Tissue: All Network: BIANA Method: NetScore (nRepetition=3, nIteration=2)	**breast_cancer**
"breast cancer"	Seeds: DisGeNET and OMIM Organism: Homo sapiens Tissue: All Network: BIANA Method: NetScore (nRepetition=3, nIteration=2)	**breast_cancer_omim_disgenet**

afatinib	Seeds: all genes Organism: Homo sapiens Tissue: All Network: BIANA Method: NetScore (nRepetition=3, nIteration=2)	**afatinib**
ceritinib	Seeds: all genes Organism: Homo sapiens Tissue: All Network: BIANA Method: NetScore (nRepetition=3, nIteration=2)	**afatinib**
crizotinib	Seeds: all genes Organism: Homo sapiens Tissue: All Network: BIANA Method: NetScore (nRepetition=3, nIteration=2)	**crizotinib**
erlotinib	Seeds: all genes Organism: Homo sapiens Tissue: All Network: BIANA Method: NetScore (nRepetition=3, nIteration=2)	**erlotinib**
gefitinib	Seeds: all genes Organism: Homo sapiens Tissue: All Network: BIANA Method: NetScore (nRepetition=3, nIteration=2)	**gefitinib**
palbociclib	Seeds: all genes Organism: Homo sapiens Tissue: All Network: BIANA Method: NetScore (nRepetition=3, nIteration=2)	**palbociclib**

8. Results of the case study of asthma and arthritis rheumatoid and the negative controls with breast cancer

We query “asthma” in GUILDify v2.0, obtaining 96 seeds. After the running GUILD and selecting the functionally-coherent top genes, we obtain 181 genes conforming the neighbourhood of asthma. We do the same for “rheumatoid arthritis”, retrieving 158 seeds and creating a neighbourhood of 290 functionally-coherent top genes. Between the top ranking genes of the two phenotypes there are 55 common genes (Fisher’s exact test, one-sided P-value = $5.9 \cdot 10^{-48}$) which is more significant than the 12 common genes between the seeds (P-value = $1.4 \cdot 10^{-9}$). When removing the seeds from the top ranking genes of the two phenotypes, we find an overlap of 43 genes which is even more significant than before (P-value = $3.7 \cdot 10^{-65}$).

If we focus on the functional overlap, we find 38 common enriched functions from the top ranking genes (P-value = $1.3 \cdot 10^{-58}$), 18 common enriched functions from the seeds (P-value = $1.7 \cdot 10^{-24}$), and 24 common functions removing the seed-functions from the top-functions (P-value = $3.5 \cdot 10^{-47}$).

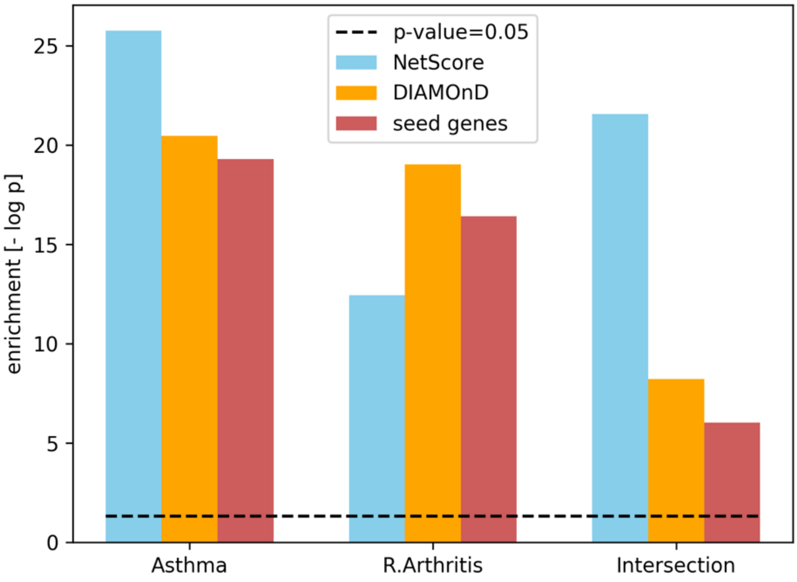
To have a negative control, we query “breast cancer” and retrieve 119 seeds. We select the 182 functionally-coherent top genes and compare the neighbourhood with the phenotypes of asthma and rheumatoid arthritis. In the case of asthma and breast cancer, we observe a significant overlap between 8 genes (P-value = $3.8 \cdot 10^{-3}$), but the functional overlap is not significant (only 1 common function). In the case of rheumatoid arthritis and breast cancer the overlap is not significant. It is important to remark that 101 of the 119 breast cancer seeds are from Uniprot and may not be as much reliable. We repeated the same analysis selecting only 28 seeds from OMIM and DisGeNET, and in general the results of the overlap are not significant neither for asthma nor rheumatoid arthritis. The results can be explored with more detail in Supplementary Table S6.

Supplementary Table S6. Results of the genetic and functional overlap between the subnetwork of genes associated to asthma and rheumatoid arthritis, asthma and breast cancer, and rheumatoid arthritis and breast cancer. Breast cancer has been calculated either using DisGeNET and OMIM seeds (D+O) or using all the seeds (all). P-values have been corrected using the Benjamini-Hochberg correction for multiple tests. Results with non-significant P-value are highlighted in red.

	Genetic overlap						**Functional overlap**					
	Top		**Top without seeds**		**Seeds**		**Top**		**Top without seeds**		**Seeds**	
	Nº	**P-val.**	**Nº**	**P-val.**	**Nº**	**P-val.**	**Nº**	**P-value**	**Nº**	**P-value**	**Nº**	**P-value**
Asthma – Rheumatoid arthritis	55	2.90E-47	43	1.80E-64	12	7.00E-09	31	4.00E-45	18	5.50E-34	18	4.60E-22
Asthma – Breast cancer (all)	8	9.50E-03	5	1.30E-04	3	9.50E-02	1	2.10E-01	1	1.40E-02	0	1
Rheumatoid arthritis – Breast cancer (all)	4	7.20E-01	2	2.30E-01	2	5.20E-01	0	1	0	1	0	1
Asthma – Breast cancer (D+O)	2	2.30E-01	0	1	2	4.50E-02	0	1	0	1	0	1
Rheumatoid arthritis – Breast cancer (D+O)	0	1	0	1	0	1	0	1	0	1	1	1

9. Comparison of the case study results of asthma and rheumatoid arthritis with DIAMOnD

We have compared the functional enrichment of top-ranking genes identified by NetScore and DIAMOnD in asthma and rheumatoid arthritis. We based the analysis in *Sharma et al*., where the authors present a comparison of DIAMOnD and several other prioritisation algorithms, showing that DIAMOnD outperforms existing algorithms in predicting asthma related genes (31). Following the procedure described in the original article, for each seed-gene, we determine the set of MSIgDB pathways (32) associated with the gene. For each pathway, we analyse its enrichment among the set of seed-genes using Fisher's exact test. The p-values are corrected using the Benjamini-Hochberg correction procedure for multiple tests. We choose the pathways with a significance level of p-value<0.01 as being associated with the set of seed-genes (enriched pathways). We calculate a Fisher's exact test between the top-ranking genes (based on functional-coherency) of the algorithm under analysis and the genes of the enriched pathways. The p-value of the Fisher's exact test gives the enrichment of the top-ranking genes. The enrichment of the pathways using top-ranking genes from NetScore and DIAMOnD as well as using the original set of seed-genes are shown in Supplementary Figure S2. When measuring the overlap between the two diseases, NetScore outperforms DIAMOnD, finding more genes involved in the pathways enriched in both diseases.



Supplementary Figure S2. Bar plot showing the enrichment of the top-ranking genes in terms of -log p-value.

10. Screening diseases to identify potential new indications of known drugs

GUILDify v2.0 introduces a “Drug Repurposing” functionality that can be accessed from the home page of the web server. This functionality takes a job ID as input, i.e. results for a drug (or a disease) and screens across a set of pre-calculated diseases (or drugs) for the significance of the overlap of genes and functions between the given job ID and the set of pre-calculated diseases (or drugs). The generation of the pre-calculated sets and the validation of the drug repurposing approach using these sets are explained below.

(1) **Set of pre-calculated diseases:** We created a list of diseases using the UMLS concept unique identifiers from DisGeNET. Specifically, we obtained all diseases with gene associations reported by curated sources (UniProt, ORPHANET, PsyGeNET and HPO). We did not include CTD because it provides several clinically ambiguous phenotypes such as “liver cirrhosis, experimental”. From this list, we selected those diseases associated to at least 10 gene from DisGeNET, obtaining a final list of 757 diseases. We ran the prioritisation using the guildifyR package and the default parameters (BIANA network, and NetScore algorithm).

(2) **Set of pre-calculated drugs:** The list of drugs was obtained by retrieving all drugs with targets stored in BIANA knowledge database. Out of this list, we selected those drugs with at least 10 targets (retrieved from at least one of the following databases in BIANA: DrugBank, ChEMBL, DGIdb, DrugCentral), obtaining a final list of 362 drugs. We run the prioritisation using the R package and default parameters (BIANA network, NetScore algorithm).

(3) **Set of drug-disease indications:** We tested the quality of the predictions of indications of drugs using as benchmark the indications of Hetionet (33). Accordingly, we used 161 drugs and 64 diseases that appeared in both Hetionet and the lists of pre-calculated drugs and diseases, producing a final set of 329 drug-disease pairs with known indications.

(4) **Finding the indication among the top-ranked results:** We plotted a histogram of the correct indications among the top

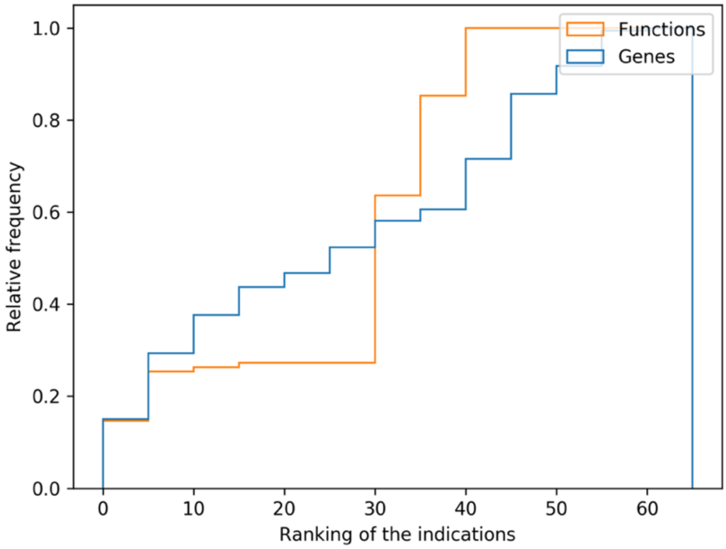
ranked (see Supplementary Figure S3). This showed that 30% of the correct indications were already among the top 10 indications selected (and 50% of correct indications appeared among the top 25 predicted indications).

We calculated how many disease indications could be guessed depending on the number of top-ranked indications for a drug. We transformed this calculation in True Positive Rate (TPR) and False Positive Rate (FPR), calculated as:

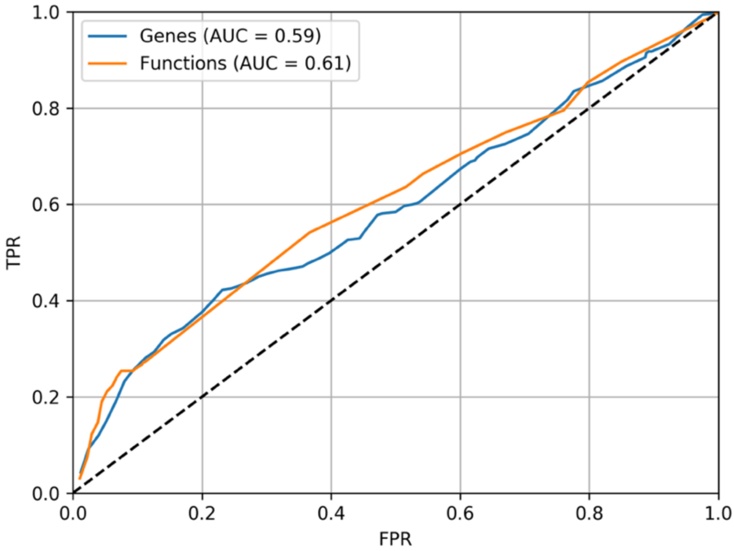
$$TPR = \frac{\#\, true\ positives}{\#\, positives} = \frac{\#\, guessed\ indications}{\#\, total\ indications}$$

$$FPR = \frac{\#\, false\ positives}{\#\, negatives} = \frac{\#\, non - indications\ in\ predictions}{\#\, non - indications}$$

Using the TPR and FPR we plotted the Receiver Operating Characteristic (ROC) curve (see Supplementary Figure S3). Using the top-scoring 1% of genes and functions, we obtained an Area Under the Curve (AUC) of 0.59 for genetic overlap and 0.61 for functional overlap.



Supplementary Figure S3. Cumulative distribution of the ranking positions achieved by the indications using the drug repurposing feature of GUILDify v2.0.



Supplementary Figure S4. Receiver Operating Characteristic (ROC) curve. The values of the Area Under the Curve (AUC) are indicated in the legend.

11. Visualization of the top-ranking subnetwork

In the results page, we provide a visualization panel to inspect in detail the interactions between the top-ranking nodes. Initially, we provide the user with an image of the subnetwork created with Matplotlib Python library (34). The users have the option to click to “activate interactive visualization”, where the top-ranking nodes and interactions are displayed in a visualization panel using the JavaScript-based network visualization library, Cytoscape.js (35). In addition to seeds (green hexagons), top-ranking proteins (yellow circles) and drugs (blue diamonds), the subnetwork includes the proteins that connect the seeds to the largest connected component induced by seeds (named “linkers” and shown as grey circles). The procedure to get the linkers is the following:

- Calculate the connected components of the top-ranking nodes.
- Check the size of the connected components and get the largest connected component (LCC). If there are more than one, we get the component with the highest mean score among all the nodes of the component.

We order the rest of the components by their size and we find the shortest paths between the LCC and the remaining components (starting from the component with the highest size) to connect them to the LCC. If there is more than one shortest path, we get the shortest path that contains the node with the maximum score. If they have the same maximum score, we get the shortest path with highest mean score between all the nodes.

Supplementary material references

1. Garcia-Garcia J, Guney E, Aragues R, Planas-Iglesias J, Oliva B. Biana: a software framework for compiling biological interactions and analyzing networks. BMC Bioinformatics. 2010;11(1):56.
2. Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, et al. The MIntAct project - IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res. 2014;42(D1):358–63.
3. Chatr-Aryamontri A, Oughtred R, Boucher L, Rust J, Chang C, Kolas NK, et al. The BioGRID interaction database: 2017 update. Nucleic Acids Res. 2017;45(D1):D369–79.
4. Salwinski L. The Database of Interacting Proteins: 2004 update. Nucleic Acids Res. 2004;32(90001):449D – 451.
5. Alanis-Lobato G, Andrade-Navarro MA, Schaefer MH. HIPPIE v2.0: Enhancing meaningfulness and reliability of protein-protein interaction networks. Nucleic Acids Res. 2017;45(D1):D408–14.
6. Li T, Wernersson R, Hansen RB, Horn H, Mercer J, Slodkowicz G, et al. A scored human protein-protein interaction network to catalyze genomic interpretation. Nat Methods. 2016;14(1):61–4.
7. Herwig R, Hardt C, Lienhard M, Kamburov A. Analyzing and interpreting genome data at the network level with ConsensusPathDB. Nat Protoc. 2016;11(10):1889–907.
8. Brown KR, Jurisica I. Unequal evolutionary conservation of human protein interactions in interologous networks. Genome Biol. 2007;8(5):R95.
9. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, et al. The STRING database in 2017 : quality-controlled protein – protein association networks , made broadly accessible. Nucleic Acids Res. 2017;45(October 2016):362–8.
10. Consortium G. Genetic effects on gene expression across human tissues. Nature. 2017;550(7675):204–13.

11. Basha O, Barshir R, Sharon M, Lerman E, Kirson BF, Hekselman I, et al. The TissueNet v.2 database: A quantitative view of protein-protein interactions across human tissues. Nucleic Acids Res. 2017;45(D1):D427–31.
12. The UniProt Consortium. UniProt: The universal protein knowledgebase. Nucleic Acids Res. 2017;45(D1):D158–69.
13. Davis AP, Grondin CJ, Johnson RJ, Sciaky D, King BL, McMorran R, et al. The Comparative Toxicogenomics Database: Update 2017. Nucleic Acids Res. 2017;45(D1):D972–8.
14. Rath A, Olry A, Dhombres F, Brandt MM, Urbero B, Ayme S. Representation of rare diseases in health information systems: The orphanet approach to serve a wide range of end users. Hum Mutat. 2012;33(5):803–8.
15. Gutiérrez-Sacristán A, Bravo À, Portero M, Valverde O, Armario A, Blanco-Gandía MC, et al. Text mining and expert curation to develop a database on psychiatric diseases and their genes. Database. 2017;1650:48–55.
16. Köhler S, Vasilevsky NA, Engelstad M, Foster E, McMurry J, Aymé S, et al. The human phenotype ontology in 2017. Nucleic Acids Res. 2017;45(D1):D865–76.
17. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an Online catalog of human genes and genetic disorders. Nucleic Acids Res. 2015;43(D1):D789–98.
18. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: Tool for the unification of biology. Vol. 25, Nature Genetics. 2000. p. 25–9.
19. The Gene Ontology Consortium. Expansion of the gene ontology knowledgebase and resources: The gene ontology consortium. Nucleic Acids Res. 2017;45(D1):D331–8.
20. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: A major update to the DrugBank database for 2018. Nucleic Acids Res. 2018;46(D1):D1074–82.

21. Cotto KC, Wagner AH, Feng Y, Kiwala S, Coffman C, Spies G, et al. DGIdb 3.0 : a redesign and expansion of the drug-gene interaction database. Nucleic Acids Res. 2018;46(November 2017):1068–73.
22. Ursu O, Holmes J, Knockel J, Bologa CG, Yang JJ, Mathias SL, et al. DrugCentral : online drug compendium. Nucleic Acids Res. 2017;45(October 2016):932–9.
23. Gaulton A, Hersey A, Patr A, Chambers J, Mendez D, Mutowo P, et al. The ChEMBL database in 2017. Nucleic Acids Res. 2017;45(November 2016):945–54.
24. Piñero J, Gonzalez-Perez A, Guney E, Aguirre-Plans J, Sanz F, Oliva B, et al. Network, Transcriptomic and Genomic Features Differentiate Genes Relevant for Drug Response. Front Genet. 2018;9:412.
25. Guney E, Oliva B. Exploiting Protein-Protein Interaction Networks for Genome-Wide Disease-Gene Prioritization. PLoS ONE. 2012;7(9):e43557.
26. Guney E, García-garcía J, Oliva B. GUILDify : A web server for phenotypic characterization of genes through biological data integration and network-based prioritization algorithms. Bioinformatics. 2014;30(12):1789–90.
27. Ghiassian SD, Menche J, Barabási AL. A DIseAse MOdule Detection (DIAMOnD) Algorithm Derived from a Systematic Analysis of Connectivity Patterns of Disease Proteins in the Human Interactome. PLoS Comput Biol. 2015;11(4):e1004120.
28. Rubio-Perez C, Guney E, Aguilar D, Piñero J, Garcia-Garcia J, Iadarola B, et al. Genetic and functional characterization of disease associations explains comorbidity. Sci Rep. 2017;7(1):6207.
29. Rolland T, Taşan M, Charloteaux B, Pevzner SJ, Zhong Q, Sahni N, et al. A proteome-scale map of the human interactome network. Cell. 2014;159(5):1212–26.
30. Luck K, Sheynkman GM, Zhang I, Vidal M. Proteome-Scale Human Interactomics. Trends Biochem Sci. 2017;42(5):342–54.
31. Sharma A, Menche J, Chris Huang C, Ort T, Zhou X, Kitsak M, et al. A disease module in the interactome explains disease heterogeneity,

drug response and captures novel pathways and genes in asthma. Hum Mol Genet. 2014;24(11):3005–20.

32. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics. 2011;27(12):1739–40.
33. Himmelstein DS, Lizee A, Hessler C, Brueggeman L, Chen SL, Hadley D, et al. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. eLife. 2017;6:e26726.
34. Hunter JD. Matplotlib: A 2D graphics environment. Comput Sci Eng. 2007;9(3):90–5.
35. Franz M, Lopes CT, Huck G, Dong Y, Sumer O, Bader GD. Cytoscape.js: A graph theory library for visualisation and analysis. Bioinformatics. 2015;32(2):309–11.

3.2. PxEA: A tool to prioritize drug repurposing candidates targeting endophenotypes

In the second article of the thesis, I present a method called Proximal pathway Enrichment Analysis (PxEA) to repurpose drugs specifically targeting the endophenotypes shared by different pathophenotypes. PxEA methodology consists in:

(1) Calculate the network distance between a drug and the pathways of the interactome by measuring the network proximity metric defined in Guney et al. (79) between the drug targets and the pathway-associated genes.

(2) Calculate a sum score based on Gene Set Enrichment Analysis (264) from which pathways are ranked by their proximity to the drug targets and a pathway set of interest (i.e., belonging to an endophenotype).

We evaluated PxEA in two steps:

(1) First, we investigated whether the drugs used in autoimmune disorders target specifically pathways associated with one disease or pathways shared across diseases. We found common pathways between almost all autoimmune disorders and drugs potentially targeting these common pathways.

(2) Second, we explored the potential endophenotypes shared by Type 2 Diabetes and Alzheimer's Disease, two diseases highly prevalent in our ageing society that are known to exhibit increased comorbidity.

PxEA paves the way for simultaneously targeting endophenotypes that manifest across various diseases, a concept which we refer to as *endopharmacology*.

Aguirre-Plans J, Piñero J, Menche J, Sanz F, Furlong LI, Schmidt HHHW, Oliva B, Guney E. Proximal Pathway Enrichment Analysis for Targeting Comorbid Diseases via Network Endopharmacology. ***Pharmaceuticals (Basel).*** 2018; 11(3): 61. DOI: 10.3390/ph11030061

Proximal Pathway Enrichment Analysis for Targeting Comorbid Diseases via Network Endopharmacology

Pharmaceuticals (Basel). 2018 Jun 22;11(3):61

Joaquim Aguirre-Plans[1], Janet Piñero[1], Jörg Menche[2], Ferran Sanz[1], Laura I. Furlong[1], Harald H. H. W. Schmidt[3], Baldo Oliva[1] and Emre Guney[1,3,*]

1- Research Programme on Biomedical Informatics, the Hospital del Mar Medical Research Institute and Pompeu Fabra University, Dr. Aiguader 88, 08003 Barcelona, Spain.

2- CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Lazarettgasse 14, AKH BT 25.3, A-1090 Vienna, Austria

3- Department of Pharmacology and Personalised Medicine, CARIM, FHML, Maastricht University, Universiteitssingel 50, 6229 ER Maastricht, The Netherlands

***Correspondence to Emre Guney**: emre.guney@upf.edu

Abstract

The past decades have witnessed a paradigm shift from the traditional drug discovery shaped around the idea of “one target, one disease” to polypharmacology (multiple targets, one disease). Given the lack of clear-cut boundaries across disease (endo)phenotypes and genetic heterogeneity across patients, a natural extension to the current polypharmacology paradigm is to target common biological pathways involved in diseases via endopharmacology (multiple targets, multiple diseases). In this study, we present proximal pathway enrichment analysis (PxEA) for pinpointing drugs that target common disease pathways towards network endopharmacology. PxEA uses the topology information of the network of interactions between disease genes, pathway genes, drug targets and other proteins to rank drugs by their interactome-based proximity to pathways shared across multiple diseases, providing unprecedented drug repurposing opportunities. Using PxEA, we show that many drugs indicated for autoimmune disorders are not necessarily specific to the condition of interest, but rather target the common biological pathways across these diseases. Finally, we provide high scoring drug repurposing candidates that can target common mechanisms involved in type 2 diabetes and Alzheimer’s disease, two conditions that have recently gained attention due to the increased comorbidity among patients.

Keywords: drug repurposing, proximal pathway enrichment analysis, network endopharmacology, systems medicine, comorbidity, autoimmune disorders, Alzheimer’s disease, type 2 diabetes.

1. Introduction

Following Paul Ehrlich’s more-than-a-century-old proposition on magic bullets (one drug, one target, one disease), the drug discovery pipeline traditionally pursues a handful of leads identified in vitro based on their potential to bind to target(s) known to modulate the disease (1). The success of the selected lead in the consequent clinical validation process relies on the prediction of a drug’s effect in vivo. Although it is often more desirable to tinker the cellular network by targeting multiple proteins (2), this is hard to achieve in practice due to the interactions of the compound and its targets with other proteins and metabolites. As a result, the characterization of drug effect has been a daunting task, yielding high pre-clinical attrition rates for novel compounds (3,4).

The high attrition rates can be attributed to the immense response heterogeneity across patients, likely stemming from a polygenic nature of most complex diseases. Consequently, researchers have turned their attention to polypharmacology, where novel therapies aim to alter multiple targets involved in the pathway cross-talk pertinent to the disease pathology, rather than single proteins (5,6). This has given rise to network-based approaches that predict the effects of individual drugs (79) as well as drug combinations (8), allowing for the repositioning of compounds for novel indications.

Over the past years, reusing existing drugs for conditions different from their intended indications has emerged as a cost effective alternative to traditional drug discovery. Various drug repurposing methods aim to mimic the most likely therapeutic and safety outcomes of candidate compounds based on similarities between compounds and diseases characterized by high-throughput omics data (9–11). Most studies so far, however, have focused on repurposing drugs for a single condition of interest, failing to recognize the cellular, genetic and ontological complexity inherent to human diseases (12,13). In reality, pathway cross-talk plays an important role in modulating the pathophysiology of diseases (14) and most comorbid diseases are interconnected to each other in the interactome through proteins belonging to similar pathways (15–19). The pathway cross-talk is especially relevant for autoimmune disorders, which have been shown to share several biological functions involved in immune and inflammatory responses (20,21). Autoimmune disorders affect around 15% of the population in the USA (22) and co-occur in the same patient more often than expected (i.e., comorbid) (23). Recent evidence suggests that endophenotypes—shared intermediate pathophenotypes—(24), such as inflammasome, thrombosome, and fibrosome play essential roles in the progression of not only autoimmune disorders but also many other diseases (25).

Here, we propose a novel drug repurposing approach, **P**ro**x**imal pathway **E**nrichment **A**nalysis (PxEA), to specifically target intertangled biological pathways involved in the common pathology of complex diseases. We first identify pathways proximal to disease genes across various autoimmune disorders. Then we use PxEA to investigate whether the drugs promiscuously used in these disorders target specifically the pathways associated with one disease or the pathways shared across the diseases. We find several examples of anti-inflammatory drugs where the pathways proximal to the drug targets in the interactome correspond to the pathways shared between two autoimmune disorders. The observed lack of specificity among these drugs points to the existence of immune system related endophenotypes, motivating us to explore shared disease mechanisms for repurposing drugs. We demonstrate that PxEA is a powerful computational strategy for targeting multiple pathologies involving common biological pathways, such as type 2 diabetes (T2D) and Alzheimer's disease (AD). Based on these findings, we argue that PxEA paves the way for simultaneously targeting endophenotypes that manifest across various diseases, a concept which we refer to as *endopharmacology*.

2. Results

2.1. Pathway Proximity Captures the Similarities between Autoimmune Disorders

Conventionally, functional enrichment analysis relies on the significance of the overlap between a set of genes belonging to a condition of interest and a list of genes involved in known biological processes (pathways). Using known pathway genes, one can identify pathways associated with the disease via a statistical test (e.g., Fisher's exact test for the overlap between genes or z-score comparing the observed number of common genes to the number of genes one would have in common if genes were randomly sampled from the data set). We start with the observation that such an approach (hereafter referred as to *conventional* approach) often misses key biological processes involved in the disease due to the limited overlap between the disease and pathway genes. To show that this is the case, we focus on nine autoimmune disorders for which we obtain genes associated with the disease in the literature and we calculate *p*-values based on the overlap between these genes and the pathway genes for each of the 674 pathways in the Reactome database (Fisher's exact test, one-sided $p \leq 0.05$). Intriguingly, **Table 1** demonstrates that this conventional approach yields less than ten pathways that are significantly enriched in five out of nine diseases, potentially underestimating the molecular underpinning of these diseases.

Table 1. Number of pathways enriched across nine autoimmune disorders based on the overlap between the pathway and disease genes (one-sided $p < 0.05$, assessed by a Fisher's exact test) and the proximity of the pathway genes to the disease genes in the interactome ($z \leq -2$, see Methods for details).

Disease	**# of Pathways**	
	Overlap	**Proximity**
celiac disease	7	143
Crohn's disease	5	116
diabetes mellitus, insulin-dependent	16	121
Graves' disease	3	92
lupus erythematosus, systemic	17	98
multiple sclerosis	12	138
psoriasis	5	50
rheumatoid arthritis	55	17
ulcerative colitis	6	138

Alternatively, the shortest distance between genes in the interactome can be used to find pathways closer than random expectation to a given set of genes (7,26), augmenting substantially the number of pathways relevant to the disease pathology. Using network-based proximity (7), we define the *pathway span* of a disease as the set of pathways significantly proximal to the disease ($z \leq -2$, see Methods). We show that the number of pathways involved in diseases increases substantially when proximity is used (**Table 1**).

To show the biological relevance of the identified pathways using interactome-based proximity, we check how well these pathways can highlight genetic and phenotypic relationships between nine autoimmune disorders. First, to serve as a background model, we build a disease network for the autoimmune disorders (diseasome) using the genes and symptoms shared between these diseases as well as the comorbidity information extracted from medical insurance claim records (see Methods). The autoimmune diseasome (**Figure 1a**) is extremely connected, covering 33 out of 36 potential links between nine diseases (with average degree $< k > = 7.3$ and clustering coefficient $CC = 0.93$). The three missing links are those between ulcerative colitis and rheumatoid arthritis, ulcerative colitis and Graves' disease, and Graves' disease and type 1 diabetes. On the other hand, several diseases such as celiac disease, Crohn's disease, systemic lupus erythematosus, and multiple sclerosis are connected to each other with multiple evidence types in the autoimmune diseasome based on genetic (shared genes) and phenotypic (shared symptoms and comorbidity) similarities, emphasizing the shared pathological components underlying these diseases.

We compare the autoimmune diseasome generated using shared genes, common symptoms and comorbidity, to the disease network in which the disease-disease connections are identified using the pathways they share. We identify the pathways enriched in the diseases using both the conventional and proximity approaches mentioned above and check whether the number of common pathways between two diseases is significant (two-tailed Fisher's exact test, $p < 0.05$). The disease network based on pathways shared across diseases using the overlap between the pathway and disease genes is markedly sparser than the original diseasome, containing 17 links (**Figure 1b**). None of the diseases share pathways with psoriasis and among the connections supported by multiple evidence in the original diseasome, the links between Crohn's disease and celiac disease as well as Crohn's disease and systemic lupus erythematosus are missing. On the contrary, the disease network based on shared pathways using proximity of the pathway genes to the disease genes consists of 34 links, where the only unconnected disease pairs are Crohn's disease and Graves' disease and type 1 diabetes and psoriasis, suggesting that it captures the connectedness of the original diseasome better than the conventional approach.

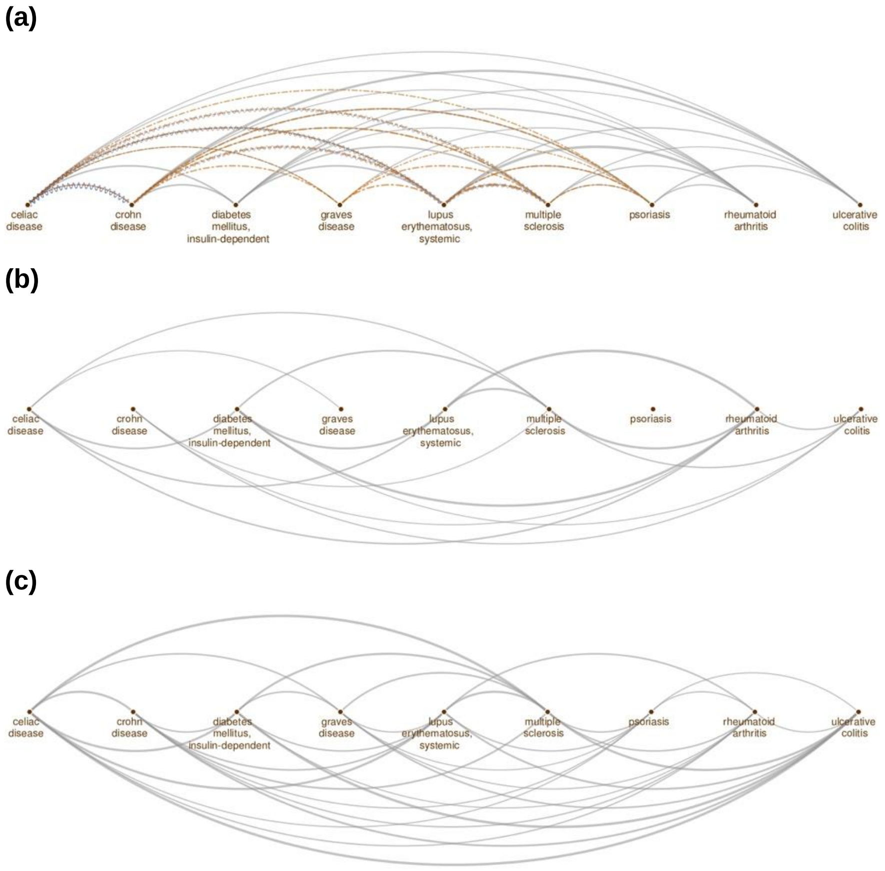


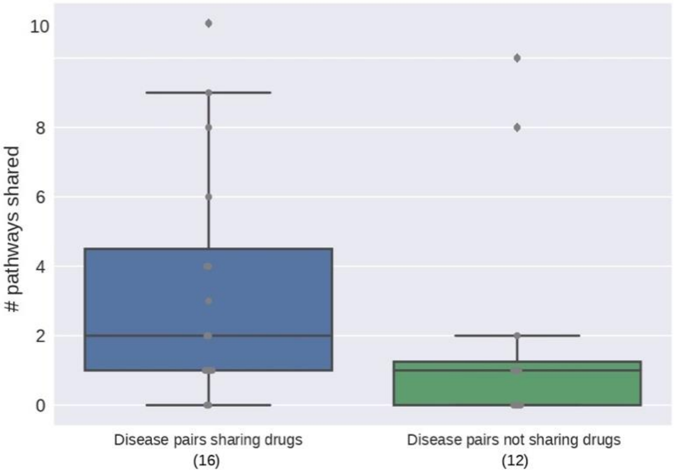
Figure 1. Genetic, phenotypic and functional overlap across autoimmune disorders. Disease relationships (links) based on **(a)** shared genes (gray solid lines), shared symptoms (orange dashed lines) and comorbidity (blue sinusoidal lines); **(b)** shared pathways (gray solid lines) using common disease and pathway genes, **(c)** shared pathways (gray solid lines) using the proximity of the pathway genes to the diseases genes in the interactome.

We next turn our attention to the shared pathways across diseases identified by both conventional and proximity based approaches and observe that most common pathways involve biological processes relevant to the immune system endophenotypes. In particular, we see that inflammasome-related pathways, such as signaling of cytokines (interferon gamma, interleukins like IL6, IL7) and lymphocytes (ZAP70, PD1, TCR, among others) are overrepresented. While conventional enrichment finds that most of these pathways are shared among only 4–5 diseases, proximity based enrichment points to the commonality of these pathways among almost all the diseases. Furthermore, the proximity based enrichment uncovers the involvement of additional interleukin (IL2, IL3, IL5) and lymphocyte (BCR) molecules ubiquitously in autoimmune disorders. These findings suggest that proximity-based pathway enrichment identifies biological processes relevant to the diseases, highlighting the common etiology across autoimmune disorders.

2.2. Diseases Targeted by the Same Drugs Exhibit Functional Similarities

Having observed that pathway proximity to diseases in the interactome captures the underlying biological mechanisms across diseases, we seek to investigate the potential implications of the connections between diseases for drug discovery. We hypothesize that a drug indicated for several autoimmune disorders would exert its effect by targeting the shared biological pathways across these diseases. To test this, we use 25 drugs that are indicated for two or more of the autoimmune disorders in Hetionet (27) and split disease pairs into two groups: (i) diseases for which a common drug exists and (ii) diseases for which no drugs are shared. We then count the number of pathways in common between two diseases for each pair in the two groups using pathway enrichment based on both the gene overlap and proximity in the interactome. We find that the diseases targeted by the same drugs tend to involve an elevated number of common pathways compared to the disease pairs that do not have any drug in common (**Figure 2**). The average number of pathways shared among diseases that are targeted by the same drug is 3.4 and 38 using overlap and proximity based enrichment, respectively, whereas, the remaining disease pairs share 2 and 31 pathways on average using the two enrichment approaches. We note that due to the relatively small sample size and potentially incomplete drug indication information, we interpret the elevated number of pathways as a trend rather than a general rule across all diseases (*p* = 0.043 and *p* = 0.066, assessed by one-tailed Mann-Whitney U test, for the overlap and proximity based approaches, respectively). Nevertheless, taken together with the high overall pathway level commonalities observed in the autoimmune disorders mentioned in the previous section, this result suggests that the drugs used for multiple indications are likely to target common pathways involved in these diseases.

(a)



(b)

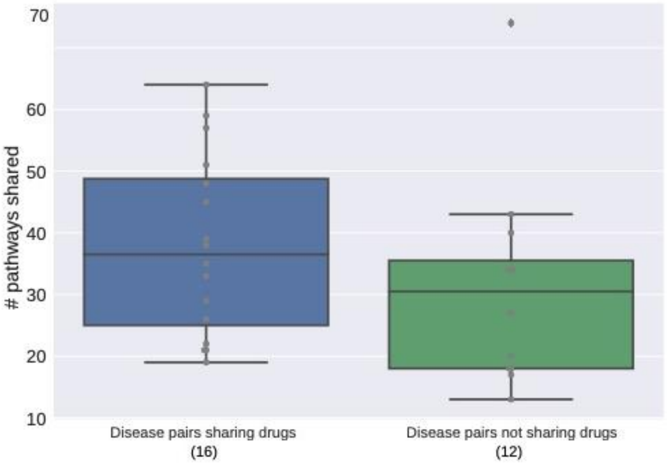


Figure 2. Number of shared pathways across disease pairs that are targeted by the same drug compared to the rest of the pairs. The pathway enrichment is calculated using **(a)** gene overlap and **(b)** proximity of genes in the interactome. The number of disease pairs in each group is given in the parenthesis below the group label in the x-axis.

2.3. Proximal Pathway Enrichment Analysis Reveals Drugs Targeting the Autoimmune Endophenotypes

The results indicating that the drugs used for multiple autoimmune disorders potentially target common pathways raise the following question: “Can pathway level commonalities between diseases be leveraged to quantify the impact of a given drug on these diseases?” To this end, we propose PxEA, a novel method for **P**ro**x**imal pathway **E**nrichment **A**nalysis that scores the likelihood of a set of pathways (e.g., targeted by a drug) to be represented among another set of pathways (e.g., disease pathways) based on the proximity of the pathway genes in the interactome. As opposed to the Gene Set Enrichment Analysis (GSEA) (28) which uses gene sets and the ranking of genes based on differential expression, PxEA uses pathway sets and the ranking of pathways based on proximity in the interactome. PxEA scores a drug based on whether or not the pathways targeted by the drug are proximal to a pathway set of interest, such as pathways shared across different diseases. For a given drug and a pair of diseases, we first identify the pathways in the pathway span of both of the diseases, then we rank the pathways with respect to the proximity of the drug targets to the pathway genes and finally we calculate a running sum statistics corresponding to the enrichment score between the drug and the disease pair (**Figure 3**, see Methods for details).

We employ PxEA to score 25 drugs indicated for at least two of the seven autoimmune disorders (there were no common drugs for celiac and Graves' diseases). For each disease, we first run PxEA using the pathways proximal to the disease and the proximity of the drugs used for that disease to these pathways. We then run PxEA for each disease pair, using the pathways proximal to both of the diseases in the pair and the drugs commonly used for the two diseases. We notice that several drugs indicated for multiple conditions score higher using common pathways between two diseases than using the pathways of the disease they are indicated for (**Figure 4**). This is not surprising considering that many of the drugs used for autoimmune disorders target common immune and inflammatory processes. For instance, sildenafil, a drug used for the treatment of erectile dysfunction and to relieve the symptoms of pulmonary arterial hypertension, is reported by Hetionet to show palliative effect on type 1 diabetes and multiple sclerosis. Actually, sildenafil is not specific to any of these two conditions and targets a number of the 57 pathways in common between type 1 diabetes and multiple sclerosis including but not limited to pathways mentioned in **Table 2**, such as “IL-3, 5 and GM CSF signaling” ($z = -1.6$), “regulation of signaling by CBL” ($z = -1.1$), “regulation of KIT signaling” ($z = -1.0$), “IL receptor SHC signaling” ($z = -1.0$), and “growth hormone receptor signaling” ($z = -1.0$).

Similarly, prednisone, a synthetic anti-inflammatory glucocorticoid agent that is indicated for six of the autoimmune disorders, is assigned a higher PxEA score using the pathways shared by Crohn’s disease and systemic lupus erythematosus compared to using the pathways involved only in Crohn’s disease, systemic lupus erythematosus, multiple sclerosis, psoriasis, rheumatoid arthritis, or ulcerative colitis. Thus, prednisone does not specifically target any of the six autoimmune disorders but rather acts on the endophenotypes that manifest across these diseases. We observe a similar trend in meloxicam, an anti-inflammatory drug that shows analgesic and antipyretic effects by inhibiting prostaglandin synthesis. Consistent with its known mechanism of action, meloxicam is proximal to “cholesterol biosynthesis” ($z = -3.5$), “fatty acid, triacylglycerol, and ketone body metabolism” ($z = -2.0$), and “prostanoid ligand receptors” ($z = -1.7$) pathways in the interactome. While meloxicam is originally indicated for rheumatoid arthritis and systemic lupus erythematosus, the higher PxEA score when common arthritis and lupus pathways are used suggests that it targets common inflammatory processes in these two diseases.

Table 2. Pathways shared by autoimmune disorders based on the overlap and proximity of genes (only pathways that appear most commonly across diseases are shown).

Pathway	**# of Shared Diseases**	
	Overlap	**Proximity**
interferon gamma signaling	5	8
costimulation by the CD28 family	5	7
cytokine signaling in immune system	5	7
translocation of ZAP-70 to immunological synapse	5	6
phosphorylation of CD3 and TCR zeta chains	5	6
PD1 signaling	5	4
IL-6 signaling	4	8
generation of second messenger molecules	4	6
TCR signaling	4	6
signaling by ILs	3	9
immune system	3	7
downstream TCR signaling	3	7
interferon signaling	3	7
adaptive immune system	3	3
regulation of KIT signaling	2	7
IL-7 signaling	2	6
CTLA4 inhibitory signaling	2	5
chemokine receptors bind chemokines	2	3
extrinsic pathway for apoptosis	2	3
MHC class II antigen presentation	2	2
IL receptor SHC signaling	-	9
IL-3, 5 and GM CSF signaling	-	9

signaling by the B cell receptor BCR	-	8
regulation of IFNG signaling	-	8
growth hormone receptor signaling	-	8
IL-2 signaling	-	8
regulation of signaling by CBL	-	8

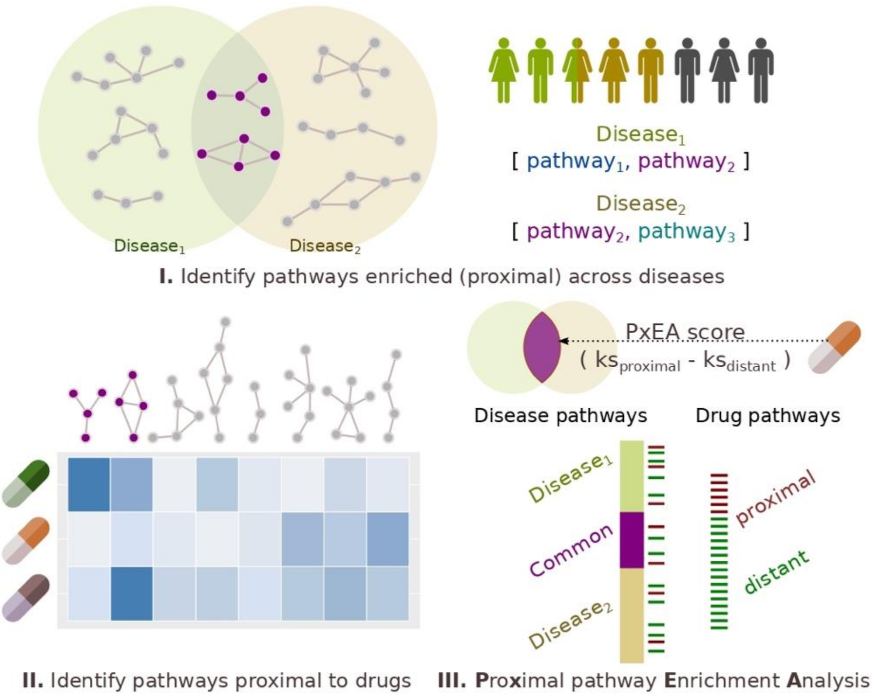

Disease$_1$
Disease$_2$
Disease$_1$
[pathway$_1$, pathway$_2$]
Disease$_2$
[pathway$_2$, pathway$_3$]
I. Identify pathways enriched (proximal) across diseases
PxEA score
(ks$_{proximal}$ - ks$_{distant}$)
Disease pathways
Drug pathways
Disease$_1$
Common
Disease$_2$
proximal
distant
II. Identify pathways proximal to drugs
III. Proximal pathway Enrichment Analysis

Figure 3. Schematic overview of proximal pathway enrichment analysis (PxEA). PxEA scores a drug with respect to its potential to target the pathways shared between two diseases. For a given drug and two diseases of interest, PxEA first identifies the common pathways between the two disease and then uses the proximity-based ranking of the pathways (i.e., average distance in the interactome to the nearest pathway gene, normalized with respect to a background distribution of expected scores) to assign a score to the drug and the disease pair.

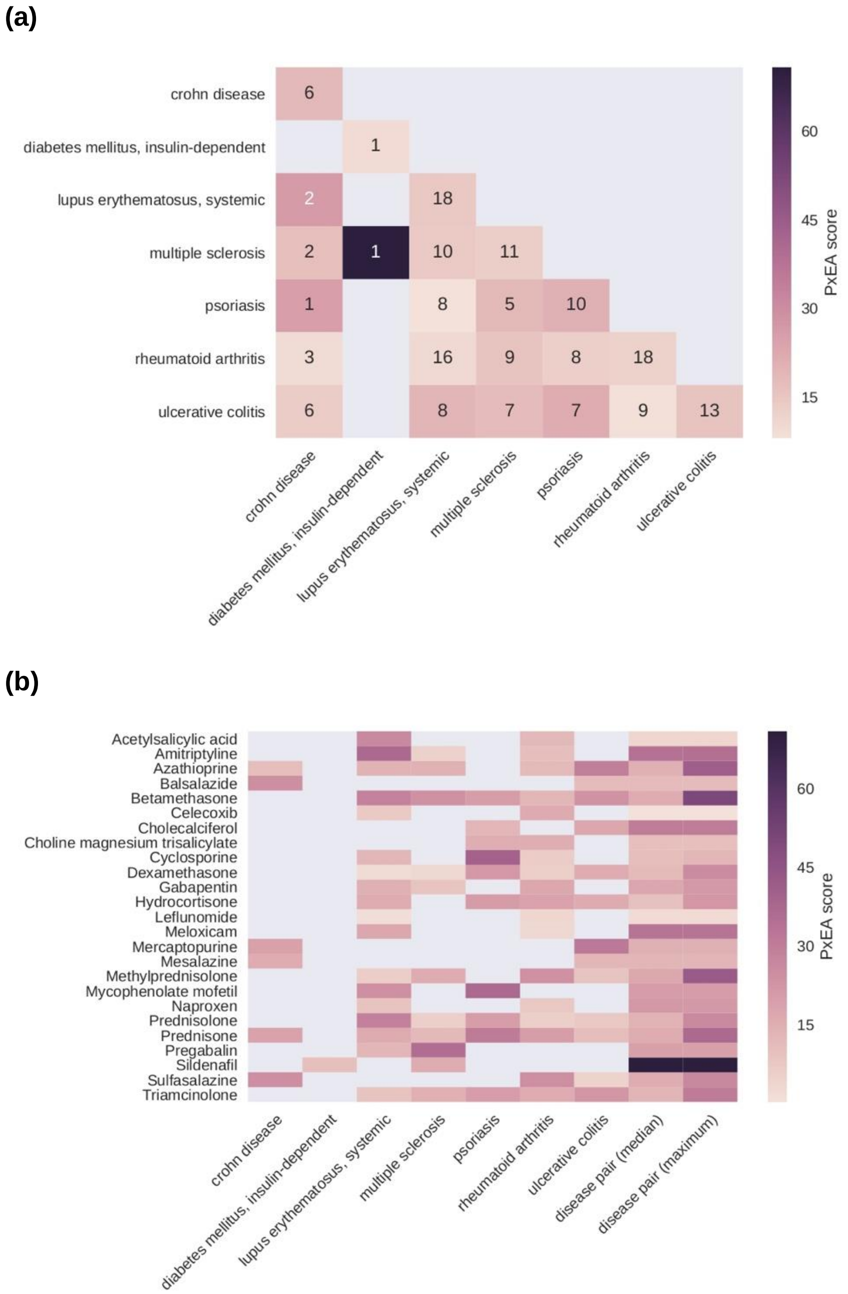


Figure 4. PxEA scores of drugs used in autoimmune disorders. (**a**) Disease-disease heatmap, in which for each disease pair, the common pathways proximal to the two diseases are used to run PxEA. Note that the diagonal contains the PxEA scores obtained when the proximal pathways

for only that disease are used. The hue of the color scales with the PxEA score. (**b**) Drug-disease heatmap, in which the PxEA is run using the pathways proximal to the pathways of the disease in the column for the drugs in the rows (25 drugs that are used at least in two diseases). The last two columns show the median and maximum values of the PxEA scores obtained for the drug among all disease pairs the drug is indicated for.

2.4. Targeting the Common Pathology of Type 2 Diabetes and Alzheimer's Disease

T2D and AD, two diseases highly prevalent to an ageing society, are known to exhibit increased comorbidity (29,30). Recently, repurposing anti-diabetic agents to prevent insulin resistance in AD has gained substantial attention due to the therapeutic potential it offers (31). Indeed, the pathway spans of T2D and AD cover 170 and 82 pathways, respectively, 35 of which are shared between two diseases, linking significantly the two diseases at the pathway level (Fisher's exact test, two-sided $p = 2.2 \times 10^{-4}$).

We use PxEA to score 1466 drugs from DrugBank using the 35 pathways involved in the common pathology of T2D and AD. When we look at the drugs ranked on the top of the list (**Table 3**), we spot orlistat, a drug indicated for obesity and T2D in Hetionet. Interestingly, existing studies also suggest a role for this drug in the treatment of AD (32). Orlistat targets extracellular communication (Ras-Raf-MEK-ERK, NOTCH, and GM-CSF/IL-3/IL-5 signaling) and lipid metabolism pathways (**Figure 5**). Several of the proteins in the pathways pertinent to the common T2D-AD pathology, such as APOA1, PSEN2, PNLIP, LPL, and IGHG1 are either orlistat's targets themselves or are in the close vicinity of the targets. The next top scoring drugs are chenodeoxycholic and obeticholic acid, biliar acids that are in clinical trials for T2D (NCT01666223) and are argued to modulate cognitive changes in AD (33).

