ALMA MATER STUDIORUM · UNIVERSITÀ DI BOLOGNA

DOTTORATO DI RICERCA IN
DATA SCIENCE AND COMPUTATION

Ciclo XXXV

Settore Concorsuale: 03/D1 - Chimica e Tecnologie Farmaceutiche, Tossicologiche e Nutraceutico-alimentari

Settore Scientifico Disciplinare: CHIM/08 - Chimica Farmaceutica

Network Analysis and Machine Learning Assist Drug Repurposing and Safety Assessment in Neurological Diseases

Presentata da: Luca Menestrina

Coordinatore Dottorato
Prof. Daniele Bonacorsi

Supervisore
Prof. Maurizio Recanatini

Co-supervisore
Prof. Andrea Cavalli

Esame Finale Anno 2024

Alla mia Famiglia

Abstract

In recent decades, two prominent trends have influenced the data modeling field, namely network analysis and machine learning. This thesis explores the practical applications of these techniques within the domain of drug research, unveiling their multifaceted potential for advancing our comprehension of complex biological systems.
The research undertaken during this PhD program is situated at the intersection of network theory, computational methods, and drug research.

Across six projects presented herein, there is a gradual increase in model complexity. These projects traverse a diverse range of topics, with a specific emphasis on drug repurposing and safety in the context of neurological diseases.
The aim of these projects is to leverage existing biomedical knowledge to develop innovative approaches that bolster drug research. The investigations have produced practical solutions, not only providing insights into the intricacies of biological systems, but also allowing the creation of valuable tools for their analysis. In short, the achievements are:

- A novel computational algorithm to identify adverse events specific to fixed-dose drug combinations.
- A web application that tracks the clinical drug research response to SARS-CoV-2.
- A Python package for differential gene expression analysis and the identification of key regulatory "switch genes".
- The identification of pivotal events causing drug-induced impulse control disorders linked to specific medications.
- An automated pipeline for discovering potential drug repurposing opportunities.
- The creation of a comprehensive knowledge graph and development of a graph machine learning model for predictions.

Collectively, these projects illustrate diverse applications of data science and network-based methodologies, highlighting the profound impact they can have in supporting drug research activities.

Table of Contents

List of Acronyms and Abbreviations

AD Alzheimer's Disease

ADR Adverse Drug Reaction

AI Artificial Intelligence

AMRR Adjusted Mean Reciprocal Rank

APCC Average Pearson Correlation Coefficient

ATC Anatomical Therapeutic Chemical (code)

AUC-ROC Areas Under the Receiver Operating Characteristic Curve

CAG Cytosine-Adenosine-Guanine

CNS Central Nervous System

CSN Chemical Space Network

DP Drug Projection

DT Drug-Target

EMA European Medicines Agency

ER Erdös and Rényi

FAERS FDA Adverse Event Reporting System

FDA Food and Drug Administration

FDR False Discovery Rate

GBM Glioblastoma

GML Graph Machine Learning

GNN Graph Neural Network

GO Gene Ontology

HD Huntington's Disease

HPO Human Phenotype Ontology

IGSEA Inverted Gene Set Enrichment Analysis

JADER Japanese Adverse Drug Event Report

KG Knowledge Graph

KGE Knowledge Graph Embedding

KGEM Knowledge Graph Embedding Model

LCWA Local Closed-World Assumption

LP Link Prediction

MLP Multilayer Perceptron

MMP Matched Molecular Pair

MRR Mean Reciprocal Rank

MS Multiple Sclerosis

NHGRI National Human Genome Research Institute

NLP Natural Language Processing

NSSA Self-Adversarial Negative Sampling

OOV Out-Of-Vocabulary

PMI Pointwise Mutual Information

PPI Protein-Protein Interaction

PPMI Positive Pointwise Mutual Information

PRAC Pharmacovigilance Risk Assessment Committee

PRR Proportional Reporting Ratio

QoL Quality of Life

QSAR Quantitative Structure-Activity Relationship

sLCWA Stocastic Local Closed-World Assumption

SMILES Simplified Molecular Input Line Entry System

SRL Statistical Relational Learning

TCGA The Cancer Genome Atlas

TP Target Projection

TSV Tab-Separated Values

VAERS Vaccine Adverse Event Reporting System

Part I

Introduction

0 Thesis Overview

The main mission of science is that of solving problems, yet complex systems, by their nature, elude an exhaustive description. Predicting the long-term behavior of these systems remains a formidable challenge for science.[1] They have: diverse components with intricate hierarchical interactions, behaviors spanning multiple scales, and complicated transition laws. They possess a remarkable sensitivity to initial conditions, and they are characterized by emergence of unpredictable phenomena, non-equilibrium dynamics, combinatorial explosion, and self-organization.[2]
Complex systems science has the ambitious task of understanding the principles of such systems, and explaining the emergent phenomena.[2]
Inherent complexity defines living matter, making it impossible to describe a living being in terms of mere few variables.[1, 3] Indeed, it can represent the prototypical exemple of a complex system.[1]

The PhD program in Data Science and Computation combined with the academic discipline of pharmaceutical chemistry[†] appears ideally suited for this topic. Indeed, during this PhD program several contemporary approaches were explored, exploiting a variety of methods to tackle the comprehension of biological complex systems of relevance for the drug research.

The content of this thesis follows the trajectory path of the PhD program and is organized into three main sections: Introduction, Projects, and Conclusions.
The first part provides the essential theoretical background and context for the thesis. It primarily focuses on two rising trends that have shaped the data modeling community in recent decades: network analysis and machine learning[4, 5], with a brief mention of graph machine learning. Additionally, it provides an overview of network applications in drug research.
The second part delves into the projects undertaken during the course of the PhD program. Each chapter takes the form of a scientific article, with some having been published during the program's duration. These six projects exhibit a progressive increase

[†] Translation of "Settore Scientifico Disciplinare: CHIM/08 - Chimica Farmaceutica" following the Consiglio Universitario Nazionale indications

in model complexity, beginning with descriptive approaches, continuing to predictive models, and culminating in a Knowledge Graph Machine Learning one. These projects encompass a diverse array of topics, with a particular focus on drug repurposing and safety for neurological diseases. Additionally, a data science project, carried out during an international research period at Chemotargets in Barcelona (Spain), involves the study of a vast pharmacovigilance dataset.

The last part summarizes the key findings, provides general conclusions, and presents possible future prospects and directions.

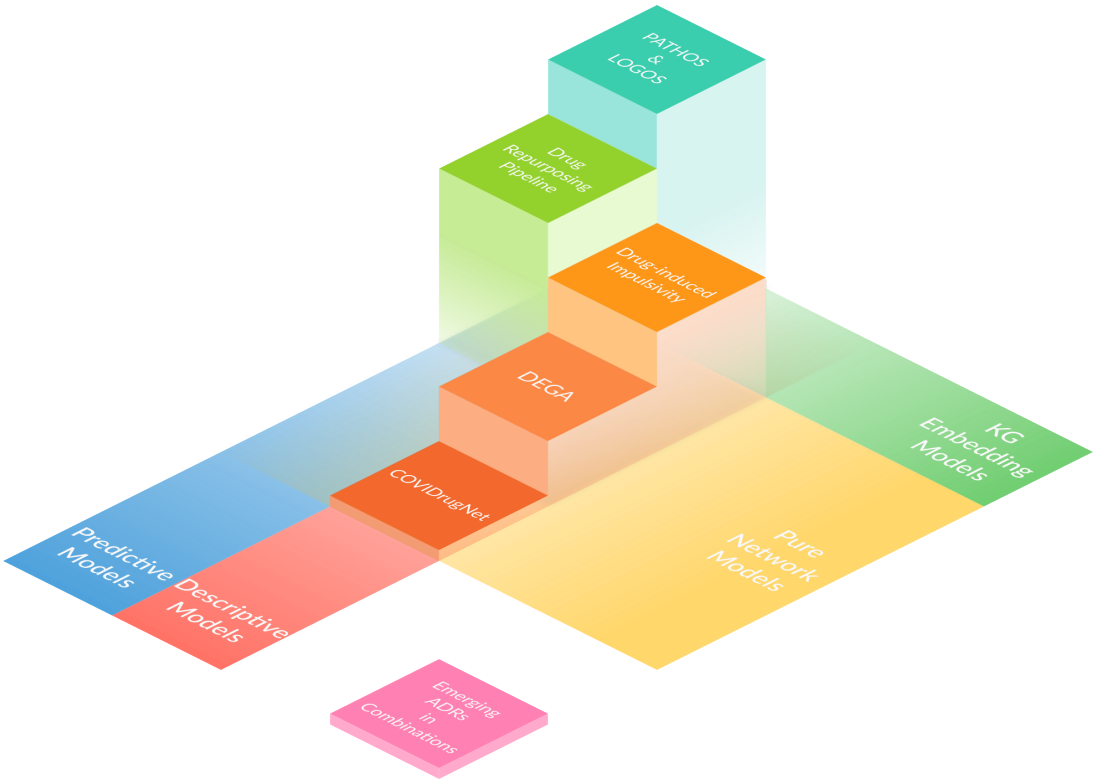


Figure 1. Visual Overview of PhD Project Progression. This diagram illustrates a progressive increase in complexity across the six projects undertaken during the PhD, starting with descriptive approaches (COVIDrugNet, DEGA, and Drug-induced Impulsivity), transitioning to a predictive model (Drug Repurposing Pipeline), and culminating in a Knowledge Graph Embedding Model (PATHOS and LOGOS).

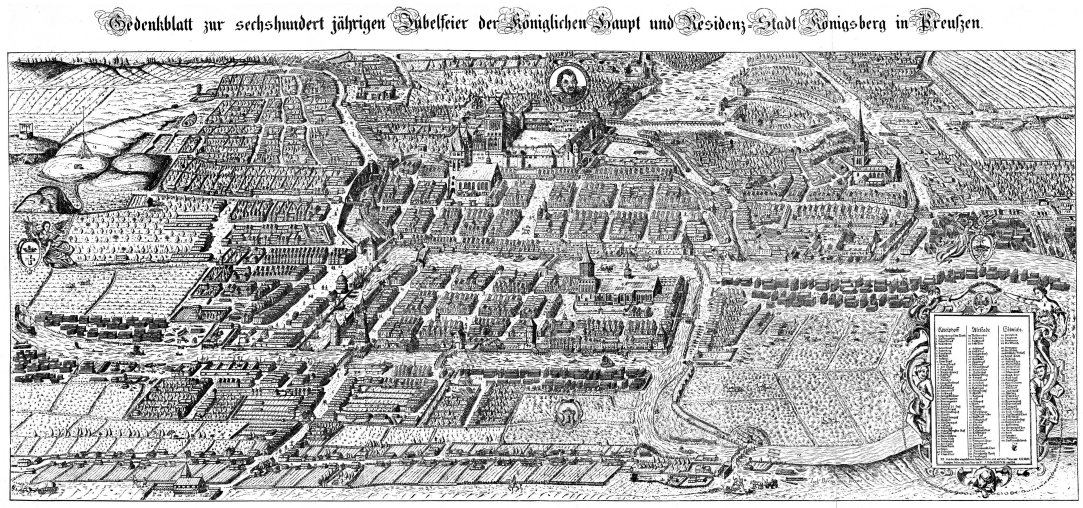
1 Network Theory

The work *"Solutio problematis ad geometriam situs pertinentis"*[6] by Leonard Euler is considered the first introduction to the notion of graphs. In it, he presents the Königsberg Bridges problem which consists in finding a path through the city of Königsberg (at Euler's time a Prussian city, now Kaliningrad, Russia) crossing each of its seven bridges exactly once. Euler's abstraction represented the different parts of the city as nodes (vertices), connected by bridges depicted as edges (links). By disregarding real distances and focusing solely on the relationships between these nodes and edges, Euler introduced the concept of graphs[†] as a powerful tool for studying complex networks. Graphs provide a mathematical representation of real-world networks, capturing the pairwise relations between various entities while discarding unnecessary details. Numerous everyday situations (systems) involve a collection of diverse elements interconnected through multiple interactions, exhibiting an underlying network structure. Frequently, it is this concealed network that holds the crucial insights to comprehend these situations effectively.[8]
This abstraction enables researchers to comprehend the underlying structures and hidden connections within intricate systems effectively. Consequently, graphs have become fundamental in analyzing diverse scenarios, from social networks and transportation systems to biological interactions and technological networks.)[9]

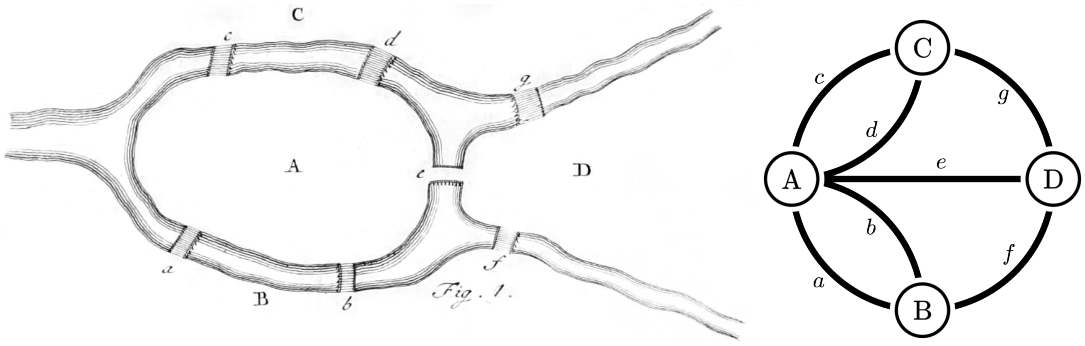
1.1 Basic Concepts

A graph is a collection of nodes (also called vertices) connected by edges (also called links).[10] Mathematically, a graph G consists of an ordered pair $(\mathcal{N}, \mathcal{E})$, comprising a set of nodes $\mathcal{N}$ and a set of unordered pairs of nodes termed edges $\mathcal{E} \subseteq \{\{i, j\} \mid i, j \in \mathcal{N}\}$. A simple graph (Figure 1.2a), by definition, allows only one edge between any two nodes and does not permit self-loops (self-interactions), which are edges connecting nodes to themselves: $\mathcal{E} \subseteq \{\{i, j\} \mid i, j \in \mathcal{N}\ and\ i \neq j\}$.
The typical mathematical notation used to denote a graph with $|\mathcal{N}|$ nodes and $|\mathcal{E}|$ edges

[†] Interestingly, it was J. J. Sylvester in 1878 who introduced the term "graph", drawing a parallel between mathematics and chemical structures.[7]



(a)



(b) **(c)**

Figure 1.1. Seven Bridges Problem. (a) Königsberg Map from 1613, **(b)** Original Euler's representation from Solutio problematis ad geometriam situs pertinentis.*Commentarii Academiae Scientiarum Imperialis Petropolitana* Tom.VIII Tab.VII p.128[6], **(c)** A graph representation of the same network.

is $G(|\mathcal{N}|, |\mathcal{E}|)$*.[9]

When working with graphs, it's a common practice to depict them using diagrams. In these representations, nodes are typically visualized as points or small circles, while edges are represented as line segments or curves connecting the respective nodes.[11, 12]

In graph theory, defining subgraphs is essential as it allows us to examine specific components within a larger network. A subgraph of graph G, denoted as B, is a graph that

* Throughout this thesis, the number of elements contained in a set $S = \{..\}$ will be denoted as $|S|$ (its cardinality), which is a common notation in the mathematical literature.

inherits a subset of nodes $\mathcal{N}_B$ from the node set $\mathcal{N}_G$ and a subset of edges $\mathcal{E}_B$ from the edge set $\mathcal{E}_G$. In simpler terms, it's a graph extracted from G, comprising selected nodes and edges while maintaining the relationships between those chosen elements.[13]

Adjacency Matrix The adjacency matrix is a fundamental representation of a graph, presented as a binary matrix where a value of 1 indicates the presence of an edge between two vertices, and 0 represents the absence of an edge.

$$\boldsymbol{A}_{ij} = \begin{cases} 1, & \text{if there is an edge between nodes } i \text{ and } j, \text{ thus } (i,j) \in \mathcal{E} \\ 0, & \text{otherwise} \end{cases} \tag{1.1}$$

The rows and columns of this matrix share the same names, serving as identifiers for the nodes in the graph.
When the adjacency matrix is empty, lacking any entries to define edges, the graph is classified as empty as well. Conversely, when the adjacency matrix is fully populated, signifying the presence of all potential edges, the graph is deemed complete. [9]
In cases where the graph is weighted, each edge is assigned a specific weight, leading to what is known as a weighted graph (Figure 1.2b).
Additionally, graphs may have directional edges (Figure 1.2c), where the order of the vertices matters, designating one vertex as the starting (source) point and the other as the ending (target) point. Consequently, the adjacency matrix loses its symmetry along the diagonal due to the distinction between source and target vertices ($\boldsymbol{A}_{ij} \neq \boldsymbol{A}_{ji}$).

Adding complexity The study of network structures has evolved to accommodate higher complexity and address real-world scenarios with greater fidelity. One illustrative example is found in multigraphs (Figure 1.2d), where a single pair of nodes can be connected by multiple edges. Another fundamental aspect of such networks is their data heterogeneity, where different types of nodes coexist, each representing distinct entities or attributes. This diversity enables the incorporation of various facets of interconnected systems, making heterogeneous (or multipartite) networks more comprehensive models of complex phenomena (Figure 1.2e).[14, 15]
Among the frameworks capable of handling this intricate network architecture are multilayer networks (Figure 1.2f). A multilayer network consists of nodes, edges, and layers, each layer serving as a distinct level of connectivity. The interpretation of these

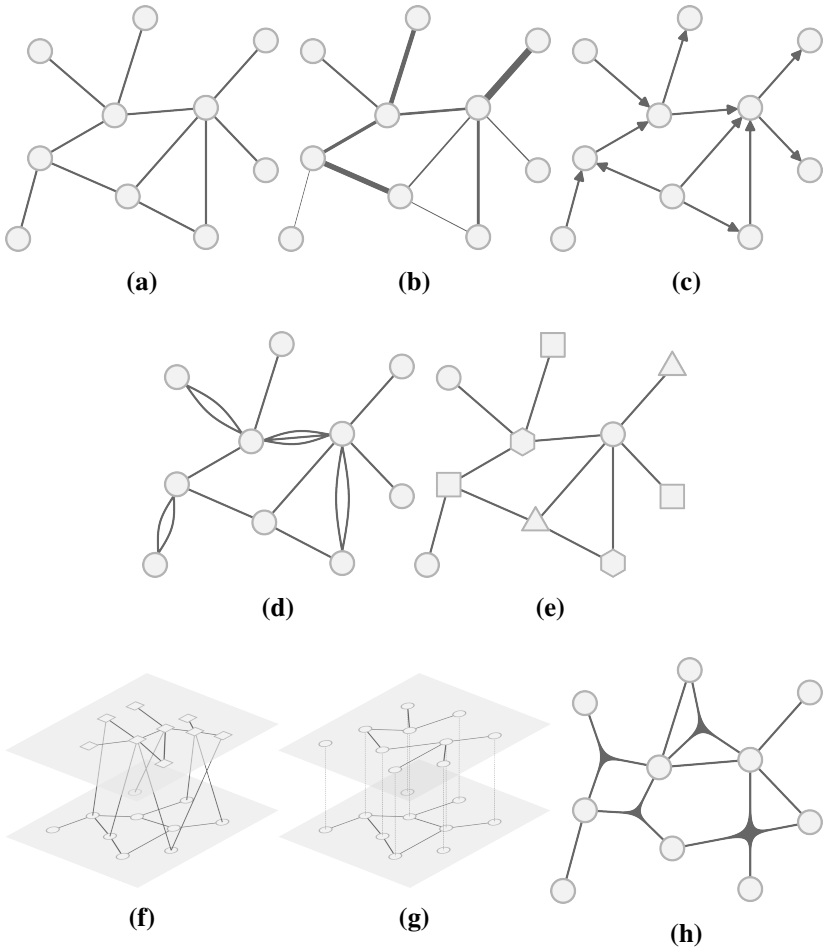


Figure 1.2. Types of Networks (a) Simple Network, **(b)** Weighted Network, **(c)** Directed Network, **(d)** Multigraph, **(e)** Heterogeneous Network, **(f)** Multilayer Network, **(g)** Multiplex, **(h)** Hypergraph.

layers relies on the specific implementation of the model, granting flexibility to capture relationships of different scales and nature.[14, 15]

A specific class of multilayer networks is the multiplex network (Figure 1.2g). Multiplex networks consist of a single set of nodes interconnected by multiple types of relations represented in separate layers. Each layer represents a unique type of link between the same set of nodes, providing a multi-faceted view of their interactions.[15]

Beyond traditional graphs and multilayer networks, another generalization is the concept of hypergraphs (Figure 1.2h). In hypergraphs, edges can connect any number of vertices, not just two, as in standard graphs. This extension offers even more expressive power to model relationships that involve multiple entities simultaneously.[15]

Knowledge Graphs Defining a knowledge graph (KG) precisely is challenging due to the presence of multiple conflicting definitions in the literature.[16, 17] Here, a specific definition that considers KGs as heterogeneous directed and labeled multigraphs will be adopted, building on top of the one given by the Schlichtkrull et al.[18] Thus, a knowledge graph could be described as a graph-based data structure that contains diverse types of vertices and edges, which can be defined as $G = (\mathcal{N}, \mathcal{E}, \mathcal{R}, \Psi)$. Each edge within the graph is characterized by a relation type ($r \in \mathcal{R}$), and represented as a triplet value $(u, r, v) \in \mathcal{E}$. The vertices, often referred to as entities, are categorized into subsets based on their type (from the set Ψ).[17]
KGs serve as structured representations of real-world information. Their ability to model complex, structured data in a machine-readable manner has led to their extensive use in various domains, including question answering, information retrieval, and content-based recommendation systems.[19]
Although the initial entity in the triple is commonly known as the head entity, linked to the tail entity through a relation, it's worth noting that, due to their significant role in encoding human reasoning and language, the components of triplets are often termed as subject, predicate, and object.[18]

1.2 Graph Properties

To effectively work with graphs, certain definitions are essential:

Order, Size and Density The primary two attributes characterizing a graph are its count of vertices and edges, often referred to as its order and size, respectively. For simple graphs, the concept of graph density becomes relevant. Graph density can be defined as the proportion of actual edges in relation to the maximum potential edges. In undirected simple graphs, this ratio is calculated as follows:

$$D = \frac{2|\mathcal{E}|}{|\mathcal{N}|(|\mathcal{N}| - 1)} \tag{1.2}$$

For directed simple graphs, instead, the maximum potential edges double in comparison to undirected graphs (due to the presence of two directions for each edge). Consequently, the formula for density in directed simple graphs becomes:

$$D = \frac{|\mathcal{E}|}{|\mathcal{N}|(|\mathcal{N}| - 1)} \tag{1.3}$$

A complete simple graph has the maximum number of edges, thus its density is equal to 1.

Degree Node degree refers to the count of edges connected to a particular node. In undirected graphs with vertices $\mathcal{N}$, the degree of node i, denoted as $k(i)$, is determined by the sum of values within its corresponding row or column in the adjacency matrix:

$$k_i = \sum_{j=1,|\mathcal{N}|} a_{ij} \tag{1.4}$$

In directed graphs, the concept of node degree gives rise to the need for distinguishing two distinct types. One relates to the number of edges directed toward a node (in-degree), while the other represents the number of edges departing from a node (out-degree). These quantities are formulated as follows:

$$k_i^{in} = \sum_{j \in \mathcal{N}} a_{ji} \qquad k_i^{out} = \sum_{j \in \mathcal{N}} a_{ij} \tag{1.5}$$

Node with high degrees hold a crucial position within the graph and are frequently referred to as hubs. Eliminating these nodes often results in the disconnection of the network.[13]

Shortest Path, Distance and Diameter The shortest path is the minimum number of edges required to travel from node i to node j if they are connected. The measure of its length is referred to as the distance d_{ij}. As a result, the neighbors of a node include all the nodes connected to it by just one edge.
The diameter D of a graph is defined as the greatest distance achievable between two nodes within the graph.

Centrality

Graphs, though abstract, mirror intricate real-world networks, offering insight into their complexity. Assessing the importance of nodes within these networks holds practical significance. Relevant examples include identifying influential individuals in social networks, pinpointing crucial nodes in digital and urban infrastructure, tracking disease super-spreaders, and uncovering interactions between genes.
Nodes with many connections (recognized by their high degree) are often considered

to have a pivotal role within graphs. Nonetheless, the concept of node importance depends on both the underlying network structure and its concrete implications. Thus, graph theory provides the more nuanced notion of node centrality. Diverse centrality measures have emerged, each providing a unique lens to assess node importance within the intricate patterns inherent to real-world networks. Noteworthy among the measures are:

Closeness Centrality The easiest approach to identify the most central node is to find the node that efficiently reaches and directly influences other nodes of the network. This is reflected in the identification of the node with the shortest average distance to all others, thus maximizing its closeness centrality,[9, 13] which, for node i, is defined as:

$$c_i^{cl} = \frac{1}{\sum_{j \in \mathcal{N}} d_{ij}} \tag{1.6}$$

The closeness centrality measures the ability of an individual in a social network to rapidly interact with others, making it essential for identifying influential people and potential opinion leaders in various social contexts, including the identification of disease spreaders in epidemiological studies.

Betweenness Centrality An alternative method for evaluating the significance of a node z is through its role as an intermediary between others, which means measuring its betweenness centrality c_z^b. This is achieved by counting how many times the shortest path (σ_{ij}) connecting nodes i and j passes through the node in question.[9, 13]

$$c_z^b = \sum_{j \in \mathcal{N}} \frac{\sigma_{ij}^z}{\sigma_{ij}} \tag{1.7}$$

Nodes of this kind play a vital role in maintaining network connectivity, as their removal leads to network fragmentation.
The betweenness centrality is crucial in transportation or telecommunication networks. It identifies vital nodes for traffic or information flow unveiling vulnerabilities.

Eigenvector Centrality Frequently, the importance of a node extends beyond its direct connections, encompassing the influence of its neighbors. When a node is connected to others of notable influence, its own impact is augmented compared to connections with less influential nodes. This is the core consideration of eigenvector centrality

c_i^{eig}. Consequently, a node has high eigenvector centrality if also its neighbors hold elevated eigenvector centrality. This implies that the eigenvector centrality of node i aligns with the average centralities of its nearest neighbors $\mathcal{H}_i = \{j \mid (i,j) \in \mathcal{E} \vee (j,i) \in \mathcal{E}\}$, as expressed in:

$$c_i^{eig} = \frac{1}{\lambda} \sum_{j \in \mathcal{H}_i} \boldsymbol{A}_{ij} c_j^{eig} \tag{1.8}$$

where λ represents a constant. Elaborating on this concept, formulating centralities as a vector $\boldsymbol{c}$ and rewriting this equation in matrix form:

$$\lambda \boldsymbol{c} = \boldsymbol{A} \cdot \boldsymbol{c} \tag{1.9}$$

reveals that $\boldsymbol{c}$ constitutes the primary eigenvector of the adjacency matrix with eigenvalue λ. The components of this leading eigenvector provide the eigenvector centrality values for each node in the network. Defined in this manner, the eigenvector centrality of each node depends on both the quantity and quality of its connections. While nodes with numerous connections retain their significance, vertices with fewer yet more impactful connections can surpass those with a greater number of less influential connections.[14, 20]
The eigenvector centrality plays a pivotal role in web search engines like Google by ranking web pages based on their connections to other highly ranked pages, enhancing the relevance and quality of search results.

Assortativity Nodes often exhibit a tendency to connect with others that share similar attributes or properties. This phenomenon is termed assortativity, and it plays a crucial role in understanding the structural patterns within a network. Assortativity can be evaluated based on various properties, each offering insights into different aspects of the network's behavior. The most common approach involves assessing assortativity in terms of node degree. In this context, nodes with high degrees typically display a preference for associating with other high-degree nodes, while low-degree nodes tend to link with fellow low-degree nodes. Conversely, a situation where high-degree nodes link with low-degree nodes is referred to as disassortativity. Degree assortativity can be quantified using the assortative coefficient, which is a particular case of the Pearson correlation coefficient. This number, ranging between -1 and 1, holds positive values in assortative networks, and negative values for those that are disassortative. A

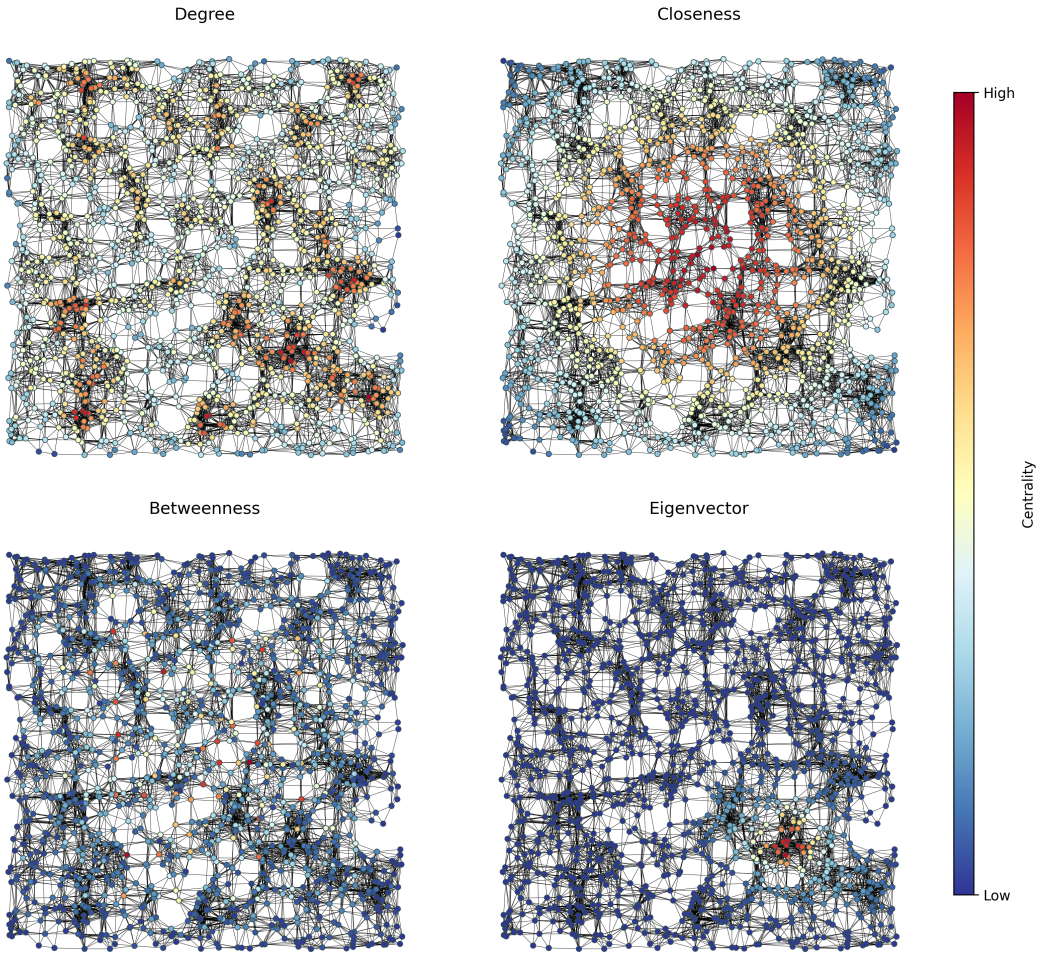


Figure 1.3. Network Centralities. Degree centrality (top left), closeness centrality (top right), betweenness centrality (bottom left), eigenvector centrality (bottom right) of the same random geometric graph.

coefficient of 0 suggests a network with random assortativity, where connections are established without any specific pattern.[9, 13, 21]

Clustering Coefficient The clustering coefficient cc is another valuable parameter for assessing node relationships within a network. This metric quantifies the degree to which a node's neighbors ($\mathcal{H}$) are interconnected, representing the average fraction of pairs of neighbors that share a direct connection.[9, 22] For a node i, with more than one connection ($k > 1$), the clustering coefficient is computed as:

$$cc_i = \frac{2e_i}{|\mathcal{H}|_i(|\mathcal{H}|_i - 1)} \tag{1.10}$$

where e_i is the count of edges connecting the neighbors of node i. Moreover, a global clustering coefficient can be defined in relation to the entire network. It is computed as the average of individual clustering coefficients, as shown in:

$$CC = \langle cc \rangle = \frac{1}{|\mathcal{N}_{k>1}|} \sum_{i \in \mathcal{N}_{k>1}} cc_i \tag{1.11}$$

Notably, nodes with a degree lower than 2 ($k < 2$) are omitted from the calculation of the average clustering coefficient.[9, 13, 22, 23]

Clique The term clique refers to a maximal subgraph containing three or more nodes, where all nodes within the group are directly connected to each other and no other node is linked to all of them.
This definition can actually be broadened by relaxing the requirement of adjacency and considering reachability instead. Accordingly, an n-clique is a maximal subgraph where the maximum shortest path between any pair of nodes is n or less. This definition aligns with the conventional one when n equals 1.[23]

Communities

Nodes in networks are often clustered into tightly connected groups with more edges directed towards nodes in the same group than nodes outside of it. These groups are usually called communities, but also clusters or modules. Many times, communities give hints about how a network is organized and what functions it serves.[24, 25]
Network modularity, a fundamental concept introduced by Newman[24], is a fundamental tool for evaluating how effectively a network can be divided into subgraphs. The objective is to optimize the internal edges within each distinct community while minimizing the connections that between different communities.
Evaluating the partition's quality necessitates the definition of a symmetric matrix (which dimensions are equal to the number of subgraphs), with entries e_{ij} indicating half the fraction of edges linking subgraph i with subgraph j in the original graph.[25]
Subsequently, the diagonal values e_{ii}, representing the proportion of edges within group i without the factor of a half, are compared with the corresponding quantity in a null case (i.e., a random partition $\left(\sum_j e_{ij}\right)^2$).[24, 25]

The modularity Q is thus expressed as:

$$Q = \sum_i \left(e_{ii} - \left(\sum_j e_{ij} \right)^2 \right) \tag{1.12}$$

Within the domain of community detection, numerous procedures are available. These methods can be systematically categorized using a nomenclature introduced by Newman: traditional computer science methods and sociological methods, as well as more contemporary divisive and agglomerative approaches.[26] The first approach, favored by graph theorists and mathematicians, centers around spectral analysis. The second, embodied in hierarchical clustering, represents a principal technique employed by sociologists. On the other and, divisive approaches employ a top-down strategy, recursively removing nodes and edges. Conversely, agglomerative methods involve building communities through recursive node grouping, utilizing a bottom-up process.[23] Illustrating these categories, some key exemplars are: Spectral Analysis, Hierarchical Clustering, Girvan-Newman, and Greedy Modularity.

Spectral Analysis In mathematics, the spectrum of a matrix is the set of its eigenvalues. In the context of graph theory, this spectrum is commonly computed for matrices such as the adjacency matrix, its Laplacian, or their normalized counterparts. For an undirected graph, the Laplacian $\boldsymbol{L}$ assumes the role of a symmetric matrix with dimensions corresponding to the number of nodes. $\boldsymbol{L}$'s diagonal elements signify node degrees, while off-diagonal elements mirror those of the adjacency matrix. Alternatively, $\boldsymbol{L}$ can be expressed as the difference between the diagonal matrix of node degrees $\boldsymbol{D}$ and the adjacency matrix $\boldsymbol{A}$:

$$\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A} \tag{1.13}$$

Rows and columns of the Laplacian sum to zero, since $\boldsymbol{D}_{ii} = \sum_j \boldsymbol{A}_{ij}$. As a consequence, the vector $\mathbf{1} = 1, 1, 1, ...$ is always an eigenvector with eigenvalue zero.[27] If a network naturally separated into distinct components, non-overlapping communities $G_k (k = 1 \ ... \ g)$, the Laplacian takes a block diagonal structure. Each block represents the Laplacian of its respective component and has an eigenvector $\boldsymbol{v}^{(k)}$ with eigenvalue zero. The elements $\boldsymbol{v}_i^{(k)}$ of this eigenvector equal 1 if node i resides in G_k, and 0 otherwise*. Thus, there will be g degenerate eigenvectors with eigenvalue 0.