Table 3. Top ten drug repurposing opportunities to target common T2D and AD pathology, where the drugs that target the same proteins according to DrugBank are grouped together in the same row and the Anatomical Therapeutic Chemical (ATC) classification and indication information within the same group is marked with the first letter of the drug in the parenthesis (if applicable).

Drug	ATC	Hetionet Indication	DrugBank Indication	PxEA score	Adjusted P-value
orlistat	A08	obesity, type 2 diabetes	obesity	94.07	<0.001
obeticholic acid, chenodeoxycholic acid	A05	primary biliary cirrhosis (C)	liver disease (O), primary biliary cholangitis (O), gallbladders (C)	74.06	<0.001
esmolol, practolol	C07	hypertension (E)	atrial fibrillation (E), noncompensatory sinus tachycardia (E), cardiac arrhythmias (P)	70.55	<0.001
clenbuterol	R03	-	asthma	70.44	<0.001
erythrityl tetranitrate	C01	-	angina	70.32	<0.001
fenoterol, arbutamine, bupranolol	R03 (F), G02 (F) C01 (A), C07 (B)	-	asthma (F); coronary artery disease (A); hypertension (B), tachycardia (B), glaucoma (B)	68.97	<0.001
dalfampridine	N07	multiple sclerosis	multiple sclerosis	68.44	<0.001
magnesium sulfate	D11, V04, A06, B05, A12	-	eclampsia, acute nephritis, acute hypomagnesemia, uterine tetany	68.27	<0.001
roflumilast, crisaborole	R03 (R)	chronic obstructive pulmonary disease (R)	chronic obstructive pulmonary disease (R), dermatitis (C), psoriasis (C)	66.33	<0.001

montelukast	R03	chronic obstructive pulmonary disease, asthma, allergic rhinitis	asthma	65.94	<0.001
orlistat	A08	obesity, type 2 diabetes	obesity	94.07	<0.001
obeticholic acid, chenodeoxycholic acid	A05	primary biliary cirrhosis (C)	liver disease (O), primary biliary cholangitis (O), gallbladders (C)	74.06	<0.001
esmolol, practolol	C07	hypertension (E)	atrial fibrillation (E), noncompensatory sinus tachycardia (E), cardiac arrhythmias (P)	70.55	<0.001
clenbuterol	R03	-	asthma	70.44	<0.001
erythrityl tetranitrate	C01	-	angina	70.32	<0.001
fenoterol, arbutamine, bupranolol	R03 (F), G02 (F) C01 (A), C07 (B)	-	asthma (F); coronary artery disease (A); hypertension (B), tachycardia (B), glaucoma (B)	68.97	<0.001
dalfampridine	N07	multiple sclerosis	multiple sclerosis	68.44	<0.001
magnesium sulfate	D11, V04, A06, B05, A12	-	eclampsia, acute nephritis, acute hypomagnesemia, uterine tetany	68.27	<0.001
roflumilast, crisaborole	R03 (R)	chronic obstructive pulmonary disease (R)	chronic obstructive pulmonary disease (R), dermatitis (C), psoriasis (C)	66.33	<0.001
montelukast	R03	chronic obstructive pulmonary disease, asthma, allergic rhinitis	asthma	65.94	<0.001

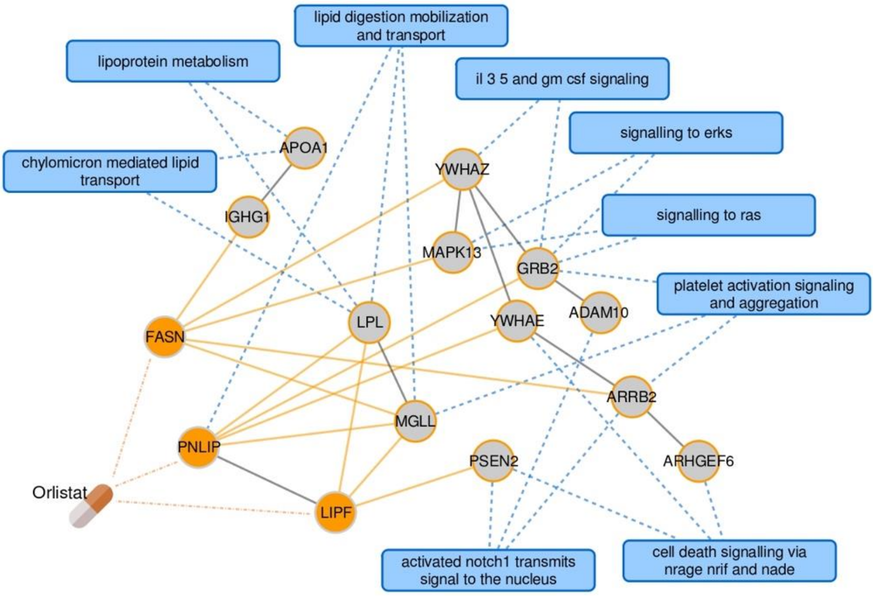

lipid digestion mobilization and transport
lipoprotein metabolism
il 3 5 and gm csf signaling
signalling to erks
chylomicron mediated lipid transport
signalling to ras
platelet activation signaling and aggregation
APOA1
YWHAZ
IGHG1
MAPK13
GRB2
ADAM10
YWHAE
LPL
FASN
ARRB2
MGLL
PNLIP
PSEN2
ARHGEF6
Orlistat
LIPF
activated notch1 transmits signal to the nucleus
cell death signalling via nrage nrif and nade

Figure 5. Orlistat from PxEA perspective. The subnetwork shows how the targets of orlistat are connected to the nearest pathway protein for the pathways shared between T2D and AD. For clarity, only the pathways that are proximal to the drug are shown. Blue rectangles represent pathways, circles represent drug targets (orange) or proteins on the shortest path to the nearest pathway gene (gray). Blue dashed lines denote pathway membership, solid lines are protein interactions. The interactions between the drug and its targets are shown in dashed orange lines and the interactions between the drug targets and their neighbors are highlighted with solid orange lines.

It is noteworthy that the top scoring drugs belong to a diverse set of Anatomical Therapeutic Chemical (ATC) classes, covering alimentary tract and metabolism drugs (A05, A06, A08, A12), blood substitutes (B05), dermatologicals (D11) as well as cardiovascular (C01, C07), genito-urinary (G02), nervous (N07), and respiratory (R03) system drugs. The diversity of the ATC classes of top scoring drugs indicates that PxEA is not biased towards any particular ATC class. We also calculate the significance of the PxEA scores by permuting the ranking of the pathways. We find that the adjusted *p*-values (corrected for multiple hypothesis testing using Benjamini–Hochberg procedure) for the top candidates are all below 1×10^{-4}, the minimum possible value (due to the 10,000 permutations used in the calculation).

3. Discussion

The past decades have witnessed a substantial increase in human life expectancy owing to major breakthroughs in translational medicine. Yet, the increase on average age and changes in life style, have given rise to a spectra of problems challenging human health like cancer, neurodegenerative disorders and diabetes. These diseases do not only limit the life expectancy but also induce a high burden on public healthcare costs. In the US alone, more than 20 and 5 million people have been affected by T2D and AD, respectively, ranking these diseases among the most prevalent health problems (29).

Mainly characterized by hyperglycemia due to resistance to insulin, the disease mechanism of T2D involves a combination of multiple genetic and dietary factors. On the other hand, AD is relatively less understood and several hypotheses have been proposed for its cause: reduced synthesis of neurotransmitter acetylcholine, accumulation of amyloid beta plaques and/or tau protein abnormalities, giving rise to neurofibrillary tangles. Accordingly, most available treatments in AD are palliative (treating symptoms rather than the cause). Given the comorbidity between T2D and AD (29,30) several studies have recently suggested repurposing diabetes drugs for AD (31). However, to our knowledge, currently there is no systematic method that can pinpoint drugs that could be useful to target common disease pathology such as the one between T2D and AD.

In this study, we first show that diseases that share drugs also tend to share biological pathways and hypothesize that these pathways can be targeted to exploit novel drug repurposing opportunities. We introduce PxEA, a method based on (i) pathways that are proximal to diseases and (ii) the ranking of the pathways targeted by a drug using the topology information encoded in the human interactome. We show that PxEA picks up whether drugs target specifically the pathways associated with a disease or common pathways shared across various conditions. We observe that many anti-inflammatory drugs are not specific to the condition they are used for and likely to target pathways involved in the autoimmune endophenotypes.

To further explore shared disease mechanisms for repurposing drugs, we use PxEA and rank drugs for their therapeutic potential in targeting the common disease pathology between T2D and AD. We identify orlistat, a semisynthetic derivative of lipstatin that inhibits lipase—a pancreatic enzyme that breaks down fat—as the top repurposing candidate. Orlistat inhibits hydrolysis of triglycerides, which in turn, reduces the absorption of monoaclglycerides and free fatty acids (34). Recent evidence indicates that perturbations in unsaturated fatty acid metabolism are tightly coupled to neuritic plaque and neurofibrillary tangle formation in AD patients (35). Thus, orlistat might help slowing down the plaque and tangle formation due to its effect on the fatty acid metabolism. Targeting of fatty acid metabolism for improving the cognitive performance presents a novel therapeutic approach and is further supported by experiments in mouse models (36).

PxEA can suggest rather counter-intuitive repositioning opportunities such as the use of clenbuterol, an asthmatic drug, in the treatment of metabolic and neurodegenerative diseases such as T2D and AD. In fact, the potential use of clenbuterol in these diseases is not too far fetched: it enhances cognitive performance in aging rats and monkeys (37), improves memory deficit in mice (38), and reduces the insulin resistance in obese rats (39). On the flip side, while PxEA provides a cellular network based perspective to recommend drugs, it does not take into account dosage-related effects of drugs, potential adverse events, or the genetic background of the patients. For instance, practolol, a beta-adrenergic antagonist that stands out among the T2D-AD candidates, has been withdrawn from the market due to its high toxicity, limiting its potential therapeutic use in the clinical setting. Despite the limitations of PxEA, such as the incompleteness in the drug target, disease and pathway genes, lack of consideration of dosage-related effects or genetic heterogeneity, we believe PxEA is the first step towards achieving endopharmacology, that is, targeting endophenotypes involved across multiple diseases.

4. Material and Methods

4.1. Protein Interaction Data and Interactome-Based Proximity

To define a global map of interactions between human proteins, we obtained the physical protein interaction data from a previous study that integrated various publicly available resources (16). We downloaded the supplementary data accompanying the article to generate the human protein interaction network (interactome) containing data from MINT (40), BioGRID (41), HPRD (42), KEGG (43), BIGG (44), CORUM (45), and PhosphoSitePlus (46). We used the largest connected component of the interactome in our analyses, which covered 141,150 interactions between 13,329 proteins (represented by ENTREZ gene ids).

Network-based proximity is a graph theoretic approach that incorporates the interactions of a set of genes (i.e., disease genes or drug targets) with other proteins in the human interactome and contextual information as to where the genes involved in pathways reside with respect to the original set of genes (7). To quantify interactome-based proximity between two gene sets (such as drug targets, pathway genes or disease genes), we used the average shortest path length from the first set to the nearest protein in the second set following the definition in the original study (7). Accordingly, the proximity from nodes S to nodes T in a network $G(V, E)$, is defined as:

$$d(S,T) = \frac{1}{\parallel S \parallel} \sum_{u \in S} \min_{v \in T} d(u,v)$$

where $d(u,v)$ is the shortest path length between nodes u and v in G. We then calculated a z-score based on the distribution of the average shortest path lengths across random gene sets S_{random} and T_{random} ($d_{random}(S,T) = d(S_{random}, T_{random})$) as follows:

$$z(S,T) = \frac{d(S,T) - \mu_{d_{random}(S,T)}}{\sigma_{d_{random}(S,T)}}$$

where $\mu_{d_{random}(S,T)}$ and $\sigma_{d_{random}(T,S)}$ are the mean and the standard deviation of the $d_{random}(S,T)$, respectively obtained using 1,000 realizations of random sampling of gene sets that match the original sets in size and degree. We refer to the pathways that are significantly proximal ($z \leq -2$) to a disease as the *pathway span* of the disease throughout text.

Note that, instead of average shortest path distances, one can also use random-walk based distances to calculate proximity between gene sets (26). However, random walks in the networks are inherently biased towards high-degree nodes (47,48) and require additional statistical adjustment (26,48). Sampling based on size and degree matched gene sets has been shown to be robust against data-incompleteness in the interactome and in the known pathway annotations (7,48).

To investigate the effect of noise in the pathway data, following the procedure proposed in (49), we created a synthetic pathway data set, in which we defined pathways using a certain percentage *k* of known disease genes in T2D and AD (*k* = 10, 25, 50, 75, 90). Hence, for each value of *k*, we created 10 groups of genes, containing a random sampling of *k*% of the T2D-associated genes. We repeated the procedure using the AD-associated genes, yielding 100 gold standard pathways (10 for each disease across 5 different values of *k*) that were subsets of the known disease genes. For each gold standard pathway, we then generated so called control pathway, that is, randomly selected group of genes in the interactome that match the size of the gold standard pathway under consideration. Next, we assessed the shortest path distance based proximity between the gold standard pathways and the disease genes (proximity of the gold standard T2D pathways to the T2D disease genes and of the gold standard AD pathways to the AD disease genes) and compared it to the proximity of the control pathways to the same disease genes. We also calculated the proximity using random walk scores as proposed in a previous study (50). We used the random walk implementation in GUILD software package (51) with the default parameters. As one would expect, the gold standard pathways were significantly more proximal ($z \leq -2$) to the disease genes than the control pathways using both proximity calculation approaches (**Figure 6**). On the other hand, the shortest path distance based proximity distinguished better the overlap between the gold standard pathway genes and the disease genes by providing lower values than the random walk based proximity as the noise in the pathway information decreased (higher values of *k* in the gold pathways).

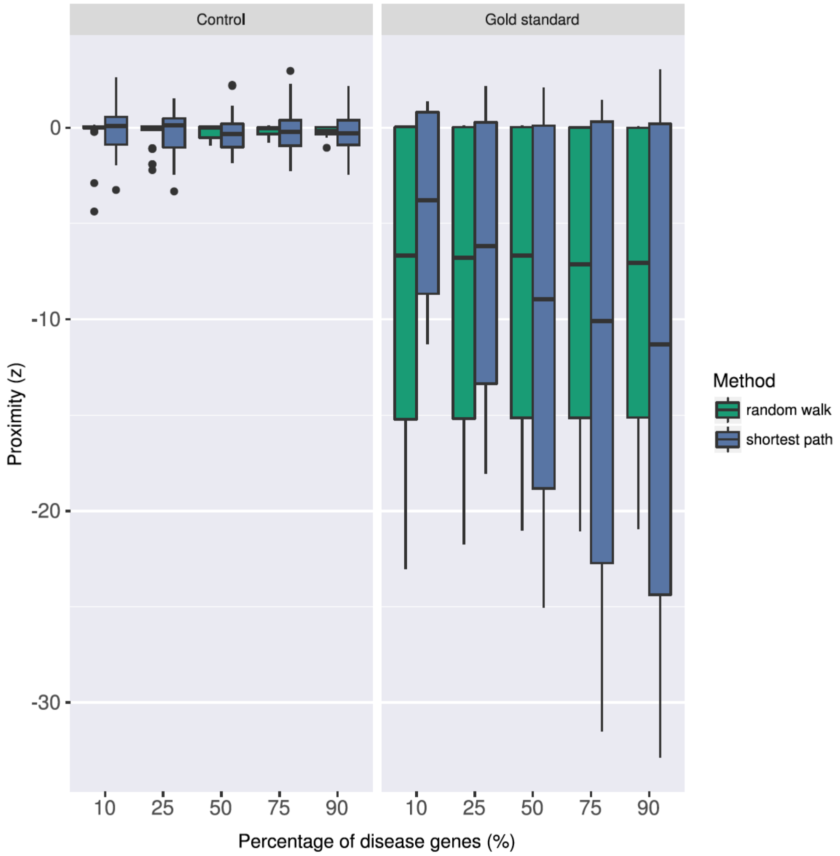


Figure 6. Effect of noise in the pathway data on the random walk and shortest path based proximity calculation. To assess the robustness of the interactome-based proximity in regards to noise in the pathway data, we generated synthetic gold standard pathways containing a certain proportion (k%) of the known disease genes in T2D and AD (see text for details). We compared the proximity between these gold standard pathways and the disease genes to the proximity between the control pathways (random groups of gene in the interactome) and the disease genes. The proximity values using random walk and the shortest path for increasing k values are shown for the control and gold standard pathways.

4.2. Disease-Gene, Drug and Pathway Information

We compiled genes associated with nine autoimmune disorders listed in **Table 4** using disease-gene annotations from DisGeNET (52). We downloaded curated disease-gene associations from DisGeNET that contained information from UniProt (53), ClinVar (54), Orphanet (55), GWAS Catalog (56) and CTD (57). To ensure that the disease-gene associations were of high confidence, we kept only the associations that were also provided in a previous large-scale analysis of human diseases (16).

We retrieved drug target information from DrugBank for 1489 drugs in the version 5.0.6 of the database (58), 1466 of which had at least a target in the interactome. UniProt ids from DrugBank were mapped to ENTREZ gene ids using UniProt id mapping file (retrieved on October 2017). We used drug indication information from Hetionet (compound treats or palliates disease edges) that compiled data from publicly available resources (27). We focused on 78 drugs that were indicated for nine autoimmune disorders above. We created a subset of drugs used for two or more of the autoimmune disorders, yielding 25 drugs across seven conditions (there were no indications for celiac disease, and the two drugs used for Graves' disease were not used in any other disease).

The ENTREZ gene ids of the proteins involved in biological pathways were taken from the version 5.0 of MSigDB curated gene sets (59). In our analysis, we used 674 Reactome (60) pathways and the genes associated with these pathways in the MSigDB.

Table 4. Disease-gene associations for the nine autoimmune disorders used in this study.

Drug	ATC	Hetionet Indication
celiac disease	11	IL21 CCR4 HLA-DQA1 BACH2 RUNX3 ICOSLG SH2B3 CTLA4 MYO9B ZMIZ1 ETS1
Crohn's disease	19	DNMT3A IL12B IRGM IL10 CCL2 FUT2 SMAD3 TYK2 ATG16L1 BACH2 IL2RA NKX2-3 PTPN2 NOD2 TAGAP MST1 DENND1B IL23R ERAP2
diabetes mellitus, insulin-dependent	18	IL10 GLIS3 HLA-DQA1 HLA-DRB1 PTPN22 SLC29A3 INS BACH2 CLEC16A PAX4 HLA-DQB1 IL2RA CD69 IL27 HNF1A CTSH SH2B3 C1QTNF6
Graves' disease	4	RNASET2 CTLA4 FCRL3 TSHR
lupus erythematosus, systemic	29	IKZF1 CFB RASGRP3 PDCD1 RASGRP1 DNASE1 HLA-DRB1 PTPN22 ETS1 TNIP1 FCGR2B TNFSF4 IRF5 C2 PRDM1 PXK TLR5 TREX1 TNFAIP3 SLC15A4 PHRF1 HLA-DQA1 STAT4 ITGAX ITGAM BLK C4A BANK1 CR2
multiple sclerosis	15	CD58 CD6 IRF8 HLA-DQB1 CBLB HLA-DRA KIF1B IL2RA TNFSF14 VCAM1 IL7R HLA-DRB1 CD24 TNFRSF1A PTPRC
psoriasis	15	IL12B TNIP1 LCE3D IL13 IL23R TYK2 HLA-DQB1 HLA-C FBXL19 ERAP1 TRAF3IP2 TNFAIP3 TNF REL NOS2
rheumatoid arthritis	23	MIF CD40 ANKRD55 HLA-DRB1 PTPN22 RBPJ IL2RA AFF3 CCL21 REL SLC22A4 CCR6 IRF5 SPRED2 CTLA4 PADI4 TNFAIP3 NFKBIL1 HLA-DQA2 STAT4 IL6 BLK TRAF1
ulcerative colitis	24	IL12B JAK2 ICOSLG IL1R2 LSP1 CXCR2 IL10 IL7R CXCR1 DAP NKX2-3 CARD9 GNA12 IRF5 PRDM1 HNF4A CCNY SLC26A3 FCGR2A IL23R IL17REL MST1 TNFSF15 CDH3

4.3. Genetic, Phenotypic and Functional Relationships across Diseases

To identify relationships across disease pairs (autoimmune diseasome), we used the similarities between diseases in terms of the genes and symptoms they share. We assessed the significance of the overlap between genes (or symptoms) associated with two diseases using Fisher's exact test. An alpha value of 0.05 was set to deem the connections significant (two-sided test $p \leq 0.05$). The disease symptom information was taken from a previous study based on text mining of PubMed abstracts (61). In this study, the number of times a symptom appears in a PubMed abstract was adjusted by the frequency of the symptom in the whole corpus using time frequency-inverse document frequency approach (TF-IDF). To ensure that the disease-symptom associations are of high quality, we considered associations with TF-IDF score higher than 3.5 as suggested in the original study.

Comorbidity relationships across diseases were inferred using data from medical insurance claims, where we assessed whether two diseases occurred more often in the same patient compared to the rest using the relative risk score (62). Relative risk score relies on the relative occurrence frequencies of diseases across patients, adjusting for the prevalence of the diseases. We mapped the ICD9 codes to MeSH identifiers using the annotations provided by Disease Ontology (63) and we considered the disease pairs with a relative risk score higher than 1 as potential comorbidity links.

To identify pathways enriched in diseases, we used the significance (i) of the overlap between the pathway and disease genes assessed by a one-tailed Fisher's exact test and (ii) of the proximity between the pathway and disease genes in the interactome. We considered the pathways that had $p \leq 0.05$ and $z \leq -2$, respectively, as the pathways that were enriched in a given disease using the two approaches. The pathway information was taken from Reactome and the proximity was calculated as explained above.

4.4. PxEA: Proximal Pathway Enrichment Analysis

Toward the goal of pathway level characterization of the common pathology of diseases and to evaluate the therapeutic potential of drugs based on their impact on the common pathways, we developed **P**ro**x**imal pathway **E**nrichment **A**nalysis (PxEA), a novel method that scores drugs based on the proximity of drug targets to pathway genes in the interactome. PxEA uses a GSEA-like running sum score (28), where the pathways are ranked with respect to the proximity of drug targets to the pathways and each pathway is evaluated to see whether or not it appears among the pathways of interest (e.g., common pathways between two diseases). Given *D*, the pathways ranked with respect to their proximity to drug targets, p_i, the pathway in consideration within *D*, and *C*, the set of pathways of interest, the running score is defined as follows (64):

$$ES(D, C) = \sum_{p_i \in P} X_i$$

where,

$$X_i = \{ \begin{matrix} \sqrt{\frac{|D| - |C|}{|C|}}, & if\ p_i \in C \\ -\sqrt{\frac{|C|}{|D| - |C|}}, & otherwise \end{matrix}$$

To calculate P-values for the case study, we repeat the procedure above 10,000 times, shuffling randomly D to calculate the expected enrichment score $ES\left(D^{random}, C\right)$. We then calculate the P-value for the enrichment using

$$P = \frac{\left|ES(D, C) < ES\left(D^{random}, C\right)\right|}{10{,}000}$$

The P-values were corrected for multiple hypothesis testing using Benjamini-Hochberg procedure (65).

4.5. Implementation Details and Code Availability

We used the toolbox Python package for running PxEA, available at github.com/emreg00/toolbox. The proximity was calculated using networkx package that implements Dijkstra's shortest path algorithm. The statistical tests were conducted in R (www.R-project.org) and Python (www.python.org). The network visualizations were generated using Cytoscape (66) and the plots were drawn using either Seaborn python package (67) or ggplot2 R package (68).

Abbreviations

The following abbreviations are used in this manuscript:

AD: Alzheimer’s disease
ATC: Anatomical Therapeutic Chemical
GSEA: Gene set enrichment analysis
PxEA: Proximal pathway enrichment analysis
T2D: Type 2 diabetes
TF-IDF: Time frequency-inverse document frequency approach

References

1. Strebhardt K, Ullrich A. Paul Ehrlich’s magic bullet concept: 100 years of progress. Nat Rev Cancer. 2008 Jun;8(6):473–80.
2. Csermely P, Korcsmáros T, Kiss HJM, London G, Nussinov R. Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. Pharmacol Ther. 2013 Jun;138(3):333–408.
3. Allison M. Reinventing clinical trials. Nat Biotechnol. 2012 Jan 9;30(1):41–9.
4. Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J. Clinical development success rates for investigational drugs. Nat Biotechnol. 2014 Jan;32(1):40–51.
5. Hopkins AL. Network pharmacology: the next paradigm in drug discovery. Nat Chem Biol. 2008 Nov;4(11):682–90.
6. Pujol A, Mosca R, Farrés J, Aloy P. Unveiling the role of network and systems biology in drug discovery. Trends Pharmacol Sci. 2010 Mar;31(3):115–23.

7. Guney E, Menche J, Vidal M, Barábasi A-L. Network-based in silico drug efficacy screening. Nat Commun. 2016 Feb 1;7:10331.
8. Jaeger S, Igea A, Arroyo R, Alcalde V, Canovas B, Orozco M, et al. Quantification of Pathway Cross-talk Reveals Novel Synergistic Drug Combinations for Breast Cancer. Cancer Res. 2017 Jan 15;77(2):459–69.
9. Jin G, Wong STC. Toward better drug repositioning: prioritizing and integrating existing methods into efficient pipelines. Drug Discov Today. 2014 May;19(5):637–44.
10. Hodos RA, Kidd BA, Shameer K, Readhead BP, Dudley JT. In silico methods for drug repurposing and pharmacology. Wiley Interdiscip Rev Syst Biol Med. 2016 May;8(3):186–210.
11. Vilar S, Hripcsak G. The role of drug profiles as similarity metrics: applications to repurposing, adverse effects detection and drug-drug interactions. Brief Bioinform. 2017 Jul 1;18(4):670–81.
12. Loscalzo J, Kohane I, Barabasi A-L. Human disease classification in the postgenomic era: a complex systems approach to human pathobiology. Mol Syst Biol. 2007;3:124.
13. Duran-Frigola M, Mateo L, Aloy P. Drug repositioning beyond the low-hanging fruits. Curr Opin Syst Biol. 2017 Jun 1;3:95–102.
14. Li Y, Agarwal P, Rajagopalan D. A global pathway crosstalk network. Bioinforma Oxf Engl. 2008 Jun 15;24(12):1442–7.
15. Garcia-Garcia J, Guney E, Aragues R, Planas-Iglesias J, Oliva B. Biana: a software framework for compiling biological interactions and analyzing networks. BMC Bioinformatics. 2010 Jan 27;11(1):56.
16. Menche J, Sharma A, Kitsak M, Ghiassian S, Vidal M, Loscalzo J, et al. Uncovering disease-disease relationships through the incomplete human interactome. Science. 2015 Feb 20;347(6224):1257601.
17. Ko Y, Cho M, Lee J-S, Kim J. Identification of disease comorbidity through hidden molecular mechanisms. Sci Rep. 2016 19;6:39433.
18. Rubio-Perez C, Guney E, Aguilar D, Piñero J, Garcia-Garcia J, Iadarola B, et al. Genetic and functional characterization of disease associations explains comorbidity. Sci Rep. 2017 24;7(1):6207.

19. Cuadrado A, Manda G, Hassan A, Alcaraz MJ, Barbas C, Daiber A, et al. Transcription Factor NRF2 as a Therapeutic Target for Chronic Diseases: A Systems Medicine Approach. Pharmacol Rev. 2018 Apr;70(2):348–83.
20. Toro-Domínguez D, Carmona-Sáez P, Alarcón-Riquelme ME. Shared signatures between rheumatoid arthritis, systemic lupus erythematosus and Sjögren's syndrome uncovered through gene expression meta-analysis. Arthritis Res Ther. 2014 Dec 3;16(6):489.
21. Luan M, Shang Z, Teng Y, Chen X, Zhang M, Lv H, et al. The shared and specific mechanism of four autoimmune diseases. Oncotarget. 2017 Dec 12;8(65):108355–74.
22. American Autoimmune Related Diseases Association Autoimmune Disease Statistics [Internet]. American Autoimmune Related Diseases Association Autoimmune Disease Statistics. [cited 2018 Jun 13]. Available from: www.aarda.org/news-information/statistics
23. Baranzini SE. Chapter 70-Autoimmune Disorders. In: Genomic and Personalized Medicine. 2nd edition. Cambridge, MA, USA: Academic Press; 2013. p. 822–38.
24. Gottesman II, Gould TD. The endophenotype concept in psychiatry: etymology and strategic intentions. Am J Psychiatry. 2003 Apr;160(4):636–45.
25. Ghiassian SD, Menche J, Chasman DI, Giulianini F, Wang R, Ricchiuto P, et al. Endophenotype Network Models: Common Core of Complex Diseases. Sci Rep. 2016 09;6:27414.
26. Glaab E, Baudot A, Krasnogor N, Schneider R, Valencia A. EnrichNet: network-based gene set enrichment analysis. Bioinforma Oxf Engl. 2012 Sep 15;28(18):i451–7.
27. Himmelstein DS, Lizee A, Hessler C, Brueggeman L, Chen SL, Hadley D, et al. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. eLife. 2017 Sep 22;6.
28. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, et al. The Connectivity Map: using gene-expression signatures to connect

small molecules, genes, and disease. Science. 2006 Sep 29;313(5795):1929–35.

29. Sims-Robinson C, Kim B, Rosko A, Feldman EL. How does diabetes accelerate Alzheimer disease pathology? Nat Rev Neurol. 2010;6(10):551–9.
30. Hiltunen M, Khandelwal VKM, Yaluri N, Tiilikainen T, Tusa M, Koivisto H, et al. Contribution of genetic and dietary insulin resistance to Alzheimer phenotype in APP/PS1 transgenic mice. J Cell Mol Med. 2012 Jun;16(6):1206–22.
31. Yarchoan M, Arnold SE. Repurposing diabetes drugs for brain insulin resistance in Alzheimer disease. Diabetes. 2014 Jul;63(7):2253–61.
32. Du J, Wang Z. Therapeutic potential of lipase inhibitor orlistat in Alzheimer's disease. Med Hypotheses. 2009 Nov;73(5):662–3.
33. MahmoudianDehkordi S, Arnold M, Nho K, Ahmad S, Jia W, Xie G, et al. Altered Bile Acid Profile Associates with Cognitive Impairment in Alzheimer's Disease – An Emerging Role for Gut Microbiome. bioRxiv. 2018 Mar 17;281956.
34. Guerciolini R. Mode of action of orlistat. Int J Obes Relat Metab Disord J Int Assoc Study Obes. 1997 Jun;21 Suppl 3:S12-23.
35. Snowden SG, Ebshiana AA, Hye A, An Y, Pletnikova O, O'Brien R, et al. Association between fatty acid metabolism in the brain and Alzheimer disease neuropathology and cognitive performance: A nontargeted metabolomic study. PLoS Med. 2017 Mar;14(3):e1002266.
36. Daugherty D, Goldberg J, Fischer W, Dargusch R, Maher P, Schubert D. A novel Alzheimer's disease drug candidate targeting inflammation and fatty acid metabolism. Alzheimers Res Ther. 2017 Jul 14;9(1):50.
37. Ramos BP, Colgan LA, Nou E, Arnsten AFT. Beta2 adrenergic agonist, clenbuterol, enhances working memory performance in aging animals. Neurobiol Aging. 2008 Jul;29(7):1060–9.

38. Chai G-S, Wang Y-Y, Yasheng A, Zhao P. Beta 2-adrenergic receptor activation enhances neurogenesis in Alzheimer's disease mice. Neural Regen Res. 2016 Oct;11(10):1617–24.
39. Pan SJ, Hancock J, Ding Z, Fogt D, Lee M, Ivy JL. Effects of clenbuterol on insulin resistance in conscious obese Zucker rats. Am J Physiol Endocrinol Metab. 2001 Apr;280(4):E554-561.
40. Ceol A, Chatr Aryamontri A, Licata L, Peluso D, Briganti L, Perfetto L, et al. MINT, the molecular interaction database: 2009 update. Nucleic Acids Res. 2010 Jan;38(Database issue):D532-539.
41. Stark C, Breitkreutz B-J, Chatr-Aryamontri A, Boucher L, Oughtred R, Livstone MS, et al. The BioGRID Interaction Database: 2011 update. Nucleic Acids Res. 2011 Jan;39(Database issue):D698-704.
42. Prasad TSK, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, et al. Human Protein Reference Database--2009 update. Nucleic Acids Res. 2009 Jan;37(Database issue):D767-772.
43. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, et al. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D354–7.
44. Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, et al. Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proc Natl Acad Sci U S A. 2007 Feb 6;104(6):1777–82.
45. Ruepp A, Brauner B, Dunger-Kaltenbach I, Frishman G, Montrone C, Stransky M, et al. CORUM: the comprehensive resource of mammalian protein complexes. Nucleic Acids Res. 2008 Jan;36(Database issue):D646-650.
46. Hornbeck PV, Kornhauser JM, Tkachev S, Zhang B, Skrzypek E, Murray B, et al. PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. Nucleic Acids Res. 2012 Jan;40(Database issue):D261-270.

47. Leskovec J, Faloutsos C. Sampling from large graphs. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining [Internet]. New York, NY, USA: Association for Computing Machinery; 2006 [cited 2021 May 20]. p. 631–6. (KDD ’06). Available from: https://doi.org/10.1145/1150402.1150479
48. Erten S, Bebek G, Ewing RM, Koyutürk M. DADA: Degree-Aware Algorithms for Network-Based Disease Gene Prioritization. BioData Min. 2011 Jun 24;4:19.
49. Guney E, Oliva B. Analysis of the robustness of network-based disease-gene prioritization methods reveals redundancy in the human interactome and functional diversity of disease-genes. PloS One. 2014;9(4):e94686.
50. Guney E. Investigating Side Effect Modules in the Interactome and Their Use in Drug Adverse Effect Discovery. In: Complex Networks VIII [Internet]. Springer, Cham; 2017 [cited 2021 May 6]. p. 239–50. Available from: https://link-springer-com.sare.upf.edu/chapter/10.1007/978-3-319-54241-6_21
51. Guney E, Oliva B. Exploiting Protein-Protein Interaction Networks for Genome-Wide Disease-Gene Prioritization. PLoS ONE. 2012 Sep 21;7(9).
52. Piñero J, Bravo À, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, Centeno E, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res. 2017 04;45(D1):D833–9.
53. UniProt Consortium. UniProt: a hub for protein information. Nucleic Acids Res. 2015 Jan;43(Database issue):D204-212.
54. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. 2016 Jan 4;44(D1):D862-868.
55. Rath A, Olry A, Dhombres F, Brandt MM, Urbero B, Ayme S. Representation of rare diseases in health information systems: the

Orphanet approach to serve a wide range of end users. Hum Mutat. 2012 May;33(5):803–8.

56. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014 Jan;42(Database issue):D1001-1006.
57. Davis AP, Grondin CJ, Lennon-Hopkins K, Saraceni-Richards C, Sciaky D, King BL, et al. The Comparative Toxicogenomics Database's 10th year anniversary: update 2015. Nucleic Acids Res. 2015 Jan;43(Database issue):D914-920.
58. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2018 Jan 4;46(D1):D1074–82.
59. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinforma Oxf Engl. 2011 Jun 15;27(12):1739–40.
60. Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, et al. The Reactome pathway knowledgebase. Nucleic Acids Res. 2014 Jan;42(Database issue):D472-477.
61. Zhou X, Menche J, Barabási A-L, Sharma A. Human symptoms-disease network. Nat Commun. 2014 Jun 26;5:4212.
62. Hidalgo CA, Blumm N, Barabási A-L, Christakis NA. A Dynamic Network Approach for the Study of Human Phenotypes. PLoS Comput Biol [Internet]. 2009 Apr 10 [cited 2019 May 31];5(4). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2661364/
63. Kibbe WA, Arze C, Felix V, Mitraka E, Bolton E, Fu G, et al. Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. Nucleic Acids Res. 2015 Jan;43(Database issue):D1071-1078.
64. Clark NR, Ma'ayan A. Introduction to statistical methods for analyzing large data sets: gene-set enrichment analysis. Sci Signal. 2011 Sep 6;4(190):tr4.

65. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B Methodol. 1995;57(1):289–300.

66. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003 Nov;13(11):2498–504.

67. Vanderplas J. Python Data Science Handbook. Sebastopol, CA, USA: O'Reilly Media, Inc.; 2018.

68. Wickham H. ggplot2: Elegant Graphics for Data Analysis (Use R!). New York, NY, USA: Springer; 2009.

3.3. Network medicine tools applied to modelling the response of a drug combination in prototype-patients

In the third article of the thesis, I present a collaboration with Anaxomics Biotech S.L. where we apply the network medicine tools TPMS and GUILDify v2.0 to model the response of the drug combination sacubitril/valsartan towards the phenotypes of heart failure and macular degeneration using theoretical models called “prototype-patients”:

(1) TPMS allows to model all the possible mechanisms of action between the targets of the drug and the proteins modulated by a disease or side effect. TPMS simulates the transmission of the perturbation of the drug through the PPI network from the stimulus (the drug targets) until the response (the disease-associated proteins). The simulation is carried out by a Multilayer Perceptron algorithm, and the models are trained using restrictions from gene expression datasets.

(2) GUILDify v2.0 applies diffusion-based algorithms to identify the modules associated to a disease or side effect using a different network from TPMS, ideal to compare the results of both methods.

I carried out this work together with Guillem Jorba, the other first co-author of the publication. We employed TPMS to stratify different types of prototype-patients depending on how the intake of sacubitril/valsartan modulates the proteins associated to the phenotypes heart failure and macular degeneration. We identified biomarker proteins that allowed to differentiate such prototype-patients. We applied GUILDify v2.0 to identify the disease modules of heart failure and macular degeneration, assess how the target proteins of the drug combination were overlapping them and search the biomarker proteins identified by TPMS in this context.

Jorba G*, **Aguirre-Plans J***, Junet V, Segú-Vergés C, Ruiz JL, Pujol A, Fernandez-Fuentes N, Mas JM, Oliva B. In-silico simulated prototype-patients using TPMS technology to study a potential adverse effect of sacubitril and valsartan. ***PLoS One*.** 2020; 15(2): e0228926. DOI: 10.1371/journal.pone.0228926
*co-first authors

In-silico simulated prototype-patients using TPMS technology to study a potential adverse effect of sacubitril and valsartan

PLoS One. 2020 Feb 13;15(2):e0228926

Guillem Jorba[1,2,#], Joaquim Aguirre-Plans[2,#], Valentin Junet[1,3], Cristina Segú-Vergés[1], José Luis Ruiz[1], Albert Pujol[1], Narcis Fernandez-Fuentes[4], José Manuel Mas[1,*], Baldo Oliva[2,*]

[1]**Anaxomics Biotech SL**, Barcelona 08008, Catalonia, Spain
[2]**Structural Bioinformatics Group**, Research Programme on Biomedical Informatics, Department of Experimental and Health Science, Universitat Pompeu Fabra, Barcelona 08003, Catalonia, Spain
[3]**Institute of Biotechnology and Biomedicine**, Universitat Autònoma de Barcelona, Cerdanyola del Vallès 08193, Catalonia, Spain
[4]**Department of Biosciences**, U Science Tech, Universitat de Vic-Universitat Central de Catalunya, Vic 08500, Catalonia, Spain

[#]These authors contributed equally to this work and are considered to be co-first authors
***Corresponding authors**

Abstract

Unveiling the mechanism of action of a drug is key to understand the benefits and adverse reactions of a medication in an organism. However, in complex diseases such as heart diseases there is not a unique mechanism of action but a wide range of different responses depending on the patient. Exploring this collection of mechanisms is one of the clues for a future personalized medicine. The Therapeutic Performance Mapping System (TPMS) is a Systems Biology approach that generates multiple models of the mechanism of action of a drug. Each molecular mechanism generated could be associated to particular individuals, here defined as prototype-patients, hence the generation of models using TPMS technology may be used for detecting adverse effects to specific patients. TPMS operates by (1) modelling the responses in humans with an accurate description of a protein network and (2) applying a Multilayer Perceptron-like and sampling strategy to find all plausible solutions. In the present study, TPMS is applied to explore the diversity of mechanisms of action of the drug combination sacubitril/valsartan. We use TPMS to generate a wide range of models explaining the relationship between sacubitril/valsartan and heart failure (the indication), as well as evaluating their association with macular degeneration (a potential adverse effect). Among the models generated, we identify a set of mechanisms of action associated to a better response in terms of heart failure treatment, which could also be associated to macular degeneration development. Finally, a set of 30 potential biomarkers are proposed to identify mechanisms (or prototype-patients) more prone of suffering macular degeneration when presenting good heart failure response. All prototype-patients models generated are completely theoretical and

therefore they do not necessarily involve clinical effects in real patients. Data and accession to software are available at http://sbi.upf.edu/data/tpms/.

Introduction

Systems biology methods are an increasingly recurring strategy to understand the molecular effects of a drug in complex clinical settings (1). Some of these methods apply computer science techniques and mathematical approaches to simulate the responses of a drug. In 2005, the Virtual Physiological Human initiative was founded with the objective of developing computational models of patients (2). Later, they defined the concept of *In Silico* Clinical Trials as "the use of individualized computer simulation in the development or regulatory evaluation of a medicinal product, medical device, or medical intervention" (3). Since then, *In Silico* Clinical Trials have been adopted in several occasions in preclinical and clinical trials (1).

However, current methodologies do not consider the inter-patient variability intrinsic to pharmacological treatments, missing relevant information that should be incorporated into the models. Indeed, there are many parameters influencing the Mechanisms of Action (MoA) in such therapies, including demographic data of the patient, co-treatments or clinical history. Thus, by modelling all molecular mechanisms affected by the drug, the diversity of responses observed in patients during or after the treatment could be explained.

The Therapeutic Performance Mapping System (TPMS) (4) is a method used to elucidate all the possible MoAs that could exist between an input drug and a pathology or adverse effect. It is a systems biology approach based on the simulation of patient-specific protein-protein interaction networks. TPMS incorporates data from different resources and uses the information from the drugs and diseases under study to generate multiple models of potential MoAs. In the last years, TPMS has been broadly used in different clinical areas and with different objectives (5–12), in some cases being validated in the posterior experiments (6,11,12). Our working hypothesis is that a set of MoAs can represent the different responses to a drug in cells and that a real population of patients is the result of a myriad of cell responses. Thus, we define a prototype-patient as an abstract case with all cells responding to a single MoA.

Here, we propose the application of TPMS and protein-network approaches in the specific case study of the drug combination sacubitril/valsartan, used for the treatment of Heart Failure (HF). HF is becoming a major health problem in the western world due to its increasing hospitalization rates (13), with a prevalence being influenced by many factors like age, nutritional habits, lifestyles or genetics. This complicates the development of treatments and the identification of universal biomarkers to stratify the population. To facilitate this segmentation, it is necessary to understand the molecular details of the treatment and the pathology. Sacubitril/valsartan (marketed by Novartis as Entresto®) is a drug combination that shows better results than conventional treatments by reducing cardiovascular deaths and heart failure (HF) readmissions (14). In pharmacological terms, it is an angiotensin receptor-neprilysin inhibitor. Consequently, it triggers the natriuretic peptide system by inhibiting neprilysin (NEP) and inhibits renin-angiotensin-aldosterone system by blocking the type-1 angiotensin II receptor (AT1R) (15). In a previous work, TPMS was already applied to unveil the MoA of sacubitril/valsartan synergy, revealing its effect against two molecular processes (9): the left ventricular extracellular matrix remodeling, mediated by proteins like gap junction alpha-1 protein or matrix metalloproteinase-9; and the cardiomyocyte apoptosis, through modulation of glycogen synthase kinase-3 beta. However, several publications warned about the potential long-term negative implications of using a neprilysin inhibitor like sacubitril (15–19). Neprilysin plays a critical role at maintaining the amyloid-β homeostasis in the brain, and the alteration of amyloid-β levels has been linked to a potential long-term development of Alzheimer’s disease or Macular Degeneration (MD) (15,17,19–21). During the clinical trials PARADIGM-HF and PARAGON-HF with sacubitril/valsartan no serious effects were detected (14,22). Still, their patient follow-up was relatively short and not specialized in finding neurodegenerative specific symptoms. For this reason, in a forthcoming PERSPECTIVE trial (NCT02884206) a battery of cognitive tests was taken (18). In line with this, the application of systems biology methods may shed light to the potential relationship between the treatment and the adverse effect.

In this study, we used TPMS and GUILDify v2.0 to analyze the relationship between sacubitril/valsartan, HF and MD in entirely theoretical models, which could not necessarily involve clinical effects in real patients. We analyzed a population of MoAs that describe the possible protein links from a sacubitril/valsartan treatment to HF and MD phenotypes. We clustered the MoAs in groups according to their response intensity and labelled them as high or low efficacy of treating HF and possibility of causing MD. We then compared these sets of MoAs and proposed a list of biomarkers to identify potential cases of MD when using sacubitril/valsartan. Simultaneously, we used GUILDify v2.0 web server (23) as an alternative approach to compare the biomarkers proposed by TPMS and reinforce the results.

Materials and Methods

1. Biological Effectors Database (BED) to molecularly describe specific clinical conditions

Biological Effectors Database (BED) (5,24) describes more than 300 clinical conditions as sets of genes and proteins (effectors) that can be “active”, “inactive” or “neutral”. For example, in a metabolic protein-like network, an enzyme will become “active” in the presence of a catalyst, or become inactivated when interacting with an inhibitor (see further details in supplementary material).

2. TPMS modelling

The Therapeutic Performance Mapping System (TPMS) is a tool that creates mathematical models of the protein pathways underlying a drug/pathology to explain a clinical outcome or phenotype (4–10). These models find MoAs that explain how a *Stimulus* (i.e. proteins activated or inhibited by a drug) produces a *Response* (i.e. proteins active or inhibited in a phenotype). In the present case study, we applied TPMS to the drug-indication pair sacubitril/valsartan and HF.

Regarding the drug, we retrieved the sacubitril/valsartan targets from DrugBank (25), PubChem (26), STITCH (27), SuperTarget (28) and hand curated literature revision. As for the indication, we retrieved the proteins associated with the phenotype from the BED (5,24).

2.1. Building the Human protein network (HPN)

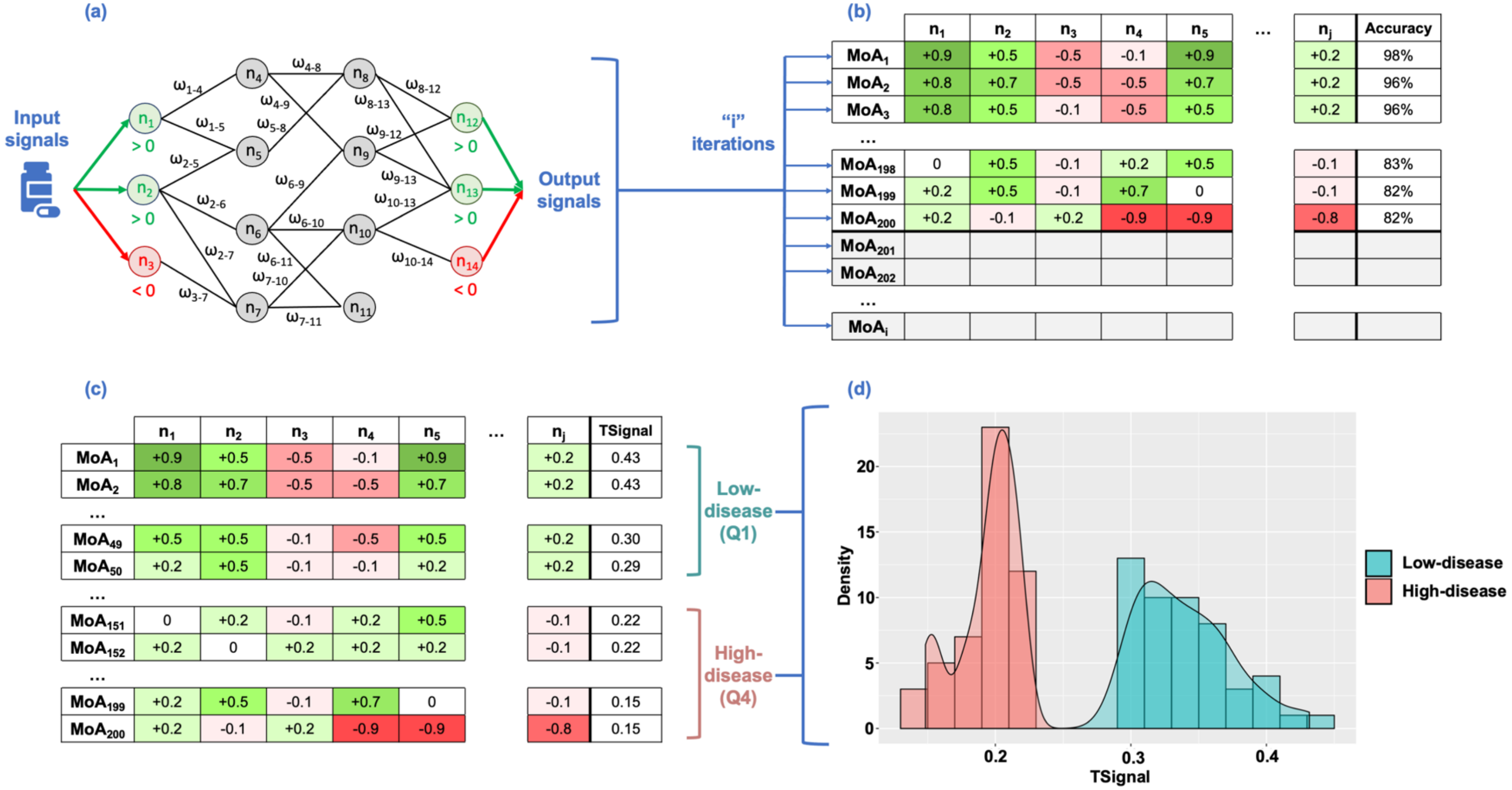
To apply the TPMS approach and create the mathematical models of MoAs, a Human Protein Network (HPN) is needed beforehand. In this study, we used a protein-protein interactions network created from the integration of public and private databases: KEGG (29), BioGRID (30), IntAct (31), REACTOME (32), TRRUST (33), and HPRD (34). In addition, information extracted from scientific literature, which was manually curated, was also included and used for trimming the network. The resulting HPN considers interactions corresponding to different tissues to take into account the effect of the *Stimulus* in the whole body.

2.2. Defining active/inactive nodes

We define the state of human proteins as active or inactive for a particular phenotype, including its expression (as active) or repression (as inactive) extracted from the GSE57345 gene expression dataset (35) as in *Iborra-Egea et al* (9) (see further details in supplementary material).

2.3. Description of the mathematical models

The algorithm of TPMS takes as input signals the activation (+1) and inactivation (-1) of the drug target proteins, and as output the BED protein states of the pathology. It then optimizes the paths between both protein sets and computes the activation and inactivation values of all proteins in the HPN. Each node of the protein network receives as input the output of the incoming connected nodes and every link is given a weight (ω_l). The sum of inputs is transformed by a hyperbolic tangent function that generates a score for every node, which becomes the “output signal” towards the outgoing connected nodes. The ω_l parameters are obtained by optimization, using a Stochastic Optimization Method based on Simulated Annealing (36). The models are then trained by using the general restrictions (i.e. defined as edges and nodes with the property of being active or inactive) and the specific conditions set by the user. Details of the approach are shown in **Fig 1** and supplementary material.