* $\boldsymbol{v}_i^{(k)} = \begin{cases} 1, & \text{if } i \in G_k \\ 0, & \text{otherwise} \end{cases}$

In cases where a network doesn't divide perfectly into communities, due to a handful of edges that don't align perfectly with the block-diagonal pattern, the previously stated perfect scenario no longer holds true. Instead, a different scenario emerges. In general, there will be the primary eigenvector $\mathbf{1}$ with an eigenvalue of zero, accompanied by $g-1$ eigenvalues that slightly deviate from zero*. The corresponding eigenvectors will be approximately represented as linear combinations of the previously defined $\boldsymbol{v}^{(k)}$ eigenvectors. By identifying eigenvalues marginally greater than zero and employing linear combinations of corresponding eigenvectors, one can theoretically approximate the underlying community blocks, even in cases of imperfect separation.[26]
Knowing the Laplacian's spectrum, the mapping of nodes to communities can be performed utilizing various vector clustering techniques. This mapping, termed spectral embedding, involves computing the Laplacian matrix, finding the first k eigenvectors associated with the k smallest eigenvalues, forming a matrix from these eigenvectors (each row defines features of a specific node), and then clustering nodes using methods like k-means[28], DBSCAN[29], etc.[27]

Hierarchical Clustering The concept underlying this method involves establishing a similarity measure (x_{ij}) for node pairs (i, j) based on the given network structure. With this measure in hand, a network initially comprising $|\mathcal{N}|$ disconnected nodes is constructed. Edges are incrementally added between pairs of nodes in descending order of similarity, starting with the most similar pair. It's important to note that the newly added edges are unrelated to those in the original network; the original network is solely employed for calculating the similarity measure. As the similarity measure (x) decreases and communities merge, the single linkage hierarchical clustering approach progressively reduces the number of communities. The algorithm begins with $|\mathcal{N}|$ components, each containing a single node, and culminates with a single component encompassing all nodes. The components at each step are fully contained within the components of the subsequent step, forming a hierarchical tree or dendrogram that represents the entire progression of the algorithm.[30]

Girvan-Newman Girvan and Newman introduced an approach that shifts the focus towards the edges that exhibit the least centrality, those that predominantly bridge dif-

* Slightly greater than zero, since all the eigenvalues of the graph Laplacian are non-negative.

ferent communities, rather than aiming to the ones most central to communities.
This approach involves a generalization of the concept of betweenness centrality, translating it from nodes to edges. Edge betweenness becomes the quantification of shortest paths between pairs of nodes traversing a given edge. This definition spots edges positioned between communities, because if communities are connected by a sparse number of intergroup edges, all shortest paths between different communities tend to traverse these select edges (those with elevated betweenness). By systematically removing these critical edges, the algorithm dissects the network, and unveils the underlying community structure.[23, 30]

Greedy Modularity To address community detection in large graphs, Newman introduced a rapid algorithm rooted in optimizing modularity.[24] As exhaustive search for all potential divisions would prove intractable due to exponential time complexity, the method employs a "greedy" approach, iteratively joining small communities. Beginning with individual-node communities, pairs of communities that determine the greatest increase (or minimal decrease) in Q are sequentially merged. Importantly, only pairs interconnected by edges are considered, since merging edge-disconnected communities does not impact Q.
Optimized versions of this algorithm have been suggested by Clauset[31] and Nguyen[32] to improve its efficiency.

Network Topologies and Models

In recent decades, the increased availability of large databases, optimized computing resources, and robust data analysis tools has facilitated the exploration of topological attributes of various real-world systems. These studies revealed that most of real networks often exhibit shared topological characteristics, such as relatively short characteristic path lengths, elevated clustering coefficients, fat-tailed degree distributions, degree correlations, and the presence of motifs and community structures.[23]
In the upcoming paragraphs, two key topological traits are explored, the small-world effect and scale-free degree distributions, while also referencing the classical mathematical framework of random graphs.

Random Graphs The systematic exploration of random graphs, inaugurated by Erdös and Rényi in 1959, led to the most extensively studied graph model: Erdös and Rényi (ER) random graphs.[33] They proposed to study the properties of graphs as a function of the increasing number of random connections, leveraging on probabilistic methods.
Erdös and Rényi proposed a model to generate random graphs ($G_{ER}(\mathcal{N}, \mathcal{E})$). Starting with $|\mathcal{N}|$ isolated nodes, the graph evolves by linking randomly chosen pairs of nodes, while prohibiting multiple connections, until edge count reaches $|\mathcal{E}|$. It is important to underline that this outcome is one realization within a statistical ensemble comprising all possible connection combinations.[23, 33]

Small-world Networks The most famous experiment on the small-world effect, is the one conducted by the psychologist Stanley Milgram in the 1960s.[34] While this experiment didn't reconstruct actual networks, it yielded insights into network structure. Participants were tasked with sending a letter through their acquaintance network to a specific target individual, allowing the experiment to explore path length distribution. Remarkably, about one-fifth of the letters reached the target, passing through approximately six individuals on average (inspiring the popular concept of "six degrees of separation"[35, 36]). This experiment stands as one of the earliest direct demonstrations of the small-world effect, illustrating that a majority of vertex pairs in networks are linked by short paths.
Watts and Strogatz proposed a method to construct graphs ($G_{WS}(\mathcal{N}, \mathcal{E})$) that exhibit both the small-world property and a high clustering coefficient. This small-world model begins with a one-dimensional regular lattice* of $|\mathcal{N}|$ nodes with periodic boundary conditions (i.e., a ring). Each node is initially connected to its neighbors within k lattice spacings, resulting in $|\mathcal{N}| \times k$ edges. Subsequently, for each node, links connected to clockwise neighbors are rewired with probability p to other nodes, ensuring no double or self-edges emerge. This rewiring process allows the small-world model to interpolate between a regular lattice and a configuration resembling a random graph. At $p = 0$, the lattice remains regular, while $p = 1$ approaches a random graph. Intermediate values of p, indeed, generate graphs with both the small-world property and

* While small-world models can be constructed on lattices of various dimensions or topologies, the most studied case is the one-dimensional version.

a non-trivial clustering coefficient. The emergence of long-range edges (shortcuts) is evident by the rapid reduction in $\langle d \rangle$* (mean shortest distance between node pairs) once p exceeds zero.[10, 22, 23, 36]

Scale-free Networks The fraction of nodes within the network that possess a degree of k is denoted as $p(k)$. Alternatively, $p(k)$ represents the probability that a randomly chosen node holds a degree of k. A plot illustrating $p(k)$ for a specific network can be generated by constructing a histogram based on the degree of nodes. This histogram constitutes the network's degree distribution.[36] In scenarios where all nodes exhibit similar topological characteristics (as seen in regular lattices or random graphs), edges would be uniformly distributed, leading to a binomial or Poisson degree distribution. However, real-world networks frequently exhibit a pronounced right-skewed distribution.† Many of them, indeed, manifest power laws tails in their degree distributions:

$$p(k) \propto k^{-\alpha} \tag{1.14}$$

The scaling exponent α, often referred to as the power law exponent, typically ranges from 2 to 3, and the degree value k must be equal to or greater than a threshold (which is always greater than 1)[37, 38]

First coined as "scale-free"‡ networks by Barabasi in 1999[38], these networks are identified by their consistent adherence to the same functional form across various scales.[23]

As previously noted, real networks frequently deviate from simple graph models. Consequently, analyzing degree distributions can become more intricate in certain cases. For instance, heterogeneous graphs exhibit multiple degree distributions, each corresponding to a specific node type. Alternatively, in the case of directed graphs, the

* For nodes i and j:

$$\langle d \rangle = \frac{1}{\frac{1}{2}|\mathcal{N}|(|\mathcal{N}|+1)} \sum_{i \geq j} d_{i,j}$$

† Non-symmetrical distribution in which most values are concentrated on the left (low degree), and there's a long tail extending to the right (a few nodes have exceptionally high degrees).

‡ Indeed, a functional form $f(x)$ is deemed scale-free when it remains unchanged, apart from a multiplicative factor, to the rescaling of the independent variable x. Thus, in this context, "power law" and "scale-free" can be considered synonymous, since only power law distributions are the solutions to $f(ax) = bf(x)$.[36]

degree distribution transforms into a function $p(j, k)$ reliant on two variables, necessitating the depiction of both in-degree and out-degree distributions.[36]
The prevalence of scale-free networks holds substantial implications for our comprehension of diverse phenomena, both natural and human-made, as power-law distributions play a pivotal role in numerous scientific contexts, influencing fields ranging from biology and sociology to technology and information systems.[39, 40]

1.3 Network-based Predictions

Inherent patterns within networks, whether representing social relationships, information flow, or intricate connections, can be leveraged for overcoming challenges that can compromise the reliability of real network representations, such as technical limitations, experimental errors, and data availability. Realizing this opens avenues for predictive insights, offering a potent tool to comprehend and forecast various elements within the structure of a network.
Predicting new edges within a network involves foreseeing the formation of connections between existing or new nodes. This could include predicting potential friendships in a social network, potential collaborations between researchers in a co-authorship network, or potential interactions between proteins in a biological network. The properties of nodes, as well as examining similarity in edge neighborhoods (first neighbors, entire modules, or even the whole network), and analyzing sequential snapshots of network topology, play a pivotal role in the prediction of connections.[41]
Similarly, predicting new nodes, means anticipating the emergence of new entities within the network, for instance the appearance of new users in a social network, new web pages in a hyperlink structure, or new proteins in a molecular interaction network. Predicting nodes is more intricate than edge prediction, especially if they bridge diverse network modules.[42]
At a broader level, there are cases where the network is so incomplete that very little is known about its structure. Nevertheless, a comprehensive understanding of the behavior of the complex system encoded by the network is often available. Unveiling the underlying network from this system behavior implies envisioning its evolution over time or under slightly different conditions, and is referred to as network inference or reconstruction.[43, 44]

In addition to predicting nodes, edges, and whole networks, the properties of these individual components, as well as the properties of entire networks, can also be foreseen.
Spanning across various domains, from social network analysis to information dissemination, predictive approaches within networks play a pivotal role in offering valuable insights into the behavior of interconnected systems going beyond mere discovery of new components.[43]

2 Machine Learning

In 1956, John McCarthy first described "the science and engineering dedicated to creating intelligent machines" with the term "artificial intelligence".[45, 46] It's crucial to emphasize that artificial intelligence (AI) doesn't rely on some sort of wizardry, but instead relies on the application of probability, statistics, and mathematics, leveraging relevant data and hardware capacities.[46] Machines actually struggle with tasks that humans find intuitive, appearing automatic and effortless, yet remain intricate to define formally.[47]

In practice, much of contemporary AI research doesn't aim to replicate human-like intelligence (also known as strong AI); instead, it focuses on automating execution and decision-making processes for specific tasks (weak AI).[46]

The first generally recognized work about AI traces back to 1943, with Warren McCulloch and Walter Pitts attempting to simulate neural networks using computational circuits (a concept now referred to as neural networks, despite only their superficial resemblance to actual neural structures).[48] McCulloch and Pitts proposed an artificial neuron model where each neuron adopts an "on" or "off" state based on stimulation from neighboring neurons. They demonstrated the ability of networks to compute various functions and logical operations, suggesting the potential for learning within appropriately designed networks.[47]

The subfield of AI exploring the ability to enhance performance through experience is recognized as machine learning.[49] Unlike traditional explicit programming, machine learning models undergo a "training" to uncover patterns within data.[46]

It's crucial to emphasize that the nodes and edges within artificial neural networks don't inherently mirror real-world connections; instead, they employ mathematical operations on inputs to generate outputs[46]. Nevertheless, owing to their capacity to learn from input data, artificial neural networks present a powerful machine learning technique for capturing nonlinear relationships among variables.[50]

The ability to process real raw data of conventional machine learning is limited. Representation learning emerged as a collection of methods enabling machines to autonomously infer the necessary representations for detection or classification directly from raw data. Deep learning, a subset of representation learning, incorporates multiple layers of representation achieved through compositions of nonlinear modules,

gradually transforming raw inputs into higher-level abstractions.[51] These encodings capture essential features or representations of data in a lower-dimensional space while preserving meaningful relationships and patterns. In the context of machine learning and data analysis, this technique is called embedding.
The rapid expansion of big data*, accompanied by advancements in computing capabilities †, alongside the improvement of machine learning algorithms, has allowed the training of complex network architectures within manageable timeframes, utilizing massive datasets. This convergence has catalyzed the current surge of AI applications.[46]

2.1 Learning Techniques

Most learning techniques can be categorized into seven distinct classes: supervised learning, unsupervised learning, semi-supervised learning, active learning, reinforcement learning, transfer learning, and multitask learning.[46] The choice of appropriate techniques should be guided by the specific requirements of the task, having each class its own strengths and weaknesses. A brief overview of these approaches follows.

Supervised Learning The core characteristic of supervised learning lies in the availability of labeled data‡. Through iterative assessments, the model's effectiveness is measured, and the feedback on the system's performance is exploited to finely adjust internal parameters, ultimately aiming to minimize deviations from the desired outcome.[46, 52]

Unsupervised Learning Unsupervised learning algorithms are employed when actual responses to the data are absent, precluding the use of feedback for evaluating the model's solutions. In this scenario, the model recognizes patterns within the data and organizes input samples into distinct clusters. Alternatively, unsupervised learning

* Big data refers to vast and intricate datasets that surpass the capabilities of conventional data analysis tools, presenting challenges in terms of volume, velocity, variety, and veracity (the "four Vs").[50]

† Such as Graphic Processing Units (GPUs) and Tensor Processing Units (TPUs).

‡ Data accompanied by a corresponding tag, annotation, or "label" that represents the desired output or outcome associated with that data point.

methods can also be utilized to effectively reduce the dimensionality of the data.[46]

Semi-supervised Learning When a substantial amount of input data is accessible, yet only a subset holds known annotations, the application of semi-supervised learning models gains importance. This approach harnesses unlabeled data to either modify or reprioritize hypotheses drawn from limited labeled data. The process generally consists of utilizing a supervised learning algorithm to initial model training using available annotated data. Following this, the trained model is applied to predict labels for unlabeled data. Subsequently, the model is refined by integrating these newly assigned pseudo-labels with the original annotated data. The effectiveness of semi-supervised learning relies on the careful selection of pertinent a priori assumptions, specifically about the distribution of unlabeled data and the function inferred from the annotated set, that align with the underlying problem structure.[46]

Transductive Learning In cases where information is limited, accurate estimation of a function within a defined data region can be achievable, while facing challenges in obtaining high accuracy across the entire dataset. Transductive learning, similar to semi-supervised learning, uses all available data: labeled and unlabeled instances. However, it focuses on generating precise predictions for a designated subset of instances (the unlabeled ones), in contrast to inductive learning (the supervised and semi-supervised approaches described earlier), which aims to generalize across the complete dataset.[4, 53, 54]

Active Learning Active learning is a specialized interpretation of semi-supervised learning in which the algorithm actively interact with a user or an alternative source of information to collect labels for unlabeled data residing in the most uncertain regions of the input space. Unlike conventional semi-supervised techniques that aim to leverage the latent structure of unlabeled data to enhance label predictions, active learning is designed to minimize the quantity of labeled examples necessary for effective learning.[46]

Reinforcement Learning Reinforcement learning revolves around discovering optimal actions to maximize a numerical reward signal. Unlike most machine learning

approaches, explicit action instructions aren't provided to the model. Instead, it explores actions to discover those yielding the highest rewards. This involves analyzing the environment, executing actions to manipulate it, and assessing the outcomes. In intriguing and complex scenarios, actions can impact not only immediate rewards but also subsequent situations, affecting all future rewards. The emphasis isn't on characterizing learning algorithms, but rather on characterizing the learning problem itself. The distinctive attributes of trial-and-error search and delayed rewards define the essence of reinforcement learning.[46, 55]

Transfer Learning In machine learning, it's widely assumed that training and test data should share the same feature space and distribution. However, transfer learning introduces a family of algorithms that question and relax this assumption. These methods learn and transfer valuable insights from established data domains (sources) to novel data domains (targets), aiming to improve predictive performance in the target domain.[46, 56]

Multitask Learning Multitask learning diverges from single-task learning by simultaneously addressing multiple tasks that share the same set of features. It leverages domain information present in the training signals from these related tasks to learn a shared internal representation that enhances generalization. This approach functions as an inductive transfer strategy, enhancing overall learning by concurrently acquiring insights from each task and utilizing a shared representation to mutually benefit the learning process of different tasks.[46, 57]

2.2 Graph Machine Learning

The impact of deep learning was revolutionary in computer vision [58] and natural language processing [59], yet its applicability remained bounded by the prerequisites of data structure regularity. The convergence of network analysis and machine learning gives rise to graph machine learning (GML), enabling the utilization of graph structures and other non-uniform datasets (such as point clouds, meshes, manifolds, etc.).[4]
At the core of GML methods lies the fundamental concept of acquiring valid feature

representations for nodes, edges, or complete graphs. Notably, graph neural networks (GNNs), specialized deep neural network architectures tailored for graph-structured data, systematically enhance the node features within a graph through iterative information propagation from neighboring nodes.[4]
The majority of machine learning techniques applied to graphs are made of two distinct components: a versatile encoder and a task-specific decoder.[60] The encoder is responsible for embedding the graph's nodes or the entire graph into a feature space of reduced dimensions. For embedding complete graphs, a common approach involves first embedding individual nodes and then applying a permutation invariant pooling function (such as sum, max, or mean) to generate a representation at the graph level. On the other hand, the decoder computes the output that pertains to the specific task under consideration.[4]
Furthermore, GML tasks, much like traditional machine learning tasks, can be classified using various dichotomies: supervised/unsupervised, inductive/transductive, and node-level/graph-level. For example, predicting the chemical properties of small molecules based on their chemical structures is a supervised (inductive) graph-level task. In this scenario, the model leverages labeled data to learn how to predict chemical properties for given chemical structures. On the other hand, the task of identifying closely associated protein groups within a protein-protein interaction (PPI) graph falls under the category of unsupervised node-level tasks. Meanwhile, predicting the biological functions of proteins based on their interactions within a PPI graph corresponds to a node-level transductive task.[4]

Knowledge Graph Machine Learning

The application of Knowledge Graphs (KGs) spans numerous sectors, both in industry and academia, fueling extensive research into large-scale information extraction from diverse sources. Despite these endeavors, it's widely recognized that even the most sophisticated KGs are not complete or perfect.[61, 62] Consequently, researchers have investigated various techniques to fix inaccuracies and fill in missing information within KGs, a task often referred to as Knowledge Graph Completion or Knowledge Graph Augmentation. KG expansion can involve extracting new facts from external sources, experimentally generating new facts, or inferring missing facts based on the existing ones within the KG.[19]

This latter approach, known as Link Prediction (LP), has become an increasingly active research domain, particularly benefiting from the surge in machine learning and deep learning techniques. The majority of LP models leverages KG components to learn low-dimensional representations, commonly referred to as Knowledge Graph Embeddings, then using them to infer new facts.[19]

Traditional machine learning algorithms usually work by taking a feature vector as input and learning a mapping from this vector to a predictive output. In contrast, incorporating object's relationships into its representation, Statistical Relational Learning (SRL) is dedicated to constructing statistical models for relational data such as the structures found in Knowledge Graphs. SRL techniques can be applied to existing KGs to develop LP models that predict new facts (triples) by exploiting the underlying information within the existing facts.[63]

Utilizing the previously described notation (Section 1.1), where each KG fact is structured as a triple (h, r, t) with h representing the head (subject), r as the relation (predicate), and t denoting the tail (object), LP techniques predict the accurate entity to complete $(h, r, ?)$ (tail prediction) or $(?, r, t)$ (head prediction).[19]

The majority of LP models rely on embeddings to establish a scoring function $f(\boldsymbol{h}, \boldsymbol{r}, \boldsymbol{t})$* for assessing the credibility of a given fact (h, r, t). During the prediction phase, when presented with an incomplete triple $(h, r, ?)$, the missing tail entity is inferred as the one that, upon inclusion in the triple, produces the highest score:[19]

$$t = \underset{e \in \mathcal{E}}{\operatorname{argmax}} f(\boldsymbol{h}, \boldsymbol{r}, \boldsymbol{e}) \tag{2.1}$$

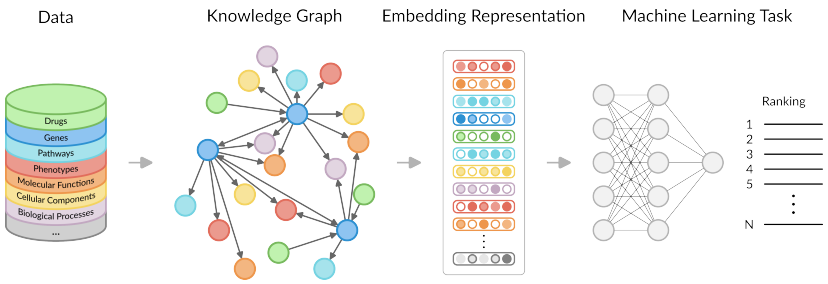


Figure 2.1. Knowledge Graph Machine Learning. The knowledge graph is constructed using the available data sources. Subsequently, vector representations (embeddings) are learned for entities and relations, which can be employed for a variety of machine learning tasks.

* Where $\boldsymbol{h}$, $\boldsymbol{r}$, and $\boldsymbol{t}$ are the embedding vectors of h, r, and t.

3 Networks in Drug Research

Network theory represents a potent analytical tool for exploring complex systems, particularly in the context of living organisms. Embracing the principles of network science, this approach captures the holistic behavior of a system and highlights emergent properties that result from intricate interactions among its components.[13] This methodology is useful in understanding the pathophysiological basis of diseases, conceptualized as emergent properties arising from complex interplays within living systems.[3] Recently, network models have played a pivotal role in characterizing drug-disease relationships and shedding light on various aspects of drug research.[3] These models depict nodes representing entities connected either physically or conceptually, forming a dynamic framework to untangle the intricacies of these complex systems.[43]

3.1 Network Biology, Pharmacology, and Medicine

In 1957, C. H. Waddington diverged from the prevailing Mendelian "one gene - one phenotype" model, pioneering a new perspective that highlighted the influence of gene networks on cellular states and developmental outcomes.[64] He proposed that phenotypes arise from stable conditions (states) of gene networks and the transitions between them.[64, 65] This concept lays the foundation for network biology, a discipline dedicated to unraveling the complex interactions among molecules constituting vital functional units driving physiological functions at the cellular, tissue, and organ levels.[66, 67]

A parallel transformation can be observed in the traditional target-centered approach to treating diseases, which traces its origins to the pioneering work of Paul Ehrlich and his renowned statement *"Corpora non agunt nisi fixata"*.[68] Despite being inherently reductionist, the "one disease-one target-one drug" model drove scientists to focus on molecules and their interactions, and served as the foundation for much of the research that led to the development of many of the medicines used nowadays.[3] Nevertheless, it has become increasingly evident that numerous effective drugs exert their effects on multiple targets rather than a singular one, a phenomenon called polypharmacol-

ogy.[69] Indeed, the integration of systems biology and the expansion of the "omics" technologies has prompted a shift in the drug discovery paradigm towards what is now known as network pharmacology.[3, 69]
This strategy finds its place within the framework of network medicine, an established discipline rooted in the application of principles of network theory and network biology to disease mechanisms and pharmacotherapy [70, 71]. The key concept of network medicine is that the comprehension of drug actions and the design of innovative pharmacological treatments extends beyond the consideration of isolated protein targets directly associated with a disease. Instead, it necessitates the consideration of the subnetwork of proteins interacting with the specific target(s) implicated in the disease; this interconnected group is commonly referred to as the "disease module".[3]

3.2 Network Types

Outlined below are some notable network types that hold significant relevance in the domain of drug research.

Signaling and Metabolic Networks The architecture of a signaling network consists of two main components: upstream signaling pathways and downstream gene regulatory networks. The upstream component is responsible for transmitting extracellular information to transcription factors through receptors and mediators. Its pathways are structured and characterized by cross-talks, which are directed interactions between them. Cross-talks, while enhancing functional diversity and adaptability, require precise regulation to maintain output specificity and input fidelity. Complementing this, in the downstream section, microRNAs play a crucial role in gene expression modulation by attaching to complementary sequences of DNA transcription factor binding sites.[43]
In metabolic networks, nodes represent metabolites, connected by edges indicating possible biochemical conversions. Edges depict reactions and, when the reaction is not spontaneous, the associated enzymes. While metabolites are relatively universal, the configuration of specific biochemical reactions linking them tends to be organism-specific.[43]
Analyzing signaling and metabolic networks that have undergone pathological changes

can reveal potential drug targets, whose dysfunction contributes to the development of specific diseases.[43]

Protein-protein Interaction Networks Protein-protein interaction (PPI) networks, often referred to as interactomes, depict relationships between proteins, with nodes representing proteins and edges denoting their direct physical interactions. Interactome data is acquired using diverse techniques like high-throughput methods, text mining, and predictions. Estimated at approximately 650,000 interactions,[72] the full complexity of the human interactome remains to be unveiled.
Data quality presents a key challenge due to sampling bias, missing interactions, false positives, and data coherence issues. For instance, large interactome hubs might emerge from aggregated data, neglecting critical protein attributes like conformations, posttranslational modifications, isoforms, expression differences and localizations.[73] It is worth highlighting the tendency for soluble proteins to have more connections compared to membrane proteins,[74] and that steric hindrances impose limitations on the maximum number of simultaneous interactions.[43]
Furthermore, it's noteworthy that interactome modules often align with significant cellular functions.[75] Additionally, essential proteins exhibit a higher degree compared to the average, and they tend to participate in a diverse range of functions.[43]
In drug research, specific network properties of PPI networks offer valuable insights. Disease-associated proteins typically don't serve as interactome hubs, except in cases like cancer, where tightly interconnected hubs emerge due to somatic mutations.[76–78] Potential drug targets often bridge multiple network modules and exhibit more connections than the average protein node.[43] To minimize side effects, it's crucial to target non-hub nodes with intermediate neighbor counts, ensuring controlled perturbations in the interactome.[43]

Protein Structure Networks In protein structure network representations (also called protein contact networks, amino acid networks, residue interaction networks, or protein meta-structures), nodes depict amino acid side chains, and edges are shaped

by physical distances between these side chains (measured between α or β carbons* or the side chains' center of mass). In unweighted networks, edges are drawn when two side chains are closer than a threshold distance (usually ranging from 4 to 8.5 Ångström†).[43, 79]
Protein structure networks exhibit "small world" properties, facilitating rapid drug-induced conformational changes through efficient communication among amino acids.[80–82] Protein structure networks have modules, which often encode protein domains. High-centrality, inter-modular bridges play a key role in the transmission of allosteric changes. Protein structure networks serve as efficient tools for identifying crucial amino acids involved in intra-protein signal transmission.[83]

Drug-Target Networks In a drug-target (DT) network, if a protein is recognized as a target for a drug, a connection is established between them. This network configuration, represented as a bipartite graph, enables the extraction of two biologically insightful network projections, each encompassing distinct node types: drugs or targets. The drug projection (DP) network, exclusively comprises drug nodes, where connections link two drugs sharing a common target. Similarly, the complementary target projection (TP) network features protein nodes, and edges are drawn when two proteins are targeted by at least one mutual drug. The visualization of drug-protein associations within this network framework offers a valuable overview of the ongoing landscape of drug discovery.[37, 84]

Drug-drug Interaction Networks Within drug-drug interaction networks, nodes correspond to drugs, while edges depict the interactions between them. These interactions arise when one drug's pharmacological effects is modified by another drug, frequently resulting in clinical outcomes that are difficult to predict, including adverse drug reactions. Such interactions can lead to diverse effects, wherein one drug may either amplify or attenuate the intended effect of the other drug, and in some cases,

* In organic chemistry, carbon atoms are designated with Greek letters, starting from the carbon atom immediately following the one bonded to the functional group. In the context of amino acids and proteins, the α-carbon is the carbon atom in the backbone (which is the central chain of atoms in the amino acid structure) adjacent to the carbonyl carbon, and the β-carbon occupies the subsequent position as the second one in line.

† 1 Ångström $= 10^{-10}$ meters.

even trigger an unexpected response.[85]

Chemical Structure Networks The structure of chemical compounds can be visualized as a network, often referred to as a chemical graph.[7] Here, atoms within the molecule serve as nodes, and the covalent bonds that link them together form the edges. This network representation can accommodate multiple edges to depict various bonding interactions. The descriptive attributes from this network framework are valuable assets in quantitative structure-activity relationship (QSAR) models.[43]

Chemical Space Networks Chemicals can be described based on diverse properties, such as molecular mass, lipophilicity, and topological features. These descriptors collectively define a multi-dimensional space, often termed chemical space.[86] This concept holds growing significance across various chemical domains, particularly in medicinal chemistry and chemical biology.[87] However, conventional representations of chemical space suffer the "curse of dimensionality"*.[88, 89] This challenge can be effectively addressed by adopting a network-based approach, allowing a more manageable representation.[87]
Indeed, chemical space networks (CSNs) have emerged as a valuable approach to visualize and comprehend relationships within datasets of small molecules.[87, 90] In a typical CSN, nodes represent compounds and are linked by edges that can convey diverse relationships between them.[90] These relationships encompass fingerprint-based Tanimoto similarity[91], substructure-based similarity[92], or asymmetric Tversky similarity[93], among others. When CSN edges reflect a continuous spectrum of similarity values, such as fingerprint-based Tanimoto similarity, the incorporation of a minimum threshold value permits to control the number of edges. Specifically, an edge is drawn if the relationship value between compounds matches or surpasses the specified threshold. Conversely, in scenarios like CSNs based on matched molecular pair (MMP)†, edges indicate binary relationships, connecting only MMPs.[92, 95]

* As the number of dimensions increases, the volume of the space expands rapidly, leading to sparse data distribution. Ensuring reliable outcomes often demands exponentially larger datasets as dimensionality grows.

† A pair of compounds that differ from each other by a single modification at a specific structural site.[94]

Knowledge Graphs Basic network models have long been employed to represent complex interactions in biomedical systems. While these models have shown impressive performance, they often struggle to capture the semantic richness of diverse relationships among biomedical entities. To overcome this limitation, recent approaches have shifted towards employing multi-relational networks, exemplified by knowledge graphs.[96] Knowledge graphs integrate information from expert-derived sources into a graph structure, where nodes represent biomedical entities and edges symbolize relationships between them.[16]
KGs find extensive applications in various biomedical challenges, spanning from elucidating protein functions to prioritizing disease-associated genes and suggesting safer drug alternatives for patients. They offer valuable insights into drug properties, such as predicting interactions between drugs, identifying potential molecular targets for drugs, and discovering novel therapeutic possibilities for established drugs.[16]
One prevalent approach is KG embedding (KGE), showcased in projects like Hetionet[97] and PharmaKG[96], which maps the intricate graph structure into a lower-dimensional space while retaining essential topological features.[96]

3.3 Network Applications

The process that leads from the identification of a disease or clinical condition to the marketing of a drug treating it, is long and complex.[98] It starts with target identification and validation, followed by hit* discovery and the selection of lead molecules†. These leads then undergo optimization through chemoinformatics, drug efficiency assessments, and ADMET (drug absorption, distribution, metabolism, excretion and toxicity) studies, ensuring desirable drug candidate characteristics.[4, 43, 98] Subsequently, pre-clinical research evaluates toxicity, pharmacokinetics, pharmacodynamics, and efficacy through in vitro and in vivo studies. With compelling preclinical data, the drug advances to three successive phases of human clinical trials.[4, 98]
In the review titled "Network modeling helps to tackle the complexity of drug–disease

* A primary active compound, demonstrating non-promiscuous binding behavior and surpassing a specified threshold value in a given assay or assays.[99]

† Prototypical compounds exhibiting activity and selectivity in a pharmacological or biochemically relevant screening process.[99]

systems"[3], various methods for modeling complex drug-disease systems were examined. Below, several crucial stages within the drug discovery and development process that can notably benefit from the application of network tools are presented.

Molecular Properties Prediction Selecting molecules with heterogeneous yet desirable chemical properties can help avoid the need to screen millions of compounds, resulting in reduced costs and time for drug discovery.[4, 46] Recent research has demonstrated that utilizing graph-based representations results in superior predictive accuracy for molecular property tasks when compared to non-graph methods, while also providing enhanced model interpretability.[4, 100]

Target Identification Target identification involves the search for a key molecular target pivotal to the pathophysiology of a disease, susceptible to modulation for therapeutic impact, hopefully reversing the disease condition. Systems biology employs a network approach, bridging disease biology and genetic influences to uncover such intervention points. These targets, then, necessitate experimental validation and comprehensive evaluation, including factors like accessibility and efflux pumps, to ensure their viability.[4]

De Novo Design In drug design, generating molecules with a high likelihood of interacting effectively with a specific target is vital.[44] The approach of in silico de novo drug design has emerged as an effective strategy for narrowing down the chemical space, facilitating the discovery of compounds for chemogenomic research and initiating hit-to-lead optimization.[46] This method revolves around producing new or modified molecular structures that possess desired properties. Traditionally, this challenge is tackled through inverse QSAR problems, generating structures within the constraints of established QSAR models. This involves defining an inverse-mapping function that translates molecular activity into chemical descriptors, which in turn guides the creation of novel compounds.[101]
Recent years have witnessed the rise of AI, particularly deep learning models, as potent tools for de novo molecular design.[44, 101] Initial AI-driven drug discovery pre-

dominantly employed SMILES* strings as input representations, capitalizing on pre-existing sequence learning architectures that were adapted for this purpose.[44] While successful in generating credible molecules, the SMILES approach harbors inherent limitations compared to graph-based methods. SMILES models must simultaneously accommodate both chemical rules and grammar, thus wasting valuable representation resources. Graph-based models, instead, circumvent the need to learn grammar rules, therefore conserving representation capacity for the primary learning task and offering an inherently more efficient alternative to SMILES representations.[44] Additionally, subgraphs within this approach can be interpreted as molecular fragments, facilitating more intuitive analysis. Furthermore, graph-based models enable the explicit incorporation of chemical constraints on complete molecules and fragments.[44]

Drug Repurposing In response to escalating drug development costs, researchers are exploring the utilization of existing drugs, whether approved or in development as therapies, for indications beyond their original intended use.[4, 46] The rationale behind this is that recommending an existing molecule is arguably simpler than designing one from scratch.[4]
Repurposed treatments are estimated to constitute around 30% of newly FDA-approved drugs and their associated revenues, with the potential to repurpose up to 75% of existing entities.[4] Repurposing is possible due to many drugs having multiple targets, and the presence of shared characteristics among diverse diseases, including genetic factors, molecular pathways, and clinical features.[46] An important incentive for drug repurposing is addressing unmet therapeutic needs of rare or neglected diseases.[46]
Computational drug repurposing has evolved from traditional approaches like chemical similarity assessment and molecular docking to innovative methodologies rooted in systems biology.[46]

Drug Combinations Combination drugs have proven particularly effective in diseases characterized by complex aetiology or evolutionary components, which commonly give rise to treatment resistance. They can enhance convenience and compliance through fixed-dose formulations, achieve synergistic effects, broaden the thera-

* The Simplified Molecular Input Line Entry System is an efficient chemical notation system designed for chemical information processing, which employs a compact and intuitive grammar based on molecular graph theory.[102]

peutic spectrum, and combating disease resistance, among other benefits.[4]
The vast number of possible pairwise combinations involving just two drugs presents a daunting combinatorial challenge for a brute force empirical laboratory testing approach. To put it into perspective, considering approximately 4,000 approved drugs, it would be necessary to conduct $\sim$8 million experiments to assess all conceivable two-drug combinations at a single dose. Additionally, these combinations would need to be evaluated across roughly 3,000 human diseases. Moreover, the potential variations in dosage, timing of treatments, and delivery methods are virtually limitless.[4]
Efforts to systematically identify drug combinations that deliver both substantial clinical efficacy and minimal toxicity often rely on intuition and experience rather than well-defined principles.[103] Network-based methodologies have provided a promising foundation to move beyond the conventional "one-drug, one-target" mindset, enabling the exploration of the "multiple-drugs, multiple-targets" paradigm by targeting multiple disease proteins within a disease module, while also aiming to mitigate toxicity.[4]

Drug Safety The safety profile of a compound is determined by its adverse impact on an organism or its components, which could encompass cells, organs, and other substructures. Assessing toxicity stands as one of the most central and demanding stages within the drug discovery and development process. Since developing reliable high-throughput assays for extensive in vivo and in vitro bioassays is a costly and time-consuming endeavor, there is a strong need for rapid, cost-effective, and consistent computational alternatives.[46]
The efficacy of deep learning applications in this context is closely tied to the quality of encoding functions used to map molecular structural information into fixed-size vectors. To address this limitation, various graph-based learning architectures have been introduced, which are designed to automatically extract suitable features from raw molecular graphs.[46]

4 Aim of the Work

This thesis aims to make optimal use of available biomedical knowledge investigating approaches to support drug research, with a specific focus on drug repurposing and drug safety. By integrating network theory and data science techniques, the objective is to unveil novel insights into the complex relationships between biological components and drugs. Through a series of projects ranging from descriptive analyses to predictive modeling, this thesis endeavors to advance our comprehension of drug effects, contributing to the development of safer and more efficient pharmaceutical interventions. Ultimately, this work seeks to bridge the gap between traditional pharmaceutical research and the emerging paradigms of network analysis and machine learning, offering perspectives for the future of drug discovery and healthcare.