(b)

	n_1	n_2	n_3	n_4	n_5	…	n_j	Accuracy
MoA_1	+0.9	+0.5	-0.5	-0.1	+0.9		+0.2	98%
MoA_2	+0.8	+0.7	-0.5	-0.5	+0.7		+0.2	96%
MoA_3	+0.8	+0.5	-0.1	-0.5	+0.5		+0.2	96%
…								
MoA_{198}	0	+0.5	-0.1	+0.2	+0.5		-0.1	83%
MoA_{199}	+0.2	+0.5	-0.1	+0.7	0		-0.1	82%
MoA_{200}	+0.2	-0.1	+0.2	-0.9	-0.9		-0.8	82%
MoA_{201}								
MoA_{202}								
…								
MoA_i								

(c)

	n_1	n_2	n_3	n_4	n_5	…	n_j	TSignal
MoA_1	+0.9	+0.5	-0.5	-0.1	+0.9		+0.2	0.43
MoA_2	+0.8	+0.7	-0.5	-0.5	+0.7		+0.2	0.43
…								
MoA_{49}	+0.5	+0.5	-0.1	-0.5	+0.5		+0.2	0.30
MoA_{50}	+0.2	+0.5	-0.1	-0.1	+0.2		+0.2	0.29
…								
MoA_{151}	0	+0.2	-0.1	+0.2	+0.5		-0.1	0.22
MoA_{152}	+0.2	0	+0.2	+0.2	+0.2		-0.1	0.22
…								
MoA_{199}	+0.2	+0.5	-0.1	+0.7	0		-0.1	0.15
MoA_{200}	+0.2	-0.1	+0.2	-0.9	-0.9		-0.8	0.15

Fig 1. Scheme of how to apply TPMS to find the Mechanisms of Action (MoA) of a drug. (a) Scheme of the method, transmitting information over the Human Protein Network (HPN) using a Multilayer Perceptron-like and sampling. **(b)** After a given number of iterations, we obtain a collection of Mechanisms of Actions (MoA). Rows represent the MoAs and columns the output signal values of the proteins (nodes of the network). The final column shows the accuracy of the model as a percentage of the number restrictions accomplished. **(c)** 200 MoAs are selected (coloured in the slide) and sorted by TSignal. The first quartile is defined as the Low-disease group, and the fourth quartile as High-disease group. The distribution of the output signals of the two groups of MoA are shown in **(d)** (High-disease in red and Low-disease is in blue).

3. Measures to compare sets of MoAs

To understand the relationships between all potential mechanisms we defined some measures of comparison between different sets of solutions. We expect that a drug will revert the conditions of a disease phenotype; subsequently, a drug should inactivate the active protein effectors of a pathology-phenotype and activate the inactive ones. In this section we describe the measures used in the present study to analyze and compare sets of MoAs from different views (see further details in supplementary material).

3.1. TSignal

To quantify the intensity of the response of a MoA, we defined TSignal as the average signal arriving at the protein effectors (equation in supplementary material).

3.2. Distance between two sets of MoAs

We used the modified Hausdorff distance (MHD) introduced by Dubuisson and Jain (37) as the *distance* between two or more sets of MoAs in order to determine their similarity. Details of the equations are explained in the supplementary material.

3.3. Potential biomarkers extracted from MoAs

In order to extract potential biomarkers when comparing sets of MoAs, we first defined the *best-classifier proteins*. These are proteins inside the HPN that allow to better classify between groups of models and are identified following a Data-Science strategy (see supplementary material). Best-classifier proteins are usually strongly related to the intensity of a response and are proteins with values differently distributed between the groups of MoAs analyzed. For this study, and for the sake of simplicity, we focused only on the 200 proteins (or pair of proteins) showing the higher classification accuracy. Assuming the hypothesis that the selected MoAs are representative of individual prototype-patients, these proteins could be used as biomarkers to classify a cohort of patients.

Then, we applied the Mann-Whitney *U* test to compare the distributions of the best-classifier proteins values between the groups and selected those proteins with significant difference (p-value< 0.01). We also restricted the list to proteins having an average value with opposite sign among groups (i.e. positive vs. negative or vice versa) and named them as *differential best-classifier proteins*. By following this strategy, we can identify two groups of differential best-classifier proteins: those active in the first group (positive output signal in average) and inactive in the other (negative output signal in average), and the opposite.

Results and discussion

We applied TPMS to the HPN using as input signals the drug targets of sacubitril/valsartan (NEP / AT1R) and as output signals the proteins associated with HF extracted from the BED. Out of all MoAs found by TPMS, we selected the 200 satisfying the largest number of restrictions (and at least 80% of them) to perform further analysis.

Note that TPMS was only executed once, optimizing the results to satisfy the restrictions on HF data. The values of MD are obtained by measuring the signal arriving at the MD effectors, which are part of the HPN and also receive signal. This procedure was chosen because we defined HF as the indication of the drug (sacubitril/valsartan), while MD is a potential adverse effect.

1. Stratification of MoAs

In order to compare models related to a good or bad response to the treatment, or those more prone to lead towards potential MD adverse effect, we stratified the MoAs. For HF, or treatment response, MoAs were ranked by their TSignal and then split in four quartiles. The first quartile (top 25%) contains MoAs with higher intensity of the response, which in turn corresponds to lower values of the effectors associated with HF phenotype (we named them as “Low”-disease MoAs). On the contrary, the fourth quartile (bottom 25%) collects

MoAs with lower intensity of response (thus, we named as “High”-disease MoAs) (**Supplementary Fig 1a**). On the other hand, for MD, the first quartile (top 25%) contains MoAs with higher intensity, which as an adverse event, correspond to models with high values of the effectors associated to MD (we named them as High-adverseEvent MoAs). The fourth quartile (bottom 25%) collects MoAs with lower intensity of response (thus, we named as Low- adverseEvent MoAs) (**Supplementary Fig 1b**). Note that, in the following steps and because HF and MD groups were extracted from the same 200 set of models, common MoAs between different HF and MD-defined sets could be expected.

2. Comparison of MoAs with high/low TSignal associated to HF or MD

We calculated the modified Hausdorff distance between the groups of MoAs (High-MD, Low-MD, High-HF and Low-HF) to elucidate their similarity values (**Supplementary Table 5**). In this sense, the higher the distance between the groups is, the more different they are. We used these distances to calculate a dendrogram tree (see **Supplementary Fig 2**) showing that MoAs associated with a bad response to sacubitril/valsartan for HF (high-HF) are more similar (i.e. closer) to MoAs linked to a stronger MD adverse effect (high-MD). It is remarkable that the distances between Low- and High-HF and between Low- and High-MD are larger than the cross distances between HF and MD. However, by the definition of distance (equation 3 in supplementary material), it cannot account for the dispersion among the MoAs within and between each group. Therefore, for each set we calculated the mean Euclidean distance between all the points and its center, defined by the average of all points (see **Supplementary Table 6**). As a result, all groups showed very similar dispersion values.

In order to have a global and graphical view of the distance between the individual MoAs, we generated a multidimensional scaling (MDS) plot calculated using MATLAB (see **Fig 2**). MDS plots display the pairwise distances in two dimensions while preserving the clustering characteristics (i.e. close MoAs are also close in the 2D-plot and far MoAs are also far in 2D). Focusing on the Low-HF group depicted in blue circles, we observe that there is no clear tendency to cluster with any of the MD groups. There are few cases of Low-HF MoAs coinciding in the space with Low- or High-MD MoAs. This implies that a good response to sacubitril/valsartan of HF patients would not be usually linked to the development of MD. Moreover, no clear distinction is found when plotting only the MD MoAs within the Low-HF group (see **Supplementary Fig 3a**). However, regarding the set of High-HF MoAs, we can differentiate two clusters of MoAs: one related to the High-MD group (green crosses); and the other close to MoAs of the Low-MD group (black crosses) (see **Supplementary Fig 3b**).

Assuming the hypothesis that different MoAs correspond to distinct prototype-patients, we conclude that for the specific set of patients for which sacubitril/valsartan works best reducing HF, it would be more difficult to differentiate between those presenting MD and those who do not. Instead, for the High-HF group, patients having MD could indeed be easily distinguished from those not presenting MD as side effect. However, because Low-HF group has more relevance to the clinics, specific functional analyses were performed in this specific group, as seen in following sections. Finally, we highlight that, as these distinct groups of prototype-patients are theoretical simulations, they may not be reflecting the clinical effects of the real patients.

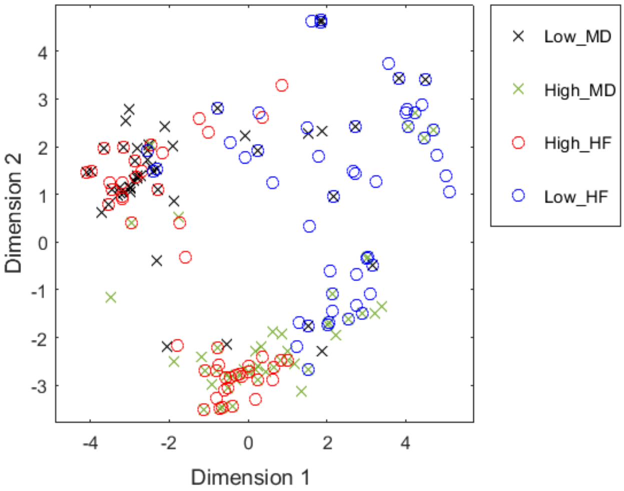


Fig 2. Multidimensional scaling plot of the distances between the Mechanisms of Action (MoA) of the four groups defined. Each point represents a MoA. Axes are defined by the most representative dimensions.

3. Identification and functional analysis of potential biomarkers

For this section, we identified the nodes (i.e. proteins) significantly differentiating two groups of models (using a Mann-Whitney *U* test) for which the average of output signals have opposite signs (see methods in 3.3). After that, the function of the identified proteins was extracted from Gene Ontology (GO).

3.1. Identification of best-classifier proteins differentiating HF responses

After comparing High- vs Low- HF groups, we found a total of 45 differential best-classifier proteins associated with the treatment response (6 Low-HF-active/High-HF-inactive and 39 Low-HF-inactive/High-HF-active) (see **Fig 3a** and **Supplementary Table 1**). To pinpoint the biological role of these proteins, we first identified the GO enriched functions (see **Supplementary Table 2)** and then searched in the literature for evidences linking them with HF. As a result, we found that the differential best-classifier proteins Low-HF-active/High-HF-inactive point towards an important role for actin nucleation and polymerization mechanisms in drug response (reflected by the functions *regulation of actin nucleation*, *regulation of Arp2/3 complex-mediated actin nucleation*, *SCAR complex*, *filopodium tip*, or *dendrite extension*). In fact, the alteration of actin nucleation and polymerization mechanisms has been reported in heart failure (38–40). Interestingly, a role for the activation of another differential best-classifier candidate, ATGR2, has been proposed to mediate some of the beneficial effects of angiotensin II receptor type 1 antagonists, such as valsartan (41,42). On the other hand, the results of the differential best-classifier proteins Low-HF-inactive/High-HF-active are linked to phosphatidylinositol kinase mediated pathways (*phosphatidylinositol-3,4-bisphosphate 5-kinase activity*) and MAP kinase mediated pathways (*MAP kinase kinase activity*, best classifier proteins MAPK1, MAPK3, MAPK11, MAPK12 or MAPK13). In this case, both signaling pathways have been associated to cardiac hypertrophy and subsequent heart failure (43,44). These outcomes clearly lead towards the idea that High-HF models are a representation of prototype-patients with a worst response to the treatment, while Low-HF models are related to more beneficial response to the medication. A more detailed explanation can be found in the supplementary material.

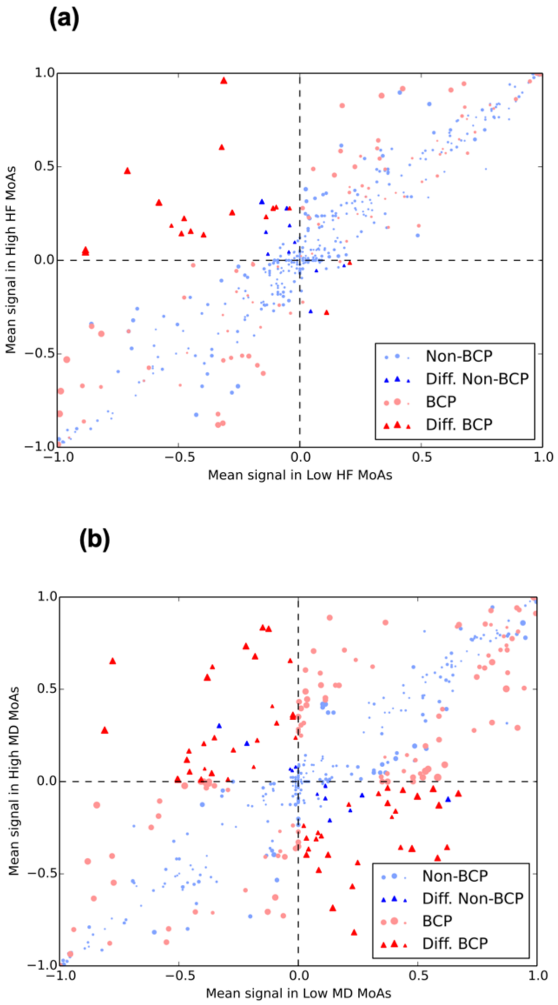


Fig 3. Scatter plot of the mean signal values of Low and High-"disease" Mechanisms of Action (MoA). Scatter plot of the mean signal values of Low-"disease" and High-"disease" MoAs for each protein using as disease Heart Failure (HF) in **(a)** and Macular Degeneration (MD) in **(b)**. The average of the output signal of each protein in High-group is presented versus its value in Low-group. Differential signals (Diff., shown as triangles) are defined as those with opposite sign when comparing High versus Low

average, and a p-value < 0.01 when calculating the Mann-Whitney *U* test between the two distributions of signals. Best-classifier proteins (BCP) are colored in red, otherwise they are blue. Sizes of markers are proportional to p-values of the Mann-Whitney *U* test.

3.2. Identification of best-classifier proteins differentiating MD responses

We identified 57 differential best-classifier proteins of MD (28 Low-MD-active/High-MD-inactive and 29 Low-MD-inactive/High-MD-active) (see **Fig 3b** and **Supplementary Table 3**). Again, we searched for relationships between these proteins and MD by identifying the GO enriched functions (see **Supplementary Table 4)** and searching for links in the literature. Some of the proteins and functions highlighted in the current analysis had been related to MD in previous works. The presence of dendritic spine development and dorsal/ventral axon guidance related proteins emphasizes the role of sacubitril/valsartan in dendritic and synaptic plasticity mechanisms, which had been previously linked to MD (45). Furthermore, valsartan treatment has been reported to promote dendritic spine development in other related neurodegenerative diseases, such as Alzheimer's disease (46). Other enriched functions are implicated in growth factor related pathways, which are known to be involved in wet MD pathogenesis (47). Moreover, neovascularization in the wet variant of MD has been linked to the signaling of some of the growth factors detected as sacubitril/valsartan-associated MD classifiers in this study, including FGF1 (47) and PDGF (48,49). A more detailed explanation can be found in the supplementary material.

3.3. Identification of potential biomarkers differentiating MD responses in Low-HF

Because of its clinical relevance, we decided to focus on analyzing the special case of prototype-patients in which the treatment reduces HF (Low-HF) but produces MD adverse effect (High-HF). In order to find these prototype-patients, we: (i) identified 13 Low-HF ∩ Low-MD MoAs and 12 Low-HF ∩ High-MD MoAs; and (ii) compared the protein signal of the two groups and proposed 30 potential biomarkers (**Table 1**). Among the proposed biomarkers, we found 16 proteins active in Low-HF ∩ Low-MD MoAs but inactive in Low-HF ∩ High-MD (15 of them shared with MD best-classifier proteins). On the other hand, 14 proteins were identified as inactive in Low-HF ∩ Low-MD and active in Low-HF ∩ High-MD MoAs (12 of them were MD best-classifier proteins). We calculated the GO enriched functions of these two groups and observed that "phosphatidylinositol bisphosphate kinase activity" is enriched among proteins that are active in Low-HF ∩ Low-MD MoAs. Instead, "fibrinolysis" was found to be enriched among proteins active in Low-HF ∩ High-MD MoAs (**Table 2**). With this, we conclude that among the group of prototype-patients for which sacubitril/valsartan improves HF treatment response, the modulation of fibrinolysis could play a role at inducing the MD adverse effect. Moreover, we propose 12 best-classifier proteins that may be considered as biomarkers for good prognosis of the side effect.

In fact, since neovascular MD development is characterized by subretinal extravasations of novel vessels derived from the choroid (CNV) and the subsequent hemorrhage into the photoreceptor cell layer in the macula region (51), it might be reasonable to think that the modulation of fibrinolysis and blood coagulation pathways could play a role. The reported implication of some fibrinolysis related classifiers, such as FGB, SERPINE1 (PAI-1), and SERPING1, in neovascular MD development seems to support this hypothesis (52–54). Besides, valsartan might be implicated in this mechanism, since it has been reported to modulate PAI-1 levels and promote fibrinolysis in different animal and human models (55,56). In addition, the presence of several other MD related classifiers in this list, such as IRS2 (57), PTGS2 (58), DCN (59) and FGF1 (60), further supports the interest of the classifiers as biomarkers of MD development in sacubitril/valsartan good responders. Still, we would like to highlight that the biomarkers have been proposed using a theoretical approach, and that the clinical effects studied may not be present in real patients.

Table 1. Potential biomarker proteins, with opposite signal in Low-HF ∩ Low-MD and Low-HF ∩ High-MD MoAs.

	Uniprot ID	Gene symbol	Gene name	$\langle LMD \rangle$	$\langle HMD \rangle$	$\sqrt{\left\lvert LMDx\ HMD \right\rvert}$	Adjusted P-value	BCP
1	**P02675**	**FGB**	**Fibrinogen beta chain**	-0.576	0.814	0.685	1.297E-03	MD
2	O43639	NCK2	Cytoplasmic protein NCK2	0.620	-0.697	0.657	1.656E-04	MD
3	P54762	EPHB1	Ephrin type-B receptor 1	0.317	-0.677	0.464	3.669E-04	HF& MD
4	Q9Y4H2	IRS2	Insulin receptor substrate 2	0.417	-0.465	0.440	8.181E-04	MD
5	O60674	JAK2	Tyrosine-protein kinase JAK2	-0.747	0.249	0.431	1.656E-04	MD
6	P06241	FYN	Tyrosine-protein kinase Fyn	0.591	-0.236	0.373	2.466E-04	HF& MD
7	P30530	AXL	Tyrosine-protein kinase receptor UFO	0.392	-0.330	0.360	2.111E-04	MD
8	**Q02297**	**NRG1**	**Pro-neuregulin-1, membrane-bound isoform**	0.672	-0.188	0.355	2.111E-04	MD
9	P32004	L1CAM	Neural cell adhesion molecule L1	-0.373	0.309	0.339	1.297E-03	HF& MD

10	Q05586	GRIN1	Glutamate receptor ionotropic, NMDA 1	-0.174	0.620	0.329	1.955E-04	MD
11	**P05230**	**FGF1**	**Fibroblast growth factor 1**	-0.152	0.688	0.323	8.181E-04	HF& MD
12	**P18084**	**ITGB5**	**Integrin beta-5**	0.436	-0.236	0.321	2.111E-04	MD
13	**P01583**	**IL1A**	**Interleukin-1 alpha**	0.174	-0.472	0.287	1.955E-04	MD
14	P10275	AR	Androgen receptor	0.349	-0.201	0.265	8.008E-04	MD
15	P15941	MUC1	Mucin-1 subunit alpha	0.099	-0.652	0.254	6.905E-04	HF& MD
16	O14757	CHEK1	Serine/threonine-protein kinase Chk1	0.436	-0.142	0.248	1.549E-03	MD
17	P15391	CD19	B-lymphocyte antigen CD19	-0.131	0.357	0.216	8.160E-03	MD
18	**P61981**	**YWHAG**	**14-3-3 protein gamma, N-terminally processed**	0.174	-0.236	0.203	2.783E-03	-
19	Q9Y478	PRKAB1	5'-AMP-activated protein kinase subunit beta-1	0.261	-0.142	0.192	5.682E-03	MD
20	P62158	CALM1 ; CALM2 ; CALM3	Calmodulin-1 {ECO:0000312|HGNC:HGNC:1442}	-0.282	0.107	0.174	9.405E-03	MD
21	**P06748**	**NPM1**	**Nucleophosmin**	0.261	-0.107	0.167	3.618E-03	MD
22	O15357	INPPL1	Phosphatidylinositol 3,4,5-trisphosphate 5-phosphatase 2	-0.261	0.094	0.157	3.618E-03	MD
23	P17081	RHOQ	Rho-related GTP-binding protein RhoQ	-0.218	0.094	0.143	9.794E-03	MD
24	P35354	PTGS2	Prostaglandin G/H synthase 2	0.044	-0.472	0.143	3.669E-04	MD
25	P42684	ABL2	Abelson tyrosine-protein kinase 2	-0.218	0.094	0.143	9.794E-03	MD
26	**Q15109**	**AGER**	**Advanced glycosylation end product-specific receptor**	-0.267	0.063	0.130	8.160E-03	-
27	P07585	DCN	Decorin	-0.044	0.236	0.101	5.682E-03	MD
28	**P05155**	**SERPING1**	**Plasma protease C1 inhibitor**	-0.044	0.236	0.101	5.682E-03	MD
29	**P05121**	**SERPINE1**	**Plasminogen activator inhibitor 1**	-0.044	0.236	0.101	5.682E-03	-
30	P14770	GP9	Platelet glycoprotein IX	0.044	-0.236	0.101	5.682E-03	MD

Highlighted cells correspond to proteins that are part of the Top-HF ∪ Top-MD ∪ Top-Drug set, the top-scoring proteins according to GUILDify. Columns show: the protein name (as UniprotID, gene-symbol and gene-name), the average of the signal in in Low-MD (<LMD>) and High-MD (<HMD>) in the selected sets of MoAs and a measure of the strength of the signal in both distributions (calculated as $\sqrt{\boldsymbol{LMDxHMD}}$), the significance (adjusted P-value) ensuring that both distributions of signals are different, and whether the protein has been considered best-classifier in MD of HF (BCP).

Table 2. Top 10 gene Ontology functions enriched from proteins with opposite signal in Low-HF ∩ Low-MD and Low-HF ∩ High-MD MoAs.

	Low-HF ∩ LMD+ HMD-			Low-HF ∩ HMD+ LMD-			Overlapped functions		
	GO name	**LOD**	**P-val.**	**GO name**	**LOD**	**P-val.**	**GO name**	**LOD**	**P-val.**
1	phosphatidylinositol-4,5-bisphosphate 3-kinase activity	1.89	0.03600	fibrinolysis	2.51	0.00050	response to stimulus	1.19	<0.00050
2	cellular response to UV	1.87	0.04200	negative regulation of wound healing	2.13	0.00050	positive regulation of transport	1.24	<0.00050
3	phosphatidylinositol bisphosphate kinase activity	1.87	0.04200	negative regulation of blood coagulation	2.12	0.00850	positive regulation of biological process	1.13	0.00051
4	vascular endothelial growth factor receptor signaling pathway	1.86	0.04200	negative regulation of hemostasis	2.12	0.00850	positive regulation of developmental process	1.18	<0.00050
5	positive regulation of protein kinase B signaling	1.70	0.01050	negative regulation of coagulation	2.10	0.01050	positive regulation of cellular process	1.04	0.00294
6	negative regulation of apoptotic signaling pathway	1.68	0.00050	platelet alpha granule lumen	1.96	0.02300	positive regulation of response to stimulus	1.04	0.00417
7	peptidyl-tyrosine phosphorylation	1.63	0.01400	regulation of epithelial cell apoptotic process	1.96	0.02300	-	-	-
8	regulation of apoptotic signaling pathway	1.63	<0.00050	regulation of blood coagulation	1.91	0.02800	-	-	-
9	peptidyl-tyrosine modification	1.62	0.01400	regulation of hemostasis	1.91	0.02800	-	-	-
10	protein tyrosine kinase activity	1.61	0.01850	regulation of coagulation	1.89	0.03450	-	-	-

Functional enrichment analysis from FuncAssociate (50).

4. Analysis of proposed biomarkers with GUILDify

In the previous section, we proposed 30 proteins that could potentially help to identify HF patients at risk of developing MD. To corroborate these biomarkers, we tested how many of them are found using a different approach also based on the use of functional networks. For this purpose, we used GUILDify v2.0 (23), a web server that extends the information of disease-gene associations through the protein-protein interactions network. GUILDify scores proteins according to their proximity with the genes associated with a disease (seeds). Using this web server, we identify a list of top-scoring proteins that are critical on transmitting the perturbation of disease genes through the network. The network used by GUILDify is completely independent from the HPN used in the TPMS, becoming an ideal, independent context to test the potential biomarkers.

Thus, we used GUILDify to indicate which of the potential biomarkers identified by TPMS may have a relevant role in the molecular mechanism of the drug. We ran GUILDify using the two targets of sacubitril/valsartan (NEP, AT1R) as seeds, and selected the top 2% scored nodes (defined as the “top-drug” set). We did the same with the phenotypes of HF and MD, using as seeds the 124 effectors of HF and 163 effectors of MD from the BED database. We merged the top scored sets of HF, MD and top-drug (“top-drug ∪ top-HF ∪ top-MD”) and studied the overlap with the set of 30 biomarkers proposed in the previous section. 10 of the candidate biomarkers are found in the merged set “top-drug ∪ top-HF ∪ top-MD” and are consequently significant (see **Supplementary Tables 10 and 11**).

Some of these candidates can be functionally linked to both diseases and the drug under study. For example, among these 10 classifiers, AGER has been implicated in both HF (61), through extracellular matrix remodeling, and MD development (62), through inflammation, oxidative stress, and basal laminar deposit formation between retinal pigment epithelium cells and the basal membrane; furthermore, this receptor is known to be modulated by AT1R (63), valsartan target. Similarly, FGF1 has been proposed to improve cardiac function after

HF (64), as well as to promote choroid neovascularization leading to MD (47). Moreover, FGF1 is regulated by angiotensin II through ATGR2 (65), another protein suggested as classifier in the current analysis that is known to mediate some of the effects of AT1R antagonists, such as valsartan (41,42). Another candidate, NRG1, has been linked to myocardial regeneration after HF (66) and is known to lessen the development of neurodegenerative diseases such as Alzheimer’s disease (67), which shares similar pathological features with MD (68). NRG1 is also linked to the expression of neprilysin (67), sacubitril target. ITGB5 has been identified as risk locus for HF (69) and its modulation has been linked to lipofucsin accumulation in MD (70). Interestingly, ATGR1 inhibitors have been reported to modulate ITGB5 expression in animal models (71). Finally, IL1A has been proposed as an essential mediator of HF pathogenesis (72,73) through inflammation modulations, and serum levels of this protein have been found increased in MD patients (74). In addition, as described in previous sections, classifiers FGB, SERPINE1, and SERPING1 have been linked to MD (52–54) and are also known to play a role in HF development (75–78). According to these findings, the 10 potential biomarkers proposed by TPMS and identified with GUILDify might be prioritized when studying good responder HF patients at risk of MD development.

Limitations

Although TPMS returns the amount of signal from the drug arriving to the rest of the proteins in the HPN, this signal is only a qualitative measure. We are not using data about the dosage of the drug or the quantity of expression of the proteins. However, we are already working to make TPMS move towards the growing tendency of Quantitative Systems Pharmacology. The quantification of the availability of drugs in the target tissue for each patient opens the opportunity to have an accurate patient simulation to do *in silico* clinical trials.

Conclusions

It exists an increasing need for new tools to get closer to real life clinical problems and the Systems Biology-based computational methods could be the solution needed. The specific case of sacubitril/valsartan stands out because of the amount of resources invested in the safety of the drug and the concern on the possible risk of inducing amyloid accumulation-associated conditions, such as macular degeneration (MD), in the long term. In this study, we applied TPMS technology to uncover different Mechanisms of Action (MoAs) of sacubitril/valsartan over heart failure (HF) and reveal its molecular relationship with MD. For this approach, we hypothesize that each MoA would correspond to a prototype-patient. The method is then used to generate a wide battery of MoAs by performing an *in silico* trial of the drug and pathology under study. TPMS computes the models by using a hand curated Human Protein Network and applying a Multilayer Perceptron-like and sampling method strategy to find all plausible solutions. After analyzing the models generated, we found different sets of proteins able to classify the models according to HF treatment efficacy or MD treatment relationship. The sets include functions such as PI3K and MAPK kinase signaling pathways, involved in HF-related cardiac hypertrophy, or fibrinolysis and coagulation processes (e.g. FGB, SERPINE1 or SERPING1) and growth factors (e.g. FGF1 or PDGF) related to MD induction. Furthermore, we propose 30 biomarker candidates to identify patients potentially developing MD under a successful treatment with sacubitril/valsartan. Out of this 30, 10 biomarkers were also found in the alternative, independent molecular context proposed by GUILDify, including some HF and MD effectors such as AGER, NRG1, ITGB5 or IL1A. Further studies might prospectively validate the herein raised hypothesis.

We would like to highlight that the models generated with TPMS are completely theoretical and thus, they do not necessarily reflect the real clinical effects. Consequently, the biomarkers proposed on the basis of these models are also theoretical and would require an experimental validation. Still, TPMS represents a huge improvement for studying the hypothetical relationship between a drug and an adverse effect. Until now, there were not enough tools that allow to perform an exhaustive study on the MoAs of an adverse effect. Now, with the MoAs and biomarkers proposed by TPMS, we provide a starting point in this type of research.

Acknowledgements

Public funders provided support for authors salaries: JAP, NFF and BO received support from the Spanish Ministry of Economy (MINECO) [BIO2017-85329-R] [RYC-2015-17519]; “Unidad de Excelencia María de Maeztu”, funded by the Spanish Ministry of Economy [ref: MDM-2014-0370]. The Research Programme on Biomedical Informatics (GRIB) is a member of the Spanish National Bioinformatics Institute (INB), PRB2-ISCIII and is supported by grant

PT13/0001/0023, of the PE I+D+i 2013-2016, funded by ISCIII and FEDER. GJ has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765912. VJ is part of a project (COSMIC; www.cosmic-h2020.eu) that has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765158.

References

1. Pappalardo F, Russo G, Tshinanu FM, Viceconti M. In silico clinical trials: concepts and early adoptions. Brief Bioinform. 2018;(March):1–10.
2. Viceconti M, Clapworthy G. The virtual physiological human: Challenges and opportunities. In: 2006 3rd IEEE International Symposium on Biomedical Imaging: From Nano to Macro - Proceedings. 2006. p. 812–5.
3. Viceconti M, Henney A, Morley-Fletcher E. In silico clinical trials: how computer simulation will transform the biomedical industry. Int J Clin Trials. 2016;3(2):37.
4. Anaxomics Biotech SL. TPMS technology [Internet]. 2018. Available from: http://www.anaxomics.com/tpms.php
5. Pujol A, Mosca R, Farrés J, Aloy P. Unveiling the role of network and systems biology in drug discovery. Trends Pharmacol Sci. 2010;31(3):115–23.
6. Herrando-Grabulosa M, Mulet R, Pujol A, Mas JM, Navarro X, Aloy P, et al. Novel Neuroprotective Multicomponent Therapy for Amyotrophic Lateral Sclerosis Designed by Networked Systems. PloS One. 2016;11(1):e0147626.

7. Gómez-Serrano M, Camafeita E, García-Santos E, López JA, Rubio MA, Sánchez-Pernaute A, et al. Proteome-wide alterations on adipose tissue from obese patients as age-, diabetes- and gender-specific hallmarks. Sci Rep. 2016;6(January):1–15.
8. Perera S, Artigas L, Mulet R, Mas JM, Sardón T. Systems biology applied to non-alcoholic fatty liver disease (NAFLD): treatment selection based on the mechanism of action of nutraceuticals. Nutrafoods. 2014;13(2):61–8.
9. Iborra-Egea O, Gálvez-Montón C, Roura S, Perea-Gil I, Prat-Vidal C, Soler-Botija C, et al. Mechanisms of action of sacubitril/valsartan on cardiac remodeling: a systems biology approach. Npj Syst Biol Appl. 2017;3(1):1–8.
10. Romeo-Guitart D, Forés J, Herrando-Grabulosa M, Valls R, Leiva-Rodríguez T, Galea E, et al. Neuroprotective Drug for Nerve Trauma Revealed Using Artificial Intelligence. Sci Rep. 2018;8:1879.
11. Lorén V, Garcia-Jaraquemada A, Naves JE, Carmona X, Mañosa M, Aransay AM, et al. Anp32e, a protein involved in steroid-refractoriness in ulcerative colitis, identified by a systems biology approach. J Crohns Colitis. 2019;13(3):351–61.
12. Iborra-Egea O, Santiago-Vacas E, Yurista SR, Lupón J, Packer M, Heymans S, et al. Unraveling the Molecular Mechanism of Action of Empagliflozin in Heart Failure With Reduced Ejection Fraction With or Without Diabetes. JACC Basic Transl Sci. 2019;4(7):831–40.
13. Van Riet EES, Hoes AW, Wagenaar KP, Limburg A, Landman MAJ, Rutten FH. Epidemiology of heart failure: The prevalence of heart failure and ventricular dysfunction in older adults over time. A systematic review. Eur J Heart Fail. 2016;18(3):242–52.
14. McMurray JJV, Packer M, Desai AS, Gong J, Lefkowitz MP, Rizkala AR, et al. Angiotensin–Neprilysin Inhibition versus Enalapril in Heart Failure. N Engl J Med. 2014;371(11):993–1004.
15. Singh JSS, Burrell LM, Cherif M, Squire IB, Clark AL, Lang CC. Sacubitril/valsartan: Beyond natriuretic peptides. Heart. 2017;103(20):1569–77.

16. Ponikowski P, Voors AA, Anker SD, Bueno H, Cleland JGF, Coats AJS, et al. 2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure. Eur Heart J. 2016;37(27):2129–2200m.
17. Feldman AM, Haller JA, DeKosky ST. Valsartan/sacubitril for heart failure: Reconciling disparities between preclinical and clinical investigations. JAMA - J Am Med Assoc. 2016;315(1):25–6.
18. Riddell E, Vader JM. Potential Expanded Indications for Neprilysin Inhibitors. Current Heart Failure Reports. 2017. p. 134–45.
19. Zhang Z lu, Li R, Yang FY, Xi L. Natriuretic peptide family as diagnostic/prognostic biomarker and treatment modality in management of adult and geriatric patients with heart failure: Remaining issues and challenges. J Geriatr Cardiol. 2018;15(8):540–6.
20. Baranello RJ, Bharani KL, Padmaraju V, Chopra N, Lahiri DK, Greig NH, et al. Amyloid-Beta Protein Clearance and Degradation (ABCD) Pathways and their Role in Alzheimer’s Disease. Curr Alzheimer Res. 2015;12(1):32–46.
21. Ohno-Matsui K. Parallel findings in age-related macular degeneration and Alzheimer’s disease. Prog Retin Eye Res. 2011;30(4):217–38.
22. Solomon SD, Rizkala AR, Gong J, Wang W, Anand IS, Ge J, et al. Angiotensin Receptor Neprilysin Inhibition in Heart Failure With Preserved Ejection Fraction: Rationale and Design of the PARAGON-HF Trial. JACC Heart Fail. 2017;5(7):471–82.
23. Aguirre-Plans J, Piñero J, Sanz F, Furlong LI, Fernandez-Fuentes N, Oliva B, et al. GUILDify v2.0: A Tool to Identify Molecular Networks Underlying Human Diseases, Their Comorbidities and Their Druggable Targets. J Mol Biol. 2019;30117–2.
24. Anaxomics Biotech SL. Biological Effectors Database [Internet]. 2018. Available from: http://www.anaxomics.com/biological-effectors-database.php

25. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: A major update to the DrugBank database for 2018. Nucleic Acids Res. 2018;46(D1):D1074–82.

26. Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, et al. PubChem substance and compound databases. Nucleic Acids Res. 2016;44(D1):D1202–13.

27. Szklarczyk D, Santos A, Von Mering C, Jensen LJ, Bork P, Kuhn M. STITCH 5: Augmenting protein-chemical interaction networks with tissue and affinity data. Nucleic Acids Res. 2016;44(D1):D380–4.

28. Hecker N, Ahmed J, von Eichborn J, Dunkel M, Macha K, Eckert A, et al. SuperTarget goes quantitative: update on drug-target interactions. Nucleic Acids Res. 2011;40(D1):D1113–7.

29. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: New perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017;45(D1):D353–61.

30. Chatr-Aryamontri A, Oughtred R, Boucher L, Rust J, Chang C, Kolas NK, et al. The BioGRID interaction database: 2017 update. Nucleic Acids Res. 2017;45(D1):D369–79.

31. Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, et al. The MIntAct project - IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res. 2014;42(D1):358–63.

32. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, et al. The Reactome Pathway Knowledgebase. Nucleic Acids Res. 2018;46(D1):D649–55.

33. Han H, Cho JW, Lee S, Yun A, Kim H, Bae D, et al. TRRUST v2: An expanded reference database of human and mouse transcriptional regulatory interactions. Nucleic Acids Res. 2018;46(D1):D380–6.

34. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, et al. Human Protein Reference Database - 2009 update. Nucleic Acids Res. 2009;37(D1):D767–72.

35. Liu Y, Morley M, Brandimarto J, Hannenhalli S, Hu Y, Ashley EA, et al. RNA-Seq identifies novel myocardial gene expression signatures of heart failure. Genomics. 2015;105(2):83–9.
36. Collet P, Rennard J-P. Stochastic Optimization Algorithms. Intell Inf Technol. 2011;1121–37.
37. Dubuisson M-P, Jain AK. A modified Hausdorff distance for object matching. Proc 12th Int Conf Pattern Recognit. 1994;1(1):566–8.
38. Patel VB, Wang Z, Fan D, Zhabyeyev P, Basu R, Das SK, et al. Loss of p47phox subunit enhances susceptibility to biomechanical stress and heart failure because of dysregulation of cortactin and actin filaments. Circ Res. 2013;112(12):1542–56.
39. Karsanov N V., Pirtskhalaishvili MP, Semerikova VJ, Losaberidze NS. Thin myofilament proteins in norm and heart failure I. Polymerizability of myocardial Straub actin in acute and chronic heart failure. Basic Res Cardiol. 1986;81(2):199–212.
40. Childers RC, Sunyecz I, West TA, Cismowski MJ, Lucchesi PA, Gooch KJ. Role of the Cytoskeleton in the Development of a Hypofibrotic Cardiac Fibroblast Phenotype in Volume Overload Heart Failure. Am J Physiol Heart Circ Physiol. 2018;316(3):H596–608.
41. Liu YH, Yang XP, Sharov VG, Nass O, Sabbah HN, Peterson E, et al. Effects of angiotensin-converting enzyme inhibitors and angiotensin II type 1 receptor antagonists in rats with heart failure: Role of kinins and angiotensin II type 2 receptors. J Clin Invest. 1997;99(8):1926–35.
42. Schrier RW, Abdallah JG, Weinberger HHD, Abraham WT. Therapy of heart failure. Kidney Int. 2000;57(4):1418–25.
43. Aoyagi T, Matsui T. Phosphoinositide-3 kinase signaling in cardiac hypertrophy and heart failure. Curr Pharm Des. 2011;17(18):1818–24.
44. Ennis I, Aiello E, Cingolani H, Perez N. The Autocrine/Paracrine Loop After Myocardial Stretch: Mineralocorticoid Receptor Activation. Curr Cardiol Rev. 2013;9(3):230–40.

45. Sullivan RKP, WoldeMussie E, Pow D V. Dendritic and synaptic plasticity of neurons in the human age-related macular degeneration retina. Invest Ophthalmol Vis Sci. 2007;48(6):2782–91.
46. Sohn YI, Lee NJ, Chung A, Saavedra JM, Scott Turner R, Pak DTS, et al. Antihypertensive drug Valsartan promotes dendritic spine density by altering AMPA receptor trafficking. Biochem Biophys Res Commun. 2013;439(4):464–70.
47. Frank RN. Growth factors in age-related macular degeneration: Pathogenic and therapeutic implications. Ophthalmic Res. 1997;29(5):341–53.
48. Glenn J V., Stitt AW. The role of advanced glycation end products in retinal ageing and disease. Biochim Biophys Acta - Gen Subj. 2009;1790(10):1109–16.
49. Grossniklaus HE, Green WR. Choroidal neovascularization. Am J Ophthalmol. 2004;137(3):496–503.
50. Berriz GF, Beaver JE, Cenik C, Tasan M, Roth FP. Next generation software for functional trend analysis. Bioinformatics. 2009;25(22):3043–4.
51. Nowak JZ. AMD-the retinal disease with an unprecised etiopathogenesis: In search of effective therapeutics. Acta Pol Pharm - Drug Res. 2014;71(6):900–16.
52. Yuan X, Gu X, Crabb JS, Yue X, Shadrach K, Hollyfield JG, et al. Quantitative Proteomics: Comparison of the Macular Bruch Membrane/Choroid Complex from Age-related Macular Degeneration and Normal Eyes. Mol Cell Proteomics. 2010;9(6):1031–46.
53. Lee AY, Kulkarni M, Fang AM, Edelstein S, Osborn MP, Brantley MA. The effect of genetic variants in SERPING1 on the risk of neovascular age-related macular degeneration. Br J Ophthalmol. 2010;94(7):915–7.
54. Higgins P. Balancing AhR-Dependent Pro-Oxidant and Nrf2-Responsive Anti-Oxidant Pathways in Age-Related Retinopathy: Is SERPINE1 Expression a Therapeutic Target in Disease Onset and Progression? J Mol Genet Med. 2015;8(2):101.

55. Miyata M, Ikeda Y, Nakamura S, Sasaki T, Abe S, Minagoe S, et al. Effects of Valsartan on Fibrinolysis in Hypertensive Patients With Metabolic Syndrome. Circ J. 2012;76(4):843–51.

56. Oubiña MP, De las Heras N, Vázquez-Pérez S, Cediel E, Sanz-Rosa D, Ruilope LM, et al. Valsartan improves fibrinolytic balance in atherosclerotic rabbits. J Hypertens. 2002;20(2):303–10.

57. Albert-Fort M, Hombrebueno JR, Pons-Vazquez S, Sanz-Gonzalez S, Diaz-Llopis M, Pinazo-Durán MD. Retinal neurodegenerative changes in the adult insulin receptor substrate-2 deficient mouse. Exp Eye Res. 2014;124:1–10.

58. Zhang R, Liu Z, Zhang H, Zhang Y, Lin D. The COX-2-selective antagonist (NS-398) inhibits choroidal neovascularization and subretinal fibrosis. PLoS ONE. 2016;11(1):e0146808.

59. Wang X, Ma W, Han S, Meng Z, Zhao L, Yin Y, et al. TGF-β participates choroid neovascularization through Smad2/3-VEGF/TNF-α signaling in mice with Laser-induced wet age-related macular degeneration. Sci Rep. 2017;7(1):9672.

60. Skeie JM, Zeng S, Faidley EA, Mullins RF. Angiogenin in age-related macular degeneration. Mol Vis. 2011;17:576–82.

61. Hegab Z, Gibbons S, Neyses L, Mamas M. Role of advanced glycation end products in cardiovascular disease. World J Cardiol. 2012;4(4):90–102.

62. Banevicius M, Vilkeviciute A, Kriauciuniene L, Liutkeviciene R, Deltuva VP. The Association Between Variants of Receptor for Advanced Glycation End Products (RAGE) Gene Polymorphisms and Age-Related Macular Degeneration. Med Sci Monit. 2018;24:190–9.

63. Pickering RJ, Tikellis C, Rosado CJ, Tsorotes D, Dimitropoulos A, Smith M, et al. Transactivation of RAGE mediates angiotensin-induced inflammation and atherogenesis. J Clin Invest. 2019;129(1):406–21.

64. Garbayo E, Gavira JJ, De Yebenes MG, Pelacho B, Abizanda G, Lana H, et al. Catheter-based intramyocardial injection of FGF1 or

NRG1-loaded MPs improves cardiac function in a preclinical model of ischemia-reperfusion. Sci Rep. 2016;6:25932.

65. Lakó-Futó Z, Szokodi I, Sármán B, Földes G, Tokola H, Ilves M, et al. Evidence for a Functional Role of Angiotensin II Type 2 Receptor in the Cardiac Hypertrophic Process in Vivo in the Rat Heart. Circulation. 2003;108(19):2414–22.
66. Galindo CL, Ryzhov S, Sawyer DB. Neuregulin as a heart failure therapy and mediator of reverse remodeling. Curr Heart Fail Rep. 2014;11(1):40–9.
67. Xu J, De Winter F, Farrokhi C, Rockenstein E, Mante M, Adame A, et al. Neuregulin 1 improves cognitive deficits and neuropathology in an Alzheimer’s disease model. Sci Rep. 2016;6:31692.
68. Kaarniranta K, Salminen A, Haapasalo A, Soininen H, Hiltunen M. Age-related macular degeneration (AMD): Alzheimer’s disease in the eye? J Alzheimers Dis. 2011;24(4):615–31.
69. Verweij N, Eppinga RN, Hagemeijer Y, Van Der Harst P. Identification of 15 novel risk loci for coronary artery disease and genetic risk of recurrent events, atrial fibrillation and heart failure. Sci Rep. 2017;7(1):2761.
70. Kaarniranta K, Sinha D, Blasiak J, Kauppinen A, Veréb Z, Salminen A, et al. Autophagy and heterophagy dysregulation leads to retinal pigment epithelium dysfunction and development of age-related macular degeneration. Autophagy. 2013;9(7):973–84.
71. Kawano H, Cody RJ, Graf K, Goetze S, Kawano Y, Schnee J, et al. Angiotensin II Enhances Integrin and α-Actinin Expression in Adult Rat Cardiac Fibroblasts. Hypertension. 2012;35(1 Pt 2):273–9.
72. Bujak M, Frangogiannis NG. The role of IL-1 in the pathogenesis of heart disease. Arch Immunol Ther Exp (Warsz). 2009;57(3):165–76.
73. Turner NA. Effects of interleukin-1 on cardiac fibroblast function: Relevance to post-myocardial infarction remodelling. Vascul Pharmacol. 2014;60(1):1–7.

74. Nassar K, Grisanti S, Elfar E, Lüke J, Lüke M, Grisanti S. Serum cytokines as biomarkers for age-related macular degeneration. Graefes Arch Clin Exp Ophthalmol. 2015;253(5):699–704.
75. Zhang YN, Vernooij F, Ibrahim I, Ooi S, Gijsberts CM, Schoneveld AH, et al. Extracellular vesicle proteins associated with systemic vascular events correlate with heart failure: An observational study in a dyspnoea cohort. PLoS ONE. 2016;11(1):e0148073.
76. Zaman AKMT, French CJ, Schneider DJ, Sobel BE. A Profibrotic Effect of Plasminogen Activator Inhibitor Type-1 (PAI-1) in the Heart. Exp Biol Med. 2009;234(3):246–54.
77. Messaoudi S, Azibani F, Delcayre C, Jaisser F. Aldosterone, mineralocorticoid receptor, and heart failure. Mol Cell Endocrinol. 2012;350(2):266–72.
78. Chakravarthy U, Wong TY, Fletcher A, Piault E, Evans C, Zlateva G, et al. Clinical risk factors for age-related macular degeneration: A systematic review and meta-analysis. BMC Ophthalmol. 2010;10:31.

Supplementary material

Extended version of materials and methods

1. Biological Effectors Database (BED) to molecularly describe specific clinical conditions

Patient-like characteristics are modelled using clinical data and/or experimental molecular data. There are many databases providing clinical data of patients, adverse drug reactions, diseases or indications (e.g. ClinicalTrials.gov, SIDER, ChEMBL, PubChem, DrugBank…). Many other databases provide molecular data defining the existing human genes and/or proteins and describing the relationships between them (IntAct, BioGRID, REACTOME…). Combining both, clinical and molecular information available, the BED describes more than 300 clinical phenotypes as sets of genes and proteins (effectors) that can be “active”, “inactive” or “neutral” (1,2). For example, in a metabolic protein-like network, an enzyme will become “active” in the presence of a catalyst, or become inactivated when interacting with an inhibitor. Alternatively, in a genetic network, genes are active when they are expressed (experimentally detected as over-expression) and inactive when they are repressed (experimentally detected as under-expression). Additionally, in protein-protein interaction (PPI) networks, some proteins carry out their interactions only when they are phosphorylated, thus becoming active, and vice versa by dephosphorylation. By default, neutral proteins remain unaffected, neither active nor inactive, for a particular phenotype.

The methodology used for assigning the protein effectors to each pathology starts by defining the pathophysiological processes (functions) according to the general definitions used by the scientists studying the disease. Then, a review of the most recent, relevant and accepted information in the field is performed through PubMed queries, starting from general pathophysiology reviews. An expansion of the effector candidate’s identification is done through reading the relevant original papers from the references or adding searches of important concepts that are not covered enough (molecularly wise) within the reviews read. The final goal of the characterization is to select proteins with an accepted functional role within the disease, and specifically within the functions that define the disease to center the analysis.

1.1. HF effectors

Regarding the molecular basis of HF BED proteins, they were characterized as described above and in *Iborra-Egea et al. (2017)* (3). The definition used of heart failure in the current study has been performed according to the indication of Entresto and to the EMA Assessment report (4). Thus, it is centered in processes associated to long term changes related to cardiac remodeling (as discussed in the paper were the models were initially presented (3)), that can be cause and consequence of heart failure, not necessarily caused by ischemic causes. The identified functions are detailed in Supplementary Table 12.

1.2. MD effectors

MD pathophysiology is tightly related to protein accumulation (5–7). However, the characterization used for the current study not only included this function, but also other processes associated to MD pathophysiology, including neovascularization, characteristics of wet Age-Related MD and changes associated to geographic atrophy (late stage dry Age-Related MD) (8). The functions are detailed in Supplementary Table 13.

2. TPMS modelling

The Therapeutic Performance Mapping System (TPMS) is a tool that creates mathematical models of a drug/pathology protein pathways to explain a clinical outcome or phenotype (2,3,9–13). These models find MoAs that explain how a *Stimulus* (i.e. proteins activated or inhibited by a drug) produces a *Response* (i.e. proteins active or inhibited in a phenotype). As an example of usage, here we applied TPMS to the drug-indication pair sacubitril/valsartan and HF. Regarding the drug, we retrieved the sacubitril/valsartan targets from DrugBank (14), PubChem (15), STITCH (16), SuperTarget (17) and hand curated literature revision. As for the indication, we retrieved the proteins whose modulations had been associated with HF from the BED (1,2). Finally, after applying the TPMS methodology, we obtained a set of connected proteins (subnetworks) with associated activities, each subnetwork with a potential explanation of the molecular mechanism of the drug in agreement with what had been previously described (i.e. a potential MoA).

2.1. Building the Human protein network (HPN)

To apply the TPMS approach and create the mathematical models of MoAs, an HPN is needed beforehand. In this study, we used a PPI network created from the integration of public and private databases: KEGG (18), BioGRID (19), IntAct (20), REACTOME (21), TRRUST (22), and HPRD (23). In addition, information extracted from scientific literature, which was manually curated, was also included and used for trimming the network. The resulting HPN considers interactions corresponding to different tissues to take into account the effect of the *Stimulus* in the whole body.

2.2. Defining model restrictions

A collection of restrictions, defined as the true set of edges and nodes with the property of being active or inactive, are used for validating the models obtained with TPMS. We define two types of restrictions depending on its specificity. The general or global restrictions are those used in all approaches and describe a wide expanse of knowledge about protein interactions and relations. This information is obtained from HPRD (23), DIP (24), TRRUST (22), INTACT (20), REACTOME (21), BIOGRID (19), SIDER (25) and DrugBank (14). These set of restrictions help indicate what proteins are active or inactive, and their interactions, in a general human being. Additionally, specific restrictions regarding the phenotype under study can also be used, usually derived from high throughput data or additional protein knowledge.

For this study, we added specific information to our models concerning the changes of gene expression induced by sacubitril/valsartan on HF patients. Specifically, we used the GSE57345 gene expression dataset (26), extracted from GEO database, as in *Iborra-Egea et al. (2017)* (3). We calculated the expression fold change of genes associated with the HPN and mapped them as activated or inhibited proteins (active if they corresponded to over-expressed genes and inactive -inhibited- for under-expressed).