Part II

Projects

5 Projects Overview

The projects undertaken during the PhD course are presented here with the structure of scientific articles, encompassing introductions, methodologies, results and conclusions.
These projects can be categorized into three main types: descriptive conceptual models, aimed to interpret the complex characteristics of biological systems; computational models designed for predictive tasks on biological systems; and a plain data analysis project.
The objectives and areas of focus of each project are:

- Exploration of the clinical drug research response to COVID-19 (Section 6.1).
- Identification of crucial regulators from gene expression data (Section 6.2).
- Inspection of drug-induced impulsivity associated to dopaminergic agents (Section 6.3).
- Selection of potential repurposable drugs and combinations to treat Huntington's disease and multiple sclerosis (Section 7.1).
- Collection and aggregation of relevant biological data into a KG to make predictions based on it (Section 7.2).
- Investigation of adverse events that are more frequent in fixed-dose drug combinations compared to the individual components (Section 8.1).

6 Descriptive Models

6.1 COVIDrugNet

The COVID-19 pandemic, caused by SARS-CoV-2, presented an unprecedented global health issue, necessitating a massive, and diversified but concerted response from the scientific and medical communities. Among the many strategies employed to combat the virus, drug development and clinical trials have played a crucial role in seeking effective therapeutic solutions. In this context, a valuable resource was introduced in the form of the web-based application COVIDrugNet, as described in the paper "COVIDrugNet: a network-based web tool to investigate the drugs currently in clinical trial to contrast COVID-19" published in *Scientific Reports* in 2021.[37] This tool offers a user-friendly interface to comprehensively and continuously monitor the evolving landscape of drugs in clinical trials for the treatment of COVID-19. The pressing need for a platform like COVIDrugNet arose from the dynamic nature of clinical research during a pandemic, where the identification of promising therapeutic candidates and understanding their biological and pharmacological implications is of paramount importance. COVIDrugNet, accessible at http://compmedchem.unibo.it/covidrugnet, provides a means for us to navigate this complex ecosystem of drug trials, facilitating real-time exploration and analysis.
In this paper, we describe the functionalities of COVIDrugNet and present illustrative examples of how this tool can be employed to gain insights into the ongoing clinical trials. By harnessing the power of network-based analysis, we demonstrate how COVIDrugNet enables users to probe the consistency of therapeutic approaches with existing biological and pharmacological evidence. Such analyses can help in comprehending the implications of proposed drug options and, ultimately, in guiding the search for more effective therapies against COVID-19.

Details

Authors Luca Menestrina, Chiara Cabrelle, Maurizio Recanatini

Type Research Article

Status Published

Title COVIDrugNet: a network-based web tool to investigate the drugs currently in clinical trial to contrast COVID-19

Journal Scientific Reports

DOI 10.1038/s41598-021-98812-0

Data Availability The full code for the collection, building and analysis of the networks is available in the GitHub repository at https://github.com/LucaMenestrina/COVIDrugNet. It is entirely written in Python. All data generated or analyzed in this study is publicly available on the GitHub repository. Furthermore, some data is easily downloadable from the web tool itself: all tables in tab-separated values (tsv) format and the networks in various formats (adjacency list, pickle, cytoscape json, graphml, gexf, edges list, multiline adjacency list, tsv, png and jpg).
Supplementary data can also be accessed at the original publication.

COVIDrugNet: a Network-based Web Tool to Investigate the Drugs Currently in Clinical Trial to Contrast COVID-19

6.1.1 Introduction

The outbreak of the COVID-19 pandemic caused by SARS-CoV-2 at the beginning of 2020 has shocked the population worldwide. A year and a half later, (August 2021) about 200 million confirmed cases of COVID-19 have been reported by WHO included more than 4.2 million deaths (https://covid19.who.int/). As expected, in such a mankind threatening situation, the scientific community put in place a great effort to help countering the spread of the virus, as evidenced among the other things by the huge number of papers dealing with various aspects of the disease appeared in the literature. For instance, the LitCovid literature hub[104] has collected around 160,000 articles as of August 2021 covering arguments categorized as general, mechanism, transmission, diagnosis, treatment, prevention, case report and forecasting.
As regards the COVID-19 treatment, the race to the vaccine against SARS-CoV-2 started immediately after the isolation of the viral genome[105] and gave the first results as soon as December 2020. Moreover, despite the exploration of different approaches like, e.g., the infusion of plasma from human survivors[106], the pharmacological option, namely small molecule drugs and antibodies, is being actively pursued. However, the route to a new drug is long and costly, and the classical drug discovery pipeline is not compatible with the need of rapid intervention on a population of millions of patients. At the moment, a viable alternative seems to be the repurposing of known drugs[107], i.e., the use for the treatment of COVID-19 of drugs currently on the market for different therapeutic purposes.
Known drugs that are currently in clinical or pre-clinical study for the treatment of COVID-19 are aimed either at inhibiting viral or human targets involved in some of the processes of viral entry and replication, or at treating inflammation and tissue injury consequent to the viral infection[108, 109]. Even though it might seem that a direct antiviral approach could lead to a straightforward solution, only few of the existing antivirals have performed well in the clinic so far. On the other hand, a number of drugs used for the most disparate therapeutic indications and entered into clinical

trials even with an uncertain rationale[110] are showing preliminary promising results. However, as it has been observed[111], a real "repurposing tsunami" has invested the biomedical community, so much so that today it is difficult not only to keep track of the results of the trials, but also to follow the new proposals.

With the aim of helping researchers navigate the sea of outcomes and reports coming from the studies on COVID-19, some institutions and companies have developed online platforms that collect and organize both literature and data, eventually providing free access to the latter. For example, the already mentioned LitCovid hub[104] (https://www.ncbi.nlm.nih.gov/research/coronavirus/) is a daily updated source of relevant articles retrieved from PubMed. Other platforms dealing with data on drugs and chemicals, like, e.g., CHEMBL[112] (https://www.ebi.ac.uk/chembl/), PubChem[113] (https://pubchem.ncbi.nlm.nih.gov/), or DrugBank[114] (https://www.drugbank.ca/), have introduced special sections dedicated to COVID-19-related information. In addition, more specialized resources have appeared on the web to help accessing and analyzing COVID-19 data, mainly in the fields of epidemiology, genomics, interactomics, and, to a lesser extent, pharmacology. In this class of web tools, it is worth mentioning CORDITE (CORona Drug InTEractions database)[115], a web interface that provides a database of potential drugs, targets, interactions, and relative publications obtained from a manually curated selection of literature sources. With the same purpose of facilitating the data analysis, the COVID-19 Drug and Gene Set Library was built as an online collection of COVID-19 related drugs and genes.[116] A comprehensive critical review on this kind of web tools has recently been published by Mercatelli et al.[117]

Considering the great amount of valuable scientific information that has already been produced and published, and that will be presumably produced for some time more on COVID-19 related topics, it could be useful to look at the whole scenario of results, to foster the acquisition of that knowledge that can only emerge from consideration of both the totality and the complexity of data. In other words, and limiting ourselves to the pharmacological treatment issue, one might think of presenting and analyzing the information on proposed drugs in a way that takes into account not only the different types of data (chemical, biological, genomic, etc.), but also the relationships among them, that is on a network basis. The context is that of network medicine.[118] An attempt in this direction has recently been proposed by Korn et al.[119], who devel-

oped a knowledgebase and an online platform (COVID-KOP) to integrate the existing biomedical information with the newly acquired knowledge on COVID-19. By means of this web tool, one can easily produce an aggregate graph connecting, e.g., COVID-19 phenotypic features to a drug studied for treating the disease, through the genes known to be linked to both. Still in the context of network medicine, CoVex is another platform that offers the user the possibility to explore the SARS-CoV-2 virus-host-drug interactome for drug repurposing aims.[120] In addition, we want to mention CovMulNet19[121] that at present looks like the most thorough network-based tool allowing to integrate the available genotypic and phenotypic information on COVID-19, like, SARS-CoV-2 proteins, their human partners, as well as symptoms, diseases, and drugs. Finally, Coronavirus canSAR[122] is a freely available resource that offers druggable interactomes of SARS-CoV-2 proteins and human proteins, as well as reports about 3D structures, drugs, and clinical trials.

In a specifically drug-focused context, the network medicine approach assumes the overcoming of the old "one drug, one target, one disease" concept in favor of a more outright "multi-drug, multi-target, multi-disease" approach.[69] The exploration of a such complex system of interactions can be aided by the construction of a drug-target network.[84] In reference to the COVID-19 case, drug-target networks based on host-virus protein-protein interactions (PPIs) have already been built and examined[123–125] with the aim of repurposing already approved drugs.

Here, we present the COVID-19 Drugs Networker (COVIDrugNet: http://compmedchem.unibo.it/covidrugnet), a web application that offers a different point of view on anti-COVID-19 drugs by allowing a network-based analysis of the DrugBank dataset of potential repurposed drugs currently in clinical trial. The freely accessible application automatically retrieves the data from DrugBank, builds the drug-target network, and allows the user to carry out some basic network analysis. Moreover, we show how, using COVIDrugNet, some peculiar aspects of the proposed pharmacological options against COVID-19, in terms of substances, targets, and their interrelationships can be revealed. Although what is reported here is an instant analysis based on current data, the continuous updating of COVIDrugNet will allow us to follow the future development of the drugs proposed for the treatment of the disease, thus providing an always updated view of the COVID-19 system pharmacology.

6.1.2 Results and Discussion

6.1.2.1 COVID-19 Drugs Networker

The COVID-19 Drug Networker (COVIDrugNet, Figure 6.1) is a web tool designed for the exploration of the landscape of the drugs currently in clinical trials to combat the SARS-CoV-2 infection. The web app is based on a network approach that supports both visualization and analysis of the complex scenario of repurposed drugs for the COVID-19 and related conditions. The core of the web tool are the interactive graphs and the additional features that allow one to explore drug and target data, as well as networks properties. The main graph represents a bipartite Drug-Target network (DT network, Figure 6.2a), where the nodes are drugs and targets that are connected if a relation between them is reported in DrugBank. Since bipartite networks are usually investigated by compressing their information into two monopartite networks called projections[126], COVIDrugNet provides two of such networks only having drugs or targets as nodes: in the following, we refer to them as Drug and Target projections, (DP and TP, respectively; Figure 6.2b and 6.2c).

As regards the user interface, it is basically divided into the main and the *Advanced Tools* blocks. The first one allows users to immediately access the main body of information, capturing the holistic view of the current drug repurposing status for COVID-19. However, a more in-depth examination of the data is possible, by taking advantage of some more specialized graph analysis tools provided in *Advanced Tools*.

In detail, the main block includes the graph, and the *Charts and Plots* and *Graph Properties* sections (Figure 6.1). As mentioned before, the heart of each web page is the interactive graph with its related information box (*Node Info*) that provides a summary documentation of single drug/target nodes hovered over or individually selected. The box contains links to some databases providing the available information related to individual properties of both drugs and targets. In addition, a multiple node selection brings up the *Inspected Data* hidden section that displays detailed information of the selected nodes in a tabular format. By the way, networks and tables can be downloaded in different formats to allow an external analysis of the data.

Node coloring options are provided, useful to visualize some node attributes related to therapeutic, biological, or network-based features. For instance, in the DP graph (Figure 6.2b) the user can decide to color the nodes according to the Anatomical Ther-

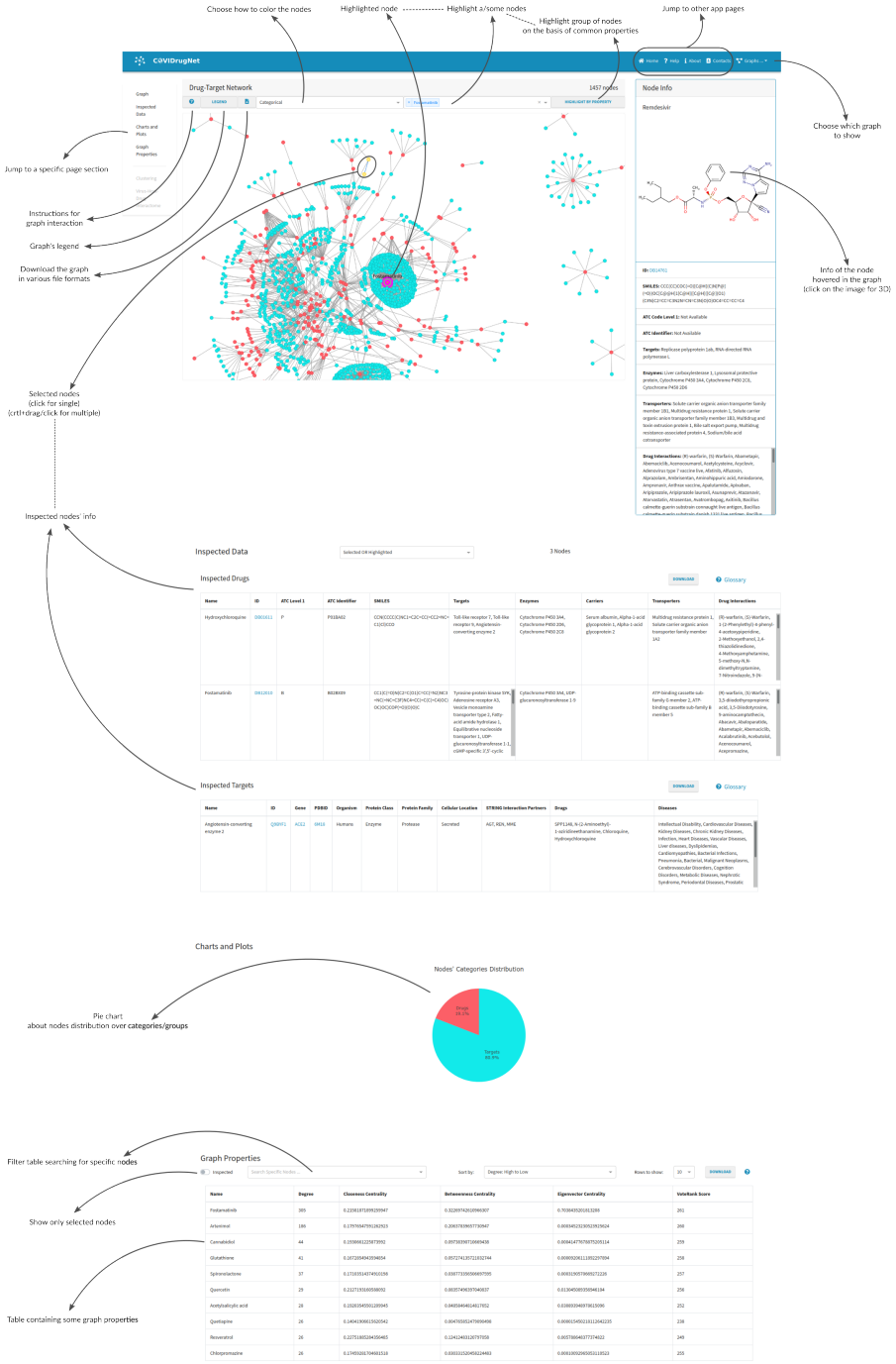


Figure 6.1. The COVIDrugNet Web Tool. A screenshot of the main block of the Drug-Target Network page. It displays the fundamental features accessible in the web tool that allow the user to inspect the network and its properties.

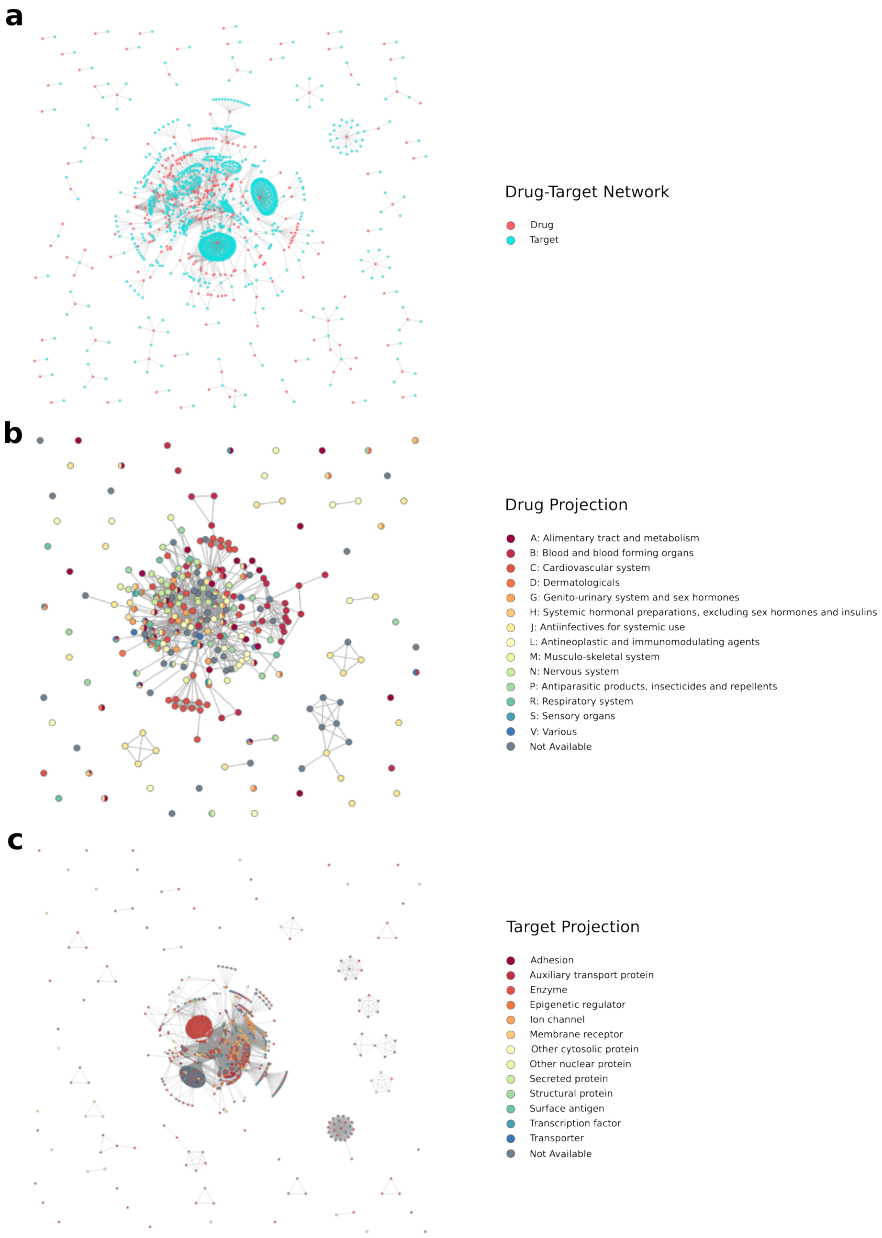


Figure 6.2. COVIDrugNet Networks. The three networks generated and available for inspection in COVIDrugNet. (a) Drug-Target Bipartite Network. It is the main network, and it is built connecting drugs currently in clinical trial present in the COVID-19 Dashboard of DrugBank[114] and their reported targets. The red nodes are drugs, and the light blue ones are targets. (b) Drug Projection. It is built from the Drug-Target network and contains only drugs. The nodes are color coded on the basis of their first level ATC codes (retrieved from DrugBank[114]). (c) Target Projection. It is built from the Drug-Target network and contains only targets. The nodes are color coded according to their protein class (retrieved from ChEMBL[112]). The networks were generated by means of the Python package NetworkX[127].

apeutic Chemical (ATC) code or the clinical trial phase, while in the TP graph (Figure 6.2c) the color coding allows one to spot protein family, protein class or cellular location. Moreover, in the DT network and in both projections, it is possible to color the nodes based on some network attributes - i.e., degree, centrality measures or node grouping - considering the entire graph or the major component. To examine all these properties at a glance, the web tool also provides the *Chart and Plot* section, in which the pie charts - or bar chart in the case of the ATC code coloring option - are updated accordingly to the node coloring option to show the relative proportions between the values of that property. In this area of the projection web pages, the web tool also provides the plot of the nodes degree distribution. Among the graph interactive features, the *Highlight a node* dropdown menu is useful to find nodes by name, and the button *HIGHLIGHT BY PROPERTY* allows a customized filtering on node properties to highlight and/or download a specific nodes selection. In the *Graph Properties* section, some centrality measures useful to analyze the network topology are displayed in a downloadable table. A short explanation of each computed property is provided in a Glossary in the Help page.

Regarding the *Advanced Tools* block, it contains three sections: *Clustering*, *Advanced Degree Distribution* and *Current Virus-Host-Drug Interactome*. The *Clustering* section is dedicated to the node grouping analysis carried out through different methods (see Nodes Grouping in Section 6.1.5.3). In particular, we thought it could be of interest to examine the grouping of the nodes in the projection graphs, as, e.g., in perspective it might reveal possible trends in the selection of drugs to be repurposed or privileged areas of intervention in the biology of the infected cells. To this aim, the web app allows for three different techniques of investigation of the networks partitioning: spectral analysis combined with K-means clustering[127], Girvan-Newman[27] and greedy modularity community detection[30] methods. The plot in this *Clustering* section reports either the eigenvalues distribution used in the application of the spectral clustering method, or the modularity trend in the Girvan-Newman community detection method. Both plots are interactive and allow the user to choose the level (number) of grouping.

The *Advanced Degree Distribution* section presents an interactive chart of the degree distribution and some of its possible distribution fittings compared to those of an Erdős-Rényi equivalent graph (see Degree Distribution Fitting in Section 6.1.5.3).

Finally, the *Current Virus-Host-Drug Interactome* section displays a bipartite network built on the basis of experimental studies and checked for protein targets present in the DT network (see below for details). As mentioned before, the network table is downloadable, to provide interested users with the possibility of rebuilding and manipulating the graph.

6.1.2.2 Graphs Analysis

In Figure 6.2, the graphs representing the networks generated by COVIDrugNet are displayed. The DT network is a disconnected network with a large connected component accounting for 85.1% of nodes (1248 out of 1466). This structure reminds that of the general drug-target network reported elsewhere[84], where most drugs have more than one target and several drugs can share the same target(s). However, from inspection of the graph, it immediately appears that there are two drug nodes that heavily affect the network topology by showing an exceedingly high degree compared to all other nodes: Fostamatinib and Artenimol, having 305 and 186 direct neighbors, respectively. For both drugs, this reflects a number of reported targets that is considerably higher than the average (<7), being 6.9 and 4.2 times higher, respectively, than that of Cannabidiol that, with 44 targets, is the third in rank for the highest number of neighbors in the DT network. Indeed, these two drugs show a peculiar behavior strongly affecting the network structure not only in the DT, but consequently also in the TP graph where they cause the formation of two highly intra-connected clumps of nodes. To take this aspect under consideration and possibly clarify its role in respect to the topology of both the whole drug-target network and the projections, in the following, we compared the results of the network analyses carried out on the entire networks and on the graphs containing all nodes except Artenimol, Fostamatinib and their exclusive direct neighbors.

As a first step in the analysis, we tried to assess the character of the monopartite projection networks DP (290 nodes) and TP (1176 nodes), i.e., whether they belong to the random network category or are scale-free. Scale-free networks have a characteristic organization, in which there is a limited number of nodes with a high number of neighbors (called hubs) and an abundance of nodes having a low degree.[24] This arrangement can be found in plenty of real-world networks, from the World Wide Web to citations in science, from social interactions to metabolic maps.[24, 38] Both DP

and TP show a significant difference from an equivalent (same number of nodes and probability of edge creation) Erdős-Rényi graph[9] (Figure A.1). To further investigate on the scale-freeness of the networks, we considered three properties for each graph: the degree distribution, the relationship between clustering coefficient and degree, and the ability to withstand targeted attacks compared to random failures.
In order to address the scale-free character of both networks by evaluating the fitness of the degree distribution to a power-law, we employed the approach reported by Broido et al.[33], which applied a previously defined rigorous method.[128] This analysis was carried out on both the entire DP and TP networks and also in cases where Artenimol and Fostamatinib as well as their exclusive direct neighbors were removed.
In the DP network, the degree distribution could be described by a power-law, suggesting that these networks are plausibly scale-free (Figure A.2a,b). However, other heavy-tailed distributions cannot be ruled out (Figure A.2c,d). The situation for the TP network is less clear-cut, at least in the case of the entire network. To advance an explanation for these results, we observe that, these networks are small, such that they would probably not provide enough data for clearly electing a distribution form. Still, they are unequivocally dissimilar to random networks.
The inspection of both the clustering coefficient and the robustness evaluation is best illustrated considering the two projections one at a time.
Looking at the DP network and specifically at its clustering coefficient, it shows a tendency to decrease as the degree increases (Figure A.3a,b), implicating the existence of a few hubs connecting peripheral nodes of high degree. Also, there is an evident distinction between the response to a targeted attack and to a random failure[39] (Figure A.4a). In the first case, nodes with the highest degree are progressively removed from the network, causing it to break apart quickly. On the other hand, if the nodes to be dismissed are chosen randomly, the connectedness of the network is almost unaffected. Notably, these findings are strengthened by the fact that carrying out the same investigation on a network from which Artenimol and Fostamatinib are excluded, leads to almost identical results (Figure A.4b).
The same examination carried out on the TP network does not yield equally unambiguous conclusions. As stated above, the targets linked to Artenimol and Fostamatinib compose two almost-clique aggregations, which distort the morphology of the network. The relationship between clustering coefficient and degree is strongly de-

pendent on the presence of these two exceptionally connected drugs (Figure A.3c,d). When they are not taken into account, the inverse proportionality is fairly visible. Nevertheless, if they are considered, the scatterplot displaying this relationship is warped, due to the formation of two separate but remarkably dense groups representing the targets connected to Artenimol and Fostamatinib. The check of the robustness of the network by comparing the responses to targeted attacks or random failures gives a result that agrees with that obtained from the DP network. The communities related to the two "super-spreaders" simply introduce a delay in the fragmentation of the network, since they are made of a multitude of nodes with equally high degree (Figure A.4c,d). Anyhow, this shift does not alter the network robustness to random failures and the susceptibility to targeted attacks.

As a final remark on the networks organization, we stress that all results and conclusions presented here are just a snapshot of the continuously evolving COVID-19 drug repurposing scene, and that it will be worthwhile to follow the time progression of this system. For instance, in the future, the growth of the network could smooth out or even hide the effects of Artenimol and Fostamatinib that now we observe so evidently. In respect of this, we recognize a different response of the DP and TP networks to the influence of these nodes. The former is less affected, since the vast majority of the targets related to both drugs are not shared by others, such that the information related to these proteins vanishes in the projection process. On the contrary, the latter suffers a huge impact, showing a situation that is antithetical to the previous one. Here, the proteins amass together constituting two highly intra-connected jumbles, which are poorly linked to the rest of the network. A continuous growth and the ability of self-organizing are two key features of scale-free networks, which frequently describe real complex systems[24]. These characteristics are shown by both projections, and indeed their scale-freeness is supported by their degree distribution, the relation of clustering coefficient to degree, and their robustness. Mainly due to the influence of Artenimol and Fostamatinib, these properties are manifest in the DP network, but not so neat in the TP one.

6.1.2.3 Applications to COVID-19 Repurposed Drugs: Network-based Inferences

To illustrate the capabilities of COVIDrugNet, in the following we report some example considerations that can be derived from the analysis of the DT network, and of the projection graphs relating to drugs (DP) and targets (TP).

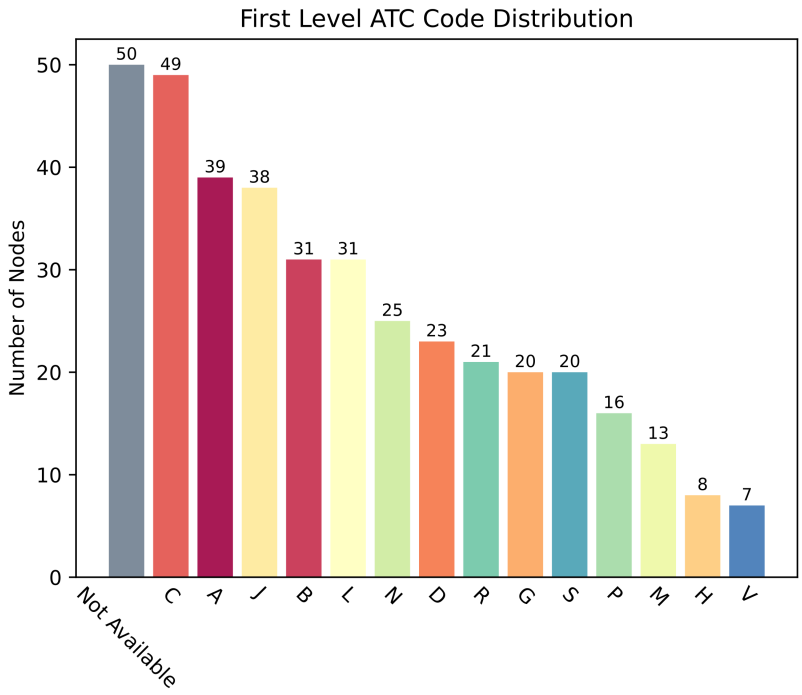


Figure 6.3. First Level ATC Code Distribution. A bar chart displaying the count of nodes for every first level ATC code (anatomical/pharmacological main group). The total count is higher than the number of nodes in the DP, because more than one ATC code can be assigned to a single drug.

Drugs Examining the DP network with nodes colored by ATC code (https://www.whocc.no/atc/structure_and_principles/) (Figure 6.2b) can reveal at a glance which therapeutic areas are mostly covered by the repurposed drugs presently in clinical trials. In the *Charts and Plots* section of the COVIDrugNet Drug Projection page, the nodes categories distribution is shown, from which it appears that all the 14 main anatomical/pharmacological groups (1^{st} level codes) are represented, even though with different numbers of drugs. Not taking into consideration the 50 substances for which an ATC code is not yet reported, the remaining 240 drugs are distributed in three top ranked groups: C (Cardiovascular system), A (Alimentary tract and metabolism), and

J (Antiinfectives for systemic use) comprising 49, 39, and 38 active substances, respectively (Figure 6.3). Then, two other highly populated ATC groups follow: B (Blood and blood forming organs), and L (Antineoplastic and immunomodulating agents) both counting 31 drugs. By considering the composition of the bars that reports the distribution of drugs in the 3rd level groups for each 1st level ATC code (visible in the web tool), one can have a more detailed picture of the actual pharmacological approaches to COVID-19 treatment. First, it is worth noting that the drugs belonging to the J group are located mostly out of the main connected component of the graph, accordingly to the fact that they share a target with a very small number of other drugs. Conversely, substances of the A and C groups mostly populate the main connected component, indicating a high level of promiscuity among them as regards the targets. Also, we observe that most drugs classified in the C, A, J, B, L, N, and P groups show just one ATC code, while drugs in D, G, R, S, M, and H belong to more than one ATC group. Even though the ATC system is not aimed at providing direct therapeutic indications and considering also that more than one code can be assigned to individual medicines, the landscape of pharmacological interventions against the SARS-CoV-2 infection emerging form the DP network appears rather intricate. Overall, it mostly confirms that the drugs in clinical trials are aimed at contrasting both the viral infection process (antivirals in J group, agents acting on the renin-angiotensin system in C group), and its pathological consequences at systemic level (substances in A, B, L, and other groups). These approaches are in line with evidence recognizing that, as the severity of the COVID-19 increases - apparently in consequence of a dysregulated host immune response - various pathophysiological mechanisms are activated leading to hematological (mainly thromboembolic) manifestations and, eventually, multi-organ dysfunctions.[129, 130] In addition, bacterial superinfections have been reported in COVID-19 patients, and even though the issue is still debated[131], antibiotics belonging to the J group are actually in the current treatment guidelines.[132]

Indeed, even this brief analysis of the ATC codes distribution among the substances currently in clinical trials highlights a complex and multifaceted drug repurposing scenario consequent to the fact that the COVID-19 is a multi-systemic disease requiring a well-equipped therapeutic armamentarium and possibly a combined polypharmacological intervention.[133]

To provide an example of using COVIDrugNet focused on a group of drugs, we could

take into consideration the inhibition of the virus attachment and entry into the host cell. It is believed that SARS-CoV-2 enters the target cell mainly through an endocytic pathway that exploits the ability of its Spike (S) protein to bind the human Angiotensin-converting enzyme 2 (ACE2) receptor. Subsequently, S is cleaved by the Transmembrane protease serine 2 (TMPRSS2) to provide the S2 subunit necessary for the membrane fusion.[134, 135] The drug repurposing activity aimed at preventing this step of the viral infection points to blocking the protein targets ACE2 and TMPRSS2, or to raising the endosomal pH in order to prevent the S processing.[109] To retrieve information on drugs in clinical trial for this purpose, we can highlight the target nodes Angiotensin-converting enzyme 2 and Transmembrane protease serine 2 in the DT network of COVIDrugNet and check the "Inspected targets" table below for drugs reported to bind those targets. Here, we can find among others Chloroquine (CQ), Hydroxychloroquine (HCQ), and Bromhexine that are reported as ACE2 binders, and Camostat and Bromhexine reported as TMPRSS2 inhibitors. Notably, it is known that CQ and HCQ are also able to raise the endo-lysosomal pH thus inhibiting the protease activities and preventing the cleavage of S protein.[135] In addition, recent evidence suggests the combined use of Camostat and CQ (together with another drug, arbidol, an inhibitor of the virus-host cell membrane fusion with no known targets) to contrast the entry routes of SARS-CoV-2.[136] Finally, in the DP graph, one can select all the mentioned drugs and check the status of the clinical trials in which they are involved in the "Node Info" box on the right.