2.3. Description of the mathematical models

The algorithm of TPMS for generating the models is similar to a Multilayer Perceptron of an Artificial Neural Network over the HPN (where neurons are the proteins and the edges of the network are used to transfer the information). It takes as input signals the activation (+1) and inactivation (-1) of the drug target proteins and as output the BED protein states of the pathology phenotype. The network is limited to only interactions that connect the drug targets with any other protein in the HPN in a maximum of three steps to avoid signal noisiness. Once set, the algorithm optimizes the paths between both input and output protein sets and computes the activation and inactivation values of the all proteins in the HPN. The parameters to solve are the weights associated to the links between every node pair (ω_l). Each node of the protein network receives as input the output of the incoming connected nodes, which are weighted by each link weight. The sum of inputs is transformed by a hyperbolic tangent function to generate the score of the node (neuron), which become the "output signal" of the current node towards outgoing nodes. Details of the approach are shown in **Fig 1a**, where n_5 is linked to n_1 and n_2. The output signal of n_5 is $n_5 = \tanh(n_1 \cdot \omega_{1-5} + n_2 \cdot \omega_{2-5})$. The ω_l parameters are obtained by optimization, using a Stochastic Optimization Method based on Simulated Annealing (27), such that the values of the effector nodes are the closest to their expected values, and always adjusting to the maximum of the restrictions mentioned above. The iterative process of optimization usually requires between 10^6 and 10^9 iterations, until satisfying at least the 80% of the restrictions and the values of the effectors. However, the number of ω_l parameters can be very high (between 100,000 and 400,000 depending on the size of the subnetwork) and the size of the collection of restrictions (approximately 10^7) is usually not enough to find a unique solution. For that, many final models can be obtained and manual curation can be applied to select and modify the network and reduce the space of exploration.

3. Measures to compare sets of MoAs

TPMS returns a set of MoAs describing potential relationships between the targets of a drug and the biological protein effectors of a disease. We hypothesize that TPMS solutions represent MoAs in different prototype-patients. Therefore, we needed to define some comparison measures in order to understand the relationships between all potential mechanisms and compare sets of MoAs from different views.

3.1. Intensity of the response

We defined the “intensity” of the response as a pair: 1) the number of protein effectors (#) achieving an expected signal sign; and 2) a measure of the strength of the output signal of the effectors (i.e. a global measure of the output signal, named TSignal). For the present study, however, only the TSignal was used.

Assuming y_i as the value achieved by a protein effector “i”, while v_i is the effector sign according to the BED (active or inactive) and n is the total number of effectors described for a phenotype, we define:

- **Number of effectors achieving the expected sign**: We expect that a drug will revert the conditions of a disease phenotype, while it may reach the effectors of an adverse event. Consequently, a drug should inactivate the active protein effectors of a pathology-phenotype and activate the inactive ones, but it could activate/inhibit other adverse event effectors with the same sign as described in the BED. Using Dirac’s δ (i.e. $\delta(0)=1$, and zero otherwise), for drug indications the formula is defined as following:

$$\#_{indication} = \sum_{i=1}^{n} \delta\left(v_i + \frac{y_i}{|y_i|}\right) \qquad \textbf{[Equation 1a]}$$

Therefore, in the case of the disease effectors we only count the effectors with a BED value of opposite sign to the signal arriving from the drug.

However, for adverse events, the formula changes because we count the effectors that are affected by the drug, such that the signal arriving from the drug has the same sign as in the BED:

$$\#_{adverse\ event} = \sum_{i=1}^{n} \delta\left(v_i - \frac{y_i}{|y_i|}\right) \qquad \textbf{[Equation 1b]}$$

- **TSignal**: The average of the output values of the protein effectors such that the proteins with correct sign are considered as positive signal, and the ones with the incorrect sign considered as negative signal. For a drug affecting the phenotype of a disease, this implies that v_i and y_i have opposite sign and we need to change the sign:

$$TSignal_{indication} = -\frac{1}{n}\sum_{i=1}^{n} v_i y_i \qquad \textbf{[Equation 2a]}$$

On the contrary, to test if a drug induces an adverse event, we check if the output signal has the same sign as the effectors of the desired phenotype, and therefore TSignal is defined as:

$$TSignal_{adverse\ event} = \frac{1}{n}\sum_{i=1}^{n} v_i y_i \qquad \textbf{[Equation 2b]}$$

3.2. Distance between two sets of MoAs

We used the modified Hausdorff distance (MHD) introduced by Dubuisson and Jain (28) as the *distance* between two or more sets of MoAs in order to determine their similarity. We used the distance measures between two (finite) point sets A and B as following:

$$\text{For } a \in A, \qquad d(a,B) \coloneqq \min_{b \in B} d(a,b),$$

$$\text{and} \qquad d_A(B) \coloneqq \frac{1}{|A|} \sum_{a \in A} d(a,B),$$

Where |A| is the number of elements in A, d(·,·) is the Euclidean distance and “a” and “b” are n-tuples of the activities (output signals) of the nodes of two MoAs (a in A and b in B). Then, we defined the MHD as:

$$d_{\text{MHD}}(A,B) \coloneqq \max\left(d_A(B), d_B(A) \right) \qquad \textbf{[Equation 3]}$$

Note that the MHD is a semimetric and not a metric, since the triangular inequality does not hold.

3.3. Potential biomarkers extracted from MoAs

3.3.1. Identification of Best-Classifier Proteins

In order to extract potential biomarkers from comparing sets of MoAs, we first defined the *best-classifier proteins*, specific proteins helping us to infer biological associations and distinguish the responses of drugs on a population (i.e. potential biomarkers). Best-classifier proteins (single or pairs) are the proteins inside the HPN that allow to better classify samples between groups of MoAs. These classifiers are determined by a Data-Science strategy, which is based on a set of Feature Selection algorithms combined with several Base Classifiers. The feature selection used for single proteins was brute force (29), so analyzing one feature or protein at a time, while for protein pairs the following selection methods were used: elastic net (30); entropy and correlation (31); LASSO (32); random forest (33); GLM random sets (34); ReliefF (35); Ridge regression (36); simple regression (37); Wilcoxon test (38); and Wilcoxon test with correlation (38). Several base classifiers were applied to distinguish the two groups using the selected features: optimal threshold; linear regression (37); Multilayer Perceptron Network (39); Generalized Linear Model (34); elastic net (40); and optimal quadratic threshold (41). Finally, after a k-fold cross-validation (k=10) (42) was applied, the proteins were sorted by the balanced accuracy (43) of the classification. For this study, only the 200 proteins (or pair of proteins) with highest balanced accuracy were selected as best-classifier proteins. Assuming the hypothesis that the selected MoAs are representative of individual prototype-patients, these proteins could then be used as biomarkers to classify a cohort of patients by the activity or absence of activity of the proteins.

3.3.2. Identification of differential Best-Classifier Proteins

Each best-classifier protein has a specific distribution of signal values corresponding to each group of MoAs. We applied the Mann-Whitney *U* test to compare the two distributions and selected those proteins having a significantly different distribution (p-value< 0.01).

We also restricted the list to proteins having an average value with opposite sign among groups (i.e. positive vs. negative or vice versa), and named them as *differential best-classifier proteins*. By following this strategy, we can identify two groups of differential best-classifier proteins: those active in the first group (positive output signal in average) and inactive in the other (negative output signal in average), and the opposite.

3.3.3. Types of proteins not considered

- **Non-differential Best-Classifier Proteins**: Those are proteins in which, even if the mean signal in both groups is very similar, the machine learning algorithms are still able to differentiate High- and Low- MoAs based on their distribution values. For example, in the upper right corner of Figure 2a we find the protein P29353, the 181st best protein to classify High- and Low- HF models (cross-validation AUC = 0.67, P-value = 1.12·10-4). P29353 has a Low-HF mean signal of 0.99999999948 and a High-HF mean signal of 0.9999999985. As showed in **Supplementary Fig 4a**, the High- and Low- HF signals values are both very close to each other. However, if we explore the distribution of signals considering all the decimals given by TPMS (**Supplementary Fig 4b**), we can observe a slight difference between the two distributions. This fact allowed the machine learning algorithms to include the protein as a best-classifier protein, but was then rejected as a differential best-classifier protein after applying the Mann-Whitney *U* test.

- **Differential non-Best-Classifier Proteins**: Those are proteins that, when comparing the signals between groups, they have significantly opposite sign. However, they are not considered Best-

Classifier Proteins because they are not among the top 200 proteins selected by the machine learning algorithms. For example, the protein P40763 is the 241st best feature on distinguishing High- and Low- Heart Failure Mechanisms of Action (cross-validation AUC = 0.66, P-value = $1.22 \cdot 10^{-3}$). The distribution of High- and Low signals are represented in **Supplementary Fig 5**. In the figure we can appreciate how the distributions of High- and Low- signals are overlapped, complicating their differentiation. Still, the p-value of the cross-validation is below 0.05, reflecting the potential of this feature to differentiate the distinct types of Mechanisms of Action.

Extended version of results and discussion

We applied TPMS to the HPN using as input signals the drug targets of sacubitril/valsartan (NEP / AT1R) and as output signals the proteins associated with HF extracted from the BED. Out of all MoAs found by TPMS, we selected the 200 satisfying the largest number of restrictions (and at least 80% of them) to perform further analysis.

Note that TPMS was only executed once, optimizing the results to satisfy the restrictions on HF data. The values of MD are obtained by measuring the signal arriving at the MD effectors, which are part of the HPN and also receive signal. This procedure was chosen because we defined HF as the indication of the drug (sacubitril/valsartan), while MD is a potential adverse effect.

1. Stratification of MoAs

In order to compare models related to a good or bad response to the treatment, or those more prone to lead towards potential MD adverse effect, we stratified the MoAs. For HF, or treatment response, MoAs were ranked by their TSignal and then split in four quartiles. The first quartile (top 25%) contains MoAs with higher intensity of the response, which in turn corresponds to lower values of the effectors associated with HF phenotype (we named them as “Low”-disease MoAs). On the contrary, the fourth quartile (bottom 25%) collects MoAs with lower intensity of response (thus, we named as “High”-disease MoAs) (**Supplementary Fig 1a**). On the other hand, for MD, the first quartile (top 25%) contains MoAs with higher intensity, which as an adverse event, correspond to models with high values of the effectors associated to MD (we named them as High-adverseEvent MoAs). The fourth quartile (bottom 25%) collects MoAs with lower intensity of response (thus, we named as Low- adverseEvent MoAs) (**Supplementary Fig 1b**). Note that, in the following steps and because HF and MD groups were extracted from the same 200 set of models, common MoAs between different HF and MD-defined sets could be expected.

2. Comparison of MoAs with high/low TSignal associated to HF or MD

We calculated the modified Hausdorff distance between the groups of MoAs (High-MD, Low-MD, High-HF and Low-HF) to elucidate their similarity values (**Supplementary Table 5**). In this sense, the higher distance between the groups, the more different they are. We used these distances to calculate a dendrogram tree (see

Supplementary Fig 2) showing that MoAs associated with a bad response to sacubitril/valsartan for HF (high-HF) are more similar (i.e. closer) to MoAs linked to a stronger MD adverse effect (high-MD). It is remarkable that the distances between Low-HF and High-HF and between Low-MD and High-MD are larger than the cross distances between HF and MD. However, by the definition of distance (equation 3 in supplementary material), we cannot account for the dispersion among the MoAs within and between each group. Therefore, for each set we calculated the mean Euclidean distance between all the points and its center, defined by the average of all points (see **Supplementary Table 6**). As a result, all groups showed very similar dispersion values.

In order to have a global and graphical view of the distance between the individual MoAs, we generated a multidimensional scaling (MDS) plot calculated using MATLAB (see **Fig 2**). MDS plots display the pairwise distances in two dimensions while preserving the clustering characteristics (i.e. close MoAs are also close in the 2D-plot and far MoAs are also far in 2D). Focusing on the Low-HF group depicted in blue circles, we observe that there is no clear tendency to cluster with any of the MD groups. There are few cases of Low-HF MoAs coinciding in the space with Low- or High-MD MoAs. This implies that a good response to sacubitril/valsartan of HF patients would not be usually linked to the development of MD. Moreover, no clear distinction is found when plotting only the MD MoAs within the Low-HF group (see **Supplementary Fig 3a**). However, regarding the set of High-HF MoAs, we can differentiate two clusters of MoAs: one related to the High-MD group (green crosses); and the other close to MoAs of the Low-MD group (black crosses) (see **Supplementary Fig 3b**).

Assuming the hypothesis that different MoAs correspond to distinct prototype-patients, we conclude that for the specific set of patients for which sacubitril/valsartan works best reducing HF, it would be more difficult to differentiate between those presenting MD and those who do not. Instead, for the High-HF group, patients having MD could indeed be easily distinguished from those not presenting MD as side effect. However, because Low-HF group has more relevance to the clinics, specific functional analyses were performed in this specific group, as seen in following sections.

3. Identification and functional analysis of potential biomarkers

For this section, we identified the nodes (i.e. proteins) significantly differentiating two groups of models (using a Mann-Whitney *U* test) for which the average of output signals have opposite signs (see methods in 3.3). After that, the function of the identified proteins was extracted from Gene Ontology (GO).

3.1. Identification of best-classifier proteins differentiating HF responses

After the model stratification regarding the HF groups, we selected the 200 best-classifier proteins to differentiate the two groups of MoAs. Among these proteins, we identified the differential best-classifier proteins as explained in the methodology, and ended up with two groups: those active in Low-HF (the average of output signals in Low-HF MoAs is positive) and inactive in High-HF (the average of output signals in High-HF MoAs is negative); and those active in High-HF but inactive in Low-HF. Out of the starting 200 best-classifier proteins, we found a total of 45 differential best-classifier proteins associated with the treatment response (6 in the first group and 39 in the second) (see **Supplementary Table 1**). **Fig 3a** displays all the proteins average signal values for the MoAs of Low-HF vs High-HF. Most of the proteins with opposite signs between the two cohorts were also selected as differential best-classifier proteins.

To pinpoint the biological role of these proteins, we first identified the GO enriched functions (see **Supplementary Table 2)** and then searched in the literature for evidences linking them with HF. The enrichment used for this proceeding was calculated using the software FuncAssociate (44). Among the enriched functions, we found processes associated with the SCAR complex, the positive regulation of actin nucleation, the regulation of neurotrophin TRK receptor and dendrite extension. We used the same procedure to extract the GO functions associated to the differential best-classifier proteins that are inactive in Low-HF but active in High-HF. We detected functions such as phosphatidylinositol kinase activity, MAP kinase activity, DNA damage induced protein phosphorylation and superoxide anion generation. Although some enriched functions are shared by both sets, such as Fc gamma receptor signaling, the majority of functions identified are different (see **Supplementary Table 2)**.

Some of the proteins and functions highlighted in the current analysis have been related to myocardial function. On the one hand, our findings show that differential best-classifier proteins Low-HF-active/High-HF-inactive point towards an important role for actin nucleation and polymerization mechanisms in drug response (reflected by the functions *regulation of actin nucleation*, *regulation* *of Arp2/3 complex-mediated actin nucleation*, *SCAR complex*, *filopodium tip*, or *dendrite extension*). In fact, the alteration of actin nucleation and polymerization mechanisms has been reported in heart failure (45–47). Interestingly, a role for the activation of another differential best-classifier candidate, ATGR2, has been proposed to mediate some of the beneficial effects of angiotensin II receptor type 1 antagonists, such as valsartan (48,49).

On the other hand, the results of the differential best-classifier proteins Low-HF-inactive/High-HF-active are linked to phosphatidylinositol kinase mediated pathways (*phosphatidylinositol-3,4-bisphosphate 5-kinase activity*) and MAP kinase mediated pathways (*MAP kinase kinase activity*, best classifier proteins MAPK1, MAPK3, MAPK11, MAPK12 or MAPK13). In this case, both signaling pathways have been associated to cardiac hypertrophy and subsequent heart failure (50,51). These outcomes clearly leads towards the idea that High-HF models are a representation of prototype-patients with a worst response to the treatment, while Low-HF models are related to more beneficial response to the medication.

3.2. Identification of best-classifier proteins differentiating MD responses

We similarly classified MoAs in High-MD and Low-MD identified the differential best-classifier proteins active in Low-MD but inactive in High-MD, and vice versa. As before, we compared the distributions of Low-MD and High-MD output signals of the best-classifier proteins and calculate the average of the signal in all MoAs in Low- and High-MD. Out of 200 best-classifier proteins, we identified 28 Low-MD-active/High-MD-inactive and 29 Low-MD-inactive/High-MD-active (see **Supplementary Table 3**). **Fig 3b** shows the plot for all proteins classified by their average output signal in Low-MD and High-MD models.

Again, we calculated the GO enriched functions for these groups of proteins (see **Supplementary Table 4**). For the first group (Low-MD-active/High-MD-inactive) we obtained unique functions such as dendritic spine development, positive regulation of vascular endothelial growth factor production and phosphotyrosine binding. For the second group (Low-MD-inactive/High-MD-active), we found functions such as dorsal/ventral axon guidance, fibroblast growth factor receptor binding and response to toxic substance. However, phosphatidylinositol bisphosphate kinase activity showed up as enriched function in both groups.

Some of the proteins and functions underlined in the current analysis had previously been related to MD. The presence of dendritic spine development and dorsal/ventral axon guidance related proteins among the differential best-classifiers points towards a role for sacubitril/valsartan-associated MD in dendritic and synaptic plasticity mechanisms, which had been previously linked to the condition (52). Furthermore, valsartan treatment has been reported to promote dendritic spine development in other related neurodegenerative diseases, such as Alzheimer’s disease (53). Other functions enriched within the differential best-classifier proteins (Low-MD-inactive/High-MD-active) are implicated in growth factor related pathways, which are known to be involved in wet MD pathogenesis (54). Moreover, neovascularization in the wet variant of MD has been linked to the signaling of some of the growth factors detected as sacubitril/valsartan-associated MD classifiers in this study, including FGF1 (54) and PDGF (55,56).

3.3. Identification of potential biomarkers differentiating MD responses in Low-HF

We previously mentioned that some MoAs could be shared between the different groups of HF and MD (**Supplementary Table 7**). Knowing that, we focused on the shared MoAs between Low-HF and High-MD to analyze the special case comprising prototype-patients in which the treatment best reduces HF disease but increases MD adverse effect. In order to identify these patients, we compared the Low-HF ∩ Low-MD with Low-HF ∩ High-MD MoAs; **Table 1** shows the 30 biomarkers identified. On the one hand, we found 16 proteins active in Low-HF ∩ Low-MD MoAs but inactive in Low-HF ∩ High-MD (15 of them shared with MD best-classifier proteins). On the other hand, 14 proteins were identified as inactive in Low-HF ∩ Low-MD and active in Low-HF ∩ High-MD MoAs (12 of them were MD best-classifier proteins). We calculated the GO enriched functions of these two groups and observed that "phosphatidylinositol bisphosphate kinase activity" is enriched among proteins that are active in Low-HF ∩ Low-MD MoAs. Instead, "fibrinolysis" was found to be enriched among proteins active in Low-HF ∩ High-MD MoAs (**Table 2**). With this, we conclude that among the group of prototype-patients for which sacubitril/valsartan improves HF treatment response, the modulation of fibrinolysis could play a role at inducing the MD adverse effect. Moreover, we propose 12 best-classifier proteins that may be considered as biomarkers for good prognosis of the side effect.

In fact, since neovascular MD development is characterized by subretinal extravasations of novel vessels derived from the choroid (CNV) and the subsequent hemorrhage into the photoreceptor cell layer in the macula region (8), it might be reasonable to think that the modulation of fibrinolysis and blood coagulation pathways could play a role. The reported implication of some fibrinolysis related classifiers, such as FGB, SERPINE1 (PAI-1), and SERPING1, in neovascular MD development seems to support this hypothesis (57–59). Besides, valsartan might be implicated in this mechanism, since it has been reported to modulate PAI-1 levels and promote fibrinolysis in different animal and human models (60,61).

In addition, the presence of several other MD related classifiers in this list, such as IRS2 (62), PTGS2 (63), DCN (64) and FGF1 (65), further supports the interest of the classifiers as biomarkers of MD development in sacubitril/valsartan good responders.

4. Analysis of proposed biomarkers with GUILDify

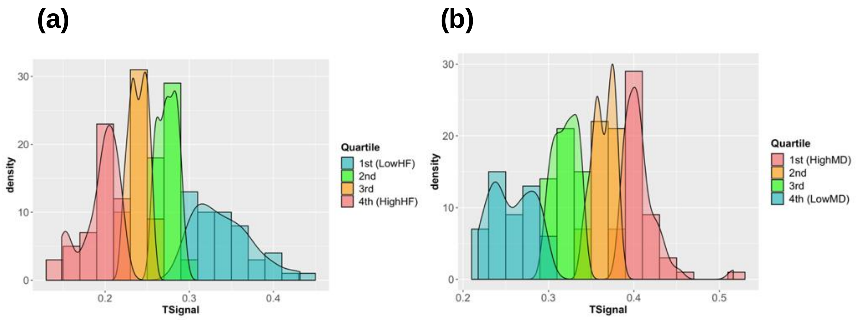
In the previous section, we proposed 30 proteins that could potentially help to identify HF patients at risk of developing MD. To corroborate these biomarkers, we tested how many of them are found using a different approach also based on the use of functional networks. For this purpose, we used GUILDify v2.0 (66), a web server that extends the information of disease-gene associations through the protein-protein interactions network. GUILDify scores proteins according to their proximity with the genes associated with a disease (seeds). Using this web server, we identify a list of top-scoring proteins that are critical on transmitting the perturbation of disease genes through the network. The network used by GUILDify is completely independent from the HPN used in the TPMS, becoming an ideal, independent context to test the potential biomarkers.

Thus, we used GUILDify to indicate which of the potential biomarkers identified by TPMS may have a relevant role in the molecular mechanism of the drug. We ran GUILDify using the two targets of sacubitril/valsartan (NEP, AT1R) as seeds, and selected the top 2% scored nodes (defined as the "top-drug" set). We did the same with the phenotypes of HF and MD, using as seeds the 124 effectors of HF and 163 effectors of MD from the BED database. We merged the top scored sets of HF, MD and top-drug ("top-drug ∪ top-HF ∪ top-MD") and studied the overlap with the set of differential best-classifier proteins associated with MD and HF. **Supplementary Table 8** shows the result of this analysis, with a significant representation of best-classifier proteins in most of the sets, especially on MD best-classifier proteins. **Supplementary Table 9** shows the list of 13 proteins involved in this overlap. We have also checked the overlap with the 30 biomarkers proposed in the previous section, of which 10 are found in the merged set "top-drug ∪ top-HF ∪ top-MD" and are consequently significant (see **Supplementary Tables 10 and 11**).

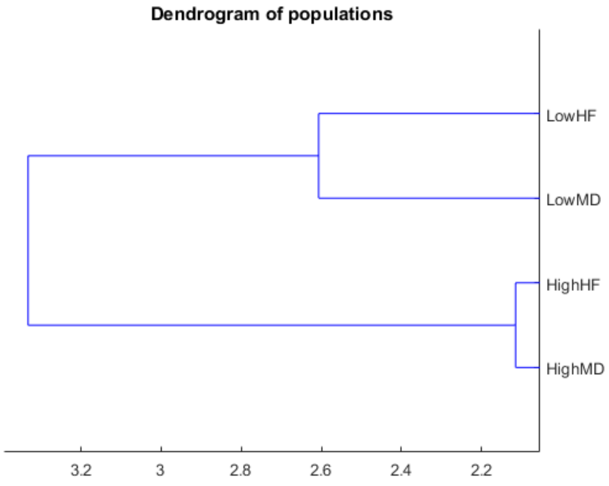
Some of these candidates can be functionally linked to both diseases and the drug under study. For example, among these 10 classifiers, AGER has been implicated in both HF (67), through extracellular matrix remodeling, and MD development (68), through inflammation, oxidative stress, and basal laminar deposit formation between retinal pigment epithelium cells and the basal membrane; furthermore, this receptor is known to be modulated by AT1R (69), valsartan target.

Similarly, FGF1 has been proposed to improve cardiac function after HF (70), as well as to promote choroid neovascularization leading to MD (54). Moreover, FGF1 is regulated by angiotensin II through ATGR2 (71), another protein suggested as classifier in the current analysis that is known to mediate some of the effects of AT1R antagonists, such as valsartan (48,49). Another candidate, NRG1, has been linked to myocardial regeneration after HF (72) and is known to lessen the development of neurodegenerative diseases such as Alzheimer's disease (73), which shares similar pathological features with MD (74). NRG1 is also linked to the expression of neprilysin (73), sacubitril target. ITGB5 has been identified as risk locus for HF (75) and its modulation has been linked to lipofucsin accumulation in MD (76). Interestingly, ATGR1 inhibitors have been reported to modulate ITGB5 expression in animal models (77). Finally, IL1A has been proposed as an essential mediator of HF pathogenesis (78,79) through inflammation modulations, and serum levels of this protein have been found increased in MD patients (80). In addition, as described in previous sections, classifiers FGB, SERPINE1, and SERPING1 have been linked to MD (57–59) and are also known to play a role in HF development (81–84). According to these findings, the 10 potential biomarkers proposed by TPMS and identified with GUILDify might be prioritized when studying good responder HF patients at risk of MD development.

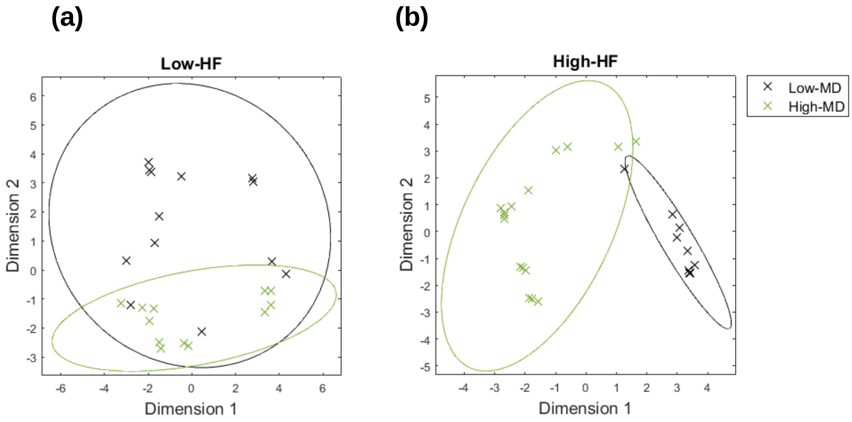
Supplementary Figures



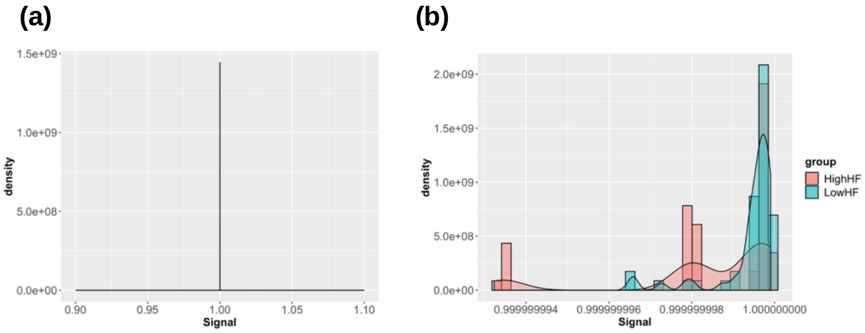
Supplementary Fig 1: Histogram of the number of models belonging to High- and Low- (HF in **(a)** and MD in **(b)**) in a range of TSignal values. The models are divided in four quartiles, the 1st and 4th corresponding to the Low- and High- groups for HF and vice versa for MD.



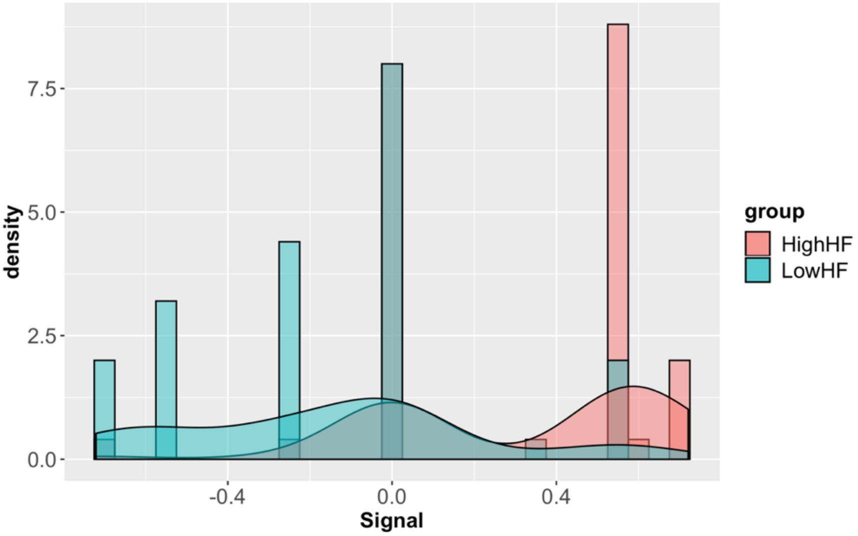
Supplementary Fig 2: Dendrogram plot of the pairwise modified Hausdorff distance (MHD) between the four groups of mechanisms of action (MoAs): LowHF, HighHF, LowMD, HighMD.



Supplementary Fig 3: Multidimensional scaling plot of the distances between the Mechanisms of Action of Low/High-Macular Degeneration (MD) and Heart Failure (HF). In **(a)** we find the models of both MD groups and Low-HF, whereas in **(b)** the models are for both MD groups and High-HF.



Supplementary Fig 4: Histogram of the signal of protein P29353 (non-differential Best-Classifier Protein) in the Mechanisms of Action belonging to High-HF and Low-HF. In **(a)** a general vision of the signal, in **(b)** a detailed vision of the signal.



Supplementary Fig 5: Histogram of the signal of protein P40763 (differential non-Best-Classifier Protein) in the Mechanisms of Action belonging to High-HF and Low-HF.

Supplementary Tables

Supplementary Table 1: Differential best-classifier proteins with opposite signal in Low-HF (LHF) and High-HF (HHF). "+" stands for active, while "-" stands for inactive. Highlighted cells correspond to proteins that are part of the Top-HF ∪ Top-MD ∪ Top-Drug set, the top-scoring proteins according to GUILDify.

	Uniprot ID	**Gene symbol**	**Gene name**	$\langle LHF \rangle$	$\langle HHF \rangle$	$\sqrt{\left\lvert \frac{LMDx}{HMD} \right\rvert}$	**Adjusted P-value**
LHF + HHF -	Q96F07	CYFIP2	Cytoplasmic FMR1-interacting protein 2	0.110	-0.278	0.175	6.388E-07
	P55160	NCKAP1L	Nck-associated protein 1-like {ECO:0000305}	0.110	-0.278	0.175	6.388E-07
	Q7L576	CYFIP1	Cytoplasmic FMR1-interacting protein 1	0.110	-0.278	0.175	6.388E-07

	Q9NYB9	ABI2	Abl interactor 2	0.110	-0.278	0.175	6.388E-07
	Q9Y2A7	NCKAP1	Nck-associated protein 1	0.110	-0.278	0.175	6.388E-07
	P50052	AGTR2	Type-2 angiotensin II receptor	0.205	-0.013	0.051	1.852E-05
LHF - HHF +	**P28482**	**MAPK1**	**Mitogen-activated protein kinase 1**	-0.710	0.479	0.584	1.854E-14
	P27361	**MAPK3**	**Mitogen-activated protein kinase 3**	-0.313	0.962	0.549	6.366E-15
	P47900	P2RY1	P2Y purinoceptor 1	-0.322	0.605	0.441	9.855E-10
	Q92558	WASF1	Wiskott-Aldrich syndrome protein family member 1	-0.580	0.309	0.424	8.494E-13
	Q9Y6W5	WASF2	Wiskott-Aldrich syndrome protein family member 2	-0.580	0.309	0.424	8.494E-13
	O00401	WASL	Neural Wiskott-Aldrich syndrome protein	-0.580	0.309	0.424	8.494E-13
	P02751	**FN1**	**Fibronectin**	-0.476	0.224	0.327	2.152E-09
	O60229	KALRN	Kalirin {ECO:0000250|UniProtKB:P97924}	-0.529	0.185	0.313	3.239E-05
	Q8TEW0	PARD3	Partitioning defective 3 homolog	-0.279	0.255	0.267	3.577E-09
	P41743	PRKCI	Protein kinase C iota type	-0.279	0.255	0.267	3.577E-09
	Q13576	IQGAP2	Ras GTPase-activating-like protein IQGAP2	-0.279	0.255	0.267	3.577E-09
	O75914	PAK3	Serine/threonine-protein kinase PAK 3	-0.279	0.255	0.267	3.577E-09
	Q9P286	PAK5	Serine/threonine-protein kinase PAK 5 {ECO:0000305}	-0.279	0.255	0.267	3.577E-09
	O96013	PAK4	Serine/threonine-protein kinase PAK 4	-0.279	0.255	0.267	3.577E-09
	Q86VI3	IQGAP3	Ras GTPase-activating-like protein IQGAP3	-0.279	0.255	0.267	3.577E-09
	Q9NPB6	PARD6A	Partitioning defective 6 homolog alpha	-0.279	0.255	0.267	3.577E-09
	Q16584	MAP3K11	Mitogen-activated protein kinase kinase kinase 11	-0.279	0.255	0.267	3.577E-09
	Q9NQU5	BUB1B-PAK6; PAK6	Serine/threonine-protein kinase PAK 6	-0.279	0.255	0.267	3.577E-09

Q9BYG5	PARD6B	Partitioning defective 6 homolog beta	-0.279	0.255	0.267	3.577E-09
Q9BYG4	PARD6G	Partitioning defective 6 homolog gamma	-0.279	0.255	0.267	3.577E-09
P84095	RHOG	Rho-related GTP-binding protein RhoG	-0.449	0.156	0.265	1.935E-09
Q96PN6	ADCY10	Adenylate cyclase type 10	-0.488	0.144	0.265	2.421E-09
Q96JJ3	ELMO2	Engulfment and cell motility protein 2	-0.396	0.137	0.233	1.935E-09
Q15759	MAPK11	Mitogen-activated protein kinase 11	-0.883	0.056	0.223	6.366E-15
O15264	MAPK13	Mitogen-activated protein kinase 13	-0.884	0.045	0.198	6.366E-15
P53778	MAPK12	Mitogen-activated protein kinase 12	-0.884	0.045	0.198	6.366E-15
P54764	EPHA4	Ephrin type-A receptor 4	-0.139	0.233	0.180	6.388E-07
Q99755	PIP5K1A	Phosphatidylinositol 4-phosphate 5-kinase type-1 alpha	-0.110	0.278	0.175	6.388E-07
Q9UQB8	BAIAP2	Brain-specific angiogenesis inhibitor 1-associated protein 2	-0.110	0.278	0.175	6.388E-07
O14986	PIP5K1B	Phosphatidylinositol 4-phosphate 5-kinase type-1 beta	-0.110	0.278	0.175	6.388E-07
Q9Y5S8	NOX1	NADPH oxidase 1	-0.110	0.278	0.175	6.388E-07
P46734	MAP2K3	Dual specificity mitogen-activated protein kinase kinase 3	-0.110	0.278	0.175	6.388E-07
Q15080	NCF4	Neutrophil cytosol factor 4	-0.110	0.278	0.175	6.388E-07
Q9HBY0	NOX3	NADPH oxidase 3	-0.110	0.278	0.175	6.388E-07
Q9UPY6	WASF3	Wiskott-Aldrich syndrome protein family member 3	-0.110	0.278	0.175	6.388E-07
P52564	MAP2K6	Dual specificity mitogen-activated protein kinase kinase 6	-0.110	0.278	0.175	6.388E-07
O14733	MAP2K7	Dual specificity mitogen-activated protein kinase kinase 7	-0.110	0.278	0.175	6.388E-07
Q9Y4K3	TRAF6	TNF receptor-associated factor 6	-0.097	0.285	0.166	1.917E-06
P19878	NCF2	Neutrophil cytosol factor 2	-0.042	0.278	0.108	3.050E-05

Supplementary Table 2: Top 10 gene Ontology functions enriched from best-classifier proteins with opposite signal in Heart Failure (HF) MoAs. Functional enrichment analysis from FuncAssociate.

	Low-HF active / High-HF inactive			**Low-HF inactive / High-HF active**			**Overlapped functions**		
	GO name	**LOD**	**P-val.**	**GO name**	**LOD**	**P-val.**	**GO name**	**LOD**	**P-val.**
1	SCAR complex	3.89	<0.00050	1-phosphatidylinositol-3-phosphate 5-kinase activity	3.41	0.01700	Rac protein signal transduction	2.54	<0.00050
2	positive regulation of Arp2/3 complex-mediated actin nucleation	3.64	<0.00050	1-phosphatidylinositol-5-kinase activity	3.41	0.01700	vascular endothelial growth factor receptor signaling pathway	2.31	<0.00050
3	positive regulation of neurotrophin TRK receptor signaling pathway	3.49	0.00150	phosphatidylinositol-3,4-bisphosphate 5-kinase activity	2.94	0.04000	immune response-regulating cell surface receptor signaling pathway involved in phagocytosis	1.95	<0.00050
4	regulation of neurotrophin TRK receptor signaling pathway	3.36	<0.00050	proteolysis in other organism	2.73	0.00250	Fc-gamma receptor signaling pathway involved in phagocytosis	1.95	<0.00050
5	positive regulation of actin nucleation	3.27	<0.00050	MAP kinase kinase activity	2.60	<0.00050	Fc receptor mediated stimulatory signaling pathway	1.95	<0.00050
6	regulation of Arp2/3 complex-mediated actin nucleation	3.23	<0.00050	NADPH oxidase complex	2.58	<0.00050	Fc-gamma receptor signaling pathway	1.94	<0.00050
7	dendrite extension	3.16	0.00350	DNA damage induced protein phosphorylation	2.53	0.00450	Fc receptor signaling pathway	1.74	<0.00050
8	regulation of actin nucleation	2.96	<0.00050	MAP kinase activity	2.52	<0.00050	Ras protein signal transduction	1.79	<0.00050
9	filopodium tip	2.84	0.01200	superoxide-generating NADPH oxidase activity	2.34	0.00600	regulation of actin cytoskeleton organization	1.62	0.00072
10	developmental cell growth	2.42	0.00150	superoxide anion generation	2.25	0.00850	lamellipodium	1.71	0.00086

Supplementary Table 3: Differential best-classifier proteins with opposite signal in Low-MD (LMD) and High-MD (HMD). “+” stands for active, while “-” stands for inactive. Highlighted cells correspond to proteins that are part of the Top-HF ∪ Top-MD ∪ Top-Drug set, the top-scoring proteins according to GUILDify.

	Uniprot ID	Gene symbol	Gene name	$\langle LMD \rangle$	$\langle HMD \rangle$	$\sqrt{\left\lvert \frac{LMDx}{HMD} \right\rvert}$	Adjusted P-value
LMD+ HMD-	Q9Y4H2	IRS2	Insulin receptor substrate 2	0.583	-0.414	0.491	1.297E-13
	O43639	NCK2	Cytoplasmic protein NCK2	0.623	-0.355	0.471	5.744E-11
	Q13153	PAK1	Serine/threonine-protein kinase PAK 1 {ECO:0000303|PubMed:8805275}	0.233	-0.817	0.437	2.266E-12
	P30530	AXL	Tyrosine-protein kinase receptor UFO	0.476	-0.362	0.415	5.509E-16
	P42081	CD86	T-lymphocyte activation antigen CD86	0.428	-0.356	0.391	2.073E-08
	P18825	ADRA2C	Alpha-2C adrenergic receptor	0.226	-0.568	0.358	2.079E-10
	Q13177	PAK2	Serine/threonine-protein kinase PAK 2	0.249	-0.439	0.330	3.753E-09
	P54762	EPHB1	Ephrin type-B receptor 1	0.144	-0.685	0.314	2.916E-14
	P15498	VAV1	Proto-oncogene vav	0.392	-0.192	0.274	8.020E-05
	P06241	FYN	Tyrosine-protein kinase Fyn	0.589	-0.127	0.274	7.322E-15
	O75787	**ATP6AP2**	**V-ATPase M8.9 subunit**	0.407	-0.160	0.255	2.741E-08
	P01583	**IL1A**	**Interleukin-1 alpha**	0.125	-0.396	0.222	2.087E-12
	P06748	**NPM1**	**Nucleophosmin**	0.374	-0.116	0.208	2.266E-12
	Q02297	**NRG1**	**Pro-neuregulin-1, membrane-bound isoform**	0.670	-0.064	0.207	5.208E-14
	P15941	MUC1	Mucin-1 subunit alpha	0.085	-0.479	0.202	1.676E-11
	P18084	**ITGB5**	**Integrin beta-5**	0.498	-0.079	0.199	1.214E-15
	P03372	ESR1	Estrogen receptor	0.096	-0.294	0.169	6.103E-08
	P01138	NGF	Beta-nerve growth factor	0.211	-0.124	0.162	6.954E-07
	P43405	SYK	Tyrosine-protein kinase SYK	0.075	-0.310	0.152	1.618E-07
	Q08722	CD47	Leukocyte surface antigen CD47	0.082	-0.277	0.151	8.239E-07
	P54764	EPHA4	Ephrin type-A receptor 4	0.336	-0.065	0.148	4.859E-08
	Q9BYF1	ACE2	Processed angiotensin-converting enzyme 2	0.565	-0.039	0.148	7.333E-15
	P10275	AR	Androgen receptor	0.438	-0.045	0.141	1.014E-11
	P38398	BRCA1	Breast cancer type 1 susceptibility protein	0.043	-0.365	0.125	9.363E-08
	P35354	PTGS2	Prostaglandin G/H synthase 2	0.034	-0.396	0.116	2.482E-12
	Q9Y478	PRKAB1	5'-AMP-activated protein kinase subunit beta-1	0.374	-0.034	0.113	5.744E-11
	P14770	GP9	Platelet glycoprotein IX	0.034	-0.306	0.102	1.190E-08
	P14138	EDN3	Endothelin-3	0.023	-0.239	0.074	3.509E-06

LMD-HMD+	**P02675**	**FGB**	**Fibrinogen beta chain**	-0.778	0.654	0.713	3.040E-14
	O60674	JAK2	Tyrosine-protein kinase JAK2	-0.811	0.279	0.476	2.749E-16
	P04085	**PDGFA**	**Platelet-derived growth factor subunit A**	-0.359	0.622	0.473	1.263E-07
	Q05586	GRIN1	Glutamate receptor ionotropic, NMDA 1	-0.381	0.565	0.464	1.049E-15
	P05230	**FGF1**	**Fibroblast growth factor 1**	-0.219	0.734	0.401	1.528E-14
	Q15768	EFNB3	Ephrin-B3	-0.149	0.835	0.353	8.307E-13
	Q14451	GRB7	Growth factor receptor-bound protein 7	-0.181	0.679	0.351	5.106E-13
	P08581	MET	Hepatocyte growth factor receptor	-0.124	0.828	0.321	2.615E-13
	Q08289	CACNB2	Voltage-dependent L-type calcium channel subunit beta-2	-0.351	0.238	0.289	5.549E-09
	P63244	RACK1	Receptor of activated protein C kinase 1, N-terminally processed	-0.395	0.206	0.285	4.410E-08
	Q00987	MDM2	E3 ubiquitin-protein ligase Mdm2	-0.458	0.166	0.275	2.789E-08
	P32004	L1CAM	Neural cell adhesion molecule L1	-0.466	0.118	0.235	2.519E-12
	P15391	CD19	B-lymphocyte antigen CD19	-0.272	0.171	0.216	4.123E-08
	P07948	LYN	Tyrosine-protein kinase Lyn	-0.109	0.408	0.211	3.099E-04
	O14745	SLC9A3R1	Na(+)/H(+) exchange regulatory cofactor NHE-RF1	-0.172	0.224	0.196	4.627E-07
	O43559	FRS3	Fibroblast growth factor receptor substrate 3	-0.091	0.317	0.170	3.717E-08
	P43146	DCC	Netrin receptor DCC	-0.392	0.070	0.165	5.835E-04
	P62158	CALM1 ; CALM2 ; CALM3	Calmodulin-1 {ECO:0000312|HGNC:HGNC:1442}	-0.455	0.054	0.156	1.670E-10
	P42574	**CASP3**	**Caspase-3 subunit p12**	-0.034	0.656	0.149	8.050E-08
	P42684	ABL2	Abelson tyrosine-protein kinase 2	-0.362	0.045	0.128	1.676E-11
	P17081	RHOQ	Rho-related GTP-binding protein RhoQ	-0.362	0.045	0.128	1.676E-11
	Q13905	RAPGEF1	Rap guanine nucleotide exchange factor 1	-0.187	0.080	0.122	2.094E-04
	P05155	**SERPING1**	**Plasma protease C1 inhibitor**	-0.023	0.362	0.091	1.014E-11
	Q92793	CREBBP	CREB-binding protein	-0.506	0.015	0.089	4.511E-11
	P07585	DCN	Decorin	-0.023	0.351	0.089	2.430E-11
	P12830	CDH1	Cadherin-1	-0.503	0.011	0.076	1.487E-14
	Q07157	TJP1	Tight junction protein ZO-1	-0.407	0.011	0.068	2.640E-12
	Q92990	GLMN	Glomulin	-0.294	0.011	0.058	2.056E-07
	P55075	FGF8	Fibroblast growth factor 8	-0.011	0.238	0.052	1.884E-05

Supplementary Table 4: Top 10 gene Ontology functions enriched from best-classifier proteins with opposite signal in Macular Degeneration (MD) MoAs. Functional enrichment analysis from FuncAssociate.

	Low-MD active / High-MD inactive			**Low-MD inactive / High-MD active**			**Overlapped functions**		
	GO name	**LOD**	**P-val.**	**GO name**	**LOD**	**P-val.**	**GO name**	**LOD**	**P-val.**
1	dendritic spine development	2.41	0.00150	dorsal/ventral axon guidance	3.07	0.01950	phosphatidylinositol-4,5-bisphosphate 3-kinase activity	1.89	<0.00050
2	positive regulation of vascular endothelial growth factor production	2.04	0.03000	fibroblast growth factor receptor binding	2.06	0.02000	phosphatidylinositol bisphosphate kinase activity	1.87	<0.00050
3	regulation of intracellular estrogen receptor signaling pathway	2.00	0.00150	platelet-derived growth factor receptor signaling pathway	1.95	0.03350	phosphatidylinositol 3-kinase activity	1.84	<0.00050
4	regulation of systemic arterial blood pressure	1.97	0.03300	non-membrane spanning protein tyrosine kinase activity	1.88	0.04000	phosphatidylinositol phosphorylation	1.72	<0.00050
5	regulation of vascular endothelial growth factor production	1.96	0.04450	growth factor receptor binding	1.85	<0.00050	single-organism cellular process	1.53	<0.00050
6	peptide hormone processing	1.96	0.04450	regulation of blood coagulation	1.68	0.01050	lipid phosphorylation	1.67	<0.00050
7	phosphotyrosine binding	1.91	0.04900	regulation of hemostasis	1.68	0.01050	positive regulation of protein kinase B signaling	1.60	<0.00050
8	neutrophil chemotaxis	1.90	0.05000	regulation of coagulation	1.66	0.01050	biological regulation	1.42	<0.00050
9	regulation of vasoconstriction	1.89	0.00150	regulation of phosphatidylinositol 3-kinase signaling	1.58	0.02350	protein binding	1.41	<0.00050
10	vascular endothelial growth factor receptor signaling pathway	1.83	<0.00050	response to toxic substance	1.53	0.00350	regulation of response to stimulus	1.24	<0.00050

Supplementary Table 5: Modified Hausdorff distance between the 4 groups of MoAs defined.

	LowMD	HighMD	HighHF	LowHF
LowMD	0	4.00226983	2.7537393	2.6068664
HighMD	4.00226983	0	2.1150102	2.55445687
HighHF	2.7537393	2.1150102	0	4.01919608
LowHF	2.6068664	2.55445687	4.01919608	0

Supplementary Table 6: Mean Euclidean distance between each one of the points of every group of MoAs and its centre.

	Mean distance from center
LowMD	3.137818031
HighMD	3.171767895
HighHF	3.298746704
LowHF	3.523965485

Supplementary Table 7: Number of common MoAs between the 4 groups of MoAs defined.

	LowMD	HighMD	HighHF	LowHF
LowMD	50	0	9	13
HighMD	0	50	17	12
HighHF	9	17	50	0
LowHF	13	12	0	50

Supplementary Table 8: Intersection of several set of proteins defined with GUILDify with the best-classifier proteins (BCP) obtained from the TPMS analysis. The p-values are calculated using a Fisher's exact test. The p-values above 0.05 are remarked in red.

Sets of proteins	**# LHF+ HHF-**	**P-value**	**# LHF- HHF+**	**P-value**	**# LMD+ HMD-**	**P-value**	**# LMD- HMD+**	**P-value**
Drug seeds	0	1.00E+00	0	1.00E+00	0	1.00E+00	0	1.00E+00
Top-Drug	0	1.00E+00	0	1.00E+00	2	1.11E-01	0	1.00E+00
HF seeds	0	1.00E+00	3	6.32E-03	1	2.32E-01	1	2.39E-01
Top-HF	0	1.00E+00	3	4.34E-02	2	1.02E-01	1	4.35E-01
MD seeds	0	1.00E+00	1	4.02E-01	2	4.81E-02	5	2.76E-05
Top-MD	0	1.00E+00	1	5.60E-01	3	1.81E-02	5	2.51E-04
Top-HF∪Top-MD∪Top-Drug	**0**	1.00E+00	**3**	3.70E-01	**5**	1.53E-02	**5**	1.77E-02

Supplementary Table 9: Best-classifier proteins found in the Top-HF ∪ Top-MD ∪ Top-Drug set.

	Uniprot ID	**Gene symbol**	**Gene name**
LHF- HHF+	P28482	MAPK1	Mitogen-activated protein kinase 1
	P27361	MAPK3	Mitogen-activated protein kinase 3
	P02751	FN1	Fibronectin
LMD+ HMD-	P18084	ITGB5	Integrin beta-5
	O75787	ATP6AP2	V-ATPase M8.9 subunit
	Q02297	NRG1	Pro-neuregulin-1, membrane-bound isoform
	P06748	NPM1	Nucleophosmin
	P01583	IL1A	Interleukin-1 alpha
LMD- HMD+	P04085	PDGFA	Platelet-derived growth factor subunit A
	P02675	FGB	Fibrinogen beta chain

	P05155	SERPING1	Plasma protease C1 inhibitor
	P05230	FGF1	Fibroblast growth factor 1
	P42574	CASP3	Caspase-3 subunit p12

Supplementary Table 10: Intersection of several set of proteins defined with GUILDify with the biomarkers obtained from the TPMS analysis. The p-values are calculated using a Fisher's exact test. The p-values above 0.05 are remarked in red.

Sets of proteins	**LHF ∩ LMD+ HMD-**	**P-value**	**LHF ∩ LMD- HMD+**	**P-value**
Drug seeds	0	1.00E+00	0	1.00E+00
Top-Drug	1	2.90E-01	0	1.00E+00
HF seeds	1	1.45E-01	1	1.28E-01
Top-HF	**2**	4.03E-02	1	2.48E-01
MD seeds	**2**	1.80E-02	**4**	2.54E-05
Top-MD	**4**	2.75E-04	**4**	1.56E-04
Top-HF∪Top-MD∪Top-Drug	**5**	1.37E-03	**5**	6.89E-04

Supplementary Table 11: Biomarkers from the TPMS analysis found in the Top-HF ∪ Top-MD ∪ Top-Drug set.

	Uniprot ID	**Gene symbol**	**Gene name**
LHF ∩ LMD+ HMD-	Q02297	NRG1	Pro-neuregulin-1, membrane-bound isoform
	P06748	NPM1	Nucleophosmin
	P01583	IL1A	Interleukin-1 alpha
	P61981	YWHAG	14-3-3 protein gamma, N-terminally processed
	P18084	ITGB5	Integrin beta-5
LHF ∩ LMD- HMD+	P05121	SERPINE1	Plasminogen activator inhibitor 1
	P02675	FGB	Fibrinogen beta chain

	P05230	FGF1	Fibroblast growth factor 1
	Q15109	AGER	Advanced glycosylation end product-specific receptor
	P05155	SERPING1	Plasma protease C1 inhibitor

Supplementary Table 12: Pathophysiological processes present in Heart Failure characterization used in the study.

Pathophysiological processes	# proteins
1- Cardiomyocyte cell death (including apoptosis and necrosis)	46
2- Left ventricle extracellular matrix remodelling	37
3- Impaired myocyte contractility	35
4- Hypertrophy	33

Supplementary Table 13: Pathophysiological processes present in Macular Degeneration characterization used in the study.

Pathophysiological processes	# proteins
1- Light and oxidative stress: lipid oxidation, lipofuscin, advanced glycation end products (AGEs)	46
2- Debris accumulation: Drusen and protein aggregation	37
3- Disturbance of lysosomal clearance	35
4- Autophagy dysregulation	33
5- Immunological processes: chronic inflammation	46
6- Mitochondrial defects	37
7- Extracellular matrix remodelling and Bruch's membrane thickening	35
8- Lipoprotein/lipid metabolism	33
9- Retinal cell death	35
10- Choroidal neovascularization (wet Age-Related MD)	33

Supplementary material references

1. Anaxomics Biotech SL. Biological Effectors Database [Internet]. 2018. Available from: http://www.anaxomics.com/biological-effectors-database.php
2. Pujol A, Mosca R, Farrés J, Aloy P. Unveiling the role of network and systems biology in drug discovery. Trends Pharmacol Sci. 2010;31(3):115–23.
3. Iborra-Egea O, Gálvez-Montón C, Roura S, Perea-Gil I, Prat-Vidal C, Soler-Botija C, et al. Mechanisms of action of sacubitril/valsartan on cardiac remodeling: a systems biology approach. Npj Syst Biol Appl. 2017;3(1):1–8.
4. European Medicines Agency. Entresto: EPAR - Public assessment report. London; 2015.
5. Leger F, Fernagut PO, Canron MH, Léoni S, Vital C, Tison F, et al. Protein aggregation in the aging retina. J Neuropathol Exp Neurol. 2011;70(1):63–8.
6. Hyttinen JMT, Amadio M, Viiri J, Pascale A, Salminen A, Kaarniranta K. Clearance of misfolded and aggregated proteins by aggrephagy and implications for aggregation diseases. Ageing Research Reviews. 2014. p. 16–28.
7. Chiras D, Kitsos G, Petersen MB, Skalidakis I, Kroupis C. Oxidative stress in dry age-related macular degeneration and exfoliation syndrome. Critical Reviews in Clinical Laboratory Sciences. 2015. p. 12–27.
8. Nowak JZ. AMD-the retinal disease with an unprecised etiopathogenesis: In search of effective therapeutics. Acta Pol Pharm - Drug Res. 2014;71(6):900–16.
9. Anaxomics Biotech SL. TPMS technology [Internet]. 2018. Available from: http://www.anaxomics.com/tpms.php
10. Herrando-Grabulosa M, Mulet R, Pujol A, Mas JM, Navarro X, Aloy P, et al. Novel Neuroprotective Multicomponent Therapy for

Amyotrophic Lateral Sclerosis Designed by Networked Systems. PloS One. 2016;11(1):e0147626.

11. Gómez-Serrano M, Camafeita E, García-Santos E, López JA, Rubio MA, Sánchez-Pernaute A, et al. Proteome-wide alterations on adipose tissue from obese patients as age-, diabetes- and gender-specific hallmarks. Sci Rep. 2016;6(January):1–15.
12. Perera S, Artigas L, Mulet R, Mas JM, Sardón T. Systems biology applied to non-alcoholic fatty liver disease (NAFLD): treatment selection based on the mechanism of action of nutraceuticals. Nutrafoods. 2014;13(2):61–8.
13. Romeo-Guitart D, Forés J, Herrando-Grabulosa M, Valls R, Leiva-Rodríguez T, Galea E, et al. Neuroprotective Drug for Nerve Trauma Revealed Using Artificial Intelligence. Sci Rep. 2018;8:1879.
14. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: A major update to the DrugBank database for 2018. Nucleic Acids Res. 2018;46(D1):D1074–82.
15. Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, et al. PubChem substance and compound databases. Nucleic Acids Res. 2016;44(D1):D1202–13.
16. Szklarczyk D, Santos A, Von Mering C, Jensen LJ, Bork P, Kuhn M. STITCH 5: Augmenting protein-chemical interaction networks with tissue and affinity data. Nucleic Acids Res. 2016;44(D1):D380–4.
17. Hecker N, Ahmed J, von Eichborn J, Dunkel M, Macha K, Eckert A, et al. SuperTarget goes quantitative: update on drug-target interactions. Nucleic Acids Res. 2011;40(D1):D1113–7.
18. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: New perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017;45(D1):D353–61.
19. Chatr-Aryamontri A, Oughtred R, Boucher L, Rust J, Chang C, Kolas NK, et al. The BioGRID interaction database: 2017 update. Nucleic Acids Res. 2017;45(D1):D369–79.
20. Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, et al. The MIntAct project - IntAct as a common curation

platform for 11 molecular interaction databases. Nucleic Acids Res. 2014;42(D1):358–63.

21. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, et al. The Reactome Pathway Knowledgebase. Nucleic Acids Res. 2018;46(D1):D649–55.
22. Han H, Cho JW, Lee S, Yun A, Kim H, Bae D, et al. TRRUST v2: An expanded reference database of human and mouse transcriptional regulatory interactions. Nucleic Acids Res. 2018;46(D1):D380–6.
23. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, et al. Human Protein Reference Database - 2009 update. Nucleic Acids Res. 2009;37(D1):D767–72.
24. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. The Database of Interacting Proteins: 2004 update. Nucleic Acids Res. 2004;32(90001):449D – 451.
25. Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. Nucleic Acids Res. 2016;44(D1):D1075–9.
26. Liu Y, Morley M, Brandimarto J, Hannenhalli S, Hu Y, Ashley EA, et al. RNA-Seq identifies novel myocardial gene expression signatures of heart failure. Genomics. 2015;105(2):83–9.
27. Collet P, Rennard J-P. Stochastic Optimization Algorithms. Intell Inf Technol. 2011;1121–37.
28. Dubuisson M-P, Jain AK. A modified Hausdorff distance for object matching. Proc 12th Int Conf Pattern Recognit. 1994;1(1):566–8.
29. Burnett M. Blocking Brute Force Attacks. UVA Comput Sci. 2007;
30. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B Stat Methodol. 2005;67(2):301–20.
31. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
32. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. J R Stat Soc Ser B Methodol. 1996;58(1):267–88.

33. Ho TK. Random decision forests. In: Proceedings of the International Conference on Document Analysis and Recognition, ICDAR. 1995. p. 278–82.

34. Madsen H, Thyregod P. A Generalized linear Model with binomial distribution and probit link function has been used as classifier. In: Introduction to General and Generalized Linear Models. Chapman & Hall/CRC; 2011.

35. Kira K, Rendell LA. Feature selection problem: traditional methods and a new algorithm. In: Proceedings Tenth National Conference on Artificial Intelligence. 1992.

36. Xuan G, Zhu X, Chai P, Zhang Z, Shi YQ, Fu D. Feature selection based on the Bhattacharyya distance. In: Proceedings - International Conference on Pattern Recognition. 2006.

37. Keinosuke Fukunaga. Introduction to statistical pattern recognition 2nd edition. Academic Press. 1990.

38. Christin C, Hoefsloot HCJ, Smilde AK, Hoekman B, Suits F, Bischoff R, et al. A critical assessment of feature selection methods for biomarker discovery in clinical proteomics. Mol Cell Proteomics. 2013;12(1):263–76.

39. Haykin S. Neural networks: a comprehensive foundation. The Knowledge Engineering Review. 1994.

40. Gorban AN, Zinovyev AY. Principal Graphs and Manifolds. In: Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods and Techniques. Information Science Reference; 2009. p. 28–59.

41. Shimizu K, Short DA, Kedem B. Single- and Double-Threshold Methods for Estimating the Variance of Area Rain Rate. J Meteorol. 1993;71(6).

42. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. Proc 14th Int Jt Conf Artif Intell - Vol 2. 1995;2(12):1137–43.

43. BIPM. Guides in metrology, Guide to the Expression of Uncertainty in Measurement (GUM) and International Vocabulary of Metrology (VIM). 2008.

44. Berriz GF, Beaver JE, Cenik C, Tasan M, Roth FP. Next generation software for functional trend analysis. Bioinformatics. 2009;25(22):3043–4.

45. Patel VB, Wang Z, Fan D, Zhabyeyev P, Basu R, Das SK, et al. Loss of p47phox subunit enhances susceptibility to biomechanical stress and heart failure because of dysregulation of cortactin and actin filaments. Circ Res. 2013;112(12):1542–56.

46. Karsanov N V., Pirtskhalaishvili MP, Semerikova VJ, Losaberidze NS. Thin myofilament proteins in norm and heart failure I. Polymerizability of myocardial Straub actin in acute and chronic heart failure. Basic Res Cardiol. 1986;81(2):199–212.

47. Childers RC, Sunyecz I, West TA, Cismowski MJ, Lucchesi PA, Gooch KJ. Role of the Cytoskeleton in the Development of a Hypofibrotic Cardiac Fibroblast Phenotype in Volume Overload Heart Failure. Am J Physiol Heart Circ Physiol. 2018;316(3):H596–608.

48. Liu YH, Yang XP, Sharov VG, Nass O, Sabbah HN, Peterson E, et al. Effects of angiotensin-converting enzyme inhibitors and angiotensin II type 1 receptor antagonists in rats with heart failure: Role of kinins and angiotensin II type 2 receptors. J Clin Invest. 1997;99(8):1926–35.

49. Schrier RW, Abdallah JG, Weinberger HHD, Abraham WT. Therapy of heart failure. Kidney Int. 2000;57(4):1418–25.

50. Aoyagi T, Matsui T. Phosphoinositide-3 kinase signaling in cardiac hypertrophy and heart failure. Curr Pharm Des. 2011;17(18):1818–24.

51. Ennis I, Aiello E, Cingolani H, Perez N. The Autocrine/Paracrine Loop After Myocardial Stretch: Mineralocorticoid Receptor Activation. Curr Cardiol Rev. 2013;9(3):230–40.

52. Sullivan RKP, WoldeMussie E, Pow D V. Dendritic and synaptic plasticity of neurons in the human age-related macular degeneration retina. Invest Ophthalmol Vis Sci. 2007;48(6):2782–91.
53. Sohn YI, Lee NJ, Chung A, Saavedra JM, Scott Turner R, Pak DTS, et al. Antihypertensive drug Valsartan promotes dendritic spine density by altering AMPA receptor trafficking. Biochem Biophys Res Commun. 2013;439(4):464–70.
54. Frank RN. Growth factors in age-related macular degeneration: Pathogenic and therapeutic implications. Ophthalmic Res. 1997;29(5):341–53.
55. Glenn J V., Stitt AW. The role of advanced glycation end products in retinal ageing and disease. Biochim Biophys Acta - Gen Subj. 2009;1790(10):1109–16.
56. Grossniklaus HE, Green WR. Choroidal neovascularization. Am J Ophthalmol. 2004;137(3):496–503.
57. Yuan X, Gu X, Crabb JS, Yue X, Shadrach K, Hollyfield JG, et al. Quantitative Proteomics: Comparison of the Macular Bruch Membrane/Choroid Complex from Age-related Macular Degeneration and Normal Eyes. Mol Cell Proteomics. 2010;9(6):1031–46.
58. Lee AY, Kulkarni M, Fang AM, Edelstein S, Osborn MP, Brantley MA. The effect of genetic variants in SERPING1 on the risk of neovascular age-related macular degeneration. Br J Ophthalmol. 2010;94(7):915–7.
59. Higgins P. Balancing AhR-Dependent Pro-Oxidant and Nrf2-Responsive Anti-Oxidant Pathways in Age-Related Retinopathy: Is SERPINE1 Expression a Therapeutic Target in Disease Onset and Progression? J Mol Genet Med. 2015;8(2):101.
60. Miyata M, Ikeda Y, Nakamura S, Sasaki T, Abe S, Minagoe S, et al. Effects of Valsartan on Fibrinolysis in Hypertensive Patients With Metabolic Syndrome. Circ J. 2012;76(4):843–51.
61. Oubiña MP, De las Heras N, Vázquez-Pérez S, Cediel E, Sanz-Rosa D, Ruilope LM, et al. Valsartan improves fibrinolytic balance in atherosclerotic rabbits. J Hypertens. 2002;20(2):303–10.

62. Albert-Fort M, Hombrebueno JR, Pons-Vazquez S, Sanz-Gonzalez S, Diaz-Llopis M, Pinazo-Durán MD. Retinal neurodegenerative changes in the adult insulin receptor substrate-2 deficient mouse. Exp Eye Res. 2014;124:1–10.

63. Zhang R, Liu Z, Zhang H, Zhang Y, Lin D. The COX-2-selective antagonist (NS-398) inhibits choroidal neovascularization and subretinal fibrosis. PLoS ONE. 2016;11(1):e0146808.

64. Wang X, Ma W, Han S, Meng Z, Zhao L, Yin Y, et al. TGF-β participates choroid neovascularization through Smad2/3-VEGF/TNF-α signaling in mice with Laser-induced wet age-related macular degeneration. Sci Rep. 2017;7(1):9672.

65. Skeie JM, Zeng S, Faidley EA, Mullins RF. Angiogenin in age-related macular degeneration. Mol Vis. 2011;17:576–82.

66. Aguirre-Plans J, Piñero J, Sanz F, Furlong LI, Fernandez-Fuentes N, Oliva B, et al. GUILDify v2.0: A Tool to Identify Molecular Networks Underlying Human Diseases, Their Comorbidities and Their Druggable Targets. J Mol Biol. 2019;30117–2.

67. Hegab Z, Gibbons S, Neyses L, Mamas M. Role of advanced glycation end products in cardiovascular disease. World J Cardiol. 2012;4(4):90–102.

68. Banevicius M, Vilkeviciute A, Kriauciuniene L, Liutkeviciene R, Deltuva VP. The Association Between Variants of Receptor for Advanced Glycation End Products (RAGE) Gene Polymorphisms and Age-Related Macular Degeneration. Med Sci Monit. 2018;24:190–9.

69. Pickering RJ, Tikellis C, Rosado CJ, Tsorotes D, Dimitropoulos A, Smith M, et al. Transactivation of RAGE mediates angiotensin-induced inflammation and atherogenesis. J Clin Invest. 2019;129(1):406–21.

70. Garbayo E, Gavira JJ, De Yebenes MG, Pelacho B, Abizanda G, Lana H, et al. Catheter-based intramyocardial injection of FGF1 or NRG1-loaded MPs improves cardiac function in a preclinical model of ischemia-reperfusion. Sci Rep. 2016;6:25932.

71. Lakó-Futó Z, Szokodi I, Sármán B, Földes G, Tokola H, Ilves M, et al. Evidence for a Functional Role of Angiotensin II Type 2 Receptor in the Cardiac Hypertrophic Process in Vivo in the Rat Heart. Circulation. 2003;108(19):2414–22.
72. Galindo CL, Ryzhov S, Sawyer DB. Neuregulin as a heart failure therapy and mediator of reverse remodeling. Curr Heart Fail Rep. 2014;11(1):40–9.
73. Xu J, De Winter F, Farrokhi C, Rockenstein E, Mante M, Adame A, et al. Neuregulin 1 improves cognitive deficits and neuropathology in an Alzheimer's disease model. Sci Rep. 2016;6:31692.
74. Kaarniranta K, Salminen A, Haapasalo A, Soininen H, Hiltunen M. Age-related macular degeneration (AMD): Alzheimer's disease in the eye? J Alzheimers Dis. 2011;24(4):615–31.
75. Verweij N, Eppinga RN, Hagemeijer Y, Van Der Harst P. Identification of 15 novel risk loci for coronary artery disease and genetic risk of recurrent events, atrial fibrillation and heart failure. Sci Rep. 2017;7(1):2761.
76. Kaarniranta K, Sinha D, Blasiak J, Kauppinen A, Veréb Z, Salminen A, et al. Autophagy and heterophagy dysregulation leads to retinal pigment epithelium dysfunction and development of age-related macular degeneration. Autophagy. 2013;9(7):973–84.
77. Kawano H, Cody RJ, Graf K, Goetze S, Kawano Y, Schnee J, et al. Angiotensin II Enhances Integrin and α-Actinin Expression in Adult Rat Cardiac Fibroblasts. Hypertension. 2012;35(1 Pt 2):273–9.
78. Bujak M, Frangogiannis NG. The role of IL-1 in the pathogenesis of heart disease. Arch Immunol Ther Exp (Warsz). 2009;57(3):165–76.
79. Turner NA. Effects of interleukin-1 on cardiac fibroblast function: Relevance to post-myocardial infarction remodelling. Vascul Pharmacol. 2014;60(1):1–7.
80. Nassar K, Grisanti S, Elfar E, Lüke J, Lüke M, Grisanti S. Serum cytokines as biomarkers for age-related macular degeneration. Graefes Arch Clin Exp Ophthalmol. 2015;253(5):699–704.

81. Zhang YN, Vernooij F, Ibrahim I, Ooi S, Gijsberts CM, Schoneveld AH, et al. Extracellular vesicle proteins associated with systemic vascular events correlate with heart failure: An observational study in a dyspnoea cohort. PLoS ONE. 2016;11(1):e0148073.
82. Zaman AKMT, French CJ, Schneider DJ, Sobel BE. A Profibrotic Effect of Plasminogen Activator Inhibitor Type-1 (PAI-1) in the Heart. Exp Biol Med. 2009;234(3):246–54.
83. Messaoudi S, Azibani F, Delcayre C, Jaisser F. Aldosterone, mineralocorticoid receptor, and heart failure. Mol Cell Endocrinol. 2012;350(2):266–72.
84. Chakravarthy U, Wong TY, Fletcher A, Piault E, Evans C, Zlateva G, et al. Clinical risk factors for age-related macular degeneration: A systematic review and meta-analysis. BMC Ophthalmol. 2010;10:31.

3.4. Modelling of drug-induced liver injury based on multi-omics integration and machine learning prediction

In the fourth article of the thesis, I present the participation of our laboratory in collaboration with several members of the TransQST project in the Connectivity Map (CMap) Drug Safety Challenge of the International Conference on Critical Assessment of Massive Data Analysis (CAMDA) of 2019. The aim of the challenge was to make *in silico* methods to predict Drug-Induced Liver Injury (DILI), an adverse reaction caused by the intake of drugs that produces liver damage. In special, the organizers propose the use of CMap gene expression data in combination with other sources of data such as chemical structures and cellular images to predict the adverse reaction.

Our group employed several network medicine methods to guide the identification of DILI gene signatures, which are used as features to train a machine learning algorithm. We described and assessed the usage of different types of biological data as features separately and in combination, and compared them with the state-of-the-art results from previous editions.

Aguirre-Plans J, Piñero J, Souza T, Callegaro G, Kunnen SJ, Sanz F, Fernandez-Fuentes N, Furlong LI, Guney E, Oliva B. An ensemble learning approach for modeling the systems biology of drug-induced injury. ***Biol Direct*.** 2021; 16(1):5.
DOI: 10.1186/s13062-020-00288-x

An ensemble learning approach for modeling the systems biology of drug-induced injury

Biol Direct. 2021 Jan 12;16(1):5

Joaquim Aguirre-Plans[1], Janet Piñero[1], Terezinha Souza[2], Giulia Callegaro[3], Steven J. Kunnen[3], Ferran Sanz[1], Narcis Fernandez-Fuentes[4,5,*], Laura I. Furlong[1,*], Emre Guney[1,*], Baldo Oliva[1,*]

[1]**Research Programme on Biomedical Informatics (GRIB)**, Hospital del Mar Medical Research Institute (IMIM), DCEXS, Pompeu Fabra University (UPF), Barcelona, Spain
[2]**Department of Toxicogenomics**, Maastricht University, The Netherlands
[3]**Leiden Academic Centre for Drug Research**, Leiden University, Leiden, The Netherlands
[4]**Department of Biosciences**, U Science Tech, Universitat de Vic-Universitat Central de Catalunya, Vic, Spain
[5]**Institute of Biological, Environmental and Rural Sciences**, Aberystwyth University, Aberystwyth, UK.

***Co-senior authors contributed equally**

Abstract

Background: Drug-induced liver injury (DILI) is an adverse reaction caused by the intake of drugs of common use that produces liver damage. The impact of DILI is estimated to affect around 20 in 100,000 inhabitants worldwide each year. Despite being one of the main causes of liver failure, the pathophysiology and mechanisms of

DILI are poorly understood. In the present study, we developed an ensemble learning approach based on different features (CMap gene expression, chemical structures, drug targets) to predict drugs that might cause DILI and gain a better understanding of the mechanisms linked to the adverse reaction.

Results: We searched for gene signatures in CMap gene expression data by using two approaches: phenotype-gene associations data from DisGeNET, and a non-parametric test comparing gene expression of DILI-Concern and No-DILI-Concern drugs (as per DILIrank definitions). The average accuracy of the classifiers in both approaches was 69%. We used chemical structures as features, obtaining an accuracy of 65%. The combination of both types of features produced an accuracy around 63%, but improved the independent hold-out test up to 67%. The use of drug-target associations as feature obtained the best accuracy (70%) in the independent hold-out test.

Conclusions: When using CMap gene expression data, searching for a specific gene signature among the landmark genes improves the quality of the classifiers, but it is still limited by the intrinsic noise of the dataset. When using chemical structures as a feature, the structural diversity of the known DILI-causing drugs hampers the prediction, which is a similar problem as for the use of gene expression information. The combination of both features did not improve the quality of the classifiers but increased the robustness as shown on independent hold-out tests. The use of drug-target associations as feature improved the prediction, specially the specificity, and the results were comparable to previous research studies.

Keywords: CAMDA; Cmap; Drug safety; Drug structure; Drug-induced liver injury; Hepatotoxicity; Machine learning; Systems biology.

Background

Drug safety is one of the main reasons of drug attrition during development (1,2). Although the causes of drug failure due to lack of safety are several, hepatic adverse reactions are among the most important, particularly at late drug development stages (3,4). Drug-induced liver injury (also named DILI) is an adverse reaction caused by the intake of drugs of common use that produces liver damage. DILI has a relatively high incidence rate, estimated to affect around 20 in 100,000 inhabitants worldwide each year (5). Many drugs ranging from pain killers to anti-tuberculous treatments can cause DILI (6). Despite DILI being one of the leading causes of acute liver failure, the pathophysiology and etiology of DILI is poorly understood and pinpointing the toxicity of compounds in human liver remains a non-trivial task (7).

Several *in-silico* methods have been proposed to predict hepatotoxicity of drugs. Among these, machine learning models trained using drug structural features have shown a good accuracy (8–10). Furthermore, incorporating gene- and pathway-level signatures from transcriptomics data has shown a high predictive accuracy using Deep Neural Networks (11). With the recent increased interest on machine learning methods to predict drug-induced toxicity, the International Conference on Critical Assessment of Massive Data Analysis (CAMDA) has been organizing the Connectivity Map (CMap) Drug Safety Challenge since 2018. The aim of the challenge was to assess the state-of-the-art on DILI prediction methods using different sources of data such as transcriptomics data, chemical structures, and cellular images. In the first edition (CAMDA 2018), the two published studies applied various machine learning methods for DILI prediction on the CMap gene expression data provided (in MCF7 and PH3 cell lines), obtaining poor predictive results (12,13). *Sumsion et al.* (12) evaluated 7 different classification algorithms and built a soft-voting classifier that combined all classifiers. Still, the accuracy results of the best performing classifiers (random forest and soft-voting) were around 70%, obtaining high sensitivity (77%) but low specificity (13-19%). They also explored different strategies to improve the results, such as normalizing gene expression data across samples, feature selection methods, adjusting class imbalance or improving the voting-based classifier. Still, the improvement of the results with each of these solutions was limited. *Chierici et al.* (13) used three deep learning classifiers and compared them with random forest and multi-layer perceptron classifiers. They also tested several strategies for balancing data and alternative train/test splits. However, the different strategies gave an overall poor performance, in which the Matthews correlation coefficient (MCC) values ranged from −0.04 to 0.21 in cross-validation and −0.16 to 0.11 in the independent hold-out test set. In both *Sumsion et al.* (12) and *Chierici et al.* (13), the limited results were attributed to having a small and highly imbalanced gold standard of 190 drugs for training (160 DILI-causing) and 86 drugs for an independent hold-out test. This problem is still present in the current edition of CAMDA (2019), as the size of the gold standard is still limited. The organizers provided a gold standard (from DILIrank dataset (14)) composed of 175 drugs for training and 55 for an independent hold-out test. They also provided a dataset of CMap L1000 gene expression responses for 1,314 compounds (15) (including the 230 drugs of the gold standard), the chemical structures (SMILES codes) of the drugs and annotated images from cell perturbation assays for a subset of 826 compounds (156 from DILIrank) (16).

In this study, we implemented an ensemble learning approach to predict drugs that can cause DILI in human liver. We experimented the inclusion in the classifiers of several features derived from transcriptomics, drug-target associations and structural data either separately or combined (**Table 1**). We investigated whether it was feasible to find a DILI gene signature using phenotype-gene associations, protein-protein interactions and gene expression data. We observed that finding a meaningful gene signature can improve the quality of the classifier instead of using all landmark genes defined in the CMap platform (i.e. the subset of 978 genes whose gene expression has been determined as informative enough to characterize the whole transcriptome (15)). We also analyzed the accuracy of the prediction when using chemical structures, drug-target information, and the combination of these together with transcriptomics data. We compared the quality of the classifiers made from these features in a robust machine learning pipeline and presented a list of conclusions that might serve as starting points for further studies.

Table 1. Summary of the features used in the classification task.

Type of feature	Name	Description
Gene expression features	Landmark genes	978 genes directly measured from the L1000 datasets
	DisGeNET DILI genes	Curated genes associated to 9 phenotypes related with DILI from DisGeNET database
	GUILDify DILI genes	Genes associated through the protein interactions network to 6 phenotypes related with DILI using GUILDify
	DILI landmark genes	66 landmark genes selected using non-parametric test for each gene across all samples of Most/Less- vs. No-DILI-Concern drugs (P-value<0.05)
Structural features	SMILES	Line notation describing the chemical structure of drugs
Drug target genes	Set of targets	1,664 drug targets retrieved from DGIdb, HitPick and SEA

Methods

1. Gold standard data on drugs causing DILI

The CAMDA challenge provided the DILIrank dataset (14) as the gold standard data of known DILI compounds. DILIrank is a dataset that classifies the drugs in three levels of DILI severity: “Most-DILI-Concern” when the drug was withdrawn for DILI-related causes or labelled with severe DILI indication; “Less-DILI-Concern” when the drug was labelled with mild DILI indication or adverse reactions; and “No-DILI-Concern” when no DILI was indicated in any of the labelling sections. Moreover, these levels of severity were verified using the standardized clinical causality assessment system, and the drugs that were not meeting the expected severity were reclassified as “Ambiguous-DILI-Concern”. Among all the drugs categorized in DILIrank, the CAMDA challenge provided data for 230 drugs: 37 Most-DILI-Concern, 87 Less-DILI-Concern, 51 No-DILI-Concern and 55 Ambiguous-DILI-Concern. Additionally, the US Food and Drug Administration classified the remaining 55 Ambiguous-DILI-Concern drugs as DILI or No-DILI-Concern. These 55 drugs served as a dataset for an independent hold-out test, because the actual severity category of the drug remained hidden.

2. Data collection

CMap gene expression

The gene expression data used in this study was gathered from the CMap L1000 Assay Platform (15). The L1000 Assay Platform provides more than one million gene expression profiles from a wide range of cell lines treated with different drugs at different doses and treatment durations. Assuming that gene expression is highly correlated, the Platform features a subset of approximately 1000 landmark genes to derive profiles that serve to infer the expression of the rest of genes. We used CMAP L1000 level 5 data which contained z-score values corresponding to the normalized differential expression between the drug treatment and control across different conditions.

Genes associated to DILI related phenotypes

We manually curated a list of phenotypes closely related with DILI and identified the genes associated with these phenotypes using the DisGeNET database v6.0 (17) (**Table 2**). We restricted disease-gene associations solely to expertly curated repositories: UniProt (18), the Comparative Toxicogenomics Database (CTD) (19), ORPHANET (20), the Clinical Genome Resource (CLINGEN) (21), the Genomics England PanelApp (22) and the Cancer Genome Interpreter (CGI) (23). We kept only the phenotypes with at least 10 curated gene associations. The full list of associations between DILI phenotypes and genes can be found at **Supplementary Table 1**.

Table 2. List of manually selected phenotypes related with DILI. The selected phenotypes were required to have 10 gene associations or more. The genetically redundant phenotypes have been merged in the same term. The empty cells correspond to phenotypes for which the expansion through the network using GUILDify was not functionally coherent.

DILI Phenotypes	UMLS	Number of genes associated in DisGeNET	Number of genes associated in GUILDify
Biliary cirrhosis	C0023892	33	86

Hepatitis, Drug-Induced;
Drug-Induced Liver Disease;
Drug-Induced Acute Liver Injury
	C1262760;
	C0860207;
	C3658290
		315	

Hyperammonemia	C0220994	104	148
Liver Cirrhosis; Fibrosis, Liver	C0023890; C0239946	97	145
Liver Cirrhosis, Alcoholic	C0023891	30	
Liver Dysfunction; Liver diseases	C0086565; C0023895	67	
Liver Failure, Acute	C0162557	22	118
Nonalcoholic Steatohepatitis; Non-alcoholic Fatty Liver Disease	C3241937; C0400966	42	67
Steatohepatitis; Fatty Liver	C2711227; C0015695	86	167
Number of different genes associated to DILI phenotypes		641	805

Drug chemical structure

The chemical structures of the drugs considered in the study were provided by the CAMDA challenge in the form of Simplified molecular-input line-entry system (SMILES) string. In order to use this type of data, we calculated the similarity between all compounds, creating a matrix of chemical similarity. Specifically, we used the R package *RxnSim* (24) to calculate the similarity matrix using the Tanimoto distance (25). We used the function *ms.compute.sim.matrix* (default parameters), which identifies the fingerprints of the SMILES and computes the fingerprint similarity between pairs of SMILES. The full list of SMILES is provided in **Supplementary Table 2**, and the matrix of Tanimoto similarity between SMILES in **Supplementary Table 3**.

Drug-target association

The targets of the compounds considered in the study were retrieved from three different databases: DGIdb (26), HitPick (27) and SEA (28). DGIdb gathers validated drug targets, whereas HitPick and SEA additionally provide predicted targets based on chemical similarity. We used the names of the drugs to retrieve the drug-protein associations from DGIdb, whereas the SMILES strings were used in the case of HitPick and SEA web servers. Any drug-protein pair that had been provided either by the database or predicted to interact by the web servers were included among the drug-target associations. This implies that there are no differences between validated and predicted targets. However, this allowed us to increase the number of input drugs and extended the potential recall of our method. After collecting all targets, a matrix was created with all the drugs in rows and all the target proteins in columns. The cells of the matrix had values 1 (if the drug targeted the protein) and 0 (otherwise). There are three drugs from the DILIrank dataset (alaproclate, fluvastatin and tenofovir) and two drugs from the independent hold-out test dataset (entecavir and vinorelbine) without any targets in these databases. These drugs have not been used neither for training nor for testing when using drug targets as features. The full list of drug-target associations is provided in **Supplementary Table 4**.

3. Prediction pipeline

We created a supervised machine learning pipeline (**Supplementary Figure 1**) to generate predictions using the features described in **Table 1**. The pipeline was implemented using the R package *caret* (29). Briefly, we used two classifiers: the random forest classifier and the gradient boosting machine. We limited the number of classifiers because CAMDA had a limited number of independent hold-out test trials, and we tested many different features. Thus, we focused on two tree-based ensemble methods that have been widely employed in previous research (30–32).

We created a balanced dataset containing the 30% of the data for testing and the rest for training. The original dataset is comprised of 124 drugs labelled as DILI (37 as Most-DILI-Concern and 87 as Less-DILI-Concern) and 51 labelled as no DILI. To create a balanced testing dataset, as there were less drugs labelled as no DILI, we randomly picked the 30% of the 51 no DILI drugs (15 drugs), and the same number of DILI drugs, maintaining the ratio of Most-DILI-Concern (29.8%) and Less-DILI-Concern (70.2%): 4 Most-DILI-Concern drugs (the 29.8% of 15) and 11 Less-DILI-Concern drugs (the 70.2% of 15). The rest of the drugs (109 DILI drugs and 36 no DILI drugs) were used for creating multiple training datasets. In order to have balanced training datasets, while at the same time, to cover as many DILI drugs as possible, we created 10 different training datasets. All of them have the same 36 no DILI drugs (corresponding to the 70% of the initial 51 drugs), but each of the training dataset has a different subset of DILI drugs. Accordingly, among the 109 DILI drugs, we picked randomly 11 Most-DILI-Concern drugs (29.8% of 36) and 25 Less-DILI-Concern (70.2% of 36) (see **Supplementary Figure 1** for a schematic representation of the procedure, and **Supplementary Table 5** for a detailed list of the number of drugs used in each step).

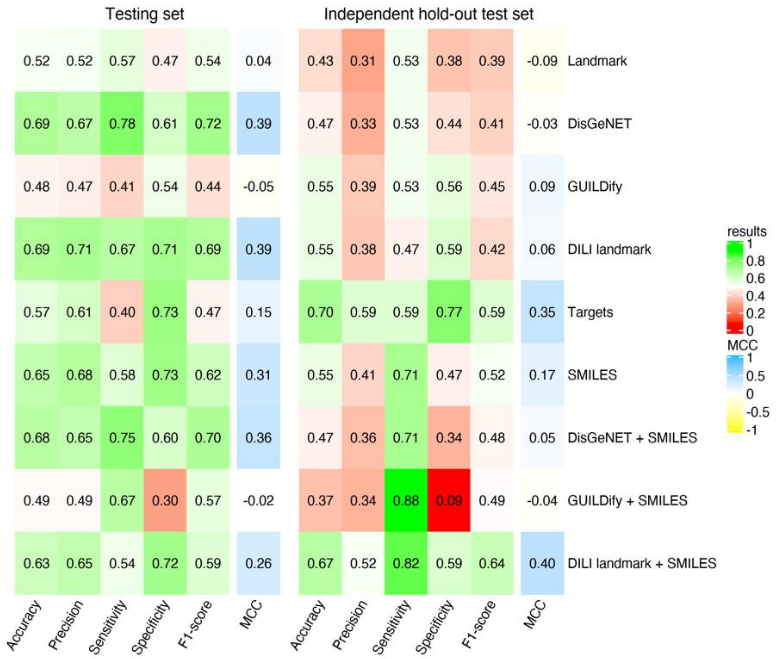
The 10 training datasets were used to train 10 different models. For each model, the hyperparameters of the machine learning classifier were tuned using the functions *trainControl* and *train* from the R package *caret* (29). Specifically, we used a 10-fold cross-validation approach, allowing resampling of the training set to avoid overfitting. The *train* function automatically tests different models using several combinations of hyperparameters and selects the model with higher accuracy. The 10 fitted models were evaluated using the testing dataset, obtaining a series of measures (accuracy, precision, sensitivity, specificity, F1-score, MCC) that indicate the quality of the model. Lastly, the testing set predictions of the 10 models were used as features to train a random forest classifier that combined them into a final model. The final model was used to classify the drugs of the independent hold-out test dataset into DILI drugs and non-DILI drugs.

Results

1. L1000 Connectivity Map data hints at transcriptomic heterogeneity of DILI compounds

CMap collects gene expression signatures obtained from cell lines upon treatments with different drug concentrations and durations. The treatment dose ranges from the drug's reported effective concentration, if known, to a relatively high concentration of 10 µM or more, often adopted in high-throughput cell based screens (33). In order to include perturbations possibly leading to adversities or able to challenge cells adaptive mechanisms, we decided to focus on drugs tested at the highest concentration and for the longest treatment duration (i.e. high coverage, high dose, and long treatments). Therefore, we focused only on the samples treated at 10 µM dose and at least for 24 hours. Furthermore, as DILI phenotypes are mainly originated and affecting the liver, we decided to study only those sets collected from the cell line "Primary Human Hepatocytes" (PHH), as to date, it is the most specific *in vitro* cellular model for liver. This produced a final set of samples with a single dose-time point from 51, 87, and 37 drugs annotated as No-DILI-Concern, Less-DILI-Concern, and Most-DILI-Concern, respectively.

As an initial exploratory analysis of the training data set, we analyzed the transcriptional response of the different drugs using k-nearest neighbor clustering algorithm (k=3,4,5) (**Supplementary Figure 2**). In the plot, we cannot distinguish the different groups of drugs based uniquely on gene expression and thus a more specific gene signature is needed. Indeed, we applied the landmark genes signature as a feature for a machine learning algorithm (as described in the **Methods** section) obtaining a mean accuracy of 52% in the testing set and 43% in the independent hold-out test set (**Figure 1**). Perhaps more relevant are the low values of MCC (0.04 in the testing set and -0.09 in the independent hold-out test set), which indicates that the level of expression of landmark genes (978) from CMap is not a predictor of DILI. In view of these results, we decided to look for alternative chemical structure, gene and phenotype based signatures. In the following sections, we explain the different strategies we developed to characterize DILI (**Figure 2**).



Testing set						Independent hold-out test set						
Accuracy	Precision	Sensitivity	Specificity	F1-score	MCC	Accuracy	Precision	Sensitivity	Specificity	F1-score	MCC	
0.52	0.52	0.57	0.47	0.54	0.04	0.43	0.31	0.53	0.38	0.39	-0.09	Landmark
0.69	0.67	0.78	0.61	0.72	0.39	0.47	0.33	0.53	0.44	0.41	-0.03	DisGeNET
0.48	0.47	0.41	0.54	0.44	-0.05	0.55	0.39	0.53	0.56	0.45	0.09	GUILDify
0.69	0.71	0.67	0.71	0.69	0.39	0.55	0.38	0.47	0.59	0.42	0.06	DILI landmark
0.57	0.61	0.40	0.73	0.47	0.15	0.70	0.59	0.59	0.77	0.59	0.35	Targets
0.65	0.68	0.58	0.73	0.62	0.31	0.55	0.41	0.71	0.47	0.52	0.17	SMILES
0.68	0.65	0.75	0.60	0.70	0.36	0.47	0.36	0.71	0.34	0.48	0.05	DisGeNET + SMILES
0.49	0.49	0.67	0.30	0.57	-0.02	0.37	0.34	0.88	0.09	0.49	-0.04	GUILDify + SMILES
0.63	0.65	0.54	0.72	0.59	0.26	0.67	0.52	0.82	0.59	0.64	0.40	DILI landmark + SMILES

Fig. 1 Results of the Classifiers in the testing set and the independent hold-out test set. The machine learning algorithm used was a Random Forest. The features that were used in the models of DisGeNET, GUILDify, DisGeNET+SMILES and GUILDify+SMILES are only from the phenotype "Biliary cirrhosis" (C0023892). The results of using different phenotypes are given in the Fig. 3. The results for gradient boosting machine classifier are given in the Supplementary Fig. 8.

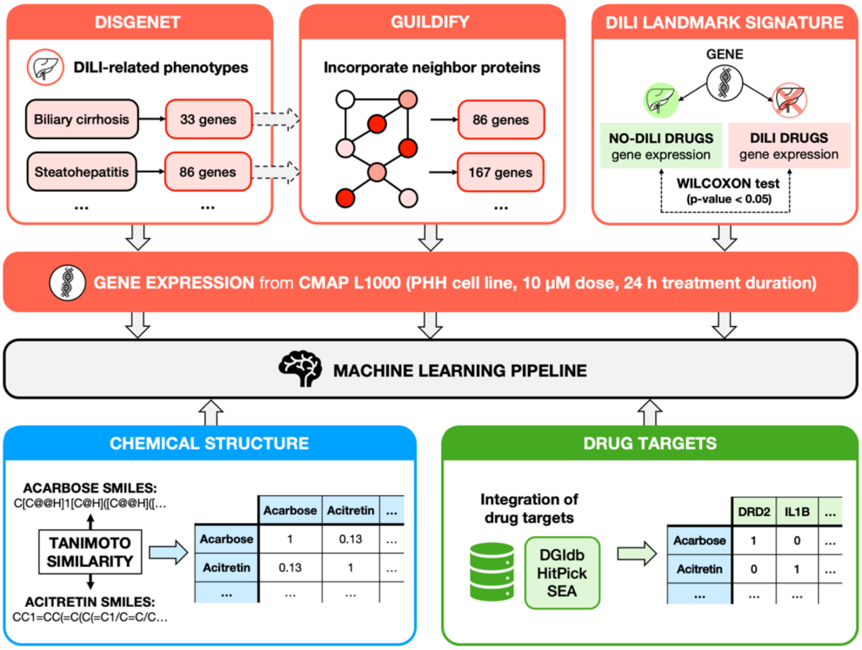
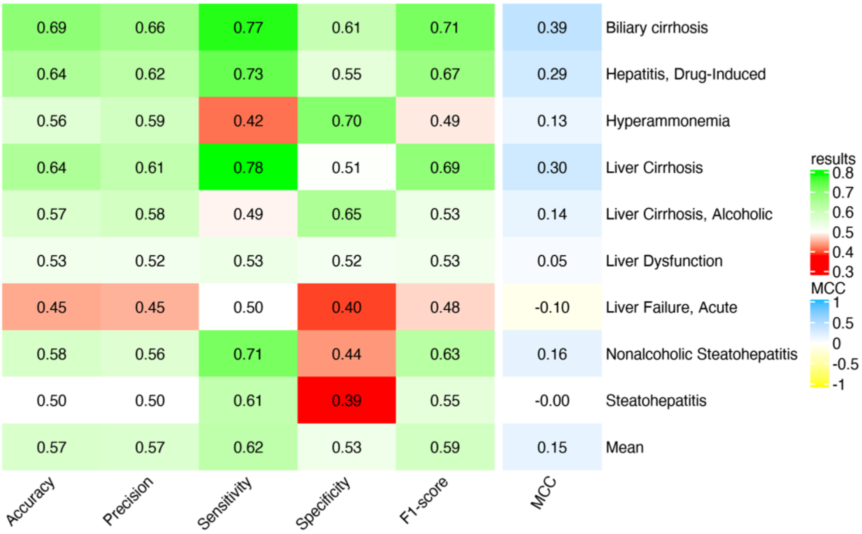

DISGENET
DILI-related phenotypes
Biliary cirrhosis
33 genes
Steatohepatitis
86 genes
...
...
GUILDIFY
Incorporate neighbor proteins
86 genes
167 genes
...
DILI LANDMARK SIGNATURE
GENE
NO-DILI DRUGS gene expression
DILI DRUGS gene expression
WILCOXON test (p-value < 0.05)
GENE EXPRESSION from CMAP L1000 (PHH cell line, 10 µM dose, 24 h treatment duration)
MACHINE LEARNING PIPELINE
CHEMICAL STRUCTURE
ACARBOSE SMILES: C[C@@H]1[C@H]([C@@H]([...
TANIMOTO SIMILARITY
ACITRETIN SMILES: CC1=CC(=C(C(=C1/C=C/C...
Acarbose
Acitretin
...
1
0.13
0.13
1
DRUG TARGETS
Integration of drug targets
DGIdb
HitPick
SEA
DRD2
IL1B
0
1

Fig. 2 General scheme of the processing of the different features.

2. Using phenotype-gene associations highlights potential connections between DILI, cirrhosis and drug induced hepatitis

To characterize genes involved in DILI that could be used as a gene signature in the classifier, we searched for specific genes associated with DILI looking into phenotype-genotype data. These data contain genes that have been described as associated to the pathophysiology or etiology of DILI, and therefore represent a suitable source to develop a list of genes representative of DILI. We manually curated a list of phenotypes closely related with DILI and identified the genes associated with these phenotypes using the DisGeNET database v6.0 (17) (**Table 2, Supplementary Table 1**). Although we might expect them to be genetically similar, the overlap of genes between the different DILI phenotypes is very small (**Supplementary Figure 3**). This fact reflects the diversity of the phenotypes considered and the challenge associated to predict DILI based solely on gene expression.

Once defined the set of genes for the different DILI-related phenotypes as annotated in DisGeNET, we retrieved their gene expression data from the CMap L1000 Assay Platform. For each DILI-related phenotype, we trained an independent machine learning model using the expression levels of their genes as features. The average accuracy obtained for the models of all DILI-related phenotypes is 57% in the testing set. This means that for some specific phenotypes the accuracy was higher than 57%. Therefore, we inspected the results for all phenotypes separately, observing those with higher accuracy than others (**Figure 3**). The phenotypes “Biliary cirrhosis”, “Hepatitis, Drug-Induced” and “Liver cirrhosis” stand out for having an accuracy between 64% and 69% and values of precision, sensitivity and specificity above 50%, and MCC above 0.3. It is worth noting that “Biliary cirrhosis” is the phenotype less genetically similar to the rest, i.e. the lowest number of shared genes (**Supplementary Figure 3**) yielding the best results of prediction. Among the genes associated with these phenotypes, some of them have been associated to hepatotoxicity by a previous study of *Peng et al. (2019)* (34), where 145 hepatotoxicity-related genes were identified. “Biliary cirrhosis” contains 5 hepatotoxicity-associated genes, “Hepatitis, Drug-Induced” has 27 and “Liver cirrhosis” 25 (**Supplementary Table 6**).



Accuracy	Precision	Sensitivity	Specificity	F1-score	MCC	
0.69	0.66	0.77	0.61	0.71	0.39	Biliary cirrhosis
0.64	0.62	0.73	0.55	0.67	0.29	Hepatitis, Drug-Induced
0.56	0.59	0.42	0.70	0.49	0.13	Hyperammonemia
0.64	0.61	0.78	0.51	0.69	0.30	Liver Cirrhosis
0.57	0.58	0.49	0.65	0.53	0.14	Liver Cirrhosis, Alcoholic
0.53	0.52	0.53	0.52	0.53	0.05	Liver Dysfunction
0.45	0.45	0.50	0.40	0.48	-0.10	Liver Failure, Acute
0.58	0.56	0.71	0.44	0.63	0.16	Nonalcoholic Steatohepatitis
0.50	0.50	0.61	0.39	0.55	-0.00	Steatohepatitis
0.57	0.57	0.62	0.53	0.59	0.15	Mean

Fig. 3 Results of the classifier based on gene sets from DisGeNET DILI phenotypes in the testing set. The machine learning algorithm used was Random Forest. Each row corresponds to the mean performance of 10 models trained using the PHH gene expression of the genes associated to each DILI phenotype. The "Mean" row corresponds to the average performance of each metric for all the phenotypes.

3. Incorporating protein-protein interactions to find a DILI signature does not improve the results of phenotype-gene associations

Our current knowledge of genotype-phenotype associations is still incomplete and therefore we might miss relevant genes associated to DILI. It has been demonstrated that the products of disease-associated genes tend to be highly connected in the protein-protein interaction network, forming the so-called disease modules (35,36). Based on this fact, network-based prioritization methods exploiting the topology of the protein-protein interactions network have been successfully applied to discover and prioritize novel disease-gene associations (37).

Using the network-based prioritization web server GUILDify (38), we extended the current knowledge of disease-associated genes obtained from DisGeNET (see above in the previous section). GUILDify uses the genes associated with DILI-related phenotypes as seeds for an algorithm that scores the proteins of the protein-protein interaction network based on their topological closeness with the seeds. Then, it selects the top-ranking genes using a functional-coherency-based cut-off: non-seed genes are iteratively included in the top-ranking set provided that they maintain the functional coherency of the seed genes (they are involved in similar biological functions). The numbers of the new associations with the DILI-related phenotypes are listed in **Table 2**.

After obtaining the new list of phenotype-gene associations, we retrieved their gene expression data from the CMap L1000 Assay Platform as shown before (i.e. PHH cell line with 10 µM dose and treatment duration of 24 hours) and used the expression level of these genes as input feature to the machine learning classifier. As shown in **Figure 1** the predictive capacity of the classifiers in the training set dropped with regards to the approach described in **Section 2**, obtaining similar values to that of when using the 978 landmark genes albeit with a slightly higher specificity (see results by phenotype at **Supplementary Figure 4**).

4. Differential comparison of gene expression does not produce a robust DILI signature

To investigate the extend the transcriptomics data on drugs with known DILI status could be used to extract a DILI gene signature, we retrieved the normalized differential expression data of the genes in PHH cell line (10 µM dose and treatment duration of 24 hours). For each gene, we checked whether the expression values were significantly different between DILI and No-DILI-Concern drugs. Therefore, for each landmark gene, we applied a two-sided Wilcoxon test, a non-parametric test comparing the expression of the gene in the samples of DILI-Concern drugs and No-DILI-Concern drugs. We selected the genes with a P-value lower than 0.05, obtaining a gene signature composed of 66 genes (referred from now on as DILI landmark gene signature) (see **Supplementary Table 7**). We chose to use marginal P-values, focusing on the ranking of genes and aiming to capture the broad transcriptomic DILI signal.

Consistent with the known heterogeneity of transcriptomics response in hepatotoxicity, the genes in the identified signature were typically perturbed only in a small subset of the samples, failing to represent a common response that could be explained by gene expression changes (**Figure 4**). However, while the gene expression of the 1000 landmark genes yielded an accuracy of 52% (43% in the independent hold-out test set), using only the 66 selected genes increased the accuracy to a 69% (55% in the independent hold-out test set). The discrepancy between the testing and independent hold-out test sets can be attributed to the gene expression signature likely fitting to the underlying biology of the training set compounds rather than representing a generalization across all potential DILI compounds. We performed a functional enrichment analysis (39) using Gene Ontology to further investigate the biological processes of these genes.

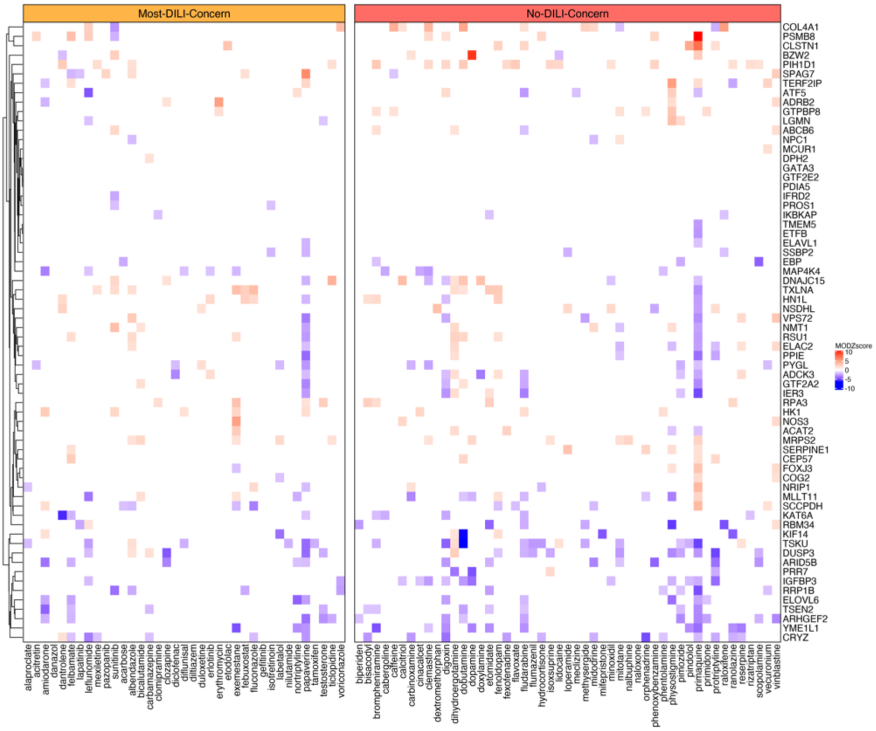


Fig. 4 Transcriptomics signatures of the DILI landmark genes. Gene expression (as Moderated Z score) of the DILI landmark genes (selected using a two-sided Wilcoxon rank sum test, *P* < 0.05) in PHH cells, for Most-DILI-Concern and No-DILI-Concern drugs. The gene expression lower than |1.5| is colored white.