Targets The TP network of Figure 6.2c is a targetome that shows the relationships among the known targets of the proposed COVID-19 drugs. Here, two nodes (proteins) are linked if they are reported as targets of at least one of the drugs in the DrugBank COVID-19 database, and in this sense it is different from a typical interactome based on PPIs. The network is made by 1176 nodes and 70873 edges and shows a main connected component comprising 1037 nodes (88.2%). Human targets are 1008 (909 in the main connected component).

Looking at this graph provides another point of view on the pharmacological approaches taken to contrast the COVID-19. The network of the targets involved in the action of the drugs in clinical trials helps one to obtain a comprehensive view of the biological processes affected by the action of drugs. Actually, from the analysis of the

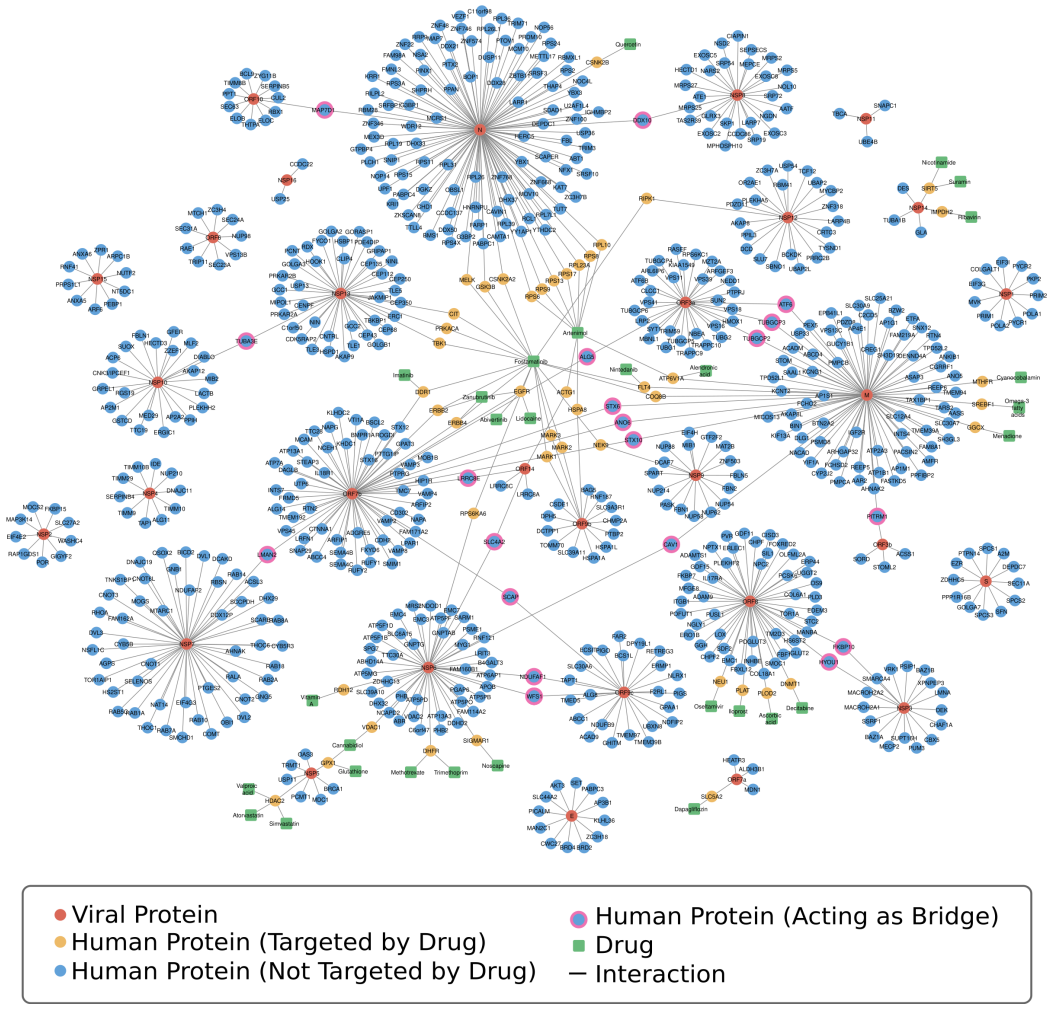


Figure 6.4. Virus–Host–Drug Interactome. The Virus–Host–Drug Interactome built on the basis of the merged datasets from Gordon et al.[124] and Chen et al.[137] Proteins (circles) are displayed in red if viral and in blue if human. The human proteins present in the TP network are shown as yellow circles, and the corresponding drugs currently in clinical trials against COVID-19 as green squares. The human proteins binding more than one viral target are highlighted as blue circles with pink contour. The network visualization was generated through Cytoscape 3.8.2.[138]

target proteins and their interactions it could be possible to trace the cellular pathways influenced by drugs. A study in this regard is currently underway. Instead, starting from the TP network, we carried out a different analysis that took into consideration both the data here presented on repurposed drugs now in clinical trials (a top-down view), and the molecular data on SARS-CoV-2 infection obtained from recent experimental studies and exploited to propose drugs to be repurposed (a bottom-up view). As regards the latter, we refer to the human-virus interactomes developed by Gordon

et al.[124] and more recently by Chen et al.[137] These interactomes are PPI networks that show which human proteins are bound directly by SARS-CoV-2 proteins to allow the virus to enter into the human cells, replicate, assemble and be released. Both research groups followed an experimental approach to identify the human proteins, using affinity purification (AP), and AP together with proximity labeling-based techniques, respectively, coupled with mass spectrometry. Merging the Gordon and Chen results, we obtained an extended list of 732 human proteins experimentally identified as interactors of the 29 viral proteins. Comparing this list with that of the human drug targets of the TP network (1008), we found that only 45 out of the 732 human proteins able to bind the viral ones are present in the TP as reported targets of drugs in clinical trials. In Figure 6.4, we show the integrated host-virus interactome (also available in the *Advanced Tools* block of COVIDrugNet), where the 45 proteins common to both lists are highlighted (yellow circles). We also checked the DT network of Figure 6.2a for drugs associated with these 45 targets and found 29 substances acting on them (Table 6.1) shown in the interactome of Figure 6.4 (green squares) linked to their targets. This is an example of how the information provided by the COVIDrugNet DT interactome can complement the one contained in human-virus PPI networks like those of Gordon and Chen. Note that the 29 substances hit direct neighbors of the viral proteins, thus interfering with the related viral processes. We see from Figure 6.4 that Artenimol and Fostamatinib, seemingly by virtue of their high target promiscuity, are able to hit simultaneously several targets, thus affecting various viral processes and allowing to foresee a better therapeutic efficacy. If confirmed by clinical results, these would be clear cases of poly-pharmacological multi-target actions exerted by single substances, a nice fit into the paradigm of network pharmacology.

Another interesting aspect emerging from inspection of the interactome of Figure 6.4 is that 20 human proteins (blue circles with pink contour) bind to two viral targets, thus acting as bridges between two node communities and playing a key role in the formation of the large connected component of the graph (Table 6.2). From a drug discovery perspective, such proteins would be ideal targets to fight the virus, as neutralizing them would help to disrupt the network of PPIs necessary to carry on the viral infection and replication processes. Unfortunately, none of these proteins appear in the TP network, implying that there is no substance targeting them among those listed in the DrugBank database of repurposed drugs presently in clinical trial. However, we

browsed some databases (DrugBank[114], DrugCentral[139] and ChEMBL[112]) in the search for bioactive substances reported to bind these 20 proteins and found that 4 of them are reported as targets of known drugs (Table 6.3). As can be seen from Table 6.3, many of the drugs listed therein have not yet been considered for therapy, while some of them are already in clinical trial for COVID-19 treatment even though their action on the proteins in the interactome is not reported in DrugBank. The former ones could be further possible candidates for COVID-19 drug repurposing in light of their ability to interfere with more than one process critical for the virus.

Gene Name	**Protein Name**	**Associated Drugs**
GSK3B	Glycogen synthase kinase-3 beta	Fostamatinib
PRKACA	cAMP-dependent protein kinase catalytic subunit alpha	Fostamatinib
DHFR	Dihydrofolate reductase	Methotrexate, Trimethoprim
ACTG1	Actin, cytoplasmic 2	Artenimol
DDR1	Epithelial discoidin domain-containing receptor 1	Imatinib, Fostamatinib
RIPK1	Receptor-interacting serine/threonine-protein kinase 1	Fostamatinib
RDH12	Retinol dehydrogenase 12	Vitamin A
COQ8B	Atypical kinase COQ8B, mitochondrial	Fostamatinib
IMPDH2	Inosine-5'-monophosphate dehydrogenase 2	Ribavirin
ERBB4	Receptor tyrosine-protein kinase erbB-4	Zanubrutinib, Fostamatinib
NEK9	Serine/threonine-protein kinase Nek9	Fostamatinib
CIT	Citron Rho-interacting kinase	Fostamatinib
HSPA8	Heat shock cognate 71 kDa protein	Artenimol
TBK1	Serine/threonine-protein kinase TBK1	Fostamatinib
HDAC2	Histone deacetylase 2	Valproic acid, Simvastatin, Atorvastatin

RPS9	40S ribosomal protein S9	Artenimol
MARK2	Serine/threonine-protein kinase MARK2	Fostamatinib
DNMT1	DNA (cytosine-5)-methyltransferase 1	Decitabine
GGCX	Vitamin K-dependent gamma-carboxylase	Menadione
SIRT5	NAD-dependent protein deacylase sirtuin-5, mitochondrial	Nicotinamide, Suramin
RPS8	40S ribosomal protein S8	Artenimol
EGFR	Epidermal growth factor receptor	Fostamatinib, Lidocaine, Zanubrutinib, Abivertinib
RPS13	40S ribosomal protein S13	Artenimol
SREBF1	Sterol regulatory element-binding protein 1	Omega-3 fatty acids
RPS6	40S ribosomal protein S6	Artenimol
MTHFR	Methylenetetrahydrofolate reductase	Cyanocobalamin
MARK3	MAP/microtubule affinity-regulating kinase 3	Fostamatinib
PLOD2	Procollagen-lysine,2-oxoglutarate 5-dioxygenase 2	Ascorbic acid
VDAC1	Voltage-dependent anion-selective channel protein 1	Cannabidiol
RPS6KA6	Ribosomal protein S6 kinase alpha-6	Fostamatinib
RPS17	40S ribosomal protein S17	Artenimol
FLT4	Vascular endothelial growth factor receptor 3	Nintedanib, Fostamatinib
PLAT	Tissue-type plasminogen activator	Iloprost
SIGMAR1	Sigma non-opioid intracellular receptor 1	Noscapine
GPX1	Glutathione peroxidase 1	Cannabidiol, Glutathione
SLC5A2	Sodium/glucose cotransporter 2	Dapagliflozin
CSNK2A2	Casein kinase II subunit alpha'	Fostamatinib
ATP6V1A	V-type proton ATPase catalytic subunit A	Alendronic acid, Artenimol

RPL23A	60S ribosomal protein L23a	Artenimol
CSNK2B	Casein kinase II subunit beta	Quercetin
RPL10	60S ribosomal protein L10	Artenimol
NEU1	Sialidase-1	Oseltamivir
MARK1	Serine/threonine-protein kinase MARK1	Fostamatinib
MELK	Maternal embryonic leucine zipper kinase	Fostamatinib
ERBB2	Receptor tyrosine-protein kinase erbB-2	Zanubrutinib, Fostamatinib

Table 6.1. Protein-Drug Associations for Common Targets between the Virus-Host Interactome and the Drug-Target Network.

6.1.3 Limitations

Our study is not exempt from some drawbacks that are common in data analysis, and regard mainly the data availability and quality. We based COVIDrugNet on the DrugBank Dashboard dedicated to COVID-19 pandemic, and although this public and free resource is known for the high reliability of the datasets, missing data or delayed updating can occur. This is evident for some drugs under clinical trial shown in Table 6.3 that have known targets yet not reported in their DrugBank file. Moreover, not all the drugs or proteins investigated here are completely characterized and classified, and this adds some uncertainty and noise to our results. Also, some bias could be incorporated in the knowledge we started from. For instance, the number of targets associated to a specific drug could considerably depend on the amount of research carried out on that medicine rather than on the actual biological interactions it has. This issue could be partially mitigated by a more extensive integration of data from a wider variety of databases. Very similar considerations can be drawn on the other databases that we exploited to retrieve auxiliary data: STRING[140], DisGeNet[141], SWISS-MODEL[142], RCSB-PDB[143], UniProt[144] and ChEMBL[112]. Furthermore, as mentioned in Section 6.1.5.1, the drug-target network was built considering only protein targets, hence nucleic acid targets were not included. However, biomolecular targets other than proteins are a minority[145] and this led us to not integrate them. Despite the massive efforts of the scientific community, SARS-CoV-2 and COVID-19

Viral Proteins	Human Proteins	Name
NSP13, NSP10	TUBA3E	Tubulin alpha-3E chain
ORF9C, NSP6	NDUFAF1	Complex I intermediate-associated protein 30, mitochondrial
NSP3, ORF8	FKBP10	Peptidyl-prolyl cis-trans isomerase FKBP10
M, ORF3a	ATF6	Cyclic AMP-dependent transcription factor ATF-6 alpha
M, ORF7b	STX10	Syntaxin-10
ORF7b, ORF14	LRRC8E	Volume-regulated anion channel subunit LRRC8E
M, ORF3a	TUBGCP3	Gamma-tubulin complex component 3
NSP6, ORF14	SLC4A2	Anion exchange protein 2
ORF8, NSP3	HYOU1	Hypoxia up-regulated protein 1
M, ORF7b	STX6	Syntaxin-6
M, ORF3a	TUBGCP2	Gamma-tubulin complex component 2
ORF10, N	MAP7D1	MAP7 domain-containing protein 1
N, NSP8	DDX10	Probable ATP-dependent RNA helicase DDX10
ORF9c, NSP6	WFS1	Wolframin
M, ORF3b	PITRM1	Presequence protease, mitochondrial
ORF7b, M	ANO6	Anoctamin-6
ORF7b, NSP7	LMAN2	Vesicular integral-membrane protein VIP36
M, NSP6	CAV1	Caveolin-1
ORF9c, ORF7b	SCAP	Sterol regulatory element-binding protein cleavage-activating protein
ORF3a, ORF7b	ALG5	Dolichyl-phosphate beta-glucosyltransferase

Table 6.2. Human Proteins that Interact with More than One Viral Protein in the Virus-Host Interactome.

continue to be largely puzzling. Experimental assays are the solid ground on which we all start to build our hypotheses, yet also these investigations may have bias and a moderate amount of uncertainty. We have to keep this into consideration, when examining the merged interactomes by Gordon et al.[124] and Chen et al.[137] given their

Human Proteins	Name	Known Drugs
TUBA3E	Tubulin alpha-3E chain	Podophyllotoxin[a], CYT997[b], Docetaxel[c], Vincristine[c], Verubulin[c], Indibulin[c], Trastuzumab–Emtansine[c], Ixabepilone[c], Sagopilone[c], Eribulin[c], Fosbretabulin[c], Mirvetux-imab–Soravtansine[c], Paclitaxel[c], Plinabulin[c], Polatuzumab–Vedotin[c], Vinblastine[c], Crolibulin[c], Fosbretabulin[c], Cabazitaxel[c], Davunetide[c], Paclitaxel–Poliglumex[c], Vinflunine[c], Lexibulin[c], **Colchicine**[dc], Vinorelbine[c]
NDUFAF1	Complex I intermediate-associated protein 30, mitochondrial	**Metformin**[de], NV-128[f], ME-344[f]
FKBP10	Peptidyl-prolyl cis-trans isomerase FKBP10	**Tacrolimus**[dg]
ATF6	Cyclic AMP-dependent transcription factor ATF-6 alpha	Pseudoephedrine[h]

Table 6.3. Known Drugs Targeting Human Proteins that Interact with more than One Viral Protein in the Virus-Host Interactome.

[a] Retrieved from DrugCentral (https://drugcentral.org/target/Q6PEY2/)
[b] Retrieved from DrugBank (https://go.drugbank.com/drugs/DB05147)
[c] Retrieved from ChEMBL (https://www.ebi.ac.uk/chembl/g/#browse/mechanisms_of_action/filter/target.target_chembl_id:CHEMBL2095182)
[d] Drugs currently in clinical trial for COVID-19
[e] Retrieved from DrugCentral (https://drugcentral.org/target/Q9Y375/)
[f] Retrieved from ChEMBL (https://www.ebi.ac.uk/chembl/g/#browse/mechanisms_of_action/filter/target.target_chembl_id\protect\protect\leavevmode@ifvmode\kern+.2222em\relaxCHEMBL2363065)
[g] Retrieved from DrugCentral (https://drugcentral.org/target/Q96AY3/)
[h] Retrieved from DrugBank (https://go.drugbank.com/drugs/DB00852)

considerable difference. Additionally, the identification of a PPI in vitro unfortunately does not guarantee that the same interaction occurs also in vivo.

6.1.4 Conclusions

The COVID-19 pandemic poses a huge problem of public health that requires the implementation of all available approaches to contrast it, and drugs are one of them. In this context, we observed an unmet need of depicting the continuous evolving scenario of the ongoing drug clinical trials through an easy-to-use freely accessible online tool. Starting from this consideration, we developed COVIDrugNet, a web app that allows one to watch and keep up to date on how the drug research is responding with its arsenal of known repurposed drugs to the health threat represented by the SARS-CoV-2 infection. We have shown some examples of how one could explore the whole landscape of medicines currently in clinical trial and try to probe the consistency of actual treatments with the biological evidence being accumulated on the virus infection and its systemic pathological consequences in humans. The complex network of protein targets affected by the repurposed drugs can be confronted with the host-virus interactome, and this may offer new hints on drugs currently in use or to be proposed for clinical investigation. From this comparison, we have been able to single out some human proteins that contact two viral counterparts, and that might be possible new targets for anti-COVID-19 drugs. Finally, given that, as already noticed by others[110], several treatments proposed for COVID-19 are still lacking a known mechanism of viral inhibition or even a pharmacological rationale, careful analyses of the drug-target data as those reported in the present work might help to understand the molecular implications of these pharmacological options, and eventually improve the search for more effective therapies.

6.1.5 Methods

6.1.5.1 Data Acquisition

The set of drugs in clinical trial for the treatment of COVID-19 (731 on August 11, 2021) was retrieved from the dedicated web page of DrugBank (https://www.drugbank.ca/covid-19). Both experimental unapproved substances, and drugs in clinical trials were considered, and duplicates were removed (more than one trial is

going on for some drugs). The set was also filtered for both the number of heavy atoms (to exclude inorganic compounds), and the availability of data (a drug was not added if it was not present in the PubChem database). This cleaning step reduced the number of drugs considered to 397. From the same site and from PubChem, we gathered some features related to structure, as well as pharmacological classification, pharmacodynamics, and pharmacokinetics of each drug (Table 6.4). Drugs for which no targets were reported in DrugBank were discarded (290 remaining). As regards the drug targets, they were retrieved from DrugBank, and in this case we collected some information on classification, biology, and pharmacology of each protein. A detailed description of the features and the data sources are reported in Table 6.5.

6.1.5.2 Networks Construction

We chose to inspect the data in the form of a graph. All networks presented in the web app and in the paper were built by means of the NetworkX software.[127]

6.1.5.3 Network Analysis

Node Attributes Some suitable node attributes (Degree, Closeness Centrality, Betweenness Centrality, Eigenvector Centrality, Clustering Coefficient, VoteRank) were calculated through NetworkX. The only property we tweaked was the result of the VoteRank because its algorithm draws up a ranking of nodes based on an iterative voting system[146] without assigning a specific value to each one of them. Thus, we translated this ranking into a score for each node on the basis of its position and the total number of nodes with the following method:

```
IF node in rank
    score = length(rank) - index_in_rank(node)
ELSE
    score = 0
```

Nodes Grouping Dividing a network into groups, clusters, or communities could be useful to unveil not trivial patterns of interaction. It is accomplished by splitting the network into subgroups that have the fewest possible number of connections between them[10]. In this work, we took advantage of (and provide access to in the webtool)

Feature	Description	Source
ID	DrugBank unique identification code	DrugBank[114]
SMILES	The chemical structure string notation for drugs. SMILES were recovered from PubChem if available, otherwise from DrugBank	PubChem[113] DrugBank[114]
ATC code level 1	The broad-based level of the ATC classification system identifying the fourteen anatomical/pharmacological groups	DrugBank[114]
ATC identifier	ATC code	DrugBank[114]
Targets	Entities to which the drug binds or interacts with, resulting in an alteration of their normal function and thus in desirable therapeutic effects or unwanted adverse effects	DrugBank[114]
Enzymes	Proteins that facilitate a metabolic reaction that transforms the drug into one or more metabolites	DrugBank[114]
Carriers	Proteins that bind to the drug and modify its pharmacokinetics, e.g., facilitating its transport in the blood stream or across cell membranes	DrugBank[114]
Transporters	Proteins that move the drug across the cell membrane	DrugBank[114]
Drug Interactions	Drugs that are known to interact, interfere or cause adverse reactions when taken with this drug	DrugBank[114]
Trials	Identifiers of clinical trials with the respective phase	DrugBank[114]

Table 6.4. Drugs Features.

three of the most common algorithms for this purpose: spectral clustering[27], Girvan-Newman community detection[30], and greedy modularity community detection[24]. The first one makes use of the spectrum of the graph Laplacian to convey the information about the graph partition.[27] The division is then carried out on this data by a k-means clustering algorithm (see the Supplementary Information for more detail). In the second case, communities are recognized employing the Girvan-Newman method.[30] It is a hierarchical method based on the progressive removal of the edges

Feature	Description	Source
Gene	Short identifier of the unique gene name	DrugBank[114]
Organism	Organism where the protein comes from	DrugBank[114]
Cellular Location	The protein cellular location	DrugBank[114]
Drugs	List of known drugs related with the protein (e.g., agonists, antagonists, inhibitors...)	DrugBank[114]
ID	UniProt unique identification code	DrugBank11
STRING Interaction Partners	Known and predicted protein-protein interactions (both physical and functional) only in Homo Sapiens and with a minimum score of 0.95	STRING[140]
Diseases	Disease groups with an Evidence Index of 1 (see https://www.disgenet.org/dbinfo#section44 for more information)	DisGeNET[141]
PDBID	Protein Data Bank identification code (the structure with the best resolution)	SWISS-MODEL[142]
Protein Classification	The first and the second level of Protein Target Classification are named Protein Class and Protein Family respectively	ChEMBL[112]

Table 6.5. Targets Features.

with the highest betweenness centrality from the graph, causing it to break into sets of smaller constituents. The partition with the best modularity is shown, but the user can manually choose an arbitrary number of communities in the web tool. The greedy modularity community detection method[24] pursues the graph division through a bottom-up approach (opposite to the previous one), by exploiting a "greedy" algorithm that progressively associates the nodes into groups that maximize the modularity. It starts with all nodes separated into single communities and recursively merges the couple of them that brings to the highest modularity increasing, until the point that joining two communities would lead to a modularity reduction.
These tasks were accomplished through in-house Python scripts, mainly making use of the packages NetworkX and Scikit-learn[147].

Degree Distribution Fitting A network is commonly considered to be scale-free if the degree distribution of its nodes follows a power-law[38], which has the form:

$$p(x) \propto x^{-\alpha} \tag{6.1}$$

where the scaling exponent α is higher than 1 (usually between 2 and 3) and the degree value x is equal or greater than x_{min} (which is always higher than 1). To the best of our knowledge, the most severe scale-freeness test is presented by Broido et al.[128] that take advantage of a rigorous mathematical procedure[39] to assess the validity of a power-law distribution to describe the investigated degrees. Here, we followed their approach probing the fitting of a power-law to the degree distributions of both projected networks DP and TP (with and without the Artenimol and Fostamatinib nodes). As a first step, the parameters of the best fitting power-law are determined (x_{min} with a standard Kolmogorov-Smirnov minimization approach, and then α with a discrete maximum likelihood estimation) employing the Python package Powerlaw[148]. Then, the fitting is evaluated considering the p-value of the Kolmogorov-Smirnov distance (computed with a semi-parametric bootstrap), and of the x_{min} and α (bootstrap). If $p \geq 0.1$, the degree distribution is considered plausibly scale-free. Lastly, the chosen power-law distribution is compared to four non-scale-free alternatives (using loglikelihood ratio tests), to evaluate if it is favored over the others. Such alternatives are the exponentially truncated power-law, the exponential, the stretched exponential (Weibull) and the log-normal. This entire procedure was carried out using an in-house Python script, with a large employment of the Python package Powerlaw. A more thorough explanation and method validation are provided in the Supplementary Information.

Robustness Scale-free networks (contrary to random Erdős-Rényi graphs) have an exceptional tolerance against random failures, but at the same time they are very vulnerable to targeted attacks.[129] We investigated the robustness of these networks evaluating their diameter (as a measure of interconnectivity) throughout a process of node removal. We took into account both targeted attacks and random failures and compared the results. In the first case, at every iteration the node with the highest degree was chosen and removed. In the other case, a node was selected randomly and eliminated. In this latter condition, the average of multiple 100 runs was considered in order to avoid misinterpretations induced by a single random choice. This procedure was carried out through an in-house Python script.

6.1.5.4 COVIDrugNet Implementation and Deployment

COVIDrugNet is mainly composed by the collector and the web tool itself. Both are written in Python, but the purpose of the former is to collect the data from web databases, build the graphs, compute some properties, and store everything in pickle format. The latter, instead, retrieves the data from the created database and sets up the front-end part of COVIDrugNet with Python Dash.[149] The web tool deployment was carried out with Apache[150] through the `mod_wsgi` interface in an Ubuntu server.

6.2 DEGA

This project originated as a response to the need for a Python-based tool to facilitate differential gene expression analysis. It later expanded its scope to incorporate a network-based methodology, which can give useful insights into the conditions in study.
DEGA, the resulting Python package, has undergone validation by closely reproducing the results of DESeq2, a well-established tool in the field. This validation was conducted using the same dataset DESeq2 was originally tested on.
Beyond its role in differential expression analysis, DEGA integrates a network-based approach to identify a subset of pivotal regulatory genes referred to as "switch genes". These switch genes exhibit significant associations with pronounced changes in various biological contexts. A case study involving glioblastoma demonstrates DEGA's capacity to unveil key regulatory genes with potential implications for understanding and treating this complex disease.

Details

Authors Luca Menestrina, Maurizio Recanatini

Type Application Note

Status In preparation

Data Availability The package is available on the Python Packaging Index (https://pypi.org/project/DEGA) and on GitHub (https://github.com/LucaMenestrina/DEGA). All the datasets and scripts for reproducing the experiment, as well as the obtained tables and figures, are provided in the `validation` folder on the GitHub repository.

DEGA: A Python Package for Differential Gene Expression Analysis with Enhanced Functionality for Dataset Exploration and Results Interpretation

6.2.1 Introduction

Differential gene expression analysis is a tool for understanding the underlying molecular changes that occur in response to various stimuli, such as disease, environmental factors, or treatments. This method is used to identify genes that have an expression significantly different between two or more biological conditions.
This task is usually carried out employing R packages[151], the most common ones being edgeR[152], limma[153] and DESeq2[154].
It is possible to call R packages from Python using rpy2 (https://rpy2.github.io/) however, this would force the user to have both languages (Python and R) installed on the machine, and to be familiar with both.
Here we introduce DEGA, a Python implementation of the widely used R package DESeq2.
The opportunity to perform a differential expression analysis directly within Python translates mainly to a more straightforward approach for Python users to such a relevant analysis. Furthermore, providing access to a wider community, it could lead to improvements both in the method and in its implementation, or even to the development of new approaches.

To illustrate the potentiality of DEGA, we present a compelling case study in the context of glioblastoma.
Glioblastoma (GBM), the most frequently diagnosed malignant brain tumor in adults, accounts for approximately 14.5% of central nervous system (CNS) cancers.[155] Patients with GBM often experience a spectrum of symptoms, such as headaches, focal neurologic deficits, confusion, memory loss, personality changes, or seizures.[156]
While there have been recent advances in understanding the genetic, epigenetic, and molecular subtypes of gliomas, the current standard-of-care treatment for GBM still revolves around surgical resection, followed by a course of radiotherapy with concurrent and adjuvant chemotherapy.[157] Nevertheless, despite these significant treatment efforts, glioblastoma remains an incurable disease, with patients facing a grim prog-

nosis of 12 to 15 months of survival (median 14.6 months).[156]
Recent research has delved into the initial molecular pathogenesis of these tumors, focusing on alterations in cellular signal transduction pathways and the development of resistance to therapy.[156]

Utilizing DEGA's enhanced functionalities, including switch gene identification[158], we aim to pinpoint pivotal genes that hold the potential to provide valuable insights into the pathogenesis of glioblastoma and serve as possible therapeutic targets. These switch genes appear to act as crucial regulators, driving significant transcriptomic transitions, and are identified through the analysis of a co-expression network. Characterized by interactions extending beyond their own community, minimal local hub connectivity, and predominantly negative correlations with their interaction partners, switch genes represent a unique group within the co-expression network landscape.

6.2.2 Methods

6.2.2.1 Differential Expression Analysis

DEGA (Differentially Expressed Genes Analysis) is a Python package, which implements the algorithm presented in DESeq2 for identifying genes having an expression significantly different between two biological conditions.
Here is a brief description of the key steps performed by DEGA. The input is a matrix of raw counts, which reflect gene abundance but are also influenced by the sequencing analysis. DEGA deals with this point normalizing the counts, dividing them by the corresponding size factor. Each sample-specific size factor is computed as the median of ratios of sample's counts (k_{ij}) to a pseudo-reference (obtained by taking the geometric mean across samples):

$$SizeFactor_j = \underset{i}{\text{median}} \frac{k_{ij}}{\left(\prod_{v=1}^{m} k_{iv} \right)^{1/m}} \tag{6.2}$$

The counts distributions are modeled by a negative binomial distribution, which is defined by two parameters: mean and dispersion (a measure of the variability of the data, defining the relationship between the variance and the mean of the counts).

$$k_{ij} \sim NB(mean = \mu_{ij}, dispersion = \alpha_i) \tag{6.3}$$

Where k_{ij} is the count for gene i and sample j, μ_{ij} is the mean, and α_i is the gene-specific dispersion parameter.
The mean is easily calculated from the observed normalized counts, but the dispersion is trickier given the usually very low number of samples. This issue is overcome assuming that the dispersions of genes having similar expression levels are similar in turn. Thus, the information coming from similarly expressed genes is used for estimating the dispersions. The procedure follows three steps. First of all, the dispersion is estimated for each gene separately, then a curve is fitted to these gene-wise dispersion estimates, and lastly, the dispersions that are not evaluated as outliers are shrunken toward the values predicted by the curve.
The subsequent step is to fit a generalized linear model in order to determine the log2 fold change and the p-value.

$$\log_2(q_{ij}) = \sum_r \Big(x_{jr} \cdot \beta_{ir} \Big) \tag{6.4}$$

Where x_{jr} is the design factor, β_{ir} is the log2 fold change between the conditions, and q_{ij} is a parameter proportional to the expected true concentration of fragments of sample j:

$$q_{ij} = \frac{\mu_{ij}}{SizeFactor_j} \tag{6.5}$$

The p-values are corrected for multiple tests (using the procedure of Benjamini-Hochberg[159], avoiding loss of power employing an independent filtering on the mean of normalized counts).
In the end, outliers are detected on the basis of Cook's distance[160] and then automatically handled.

6.2.2.2 Switch Genes Identification

Furthermore, for applications other than differential testing (e.g., clustering and machine learning analyses), DEGA offers some methods for counts transformations (variance stabilizing, regularized log, and shifted log). Some convenient data explorations (e.g., correlations, PCA, etc.) are already implemented in DEGA along with

some other useful data visualizations.
One of DEGA's valuable tools for dataset exploration and result interpretation is the identification of switch genes.
To demonstrate its capabilities, we utilized RNA-Seq data from the glioblastoma (GBM) dataset[161] of the TCGA (The Cancer Genome Atlas) repository[162].
We identified the significant differentially expressed genes using DEGA's main algorithm (considering only genes with log2 fold change higher than 1, and adjusted p-value less than 0.01, see Section 6.2.2.1). Subsequently, for the identification of switch genes, DEGA offers a modification of the co-expression network approach previously outlined by Paci et al. [158, 163]. The co-expression networks were constructed based on the Pearson correlation coefficient[164] between the expression levels of each pair of genes. Nodes in the network represent RNA transcripts, and connections between nodes indicate significant correlation or anti-correlation of gene expression. The significance of the edges was granted through a two-step filtering process applied to the correlation matrix. Initially, p-values for two-tailed t-tests were computed for all correlations and corrected for multiple testing using the Benjamini-Hochberg method[159] (DEGA also offers alternative adjustment methods). A significance threshold of 0.01 was used. Subsequently, the remaining values underwent a secondary filtering step, where a correlation threshold was chosen to retain at least 95% of the nodes within the largest component. (users can customize these thresholds within DEGA).
In contrast to the original k-means clustering method, DEGA utilized a greedy-modularity clustering approach[24, 31] to identify communities in the network associated with switch genes.
A heatmap of the nodes was generated according to their topological properties. The extracted switch genes were identified as markers of the shift from healthy to diseased patients. The plane was defined by two parameters: normalized within-module degree (a measure of the interconnectedness between a node and other nodes within its module):

$$z_g^i = \frac{k_i^{in} - \bar{k}_{C_i}}{\sigma_{C_i}} \tag{6.6}$$

(where k_i^{in} is the number of links of node i to nodes in its module C_i, $\bar{k}_{C_i}$ and σ_{C_i} are the average and standard deviation of the total degree distribution of the nodes in the module C_i), and clusterphobic coefficient (a measure of the "fear" of a node of being

confined within a cluster, drawing an analogy to claustrophobia):

$$K_{\pi}^{i} = 1 - \left(\frac{k_i^{in}}{k_i} \right)^2 \tag{6.7}$$

(where k_i^{in} and k_i are the internal and the total degree of node i, respectively)
It was divided into seven regions (R1-R7) defining specific node roles[165]:

Nonlocal hub ($z_g < 2.5$)

- R1 Ultra-periferal nodes ($K_{\pi} = 0$)
- R2 Peripheral nodes ($K_{\pi} \leq 0.625$)
- R3 Nonhub connectors ($0.62 < K_{\pi} \leq 0.8$)
- R4 Nonhub kinless nodes ($K_{\pi} > 0.8$)
  In this region, those having APCC < 0 are deemed switch genes

Local hub ($z_g \geq 2.5$)

- R5 Provincial hubs ($K_{\pi} = 0.3$)
- R6 Connector hubs ($K_{\pi} \leq 0.75$)
- R7 Kinless hubs ($K_{\pi} < 0.75$)

Nodes were colored according to their average Pearson correlation coefficient (APCC) value. Red nodes represented nodes positively correlated in expression with their interaction partners, while blue nodes exhibit an average negative correlation in expression with their interaction partners. Blue nodes (APCC < 0, being on average anti-correlated with their neighbors) falling in region R4 represented the switch genes characterized as not being a hub in their own cluster ($z_g < 2.5$) and having many links outside their own cluster ($K_{\pi} > 0.8$), indicating their predominant connections outside their module.

6.2.3 Results and Discussion

We evaluated DEGA differential expression analysis results comparing them with those obtained by DESeq2 on the Bottomly et al. dataset[166], the same DESeq2 was tested on when it was proposed[154]. This dataset contains RNA-seq data about two different strains of mice.

This validation is illustrated in a Jupyter notebook which is available at https://github.com/LucaMenestrina/DEGA/blob/main/validation/DEGA.ipynb.
Given the aim of identifying the significantly up- and down-regulated genes, the two key values obtained by the differential expression analysis are the log2 fold changes (positive if the gene is upregulated, negative otherwise) and the adjusted p-values (for statistical soundness). As it is shown in Figure 6.5a, DEGA reproduces almost perfectly the log2 fold changes of DESeq2 (Adjusted p-values are compared in Figure 6.5b).