The functional enrichment analysis of the 66 genes did not yield specific functions significantly associated with the genes. This indicates that, even though the selected genes improve the capacity of the classifiers to predict DILI-causing drugs in comparison with using all the landmark genes, they are not related with specific biological processes. Additionally, we compared these 66 genes to the 145 identified by *Peng et al. (2019)* (34) as associated to hepatotoxicity (**Supplementary Table 6**). Only 4 genes were highlighted as hepatotoxic in *Peng et al*'s study. Therefore, although the approach succeeded at improving the predictive capacity of the classifiers, the results could lead to overfitting by the available data, as the gene signature is not related to any specific biological function.

5. The use of chemical structure and drug-target associations increases the prediction accuracy

Besides using gene expression data, we investigated the incorporation of orthogonal information derived from chemical structure of the drug and its targets. The chemical structure and the molecular descriptors of the drugs have already been used in machine learning models, showing a fair predictive capacity (10). Here, we used the Tanimoto similarity between the molecular fingerprints of all the drugs in the dataset. First, we plotted the similarity between Most-DILI-Concern and No-DILI-Concern drugs in the dataset in a histogram (**Supplementary Figure 5**). We observed that drugs of the same group did not have higher similarity among them than with other groups. Furthermore, No-DILI-Concern drugs have higher similarity between themselves (mean 0.24) than Most-DILI-Concern drugs (mean 0.18). This indicates that there is a considerable structure heterogeneity within the Most-DILI-Concern group of drugs, which complicates the prediction using chemical structure. But this also suggests that probably, when combining this feature with other types of features (i.e. transcriptomics), the prediction may improve.

Eventually, we obtained a higher prediction accuracy when combining chemical structure and CMap than when applying them separately (**Figure 1**). When using solely chemical structure as feature for the machine learning classifiers the average accuracy was 65% (MCC of 0.31) in the test set and 55% (MCC of 0.17) in the independent hold-out test set. In contrast, when combining chemical structure with transcriptomics data, the prediction of the classifier in the independent dataset improved. This is remarkable when combining it with the DILI landmark gene signature derived from the nonparametric test: the average accuracy is maintained at 63% (MCC of 0.26) in the test set and increases to 67% (MCC of 0.40) in the independent hold-out test set (**Figure 1**).

Next, we explored the use of drug-target associations as a feature to predict DILI. We considered any drug-protein pair that had been reported in DGIdb (26) or predicted to interact by HitPick (27) or SEA (28). We integrated targets from these databases, creating a matrix containing drugs and target proteins (see **Methods**). We analyzed the percentage of DILI drugs, no-DILI drugs and drugs from the independent dataset associated to the targets in the matrix (see **Supplementary Table 8** and **Supplementary Figure 6**). We observed that some proteins are mostly targeted by one type of drug, hypothetically facilitating the classification of drugs. For instance, proteins such as CYP2C9 and CYP1A2, that are associated with a higher proportion of DILI drugs than to no-DILI drugs, have been previously associated to hepatotoxic effects (40,41). Thus, we used the matrix as a feature for the machine learning classifiers, obtaining an accuracy of 57% (MCC of 0.15) in the testing set and 70% (MCC of 0.35) in the independent hold-out test set (**Figure 1**). The increase of accuracy in the independent dataset is explained by the high specificity (i.e. no-DILI drugs are predicted correctly) in contrast with the low sensitivity (i.e. DILI drugs are not predicted correctly).

6. Hepatocyte cell lines provide a better context for DILI prediction than using combined expression from different cell lines

In the previous sections, when using CMap gene expression data, we selected only the samples from the PHH cell line with 10 μM dose and treatment duration of 24 hours. We focused on the drug response in liver cells. However, the data of CMap tends to have a high variation of expression between samples even for the same gene. Therefore, to avoid biases caused by the use of unrelated samples, we experimented using only the top correlated samples for each drug. This consists in computing the correlation between all the samples exposed to a drug (even if they are from different cell lines, doses and treatment durations) and selecting the ones that are more correlated between themselves. We selected the pairs of samples from different cell lines that have a Pearson correlation above 0.5, or otherwise we kept the pair that was more correlated. To use a correlation threshold of 0.5 guarantees that the expression of the samples selected is consistent enough across several cell lines. Once the correlated samples are selected, we use the median gene expression as feature. Although the approach was theoretically promising, the prediction accuracies with the use of correlated samples are generally worse than using specific conditions, obtaining MCC values ranging from -0.12 to 0.21 (see **Supplementary Figure 7**). This indicates that we are still getting noise from correlated samples and that, even if there are some samples that could be less reliable, the use of specific liver conditions in gene expression seems to be the best approach for the prediction of DILI-Concern drugs.

Discussion

In this work, we aim to predict DILI applying machine learning algorithms using a range of orthogonal types of data as input features. Indeed, we explored the use of gene expression data from different sets of selected genes (i.e. landmark, DisGeNET and GUILDify sets) alone and in combination with drug-centric information in the form of structural similarity (Tanimoto scores) and protein targets (see **Table 1** for a brief description of the features). Furthermore, we observed that the DILI landmark gene signature identified by a non-parametric test (Wilcoxon test) of differential expression between DILI and no-DILI samples from PHH cell line constituted a better feature set than the whole landmark genes in CMap.

The genes in the identified DILI landmark gene signature were typically perturbed only in a small subset of the samples, failing to represent a response that could solely be explained by gene expression changes (**Figure 4**). This finding is consistent with the known heterogeneity of transcriptomics response in hepatotoxicity (42). Also, it could be related with the diversity of outcomes of the different compounds (i.e. acute, chronic or idiosyncratic reactions). Nevertheless, as we were using data from the training set to obtain the signature, the results could lead to overfitting, which would explain why the accuracy of the prediction in the independent hold-out test worsened. Moreover, the drugs of the independent hold-out test set were originally flagged as ambiguous and for this reason are probably a more challenging set to classify. Also, the independent hold-out test could be unbalanced, worsening the results despite the classifier being trained on a balanced dataset.

We also took advantage of functional information of the genes involved in drug response, and evaluated gene expression related to liver phenotypes involved in drug response using DisGeNET resource (17). In the same way as before, limiting the number of genes to a specific signature (the genes associated to a DisGeNET phenotype) also constituted a better feature set than the whole landmark genes in CMap, but still failed to represent the whole response. The best accuracy was achieved by the phenotype “Biliary Cirrhosis”, which is one of the final stages of DILI. Since we used the data from the highest dose and time point, it makes sense that extreme phenotypes related with liver cirrhosis are better predictors. Perhaps, the biliary component is also important for the predictor. For further studies, it would be interesting to focus on lower doses in order to capture earlier events and not the final extreme phenotype. It is also important to remark that the gene expression signatures come from an *in vitro* model (primary cells, but still with the limitations of 2D, dedifferentiation, etc.). As we applied gene signatures derived from human data, this could have affected the results.

Additionally, we expanded the phenotype-gene associations retrieved from DisGeNET incorporating protein-protein interactions data from GUILDify. By applying GUILDify we could expand the number of genes associated with DILI-related phenotypes by incorporating those connected by the underlying protein interactome. Surprisingly, the quality of the prediction when adding protein interactions decreased with respect to using solely phenotype-gene associations. Our hypothesis is that when expanding the number of genes using GUILDify, (i.e. obtaining larger gene signatures), the intrinsic data noise from the CMap dataset is increased as well, hence hampering the prediction. Still, we think that using protein-protein interactions to extend our information on DILI targets and hepatotoxicity-associated genes without using gene expression data could be an interesting feature to explore in the future.

After working with transcriptomics data from CMap, we observed variability of the results depending on the pre-processing of the samples. We tried two different strategies that led to different results: (i) focusing on samples from a unique cell line and dose-time point for each drug, and (ii) selecting the most correlated samples for each drug. This is by no means comprehensive and various possible strategies such as using other cell lines and dose-time points, or discarding the samples with low correlation between replicas (‘distil_cc_q75’ < 0.2) and selecting the sample with highest transcriptional activity score (43) could be investigated further.

When focusing on samples from a unique cell line and dose-time point, we decided to use the highest concentration and the longest treatment duration. In this way, we were including perturbations possibly leading to adversities for the adaptive mechanisms of the cells. We acknowledge that focusing on increased exposure of the drug to characterize DILI is a relatively strong assumption as there could be certain compensatory mechanisms kicking in after a while depending on the specific compound and cell line. Nevertheless, we think that employing the highest dose at the longest time of exposure is likely to be a fair representation of the effect of DILI in the cells after the administration of the drug.

Apart from the limitations inherent in the CMap dataset, we detected: (i) an important genetic diversity between the diverse DILI-related phenotypes from DisGeNET (**Supplementary Figure 3**), and (ii) a great structural diversity between the drugs reported as DILI-Concern (**Supplementary Figure 5**). Both aspects hamper the prediction of DILI-Concern drugs when using transcriptomics or structural features separately and encouraged us to use and combine other sources of information.

When considering both transcriptomics and structural features together, we observed a similar predictive power of the classifiers, but a general increase when validating the classifiers with an independent dataset (**Figure 1**). The most accurate classifier was generated by the Random Forest algorithm using a combination of features that included the chemical similarity of drugs (Tanimoto coefficient calculated using SMILES) and gene expression from the landmark genes selected with a non-parametric test (DILI Landmark + SMILES). Under a benchmark scenario, the classifier was able to separate DILI-Concern drugs better than No-DILI-Concern drugs (accuracy 63%, sensitivity 54% and specificity 72%). Furthermore, on the independent dataset of ambiguous-DILI drugs re-labelled by the FDA, it reached an accuracy of 67%, the second highest among the different classifiers. In the future, it would be interesting to use the drug structures directly as features (without using their similarities) and to combine them with the other types of features, as there might be critical information within the actual molecular details of the drugs.

Lastly, we explored if the use of drug-target associations could be useful to predict DILI-causing drugs. The results showed that the targets of most DILI drugs were related with hepatoxicity (**Supplementary Figure 6**). The use of drug-target associations as a feature produced an accuracy of 57% in the testing set and 70% in the independent dataset. The observed accuracy on the independent dataset is in line with 72.5% sensitivity and 72.7% specificity of the computational model developed by *Zhang et al.* as well as with the 70.9% accuracy obtained by *Hong et al.* on the bootstrapped data set, highlighting the current limitations in predicting drug induced injury (8,9).

When comparing the results with the publications of the previous CAMDA 2018 edition (12,13), we still do not observe a clear improvement on the prediction of DILI. Although the data provided is much more extensive, including gene expression data from more cell lines, the gold standard is still very reduced and unbalanced. The results in terms of accuracy in the training set are very similar to the ones obtained by *Sumsion et al*. (12), but worse when looking at the independent hold-out test. This is probably due to the fact that the current independent dataset is based on “Ambiguous-DILI” drugs, making the task more challenging. In terms of MCC values, our results (ranging from −0.05 to 0.39 in cross-validation and −0.09 to 0.40 in independent hold-out test) are slightly better than the ones reported in *Chierici et al.* (13) (ranging from −0.04 to 0.21 in cross-validation and −0.16 to 0.11 in the independent hold-out test set). Still, while the two published approaches of the previous edition were more focused on testing and optimizing different types of machine learning classifiers, our study focused on evaluating different types of features and searching a specific DILI gene signature. Therefore, the point of view of our work has been very different and complement previous approaches.

Overall, our results pointed to a mild variation on the accuracies depending on the samples included in the training data as well as the feature set used in building the classifiers, which we attribute to various factors. First, the training data is limited to dozens of compounds with known hepatotoxicity annotation, and these are too few to get a robust classifier. Second, most compounds show a toxic effect based on the dosage (and are otherwise no-DILI), thus a global predictor categorizing drugs as simply DILI vs no-DILI might not be realistic. And, third, there is substantial heterogeneity in the transcriptomics data from CMap. There is also variation between the results of the testing set and the independent hold-out test set, that could be caused by the latter being unbalanced (as the labels remain hidden). Still, the variation between the machine learning algorithms (random forest vs gradient boosting machine) is not appreciable in most cases (see results for gradient boosting machine in **Supplementary Figures 8-9**). This indicates that even though the classifiers are different, the results are consistent because they depend on the data rather than on the algorithms. Still, future work would be required to experimentally validate the predictions of these models.

Conclusions

In this study, we developed an ensemble learning approach to investigate the mechanism of the drugs that cause DILI. We experimented with gene expression data from the CMap L1000 dataset both alone and in combination with other types of feature (chemical structure, drug targets). We observed that selecting a specific gene signature either using phenotype-gene associations data (DisGeNET) or a non-parametric test (Wilcoxon test) of differential expression between DILI and no-DILI samples constituted a better feature than the whole landmark genes in CMap. However, the accuracy of the best performing classifier is around the 70% mark (minimum 63%, maximum 76%), stating the limitations of predicting DILI. The results are very similar to previous publications (8–10,12). Additionally, we used the comparison of chemical structures as a feature to predict DILI-causing drugs, though this did not improve the accuracy substantially. When comparing the chemical structures of the drugs with the same DILI-Concern classification, we observed a large structural diversity among the DILI-Concern groups, reflected in their dissimilarity of structure. This may explain the limited accuracy prediction based on chemical structure. Combining transcriptomics data and chemical structure did not improve the accuracy of the prediction in the testing set, although this was improved in the independent hold-out test set. Specifically, the combination of using a DILI associated gene signature and chemical structures produced results of accuracy around or less than 70%, but more robust when they were validated with the independent hold-out test set. We also used drug-target associations as feature, obtaining 57% of accuracy in the testing set that improved to a 70% in the independent hold-out test set.

Summarizing, the overarching goal of this work was to evaluate a range of descriptors to predict DILI employing two commonly used classifiers to predict DILI. We have shown the limitations and advantages of different sets of data paving the way for future research in this field.

List of abbreviations

DILI: Drug-induced liver injury.
CAMDA: Critical Assessment of Massive Data Analysis.
CMap: Connectivity Map.
SMILES: Simplified molecular-input line-entry system.
PHH: Primary Human Hepatocytes.
RF: Random Forest.

References

1. Kola I, Landis J. Can the pharmaceutical industry reduce attrition rates? Nat Rev Drug Discov. 2004;3(8):711–5.
2. Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J. Clinical development success rates for investigational drugs. Nat Biotechnol. 2014 Jan;32(1):40–51.
3. Parasrampuria DA, Benet LZ, Sharma A. Why Drugs Fail in Late Stages of Development: Case Study Analyses from the Last Decade and Recommendations. AAPS J. 2018 13;20(3):46.
4. Kullak-Ublick GA, Andrade RJ, Merz M, End P, Benesic A, Gerbes AL, et al. Drug-induced liver injury: recent advances in diagnosis and risk assessment. Gut. 2017;66(6):1154–64.

5. Suk KT, Kim DJ. Drug-induced liver injury: present and future. Clin Mol Hepatol. 2012 Sep;18(3):249–57.

6. Sobhonslidsuk A, Poovorawan K, Soonthornworasiri N, Pan-ngum W, Phaosawasdi K. The incidence, presentation, outcomes, risk of mortality and economic data of drug-induced liver injury from a national database in Thailand: a population-base study. BMC Gastroenterol. 2016 Oct 28;16.

7. Thakkar S, Li T, Liu Z, Wu L, Roberts R, Tong W. Drug-induced liver injury severity and toxicity (DILIst): binary classification of 1279 drugs by human hepatotoxicity. Drug Discov Today. 2020 Jan;25(1):201–8.

8. Zhang H, Ding L, Zou Y, Hu S-Q, Huang H-G, Kong W-B, et al. Predicting drug-induced liver injury in human with Naïve Bayes classifier approach. J Comput Aided Mol Des. 2016;30(10):889–98.

9. Hong H, Thakkar S, Chen M, Tong W. Development of Decision Forest Models for Prediction of Drug-Induced Liver Injury in Humans Using A Large Set of FDA-approved Drugs. Sci Rep. 2017 11;7(1):17311.

10. Ai H, Chen W, Zhang L, Huang L, Yin Z, Hu H, et al. Predicting Drug-Induced Liver Injury Using Ensemble Learning Methods and Molecular Fingerprints. Toxicol Sci Off J Soc Toxicol. 2018 01;165(1):100–7.

11. Wang H, Liu R, Schyman P, Wallqvist A. Deep Neural Network Models for Predicting Chemically Induced Liver Toxicity Endpoints From Transcriptomic Responses. Front Pharmacol. 2019;10:42.

12. Sumsion GR, Bradshaw MS, Beales JT, Ford E, Caryotakis GRG, Garrett DJ, et al. Diverse approaches to predicting drug-induced liver injury using gene-expression profiles. Biol Direct. 2020 Jan 15;15(1):1.

13. Chierici M, Francescatto M, Bussola N, Jurman G, Furlanello C. Predictability of drug-induced liver injury by machine learning. Biol Direct [Internet]. 2020 Feb 13 [cited 2020 Oct 29];15. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7020573/

14. Chen M, Suzuki A, Thakkar S, Yu K, Hu C, Tong W. DILIrank: the largest reference drug list ranked by the risk for developing drug-induced liver injury in humans. Drug Discov Today. 2016 Apr;21(4):648–53.
15. Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, et al. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Cell. 2017 Nov 30;171(6):1437-1452.e17.
16. Bray M-A, Gustafsdottir SM, Rohban MH, Singh S, Ljosa V, Sokolnicki KL, et al. A dataset of images and morphological profiles of 30 000 small-molecule treatments using the Cell Painting assay. Gigascience. 2017 Dec;6(12):1–5.
17. Piñero J, Bravo À, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, Centeno E, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res. 2017 04;45(D1):D833–9.
18. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019 Jan 8;47(D1):D506–15.
19. Davis AP, Grondin CJ, Johnson RJ, Sciaky D, McMorran R, Wiegers J, et al. The Comparative Toxicogenomics Database: update 2019. Nucleic Acids Res. 2019 Jan 8;47(D1):D948–54.
20. Pavan S, Rommel K, Mateo Marquina ME, Höhn S, Lanneau V, Rath A. Clinical Practice Guidelines for Rare Diseases: The Orphanet Database. PLoS ONE [Internet]. 2017 Jan 18 [cited 2019 Oct 24];12(1). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5242437/
21. Rehm HL, Berg JS, Brooks LD, Bustamante CD, Evans JP, Landrum MJ, et al. ClinGen--the Clinical Genome Resource. N Engl J Med. 2015 04;372(23):2235–42.
22. Genomics England. Genomics England PanelApp [Internet]. 2019. Available from: https://panelapp.genomicsengland.co.uk
23. Tamborero D, Rubio-Perez C, Deu-Pons J, Schroeder MP, Vivancos A, Rovira A, et al. Cancer Genome Interpreter annotates the

biological and clinical relevance of tumor alterations. Genome Med. 2018 28;10(1):25.

24. Giri V, Sivakumar TV, Cho KM, Kim TY, Bhaduri A. RxnSim: a tool to compare biochemical reactions. Bioinforma Oxf Engl. 2015 Nov 15;31(22):3712–4.
25. Tanimoto TT. An Elementary Mathematical Theory of Classification and Prediction. International Business Machines Corporation; 1958. 10 p.
26. Cotto KC, Wagner AH, Feng Y-Y, Kiwala S, Coffman AC, Spies G, et al. DGIdb 3.0: a redesign and expansion of the drug-gene interaction database. Nucleic Acids Res. 2018 04;46(D1):D1068–73.
27. Hamad S, Adornetto G, Naveja JJ, Chavan Ravindranath A, Raffler J, Campillos M. HitPickV2: a web server to predict targets of chemical compounds. Bioinforma Oxf Engl. 2019 Apr 1;35(7):1239–40.
28. Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK. Relating protein pharmacology by ligand chemistry. Nat Biotechnol. 2007 Feb;25(2):197–206.
29. Kuhn M. Building Predictive Models in R Using the caret Package. J Stat Softw. 2008 Nov 10;28(1):1–26.
30. Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov. 2019;18(6):463–77.
31. Costello JC, Heiser LM, Georgii E, Gönen M, Menden MP, Wang NJ, et al. A community effort to assess and improve drug sensitivity prediction algorithms. Nat Biotechnol. 2014 Dec;32(12):1202–12.
32. Menden MP, Wang D, Mason MJ, Szalai B, Bulusu KC, Guan Y, et al. Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. Nat Commun. 2019 17;10(1):2674.
33. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, et al. The Connectivity Map: using gene-expression signatures to connect

small molecules, genes, and disease. Science. 2006 Sep 29;313(5795):1929–35.

34. Peng Y, Wu Z, Yang H, Cai Y, Liu G, Li W, et al. Insights into mechanisms and severity of drug-induced liver injury via computational systems toxicology approach. Toxicol Lett. 2019 Sep 15;312:22–33.

35. Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabási A-L. The human disease network. Proc Natl Acad Sci U S A. 2007 May 22;104(21):8685–90.

36. Menche J, Sharma A, Kitsak M, Ghiassian S, Vidal M, Loscalzo J, et al. Uncovering disease-disease relationships through the incomplete human interactome. Science. 2015 Feb 20;347(6224):1257601.

37. Guney E, Oliva B. Exploiting Protein-Protein Interaction Networks for Genome-Wide Disease-Gene Prioritization. PLoS ONE. 2012 Sep 21;7(9).

38. Aguirre-Plans J, Piñero J, Sanz F, Furlong LI, Fernandez-Fuentes N, Oliva B, et al. GUILDify v2.0: A Tool to Identify Molecular Networks Underlying Human Diseases, Their Comorbidities and Their Druggable Targets. J Mol Biol. 2019 Jun 14;431(13):2477–84.

39. Berriz GF, Beaver JE, Cenik C, Tasan M, Roth FP. Next generation software for functional trend analysis. Bioinforma Oxf Engl. 2009 Nov 15;25(22):3043–4.

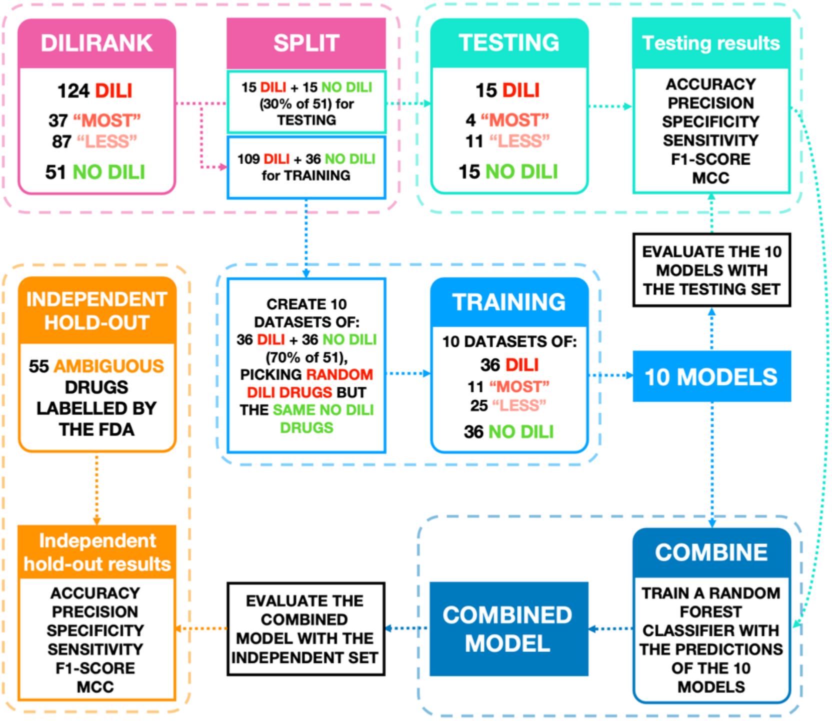
40. Zhao M, Zhang T, Li G, Qiu F, Sun Y, Zhao L. Associations of CYP2C9 and CYP2A6 Polymorphisms with the Concentrations of Valproate and its Hepatotoxin Metabolites and Valproate-Induced Hepatotoxicity. Basic Clin Pharmacol Toxicol. 2017 Aug;121(2):138–43.

41. Casley WL, Menzies JA, Mousseau N, Girard M, Moon TW, Whitehouse LW. Increased basal expression of hepatic Cyp1a1 and Cyp1a2 genes in inbred mice selected for susceptibility to acetaminophen-induced hepatotoxicity. Pharmacogenetics. 1997 Aug;7(4):283–93.

42. Jetten MJA, Kleinjans JCS, Claessen SM, Chesné C, van Delft JHM. Baseline and genotoxic compound induced gene expression profiles in HepG2 and HepaRG compared to primary human hepatocytes. Toxicol Vitro Int J Publ Assoc BIBRA. 2013 Oct;27(7):2031–40.

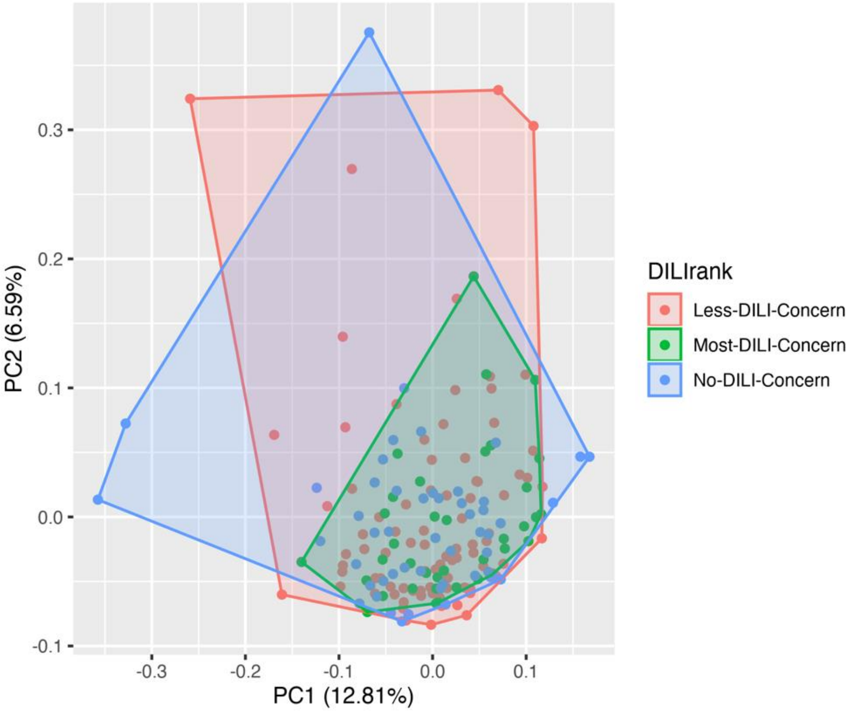
43. Duran-Frigola M, Pauls E, Guitart-Pla O, Bertoni M, Alcalde V, Amat D, et al. Extending the small molecule similarity principle to all levels of biology. bioRxiv [Internet]. 2019 Aug [cited 2020 Feb 15]; Available from: https://www.biorxiv.org/content/10.1101/745703v1

Supplementary Figure 1


DILIRANK
124 DILI
37 "MOST"
87 "LESS"
51 NO DILI
SPLIT
15 DILI + 15 NO DILI (30% of 51) for TESTING
109 DILI + 36 NO DILI for TRAINING
TESTING
15 DILI
4 "MOST"
11 "LESS"
15 NO DILI
Testing results
ACCURACY
PRECISION
SPECIFICITY
SENSITIVITY
F1-SCORE
MCC
EVALUATE THE 10 MODELS WITH THE TESTING SET
INDEPENDENT HOLD-OUT
55 AMBIGUOUS DRUGS LABELLED BY THE FDA
CREATE 10 DATASETS OF: 36 DILI + 36 NO DILI (70% of 51), PICKING RANDOM DILI DRUGS BUT THE SAME NO DILI DRUGS
TRAINING
10 DATASETS OF:
36 DILI
11 "MOST"
25 "LESS"
36 NO DILI
10 MODELS
Independent hold-out results
ACCURACY
PRECISION
SPECIFICITY
SENSITIVITY
F1-SCORE
MCC
EVALUATE THE COMBINED MODEL WITH THE INDEPENDENT SET
COMBINED MODEL
COMBINE
TRAIN A RANDOM FOREST CLASSIFIER WITH THE PREDICTIONS OF THE 10 MODELS

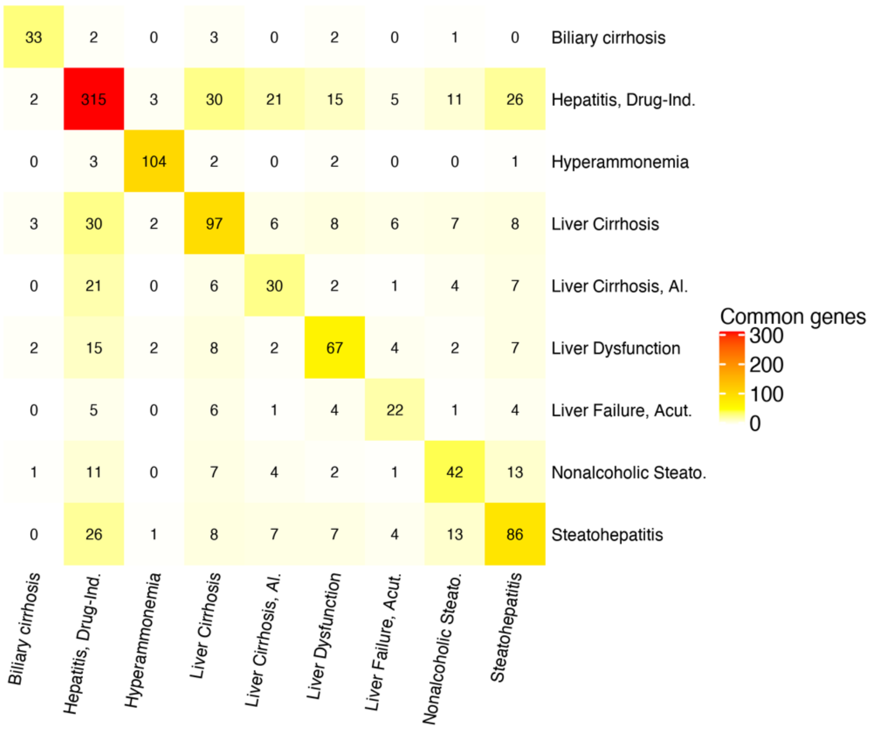
Supplementary Figure 1. Scheme of the machine learning pipeline. The DILIrank dataset is comprised of 124 drugs labelled as DILI (37 as Most-DILI-Concern and 87 as Less-DILI-Concern) and 51 labelled as no DILI. The dataset is randomly split into a balanced testing dataset made of 15 No-DILI-Concern drugs (30% of 51 drugs), and the same number of DILI drugs maintaining the ratio of Most-DILI-Concern (29.8%) and Less-DILI-Concern (70.2%): 4 Most-DILI-Concern drugs (the 29.8% of 15) and 11 Less-DILI-Concern drugs (the 70.2% of 15). The rest of the drugs (109 DILI-Concern drugs and 36 No-DILI-Concern drugs) is used to create 10 different balanced training datasets. For the 10 training datasets, we select the same 36 No-DILI-Concern drugs, but we pick randomly 36 drugs from the 109 DILI-Concern drugs: 11 Most-DILI-Concern drugs (29.8% of 36) and 25 Less-DILI-Concern (70.2% of 36). Using the 10 training datasets, we build 10 different models that are evaluated using the same testing dataset. The predictions of the 10 models are combined into a final model using a random forest algorithm. The final model is evaluated using the independent hold-out test dataset, comprising 55 drugs with hidden labels.

Supplementary Figure 2



Supplementary Figure 2. Low dimensional representation of the gene expression of the training set compounds based on their transcriptomics profiles across samples of primary human hepatocyte (PHH) cell line and their DILI category.

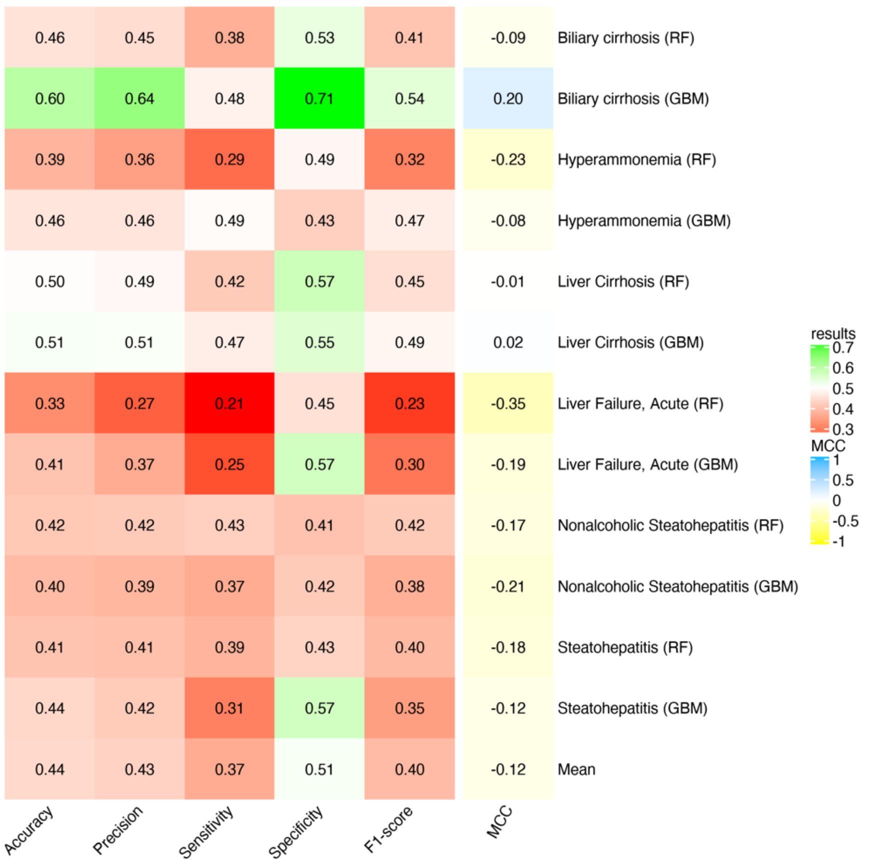
Supplementary Figure 3



	Biliary cirrhosis	Hepatitis, Drug-Ind.	Hyperammonemia	Liver Cirrhosis	Liver Cirrhosis, Al.	Liver Dysfunction	Liver Failure, Acut.	Nonalcoholic Steato.	Steatohepatitis
Biliary cirrhosis	33	2	0	3	0	2	0	1	0
Hepatitis, Drug-Ind.	2	315	3	30	21	15	5	11	26
Hyperammonemia	0	3	104	2	0	2	0	0	1
Liver Cirrhosis	3	30	2	97	6	8	6	7	8
Liver Cirrhosis, Al.	0	21	0	6	30	2	1	4	7
Liver Dysfunction	2	15	2	8	2	67	4	2	7
Liver Failure, Acut.	0	5	0	6	1	4	22	1	4
Nonalcoholic Steato.	1	11	0	7	4	2	1	42	13
Steatohepatitis	0	26	1	8	7	7	4	13	86

Supplementary Figure 3. Number of common genes between the Drug-Induced Liver Injury (DILI) phenotypes retrieved from DisGeNET.

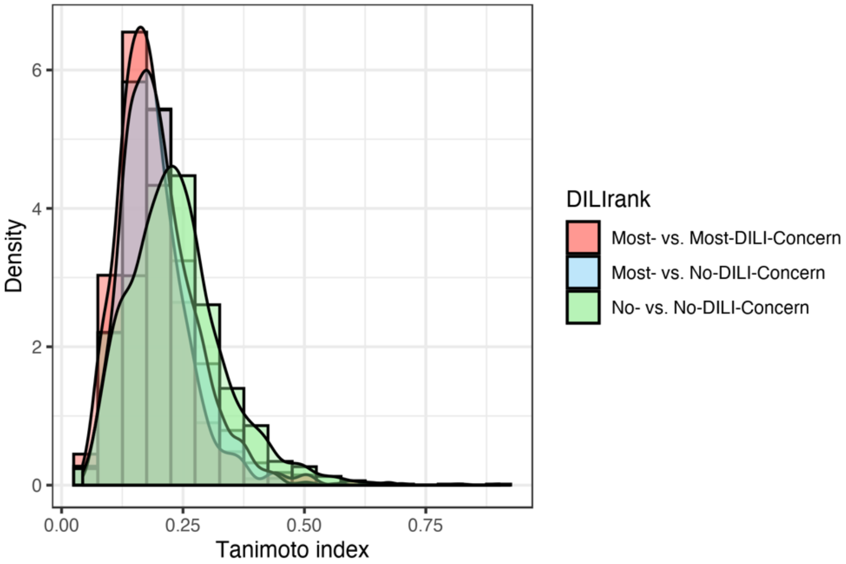
Supplementary Figure 4



	Accuracy	Precision	Sensitivity	Specificity	F1-score	MCC
Biliary cirrhosis (RF)	0.46	0.45	0.38	0.53	0.41	-0.09
Biliary cirrhosis (GBM)	0.60	0.64	0.48	0.71	0.54	0.20
Hyperammonemia (RF)	0.39	0.36	0.29	0.49	0.32	-0.23
Hyperammonemia (GBM)	0.46	0.46	0.49	0.43	0.47	-0.08
Liver Cirrhosis (RF)	0.50	0.49	0.42	0.57	0.45	-0.01
Liver Cirrhosis (GBM)	0.51	0.51	0.47	0.55	0.49	0.02
Liver Failure, Acute (RF)	0.33	0.27	0.21	0.45	0.23	-0.35
Liver Failure, Acute (GBM)	0.41	0.37	0.25	0.57	0.30	-0.19
Nonalcoholic Steatohepatitis (RF)	0.42	0.42	0.43	0.41	0.42	-0.17
Nonalcoholic Steatohepatitis (GBM)	0.40	0.39	0.37	0.42	0.38	-0.21
Steatohepatitis (RF)	0.41	0.41	0.39	0.43	0.40	-0.18
Steatohepatitis (GBM)	0.44	0.42	0.31	0.57	0.35	-0.12
Mean	0.44	0.43	0.37	0.51	0.40	-0.12

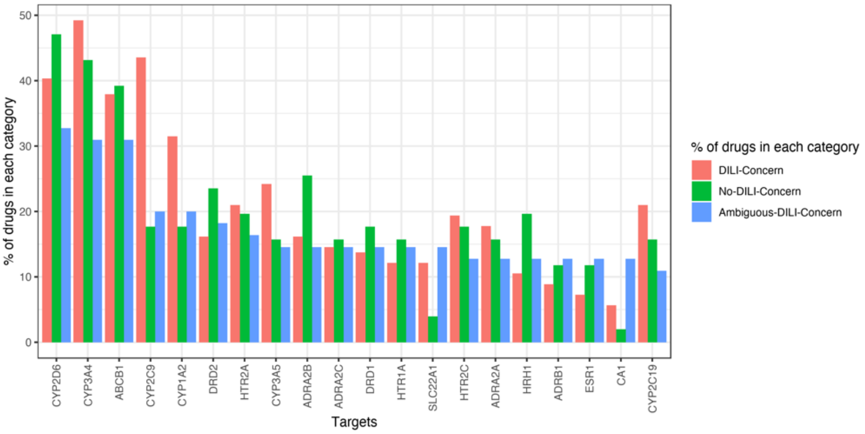
Supplementary Figure 4. Result of the classifiers based on gene sets from GUILDify DILI phenotypes in the testing set. The machine learning algorithms used are either Random Forest (RF) or Gradient Boosting Machine (GBM). Each row corresponds to the mean performance of 10 models trained using the PHH gene expression of the genes associated to each DILI phenotype. The “Mean” row corresponds to the average performance of each metric for all the phenotypes.

Supplementary Figure 5



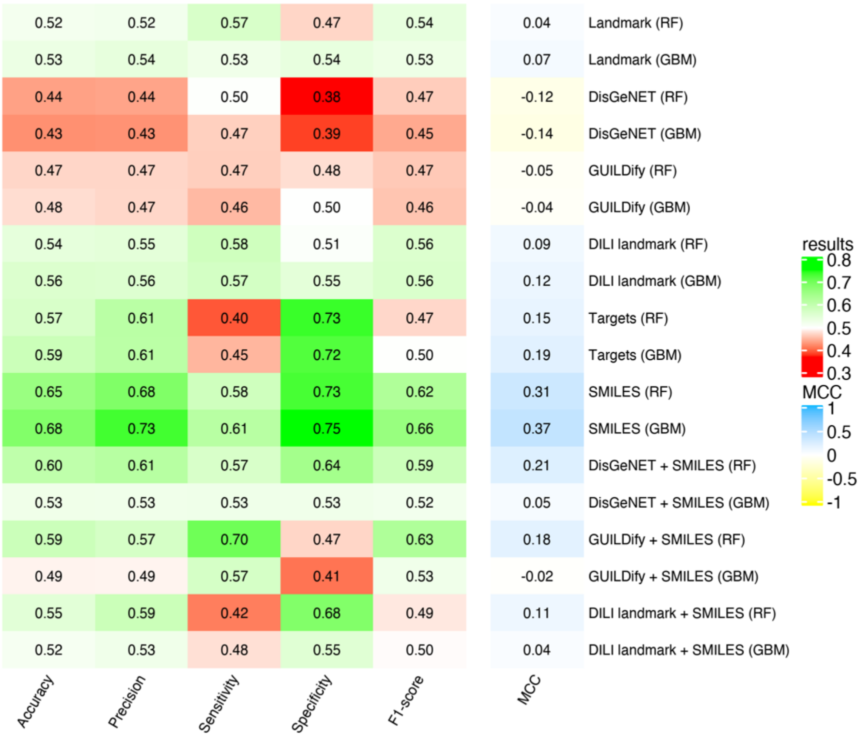
Supplementary Figure 5. Tanimoto similarity between the drugs in the DILI severity categories "Most-DILI-Concern" (Most-) and "No-DILI-Concern" (No-).

Supplementary Figure 6



Supplementary Figure 6. Percentage of drugs in each DILIrank category that interact with a selection of 20 target proteins (Supplementary Table 8). Proteins in Supplementary Table 8 were selected as those targeted by the largest number of drugs in the independent hold-out test dataset.

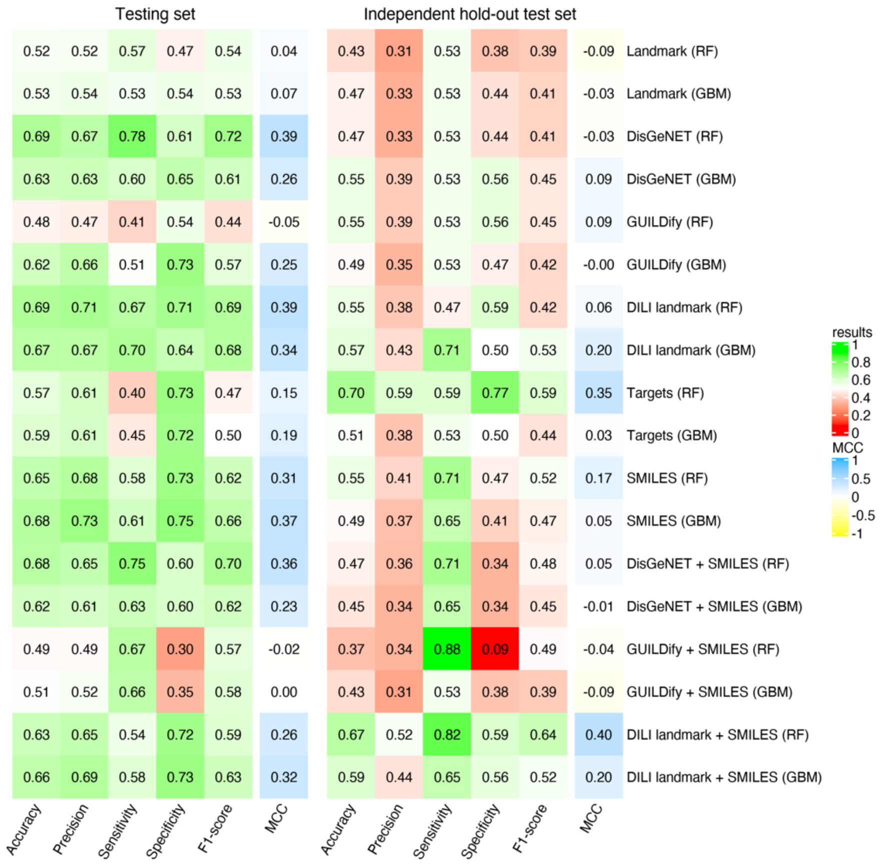
Supplementary Figure 7



	Accuracy	Precision	Sensitivity	Specificity	F1-score	MCC
Landmark (RF)	0.52	0.52	0.57	0.47	0.54	0.04
Landmark (GBM)	0.53	0.54	0.53	0.54	0.53	0.07
DisGeNET (RF)	0.44	0.44	0.50	0.38	0.47	-0.12
DisGeNET (GBM)	0.43	0.43	0.47	0.39	0.45	-0.14
GUILDify (RF)	0.47	0.47	0.47	0.48	0.47	-0.05
GUILDify (GBM)	0.48	0.47	0.46	0.50	0.46	-0.04
DILI landmark (RF)	0.54	0.55	0.58	0.51	0.56	0.09
DILI landmark (GBM)	0.56	0.56	0.57	0.55	0.56	0.12
Targets (RF)	0.57	0.61	0.40	0.73	0.47	0.15
Targets (GBM)	0.59	0.61	0.45	0.72	0.50	0.19
SMILES (RF)	0.65	0.68	0.58	0.73	0.62	0.31
SMILES (GBM)	0.68	0.73	0.61	0.75	0.66	0.37
DisGeNET + SMILES (RF)	0.60	0.61	0.57	0.64	0.59	0.21
DisGeNET + SMILES (GBM)	0.53	0.53	0.53	0.53	0.52	0.05
GUILDify + SMILES (RF)	0.59	0.57	0.70	0.47	0.63	0.18
GUILDify + SMILES (GBM)	0.49	0.49	0.57	0.41	0.53	-0.02
DILI landmark + SMILES (RF)	0.55	0.59	0.42	0.68	0.49	0.11
DILI landmark + SMILES (GBM)	0.52	0.53	0.48	0.55	0.50	0.04

Supplementary Figure 7. Results of the Classifiers in the testing set when using transcriptomics features from the most correlated samples of each drug. The machine learning algorithms used are either Random Forest (RF) or Gradient Boosting Machine (GBM). Results of DisGeNET, GUILDify, DisGeNET+SMILES and GUILDify+SMILES are the mean of all the phenotypes' results.

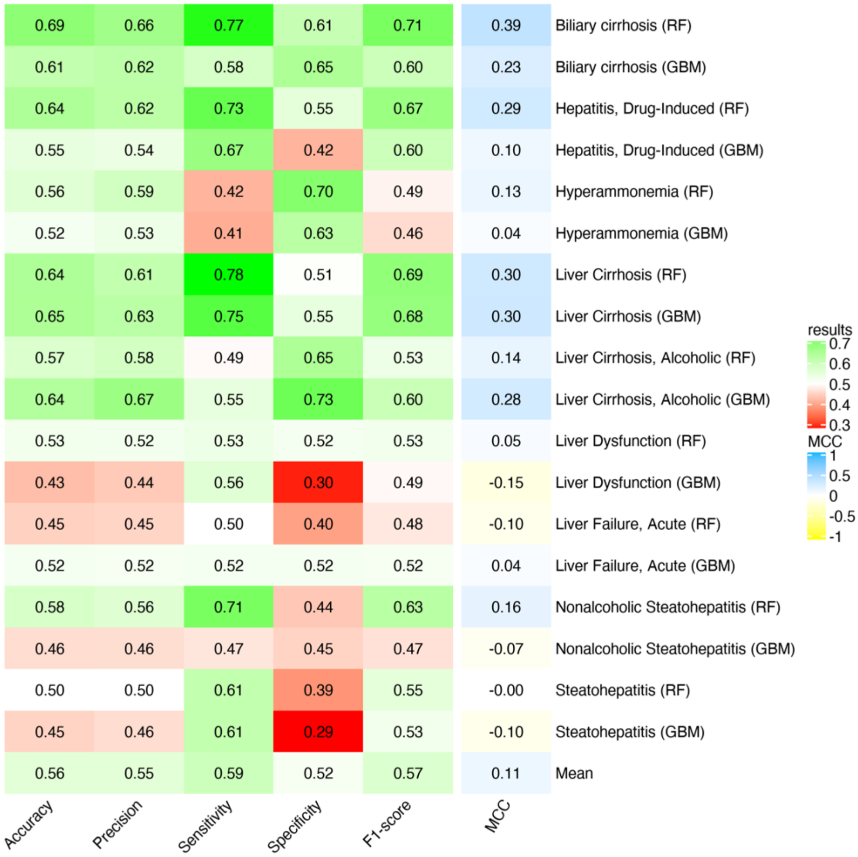
Supplementary Figure 8



	Testing set						Independent hold-out test set					
	Accuracy	Precision	Sensitivity	Specificity	F1-score	MCC	Accuracy	Precision	Sensitivity	Specificity	F1-score	MCC
Landmark (RF)	0.52	0.52	0.57	0.47	0.54	0.04	0.43	0.31	0.53	0.38	0.39	-0.09
Landmark (GBM)	0.53	0.54	0.53	0.54	0.53	0.07	0.47	0.33	0.53	0.44	0.41	-0.03
DisGeNET (RF)	0.69	0.67	0.78	0.61	0.72	0.39	0.47	0.33	0.53	0.44	0.41	-0.03
DisGeNET (GBM)	0.63	0.63	0.60	0.65	0.61	0.26	0.55	0.39	0.53	0.56	0.45	0.09
GUILDify (RF)	0.48	0.47	0.41	0.54	0.44	-0.05	0.55	0.39	0.53	0.56	0.45	0.09
GUILDify (GBM)	0.62	0.66	0.51	0.73	0.57	0.25	0.49	0.35	0.53	0.47	0.42	-0.00
DILI landmark (RF)	0.69	0.71	0.67	0.71	0.69	0.39	0.55	0.38	0.47	0.59	0.42	0.06
DILI landmark (GBM)	0.67	0.67	0.70	0.64	0.68	0.34	0.57	0.43	0.71	0.50	0.53	0.20
Targets (RF)	0.57	0.61	0.40	0.73	0.47	0.15	0.70	0.59	0.59	0.77	0.59	0.35
Targets (GBM)	0.59	0.61	0.45	0.72	0.50	0.19	0.51	0.38	0.53	0.50	0.44	0.03
SMILES (RF)	0.65	0.68	0.58	0.73	0.62	0.31	0.55	0.41	0.71	0.47	0.52	0.17
SMILES (GBM)	0.68	0.73	0.61	0.75	0.66	0.37	0.49	0.37	0.65	0.41	0.47	0.05
DisGeNET + SMILES (RF)	0.68	0.65	0.75	0.60	0.70	0.36	0.47	0.36	0.71	0.34	0.48	0.05
DisGeNET + SMILES (GBM)	0.62	0.61	0.63	0.60	0.62	0.23	0.45	0.34	0.65	0.34	0.45	-0.01
GUILDify + SMILES (RF)	0.49	0.49	0.67	0.30	0.57	-0.02	0.37	0.34	0.88	0.09	0.49	-0.04
GUILDify + SMILES (GBM)	0.51	0.52	0.66	0.35	0.58	0.00	0.43	0.31	0.53	0.38	0.39	-0.09
DILI landmark + SMILES (RF)	0.63	0.65	0.54	0.72	0.59	0.26	0.67	0.52	0.82	0.59	0.64	0.40
DILI landmark + SMILES (GBM)	0.66	0.69	0.58	0.73	0.63	0.32	0.59	0.44	0.65	0.56	0.52	0.20

Supplementary Figure 8. Results of the classifiers in the testing set and the independent hold-out test set. The machine learning algorithms used are either Random Forest (RF) or Gradient Boosting Machine (GBM). Results of DisGeNET, GUILDify, DisGeNET+SMILES and GUILDify+SMILES are from the phenotype “Biliary cirrhosis” (C0023892).