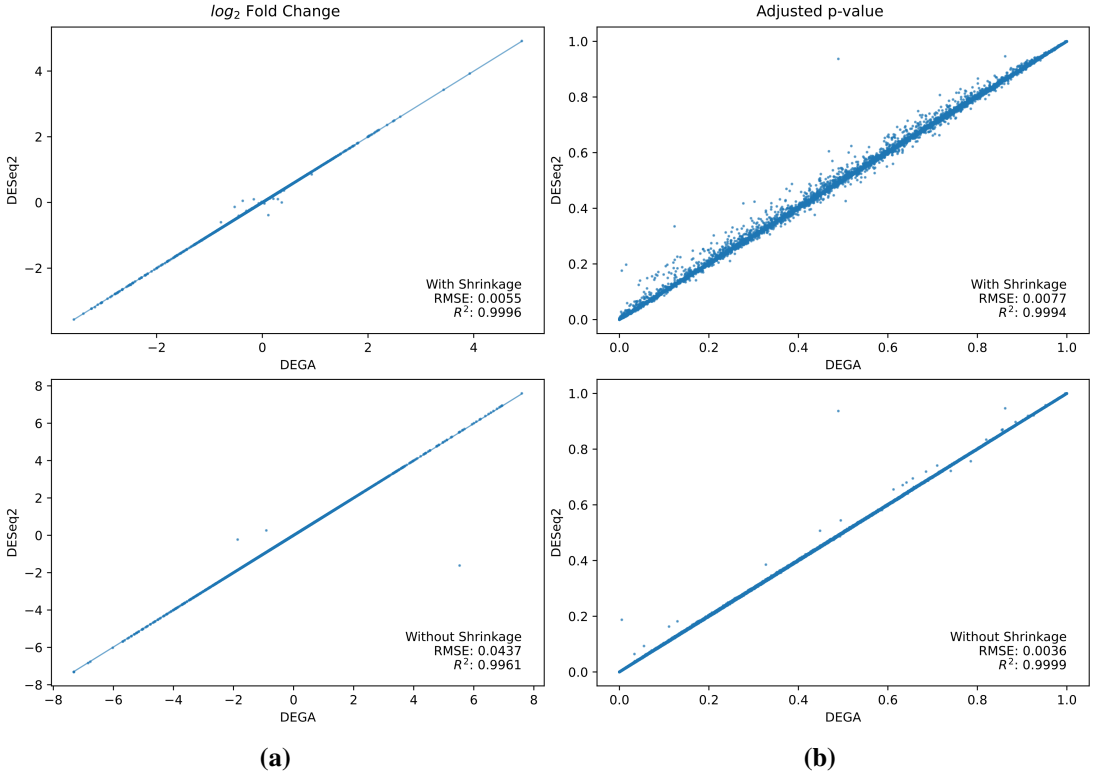


Figure 6.5. DEGA Results Compared to DESeq2 Ones. Comparison of log2 fold changes **(a)** and adjusted p-values **(b)** (with and without shrinkage, top and bottom, respectively) obtained by DESeq2 (vertical axis) and DEGA (horizontal axis).

After validating DEGA's differential expression analysis using a benchmark dataset, we proceeded to evaluate its performance in a real-life application, analyzing the glioblastoma dataset sourced from TCGA. We first identified differentially expressed genes (1066 were upregulated and 734 downregulated) and subsequently

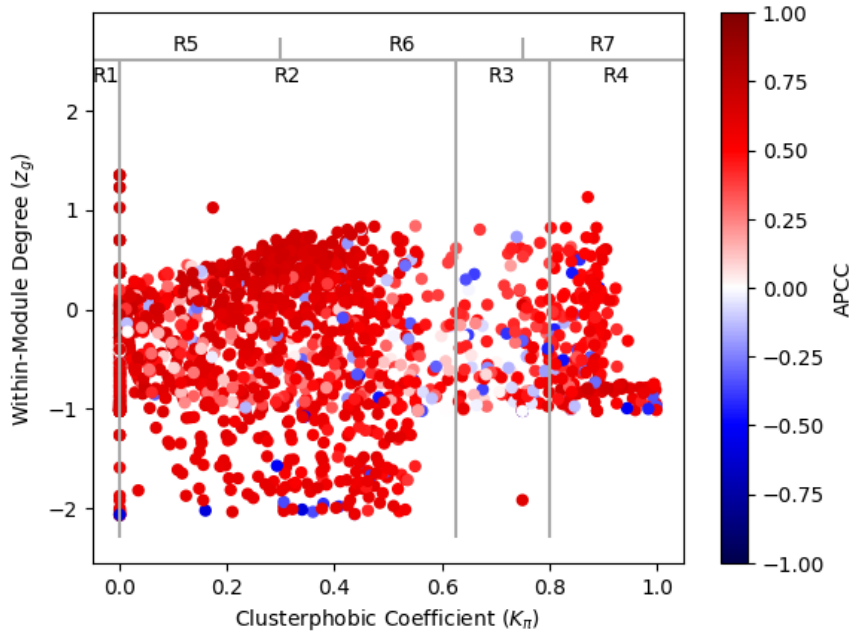


Figure 6.6. Heatmap of Nodes in GBM Network. A heatmap representing nodes within the glioblastoma correlation network. The dots in the plot, which represent nodes in the correlation network, are categorized into seven regions (labeled as R1 to R7) based on their clusterphobic coefficient (K_π) on the x-axis. The clusterphobic coefficient measures how likely a node is to interact with nodes in different modules. The y-axis represents their within-module degree of communication (z_g). The color of each dot corresponds to the APCC (Average Pearson Correlation Coefficien) value, which is indicated on a color scale from red (positive APCC) to blue (negative APCC). Switch genes are blue dots in region R4.

employed DEGA's functionality to find switch genes.

DEGA identified a set of 24 switch genes (Figure 6.6, blue dots in region R4): PTBP1, VIM, HMG20B, GNAI3, CCDC80, RBBP8, SYDE1, TGIF2, TRIP10, BCL2L12, TNFRSF19, SMO, TRIM5, TGFB1I1, CMTM3, RAB13, PDIA4, WEE1, FAM111A, SHOX2, PLEKHF2, TGIF1, F2R, EPHB4.

While the majority of these genes are directly associated with GBM, a select few deserve individual attention for their significant roles in the context of GBM:

TBP1 knockdown promotes neural differentiation of glioblastoma cells through its interaction with the UNC5B receptor, leading to the suppression of cancer cell proliferation both in vitro and in vivo.[167]

BCL2L12, known for its multifunctional nature, plays a critical role in promoting intense therapeutic resistance in GBM. It operates on two key nodes of cytoplasmic

and nuclear signaling cascades. Additionally, it is a target of a drug that has advanced to phase 0 clinical trials, underscoring its therapeutic potential. [168, 169]

TNFRSF19 significantly contributes to GBM by promoting migration, invasion, and resistance. It exerts its effects through modulating various factors, including Pyk2-Rac1, JAK1-STAT3, and RKIP.[170–172]

PDIA4 plays a regulatory role in the progression and angiogenesis of GBM.[173, 174]

WEE1 is a major regulator of the G_2 checkpoint. Its inhibition abrogates the G_2 arrest phase, impeding DNA repair processes and leading to mitotic catastrophe and subsequent cell death.[175, 176]

The inclusion of BCL2L12 and WEE1 among the switch genes is particularly significant. These genes, which are subjects of ongoing phase 0 clinical trials, not only validate our results but also underscore the potential of this approach in revealing novel, disease-specific therapeutic targets.

The GitHub repository includes also a Jupyter notebook that illustrates this case study, which is accessible at https://github.com/LucaMenestrina/DEGA/blob/main/case_study/switch_GBM.ipynb.

6.2.4 Limitations

While DEGA presents a promising approach for differential gene expression analysis and network-based insights, it is important to recognize areas for potential growth. While DEGA's computational performance may currently be slower compared to established tools like DESeq2, its innovative features offer unique opportunities for analysis. Additionally, while DEGA does not encompass the entirety of DESeq2's functionalities, it provides a solid foundation upon which future enhancements can be built. Moving forward, there is exciting potential for DEGA to optimize its computational efficiency and expand its feature set, further solidifying its position as a valuable tool.

6.2.5 Conclusions

We implemented DEGA, a Python package that performs differential expression analyses. Our package demonstrated its validity by closely replicating results obtained

with the widely recognized DESeq2 tool when applied to the Bottomly et al. dataset. This confirms the reliability and effectiveness of DEGA in conducting differential expression analysis.

Our application of DEGA to the Glioblastoma (GBM) dataset from the TCGA repository has provided valuable insights into this complex disease. Leveragning DEGA's functionality, we have identified a selection of switch genes that play pivotal roles in the context of glioblastoma. These findings contribute to improving our understanding of the underlying molecular mechanisms and potential therapeutic targets for this deadly brain tumor.

DEGA is designed to be easily accessible by the vast Python user community, ensuring that researchers can effortlessly perform differential expression analyses. By extending its accessibility to a broader user base, we aim to foster a collaborative environment, enabling further refinements and enhancements in both the method and its implementation.

6.3 Drug-induced Impulsivity

This research developed from the pressing demand for more understanding in guiding pharmacological treatments, vigilant monitoring, and precise interventions to alleviate the distressing impact of drug-induced impulsivity. Characterized by a loss of behavioral control, drug-induced impulsivity encompasses behaviors as significant as pathological gambling, hypersexuality, compulsive shopping, and hyperphagia. Nevertheless, despite its profound effects, its impact is not comprehensively addressed by existing tools.

This extensive analysis resulted from a collaborative endeavor involving a multidisciplinary team of experts in psychiatry, pharmacology, pharmacovigilance, statistics, and network analysis. The primary objective is to address a substantial knowledge gap concerning adverse drug reactions linked to dopaminergic agents, particularly pramipexole and aripiprazole. These medications find widespread use in the treatment of Parkinson's Disease and mood and psychotic disorders, respectively.

Leveraging disproportionality and network analyses within the FDA Adverse Event Reporting System (FAERS), events associated with impulsivity induced by these two drugs were uncovered, along with potential exacerbating factors. These findings unravel the complex landscape of drug-induced impulsivity, revealing its profound impact on patients' lives, spanning psychosocial, psychosomatic, metabolic, and sexual dimensions.

Beyond its immediate findings, this study holds significant implications for the design of future clinical investigations into the impact of drug-induced impulsivity on patients' quality of life. The goal is to provide a clearer understanding of its multifaceted burden and contribute to effective measures for its mitigation.

Details

Authors Michele Fusaroli*, Stefano Polizzi*, Luca Menestrina*, Valentina Giunchi, Luca Pellegrini, Naomi Fineberg, Daniel Weintraub, Maurizio Recanatini, Gastone Castellani, Fabrizio De Ponti, Elisabetta Poluzzi

* The first three authors equally contributed.

Type Research Article

Status Submitted, available in preprint

Title Unveiling the Hidden Toll of Drug-Induced Impulsivity: A Network Analysis of the FDA Adverse Event Reporting System.

Journal medRxiv

DOI 10.1101/2023.11.17.23298635

Data Availability The data we used comes from the FDA Adverse Event Reporting System (FAERS), and is made publicly available by the FDA as quarterly data downloadable at https://fis.fda.gov/extensions/FPD-QDE-FAERS/FPD-QDE-FAERS.html. The algorithm for cleaning FAERS data is open-source at https://github.com/fusarolimichele/DiAna, and the cleaned database is available on an OSF repository (https://osf.io/k9v6s/) and through the R package DiAna.

Unveiling the Hidden Toll of Drug-Induced Impulse Control Disorders: A Network Analysis of the FDA Adverse Event Reporting System.

6.3.1 Abstract

Introduction: Adverse drug reactions significantly impact patients' lives, yet their influence is often underestimated in treatment decisions and monitoring. Impulsivity induced by dopaminergic agents can lead to impaired social functioning and quality of life.

Aim: This study assesses impulsivity burdens from pramipexole and aripiprazole, pinpointing key symptoms for targeted mitigation.

Method: Leveraging data from the FDA Adverse Event Reporting System (January 2004 - March 2022), we employed the Information Component to identify the syndrome of signs and symptoms disproportionately co-reported with drug-induced impulsivity. Using composite network analyses (PPMI, Ising, ϕ) we characterized clusters of co-reported events (i.e., subsyndromes). Finally, we assessed the secondary impact of drug-induced impulsivity modeling our dataset as a chain of directed connections (Bayesian network).

Results: The drug-induced impulsivity syndrome (respectively 56 and 107 events in pramipexole and aripiprazole recipients), primarily encompassed psychiatric, social, and metabolic events, segregated into subsyndromes such as delusional jealousy and dopamine dysregulation syndromes among pramipexole recipients, and obesity-hypoventilation syndrome and social issues among aripiprazole recipients. Anxiety and economic problems emerged as pivotal nodes in the exacerbation of the syndromes.

Conclusions: Drug-induced impulsivity places a substantial burden on patients and their families, with manifestations shaped by the underlying disease. Network approaches, exploring intricate symptom connections and identifying pivotal symptoms, complement traditional techniques and clinical judgment, providing a foundation for informed prescription and targeted interventions to alleviate the burden of adverse drug reactions.

6.3.2 Introduction

Adverse drug reactions (ADRs) significantly impact patients' well-being[177] by extending beyond organic diseases and trespassing into psychological illness[178] and social sickness[179]. For instance, immunodeficiency perturbs social activities, dysphonia hinders teaching roles, and sexual dysfunction intricately affects relationships and personal identity. Despite their profound effects, ADRs are often inadequately recognized, resulting in compromised patient-doctor relationships[180], prolonged hospitalization[181], and a pervasive decline in Quality of Life (QoL)[182]. This disregard extends to patient-reported outcomes, crucial for QoL assessment and patient-centered care[183], often relegated to the margins in prescribing information or package inserts[184]

Drug-induced impulsivity, classified as "impulse control disorders induced by other specified psychoactive substance (6C4E.73)" in the International Classification of Diseases (ICD-11) category of disorders due to substance use, represents a distressing group of conditions marked by a loss of behavioral control. This pathological disinhibition can yield behaviors as pervasive as pathologic gambling, hypersexuality, compulsive shopping, and hyperphagia, the so-called "four knights of Impulse Control Disorder"[185]. Additionally, behaviors like stealing, hair pulling, and compulsive hoarding[186, 187] contribute to the intricate tapestry of drug-induced impulsivity. The first reports of drug-induced impulsivity were linked to dopamine receptor agonists like pramipexole, ropinirole, rotigotine, licensed for treating Parkinson's disease (PD)[188] and restless legs syndrome (RLS)[189, 190]. More recently, the role of partial dopamine agonists like aripiprazole, brexpiprazole, cariprazine, licensed for treating psychosis and mood disorders, has also emerged[191].

Within the landscape of PD, drug-induced impulsivity unfurls a complex narrative. Initially, it may manifest as heightened motivation and hobbyism, known as "honeymoon period"[192]. However, even when concealed in subclinical forms[193], drug-induced impulsivity holds the potential to significantly erode patients' QoL[194]. This erosion, appraised through metrics like the PDQ-39 scale[195], encompasses diverse neuropsychiatric and somatic domains including mobility, daily activities, stigma, social support, communication[196], urinary and sexual function, sleep, attention, and cardiovascular symptoms[197]. The impact extends beyond patients to affect caregivers, who grapple with their own set of health issues, depression, and

social impediments[198].

Nevertheless, conventional evaluations frequently fall short in encompassing neuropsychiatric symptoms, altered behavior patterns, financial hardships, and legal entanglements[199], failing to capture the full spectrum of these disorders. This underscores the crucial need for an integrative approach, considering the perspectives of both patients and caregivers and acknowledging the complex interconnections between symptoms[200].

The US FDA Adverse Event Reporting System (FAERS), a public global repository gathering spontaneous reports on suspected ADRs from patients and healthcare professionals, stands as a powerful data source for this purpose[201]. Remarkably, patients offer unique insights into the experiences and impacts of ADRs on QoL, surpassing the information provided by healthcare professionals[202–206].

Moreover, network analyses, providing the means to investigate complex systems consisting of multiple interacting entities, present a promising avenue. Specifically, they enable the analysis and visualization of ADRs as interwoven symptoms and signs rather than isolated events[207], a composite syndrome encompassing psychosocial implications.

Our investigation into the intricacies of drug-induced impulsivity aligns with three overarching goals. The first is to untangle the components of the syndrome of drug-induced impulsivity. Through examining the interplay of symptoms and their consequences, we aim to gain a deeper understanding of how these syndromes manifest and affect patients' lives.

Our second goal is to identify distinct sub-syndromes within the broader spectrum of drug-induced impulsivity, i.e., whether symptoms fall into distinct clusters, with different clinical consequences.

Our third goal is to identify pivotal symptoms that centrally contribute to exacerbating the syndrome of drug-induced impulsivity. By pinpointing these key symptoms, we aim to pave the way for targeted interventions that alleviate the adverse effects on patients' lives.

In pursuing these goals, our study focuses on recipients of pramipexole and aripiprazole, chosen as representative instances: pramipexole, a dopamine agonist used in neurological conditions, and aripiprazole, a dopamine partial agonist employed in psychiatric disorders. Through this focused investigation, we aim to illuminate

the complexities of drug-induced impulsivity, exploring how the same ADR might manifest differently in various populations. Specifically, we anticipate that younger and more stigmatized individuals taking aripiprazole may bear a heavier burden compared to older individuals with stronger social support prescribed pramipexole. A better understanding of the impact of these ADRs on quality of life could contribute to informed decision-making for patients and caregivers, laying the foundation for interventions capable of alleviating the toll exacted by impulsivity.

6.3.3 Materials and Methods

6.3.3.1 Study Design

The study design (Figure 6.7) involved downloading and cleaning FAERS reports (Step 1), identifying aripiprazole and pramipexole recipients, and selecting cases recording drug-induced impulsivity within these two populations (Step 2). Disproportionality analysis defined the syndrome as events statistically co-reported with drug-induced impulsivity, rather than with other suspected reactions of the same drug (Step 3). Subsequently, three parallel network analyses identified sub-syndromes as clusters of co-reported events (Step 4). Finally, a Bayesian Network (Step 5) provided insights into the potential direction of associations and the secondary impact of drug-induced impulsivity (Step 5).

Step 1 - Data Preprocessing We downloaded FAERS quarterly data (January 1st, 2004, to March 31st, 2022) in ASCII format[208]. Adverse events were coded to the Medical Dictionary for Regulatory Activities (MedDRA®, version 25.0) preferred terms (PTs)[209], while drugs were standardized to active ingredients[210, 211]. The latest report version was retained, and rule-based deduplication was applied to reduce redundancy (cfr. https://github.com/fusarolimichele/DiAna).

Step 2 - Case Retrieval Analyzing aripiprazole and pramipexole recipients separately, we identified cases as reports recording impulsivity. Following FAERS coding of events to MedDRA, we employed PTs that were specifically curated for investigating drug-induced impulsivity within the FAERS database[186]. These PTs encompassed a range of manifestations, including gambling, hypersexuality, compulsive shopping,

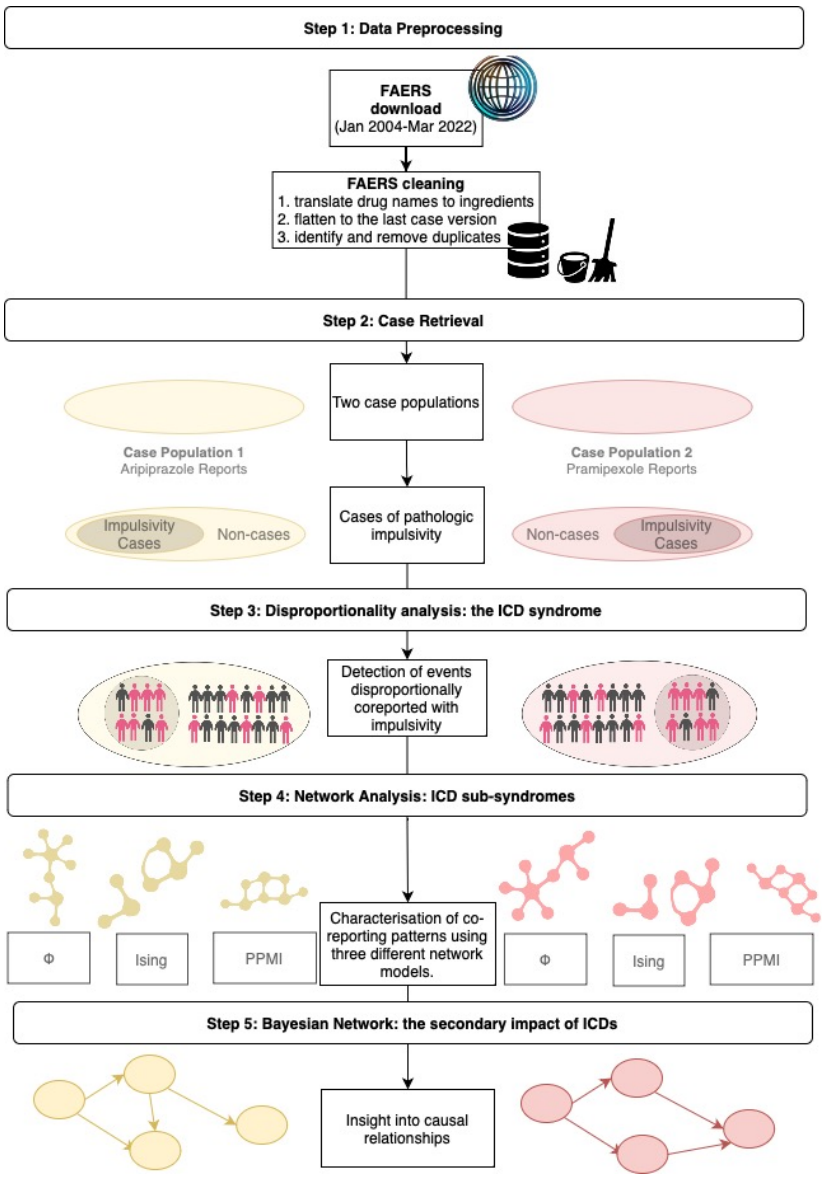


Figure 6.7. Pipeline of the Study. Showing step-by-step the study design.

hyperphagia, gaming, setting fires, stealing, hoarding, excessive exercise, overwork, compulsive wandering, body-focused repetitive behaviors, stereotypy, and impulsivity[186]. A cautious approach is imperative during interpretation, as MedDRA terms used for reporting suspected ADRs may not align directly with terms in other frameworks like the Diagnostic and Statistical Manual of mental disorders (DSM-5-TR) and ICD-11, which may refer to idiopathic conditions.

To explore potential risk factors for impulsivity, demographic characteristics, out-

comes, and reporter contributions (e.g., healthcare practitioners, patients, lawyers) were compared between cases and non-cases within each population, using the Chi-square test for categorical and Mann-Whitney test for continuous variables. To address multiple testing, we applied the Holm-Bonferroni correction with a significance level of 0.05.

	Event (y)	**Other events**	**Sum**
ICDs (x)	$n_{n_1E_1}$	$n_{n_1E_0}$	n_{n_1}
No ICDs	$n_{n_0E_1}$	$n_{n_0E_0}$	n_{n_0}
Sum	n_{E_1}	n_{E_0}	N

Table 6.6. 2-way Contingency Table. The table shows the different instances that can be observed when considering pathologic impulsivity and a specific event. Legend: E = event; I = ICDs; 1 = presence; 0 = lack; N = total.

Step 3 - Disproportionality Analysis: The Drug-induced Impulsivity Syndrome We conducted a disproportionality analysis to identify events frequently co-reported with drug-induced impulsivity, separately for aripiprazole and pramipexole recipients (see Table 6.6). Disproportionate reporting was assessed using the Information Component (IC)[212], also known as pointwise mutual information (PMI) in information theory[213, 214]. The IC compares the actual co-reporting of two events x (i.e., drug-induced impulsivity) and y (i.e., any specific event) with their expected co-reporting if their probability were independent[213]. To mitigate the risk of false positives for infrequent events[215], a shrinkage or smoothing approach was applied by adding $k = 0.5$ to both the numerator and denominator. Significance was determined using $IC_{025} > 0$ ($p(y, x) > p(x) \cdot p(y)$).

$$IC(x, y) = PMI(x, y) = \log_2 \frac{p(y, x)}{p(x), p(y)} = \log_2 \frac{n_{I_1E_1} \cdot N}{n_{I_1} \cdot n_{E_1}} \approx \log_2 \frac{n_{I_1E_1} + 0.5}{\frac{n_{I_1} \cdot n_{E_1}}{N} + 0.5} \tag{6.8}$$

$$IC(x, y)_{025} = IC - 3.3 \cdot (n_{I_1E_1} + 0.5)^{1/2} - 2 \cdot (n_{I_1E_1} + 0.5)^{3/2} \tag{6.9}$$

$$IC(x, y)_{975} = IC + 2.4 \cdot (n_{I_1E_1} + 0.5)^{-1/2} - 0.5 \cdot (n_{I_1E_1} + 0.5)^{-3/2} \tag{6.10}$$

Step 4 - Network Analysis: Sub-syndromes Building on insights from prior studies on drugs[216] and events[207, 217, 218] co-occurrence, our network analysis aimed to

unveil sub-syndromes. Using three established network estimations as distinct mathematical representations (PPMI, Ising, ϕ) and excluding negative links (i.e., potential mutually exclusive events), we explored co-reporting patterns.

The positive pointwise mutual information (PPMI) focused on cases where events were reported together (the n_{11} case), applying additive smoothing ($k = 1$, $d = N^{\circ}$ events)[219].

Bootstrap and Bonferroni adjustments assessed statistical significance, with a 0.01 threshold.

$$PPMI_{x,y} = \max\left(\log_2 \frac{(n_{11}+k)\cdot(N+kd)}{(n_{1*}+k)\cdot(n_{*1}+k)}, 0\right) \tag{6.11}$$

The Ising model computed partial logistic regression coefficients (β) considering the impact of all other events[220]. Positive coefficients meant two events tend to be reported together. A LASSO method pruned out weak links, eliminating spurious associations but potentially losing weak genuine relationships[221].

$$Ising_{x,y} = \max\left(\frac{1}{2}\beta_{x,y}\beta_{y,x}, 0\right) \tag{6.12}$$

The β coefficient[222], akin to the conventional correlation coefficient, is close to one when two events tend to be reported together, to zero if they are indifferently concomitant or mutually exclusive. The Bonferroni adjustment was applied with a significance threshold of 0.01. The p-value was computed by using a χ^2 probability distribution with one degree of freedom[222].

$$\phi_{x,y} = \max\left(\frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_{1*}n_{0*}n_{*1}n_{*0}}}, 0\right) \tag{6.13}$$

For each population, the three networks shared identical nodes but varied in links. We used modularity maximization[223] and the greedy modularity algorithm[24] to detect clusters of co-reported signs and symptoms. Between networks we compared degree of link overlap (Jaccard similarity index[224]), goodness of partitioning (clustering modularity); cluster agreement (Purity index[225–227]), link density (ratio between actual links and possible links), and interconnectedness among neighbors (small worldness[228]).

Step 5 - Bayesian Network: The Secondary Impact of Drug-induced Impulsivity

To uncover potential ramifications following drug-induced impulsivity, we estimated

the conditional probabilities of chained events. The resulting Bayesian network is both directed, offering insights into plausible causal relationships, and acyclic, ensuring that any chain originating from a node does not loop back to itself. The network was derived through 1000 bootstraps, optimizing the BIC score with the hill climbing algorithm. We computed the average network retaining links exceeding a threshold computed via L_1 minimization.
Evaluation focused on nodes with the highest out-degree centrality and the main manifestations of drug-induced impulsivity.

6.3.4 Results

6.3.4.1 Case Retrieval

After preprocessing the FAERS quarterly data, we retrieved 12,030,756 distinct reports: 27,601 pramipexole recipients and 80,238 aripiprazole recipients. Suspected drug-induced impulsivity was documented in 7.49% pramipexole recipients (n=2,066; mainly gambling disorder: 1,345, 4.87%; hypersexuality: 612, 2.22%; impulsivity: 453, 1.64%; compulsive shopping: 384, 1.39%;, and hyperphagia: 334, 1.21%) and in 4.50% aripiprazole recipients (n=3,609; mainly gambling disorder: 2,067, 2.58%; hypersexuality: 1,077, 1.34%; compulsive shopping: 1,029, 1.28%; hyperphagia: 868, 1.08%; and impulsivity: 730, 0.91%).
Among pramipexole recipients, drug-induced impulsivity was more frequently reported in males (57.42% vs. 36.99%, p<0.001), with lower median age (56 vs. 67, p<0.001), often non-serious outcomes (i.e., no death, disability, or hospitalization; 44.87% vs. 33.58%, p<0.001), and PD as indication (see Figure 6.8). Similarly, among aripiprazole recipients, drug-induced impulsivity was more common in males (48.59% vs. 40.72%, p<0.001), but with hospitalization more common (33.39% vs. 23.39%, p<0.001), and an important portion of reports submitted by lawyers (34.08% vs. 1.10%, p<0.001).

6.3.4.2 Disproportionality Analysis: The Drug-induced Impulsivity Syndrome

A total of 56 events were disproportionally reported with pramipexole-related impulsivity. The highest IC was found for obsessive-compulsive disorder (OCD, reporting rate = 26.77%; IC median = 3.47, 95% CI = 3.33-3.57), emotional distress

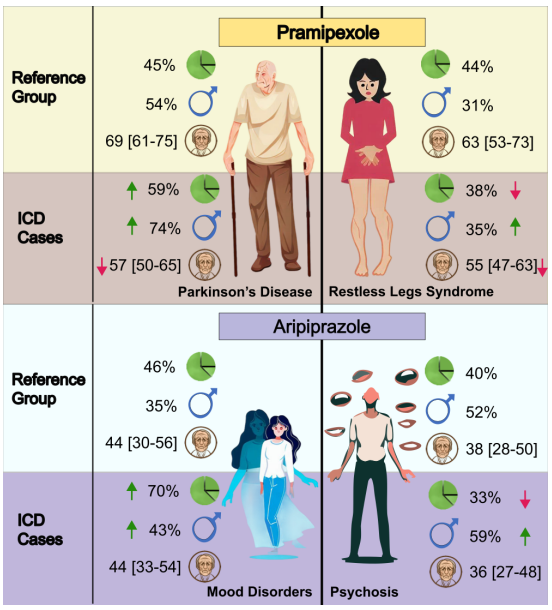


Figure 6.8. Characteristics of the Investigated Populations. The figure presents information about two populations extracted from the deduplicated FAERS database - one consisting of reports related to pramipexole and the other consisting of reports related to aripiprazole. Within these populations, cases of pathologic impulsivity were identified. The figure compares drug-induced impulsivity cases and the reference group (other reports recording the drug), considering the indication for use. Only the two most prevalent indications were taken into account. For each drug and indication, the caption describes the percentage of reports with the specified indication, the percent of reports involving males, and the median and interquartile range of ages. In the drug-induced impulsivity cases sections, green and red arrows indicate variables that are respectively higher or lower than expected based on the reference group.

(21.35%; 3.42, 3.26-3.54), marital problem (1.11%; 3.30, 2.61-3.79), dependence (2.37%; 3.26, 2.79-3.6), economic problems (6.05%; 3.15, 2.85-3.36), compulsions (1.74%; 3.05, 2.49-3.44), fear (4.65%; 2.95, 2.61-3.19), eating disorder (2.47%; 2.95, 2.49-3.28), personality change (2.66%; 2.93, 2.49-3.26), and suicide attempt (5.28%; 2.74, 2.43-2.97).

A total of 107 events were disproportionally reported with aripiprazole-related impulsivity. The highest IC was found for bankruptcy (10.58%; 4.43, 4.26-4.55), divorce (7.59%; 4.38, 4.19-4.53), homeless (6.93%; 4.37, 4.16-4.52), shoplifting (5.02%; 4.37, 4.12-4.54), neuropsychiatric symptoms (4.74%; 4.35, 4.1-4.53), loss of employment (12.64%; 4.33, 4.18-4.44), theft (5.79%; 4.32, 4.09-4.48), economic problems (37.85%; 4.28, 4.19-4.34), sexually transmitted disease (3.05%; 4.24, 3.93-4.47), and OCD (33.19%; 4.16, 4.07-4.23) (see Figure 6.9).

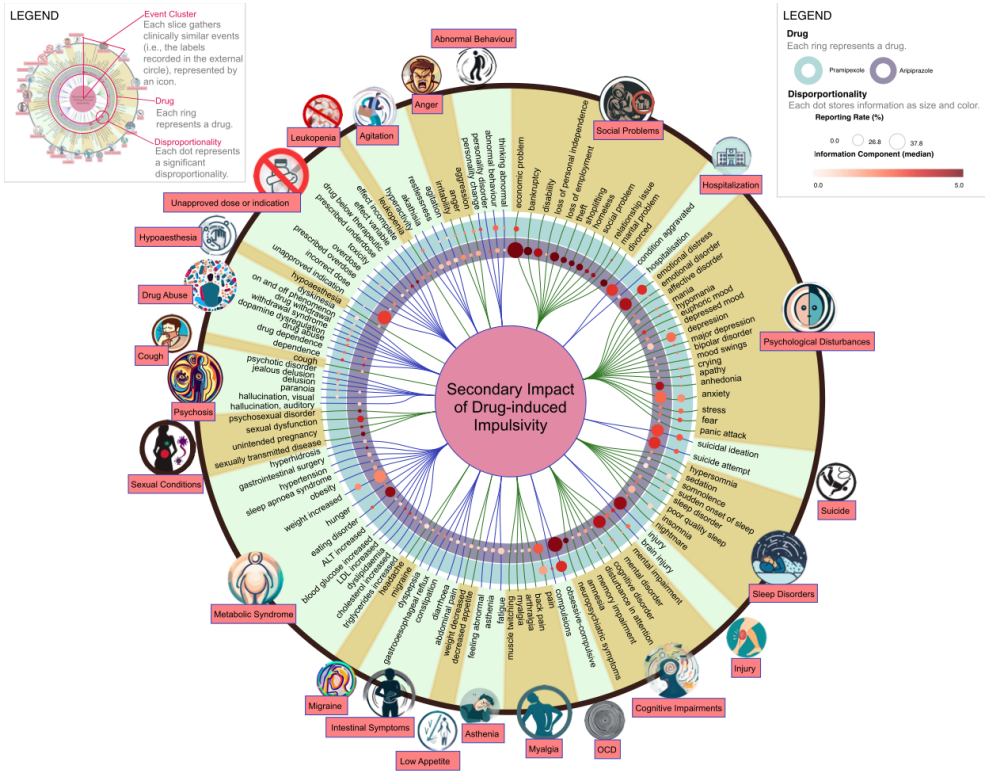


Figure 6.9. Secondary Impact of Drug-induced Impulsivity. The dendrogram shows the events disproportionally reported with aripiprazole and pramipexole-related impulsivity. Events are gathered by clinical similarity in alternately colored slices, labeled on the outer border with a name and an icon. Disproportionalities are shown as dots organized in two colored rings, each representing a drug/case population. The dots' size is proportional to the percent of reports showing the event, the color is darker for stronger disproportionality (higher median Information Component).