Supplementary Figure 9



Accuracy	Precision	Sensitivity	Specificity	F1-score	MCC	
0.69	0.66	0.77	0.61	0.71	0.39	Biliary cirrhosis (RF)
0.61	0.62	0.58	0.65	0.60	0.23	Biliary cirrhosis (GBM)
0.64	0.62	0.73	0.55	0.67	0.29	Hepatitis, Drug-Induced (RF)
0.55	0.54	0.67	0.42	0.60	0.10	Hepatitis, Drug-Induced (GBM)
0.56	0.59	0.42	0.70	0.49	0.13	Hyperammonemia (RF)
0.52	0.53	0.41	0.63	0.46	0.04	Hyperammonemia (GBM)
0.64	0.61	0.78	0.51	0.69	0.30	Liver Cirrhosis (RF)
0.65	0.63	0.75	0.55	0.68	0.30	Liver Cirrhosis (GBM)
0.57	0.58	0.49	0.65	0.53	0.14	Liver Cirrhosis, Alcoholic (RF)
0.64	0.67	0.55	0.73	0.60	0.28	Liver Cirrhosis, Alcoholic (GBM)
0.53	0.52	0.53	0.52	0.53	0.05	Liver Dysfunction (RF)
0.43	0.44	0.56	0.30	0.49	-0.15	Liver Dysfunction (GBM)
0.45	0.45	0.50	0.40	0.48	-0.10	Liver Failure, Acute (RF)
0.52	0.52	0.52	0.52	0.52	0.04	Liver Failure, Acute (GBM)
0.58	0.56	0.71	0.44	0.63	0.16	Nonalcoholic Steatohepatitis (RF)
0.46	0.46	0.47	0.45	0.47	-0.07	Nonalcoholic Steatohepatitis (GBM)
0.50	0.50	0.61	0.39	0.55	-0.00	Steatohepatitis (RF)
0.45	0.46	0.61	0.29	0.53	-0.10	Steatohepatitis (GBM)
0.56	0.55	0.59	0.52	0.57	0.11	Mean

Supplementary Figure 9. Result of the classifiers based on gene sets from DisGeNET DILI phenotypes in the testing set. The machine learning algorithms used are either Random Forest (RF) or Gradient Boosting Machine (GBM). Each row corresponds to the mean performance of 10 models trained using the PHH gene expression of the genes associated to each DILI phenotype. The "Mean" row corresponds to the average performance of each metric for all the phenotypes.

Supplementary Table 1

Supplementary Table 1: List of associations between DILI phenotypes and genes from DisGeNET and GUILDify. The number "1" in the columns DisGeNET or GUILDify indicates that the phenotype-gene association comes from this source, and the number "0" indicates the opposite.
The table is provided online at:
https://github.com/structuralbioinformatics/CAMDA2019-DILI/blob/master/outputs/tables/SupplementaryTable1.tsv

Supplementary Table 2

Supplementary Table 2: List of SMILES from the drugs of the analysis.
The table is provided online at:
https://github.com/structuralbioinformatics/CAMDA2019-DILI/blob/master/outputs/tables/SupplementaryTable2.tsv

Supplementary Table 3

Supplementary Table 3: Tanimoto distance matrix between the drugs of the analysis.
The table is provided online at:
https://github.com/structuralbioinformatics/CAMDA2019-DILI/blob/master/outputs/tables/SupplementaryTable3.tsv

Supplementary Table 4

Supplementary Table 4: List of drug-target associations used in the analysis. The drug-target associations are retrieved from DGIdb, HitPick and SEA. The number “1” indicates a drug-target association, and the number “0” indicates the opposite.
The table is provided online at:
https://github.com/structuralbioinformatics/CAMDA2019-DILI/blob/master/outputs/tables/SupplementaryTable4.tsv

Supplementary Table 5

Supplementary Table 5: Number of the drugs used in each step of the machine learning process. In parenthesis, the number of drugs when using “Targets” feature.

Type of drug	Number of drugs		
	Complete dataset	**Training**	**Testing**
DILIrank drugs	175 (172)	72	30
DILI-Concern drugs	124 (121)	36	15
Most-DILI-Concern drugs	37 (36)	11	4
Less-DILI-Concern drugs	87 (85)	25	11
No-DILI-Concern drugs	51 (51)	36	15
Independent hold-out test dataset drugs	55 (53)		

Supplementary Table 6

Supplementary Table 6: List of hepatotoxic genes from the study of *Peng et al. (2019)* and their overlap with the datasets of the article.

The table is provided online at:

https://github.com/structuralbioinformatics/CAMDA2019-DILI/blob/master/outputs/tables/SupplementaryTable6.tsv

Supplementary Table 7

Supplementary Table 7: List of genes from the DILI Landmark gene signature (obtained from a non-parametric Wilcoxon test).

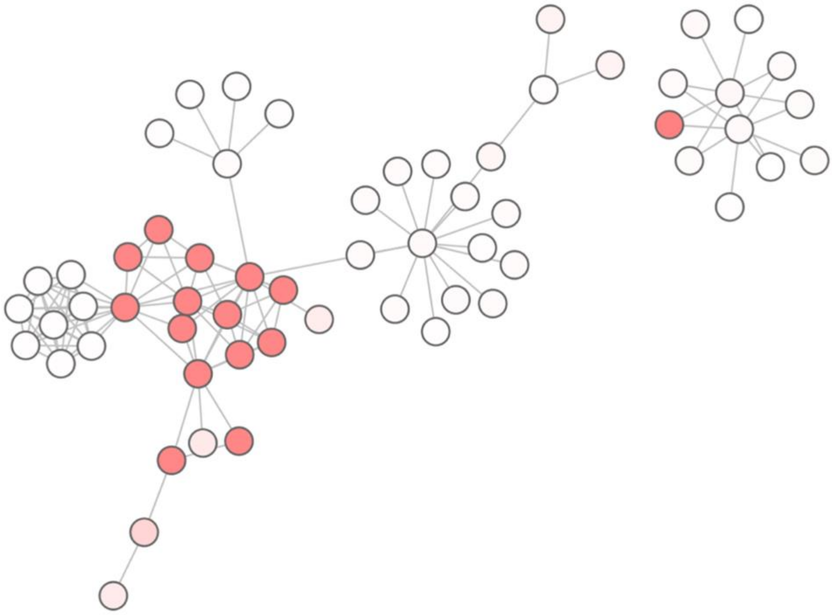
The table is provided online at:

https://github.com/structuralbioinformatics/CAMDA2019-DILI/blob/master/outputs/tables/SupplementaryTable7.txt

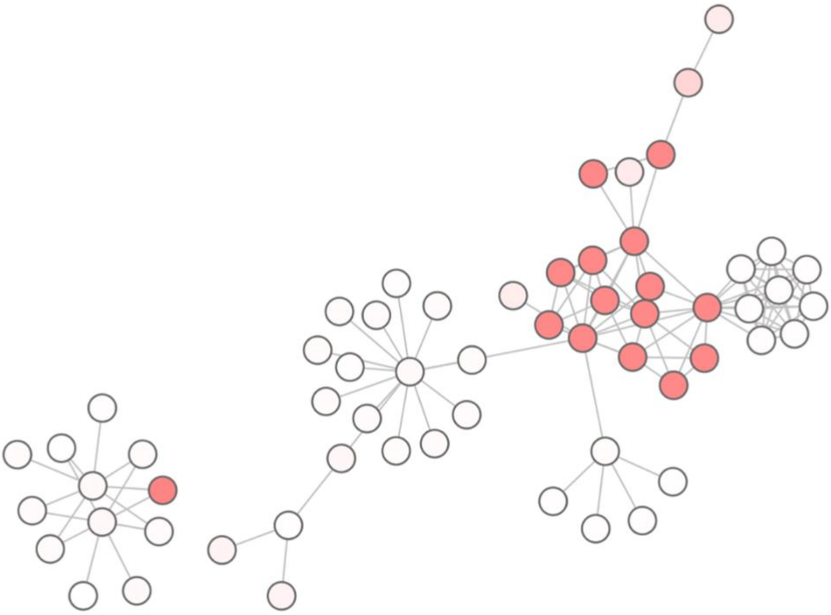
Supplementary Table 8

Supplementary Table 8: List of 20 target proteins that are targeted by the largest number of drugs from the independent hold-out dataset (Ambiguous-DILI drugs). In the table, we provide the number and percentage of interacting drugs from the independent hold-out dataset, and the number and percentage of interacting in total (from the 230 drugs of the dataset).

Target name	Num. drugs interacting (Ambiguous-DILI)	% drugs interacting (Ambiguous-DILI)	Num. drugs interacting (Total)	% drugs interacting (Total)
CYP2D6	18	32.7	92	40.0
CYP3A4	17	30.9	100	43.5
ABCB1	17	30.9	84	36.5
CYP2C9	11	20.0	74	32.2
CYP1A2	11	20.0	59	25.7
DRD2	10	18.2	42	18.3
HTR2A	9	16.4	45	19.6
CYP3A5	8	14.5	46	20.0
ADRA2B	8	14.5	41	17.8
DRD1	8	14.5	34	14.8
ADRA2C	8	14.5	34	14.8
HTR1A	8	14.5	31	13.5
SLC22A1	8	14.5	25	10.9
ESR1	7	12.7	22	9.6
CA1	7	12.7	15	6.5
HTR2C	7	12.7	40	17.4
ADRA2A	7	12.7	37	16.1
HRH1	7	12.7	30	13.0
ADRB1	7	12.7	24	10.4
CYP2C19	6	10.9	40	17.4



4. DISCUSSION



The inner workings of a human organism are mediated by different interconnected biomolecules that interact between themselves to carry out their functions. In the era of big data, we have millions of records of biological information about these interactions available in public databases, but this information is spread and unorganized, making it impossible to comprehend. Here it is where network medicine emerges, providing tools to organize the biological information in networks, as well as algorithms to analyze it and understand better the molecular mechanisms of the human organism.

My thesis has focused on developing network medicine tools and approaches for a better understanding of diseases and polypharmacology. From this point forward, I will proceed with the discussion of this thesis by explaining the contribution of my research to the field, its limitations and consider future developments.

4.1. The identification of disease modules and their associated functions is key for the understanding of disease complexity

Understanding the molecular complexity of human disease is one of the main objectives of this thesis. For this purpose, we developed GUILDify v2.0 (**Article 3.1**), a method to identify the modules associated to diseases in the PPI network and understand their molecular mechanisms and relationships. For the identification of disease modules, GUILDify uses a network-based diffusion algorithm that extends the knowledge on disease-gene associations to calculate the impact of a disease through the PPI network. In this way, when assessing the molecular mechanism of a disease, the focus is not limited to the disease-gene associations (which are incomplete), but the effect to their interactions is also considered.

Still, our knowledge on PPIs is incomplete: many interactions between proteins are yet to be discovered (4). For this reason, the use of network-based algorithms to assess the impact of a disease in the interactome might as well be inaccurate. To solve this problem, the new version of GUILDify also considers how the disease module perturbates some biological functions and pathways. Thanks to this addition, the information predicted by GUILDify is not limited to PPIs but extended to a functional level.

Apart from inspecting the molecular mechanisms of diseases, the versatility of the algorithms of GUILDify make them useful for different applications. GUILDify has been previously applied to (i) find comorbidities across genetic diseases (92), (ii) construct PPI networks specific to breast cancer metastasis to the lung and brain (265), (iii) identify candidate genes for body size in sheep (266) and (iv) prioritize preeclampsia pathogenesis (267). Precisely, the methodology applied in Rubio et al. (92) has been implemented inside the new version of GUILDify web server to facilitate the study of comorbidities.

Moreover, GUILDify has also been applied in other publications of this thesis, becoming a useful tool in a wide range of molecular contexts. In **Appendix 6.3**, GUILDify was used as a support tool to investigate the effect of Δ9-tetrahydrocannabinol in the mouse brain.

The GUILDify diffusion algorithms were used to identify proteins and pathways in the mouse interactome with the greatest association to proteins modulated by an amnesic dose Δ9-tetrahydrocannabinol. The analysis indicated a significant alteration of the proteasome function, since top scoring proteins were related to the proteasome system, a protein complex involved in ATP-dependent protein degradation. In **Article 3.3** and **Appendix 6.6**, GUILDify was used as an alternative method to corroborate the mechanisms of action of drugs predicted by TPMS for specific diseases. In **Article 3.3**, GUILDify was used to identify the network modules associated to the diseases heart failure, macular degeneration, and the drug combination sacubitril/valsartan. 10 of the 30 proteins proposed by TPMS to identify heart failure patients at risk of developing macular degeneration were found among the union between the three network modules, corroborating the molecular context predicted by TPMS. In **Appendix 6.6**, GUILDify was used to calculate the network modules associated to the mechanisms by which SARS-CoV-2 enters an organism and produces the infection (entry points), the effects produced by SARS-CoV-2 infection (acute respiratory distress), and their overlap with the proteins affected by the combination of the drugs melatonin and pirfenidone. We confirmed the effect of the combination in the entry points of the SARS-CoV-2 infection, specifically the neighbors of furin and GRP-78, and some proteins associated with acute respiratory distress. In **Article 3.4**, GUILDify was used as a support tool to identify gene expression signatures associated to Drug-Induced Liver Injury (DILI) (see **Chapter 4.5** for a more detailed discussion). Finally, in **Appendix 6.4**, GUILDify was integrated as part of the InteractoMIX Galaxy platform, as part of different pipelines to facilitate the study of the interactome.

4.2. Endopharmacology: a promising field to repurpose drugs targeting shared pathways

Network medicine not only focuses on improving our molecular knowledge on diseases but also on proposing specific treatments for them. In a work by Guney et al. (79), it was proposed that a drug was more prone to be effective against a disease if it was targeting the proteins within or in the immediate neighborhood of the corresponding disease module. The authors proposed a network-based drug-disease proximity measure where they quantified the distance between the proteins in the disease module and the targets of the drug of interest, as described in **Chapter 1.4.3.2**. In **Article 3.2**, we showed that using the proximity measure, we were able to uncover a higher number of pathways involved in autoimmune diseases than using conventional approaches such as gene or pathway overlap. The measure was also useful to reveal the relationships between these diseases.

As the proximity measure proposed by Guney et al. (79) targets specific disease modules, in **Article 3.2** we investigated if the approach could also be applied to target specific pathways shared by several disease modules. Biological pathways tend to crosstalk, i.e., they interact or influence each other. Therefore, pathway crosstalk plays an important role in modulating the pathophysiology of diseases (175) and many comorbid diseases are connected to each other in the interactome through proteins belonging to related pathways (77,92,176). These intermediate pathways shared among diseases (e.g., on comorbid diseases) are called endophenotypes (177).

Endophenotypes have recently emerged as an attractive way to study the shared mechanisms between several diseases. Ghiassian et al. (179) addressed the endophenotypes of the inflammasome, thrombosome, and fibrosome, and described their roles in the progression of cardiovascular diseases. The characteristics of endophenotypes also make them the target options to treat comorbidities, but their role in pharmacology has not been properly investigated yet.

Inspired by the study of Guney et al. (79), but targeting specifically the endophenotypes shared by disease modules, we have proposed a new drug repurposing approach called PxEA (**Article 3.2**). PxEA scores drugs based on the network-based proximity of their targets to the proteins of the pathways of interest (i.e., the common pathways of two diseases). The approach was first applied to identify drug repurposing candidates for autoimmune disorders. We investigated whether the drugs promiscuously used in these disorders target specifically the pathways associated with one disease or the pathways shared across the diseases. Using PxEA, we found common pathways between almost all autoimmune disorders and drugs potentially targeting these common pathways. Second, we also explored the potential endophenotypes shared by type 2 diabetes and Alzheimer’s disease, two diseases highly prevalent in our ageing society that are known to exhibit increased comorbidity (268,269). Among the top scoring drugs proposed by PxEA, we found orlistat, a drug indicated by type 2 diabetes which has been suggested for the treatment of Alzheimer’s disease (270).

The role of endophenotypes in different diseases has been previously studied using a network medicine vision (179), but this is the first study proposing actual therapeutic options to target endophenotypes. Thus, with PxEA we are paving the way towards endopharmacology, a new field inside network medicine focused on understanding the molecular mechanisms of endophenotypes and proposing drug candidates that target them.

4.3. Limitations of network medicine tools when representing a perturbation in the network

The main limitation in the previous approaches is to omit that, in some cases, the PPI network has an inherent directionality. Although the model of PPI network considered in most of network medicine studies is undirected, in reality there is a direction in some interactions of the network (271). For example, the perturbation of the elements of the network caused by internal or external factors (e.g., the perturbation of proteins caused by a disease or the interactions of target proteins with a drug) can provoke their activation or inhibition, consequently causing a signal that may affect the other elements of the network. These perturbations can be mimicked by diffusion, clustering or proximity algorithms, but usually these algorithms do not take into account which elements are activated and which ones are inhibited, failing to represent the type of perturbation of these elements. Guney et al. (79) already discussed about this limitation, stating that the network proximity algorithm does not imply that a proximal drug will improve the corresponding disease. The drug could even induce the disease state instead of inhibiting it.

The same limitation applies to the **Articles 3.1 and 3.2** of this thesis. In the case of PxEA (**Article 3.2**), the identification of drugs that are close to endophenotypes does not imply that the drugs will be effective to treat the diseases associated to the endophenotypes. It implies that the drugs will be more likely to perturbate the crosstalk between pathways in the endophenotypes, but not the way in which this perturbation may affect a disease. In the case of GUILDify (**Article 3.1**), the limitation is similar: the diffusion algorithms of GUILDify can simulate a perturbation that starts from the disease-associated genes, but it is not possible to know the type of perturbation in each protein, and therefore the transmission of the perturbation may not be correct.

To overcome this limitation, we can use other types of data that guide the effect of the perturbation in the network. Gene expression data from perturbation samples may be a useful resource to evaluate the impact of a drug in a set of genes. Projects such as the L1000 platform of CMap (224) facilitate this task, providing an extensive dataset of perturbation cell line samples from thousands of compounds. Do Valle et al. (272) have started incorporating gene expression data from CMap to validate network proximity predictions between polyphenol targets and disease modules. In their study, the authors used an enrichment score to measure the overrepresentation of disease genes among the most perturbated genes by polyphenols according to gene expression samples. They observed that the diseases that are more proximal to polyphenol targets show higher perturbation values in gene expression samples than distal diseases. Still, they only use gene expression data as a validation of the proximity measure, but they do not check if the expression of the polyphenols counteracts the expression of the disease-associated proteins. This is still very limited by the quantity and quality of gene expression data available on disease and drug perturbations.

4.4. Modelling perturbations with TPMS: transmitting a signal of activation or inactivation through the network

One of the solutions to this limitation is given in **Article 3.3** by the approach of TPMS. TPMS simulates how a *stimulus* (i.e., proteins activated or inhibited by a drug) produces a *response* (i.e., counteraction of proteins induced or inhibited by a disease) in the PPI network. The data from the drug is retrieved from drug target databases such as DrugBank (199), PubChem (214), STITCH (215) and SuperTarget (212), whereas the data for the phenotype is retrieved from the private repository Biological Effectors Database (BED) (273,274). The algorithm transmits the perturbation from the drug targets to the disease proteins by mimicking a neural network (where the proteins are the neurons, and the edges of the network are used to transfer the signal). Therefore, it assigns a signal value between -1 (inhibition) and +1 (activation) to all the proteins in between the input and output signals. These values are optimized following an iterative process, obtaining in the end a final set of potential solutions. By defining a set of restrictions that come both from the topology of the network and from gene expression datasets, these solutions are evaluated and ranked. Using this approach, we are able to simulate the mechanism of action of a drug taking into account the type of perturbations provoked by the drug, and how they modulate and counteract the disease-associated proteins.

TPMS not only simulates the signal, by considering the activation or inactivation of proteins by the drug, but also predicts all possible mechanisms of action according to the restrictions. Therefore, TPMS is simulating all the possible responses of a drug that we could find in real life without using data from patients. The main limitation, though, is that we are not able to know if these mechanisms of action are happening or not in real patients. It could be the case that some of these mechanisms of action are never represented, whereas others are more common. But this is impossible to know without using data from real patients. Another limitation of TPMS is that the signal transmitted through the network is not a real measure. It is just a qualitative measure to predict which proteins could be activated or inactivated after the intake of the drug. It would be interesting for the future to incorporate real quantitative measures such as the dosage of the drug or the quantity of expression of proteins in the models.

4.5. Using network medicine as a support tool to identify gene signatures and model drug adverse reactions

As seen during the thesis, network medicine can create reliable models of the molecular mechanisms of diseases or drugs. In a similar way, it can also be a tool to understand better drug-adverse reactions. In **Article 3.4**, we developed a machine learning approach to predict Drug-Induced Liver Injury (DILI) using an ensemble of different types of data (gene expression, structural features, drug-target associations). We leveraged gene expression data from CMap to find a specific DILI gene signature that could be used as a feature to predict the drugs producing the adverse reaction. Here, network medicine was not directly applied to model the drug adverse reaction, but it was used as a tool to identify a specific DILI gene signature. Specifically, we applied GUILDify to extend the current knowledge of DILI-associated genes obtained from DisGeNET.

The results of the majority of approaches tested were comparable with the ones from previous publications (260,261). However, the accuracy of the best performing classifier was around the 70% mark, stating the limitations of predicting DILI. This may be explained by different factors:

(1) **The inherent variability and noise of the gene expression dataset:** CMap collects gene expression signatures obtained from cell lines upon treatments with different drug concentrations and durations. We decided to focus on the drugs tested at the highest concentration (10 µM) and for the longest treatment duration (24 h) in the “Primary Human Hepatocytes” cell line. Still, the heterogeneity of transcriptomics response in DILI is very high, hindering the predictions solely based on gene expression.

(2) **The reduced size and imbalance of the gold standard:** The gold standard is comprised of 124 drugs labelled as DILI (37 as Most-DILI-Concern and 87 as Less-DILI-Concern) and 51 labelled as no DILI. The reduced size of the training

and testing datasets, and the imbalance between different DILI labels makes the prediction more challenging.

(3) **The important genetic diversity between the different DILI-related phenotypes:** We manually identified a list of 9 phenotypes closely related with DILI. When comparing their phenotype-gene associations, we observe a very small overlap of genes between them, reflecting the diversity of phenotypes considered inside the DILI term, and the challenge associated to predict DILI based solely on gene expression.

(4) **The great structural diversity between the drugs reported as DILI-Concern:** When plotting the drug-drug structural similarity between the drugs of the gold standard, we observe that drugs of the same group did not have higher similarity among them than with other groups. This indicates that there is a considerable structure heterogeneity within the Most-DILI-Concern group of drugs, which complicates the prediction using chemical structure.

In the case of the gene signature obtained through network medicine (by means of the GUILDify web server), the quality of the prediction decreased with respect to using solely phenotype-gene associations. This is possibly because when expanding the number of genes in the signature using GUILDify, the intrinsic noise in the gene expression dataset increases as well, complicating the prediction.

Still, here we show how network medicine applications can be used to support other approaches, as they improve the understanding of the molecular context. Moreover, we think that the modelling and prediction of the DILI could be improved in the future if instead of focusing on gene expression signatures, we apply network medicine approaches such as the detection of disease modules or proximity measure between the adverse drug reaction and the drug targets.

4.6. The integration of molecular interactions is the basis of network medicine research

Network medicine applications usually rely in a model of the interactome (usually a PPI network) as the basis from which the analyses are performed and conclusions are extracted (see **Articles 3.1 to 3.4**). The incompleteness of the PPI network is yet one of the main limitations in network medicine studies (76). The more complete the model of the interactome is, the more accurate the predictions associated with the model will be. Therefore, the first step in network medicine research is to accurately model the PPI network of the species of interest. To do so, it is necessary to integrate PPI data by taking into account the five main challenges explained in **Chapter 1.1.6**: (1) PPI data is spread across multiple repositories (e.g. databases and publications); (2) the nomenclature of the proteins is different; (3) there are different formats to store PPIs; (4) the reliability of the PPI data varies on each experiment; and (5) tissue-specific interactions are not generally considered.

To study these challenges, we used the software BIANA (62), which gives flexibility to integrate biological molecules and interactions from different databases and derive networks. BIANA has been updated three times during the thesis period (June of 2017, May of 2018 and April of 2020). The updates included the parsing of the following types of biological data:

- **Protein information** (including gene and protein names, identifiers, sequence, structure, family, species, pathways) from Uniprot (109), NCBI Gene (275), HGNC (276), Gene Ontology (277) and Reactome (278).
- **Protein-protein interactions** from IntAct (51), BioGRID (52), HIPPIE (64), ConsensusPathDB (59), I2D (61) and InnateDB (279).
- **Drug information** (including drug name, indication, drug-target associations, drug-drug associations, ATC) from DrugBank (199), DrugCentral (200), DGIdb (201), ChEMBL (202) and TTD (205).
- **Disease-gene associations** from DisGeNET (110).

BIANA allowed the creation of the biological networks that served as basis of the research projects presented during this thesis.

4.7. Studying the interactome with atomistic and structural level of detail

For an accurate representation of the interactome, it is key to provide tools to study the proteins and their interactions in a structural level of detail. The structures of proteins and PPIs permit a more precise understanding of their functioning. During this thesis, I had the opportunity to contribute to several side-projects related with the identification and evaluation of the structure of proteins and PPIs. These projects nurtured my knowledge on proteomics and interactomics, which are closely related with network medicine.

First, I contributed to the publication of the stand-alone program MODPIN (**Appendix 6.5**) for the prediction of PPI structures based on comparative modelling. MODPIN uses comparative modelling to obtain an ensemble of structural models of the PPI. These models are clustered according to common structural elements in their interfaces and evaluated using scoring functions.

I also participated in the publication of BADock (**Appendix 6.1**), a tool to predict the binding affinity of PPIs without requering the structure of the complex. Normally, most of the methods to predict the binding affinity of a PPI require the structure of the complex. BADock only requires as input the unbound structure of the proteins involved in the interaction. BADock uses docking techniques to explore all the potential conformations of the interaction and employs a regression-based classifier to infer the binding affinity of the protein complex.

Additionally, I was the first author in the publication of the SPServer (**Appendix 6.7**), a web server to evaluate the quality of protein folds and PPI structures based on knowledge-based potentials (see **Chapter 1.1.5.3** for a more detailed explanation). The innovative point in SPServer is its accessibility: an easy-to-use interface designed to facilitate its use and interpretation of the results. The resulting scores are displayed as interactive graphics that permit the user to compare the quality of multiple structures at the same time.

Finally, I also contributed to the construction of InteractoMIX (**Appendix 6.4**), a Galaxy-based platform for the study of PPI data. In InteractoMIX, we integrated several bioinformatics tools such as GUILDify, BIANA, MODPIN or BADock, and designed several pipelines to facilitate and communicate their use on the study of the interactome.

4.8. Facilitate the access to network medicine tools is one of the keys of this thesis

One of the main problems of bioinformatics in general and network medicine in particular is that their methods are not user-friendly. Most of the approaches are theoretical approximations or standalone programs, developed by specialists whose use is not trivial to non-expert users.

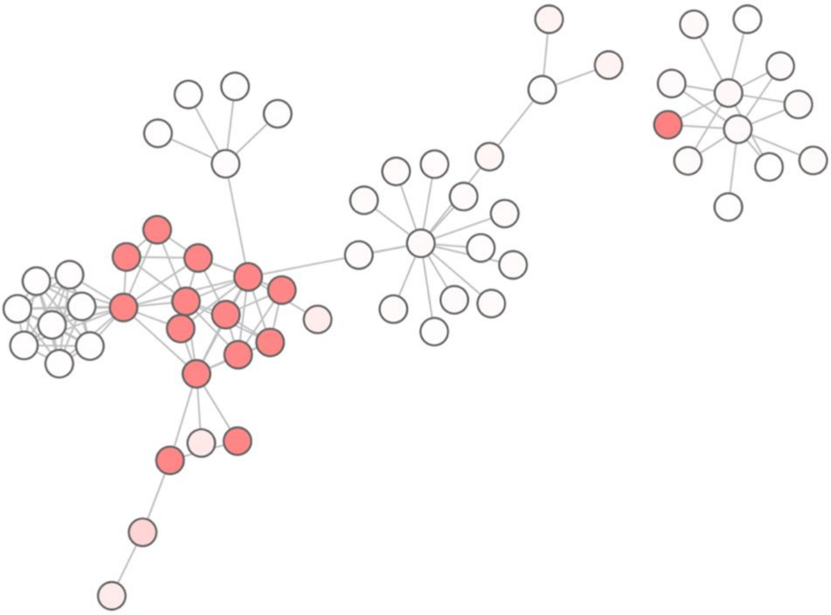
To solve this demand, GUILDify (**Article 3.1**) emerges as an easy-to-use network medicine web server, permitting an accessible identification of disease modules, disease-disease relationships and drug repurposing. The interface of GUILDify is designed to be simple and intuitive. For a new search, the user only has to introduce a disease or drug name in a Google-like search bar, and the web server directly provides the user with disease-associated genes or drug targets associated with the input. Then, the user needs to select the genes of interest and continue, and the web server calculates the module associated to the disease or drug and provides an intuitive results page. The results page includes a network visualization powered by cytoscape.js (263), and detailed tables of the top-scoring proteins conforming the network module and the enriched biological functions and pathways. Thus, the web allows to perform a complex network medicine analysis in a few easy steps without complications.

Following the same concept of user-friendliness as GUILDify, we also designed the SPServer (**Appendix 6.7**) as an easy-to-use web server to evaluate models of protein folds and PPIs. Many scientists in the field of proteomics and interactomics need to deal with structural models of proteins and complexes, and there are very few web servers offering an easy, interactive evaluation of models. With this idea in mind, we developed SPServer, so that a non-specialized user can perform an evaluation of models in few steps and obtain interactive graphics that facilitate the interpretation and comparison of results.

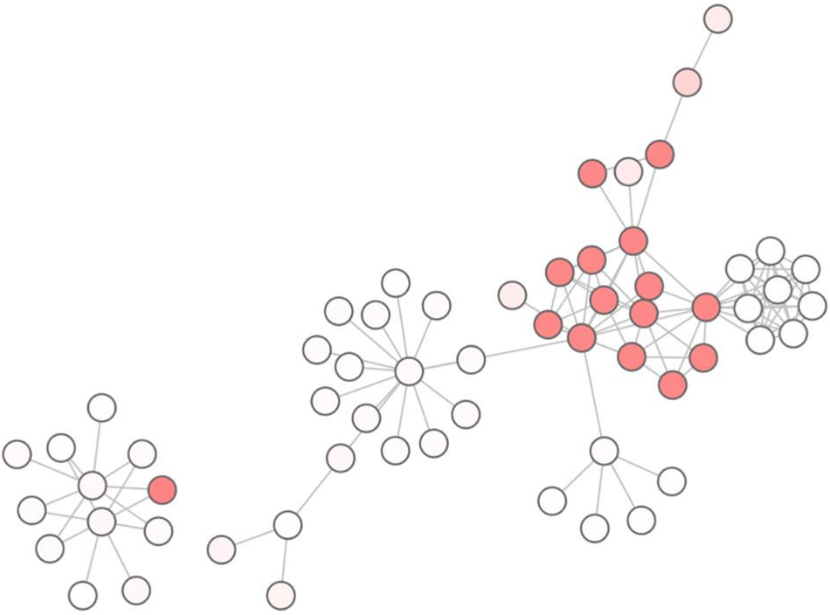
Finally, some of these tools have been integrated in the Galaxy-based platform InteractoMIX (**Appendix 6.4**), for an easier integration. Galaxy is a web-based platform to integrate bioinformatics tools and analyze large biomedical datasets (50). During the last few years, Galaxy is becoming the reference platform for accessibility and reproducibility of bioinformatics tools. Therefore, the inclusion of network medicine tools such as GUILDify or BIANA, and other structural bioinformatics tools such as MODPIN or BADock in a Galaxy platform is key for its accessibility to a wider and non-specialized audience.

4.9. Future perspectives in the field of network medicine

Network medicine emerges as the field that organizes the biological interactions of the organism in networks, with the objective of understanding the molecular complexity of diseases and find better treatments. Since the birth of the field, in 2007 (280), network medicine has evolved rapidly. It is merging with multiple other fields such as machine learning, network pharmacology or pharmacogenomics, and giving place to important consortiums such as the International Network Medicine Consortium (181). However, network medicine is still a very young field that needs to address some limitations to mature and expand. For the sake of examples, we may consider: (1) how to integrate different types of data to represent an interactome as complete as possible; (2) how to represent the signal of the network perturbations provoked by drugs and diseases; and (3) how to personalize the models for specific patients. Finally, novel approaches, such as the construction of multi-layered networks or the analysis of multi-omics datasets are leading the expansion of network medicine towards more precise and personalized models.



5. CONCLUSIONS



In this thesis, we have developed a series of network medicine *in silico* tools and studies with the aim to understand better the molecular mechanisms of diseases and polypharmacology. We have reached the following conclusions:

- The incorporation of biological functions and pathways as an extension of the analysis of the molecular interactions perturbated by drugs and diseases improves the understanding of their molecular mechanisms and relationships.

- GUILDify v2.0 pioneers in being one of the most user-friendly network medicine tools, allowing complex network medicine analyses such as the identification of disease modules, disease-disease relationships and drug repurposing in a few steps.

- The arsenal of functionalities provided by the update of GUILDify has been proofed useful in a wide range of applications:

    - Identify proteins and pathways in the mouse interactome with the greatest association to proteins modulated by an amnesic dose Δ9-tetrahydrocannabinol (**Appendix 6.3**).

    - Contextualize the predictions of mechanisms of action of the drug made by alternative network medicine tools (**Article 3.3** and **Appendix 6.6**).

  - Guide the prediction of gene signatures associated to DILI (**Article 3.4**).

- PxEA permits to repurpose drugs targeting endophenotypes. We demonstrated its application in two different scenarios:

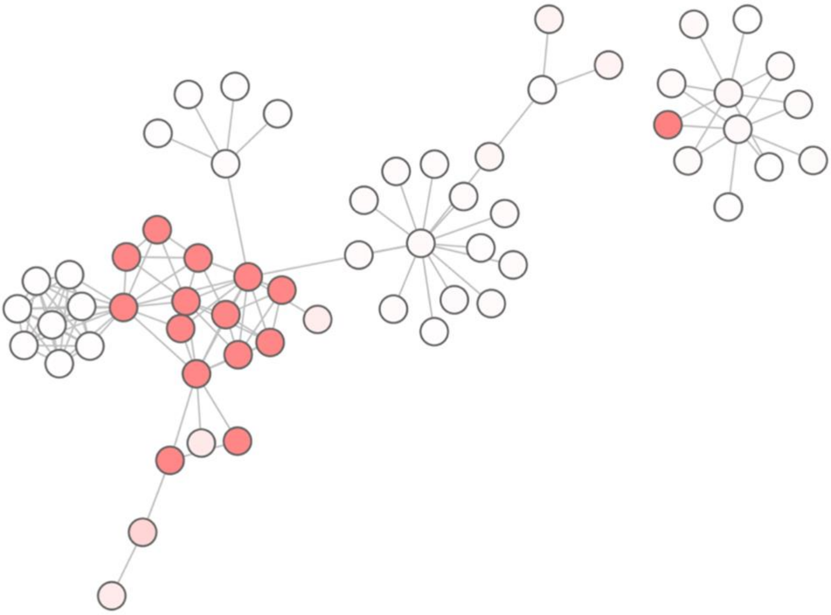
  - Identifying the common pathways shared by almost all autoimmune diseases and proposing drug candidates targeting them.

  - Exploring the pathways shared by type 2 diabetes and Alzheimer's disease, and finding orlistat among the top scoring drugs, which is a drug indicated by type 2 diabetes that has been suggested by the treatment of Alzheimer's disease.

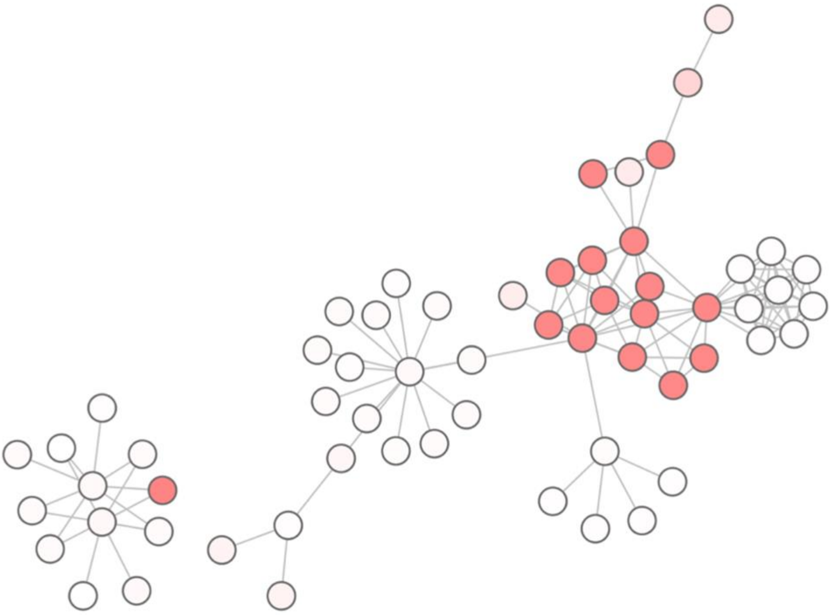
- In co-authorship with Guillem Jorba from Anaxomics Biotech S.L., we used TPMS to identify a list of mechanisms of action of the drug combination sacubitril/valsartan to treat heart failure and/or produce macular degeneration.

- We developed a methodology to associate the mechanisms of action identified by TPMS to different classes of prototype-patients and identify biomarker proteins that allow the differentiation of prototype-patients.

- We integrated different types of omics data (gene expression, structural features, drug-target associations) as features for a machine learning ensemble and assessed their accuracy in predicting DILI.



6. APPENDIX



In this section, I include other publications where I contributed during my thesis.

6.1. On the mechanisms of protein interactions: predicting their affinity from unbound tertiary structures

Most of the computational methods to predict the binding affinity of PPIs require the structure of the protein complex, which in many cases is difficult to obtain. In Marín-López et al., we presented a novel method called BADock to predict the binding affinity of PPIs only based on the structure of the unbound protein structures, thus overcoming this limitation.

In this article, I contributed to the assessment of the method using different benchmarks, and also in the preparation of a web that permits to use BADock in an intuitive way.

Marín-López MA, Planas-Iglesias J, **Aguirre-Plans J**, Bonet J, Garcia-Garcia J, Fernandez-Fuentes N, Oliva B. On the mechanisms of protein interactions: predicting their affinity from unbound tertiary structures. ***Bioinformatics*.** 2018; 34(4):592-598. DOI: 10.1093/bioinformatics/btx616

6.2. Network, Transcriptomic and Genomic Features Differentiate Genes Relevant for Drug Response

In Piñero et al., we compiled drug-target associations to characterize their transcriptomics, genomics and network features. We classified the drug-target associations in three classes: (i) TARGET, if they mediated the therapeutic effects of drugs; (ii) METAB, if they acted as drug transporters, carriers, or enzymes, involved in the drug absorption, distribution and metabolism; and (iii) TOXPROT, if they were associated to side effects or toxicity phenotypes of drugs. We explored the properties of these proteins within different global or organ-specific interactomes using multi-scale network features.

I compilated and integrated data to create several of the global and organ-specific interactomes, and participated in the review and discussion of the results.

Piñero J, Gonzalez-Perez A, Guney E, **Aguirre-Plans J**, Sanz F, Oliva B, Furlong LI. Network, Transcriptomic and Genomic Features Differentiate Genes Relevant for Drug Response. ***Front Genet*.** 2018; 9:412. DOI: 10.3389/fgene.2018.00412

6.3. Δ9-tetrahydrocannabinol modulates the proteasome system in the brain

In Salgado-Mendialdúa et al., we analyzed, through a proteomic screening of hippocampal synaptosomal fractions, those proteins and pathways modulated 3 hours after a single administration of an amnesic dose of Δ9-tetrahydrocannabinol

I contributed to the article by applying GUILDify v2.0 to identify the proteins and pathways in the mouse interactome with the greatest association to the proteins modulated by the amnesic dose of Δ9-tetrahydrocannabinol. A functional enrichment analysis of the top-scoring proteins by the GUILDify algorithm showed a significant over-representation of metabolic processes involving mitochondrial physiology and cellular respiration, as well as cytoskeletal reorganization pathways. It also pinpointed the proteasome complex among the pathways enriched in down-regulated proteins.

Salgado-Mendialdúa V, **Aguirre-Plans J**, Guney E, Reig-Viader R, Maldonado R, Bayés À, Oliva B, Ozaita A. Δ9-tetrahydrocannabinol modulates the proteasome system in the brain. ***Biochem Pharmacol*.** 2018; 157:159-168. DOI: 10.1016/j.bcp.2018.08.026

6.4. Galaxy InteractoMIX: An Integrated Computational Platform for the Study of Protein-Protein Interaction Data

Mirela-Bota et al. describes InteractoMIX, a Galaxy-based platform for the study of PPI data. In InteractoMIX, we integrated several genomics, proteomics and interactomics tools and designed several pipelines to facilitate and communicate their use on the study of the interactome.

This study was made in collaboration with the current and former methods of the Bioinsilico (UVic) and Structural Bioinformatics (UPF) groups. My contribution was to allow the use of the network medicine tools BIANA and GUILDify through the platform. I also participated in the preparation of the manuscript.

Mirela-Bota P, **Aguirre-Plans J**, Meseguer A, Galletti C, Segura J, Planas-Iglesias J, Garcia-Garcia J, Guney E, Oliva B, Fernandez-Fuentes N. Galaxy InteractoMIX: An Integrated Computational Platform for the Study of Protein-Protein Interaction Data. ***J Mol Biol.*** 2020; 166656. DOI: 10.1016/j.jmb.2020.09.015

6.5. Using collections of structural models to predict changes of binding affinity caused by mutations in protein–protein interactions

Meseguer et al. describes MODPIN, a tool to model the atomic structure of PPI complexes through comparative modelling. MODPIN automatizes the comparative modelling process to obtain an ensemble of structural models of the PPI of interest. These models are clustered according to common structural elements in their interfaces and evaluated using different scoring functions. MODPIN has been applied to predict changes of binding affinity caused by mutations affecting PPIs.

I contributed to the study by compiling one of the datasets used for the testing of the method, and I also discussed and reviewed the results of the different analyses.

Meseguer A, Dominguez L, Bota PM, **Aguirre-Plans J**, Bonet J, Fernandez-Fuentes N, Oliva B. Using collections of structural models to predict changes of binding affinity caused by mutations in protein–protein interactions. ***Protein Sci*.** 2020; 29(10):2112-2130. DOI: 10.1002/pro.3930

6.6. 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

In Artigas et al., we described the use of the network medicine tools TPMS and GUILDify to predict drug repurposing candidates for the treatment of COVID-19. TPMS was used to unveil the mechanisms of action of different drugs on proteins associated to different mechanisms by which SARS-CoV-2 enters an organism, produces the infection and the adverse reactions associated. We identified the drug combination of melatonin and pirfenidone as a promising candidate because of its potential to reduce the infection of the virus and its good safety profile.

I contributed to the study by applying GUILDify to identify the disease modules associated to the infection of SARS-CoV-2 (entry points), the effect of the infection (acute respiratory distress) and their overlap with the proteins targeted by the drugs melatonin and pirfenidone. The network medicine study confirmed a potential effect of the combination of pirfenidone and melatonin in the entry points of the SARS-CoV-2 infection, specifically the neighbors of furin and GRP-78, and some proteins associated with ARD. I also participated actively in the writing and reviewing of the manuscript and the analysis of results.

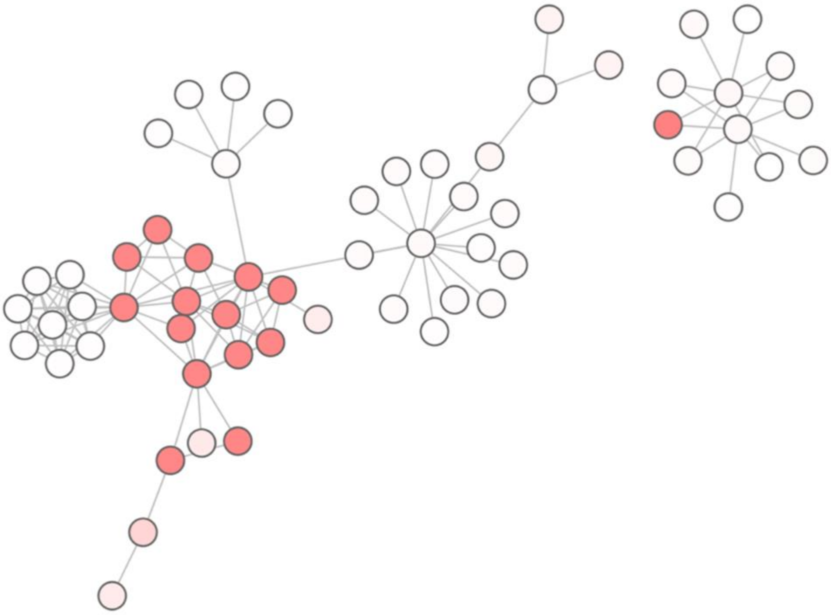
Artigas L, Coma M, Matos-Filipe P, **Aguirre-Plans J**, Farrés J, Valls R, Fernandez-Fuentes N, de la Haba-Rodriguez J, Olvera A, Barbera J, Morales R, Oliva B, Mas JM. In-silico drug repurposing study predicts the combination of pirfenidone and melatonin as a promising candidate therapy to reduce SARS-CoV-2 infection progression and respiratory distress caused by cytokine storm. ***PLoS One*.** 2020; 15(10):e0240149. DOI: 10.1371/journal.pone.0240149

6.7. SPServer: split-statistical potentials for the analysis of protein structures and protein-protein interactions

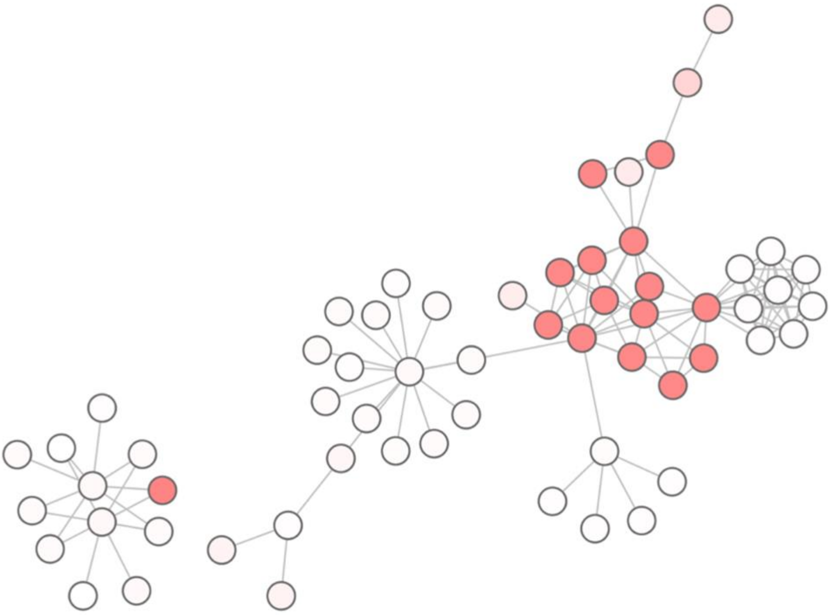
This article describes the SPServer, a web server that applies knowledge-based potentials to score and analyze the structure of protein folds and PPIs. SPServer integrates the analysis of protein folds and PPIs in a unique, easy-to-use web.

As a first author of this publication, I have contributed to the development of the web server, testing of the method, analysis of results and writing of the manuscript. As this publication is closer to a structural bioinformatics perspective and more distant from the network medicine field, I decided to include this publication here in the appendix of the thesis rather than in the main results.

Aguirre-Plans J, Meseguer A, Molina-Fernandez R, Marín-López MA, Jumde G, Casanova K, Bonet J, Fornes O, Fernandez-Fuentes N, Oliva B. SPServer: split-statistical potentials for the analysis of protein structures and protein-protein interactions. ***BMC Bioinformatics*.** 2021; 22(1):4. DOI: 10.1186/s12859-020-03770-5



7. BIBLIOGRAPHY



1. NCBI. GRCh38.p13 [Internet]. GRCh38.p13 - Genome - Assembly - NCBI. 2019 [cited 2020 Nov 16]. Available from: https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.39
2. Yadav SP. The Wholeness in Suffix -omics, -omes, and the Word Om. J Biomol Tech. 2007 Dec;18(5):277.
3. Ghiassian SD. Network Medicine: A Network-based Approach to Human Diseases. [Boston]: Northeastern University; 2015.
4. Stumpf MPH, Thorne T, de Silva E, Stewart R, An HJ, Lappe M, et al. Estimating the size of the human interactome. Proc Natl Acad Sci U S A. 2008 May 13;105(19):6959–64.
5. Aloy P, Russell RB. Ten thousand interactions for the molecular biologist. Nat Biotechnol. 2004 Oct;22(10):1317–21.
6. Garcia-Garcia J, Bonet J, Guney E, Fornes O, Planas J, Oliva B. Networks of Protein-Protein Interactions: From Uncertainty to Molecular Details. Molecular Informatics. 2012 May;31(5):342–62.
7. Braun P, Tasan M, Dreze M, Barrios-Rodiles M, Lemmens I, Yu H, et al. An experimentally derived confidence score for binary protein-protein interactions. Nat Methods. 2009 Jan;6(1):91–7.
8. Peng X, Wang J, Peng W, Wu F-X, Pan Y. Protein-protein interactions: detection, reliability assessment and applications. Briefings in Bioinformatics. 2017 01;18(5):798–819.
9. Mahdavi MA, Lin Y-H. False positive reduction in protein-protein interaction predictions using gene ontology annotations. BMC Bioinformatics. 2007 Jul 23;8:262.
10. Costes SV, Daelemans D, Cho EH, Dobbin Z, Pavlakis G, Lockett S. Automatic and Quantitative Measurement of Protein-Protein Colocalization in Live Cells. Biophys J. 2004 Jun;86(6):3993–4003.
11. De Las Rivas J, Fontanillo C. Protein-protein interactions essentials: key concepts to building and analyzing interactome networks. PLoS computational biology. 2010 Jun 24;6(6):e1000807.
12. Valencia A, Pazos F. Computational methods for the prediction of protein interactions. Curr Opin Struct Biol. 2002 Jun;12(3):368–73.
13. Aloy P, Russell RB. Structural systems biology: modelling protein

interactions. Nature Reviews Molecular Cell Biology. 2006 Mar;7(3):188–97.

14. Liu Q, Remmelzwaal S, Heck AJR, Akhmanova A, Liu F. Facilitating identification of minimal protein binding domains by cross-linking mass spectrometry. Sci Rep [Internet]. 2017 Oct 18 [cited 2020 Dec 16];7. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5647383/

15. Boxem M, Maliga Z, Klitgord N, Li N, Lemmens I, Mana M, et al. A protein domain-based interactome network for C. elegans early embryogenesis. Cell. 2008 Aug 8;134(3):534–45.

16. Waaijers S, Koorman T, Kerver J, Boxem M. Identification of human protein interaction domains using an ORFeome-based yeast two-hybrid fragment library. J Proteome Res. 2013 Jul 5;12(7):3181–92.

17. Chatham JC, Blackband SJ. Nuclear magnetic resonance spectroscopy and imaging in animal research. ILAR J. 2001;42(3):189–208.

18. Carter R, Luchini A, Liotta L, Haymond A. Next-Generation Techniques for Determination of Protein-Protein Interactions: Beyond the Crystal Structure. Curr Pathobiol Rep. 2019 Sep 1;7(3):61–71.

19. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Research. 2000 Jan 1;28(1):235–42.

20. Mosca R, Céol A, Stein A, Olivella R, Aloy P. 3did: a catalog of domain-based interactions of known three-dimensional structure. Nucleic Acids Res. 2014 Jan;42(Database issue):D374-379.