6.3.4.3 Network Analysis: Sub-syndromes

In the second step we estimated the networks using three different approaches and identifying clusters of events. We included a total of 120 nodes (107 events disproportionally reported with impulsivity + 13 impulsivity-related terms) and 70 nodes (56 + 14) for the aripiprazole and pramipexole network, respectively. Although the nodes remained constant, edges, clusters, and network properties were different (Table 6.7). The most central nodes (degree centrality) were the ones with the highest occurrence in Ising and the lowest occurrence in PPMI. The Jaccard similarity was higher for Ising-ϕ, while for ϕ-PPMI half of the links were different. The clustering was more overlapping between ϕ-PPMI, followed by ϕ-Ising and PPMI-Ising, as captured by the purity index. ϕ and PPMI tended to group together multiple Ising clusters. For

both drugs, the central clusters included cognitive disorders (e.g., cognitive impairment, memory impairment), bipolar disorder, and depression. Among aripiprazole recipients (Figure 6.10), the prominent cluster also includes stress and its psychophysical manifestations (irritability, headache, sleep disorders, decreased appetite, weight loss, constipation, and myalgia), together with panic attack and auditory hallucinations. Another stress-related sub-syndrome included migraine, nightmares, back and abdominal pain, arthralgia, reflux, diarrhoea, and hyperidrosis. Gambling and shopping were strictly related to social issues (hoarding, unemployment, homeless, bankruptcy, divorce, theft), suicide attempts, and through hypersexuality with unintended pregnancy and sexually transmitted diseases. Hyperphagia was co-reported with obesity, sleep apnoea syndrome, sedation, amnesia, and hypertension. Blood alterations, such as increased lipids, transaminases, and glucose in the blood, were reported together.

Among pramipexole recipients (Figure 6.11), the prominent cluster also includes apathy, delusion, and economic problems. A sub-syndrome included terms related to dopamine dysregulation syndrome (a manifestation of pathological impulsivity marked by excessive levodopa use[229], which can be co-administered with dopamine agonists to better control motor symptoms), such as drug dependence and withdrawal, and on and off phenomenon. Hallucinations, irritability, and crying were reported with delusional jealousy, hypersexuality, and marital problems. Hyperphagia was associated with weight increase, somnolence, insomnia, and disturbance in attention. Fear, anxiety, pain, stress, and depression were associated with suicide attempts. Finally, body-focused repetitive behaviors and stealing behaviors showed strong co-reporting.

6.3.4.4 Bayesian Network: The Secondary Impact of Drug-induced Impulsivity

The Bayesian Network yielded insights into the directional associations between co-reported events (see Figure 6.12). High out-degree centrality identified pivotal events that likely heightened the likelihood of reporting other events. Since this directed network only generates hypotheses, we preferred temporal terminology (i.e., preceding and following) to causal terminology even if no temporality was taken into account.

In pramipexole recipients, anxiety (3.55), emotional distress (2.92), and gambling (2.30) attained the highest out-degree centrality. Anxiety preceded insomnia (with irri-

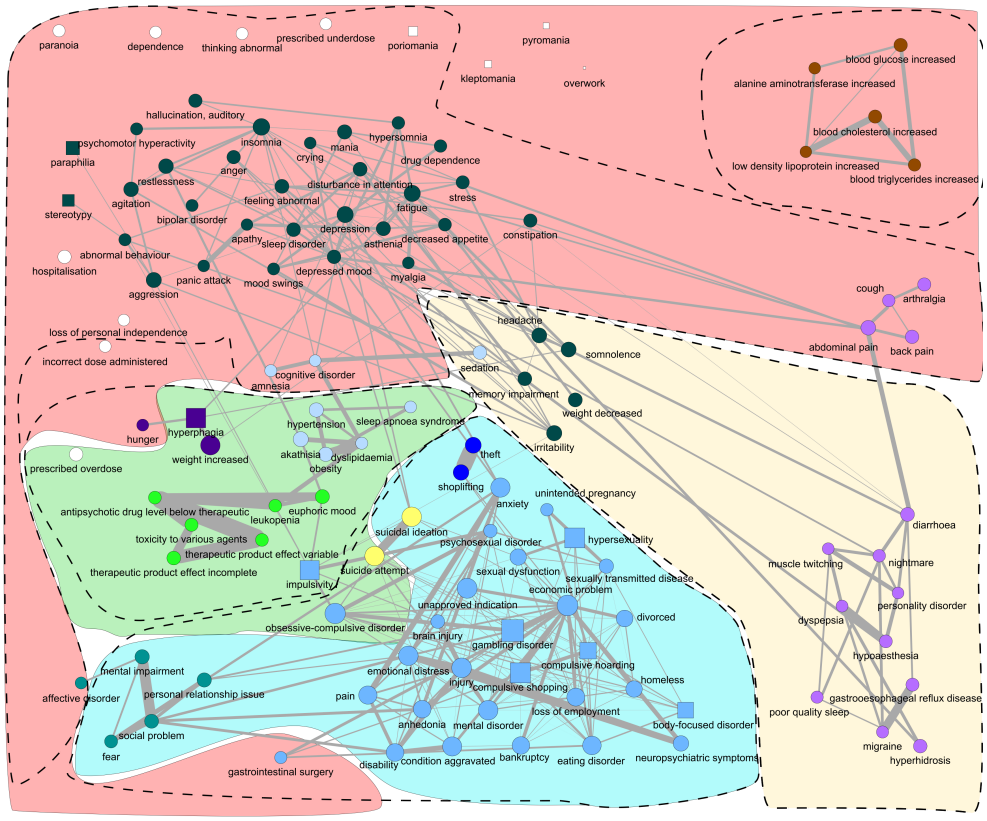


Figure 6.10. The Secondary Impact of Aripiprazole-induced Impulsivity. The network shows the events disproportionally reported with aripiprazole-related impulsivity and their pattern of co-reporting. Drug-induced impulsivity manifestations are shown as squares and other events as circles. Node colors identify clusters from the Ising estimation, dashed contours for the ϕ estimation, and colored contours for the PPMI estimation. The link width represents the weight of the links of the Ising, here chosen over the others because they are fewer and more conservative. The layout has been manually adjusted to reduce the overlapping. The layout calculated using a spring model with, as weight, the weights from the individual networks and the average of the weights of the three networks, after rescaling them from 0 to 1, is shown in the supplementary material.

tability, somnolence, and attention disturbances), stress and depression (with suicide), fear, OCD, and emotional distress. Emotional distress preceded pain and injury (with major depression and economic problems), abnormal thinking and behavior, weight gain, and pathologic gambling. Furthermore, hypersexuality preceded delusional jealousy and marital difficulties, compulsive shopping stealing behaviors, and hyperphagia weight increase.

In aripiprazole recipients, economic problems (5.97), gambling (4.15), and hyperphagia (2.33) attained the highest out-degree centrality. Economic problems preceded theft, hoarding, divorce, loss of employment, homelessness, suicide, sex dysfunc-

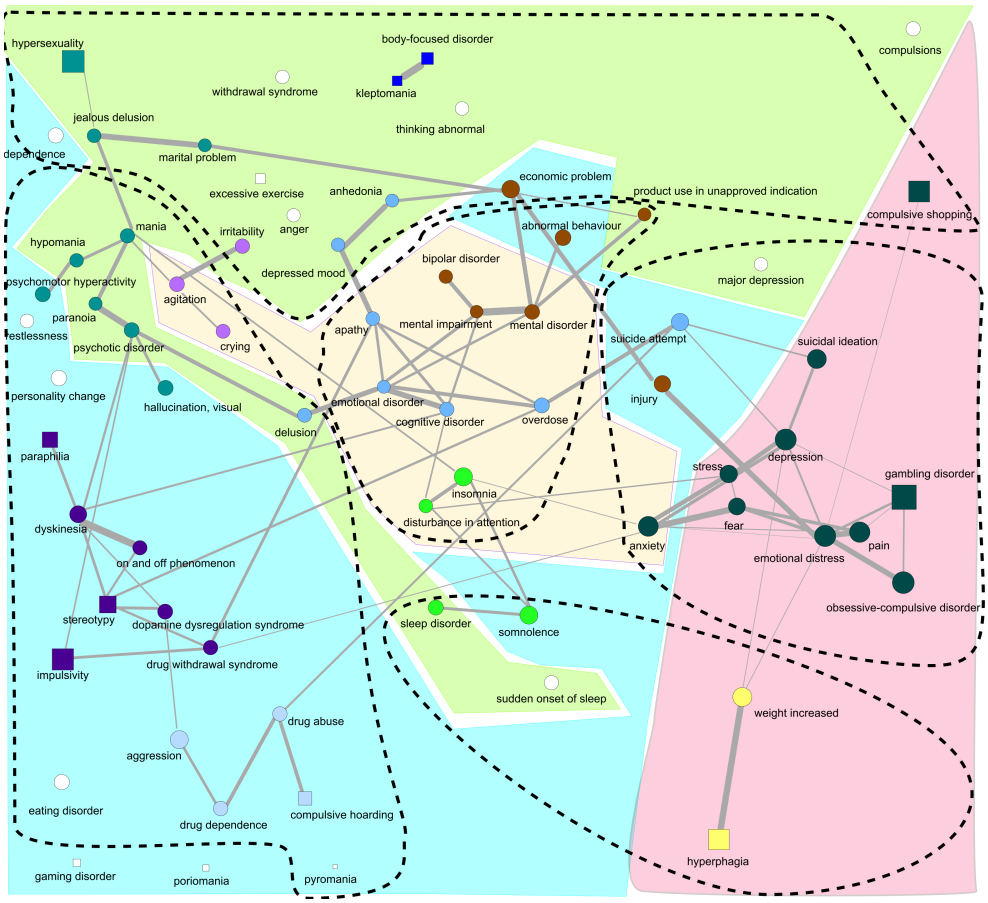


Figure 6.11. The Secondary Impact of Pramipexole-induced Impulsivity. The network shows the events disproportionally reported with pramipexole-related impulsivity and their pattern of co-reporting. Drug-induced impulsivity manifestations are shown as squares and other events as circles. Node colors identify clusters from the Ising estimation, dashed contours for the ϕ estimation, and colored contours for the PPMI estimation. The link width represents the weight of the links of the Ising, here chosen over the others because they are fewer and more conservative. The layout has been manually adjusted to reduce the overlapping. The layout calculated using a spring model with, as weight, the weights from the individual networks and the average of the weights of the three networks, after rescaling them from 0 to 1, is shown in the supplementary material.

tion, sexually transmitted diseases, and eating disorder. Gambling preceded aggressivity, suicide, cognitive disorders, hyperphagia, and paraphilia. Hyperphagia preceded somnolence and fatigue (with stress, attention disturbances, myalgia, cough), hunger, weight increase (with constipation), obesity (with hypertension), compulsive wandering, and paraphilic disorders. Anxiety preceded depression (with sleep disorders and suicide), fear and panic attacks (with relationship issues), pain and injury (with emotional distress, disability, anhedonia, and economic problems). Furthermore,

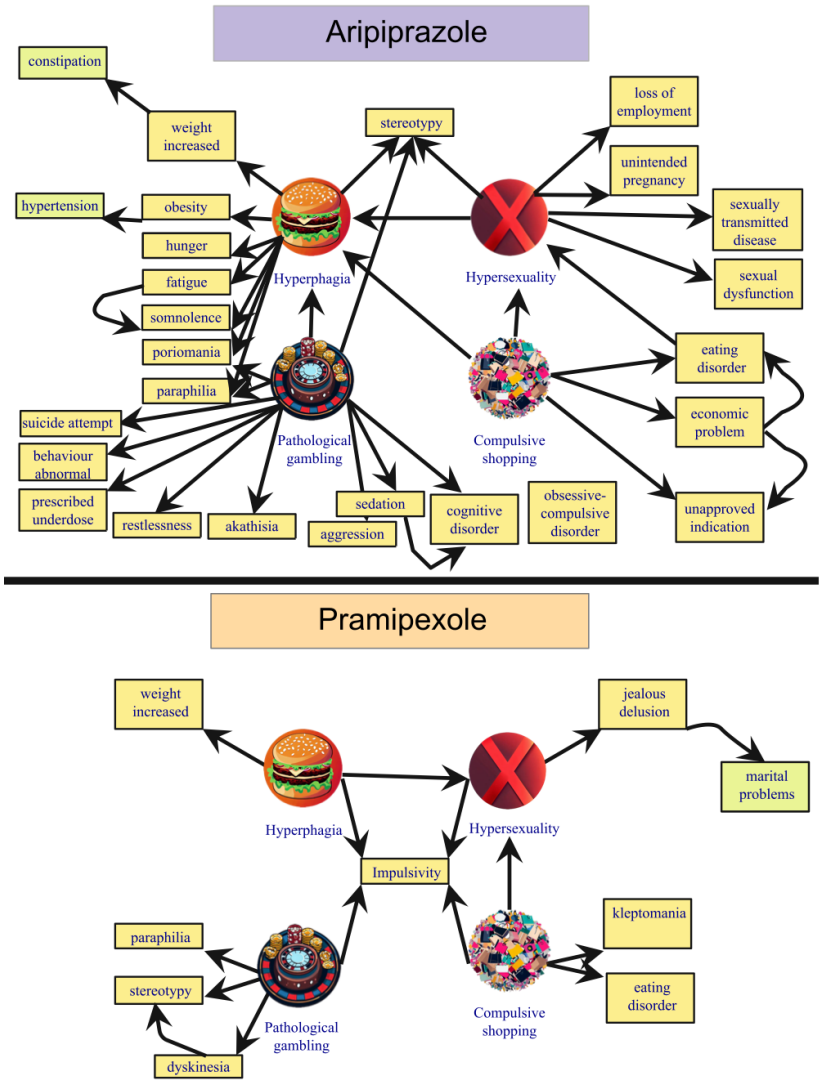


Figure 6.12. The Secondary Impact of the Main Drug-induced Impulsivity, Aripiprazole and Pramipexole. The ego-networks extracted from the Bayesian Network show the potential direction of the co-reporting relationships between the events, thus providing insight into the direct and indirect impact of drug-induced impulsivity. Nodes linked to hyperphagia, hypersexuality, pathological gambling, and compulsive shopping are represented. Only out-neighbors of order equal or less than 1 are shown here, together with out-neighbors of order 2 considered relevant for clinical interpretation.

hypersexuality preceded sexual dysfunction, sexually transmitted diseases, unintended pregnancy, and loss of employment, compulsive shopping eating disorders and economic problems.

6.3.5 Discussion

6.3.5.1 Summary and Key Results

Patients and caregivers should be informed about the potential impact of drugs inducing impulsivity on their QoL. Investigating aripiprazole and pramipexole, we captured the main clinical scenarios at risk of drug-induced impulsivity. Disproportionality analysis revealed features of the impulsivity syndrome for each scenario, encompassing mainly psychosocial events but also organic conditions. Network analysis identified sub-syndromes such as delusional jealousy (also known as Othello syndrome[230] and dopamine dysregulation syndrome (i.e., the excessive use of levodopa) in pramipexole recipients, and obesity-hypoventilation syndrome (historically Pickwickian syndrome) and social issues in aripiprazole recipients. The Bayesian Network highlighted directional associations, potentially suggesting secondary consequences of drug-induced impulsivity. Anxiety and economic problems emerged as pivotal events that could be potentially targeted to disrupt the chain of events and alleviate the burden of drug-induced impulsivity: for instance, monitoring and effectively managing anxiety or providing financial guidance or legal guardianship to prevent wasteful spending. Since marital problems affect caregivers' QoL and increase the risk of early placement in nursing homes[198], addressing delusional jealousy and economic problems, identified as factors preceding marital problems, may be critical for preserving wellbeing in pramipexole recipients.
While aripiprazole and pramipexole offer clear benefits, the substantial impact on patients' and caregivers' QoL should be acknowledged and considered in the monitoring and management of dopamine agonist therapies.

6.3.5.2 Case Retrieval

Our findings align with established risk factors, including male gender and younger age[231, 232], Parkinson's Disease (PD)[233, 234] and depression[235]. Commonly reported impulsivity manifestations included the "four knights"[185] (i.e., gambling, shopping, hyperphagia, and hypersexuality), garnering special attention due to their pronounced impact on QoL. Other manifestations were body-focused repetitive behaviors, paraphilic disorders, and hoarding.

6.3.5.3 Disproportionality Analysis: The Drug-induced Impulsivity Syndrome

Pramipexole and aripiprazole recipients differ significantly. Pramipexole is primarily administered to older patients with hypodopaminergic conditions, characterized by motor impairment and reduced motivational drive. These patients, well managed and supported by caregivers because of the later onset and clear neurologic origin of the disease may experience a mitigated drug-induced impulsivity burden. Conversely, aripiprazole is prescribed to younger patients with mood and psychotic disorders, often linked to hyperdopaminergic states and a pre-existing diathesis for impulsivity. Challenges for caregivers and social support are heightened in these cases due to earlier onset, psychiatric origins, and stigma, potentially leading to a greater burden. Over a third of aripiprazole cases were submitted by lawyers, suggesting potential overreporting for legal compensation (cfr., Abilify lawsuit)[236], but also a response to underdiagnosis by physicians hesitant to attribute behavioral changes to the drug in the presence of underlying psychiatric conditions. Intriguingly, an ascertainment bias may also arise because neurologists prescribing pramipexole may be less attuned to psychiatric issues than psychiatrists prescribing aripiprazole, further underscoring the contrast in the reported impact on QoL for these two drugs.
By performing the disproportionality analysis on each drug population, comparing reports involving impulsivity with those encompassing various reactions other than impulsivity, we addressed indication bias and other confounding factors. This comparative analysis served as a rigorous filter, allowing us to sift through the complex data and unveil the genuine characteristics associated with impulsivity, as well as those arising from the dynamic interaction between impulsivity and the underlying drug or disease, excluding traits tied solely to the underlying drug or disease.
This approach revealed a complex syndrome, characterized by psychosocial, cognitive, psychosomatic, and metabolic events. The burden of drug-induced impulsivity appears more pronounced in aripiprazole recipients, with functional (or psychosomatic) manifestations and social issues impacting work, relationships, and economics.

6.3.5.4 Network Analysis: Sub-Syndromes

Network analysis, employing three estimation methods, revealed potential subsyndromes associated with specific impulsivity expressions in the two populations (Fig-

ure 6.13). The Ising delineated well-defined clusters, while PPMI and ϕ emphasized inter-clusters relationships.

In both populations, cognitive and mood disorders, significant in their association with drug-induced impulsivity and contribution to disability development[237] played central role. Obesity-hypoventilation syndrome[238], involving weight gain, cognitive and sleep disorders, and sedation, was consistent in both populations, but seemingly heavier in aripiprazole recipients (also reporting obesity, sleep apnoea syndrome, hypertension, and metabolic blood alterations, highlighting the link between hyperphagia and diabetes onset[239]).

For aripiprazole recipients, the prominent cluster included sleep disorders and stress,

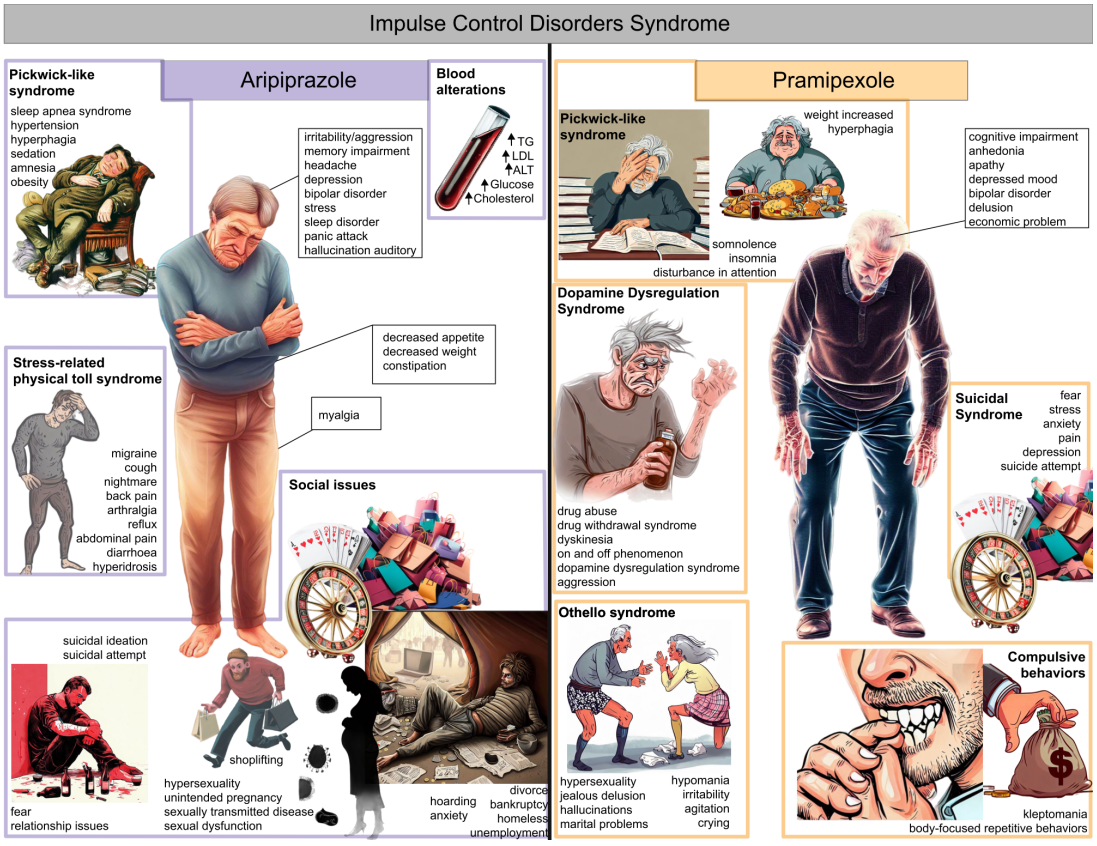


Figure 6.13. Drug-induced Impulsivity Syndrome, Aripiprazole and Pramipexole. The main syndrome, representing one or more strongly interconnected central clusters of symptoms and signs identified through network analysis, is depicted as the central figure. Other potential sub-syndromes are shown on the sides highlighted with a colored square.

connected to a psychosomatic sub-syndrome involving migraine, back and abdominal pain, reflux, diarrhoea, constipation, and hyperidrosis. Gambling and shopping

were linked to pervasive social issues, theft, and suicidal ideation (expected during hyperdopaminergic impulsive states[240]). Hypersexuality was linked to unintended pregnancy, sexually transmitted diseases, and sexual dysfunction.
Among pramipexole recipients, the prominent cluster included apathy, delusion, and economic problems. The dopamine dysregulation sub-syndrome[241], closely related to impulsivity but primarily associated with levodopa and apomorphine[242], involved on and off phenomenon (oscillations in effectiveness and motor and motivational symptoms), excessive levodopa use to avoid off phases, and dopamine agonist withdrawal syndromes (DAWS) upon discontinuation[243]. A cluster aligned with paranoid delusional jealousy (false and unwavering belief in the partner's unfaithfulness), often seen in PD with drug-induced hypersexuality[244]presented challenges in marital relationships, potentially resulting in early placement in a nursing home[245]. We also found a cluster with fear, pain, stress, anxiety, depression, and suicidal ideation, indicative of the transformation of reward-driven impulsivity into stressful risk-averting compulsivity over time[246]. Finally, the co-reporting of two archetypal compulsive symptoms, body-focused repetitive behaviors and stealing behaviors, was evident.

6.3.5.5 Bayesian Network: The Secondary Impact of Drug-induced Impulsivity

The interplay of events within the context of drug-induced impulsivity is intricate and multifaceted. Events reported alongside drug-induced impulsivity may result from impulsivity itself (like financial problems from gambling) or predispose individuals to impulsivity (e.g., bipolar disorder). Sometimes, events can both trigger and be exacerbated by drug-induced impulsivity (e.g., anxiety[247, 248]). Sometimes events are concomitantly mentioned for precision, such as in cases of semantic overlap (e.g., theft and shoplifting, or injury and brain injury). Events associated with drug-induced impulsivity may even be synonyms for well-known impulsivity expressions (e.g., restlessness, referring to excessive wandering and poriomania[244]), or could be the very reason for prescribing the drug, as seen in the off-label use of aripiprazole to prevent behavioral and cognitive decline in brain injury[249] or to address drug dependence[250, 251].
The Bayesian Network a directed acyclic graph representing our dataset, revealed potential directional associations, enabling formulation of hypotheses about clinically plausible causal sequences. Anxiety emerged as a central factor, preceding insom-

nia, irritability, cognitive impairment, stress, injury, pain (linked to disability and economic problems), depression, and even suicidal ideation. Drug-induced impulsivity manifestations appeared to exacerbate each other. Economic problems had the highest out-degree centrality among aripiprazole recipients, preceding theft, relationship difficulties, and suicidal ideation.
The Bayesian Network provides clinicians with valuable insights on the pivotal nodes that could be targeted by interventions to disrupt the cascade of events and ameliorate the secondary impact of drug-induced impulsivity. It also highlighted secondary ramifications of main impulsivity manifestations: hypersexuality precedes marital problems through delusional jealousy in pramipexole recipients, while it precedes unintended pregnancy and sexually transmitted diseases in aripiprazole recipients; hyperphagia precedes weight increase in pramipexole recipients and obesity, somnolence, and cognitive impairment in aripiprazole. Marital problems, following delusional jealousy and economic problems in pramipexole recipients, may be of particular interest since they are associated with an early placement in nursing homes[245]. Finally, the Bayesian Network seems to support the higher secondary impact of drug-induced impulsivity in aripiprazole recipients.

6.3.5.6 Limitations and Further Developments

While this study provides valuable insights into the intricate interplay of events related to drug-induced impulsivity and its subsequent implications, it is crucial to acknowledge its limitations.
Spontaneous reports, while uniquely granting access to patients' perspective, are susceptible to biases like under-reporting, missing data, and unverified reliability, preventing reliable incidence or prevalence estimates. The high contribution of reports from lawyers may have influenced the higher psychosocial impact attributed to aripiprazole-induced impulsivity. Nonetheless, this study sets the foundation for further studies and a potential score to assess the impact of ADRs on QoL.
Limitations in network analysis methodologies adopted include the Ising estimation's assumptions (pairwise interaction, linear effects, and binary variables), and the inability to account for time and severity in symptom manifestation. The incorporation of negative links could facilitate a more nuanced separation of symptoms that infrequently co-occur. The Bayesian Network lacks bidirectional relationships and cyclic

feedback loops and would require the inclusion of all shared cause between any two events (causal Markov condition) limiting its capacity to illuminate causality. These limitations could be rectified by integrating clinical longitudinal data and embedding temporal aspects into the network analysis.
Looking ahead, a broader definition of drug-induced impulsivity could improve sensitivity in case retrieval. Conditions like suicide attempts, hypersomnia, obsessive-compulsive symptoms, explosive anger, personality changes, disturbance in attention, and drug dependence might represent different expressions of this underdefined condition, warranting further exploration[186, 242].

6.3.6 Conclusion

The profound impact of drug-induced impulsivity reverberates across patients and their families, encompassing psychosocial challenges and organic complications such as metabolic syndrome (in the case of hyperphagia), and sexual health issues (in the case of hypersexuality). Recognizing these potential consequences is crucial for informed pharmacological management and diligent patient monitoring. Network analysis has revealed intriguing co-reporting patterns among adverse events, leading to their classification as sub-syndromes. Notable examples include is the emergency of obesity-hypoventilation syndrome with hyperphagia and associations of hypersexuality with delusional jealousy in pramipexole recipients and unintended pregnancy and sexually transmitted diseases in aripiprazole recipients. Our parallel approach effectively avoids the risk of disease-related diathesis compromising analytical integrity, enhancing the robustness of our findings.
For clinicians, this study emphasizes the potential burden of drug-induced impulsivity, and therefore the necessity of meticulous scrutiny into patients' medical histories. Factors such as age, gender, pre-existing mood disorders and family history should be red flags, warranting heightened vigilance. While transitioning to an alternative active ingredient may mitigate impulsivity, it is not always feasible or adequate. Monitoring for potential complications, as unveiled in our work (e.g., obesity-hypoventilation syndrome and delusional jealousy), is pivotal when such transitions are not a viable solution. For example, an overlooked delusional jealousy may result in marital problems and early nursing home placement.
Central to our findings is the pivotal realization that drug reactions rarely occur in iso-

lation; instead, they manifest as syndromes with diverse signs and symptoms. These can be direct reactions to the drug itself, secondary consequences to the reaction, risk factors for the reaction, or comorbidities. Causal chains and loops can contribute to symptom aggravation and chronicity. Identifying syndromes and sub-syndromes, combining network strategies with traditional techniques and clinical judgment, proves a potent strategy for delving into the secondary impact of adverse drug reactions and fostering heightened awareness within clinical practice.

In sum, the intricate relationships between signs and symptoms, coupled with the insights from the Bayesian Network, underscore the multifaceted nature of drug-induced impulsivity. More significantly, it equips clinicians with indispensable tools to discern intervention points, decipher causal sequences, and mitigate the cascading secondary effects associated with drug-induced impulsivity. In doing so, this study contributes to advancing our comprehension and management of drug-induced impulsivity, ultimately enhancing the well-being and care of affected patients.

Aripiprazole (120 nodes)			
gambling disorder (N=2057), economic problems (1366), obsessive-compulsive disorders (1198)			
	Ising	ϕ	**PPMI**
Links (density %)	301 (4.2%)	1185 (16.6%)	1254 (17.6%)
Central node (1°)	economic problems	irritability	overwork
Heaviest links ($1-3^\circ$)	AP below therapeutic – effect variable; theft – shoplifting; effect incomplete – effect variable	AP below therapeutic – effect variable; AP below therapeutic – toxicity; effect variable – toxicity	overwork – pyromania; kleptomania – overwork; overwork – poriomania
Clusters N	10	5	4
Clustering modularity	0.71	0.59	0.46
Small worldness (ω)[a]	0.26	0.04	0.03
Jaccard (out of max)[b]	0.25 (0.25)	0.56 (0.95)	0.21 (0.24)
Purity index[c]	0.68	0.89	0.59
Pramipexole (70 nodes)			
gambling disorder (N=1340), obsessive-compulsive disorders (553), hypersexuality (543)			
	Ising	ϕ	**PPMI**
Links (density %)	85 (3.5%)	240 (9.9%)	576 (23.9%)
Central node (1°)	gambling disorder	mental disorder	pyromania
Heaviest links ($1-3^\circ$)	body-focused disorders – kleptomania; mental impairment – mental disorder; on and off phenomenon – dyskinesia	emotional distress – pain; emotional distress – obsessive-compulsive disorder; hyperphagia – weight increased	poriomania – pyromania; gaming disorder – pyromania; gaming disorder – poriomania
Clusters N	10	6	4
Clustering modularity	0.66	0.51	0.15
Small worldness (ω)	-0.01	0.29	0.47
Jaccard (out of max)	0.34 (0.35)	0.23 (0.42)	0.12 (0.15)
Purity index	0.44	0.66	0.39

Table 6.7. Network Properties. The table shows the network properties for the three networks estimated for aripiprazole and pramipexole, respectively, and for their comparison.

[a] A small world has a $\omega = \frac{L_r}{L} - \frac{C}{Cl} \approx 0 \Rightarrow$ the shortest path length L is similar to that of an equivalent random network r and the clustering coefficient C is similar to that of an equivalent lattice network l.

[b] $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$.

[c] $Purity = \frac{1}{N} \sum_{k=1}^{K_{min}} \max_{\pi} n_{k,\pi}$, with K_{min} the minimum clusters, and $\max_{\pi} n_{k,\pi}$ the maximum elements.

7 Predictive Models

7.1 Unsupervised Pipeline for Drug Repurposing

Drug repurposing, introduced in Section 3.3, involves the identification of new therapeutic applications for established pharmaceutical agents, harnessing existing knowledge to accelerate drug development while preserving cost-effectiveness.
Within this work, as detailed in the article "An unsupervised computational pipeline identifies potential repurposable drugs to treat Huntington's disease and multiple sclerosis" published in *Artificial Intelligence in the Life Sciences* in 2022, an automated computational pipeline designed for selecting drugs for repurposing and screening their potential combinations was presented.[252] The process of drug selection is guided by the proximity of these drugs to disease-relevant genes within the protein-protein interactome, while considering the influence of these drugs on the expression of genes linked to the disorder. Furthermore, this approach incorporates the prioritization of combined therapeutic strategies, a procedure informed by the positioning of drug targets within the human interactome, also accounting for known drug-drug interactions. These efforts yielded a promising collection of molecules and potential combinations considerable for the treatment of Huntington's disease and multiple sclerosis. Notably, this pipeline extends its potential beyond these specific disorders, serving as a versatile tool for identifying novel therapeutic possibilities for other complex diseases.

Details

Authors Luca Menestrina, Maurizio Recanatini

Type Research Article

Status Published

Title An unsupervised computational pipeline identifies potential repurposable drugs to treat Huntington's disease and multiple sclerosis

Journal Artificial Intelligence in the Life Sciences

DOI 10.1016/j.ailsci.2022.100042

Data Availability The whole generated data is publicly available from the GitHub repository https://github.com/LucaMenestrina/UnsupervisedComputationalFrameworkForDrugRepurposing, as well as the full code for the collection, building and analysis. A detailed reference of the source data is provided in the file "`data/sources/sources.json`" of the aforementioned repository (for every database are reported: name, version, license, employed files, URL and date of access).
Supplementary data can also be accessed at the original publication.

An Unsupervised Computational Pipeline Identifies Potential Repurposable Drugs to Treat Huntington's Disease and Multiple Sclerosis

7.1.1 Introduction

Discovering a new drug and bringing it to the market is a process both money and time-consuming. Instead, relying on established drugs, computational drug repositioning offers a valuable alternative approach for providing promising treatments for disorders without a cure.[107, 253] In recent years, a plethora of computational approaches to drug repurposing have been proposed and applied to a wide variety of therapeutic areas.[254] Most of such approaches rely either on machine learning or on the traditional methods of computational drug design, even though some conceptually innovative ideas have brought to the light the possibility of taking new paths towards the prediction of potentially repurposable drugs. One of such ideas is based on a system view and takes the human protein-protein interactome as a reference network to quantify the relatedness between drugs and diseases by calculating the distance between drug targets and disease-associated proteins. This distance has been proposed as a suitable metrics to measure the "proximity" between drugs and diseases.[255] Recently, leveraging on the concept of drug-disease proximity,[256] novel drug indications for the treatment of cardiovascular diseases[103, 256], cancers[257], COVID-19[258], Alzheimer's disease[259] have been proposed, demonstrating how a network-based approach could successfully assist the selection of drugs to be repurposed.

In this work, we assembled an automated computational pipeline by integrating a recently developed scheme to screen repurposable drugs that combines a network-based technique with an analysis of biological and experimental data,[260, 261] with a strategy for filtering all the possible drug combinations.[103] Initially, the procedure estimates the proximity between the disease-related proteins and the drug targets on the protein-protein interactome, performing a first selection of candidates. Then, only those drugs that significantly influence the expression of disease-related genes are considered plausible for repurposing. Finally, evaluating the separation of these drugs' targets on the human interactome and taking into consideration the known drug-drug interactions, combined therapies are prioritized. The workflow of the procedure is

schematically illustrated in Figure 7.1. The entire process is automated in order to reduce human intervention, thus accelerating the whole procedure and limiting execution errors.

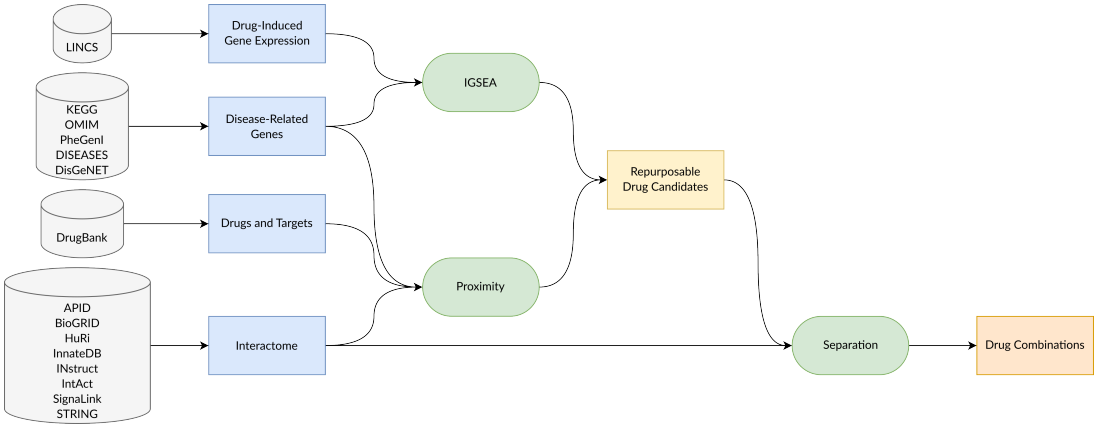


Figure 7.1. Pipeline Flowchart. The flowchart shows the sources and the steps of the automated procedure to screen repurposable drug candidates and prioritize their combinations.

We applied this pipeline to Huntington's disease (HD) and multiple sclerosis (MS) because, despite the fact that they are both neurological disorders, their different nature could represent a challenge for our strategy, and the outcomes could give us insights into its methodological strengths and limitations. HD, is reported as a typical monogenic disease, even though many other genes are known to influence its progression,[262] while for MS a single genetic cause has not been found yet, probably because many factors play an important role in the etiology. Indeed, MS fits well the definition of complex disease to be considered in the framework of network medicine. On the other hand, HD was included in our study in order to test the capabilities of the proposed method in a case where different clinical phenotypes might be related to a disease module eventually influenced by genetic modifiers leading to different pathophysiological states.[262–264]

HD is the most common monogenic neurological disorder. The onset is typically in the early stage of adult life, and it is characterized by motor dysfunction, cognitive impairment, and neuropsychiatric features.[262, 265] The autosomal dominant mutation that causes HD is located in the HTT gene, and it consists in a cytosine-adenosine-guanine trinucleotide repetition (CAG, encoding glutamine) leading to an

overexpansion of the polyglutamine (polyQ) tail in the huntingtin protein. The mutated protein tends to aggregate and accumulate, forming inclusion bodies that have deleterious consequences for the neural cell. Both the inclusion bodies and the length of the CAG expansion are proven to play an important role in the development of the disease. The clearance of the first ones slows the HD progression, while the longer the CAG expansion, the earlier the disease may manifest.[265] The remaining uncertainty on the course of HD can be ascribed to other genetic differences in the genome of the patients.[262, 264]

MS is both the most frequent non-traumatic disabling disease in young adults[266] and the commonest demyelinating disease.[267] The etiology and the mechanism causing its worsening progression are still unclear, nevertheless it has been proven that a complex interplay of genetic and environmental factors is important.[268, 269] The main known risk factors are smoking, childhood obesity, infection with the Epstein-Barr virus, and low vitamin D levels.[269]
MS is generally viewed as a two-phases autoimmune disease, in which initially focal inflammatory processes cause a relapsing-remitting form of the disease, and subsequently demyelinating plaques (lesions resulted by the previous immune response) and oligodendrocyte damage lead to neurodegeneration and non-relapsing progressive course.[267, 269] MS is commonly characterized by progressive spastic paraparesis, cognitive impairment, and sensory and cerebellar dysfunctions.[269]

Both HD and MS are still lacking resolutive treatments[270, 271], whose development needs a deeper knowledge of the underlying mechanisms[272]. To this aim, network-based models, as the ones we utilized in this study, could be adequate theoretical tools for investigating such multifactorial disorders. They would allow us to take into account the latent complex structure of these diseases without losing a comprehensive view[13]. Through the methodology presented here, we were able to collect a number of approved drugs and their plausible combinations that could be proposed for the treatment of HD and MS.

7.1.2 Methods

The workflow of this study can be outlined in the following steps (Figure 7.1): (1) collection of disease-related genes; (2) gene sets validation through enrichment

analysis; (3) collection of drugs, targets, protein interaction data, and construction of protein-protein interactome; (4) computation of drug-disease proximity on the human protein-protein interactome; (5) estimation of drug-induced gene expression signature enrichment; (6) calculation of drug-drug separation on the human interactome. Except for the collection of the disease-related genes, each operation is performed by a Python 3 script, and the entire procedure is brought together and coordinated by a main file in the same programming language.

7.1.2.1 Collection of Disease-Related Genes

For each of the two considered diseases, a set of related genes was retrieved from KEGG[273] (https://www.genome.jp/kegg/), OMIM [274] (https://www.omim.org/), PheGenI [275] (https://www.ncbi.nlm.nih.gov/gap/phegeni), DISEASES [276](https://diseases.jensenlab.org/search), and DisGeNET [141](https://www.disgenet.org/).
Briefly, for HD, 306 genes were retrieved from the KEGG Huntington Disease pathway "hsa05016"; 152 querying OMIM for "Huntington Disease"; 1 from the DISEASES database and 17 were those associated to "Huntington Disease" on DisGeNET and having an Evidence Index (https://www.disgenet.org/dbinfo#section36) of at least 0.95.
On the other hand, for MS, 160 genes were the result of querying OMIM for "Multiple Sclerosis"; 89 were collected from PheGenI with NHGRI (National Human Genome Research Institute) genome-wide association study as source and a p-value $< 1 \times 10^{-8}$; 5 were retrieved from the DISEASES database; 30 gathered from DisGeNET with the same conditions applied to HD. The genes were mapped to official gene symbols taking advantage of the NCBI database and then combined.

7.1.2.2 Ontology (GO, HPO) Enrichment Analysis

Functional enrichment analysis is often employed to perform a preliminary analysis on an investigated gene set. Examining the Gene Ontology[277] (GO, http://geneontology.org/) and the Human Phenotype Ontology[278] (HPO, https://hpo.jax.org/) associations, we gained insights on biological processes, molecular functions, cellular components and phenotypes most frequently associated to those genes. We conducted the functional enrichment analysis using the Python library GOATOOLS[279] and considered significantly enriched only those terms with a false discovery rate (FDR, p-value

corrected for multiple comparisons using the Benjamini-Hochberg procedure[159]) lower than 1×10^{-4}. We then looked at the first 20 terms ranked on the basis of their fold enrichment (computed as the ratio of the percentage of genes in the study set related to a specific term, divided by the corresponding percentage in the background, i.e., the entire human proteome).

7.1.2.3 Drugs and Targets Collection, Gene Expression Profiles Retrieval and Protein-Protein Interactome Construction

Drugs information was collected from DrugBank[114] (version 5.1.9). Only those molecules having at least one human protein as target are considered, obtaining 5,798 drugs and 2,755 corresponding targets.
Drug-induced gene expression profiles were retrieved from the Library of Integrated Network-based Cellular Signatures[280] (LINCS, profiles "GSE70138" and "GSE92742"), downloaded from Gene Expression Omnibus[281] (GEO, http://www.ncbi.nlm.nih.gov/geo). Due to the fact that we are inspecting neurological disorders, those signatures tested on neural cell lines (NEU, NPC, SHSY5Y) were examined for both diseases. Additionally, in order to consider disease-specific features, also muscular cell lines (SKB, SKL) were included for HD, and haematopoietic and lymphoid tissue cell lines (L60, JURKAT, NOMO1, PL21, SKM1, THP1, U937, WSUDLCL2) for MS. Furthermore, to guarantee maximum reliability of the results, only the data about the Best Inferred Genes (BING) in every dataset (drug signature) in these profiles was kept. The BING subset includes 978 landmark genes and 9 196 inferred genes, which are identified among the 12 328 genes in the L1000 assay by Subramanian et al.[282] evaluating the most reliable inference predictions.
Extensive interactions among proteins are a key factor in accomplishing many biological processes and functions. For this reason, we opted for a network-based approach to evaluate the correlation between drugs and diseases or drugs and other drugs. We built a human protein-protein interaction (PPI) network combining data from eight publicly available resources: Agile Protein Interactomes DataServer[283] (APID, http://cicblade.dep.usal.es:8080/APID/init.action), Biological General Repository for Interaction Datasets[284] (BioGRID, https://thebiogrid.org/), The Human Reference Interactome[285] (HuRI, http://www.interactome-atlas.org/), InnateDB[286] (https://www.innatedb.com/), INstruct[287] (http://instruct.yulab.org/), IntAct[288]

(https://www.ebi.ac.uk/intact/home), SignaLink[289] (http://signalink.org/), and Search Tool for the Retrieval of Interacting Genes/Proteins[140] (STRING, https://string-db.org/). Table B.1 gives additional info about the interactions reported in the databases and the applied filters.
The retrieved interactions were then combined, obtaining a network consisting of 20,445 nodes (genes/proteins) and 1,125,173 edges (interactions). Consistency is granted by the fact that all listed proteins are mapped to official gene symbols taking advantage of the NCBI database. Since the protein-protein interactome is the supporting pillar of the whole procedure, we assessed its validity comparing the results of the entire analyses based on two other interactomes. The first rerun was carried out on the widely recognized interactome from Cheng et al.[256] (16,677 unique proteins and 243,603 experimentally confirmed protein–protein interactions). The second one was performed on a drastically restricted version of our own interactome (16,954 proteins and 246,080 interactions), in which only interactions from low throughput studies (listing less than 20 interactions) were included.

7.1.2.4 Network Proximity

Proteins related to a specific disease are unlikely to be scattered throughout the interactome, rather, they tend to group together forming the so-called disease module.[290] The relationship between a drug and a disease could be estimated by means of an unsupervised and unbiased network-based approach[255], which quantifies the interplay of drug targets and disease-related genes measuring a network proximity. Here we used a recently modified version of such method[260] that includes a term (ω) for taking into account the degree of the drug targets directly into the distance calculation. Given G, the set of disease-related genes; T, the set of drug targets; and $d(g,t)$, the shortest path length between nodes g ($g \in G$) and t ($t \in T$) in the human protein-protein interactome; the distance $d(G,T)$ between each drug and the disease was calculated as:

$$d(G,T) = \frac{1}{|T|} \sum_{t \in T} \min_{g \in G} \big(d(g,t) + \omega\big) \tag{7.1}$$

where ω weights the targets based on their node degree in the interactome ($\omega = -\ln(D+1)$ if the target is related to the disease, $\omega = 0$ otherwise). D is the degree of the target in the PPI network.

Then, for each drug, the significance of its association to the investigated disease was assessed comparing the measured distance to that of a dummy reference distribution. This reference was obtained computing 10,000 times the distance ($d(G, R)$, defined by Equation (7.1)) between the disease-related genes and randomly selected (from the human interactome) sets of proteins (R) matching the number of the drug targets. Since the degree of the drug targets is already taken into consideration in the distance calculation, the sampling of the randomly selected proteins is facilitated having to match only the number and not also the degree distribution of the drug targets. The mean $\mu_{d(G,R)}$ and standard deviation $\sigma_{d(G,R)}$ of the reference distribution were used to normalize the observed distance into a proximity value (z-score):

$$z(G, T) = \frac{d(G, T) - \mu_{d(G,R)}}{\sigma_{d(G,R)}} \tag{7.2}$$

7.1.2.5 Inverted Gene Set Enrichment Analysis

Starting from the hypothesis that effective drugs should be able to restore the healthy expression of genes deregulated by a disease, the drugs with signatures most enriched in disease-related genes should also be the most promising ones in treating such disease. In order to gain this knowledge, an Inverted Gene Set Enrichment Analysis[260] (IGSEA) on the datasets (drug signatures) of LINCS was performed, looking for the disease-related genes under study. For each analyzed dataset, the normalized enrichment score and the p-value (estimated comparing the enrichment score with those of a null distribution generated from 100,000 permutations) were computed for measuring the enrichment magnitude and its statistical significance, respectively. The resulted p-values were then corrected for multiple comparison using the Benjamini-Hochberg procedure[159], obtaining the FDR. If the dataset was significantly enriched (FDR < 0.25), the corresponding drug was considered a potential drug candidate.

7.1.2.6 Network Separation

An important aspect in investigating drug combinations is to evaluate whether the two drug-target modules are overlapped (overlapping exposure) or separated (complementary exposure) on the human protein-protein interactome.[103] In the case of overlapping exposure, there is a higher similarity in chemical, biological, functional,

and clinical profiles. The desired combinations, instead, are those with complementary exposure, both drugs being topologically and pharmacologically distinct. In the latter case, the two drugs synergistically cooperate in treating the disease, yet each one in its own way.
As we did for computing the drug-disease proximity, also for measuring drug-drug separation s_{AB} in drug combinations, we employed a network-based approach[103, 290]:

$$s_{AB} = \langle d_{AB} \rangle - \frac{\langle d_{AA} \rangle + \langle d_{BB} \rangle}{2} \tag{7.3}$$

where A is the target module of one drug and B that of the other. Here, the mean shortest distances (calculated with Equation (7.1) with the weight ω fixed to 0) between the target modules of each drug ($\langle d_{AA} \rangle$ and $\langle d_{BB} \rangle$, computable only for drugs with at least two targets) are compared to the mean shortest distance between all possible A-B target pairs ($\langle d_{AB} \rangle$). When computing the distances between A-B target pairs, if a protein is targeted by both drugs, its distance is zero by definition. A drug combination exposure is deemed complementary if $s_{AB} \geq 0$, overlapping otherwise.

7.1.3 Results

7.1.3.1 Computational Framework

In this study, we automated the pipeline shown in Figure 7.1 for screening repurposable drug candidates and prioritizing their combinations. In order to run, the script only requires the disease name, the disease-related genes, and the cell lines of interest as inputs. This procedure consists of: collecting, cleaning and organizing the source data (disease-related genes, drugs, targets, protein interactions, drug-induced gene expression signatures); identifying repurposable drug candidates evaluating both their proximity to the disease and their effect on the expression of the disease-related genes; screening the possible drug combinations on the basis of their relative exposure and known interactions. The output of the routine is a collection of tables (tab-separated values files) and plots, recording both intermediate and final results.
Compared to previous related works[260, 261], such a systematic strategy should be more efficient and have an improved reproducibility thanks to the organization and standardization of both the overall study and results. Additionally, it takes a step forward since it evaluates also possible combined therapies.

The single steps and the outcomes of the application of the framework to HD and MS are presented and discussed in the following.

7.1.3.2 Disease-Related Genes Collection and Validation

We gathered the disease-related genes as described in the Methods section: this resulted in 451 and 217 genes associated to HD and MS, respectively. In order to evaluate whether these genes were representative of the investigated diseases, we performed an enrichment analysis on GO and HPO terms. This allowed us to check if the most enriched biological processes, molecular functions, cellular components and phenotypes were in accordance with previous knowledge.

Prior studies relate HD to dysfunctions in transcription, intracellular signaling, intracellular transport, endocytic recycling, and mitochondria.[262] This knowledge is consistent with the biological processes, cellular components and molecular functions that we found to be enriched (Figure 7.2A, Figure B.1A and B). The same holds for the phenotypes, which are associated to negativism, social and occupational deterioration, mitochondrial and nervous issues (Figure B.1C).[262, 265]

MS is an autoimmune disorder whose inflammatory infiltrates contain T-lymphocytes and B-cells, and leads to oligodendrocyte damage and demyelination.[269] This is coherent with the enriched terms in our analysis (Figure 7.2B and Figure B.2).

The fact that the obtained results were confirmed by the literature suggested that the gathered genes were representative of the diseases under study.
Furthermore, as in Menche et al. [290], the disease modules were tested to be non-random gene aggregates. The size of the largest connected component of the disease module was compared to the size of the one obtained by randomly picking the proteins (matching the number of the disease-related genes) from the interactome. For both diseases, the disease module resulted to be significantly larger than the random counterpart, allowing us to state that they cannot be attributed to a casual aggregation of genes.

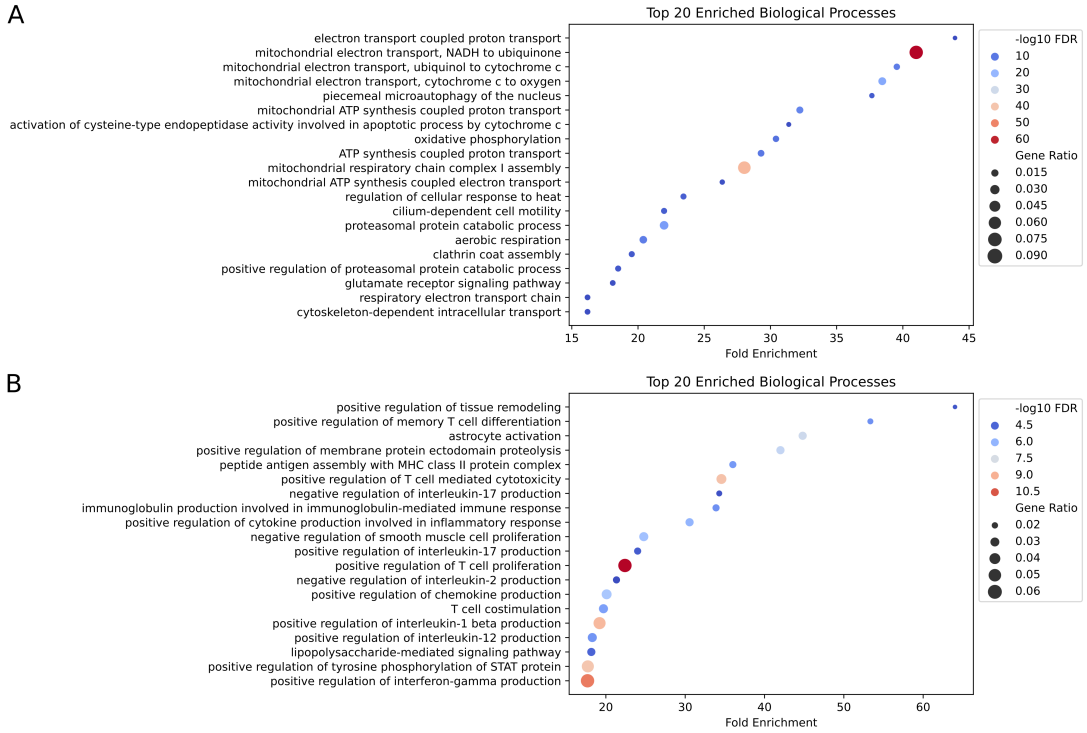


Figure 7.2. Enriched Biological Processes. The bubbleplots display the top 20 most enriched Gene Ontology terms relative to biological processes for Huntington's disease (A) and multiple sclerosis (B). On the horizontal axis, the fold enrichment is shown. The color encodes the negative of the false discovery rate logarithm, and the size represents the gene ratio (computed as the ratio of the percentage of genes in the study set related to a specific term, divided by the corresponding percentage in the background, i.e., the entire human proteome).

7.1.3.3 Repurposable Drugs Selection

The network-based proximity analysis, leveraging on the potential of a system view, could suggest valuable drugs able to interfere with the disease molecular determinants in a non-trivial way (i.e., not only directly targeting disease-related genes). The idea behind this method is that drugs proximal to the disease module should be more effective than distant ones, as shown by Guney et al. in an extensive analysis that considered known diseases and disease-associated genes, as well as drugs and their targets.[255] Following Peng's protocol[260], the procedure compares the distribution of the distances between drug targets and disease-related proteins to that of a reference collection (see Section 7.1.2 and Figure 7.3). For both diseases, it was possible to identify a distance value below which the two density curves (drugs and reference) drop dramatically. In particular, the reference density assumes negligible values for distances

below this point (Figure 7.3, green part of the plot). We elected such distance value (Figure 7.3, vertical green line) as the threshold to discriminate drugs associated to the diseases. These distances are 0.68 and 0.71 (corresponding to proximity: -0.53 and –0.98) for HD and MS, respectively. From this analysis, 685 (11.8%) out of the 5,798 drugs collected from DrugBank were considered significantly proximal medicaments for HD, and 475 (8.2%) for MS.

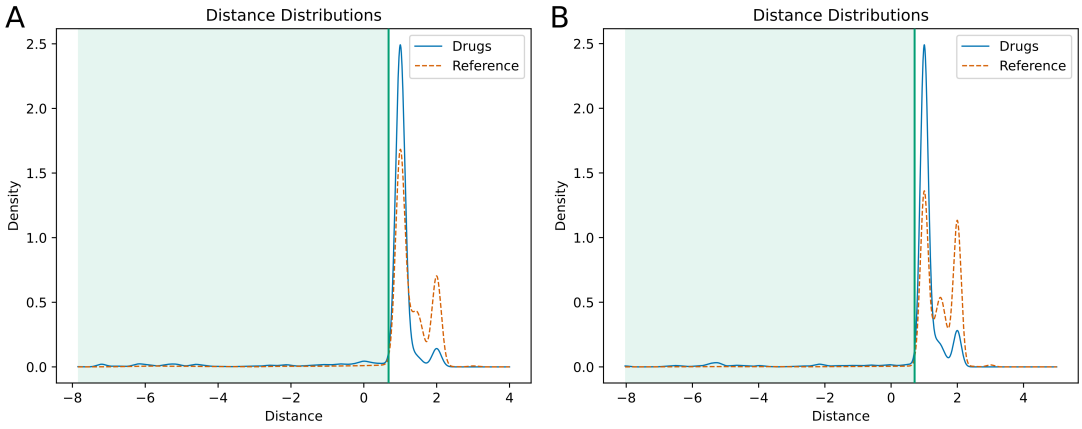


Figure 7.3. Distance Distributions. The distribution of the distance between drug targets and disease-related proteins (solid blue line) compared to that of a reference collection (dashed orange line), for Huntington's disease (A) and multiple sclerosis (B). On the vertical axis, the kernel density estimation of the distribution is shown. The plot is divided into two parts by the chosen distance threshold (green line, see Section 7.1.2).

In order to evaluate the impact of a drug on the disease, we examined the effect of its administration on the expression of the disease-related genes in relevant cell lines (see Section 7.1.2). We pursued this objective by performing an Inverted Gene Set Enrichment Analysis (IGSEA) on 896 drugs, observed in 6,212 LINCS expression datasets for HD and 960 drugs in 5,579 datasets for MS. This analysis resulted in 843 and 600 significantly enriched drugs, for HD and MS respectively.

The drugs that were both significantly enriched and proximal to the disease were deemed to be repurposable drug candidates: 138 for HD and 38 for MS. The interactions between the MS-related-genes, the drug targets, and the repurposable drug candidates are visualized in Figure 7.4 (and Figure B.3 for HD), showing how drugs can be related to the disease through their targets.
Unfortunately, only a small portion of the proximal drugs has data in the LINCS

database (21.9% for HD and 13.7% for MS). Even though the IGSEA analysis increases the reliability of the results, it dramatically reduces the number of molecules that can be investigated and possibly proposed. This has to be taken into account when evaluating the outcomes of the study.

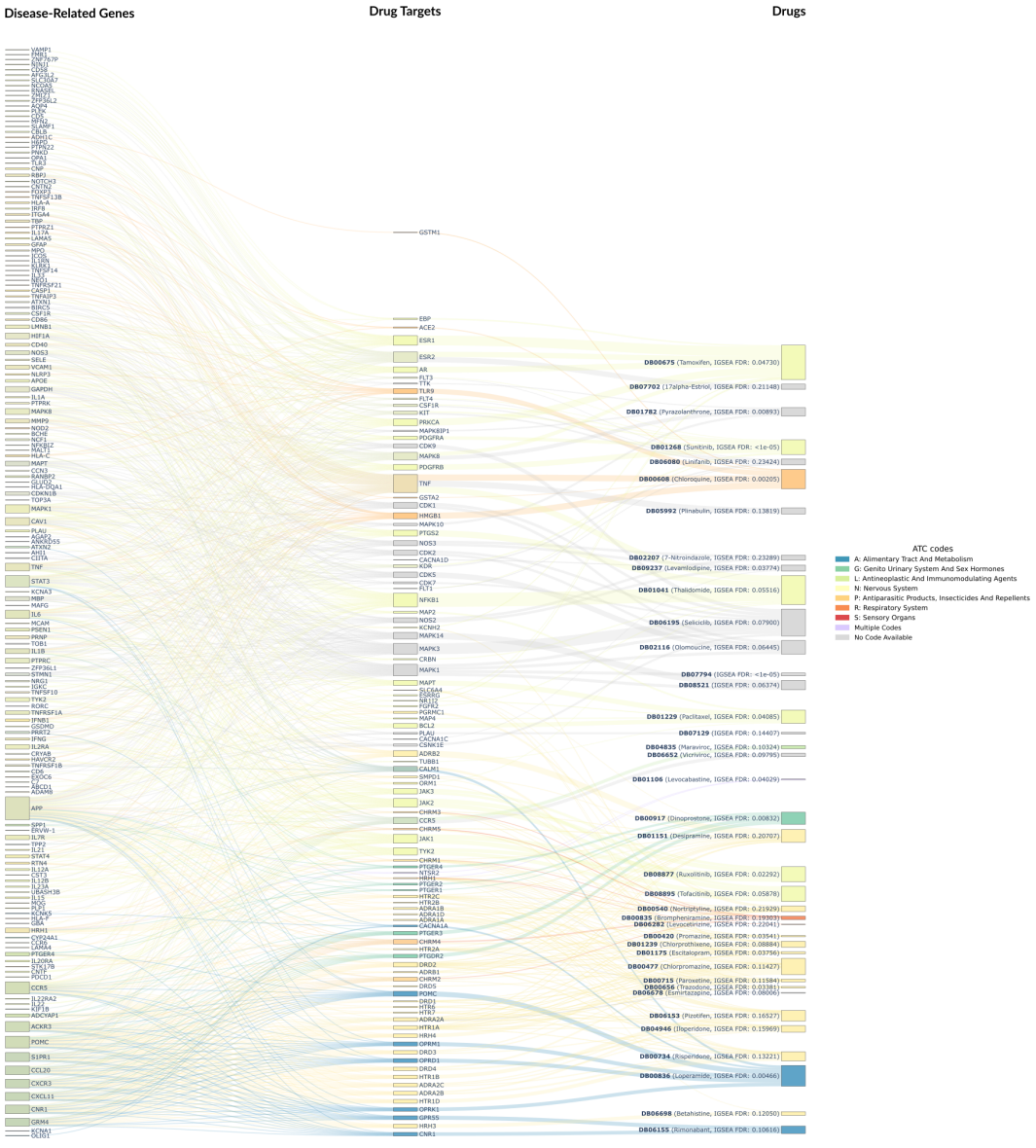


Figure 7.4. Multiple Sclerosis Gene-Target-Drug Network. The Sankey diagram illustrates the interconnections between disease-related genes, drug targets, and drugs. Each drug (right column) is connected to its reported targets (middle column), which, in turn, are proximal on the human interactome to some of the disease-associated proteins (left column). Drugs are colored by the respective ATC code, and the FDR of the IGSEA analysis (see Section 7.1.2) is reported in the label.

To be more confident on the pool of predicted repurposable drugs, we replicated the entire procedure using three different interactomes. Our original one and the two networks used to validate it differ both in size and in listed interactions (see Section 7.1.2). Despite these differences, the repurposable drug sets suggested for both investigated diseases resulted fairly consistent. In the case of HD, 138 drugs were prioritized based on the original interactome, 110 on Cheng's one, 133 on our restricted interactome. It is noteworthy that all the molecules retrieved from the two smaller interactomes are included among those of the first one. A very similar conclusion could be drawn for MS, for which the procedure predicted 39 drugs with the large interactome, 26 with Cheng's one, and 29 with the severely constrained version of our interactome.

7.1.3.4 Drug Combinations

Combined therapies and multi-targeted agents have proven to offer significant advantages over monotherapy, presenting higher efficacies and less adverse reactions.[103, 291] Due to combinatorial explosion, however, it is generally not feasible to test all theoretically possible associations. For this reason, we adopted a recent methodology proposed by Cheng et al.[103], which is based on the estimation of target neighborhoods separation on the human protein-protein interactome. Taking advantage of that, the investigated combinations may be screened on the basis of the pharmacological relationship between drugs (see Section 7.1.2). Additionally, we looked in DrugBank for approved associations and interactions that increase the risk or severity of adverse effects. In this way, we ended up having an assortment of plausible combinations and identifiable in the annotated heatmaps of Figure 7.5 for HD and of Figure B.4 for MS.

7.1.4 Discussion

7.1.4.1 Protein-Protein Interactomes

For both diseases, the repurposable drug pools predicted using the three interactomes are in reasonable concordance. However, it is significant that the execution of the pipeline on our interactome, compared to the same procedure on Cheng's interactome, improves the prediction adding 7 drugs with evidence from clinical trials, 9 from in vivo studies, 1 from in vitro experiments for HD, and 5 clinically tested drugs and

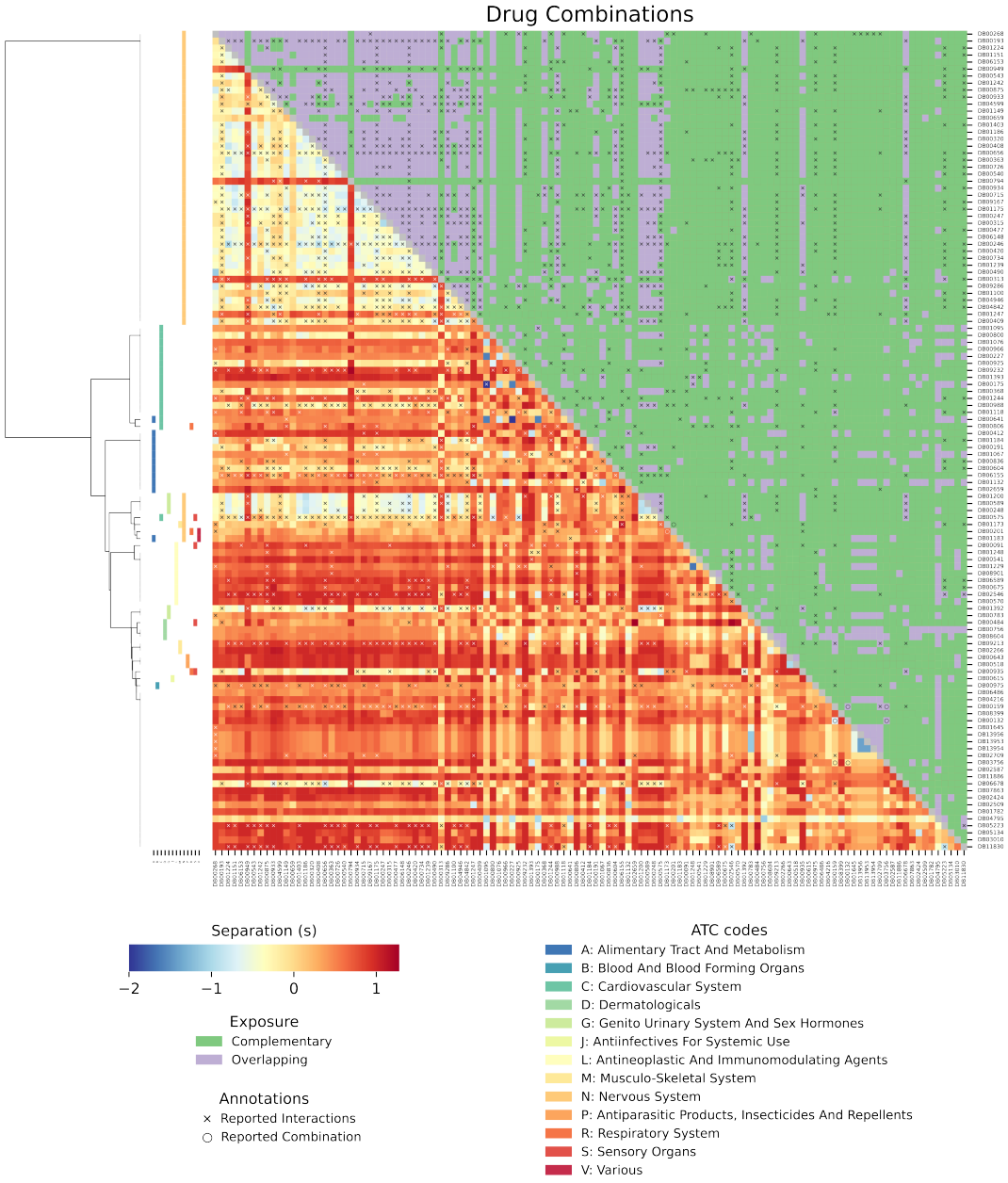


Figure 7.5. Huntington's Disease Drug Combinations. The annotated heatmap provides info about possible combinations of the selected drugs. A combination is marked with × if an interaction is reported in DrugBank, and with ○ if it is present in an approved formulation. The lower-left part of the heatmap shows the separation of the inspected drugs, color coded from blue (no separation) to red (strongly separated). The upper-right portion, instead, displays the kind of exposure: violet if overlapping and green if complementary. At the leftmost part, the ATC codes of the drugs are reported along with a dendrogram of their hierarchical clustering.

1 investigated in an animal model for MS.
This outcome seems to suggest that injecting more input data in the procedure (still maintaining high reliability standards) leads to increased performance, which is perfectly in line with the Big Data perspective.[50]

7.1.4.2 Repurposable Drugs

Among the drugs selected to tackle HD (138), several (17) have been clinically tested and suggested, many show strong evidence from in vivo tests (35) or promising results from in vitro assays (9). The most noticeable examples being selisistat[292], lisuride[293], valproic acid[294], and risperidone[270].

Selisistat was found to be safe, well tolerated, and capable of reaching a plasma concentration compatible with the SirT1 inhibition, which has been shown to restore transcriptional dysregulation in models of HD.[292] Lisuride is able to induce a temporary yet significant improvement in the motor performance of patients with hyperkinesia caused by HD.[293] Valproic acid was shown to be a possible alternative treatment for HD patients suffering from myoclonic hyperkinesia.[294] Risperidone has beneficial effects in the treatment of psychiatric manifestations and stabilization of motor symptoms in patients with HD.[270]

Inspecting the drugs screened for MS (38), we obtained a comparable outcome: 7 of them are clinically studied and 9 experimented on animal models. Most of the drugs in clinical trials aim to alleviate the symptoms, while the only one we found to be capable of reducing relapses is Escitalopram[295] for which there is evidence suggesting it may be an effective and well-tolerated treatment for preventing stress-related relapses in women with MS[295].