21. Kuhlman B, Bradley P. Advances in protein structure prediction and design. Nat Rev Mol Cell Biol. 2019 Nov;20(11):681–97.

22. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997 Sep 1;25(17):3389–402.

23. Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D,

Shen M-Y, et al. Comparative protein structure modeling using Modeller. Curr Protoc Bioinformatics. 2006 Oct;Chapter 5:Unit-5.6.

24. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, et al. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. 2018 Jul 2;46(W1):W296–303.

25. Martí-Renom MA, Stuart AC, Fiser A, Sánchez R, Melo F, Sali A. Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct. 2000;29:291–325.

26. Aloy P, Ceulemans H, Stark A, Russell RB. The relationship between sequence and interaction divergence in proteins. J Mol Biol. 2003 Oct 3;332(5):989–98.

27. Meseguer A, Dominguez L, Bota PM, Aguirre-Plans J, Bonet J, Fernandez-Fuentes N, et al. Using collections of structural models to predict changes of binding affinity caused by mutations in protein-protein interactions. Protein Sci. 2020 Oct;29(10):2112–30.

28. Bonvin AMJJ. Flexible protein-protein docking. Curr Opin Struct Biol. 2006 Apr;16(2):194–200.

29. Pagadala NS, Syed K, Tuszynski J. Software for molecular docking: a review. Biophys Rev. 2017 Apr;9(2):91–102.

30. Lensink MF, Nadzirin N, Velankar S, Wodak SJ. Modeling protein-protein, protein-peptide, and protein-oligosaccharide complexes: CAPRI 7th edition. Proteins. 2020 Aug;88(8):916–38.

31. Pierce BG, Wiehe K, Hwang H, Kim B-H, Vreven T, Weng Z. ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers. Bioinformatics. 2014 Jun 15;30(12):1771–3.

32. Segura J, Marín-López MA, Jones PF, Oliva B, Fernandez-Fuentes N. VORFFIP-driven dock: V-D2OCK, a fast and accurate protein docking strategy. PLoS One. 2015;10(3):e0118107.

33. van Zundert GCP, Rodrigues JPGLM, Trellet M, Schmitz C, Kastritis PL, Karaca E, et al. The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. J Mol Biol. 2016

Feb 22;428(4):720–5.

34. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem. 2010 Jan 30;31(2):455–61.

35. Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J. Critical assessment of methods of protein structure prediction (CASP)-Round XIII. Proteins. 2019 Oct 7;

36. AlQuraishi M. AlphaFold at CASP13. Bioinformatics. 2019 Nov 1;35(22):4862–5.

37. Aguirre-Plans J, Meseguer A, Molina-Fernandez R, Marín-López MA, Jumde G, Casanova K, et al. SPServer: split-statistical potentials for the analysis of protein structures and protein-protein interactions. BMC Bioinformatics. 2021 Jan 6;22(1):4.

38. Kryshtafovych A, Monastyrskyy B, Fidelis K, Schwede T, Tramontano A. Assessment of model accuracy estimations in CASP12. Proteins. 2018;86 Suppl 1:345–60.

39. Cheng J, Choe M-H, Elofsson A, Han K-S, Hou J, Maghrabi AHA, et al. Estimation of model accuracy in CASP13. Proteins. 2019 Jul 2;

40. Fornes O, Garcia-Garcia J, Bonet J, Oliva B. On the use of knowledge-based potentials for the evaluation of models of protein-protein, protein-DNA, and protein-RNA interactions. In: Advances in Protein Chemistry and Structural Biology. Elsevier; 2014. p. 77–120.

41. Sippl MJ. Knowledge-based potentials for proteins. Curr Opin Struct Biol. 1995 Apr;5(2):229–35.

42. Wiederstein M, Sippl MJ. ProSA-web: Interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Research. 2007;35(SUPPL.2):407–10.

43. Conway P, DiMaio F. Improving hybrid statistical and physical forcefields through local structure enumeration. Protein Science. 2016;25:1525–34.

44. Olechnovič K, Venclovas Č. VoroMQA web server for assessing three-dimensional structures of proteins and protein complexes.

Nucleic acids research. 2019 Jul 2;47(W1):W437–42.

45. Melo F, Devos D, Depiereux E, Feytmans E. ANOLEA: a www server to assess protein structures. Proceedings / . International Conference on Intelligent Systems for Molecular Biology ; ISMB International Conference on Intelligent Systems for Molecular Biology. 1997;5:187–90.

46. Eisenberg D, Lüthy R, Bowie JU. VERIFY3D: Assessment of protein models with three-dimensional profiles. Methods in Enzymology. 1997;277:396–404.

47. Benkert P, Biasini M, Schwede T. Toward the estimation of the absolute quality of individual protein structure models. Bioinformatics (Oxford, England). 2011 Feb 1;27(3):343–50.

48. Uziela K, Hurtado DM, Shu N, Wallner B, Elofsson A. ProQ3D: Improved model quality assessments using deep learning. Bioinformatics. 2017;33(10):1578–80.

49. Maghrabi AHA, McGuffin LJ. ModFOLD6: an accurate web server for the global and local quality estimation of 3D protein models. Nucleic acids research. 2017;45(W1):W416–21.

50. Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Cech M, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res. 2018 Jul 2;46(W1):W537–44.

51. Licata L, Orchard S. The MIntAct Project and Molecular Interaction Databases. Methods Mol Biol. 2016;1415:55–69.

52. Oughtred R, Stark C, Breitkreutz B-J, Rust J, Boucher L, Chang C, et al. The BioGRID interaction database: 2019 update. Nucleic Acids Res. 2019 Jan 8;47(D1):D529–41.

53. Clerc O, Deniaud M, Vallet SD, Naba A, Rivet A, Perez S, et al. MatrixDB: integration of new data with a focus on glycosaminoglycan interactions. Nucleic Acids Res. 2019 Jan 8;47(D1):D376–81.

54. Orchard S, Kerrien S, Abbani S, Aranda B, Bhate J, Bidwell S, et al. Protein interaction data curation: the International Molecular

Exchange (IMEx) consortium. Nat Methods. 2012 Apr;9(4):345–50.

55. McDowall MD, Scott MS, Barton GJ. PIPs: human protein-protein interaction prediction database. Nucleic Acids Res. 2009 Jan;37(Database issue):D651-656.

56. Zhang QC, Petrey D, Garzón JI, Deng L, Honig B. PrePPI: a structure-informed database of protein-protein interactions. Nucleic Acids Res. 2013 Jan;41(Database issue):D828-833.

57. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019 08;47(D1):D607–13.

58. Alonso-López D, Campos-Laborie FJ, Gutiérrez MA, Lambourne L, Calderwood MA, Vidal M, et al. APID database: redefining protein-protein interaction experimental evidences and binary interactomes. Database (Oxford). 2019 Jan 1;2019.

59. Kamburov A, Stelzl U, Lehrach H, Herwig R. The ConsensusPathDB interaction database: 2013 update. Nucleic Acids Res. 2013 Jan;41(Database issue):D793-800.

60. Li T, Wernersson R, Hansen RB, Horn H, Mercer J, Slodkowicz G, et al. A scored human protein-protein interaction network to catalyze genomic interpretation. Nat Methods. 2017;14(1):61–4.

61. Kotlyar M, Pastrello C, Malik Z, Jurisica I. IID 2018 update: context-specific physical protein-protein interactions in human, model organisms and domesticated species. Nucleic Acids Res. 2019 08;47(D1):D581–9.

62. Garcia-Garcia J, Guney E, Aragues R, Planas-Iglesias J, Oliva B. Biana: a software framework for compiling biological interactions and analyzing networks. BMC Bioinformatics. 2010 Jan 27;11(1):56.

63. Kamburov A, Stelzl U, Herwig R. IntScore: a web tool for confidence scoring of biological interactions. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W140-146.

64. Alanis-Lobato G, Andrade-Navarro MA, Schaefer MH. HIPPIE v2.0:

enhancing meaningfulness and reliability of protein-protein interaction networks. Nucleic Acids Res. 2017 04;45(D1):D408–14.

65. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015 May 8;348(6235):648–60.

66. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008 Jul;5(7):621–8.

67. Razick S, Magklaras G, Donaldson IM. iRefIndex: a consolidated protein interaction database with provenance. BMC Bioinformatics. 2008 Sep 30;9:405.

68. Piñero J. Computational approaches and resources to support translational research in human diseases. [Barcelona]: Universitat Pompeu Fabra; 2015.

69. Jeong H, Mason SP, Barabási AL, Oltvai ZN. Lethality and centrality in protein networks. Nature. 2001 May 3;411(6833):41–2.

70. He X, Zhang J. Why Do Hubs Tend to Be Essential in Protein Networks? PLoS Genet [Internet]. 2006 Jun [cited 2020 Nov 16];2(6). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1473040/

71. Barabási A-L, Oltvai ZN. Network biology: understanding the cell’s functional organization. Nat Rev Genet. 2004 Feb;5(2):101–13.

72. Liu C, Ma Y, Zhao J, Nussinov R, Zhang Y-C, Cheng F, et al. Computational network biology: Data, models, and applications. Physics Reports. 2020 Mar 3;846:1–66.

73. Blokh D, Segev D, Sharan R. The approximability of shortest path-based graph orientations of protein-protein interaction networks. J Comput Biol. 2013 Dec;20(12):945–57.

74. Silverbush D, Sharan R. Network orientation via shortest paths. Bioinformatics. 2014 May 15;30(10):1449–55.

75. Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabási A-L. The human disease network. Proc Natl Acad Sci USA. 2007 May 22;104(21):8685–90.

76. Barabási A-L, Gulbahce N, Loscalzo J. Network Medicine: A Network-based Approach to Human Disease. Nat Rev Genet. 2011 Jan;12(1):56–68.

77. Menche J, Sharma A, Kitsak M, Ghiassian S, Vidal M, Loscalzo J, et al. Uncovering disease-disease relationships through the incomplete human interactome. Science. 2015 Feb 20;347(6224):1257601.

78. Guney E, Oliva B. Exploiting Protein-Protein Interaction Networks for Genome-Wide Disease-Gene Prioritization. PLoS One. 2012 Sep 21;7(9).

79. Guney E, Menche J, Vidal M, Barábasi A-L. Network-based in silico drug efficacy screening. Nat Commun. 2016 Feb 1;7:10331.

80. Koschützki D, Schreiber F. Centrality Analysis Methods for Biological Networks and Their Application to Gene Regulatory Networks. Gene Regul Syst Bio. 2008 May 15;2:193–201.

81. Ma H-W, Zeng A-P. The connectivity structure, giant strong component and centrality of metabolic networks. Bioinformatics. 2003 Jul 22;19(11):1423–30.

82. Joy MP, Brock A, Ingber DE, Huang S. High-betweenness proteins in the yeast protein interaction network. J Biomed Biotechnol. 2005 Jun 30;2005(2):96–103.

83. Piñero J, Berenstein A, Gonzalez-Perez A, Chernomoretz A, Furlong LI. Uncovering disease mechanisms through network biology in the era of Next Generation Sequencing. Sci Rep. 2016 Apr 15;6:24570.

84. Chen J, Yuan B. Detecting functional modules in the yeast protein-protein interaction network. Bioinformatics. 2006 Sep 15;22(18):2283–90.

85. Huttlin EL, Ting L, Bruckner RJ, Gebreab F, Gygi MP, Szpyt J, et al. The BioPlex Network: A Systematic Exploration of the Human Interactome. Cell. 2015 Jul 16;162(2):425–40.

86. Nepusz T, Yu H, Paccanaro A. Detecting overlapping protein complexes in protein-protein interaction networks. Nature Methods. 2012 May;9(5):471–2.

87. Albert R, Jeong H, Barabási A-L. Error and attack tolerance of complex networks. Nature. 2000 Jul;406(6794):378–82.

88. Morohashi M, Winn AE, Borisuk MT, Bolouri H, Doyle J, Kitano H. Robustness as a measure of plausibility in models of biochemical networks. J Theor Biol. 2002 May 7;216(1):19–30.

89. Agrawal AA. Phenotypic plasticity in the interactions and evolution of species. Science. 2001 Oct 12;294(5541):321–6.

90. Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. Nature. 1999 Dec 2;402(6761 Suppl):C47-52.

91. Sahni N, Yi S, Taipale M, Fuxman Bass JI, Coulombe-Huntington J, Yang F, et al. Widespread macromolecular interaction perturbations in human genetic disorders. Cell. 2015 Apr 23;161(3):647–60.

92. Rubio-Perez C, Guney E, Aguilar D, Piñero J, Garcia-Garcia J, Iadarola B, et al. Genetic and functional characterization of disease associations explains comorbidity. Sci Rep. 2017 24;7(1):6207.

93. Lemke N, Herédia F, Barcellos CK, Dos Reis AN, Mombach JCM. Essentiality and damage in metabolic networks. Bioinformatics. 2004 Jan 1;20(1):115–9.

94. Guney E, Oliva B. Analysis of the robustness of network-based disease-gene prioritization methods reveals redundancy in the human interactome and functional diversity of disease-genes. PLoS ONE. 2014;9(4):e94686.

95. Piñero J, Gonzalez-Perez A, Guney E, Aguirre-Plans J, Sanz F, Oliva B, et al. Network, Transcriptomic and Genomic Features Differentiate Genes Relevant for Drug Response. Front Genet. 2018;9:412.

96. Aguirre-Plans J, Piñero J, Aguilar D, Guney E, Sanz F, Furlong LI, et al. A Review of Network Medicine Approaches to Understand Comorbidity. In: Advances in Medicine and Biology Volume 158. Leon V. Berhardt. New York: Nova Science Publishers; p. 1–42.

97. Guney E. Role of network topology based methods in discovering novel gene-phenotype associations. [Barcelona]: Universitat Pompeu Fabra; 2012.

98. Frazer KA, Murray SS, Schork NJ, Topol EJ. Human genetic variation and its contribution to complex traits. Nat Rev Genet. 2009 Apr;10(4):241–51.

99. Barton NH, Etheridge AM, Véber A. The infinitesimal model: Definition, derivation, and implications. Theor Popul Biol. 2017;118:50–73.

100. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell. 2017 Jun 15;169(7):1177–86.

101. Broeckel U, Schork NJ. Identifying genes and genetic variation underlying human diseases and complex phenotypes via recombination mapping. J Physiol. 2004 Jan 1;554(Pt 1):40–5.

102. Stranger BE, Stahl EA, Raj T. Progress and promise of genome-wide association studies for human complex trait genetics. Genetics. 2011 Feb;187(2):367–83.

103. Amberger JS, Bocchini CA, Scott AF, Hamosh A. OMIM.org: leveraging knowledge across phenotype-gene relationships. Nucleic Acids Res. 2019 Jan 8;47(D1):D1038–43.

104. Tamborero D, Rubio-Perez C, Deu-Pons J, Schroeder MP, Vivancos A, Rovira A, et al. Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. Genome Med. 2018 28;10(1):25.

105. Rehm HL, Berg JS, Brooks LD, Bustamante CD, Evans JP, Landrum MJ, et al. ClinGen--the Clinical Genome Resource. N Engl J Med. 2015 04;372(23):2235–42.

106. Davis AP, Grondin CJ, Johnson RJ, Sciaky D, Wiegers J, Wiegers TC, et al. Comparative Toxicogenomics Database (CTD): update 2021. Nucleic Acids Res. 2021 Jan 8;49(D1):D1138–43.

107. Pavan S, Rommel K, Mateo Marquina ME, Höhn S, Lanneau V, Rath A. Clinical Practice Guidelines for Rare Diseases: The Orphanet Database. PLoS One [Internet]. 2017 Jan 18 [cited 2019 Oct 24];12(1). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5242437/

108. Gutiérrez-Sacristán A, Grosdidier S, Valverde O, Torrens M, Bravo

À, Piñero J, et al. PsyGeNET: a knowledge platform on psychiatric disorders and their genes. Bioinformatics. 2015 Sep 15;31(18):3075–7.

109. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019 Jan 8;47(D1):D506–15.

110. Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, Ronzano F, Centeno E, Sanz F, et al. The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res. 2020 Jan 8;48(D1):D845–55.

111. Girvan M, Newman MEJ. Community structure in social and biological networks. PNAS. 2002 Jun 11;99(12):7821–6.

112. Alon U. Biological Networks: The Tinkerer as an Engineer. Science. 2003 Sep 26;301(5641):1866–7.

113. Kitano H. Biological robustness. Nat Rev Genet. 2004 Nov;5(11):826–37.

114. Gandhi TKB, Zhong J, Mathivanan S, Karthick L, Chandrika KN, Mohan SS, et al. Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. Nat Genet. 2006 Mar;38(3):285–93.

115. Xu J, Li Y. Discovering disease-genes by topological features in human protein-protein interaction network. Bioinformatics. 2006 Nov 15;22(22):2800–5.

116. Bauer-Mehren A, Bundschus M, Rautschka M, Mayer MA, Sanz F, Furlong LI. Gene-disease network analysis reveals functional modules in mendelian, complex and environmental diseases. PLoS ONE. 2011;6(6):e20284.

117. Park J, Lee D-S, Christakis NA, Barabási A-L. The impact of cellular networks on disease comorbidity. Mol Syst Biol. 2009 Apr 7;5:262.

118. Vlaic S, Conrad T, Tokarski-Schnelle C, Gustafsson M, Dahmen U, Guthke R, et al. ModuleDiscoverer: Identification of regulatory modules in protein-protein interaction networks. Sci Rep. 2018 11;8(1):433.

119. Ghiassian SD, Menche J, Barabási A-L. A DIseAse MOdule

Detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. PLoS Comput Biol. 2015 Apr;11(4):e1004120.

120. Oti M, Snel B, Huynen MA, Brunner HG. Predicting disease genes using protein-protein interactions. J Med Genet. 2006 Aug;43(8):691–8.

121. Clauset A, Newman MEJ, Moore C. Finding community structure in very large networks. Phys Rev E Stat Nonlin Soft Matter Phys. 2004 Dec;70(6 Pt 2):066111.

122. Ahn J, Lee DH, Yoon Y, Yeu Y, Park S. Improved method for protein complex detection using bottleneck proteins. BMC Med Inform Decis Mak. 2013 Apr 5;13(Suppl 1):S5.

123. Fortunato S. Community detection in graphs. Physics Reports. 2010 Feb 1;486(3):75–174.

124. Wang R-S, Loscalzo J. Network-Based Disease Module Discovery by a Novel Seed Connector Algorithm with Pathobiological Implications. J Mol Biol. 2018 Sep 14;430(18 Pt A):2939–50.

125. Barrenäs F, Chavali S, Alves AC, Coin L, Jarvelin M-R, Jörnsten R, et al. Highly interconnected genes in disease-specific networks are enriched for disease-associated polymorphisms. Genome Biology. 2012 Jun 15;13(6):R46.

126. Choobdar S, Ahsen ME, Crawford J, Tomasoni M, Fang T, Lamparter D, et al. Assessment of network module identification across complex diseases. Nat Methods. 2019 Sep;16(9):843–52.

127. Cao M, Zhang H, Park J, Daniels NM, Crovella ME, Cowen LJ, et al. Going the distance for protein function prediction: a new distance metric for protein interaction networks. PLoS One. 2013;8(10):e76339.

128. Cao M, Pietras CM, Feng X, Doroschak KJ, Schaffner T, Park J, et al. New directions for diffusion-based network prediction of protein function: incorporating pathways with confidence. Bioinformatics. 2014 Jun 15;30(12):i219-227.

129. Bonifaci N, Berenguer A, Díez J, Reina O, Medina I, Dopazo J, et al.

Biological processes, properties and molecular wiring diagrams of candidate low-penetrance breast cancer susceptibility genes. BMC Med Genomics. 2008 Dec 18;1:62.

130. Heiser LM, Wang NJ, Talcott CL, Laderoute KR, Knapp M, Guan Y, et al. Integrated analysis of breast cancer cell lines reveals unique signaling pathways. Genome Biol. 2009;10(3):R31.

131. Chuang H-Y, Lee E, Liu Y-T, Lee D, Ideker T. Network-based classification of breast cancer metastasis. Mol Syst Biol [Internet]. 2007 Oct 16 [cited 2021 Apr 15];3. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2063581/

132. Taylor IW, Linding R, Warde-Farley D, Liu Y, Pesquita C, Faria D, et al. Dynamic modularity in protein interaction networks predicts breast cancer outcome. Nat Biotechnol. 2009 Feb;27(2):199–204.

133. Nibbe RK, Markowitz S, Myeroff L, Ewing R, Chance MR. Discovery and Scoring of Protein Interaction Subnetworks Discriminative of Late Stage Human Colon Cancer. Mol Cell Proteomics. 2009 Apr;8(4):827–45.

134. Chang W, Ma L, Lin L, Gu L, Liu X, Cai H, et al. Identification of novel hub genes associated with liver metastasis of gastric cancer. Int J Cancer. 2009 Dec 15;125(12):2844–53.

135. Ergün A, Lawrence CA, Kohanski MA, Brennan TA, Collins JJ. A network biology approach to prostate cancer. Mol Syst Biol. 2007 Feb 13;3:82.

136. Lee E, Jung H, Radivojac P, Kim J-W, Lee D. Analysis of AML Genes in Dysregulated Molecular Networks. Summit on Translat Bioinforma. 2009 Mar 1;2009:1–18.

137. Sharma A, Menche J, Huang CC, Ort T, Zhou X, Kitsak M, et al. A disease module in the interactome explains disease heterogeneity, drug response and captures novel pathways and genes in asthma. Hum Mol Genet. 2015 Jun 1;24(11):3005–20.

138. Maiorino E, Baek SH, Guo F, Zhou X, Kothari PH, Silverman EK, et al. Discovering the genes mediating the interactions between chronic respiratory diseases in the human interactome. Nat

Commun. 2020 Feb 10;11(1):811.

139. Gysi DM, Valle Í do, Zitnik M, Ameli A, Gan X, Varol O, et al. Network medicine framework for identifying drug-repurposing opportunities for COVID-19. PNAS [Internet]. 2021 May 11 [cited 2021 Apr 28];118(19). Available from: https://www.pnas.org/content/118/19/e2025581118

140. Gordon DE, Jang GM, Bouhaddou M, Xu J, Obernier K, White KM, et al. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. Nature. 2020 Jul;583(7816):459–68.

141. Feinstein AR. The pre-therapeutic classification of co-morbidity in chronic disease. J Chronic Dis. 1970 Dec;23(7):455–68.

142. Valderas JM, Starfield B, Sibbald B, Salisbury C, Roland M. Defining comorbidity: implications for understanding health and health services. Ann Fam Med. 2009 Aug;7(4):357–63.

143. Ording AG, Sørensen HT. Concepts of comorbidities, multiple morbidities, complications, and their clinical epidemiologic analogs. Clin Epidemiol. 2013 Jul 1;5:199–203.

144. Radner H, Yoshida K, Smolen JS, Solomon DH. Multimorbidity and rheumatic conditions-enhancing the concept of comorbidity. Nat Rev Rheumatol. 2014 Apr;10(4):252–6.

145. Hu JX, Thomas CE, Brunak S. Network biology concepts in complex disease comorbidities. Nat Rev Genet. 2016;17(10):615–29.

146. Ideker T, Krogan NJ. Differential network biology. Mol Syst Biol. 2012 Jan 17;8:565.

147. Furlong LI. Human diseases through the lens of network biology. Trends Genet. 2013 Mar;29(3):150–9.

148. Roque FS, Jensen PB, Schmock H, Dalgaard M, Andreatta M, Hansen T, et al. Using electronic patient records to discover disease correlations and stratify patient cohorts. PLoS Comput Biol. 2011 Aug;7(8):e1002141.

149. Faner R, Gutiérrez-Sacristán A, Castro-Acosta A, Grosdidier S, Gan W, Sánchez-Mayor M, et al. Molecular and clinical diseasome of comorbidities in exacerbated COPD patients. Eur Respir J. 2015

Oct;46(4):1001–10.

150. Suthram S, Dudley JT, Chiang AP, Chen R, Hastie TJ, Butte AJ. Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. PLoS Comput Biol. 2010 Feb 5;6(2):e1000662.

151. van Driel MA, Bruggeman J, Vriend G, Brunner HG, Leunissen JAM. A text-mining analysis of the human phenome. Eur J Hum Genet. 2006 May;14(5):535–42.

152. Ramos EM, Hoffman D, Junkins HA, Maglott D, Phan L, Sherry ST, et al. Phenotype-Genotype Integrator (PheGenI): synthesizing genome-wide association study (GWAS) data with existing genomic resources. Eur J Hum Genet. 2014 Jan;22(1):144–7.

153. Hidalgo CA, Blumm N, Barabási A-L, Christakis NA. A Dynamic Network Approach for the Study of Human Phenotypes. PLoS Comput Biol [Internet]. 2009 Apr 10 [cited 2019 May 31];5(4). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2661364/

154. Aguirre-Plans J, Piñero J, Sanz F, Furlong LI, Fernandez-Fuentes N, Oliva B, et al. GUILDify v2.0: A Tool to Identify Molecular Networks Underlying Human Diseases, Their Comorbidities and Their Druggable Targets. J Mol Biol. 2019 Jun 14;431(13):2477–84.

155. Berry M, Brightling C, Pavord I, Wardlaw A. TNF-alpha in asthma. Curr Opin Pharmacol. 2007 Jun;7(3):279–82.

156. Catal F, Mete E, Tayman C, Topal E, Albayrak A, Sert H. A human monoclonal anti-TNF alpha antibody (adalimumab) reduces airway inflammation and ameliorates lung histology in a murine model of acute asthma. Allergol Immunopathol (Madr). 2015 Feb;43(1):14–8.

157. Sieberts SK, Zhu F, García-García J, Stahl E, Pratap A, Pandey G, et al. Crowdsourced assessment of common genetic contribution to predicting anti-TNF treatment response in rheumatoid arthritis. Nat Commun. 2016 23;7:12460.

158. Anecchino C, Fanizza C, Marino V, Romero M, DOSE Study Group. Drug outcome survey to evaluate anti-TNF treatment in rheumatoid

arthritis: an Italian observational study (the DOSE study). Clin Exp Rheumatol. 2015 Dec;33(6):779–87.

159. Umičević Mirkov M, Cui J, Vermeulen SH, Stahl EA, Toonen EJM, Makkinje RR, et al. Genome-wide association analysis of anti-TNF drug response in patients with rheumatoid arthritis. Ann Rheum Dis. 2013 Aug;72(8):1375–81.

160. de Punder YMR, Fransen J, Kievit W, Houtman PM, Visser H, van de Laar MAFJ, et al. The prevalence of clinical remission in RA patients treated with anti-TNF: results from the Dutch Rheumatoid Arthritis Monitoring (DREAM) registry. Rheumatology (Oxford). 2012 Sep;51(9):1610–7.

161. Lequerré T, Farran É, Ménard J-F, Kozyreff-Meurice M, Vandhuick T, Tharasse C, et al. Switching from an anti-TNF monoclonal antibody to soluble TNF-receptor yields better results than vice versa: An observational retrospective study of 72 rheumatoid arthritis switchers. Joint Bone Spine. 2015 Oct;82(5):330–7.

162. Rolfes MC, Juhn YJ, Wi C-I, Sheen YH. Asthma and the Risk of Rheumatoid Arthritis: An Insight into the Heterogeneity and Phenotypes of Asthma. Tuberc Respir Dis (Seoul). 2017 Apr;80(2):113–35.

163. GBD 2015 Chronic Respiratory Disease Collaborators. Global, regional, and national deaths, prevalence, disability-adjusted life years, and years lived with disability for chronic obstructive pulmonary disease and asthma, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet Respir Med. 2017 Sep;5(9):691–706.

164. Paller AS, Spergel JM, Mina-Osorio P, Irvine AD. The atopic march and atopic multimorbidity: Many trajectories, many pathways. J Allergy Clin Immunol. 2019 Jan;143(1):46–55.

165. Ballardini N, Kull I, Lind T, Hallner E, Almqvist C, Ostblom E, et al. Development and comorbidity of eczema, asthma and rhinitis to age 12: data from the BAMSE birth cohort. Allergy. 2012 Apr;67(4):537–44.

166. Garcia-Aymerich J, Benet M, Saeys Y, Pinart M, Basagaña X, Smit HA, et al. Phenotyping asthma, rhinitis and eczema in MeDALL population-based birth cohorts: an allergic comorbidity cluster. Allergy. 2015 Aug;70(8):973–84.

167. Pinart M, Benet M, Annesi-Maesano I, von Berg A, Berdel D, Carlsen KCL, et al. Comorbidity of eczema, rhinitis, and asthma in IgE-sensitised and non-IgE-sensitised children in MeDALL: a population-based cohort study. Lancet Respir Med. 2014 Feb;2(2):131–40.

168. Bousquet J, Anto JM, Wickman M, Keil T, Valenta R, Haahtela T, et al. Are allergic multimorbidities and IgE polysensitization associated with the persistence or re-occurrence of foetal type 2 signalling? The MeDALL hypothesis. Allergy. 2015 Sep;70(9):1062–78.

169. Bousquet J, Anto JM, Akdis M, Auffray C, Keil T, Momas I, et al. Paving the way of systems biology and precision medicine in allergic diseases: the MeDALL success story: Mechanisms of the Development of ALLergy; EU FP7-CP-IP; Project No: 261357; 2010-2015. Allergy. 2016;71(11):1513–25.

170. Koppelman GH, Nawijn MC. Recent advances in the epigenetics and genomics of asthma. Curr Opin Allergy Clin Immunol. 2011 Oct;11(5):414–9.

171. Aguilar D, Pinart M, Koppelman GH, Saeys Y, Nawijn MC, Postma DS, et al. Computational analysis of multimorbidity between asthma, eczema and rhinitis. PLoS ONE. 2017;12(6):e0179125.

172. Aguilar D, Lemonnier N, Koppelman GH, Melén E, Oliva B, Pinart M, et al. Understanding allergic multimorbidity within the non-eosinophilic interactome. PLoS One. 2019;14(11):e0224448.

173. Catalá-López F, Crespo-Facorro B, Vieta E, Valderas JM, Valencia A, Tabarés-Seisdedos R. Alzheimer's disease and cancer: current epidemiological evidence for a mutual protection. Neuroepidemiology. 2014;42(2):121–2.

174. Catalá-López F, Suárez-Pinilla M, Suárez-Pinilla P, Valderas JM, Gómez-Beneyto M, Martinez S, et al. Inverse and direct cancer

comorbidity in people with central nervous system disorders: a meta-analysis of cancer incidence in 577,013 participants of 50 observational studies. Psychother Psychosom. 2014;83(2):89–105.

175. Li Y, Agarwal P, Rajagopalan D. A global pathway crosstalk network. Bioinformatics. 2008 Jun 15;24(12):1442–7.

176. Ko Y, Cho M, Lee J-S, Kim J. Identification of disease comorbidity through hidden molecular mechanisms. Sci Rep. 2016 19;6:39433.

177. Gottesman II, Gould TD. The endophenotype concept in psychiatry: etymology and strategic intentions. Am J Psychiatry. 2003 Apr;160(4):636–45.

178. Loscalzo J, Kohane I, Barabasi A-L. Human disease classification in the postgenomic era: a complex systems approach to human pathobiology. Mol Syst Biol. 2007;3:124.

179. Ghiassian SD, Menche J, Chasman DI, Giulianini F, Wang R, Ricchiuto P, et al. Endophenotype Network Models: Common Core of Complex Diseases. Sci Rep. 2016 09;6:27414.

180. Aguirre-Plans J, Piñero J, Menche J, Sanz F, Furlong LI, Schmidt HHHW, et al. Proximal Pathway Enrichment Analysis for Targeting Comorbid Diseases via Network Endopharmacology. Pharmaceuticals (Basel). 2018 Jun 22;11(3).

181. Maron BA, Altucci L, Balligand J-L, Baumbach J, Ferdinandy P, Filetti S, et al. A global network for network medicine. NPJ Syst Biol Appl. 2020 Aug 31;6(1):29.

182. Hopkins AL. Network pharmacology: the next paradigm in drug discovery. Nat Chem Biol. 2008 Nov;4(11):682–90.

183. Berger SI, Iyengar R. Network analyses in systems pharmacology. Bioinformatics. 2009 Oct 1;25(19):2466–72.

184. Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J. Clinical development success rates for investigational drugs. Nat Biotechnol. 2014 Jan;32(1):40–51.

185. Harrison RK. Phase II and phase III failures: 2013-2015. Nat Rev Drug Discov. 2016 Dec;15(12):817–8.

186. Santos R, Ursu O, Gaulton A, Bento AP, Donadi RS, Bologa CG, et

al. A comprehensive map of molecular drug targets. Nat Rev Drug Discov. 2017 Jan;16(1):19–34.

187. Duffus J. Glossary for chemists of terms used in toxicology (IUPAC Recommendations 1993). Pure and Applied Chemistry. 1993 Jan 1;65(9):2003–122.

188. Schenone M, Dančík V, Wagner BK, Clemons PA. Target identification and mechanism of action in chemical biology and drug discovery. Nat Chem Biol. 2013 Apr;9(4):232–40.

189. Öztürk H, Özgür A, Ozkirimli E. DeepDTA: deep drug-target binding affinity prediction. Bioinformatics. 2018 Sep 1;34(17):i821–9.

190. Wu Z, Li W, Liu G, Tang Y. Network-Based Methods for Prediction of Drug-Target Interactions. Front Pharmacol [Internet]. 2018 Oct 9 [cited 2021 Jan 19];9. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6189482/

191. Rognan D. Structure-Based Approaches to Target Fishing and Ligand Profiling. Mol Inform. 2010 Mar 15;29(3):176–87.

192. Liu X, Ouyang S, Yu B, Liu Y, Huang K, Gong J, et al. PharmMapper server: a web server for potential drug target identification using pharmacophore mapping approach. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W609-614.

193. Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK. Relating protein pharmacology by ligand chemistry. Nat Biotechnol. 2007 Feb;25(2):197–206.

194. Hamad S, Adornetto G, Naveja JJ, Chavan Ravindranath A, Raffler J, Campillos M. HitPickV2: a web server to predict targets of chemical compounds. Bioinformatics. 2019 Apr 1;35(7):1239–40.

195. Gong J, Cai C, Liu X, Ku X, Jiang H, Gao D, et al. ChemMapper: a versatile web server for exploring pharmacology and chemical structure association based on molecular 3D similarity method. Bioinformatics. 2013 Jul 15;29(14):1827–9.

196. Campillos M, Kuhn M, Gavin A-C, Jensen LJ, Bork P. Drug target identification using side-effect similarity. Science. 2008 Jul 11;321(5886):263–6.

197. Chen X, Yan CC, Zhang X, Zhang X, Dai F, Yin J, et al. Drug-target interaction prediction: databases, web servers and computational models. Brief Bioinform. 2016 Jul;17(4):696–712.

198. Nicola G, Liu T, Gilson M. Public Domain Databases for Medicinal Chemistry. J Med Chem. 2012 Aug 23;55(16):6987–7002.

199. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2018 Jan 4;46(D1):D1074–82.

200. Avram S, Bologa CG, Holmes J, Bocci G, Wilson TB, Nguyen D-T, et al. DrugCentral 2021 supports drug discovery and repositioning. Nucleic Acids Research. 2021 Jan 8;49(D1):D1160–9.

201. Freshour SL, Kiwala S, Cotto KC, Coffman AC, McMichael JF, Song JJ, et al. Integration of the Drug-Gene Interaction Database (DGIdb 4.0) with open crowdsource efforts. Nucleic Acids Res. 2021 Jan 8;49(D1):D1144–51.

202. Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Félix E, et al. ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Res. 2019 Jan 8;47(D1):D930–40.

203. Cheng F, Desai RJ, Handy DE, Wang R, Schneeweiss S, Barabási A-L, et al. Network-based approach to prediction and population-based validation of in silico drug repurposing. Nat Commun. 2018 Jul 12;9(1):2691.

204. Cheng F, Kovács IA, Barabási A-L. Network-based prediction of drug combinations. Nat Commun [Internet]. 2019 Mar 13 [cited 2021 Jan 15];10. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6416394/

205. Wang Y, Zhang S, Li F, Zhou Y, Zhang Y, Wang Z, et al. Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics. Nucleic Acids Res. 2020 Jan 8;48(D1):D1031–41.

206. Thorn CF, Klein TE, Altman RB. PharmGKB: The Pharmacogenomics Knowledge Base. Methods Mol Biol. 2013;1015:311–20.

207. Gilson MK, Liu T, Baitaluk M, Nicola G, Hwang L, Chong J. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic Acids Res. 2016 Jan 4;44(D1):D1045-1053.

208. Armstrong JF, Faccenda E, Harding SD, Pawson AJ, Southan C, Sharman JL, et al. The IUPHAR/BPS Guide to PHARMACOLOGY in 2020: extending immunopharmacology content and introducing the IUPHAR/MMV Guide to MALARIA PHARMACOLOGY. Nucleic Acids Res. 2020 Jan 8;48(D1):D1006–21.

209. Yamanishi Y, Araki M, Gutteridge A, Honda W, Kanehisa M. Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. Bioinformatics. 2008 Jul 1;24(13):i232–40.

210. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, et al. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D354–7.

211. Jeske L, Placzek S, Schomburg I, Chang A, Schomburg D. BRENDA in 2019: a European ELIXIR core data resource. Nucleic Acids Res. 2019 Jan 8;47(D1):D542–9.

212. Hecker N, Ahmed J, von Eichborn J, Dunkel M, Macha K, Eckert A, et al. SuperTarget goes quantitative: update on drug–target interactions. Nucleic Acids Res. 2012 Jan;40(Database issue):D1113–7.

213. Tanoli Z, Alam Z, Vähä-Koskela M, Ravikumar B, Malyutina A, Jaiswal A, et al. Drug Target Commons 2.0: a community platform for systematic analysis of drug–target interaction profiles. Database (Oxford) [Internet]. 2018 Sep 13 [cited 2021 Jan 18];2018. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6146131/

214. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem in 2021: new data content and improved web interfaces. Nucleic Acids Res. 2021 Jan 8;49(D1):D1388–95.

215. Szklarczyk D, Santos A, von Mering C, Jensen LJ, Bork P, Kuhn M.

STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data. Nucleic Acids Res. 2016 Jan 4;44(D1):D380-384.

216. Ashburn TT, Thor KB. Drug repositioning: identifying and developing new uses for existing drugs. Nat Rev Drug Discov. 2004 Aug;3(8):673–83.

217. Pushpakom S, Iorio F, Eyers PA, Escott KJ, Hopper S, Wells A, et al. Drug repurposing: progress, challenges and recommendations. Nat Rev Drug Discov. 2019;18(1):41–58.

218. Breckenridge A, Jacob R. Overcoming the legal and regulatory barriers to drug repurposing. Nat Rev Drug Discov. 2019;18(1):1–2.

219. Nosengo N. Can you teach old drugs new tricks? Nature. 2016 16;534(7607):314–6.

220. Lotfi Shahreza M, Ghadiri N, Mousavi SR, Varshosaz J, Green JR. A review of network-based approaches to drug repositioning. Brief Bioinform. 2018 Sep 28;19(5):878–92.

221. Zhou Y, Hou Y, Shen J, Huang Y, Martin W, Cheng F. Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. Cell Discov. 2020;6:14.

222. Cheng F, Liu C, Jiang J, Lu W, Li W, Liu G, et al. Prediction of drug-target interactions and drug repositioning via network-based inference. PLoS Comput Biol. 2012;8(5):e1002503.

223. Stathias V, Turner J, Koleti A, Vidovic D, Cooper D, Fazel-Najafabadi M, et al. LINCS Data Portal 2.0: next generation access point for perturbation-response signatures. Nucleic Acids Res. 2020 Jan 8;48(D1):D431–9.

224. Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, et al. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Cell. 2017 Nov 30;171(6):1437-1452.e17.

225. Corsello SM, Bittker JA, Liu Z, Gould J, McCarren P, Hirschman JE, et al. The Drug Repurposing Hub: a next-generation drug library and information resource. Nat Med. 2017 Apr 7;23(4):405–8.

226. Greene CS, Krishnan A, Wong AK, Ricciotti E, Zelaya RA,

Himmelstein DS, et al. Understanding multicellular function and disease with human tissue-specific networks. Nat Genet. 2015 Jun;47(6):569–76.

227. Jia J, Zhu F, Ma X, Cao Z, Cao ZW, Li Y, et al. Mechanisms of drug combinations: interaction and network perspectives. Nat Rev Drug Discov. 2009 Feb;8(2):111–28.

228. He B, Lu C, Zheng G, He X, Wang M, Chen G, et al. Combination therapeutics in complex diseases. J Cell Mol Med. 2016 Dec;20(12):2231–40.

229. Zimmermann GR, Lehár J, Keith CT. Multi-target therapeutics: when the whole is greater than the sum of the parts. Drug Discov Today. 2007 Jan;12(1–2):34–42.

230. Cascorbi I. Drug Interactions—Principles, Examples and Clinical Consequences. Dtsch Arztebl Int. 2012 Aug;109(33–34):546–56.

231. Madani Tonekaboni SA, Soltan Ghoraie L, Manem VSK, Haibe-Kains B. Predictive approaches for drug combination discovery in cancer. Brief Bioinform. 2018 Mar 1;19(2):263–76.

232. Bansal M, Yang J, Karan C, Menden MP, Costello JC, Tang H, et al. A community computational challenge to predict the activity of pairs of compounds. Nat Biotechnol. 2014 Dec;32(12):1213–22.

233. Huang H, Zhang P, Qu XA, Sanseau P, Yang L. Systematic prediction of drug combinations based on clinical side-effects. Sci Rep. 2014 Nov 24;4:7160.

234. Sun Y, Sheng Z, Ma C, Tang K, Zhu R, Wu Z, et al. Combining genomic and network characteristics for extended capability in predicting synergistic drugs for cancer. Nat Commun. 2015 Sep 28;6:8481.

235. Yang J, Tang H, Li Y, Zhong R, Wang T, Wong S, et al. DIGRE: Drug-Induced Genomic Residual Effect Model for Successful Prediction of Multidrug Effects. CPT Pharmacometrics Syst Pharmacol. 2015 Feb;4(2):e1.

236. Li X, Xu Y, Cui H, Huang T, Wang D, Lian B, et al. Prediction of synergistic anti-cancer drug combinations based on drug target

network and drug induced gene expression profiles. Artif Intell Med. 2017 Nov;83:35–43.

237. Yang M, Jaaks P, Dry J, Garnett M, Menden MP, Saez-Rodriguez J. Stratification and prediction of drug synergy based on target functional similarity. NPJ Syst Biol Appl. 2020 Jun 2;6(1):16.

238. Zagidullin B, Aldahdooh J, Zheng S, Wang W, Wang Y, Saad J, et al. DrugComb: an integrative cancer drug combination data portal. Nucleic Acids Res. 2019 Jul 2;47(W1):W43–51.

239. Liu H, Zhang W, Zou B, Wang J, Deng Y, Deng L. DrugCombDB: a comprehensive database of drug combinations toward the discovery of combinatorial therapy. Nucleic Acids Res. 2020 Jan 8;48(D1):D871–81.

240. Kurnat-Thoma E, Baranova A, Baird P, Brodsky E, Butte AJ, Cheema AK, et al. Recent Advances in Systems and Network Medicine: Meeting Report from the First International Conference in Systems and Network Medicine. Syst Med (New Rochelle). 2020 Feb 26;3(1):22–35.

241. Ma S, Earls JC, Eddy JA, Price ND. Using Integrative -omics Approaches in Network Medicine. In: Network Medicine: Complex Systems in Human Disease and Therapeutics. Cambridge, Massachusetts: Harvard University Press; 2017. p. 448.

242. Thiele I, Palsson BØ. A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat Protoc. 2010 Jan;5(1):93–121.

243. Iskar M, Zeller G, Blattmann P, Campillos M, Kuhn M, Kaminska KH, et al. Characterization of drug-induced transcriptional modules: towards drug repositioning and functional understanding. Mol Syst Biol. 2013;9:662.

244. Oldham WM, Oliveira RKF, Wang R-S, Opotowsky AR, Rubins DM, Hainer J, et al. Network Analysis to Risk Stratify Patients with Exercise Intolerance. Circ Res. 2018 Mar 16;122(6):864–76.

245. Silverman EK, Schmidt HHHW, Anastasiadou E, Altucci L, Angelini M, Badimon L, et al. Molecular networks in Network Medicine:

Development and applications. Wiley Interdiscip Rev Syst Biol Med. 2020 Nov;12(6):e1489.

246. Kivelä M, Arenas A, Barthelemy M, Gleeson JP, Moreno Y, Porter MA. Multilayer networks. Journal of Complex Networks. 2014 Sep 1;2(3):203–71.

247. Halu A, De Domenico M, Arenas A, Sharma A. The multiplex network of human diseases. NPJ Syst Biol Appl. 2019;5:15.

248. Kinsley AC, Rossi G, Silk MJ, VanderWaal K. Multilayer and Multiplex Networks: An Introduction to Their Use in Veterinary Epidemiology. Front Vet Sci. 2020;7:596.

249. Lee D-S, Park J, Kay KA, Christakis NA, Oltvai ZN, Barabási A-L. The implications of human metabolic network topology for disease comorbidity. Proc Natl Acad Sci USA. 2008 Jul 22;105(29):9880–5.

250. Cheng F, Lu W, Liu C, Fang J, Hou Y, Handy DE, et al. A genome-wide positioning systems network algorithm for in silico drug repurposing. Nat Commun. 2019 Aug 2;10(1):3476.

251. Jorba G, Aguirre-Plans J, Junet V, Segú-Vergés C, Ruiz JL, Pujol A, et al. In-silico simulated prototype-patients using TPMS technology to study a potential adverse effect of sacubitril and valsartan. PLoS One [Internet]. 2020 Feb 13 [cited 2021 Jan 22];15(2). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7018085/

252. Iborra-Egea O, Gálvez-Montón C, Roura S, Perea-Gil I, Prat-Vidal C, Soler-Botija C, et al. Mechanisms of action of sacubitril/valsartan on cardiac remodeling: a systems biology approach. NPJ Syst Biol Appl. 2017;3:12.

253. Bender A, Scheiber J, Glick M, Davies JW, Azzaoui K, Hamon J, et al. Analysis of pharmacology data and the prediction of adverse drug reactions and off-target effects from chemical structure. ChemMedChem. 2007 Jun;2(6):861–73.

254. Pauwels E, Stoven V, Yamanishi Y. Predicting drug side-effect profiles: a chemical fragment-based approach. BMC Bioinformatics. 2011 May 18;12:169.

255. Zhao X, Chen L, Lu J. A similarity-based method for prediction of

drug side effects with heterogeneous information. Math Biosci. 2018 Dec;306:136–44.

256. Kuhn M, Al Banchaabouchi M, Campillos M, Jensen LJ, Gross C, Gavin A-C, et al. Systematic identification of proteins that elicit drug side effects. Mol Syst Biol. 2013;9:663.

257. Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. Nucleic Acids Res. 2016 Jan 4;44(D1):D1075-1079.

258. Guney E. Investigating Side Effect Modules in the Interactome and Their Use in Drug Adverse Effect Discovery. In: Complex Networks VIII [Internet]. Springer, Cham; 2017 [cited 2021 May 6]. p. 239–50. Available from: https://link-springer-com.sare.upf.edu/chapter/10.1007/978-3-319-54241-6_21

259. Alexander-Dann B, Pruteanu LL, Oerton E, Sharma N, Berindan-Neagoe I, Módos D, et al. Developments in toxicogenomics: understanding and predicting compound-induced toxicity from gene expression data. Mol Omics. 2018 Aug 1;14(4):218–36.

260. Chierici M, Francescatto M, Bussola N, Jurman G, Furlanello C. Predictability of drug-induced liver injury by machine learning. Biol Direct [Internet]. 2020 Feb 13 [cited 2020 Oct 29];15. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7020573/

261. Sumsion GR, Bradshaw MS, Beales JT, Ford E, Caryotakis GRG, Garrett DJ, et al. Diverse approaches to predicting drug-induced liver injury using gene-expression profiles. Biol Direct. 2020 Jan 15;15(1):1.

262. Aguirre-Plans J, Piñero J, Souza T, Callegaro G, Kunnen SJ, Sanz F, et al. An ensemble learning approach for modeling the systems biology of drug-induced injury. Biol Direct. 2021 Jan 12;16(1):5.

263. Franz M, Lopes CT, Huck G, Dong Y, Sumer O, Bader GD. Cytoscape.js: a graph theory library for visualisation and analysis. Bioinformatics. 2016 Jan 15;32(2):309–11.

264. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, et al. The Connectivity Map: using gene-expression signatures to connect

small molecules, genes, and disease. Science. 2006 Sep 29;313(5795):1929–35.

265. Halakou F, Kilic ES, Cukuroglu E, Keskin O, Gursoy A. Enriching Traditional Protein-protein Interaction Networks with Alternative Conformations of Proteins. Sci Rep. 2017 Aug 3;7(1):7180.

266. Kominakis A, Hager-Theodorides AL, Zoidis E, Saridaki A, Antonakos G, Tsiamis G. Combined GWAS and 'guilt by association'-based prioritization analysis identifies functional candidate genes for body size in sheep. Genet Sel Evol [Internet]. 2017 Apr 28 [cited 2021 Mar 17];49. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5408376/

267. Tejera E, Cruz-Monteagudo M, Burgos G, Sánchez M-E, Sánchez-Rodríguez A, Pérez-Castillo Y, et al. Consensus strategy in genes prioritization and combined bioinformatics analysis for preeclampsia pathogenesis. BMC Med Genomics. 2017 Aug 8;10(1):50.

268. Sims-Robinson C, Kim B, Rosko A, Feldman EL. How does diabetes accelerate Alzheimer disease pathology? Nat Rev Neurol. 2010;6(10):551–9.

269. Hiltunen M, Khandelwal VKM, Yaluri N, Tiilikainen T, Tusa M, Koivisto H, et al. Contribution of genetic and dietary insulin resistance to Alzheimer phenotype in APP/PS1 transgenic mice. J Cell Mol Med. 2012 Jun;16(6):1206–22.

270. Du J, Wang Z. Therapeutic potential of lipase inhibitor orlistat in Alzheimer's disease. Med Hypotheses. 2009 Nov;73(5):662–3.

271. Silverbush D, Sharan R. A systematic approach to orient the human protein-protein interaction network. Nat Commun. 2019 Jul 9;10(1):3015.

272. do Valle IF, Roweth HG, Malloy MW, Moco S, Barron D, Battinelli E, et al. Network medicine framework shows that proximity of polyphenol targets and disease proteins predicts therapeutic effects of polyphenols. Nature Food. 2021 Mar;2(3):143–55.

273. Pujol A, Mosca R, Farrés J, Aloy P. Unveiling the role of network and systems biology in drug discovery. Trends Pharmacol Sci. 2010

Mar;31(3):115–23.

274. Anaxomics Biotech SL. Biological Effectors Database [Internet]. Biological Effectors Database. 2020. Available from: http://www.anaxomics.com/biological-effectors-database.php

275. Brown GR, Hem V, Katz KS, Ovetsky M, Wallin C, Ermolaeva O, et al. Gene: a gene-centered information resource at NCBI. Nucleic Acids Res. 2015 Jan;43(Database issue):D36-42.

276. Braschi B, Denny P, Gray K, Jones T, Seal R, Tweedie S, et al. Genenames.org: the HGNC and VGNC resources in 2019. Nucleic Acids Res. 2019 Jan 8;47(D1):D786–92.

277. Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res. 2021 Jan 8;49(D1):D325–34.

278. Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, et al. The reactome pathway knowledgebase. Nucleic Acids Res. 2020 Jan 8;48(D1):D498–503.

279. Breuer K, Foroushani AK, Laird MR, Chen C, Sribnaia A, Lo R, et al. InnateDB: systems biology of innate immunity and beyond--recent updates and continuing curation. Nucleic Acids Res. 2013 Jan;41(Database issue):D1228-1233.

280. Barabási A-L. Network medicine--from obesity to the “diseasome.” N Engl J Med. 2007 Jul 26;357(4):404–7.