Examining the Anatomical Therapeutic Chemical (ATC) codes of the repurposable drugs, the first thing to notice is the predictable prevalence of drugs associated to the ATC code N (Nervous System) for both diseases. Apart from this, the most common codes for HD repurposable drugs are C (Cardiovascular) and L (Antineoplastic and Immunomodulating Agents). The first group is mainly represented by statins, used to cope with the cholesterol impairment typical of HD patients.[296] The immunomodulating agents are principally immunosuppressants and histone deacetylases

inhibitors, the last ones aimed at recovering from the histone hypoacetylation common in neurological disorders.[297]
For MS, instead, the second most frequent code is L (Antineoplastic and Immunomodulating Agents). Some relevant examples are ruxolitinib, paclitaxel, tamoxifen, and thalidomide, which are capable of attenuating experimental autoimmune encephalomyelitis and in inducing remyelination.[298–302]

7.1.4.3 Drug Combinations

Observing the obtained results (depicted as annotated heatmaps in Figure 7.5 for HD and Figure B.4 for MS) it is interesting to highlight that drugs that do not have ATC codes associated to them are also those with few (or nothing at all) reported interactions. This suggests that they are not sufficiently characterized and additional studies on them are needed before further consideration.
The collected plausible combinations are numerous, but the association of orphenadrine (DB01173) and caffeine (DB00201) for HD deserves to be highlighted.
These molecules are present along with acetylsalicylic acid (ASA) in an FDA approved formulation for muscular pain relief. This medication is noteworthy for many reasons. First of all, pain is a known issue in HD and could be an important non-motor symptom[303, 304] thus, its treatment should not be neglected. Furthermore, orphenadrine showed to be effective in preventing neurotoxicity in rats with a chemically-induced condition that mimics the histological and neurochemical features of HD.[305] Additionally, low dosages of caffeine showed to be beneficial in HD animal models.[306] Finally, ASA was included in the formulation for relieving pain and decreasing swelling. Even though ASA was proximal to HD, it was not included in our results because its data was not available in LINCS for the investigated cell lines. However, it is actually profitable for the present aim, since it showed to prevent protein aggregation in several neurodegenerative diseases.[307] Further assessments are needed, but this could be an interesting point where to start.
A reasonable hypothesis for treating MS might be an association of two drugs sufficiently separated from each other as escitalopram (DB01175) and tofacitinib (DB08895) or ruxolitinib (DB08877), capable of affecting complementary parts of the disease module. Figure 7.6 shows the network of the interactions among proteins associated to MS (all circles) and highlights those targeted by escitalopram and

tofacitinib (dark green and dark blue, respectively). Among these targets, two of them, namely HRH1 for escitalopram (dark green circle) and TYK2 for tofacitinib (dark blue circle), belong to the MS disease module, while those that are not directly related to MS are depicted as triangles (maintaining the same color coding). In order to better illustrate the influence on the disease module of the two drugs in terms of protein-protein interactions, the first neighbors of the drug targets are colored lighter (light green for neighbors of targets of escitalopram, and light blue for tofacitinib's ones). It can be seen that overall the targets of both escitalopram and tofacitinib or their first neighbors can influence a reasonable part of the disease module without redundantly interfering with the same MS proteins. In fact, our analysis showed that these drugs are proximal to MS and significantly influence proteins associated to this disease. Additionally, no interactions between them have been reported in DrugBank. Moreover, we found experimental evidence supporting this inference. Escitalopram is a selective serotonin re-uptake inhibitor (ATC code: N, Nervous) that in humans proved to prevent stress-related relapses.[295] Tofacitinib and ruxolitinib showed promising effects in animal models: the first one enhancing remyelination and improving myelin integrity[308], and the second one ameliorating the severity of the disease[298]. Furthermore, they are Janus kinase (JAK) inhibitors (ATC code: L, Antineoplastic and Immunomodulating Agents) and the JAK/STAT pathway is aberrantly activated in MS.[271, 309]
In the other drug combinations, which are sufficiently separated (see Section 7.1.2, green on the heatmaps) and for which no adverse interactions are reported (not annotated with an $\times$ in the heatmaps), valuable clues for polypharmacological interventions could be found. A working hypothesis might be to choose two drugs tackling different aspects of a disease, for instance belonging to distinct ATC codes.

7.1.5 Limitations

Despite our best efforts, this study is not exempt from some shortcomings that are common in data analysis, and regard mainly the data availability and quality. This could have led us to miss some promising compounds and, at the same time, it may compromise some of the analyses.
A complete characterization of all available drugs and human proteins is surely not at hand, and this has repercussions on many aspects of the study, like, e.g., the human

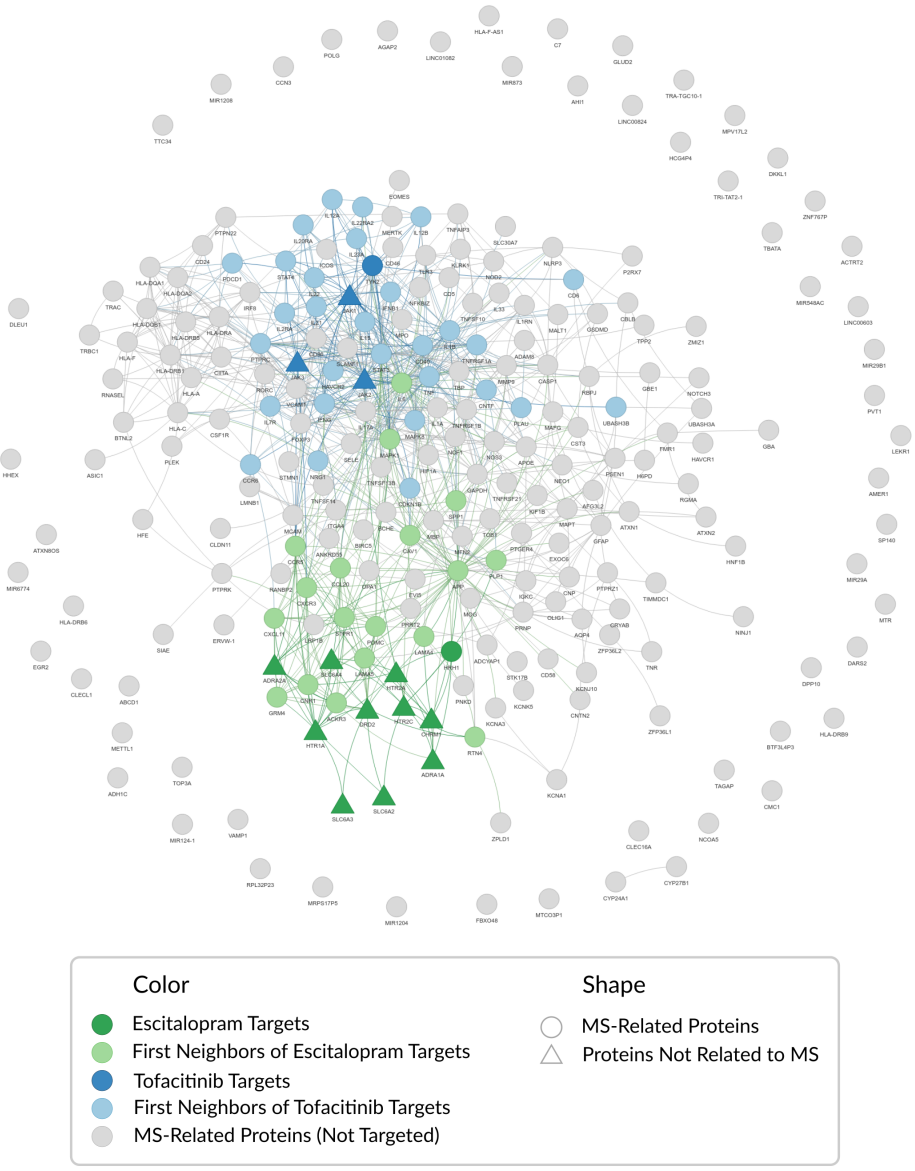


Figure 7.6. Escitalopram and Tofacitinib Complementary Exposure. The network displays the proteins associated to MS (circles) and highlights those targeted by escitalopram and tofacitinib (dark green and dark blue, respectively). Targets that are not related to MS are indicated as triangles. In order to better illustrate the influence of these two molecules given by the tight interconnection of the proteome, the first neighbors of the drug targets are depicted in a lighter color (light green for neighbors of targets of escitalopram, and light blue tofacitinib's ones).

protein-protein interactome construction, drug association to biological processes, cellular components, molecular functions and phenotypes, and drug induced gene expression profiles retrieval. Only sometimes, this issue could be partially mitigated by an extensive integration of data from a wider variety of databases. Noteworthy, puzzling examples could be the drug-target association and the availability of expression data in LINCS. The number of targets associated to a specific drug could considerably depend on the amount of research carried out on that medicine rather than on the actual biological interactions it has. This influences the drug-disease proximity evaluation. Additionally, as stated above, the LINCS database does not provide expression profiles for all the drugs selected by network proximity, limiting by far the choice space for drug repurposing.
Furthermore, if the knowledge we have about drugs is incomplete, the one we have on their combination is even sparser. This, obviously, affects our ability to screen and judge plausible associations.
Moreover, it could be argued that, even though the drug-disease proximity is evaluated with a rigorous geometrical approach, the choice of the distance threshold we use for discriminating drug efficacy is quite discretionary.

7.1.6 Conclusions

Here, we extended an unsupervised computational framework for drug repurposing with a network-based analysis for screening the possible drug combination therapies. Applying this pipeline to HD and MS, we identified several repurposable drug candidates, some of which have already been studied in humans. Eventually, we ended up with 138 potential drugs for HD and 38 for MS. Their plausible combinations are numerous, but this work can help to prioritize them. While these results are exploratory and should be experimentally verified before further consideration, they could provide valuable clues for improving the management of HD and MS.
Finally, this pipeline demonstrated to be effective on both investigated diseases, even though they have a different nature. For this reason, it could potentially provide new suggestions also for other complex disorders.

7.2 PATHOS and LOGOS

Here, it is presented the project that stands as the most complex and powerful, marking the culmination of this PhD course.
This study transcends conventional network models by embracing the vast semantic diversity of relationships among biomedical entities. Gathering vast datasets from diverse sources, spanning numerous entity types, we have integrated them to create the comprehensive knowledge graph named PATHOS (PATHologies of HOmo Sapiens). PATHOS serves as the foundation for our knowledge-driven predictions. Subsequently, we implemented LOGOS (Learning Optimized Graph-based representations of Object Semantics), a knowledge graph embedding model capable of generating predictions relevant to drug research.
The choice of the name PATHOS pays homage to ancient Greek culture. In this context, "pathos" ($\pi\acute{\alpha}\theta o\varsigma$) relates to emotions, sufferings and experiences. In the philosophical tradition, "logos" ($\lambda\acute{o}\gamma o\varsigma$) represents rationality and explanation. This aligns seamlessly with the knowledge graph embedding model's function of providing reasoned predictions based on the structured knowledge contained within the PATHOS knowledge graph. In this pairing, LOGOS complements PATHOS by offering rational insight into the depths of human sufferings.
PATHOS and LOGOS demonstrated their potential in three paradigmatic case studies (with a focus on neurological diseases): drug repurposing for Alzheimer's disease, phenotype selection for Huntington's disease, and the identification of proteins linked to multiple sclerosis.

Details

Authors Luca Menestrina, Maurizio Recanatini

Type Research Article

Status In Preparation

Data Availability The whole generated data is publicly available from the GitHub repository https://github.com/LucaMenestrina/PATHOS_LOGOS, as well as the full code for the collection, building and analysis. A detailed reference of the source data is provided in the file "`data/sources/sources.json`" of

the aforementioned repository (for every database are reported: name, version, license, employed files, URL and date of access).

Knowledge Graph and Machine Learning Help the Research of Drugs Aimed at Neurological Diseases

7.2.1 Introduction

Traditional network models have been widely used to depict intricate interactions within biomedical systems. Although these models have demonstrated impressive capabilities, they often face challenges in capturing the semantic complexity inherent in the diverse relationships among biomedical entities. In response to this limitation, recent approaches have embraced the use of multi-relational networks, such as knowledge graphs (KG).[96]. KGs incorporate insights from expert-curated sources into a structured graph format, where nodes correspond to biomedical entities, and edges denote the connections between them[16].
There are a lot of different heterogeneous biomedical pharmacological databases representable by KGs, each specializing in a specific domain. These diverse datasets serve a crucial role in progressing biomedical research, education, and diagnostic advancements. Researchers leverage these resources for a multitude of applications, including drug repurposing, as exemplified by Hetionet[97], ParmaKG[96], and PharMeBINet[310].

KGs are employed in various sectors, driving extensive research into the extraction of large-scale information from diverse sources. Nevertheless, it is widely acknowledged that even the most sophisticated KGs exhibit incompleteness or imperfections.[61, 62] Consequently, researchers have explored diverse techniques to correct inaccuracies and supplement missing information within KGs, often referred to as Knowledge Graph Completion or Knowledge Graph Augmentation. The expansion of KGs may involve extracting new facts from external sources, generating new facts through experimentation, or inferring missing facts based on the existing KG.[19]
This latter approach, known as Link Prediction (LP), has emerged as a thriving research domain, notably benefiting from the advancements in machine learning and deep learning techniques. The majority of LP models harness KG components to acquire low-dimensional representations, commonly referred to as Knowledge Graph Embeddings (KGE), which are then employed for the inference of new facts.[19]
Unlike traditional machine learning algorithms that rely solely on feature vectors,

Knowledge Graph Embedding Models (KGEMs) integrate an object's relationships into its representation. KGEMs predict new facts (triples) by leveraging the inherent information within existing facts.[63]

As explained in Chapter 2, KGEMs generally necessitate the parameterization of all nodes $n \in N$ (entities) and edge types $r \in R$ (relations).[63] Assuming vector embeddings, shallow encoders map these sets to d-dimensional vectors $f_n : n \to \mathbb{R}^d$ and $f_r : r \to \mathbb{R}^d$. Importantly, these encoders scale linearly with respect to the number of entities and relations, resulting in an entity embedding matrix with $O(|N|)$ space complexity.*

This strategy can be effective when applied to small, standard benchmark datasets such as Freebase[311] with approximately 15,000 nodes and WordNet[312] with around 40,000 nodes. However, training on more extensive graphs, such as YAGO 3-10[313] featuring 120,000 nodes or WikiKG2[314] with approximately 2.5 million nodes, presents significant computational challenges.

In a parallel with Natural Language Processing (NLP), shallow node encoding in KGs resembles shallow word embedding, which learns a vocabulary of the most frequent words while treating rarer ones as out-of-vocabulary (OOV). In NLP, the OOV issue has been addressed by enabling the creation of infinite combinations with a finite vocabulary, thanks to subword units. Inspired by this, similar strategies were explored for tokenizing entities within large knowledge graphs ($G = (\mathcal{N}, \mathcal{E}, \mathcal{R})$ constituted of $|\mathcal{N}|$ nodes, $|\mathcal{E}|$ edges, and $|\mathcal{R}|$ relation types), resulting in a substantial reduction in parameter complexity, improved generalization, and the natural representation of new, previously unseen entities using a fixed vocabulary. To achieve this, tokenization relies on atomic units analogous to subword units, rather than encompassing the entire set of nodes.

In pursuit of this goal, NodePiece, introduced by Galkin et al.[315], presents an anchor-based method for learning a fixed-size vocabulary $\mathcal{V}$ ($|\mathcal{V}| \ll |\mathcal{N}|$) applicable to any connected multi-relational graph. A selected subset of nodes (called anchors, $a \in \mathcal{A},\ \mathcal{A} \subset \mathcal{N}$) along with all relation types, constitute the set of atoms, which enables the representation of all possible nodes, with the construction of a combinatorial array of sequences from a limited atom vocabulary ($\mathcal{V} = \mathcal{A} \cup \mathcal{R}$). In contrast to

* The emphasis is typically placed on entities since the number of relations $|R|$ is usually orders of magnitude smaller than $|E|$.

shallow methods, each node n undergoes tokenization*, resulting in a unique $hash(n)^{\dagger}$ formed by k closest anchors, $z_a{}^{\ddagger}$ discrete anchor distances and m immediate relations (Figure 7.7). A crucial component for constructing a node embedding is the encoder function $enc(n) : hash(n) \rightarrow \mathbb{R}^d$, which converts the result of the tokenization of a node to its embedding, projecting it from $\mathbb{R}^{(k+m)\times d}$ to $\mathbb{R}^d$.

$$hash(n) = \left[\{a_i\}^k, \{z_{a_i}\}^k, \{r_j\}^m\right] \tag{7.4}$$

Taking advantage of this method, the overall parameter allocation is now reduced to a

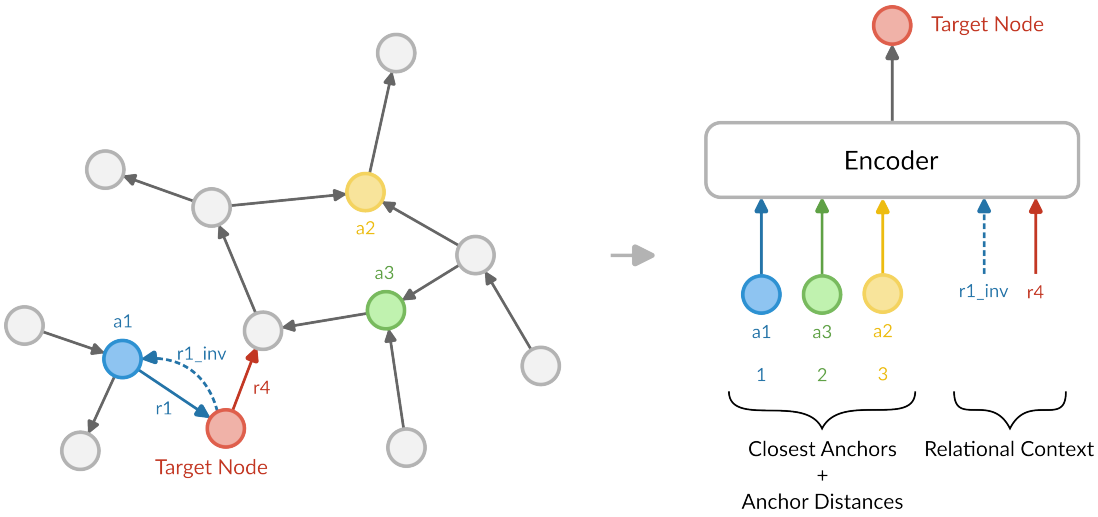


Figure 7.7. NodePiece Tokenization. Using three anchor points, a1, a2, and a3, a target node is tokenized, resulting in a hash that includes the top-k nearest anchors, their respective distances to the target node, and the relational context of outgoing relations from the target node. This sequence of hashed information is encoded, generating a distinctive embedding. The addition of inverse relations ensures network connectivity. Adapted from Galkin et al.[315]

small fixed-size atom vocabulary and the encoder function's complexity ($(k+m) \times d$ instead of $|\mathcal{N}| \times d + |\mathcal{R}| \times d$).

Galkin et al.[315] demonstrated that employing a fixed-size NodePiece vocabulary combined with a simple encoder still leads to competitive outcomes across various

* Representation of an element with a sequence of other entities called tokens. In NLP, a token is a sequence of characters in some particular document that are grouped together as a useful semantic unit for processing.

† A $hash$ function is a mathematical function that takes data of variable sizes as input and produces a fixed-size output.

‡ Given a graph G and a node n, the anchor distance $z_{a_i} \in [0, diameter(G)]$ with the anchor node a_i is defined as the shortest path distance between a_i and n in the graph G.

tasks, encompassing link prediction, node classification, and relation prediction. Additionally, the use of anchor-based hashing enables to operate effectively in both inductive settings and out-of-sample scenarios, accommodating unseen entities during the inference phase.

In this work, we created a comprehensive KG that we named PATHOS (PATHologies of HOmo Sapiens) collecting and integrating data on relevant biological entities from 24 distinct databases. Moreover, we developed LOGOS (Learning Optimized Graph-based representations of Object Semantics), a KGEM capable of providing predictions based on PATHOS.
To evaluate the capabilities of LOGOS, we carried out three crucial case studies: drug repurposing for Alzheimer's disease (AD), phenotype selection for Huntington's disease (HD), and the identification of proteins related to multiple sclerosis (MS). These studies showcase the potential of our integrated KG and KGEM to address pressing issues in the field of drug research.
Neurological disorders, like AD, HD, and MS, have profound implications for human health and well-being, resulting in significant suffering and disability. Yet, despite decades of research, effective treatments for these conditions remain elusive. Therefore, our work not only contributes to advancing scientific understanding but also holds the promise of assisting in the development of innovative solutions for these and other diseases.

7.2.2 Methods

7.2.2.1 KG Construction (PATHOS)

The analysis of biological systems, diseases, and therapies necessitates the integration of data from various sources. This integration poses unique challenges, principally consisting by the diverse formats and conflicting identifiers employed by each data source. Our approach to addressing these challenges is detailed in the subsequent sections on Data Collection and Data Integration.

Data Collection PATHOS draws upon a collection of data from 24 public databases (see Table 7.1 and Table C.2) renowned for their high-quality structured information pertaining to relevant biological entities. Our data collection exclusively focused on

Homo sapiens.

The source data, stored in a variety of formats, required the development of unique parsers for each data source to enable their transformation into a standardized file format suitable for integration.

Data Integration After standardizing the data, we conducted a merging process, during which duplicate entries were eliminated to prevent redundancy, while maintaining data integrity. This consistency was maintained by mapping all listed entities to official identifiers, including NCBI Entrez ID, MONDO, DrugBank, and more (For the complete reference, please see Table 7.1).
As a result of this integration effort, we generated a TSV (Tab-separated values) file having eight columns (subject, relation, object, subjectName, objectName, subjectType, objectType, source) and 4,487,349 rows.

7.2.2.2 KG Embedding Model (LOGOS)

We utilized PyKEEN (Python KnowlEdge EmbeddiNgs, version 1.10.1), an open-source Python package designed for KGEs.[327] This tool enables the construction of KGEMs by offering a diverse range of interaction models, training methodologies, loss functions, and the capacity to explicitly model inverse relations.
Indeed, within PyKEEN, a KGEM is constructed as a composite structure with four key components:

1. **Interaction Model**: we chose NodePiece for its reduced memory footprint and its capability of handling out-of-sample scenarios, as explained above).[315]
2. **Loss Function**: we applied the self-adversarial negative sampling (NSSA) loss function as proposed by Sun et al.[328], since negative sampling has proven quite effective for KGE[329].
3. **Training Approach**: in training under the open world assumption (OWA), triples that are not included in the KG are treated as unknown, resulting in over-generalization and poor model performance.[63] Instead, we opted for the stochastic local closed world assumption (sLCWA), which considers a random subset of head and tail generation strategies from LCWA as negative triples. This choice offers several advantages, including reduced computational load,

	Database	Content	ID Format
1	NCBI[316]	Protein	NCBI
2	APID[283]	Protein	NCBI
3	BioGRID[284]	Protein	NCBI
4	HuRI[285]	Protein	NCBI
5	InnateDB[286]	Protein	NCBI
6	INstruct[287]	Protein	NCBI
7	IntAct[288]	Protein	NCBI
8	SignaLink[289]	Protein	NCBI
9	STRING[140]	Protein	NCBI
10	HPRD[317]	Protein	NCBI
11	PINA[318]	Protein	NCBI
12	UniProt[319]	Protein	NCBI
13	HGNC[320]	Protein	NCBI
14	PRO[321]	Protein	NCBI
15	DisGeNET[141]	Protein, Disease	NCBI, MONDO
16	MONDO[322]	Disease	MONDO
17	DISEASES[276]	Disease	MONDO
18	Bgee[323]	Protein, Anatomical Entities	NCBI, Various from Uberon
19	Uberon[324]	Anatomy	Various from Uberon
20	PathwayCommons[325]	Pathway	Various from PC
21	HPO[278]	Phenotype	Various from HPO
22	GO[277]	Molecular Function, Biological Process, Cellular Location	GO
23	DrugBank[114]	Drug	DrugBank
24	DrugCentral[326]	Drug	DrugBank

Table 7.1. PATHOS Sources. List of all the databases employed for the construction of the KG PATHOS. The main type of content and the relative ID format are also reported. For a list of all the individual files see Table C.2.

minimal updates to embeddings, and flexibility for integrating new negative sampling strategies.

4. **Inverse Relations**: we included inverse triples because they are necessary for the employed version of NodePiece to maintain reachability of each node and balance in- and out-degrees.*

This extensive toolset within PyKEEN facilitated the development and fine-tuning of LOGOS, our KGEM, to meet the specific needs of our research.

7.2.2.3 NodePiece

To optimize the hyperparameters for the NodePiece model, we utilized the dedicated function offered by PyKEEN.
Among the most crucial parameters, we employed a 2-layer MLP aggregation encoder, consistent with the NodePiece original paper (ReLU activation and dropout rate of 0.1). The embedding dimension was set at 128, and we worked with 20 entity tokens and 5 relation tokens for each node. Our approach included a total of 10,000 anchors, 80% of them were the top degree nodes and 20% were randomly selected. Negative triples were generated corrupting positive triples with the Bernoulli method[330]. The training loop followed a stochastic local closed-world assumption, and the learning rate was set at 0.0001.
Further details and additional hyperparameters can be found in Appendix C.

Interaction Functions Numerous embedding models specific for knowledge graphs have been developed, primarily differing in the scoring of the plausibility of a given triple. In this section, we provide a brief overview of the interaction (scoring) functions explored in our study. These selected functions are well-established in the literature, encompass diverse methodologies, and have started to be investigated within the field of drug discovery.[331]

TransE TransE utilizes a straightforward vector summation in the latent space, where the head entity embedding is added to the relation embedding, bringing the

* For each relation $r \in \mathcal{R}$ a corresponding inverse relation r_{inv} is introduced. Consequently, the task of predicting the head entity of a (r, t)-pair (thus: $(?, r, t)$) becomes the task of predicting the tail entity of the corresponding inverse pair (t, r_{inv}) (that means: $(t, r_{inv}, ?)$).

result close to the tail embedding:[332]

$$f(h, r, t) = -||\boldsymbol{h} + \boldsymbol{r} - \boldsymbol{t}||_F \tag{7.5}$$

Where F is typically either the L1 or L2 norm.*
However, this method doesn't effectively capture one-to-many, many-to-one, and asymmetric relations in practical settings, as the embedding is accurate only when each entity and relation appears in just one fact.

DistMult DistMult employs a vector for each relation type, represented as a diagonal square matrix to significantly reduce the parameter count.[333] However, this means that it is constrained to model symmetric relations exclusively. Its scoring function is:

$$f(h, r, t) = \boldsymbol{h}^\top diag(\boldsymbol{r})\boldsymbol{t} \tag{7.6}$$

ComplEx In ComplEx, the entity and relation embeddings are complex valued ($\boldsymbol{h}, \boldsymbol{r}, \boldsymbol{t} \in \mathbb{C}^K$, K being the dimensionality of the embeddings).[329] Its scoring function becomes:

$$f(h, r, t) = Re(\boldsymbol{h} \otimes \boldsymbol{r} \otimes \boldsymbol{t}) \tag{7.7}$$

Where $Re()$ takes only the real value from the complex number, and $\otimes$ is the standard componentwise multi-linear dot product.†[334]
This allows ComplEx to handle both symmetric and asymmetric relations effectively.

RotatE Integrating concepts from various existing models, RotatE employs complex valued embeddings for entities and relations. RotatE forces the modulus of the relation vector to be 1 ($\forall i \ |r_i| = 1$), and operates in a way that the relation rotates the head to tail entities.[328]

$$f(h, r, t) = -||\boldsymbol{h} \odot \boldsymbol{r} - \boldsymbol{t}|| \tag{7.8}$$

* A norm ($||x||$), in mathematics, is a function mapping elements from a vector (x) to non-negative real numbers. In a certain way, it behaves like a distance from the origin. In general, a p-norm (Lp, with p a real number $p \geq 1$) is: $||x||_p := \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}$
Which means that a L1 norm (the taxicab or Manhattan distance) is: $||x||_1 = \sum_i |x_i|$, and the L2 (Euclidean distance) is: $||x||_2 = \sqrt{\sum_i x_i^2}$

† $(a \otimes b \otimes c) := \sum_k a_k b_k c_k$

Where $\odot$ is the Hadamard product.
This design enables RotatE to handle various relation types, including symmetric, asymmetric, inversion, and composition relations.

7.2.2.4 Learning Process

The learning process of LOGOS, as a supervised model, begins with the generation of embeddings using a Multilayer Perceptron (MLP). This MLP encodes a hash of tokenized sequences of anchors for each node in the knowledge graph (see Section 7.2.1). Subsequently, the model scores both positive and negative triples within the training dataset, with the aim of ranking true triples higher than false triples. Parameters are updated iteratively to minimize the loss during this learning process, and its performance is evaluated on the validation set.

7.2.2.5 Training Strategy

To ensure robust model performance assessment, we employed a random split of the initial dataset of triples into three distinct subsets: a training set (80% of the data), a validation set (10% of the data), and a test set (10% of the data). This stratified partitioning allowed us to perform rigorous hyperparameter optimization using the training and validation sets. In the end, the model's performance was evaluated on the test set, which it had never encountered during training.
To account for stochasticity[335], we retrained each of the models five times and evaluated their performance on the same test set. The model that on average exhibited the best performance was employed for the subsequent case studies (hyperparameters in Appendix C).

7.2.2.6 Transductive Link Prediction

The task is that of link prediction, which involves the completion of triples of interest either inserting the missing head or tail. Since no new entities are introduced in the triples (the set of nodes in the knowledge graph remains unchanged), this task is considered transductive.
During this phase, LOGOS leverages on the embeddings learned during the training phase to score the triples completed with all the possibile entities in the KG, construct-

ing a ranking based on the results.

7.2.2.7 Evaluation Protocol

In our evaluation process, we implemented a ranking procedure. To assess the model's performance, each validation or test triple underwent a corruption process, in which the head entity h_i was removed and replaced by each entity $e_i \in \mathcal{E} \backslash h_i$ in turn. Subsequently, we calculated a score for each triple, sorted these scores in ascending order to determine the rank of the correct triple (h, r, t), and then repeated this process for the tail entity as well. The overall model performance was evaluated based on the results of both head and tail corruption procedures (using the mean).
Our evaluation metrics included the mean reciprocal rank (MRR), its adjusted version (AMRR)[336], and the fraction of correct entities in the top-k rank positions (Hits@k) for various values of k: 1, 3, and 10, as well as the adjusted Hits@10.
Additionally, for downstream tasks, we incorporated another widely recognized metric: the areas under the receiver operating characteristic curve (AUC-ROC).
We followed the approach proposed by Bordes et al.[332] ensuring that all corrupted triples did not belong to the original dataset. Additionally, we considered the realistic ranking, for which the rank of an entity is calculated as the mean of the optimistic rank (where the entity is ranked first among those with equal scores) and the pessimistic rank (where the entity is ranked last among those with equal scores), following the method proposed by Berrendorf et al.[336]

7.2.2.8 Case Studies

In order to demonstrate the capabilities and versatility of PATHOS and LOGOS, we applied them to three distinct case studies in the field of neurological diseases: inferring novel drug candidates for repurposing, selecting plausible phenotypes, and identifying disease-related proteins.
For each case study, we asked LOGOS to complete a triple, either suggesting the subject or the object. In response, it generated a ranked list of entities from the knowledge graph, positioning the most promising candidates at the top. We then verified that the correct entity types were prioritized, assessed the results against the known triples available in PATHOS, and conducted a thorough literature search to find supporting evidence for the top-ranked entities.

Drug Repurposing for Alzheimer Disease We asked LOGOS to complete the triple *(?, indication, Alzheimer's disease)*. Subsequently, we conducted a literature search for evidence supporting the top 15 suggested drugs.

Huntington's Disease Phenotype Prediction For this case study, the triple to complete was: *(Huntington disease, has_phenotype, ?)*. From the top 50 selected phenotypes, we filtered out those present in the training or validation set, and ensured their consistency with known symptoms and Huntington's disease phenotypes.

Proteins Related to Multiple Sclerosis In this case study, we aimed to identify the subject of the triple *(?, related_to_disease, multiple sclerosis)*. To validate the correctness of the answer, we assessed the molecular functions and biological processes enriched in the first 100 proteins within the ranking.
For gene ontology enrichment analysis, we leveraged PANTHER (v 17.0)[337], a powerful and up-to-date tool that seamlessly integrates with the Gene Ontology (GO) website. This system is well-maintained with current GO annotations, ensuring reliable and comprehensive functional annotations. To assure robust results, only gene sets with false discovery rate (FDR) p-values below 0.05 were included in the analysis.

7.2.3 Results and Discussion

In this section, we will delve into the outcomes of our study, showcasing the twofold contributions of our work: the KG and the KGEM. First, we will provide a descriptive analysis of the knowledge graph, then we will present three compelling case studies, demonstrating the power of our model, focusing on neurological diseases.

Furthermore, we will present the results of our comparative analysis of interaction functions, providing a clear rationale for our selection of the best-performing one for the case studies.

7.2.3.1 Data Analysis

PATHOS, our comprehensive biomedical knowledge graph, represents a vast network of interconnected information, encompassing 174,367 entities categorized into 17 distinct types. These entities are linked by 4,487,349 relations, spanning 158 unique

types. Among the most prominent node types within PATHOS are proteins, biological processes, diseases, anatomical entities, molecular functions, phenotypes, and drugs (for a complete list check Table C.1). Particularly noteworthy are proteins, diseases and drugs, with 61,862 protein entities, 23,232 disease entities and 8,282 drug entities.

This multifaceted knowledge graph serves as the foundation for our data-driven investigations, enabling us to explore and extract valuable insights in the fields of drug repurposing and disease characterization.

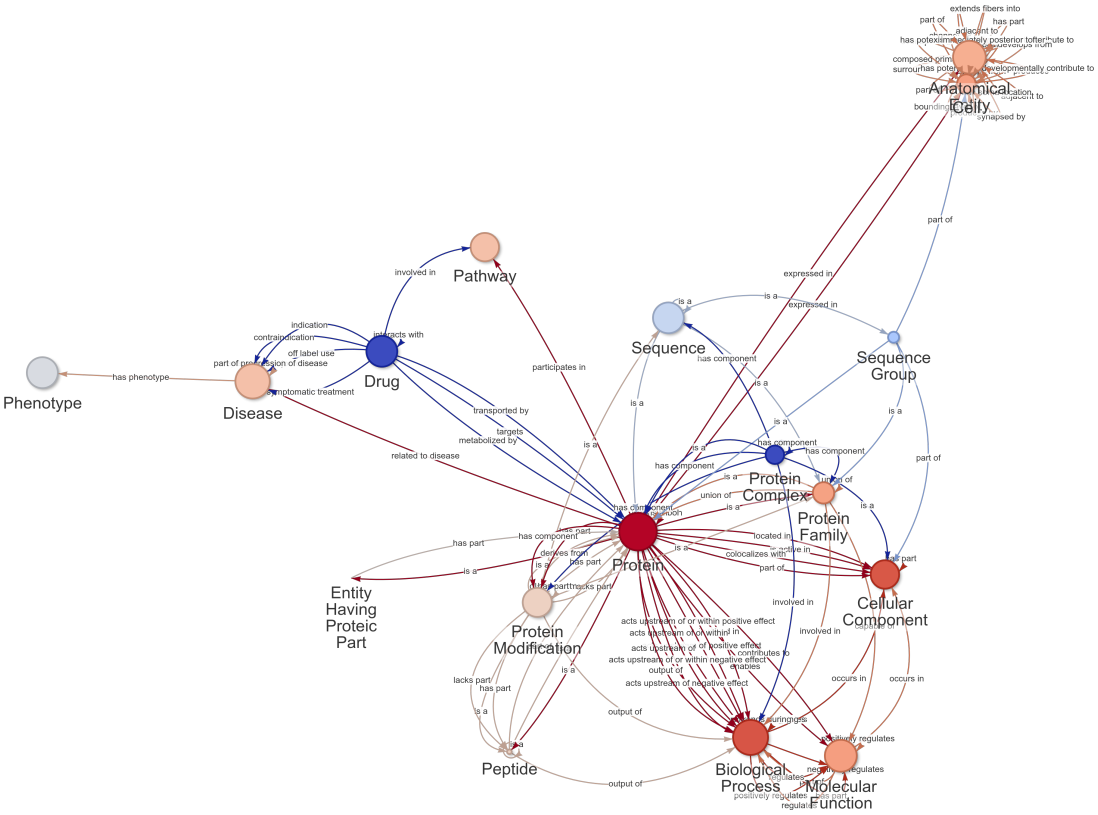


Figure 7.8. PATHOS Metagraph. A graphical representation of the relationships and connections within PATHOS

Figure 7.8 illustrates the metagraph, which is a graphical representation of the relationships and connections within PATHOS. Notably, proteins, as the central entity type with the most nodes and connections, play a pivotal role in shaping the structure of this knowledge graph.

7.2.3.2 Comparison of Interaction Functions

Table 7.2 provides an overview of the performances of several interaction functions, all the models were trained with the same hyperparameters.

	DistMult	**TransE**	**ComplEx**	**RotatE**
AMRR	0.139 ± 0.008	0.036 ± 0.004	$\mathbf{0.221 \pm 0.014}$	0.175 ± 0.008
Adjusted Hits@10	0.237 ± 0.006	0.074 ± 0.007	$\mathbf{0.369 \pm 0.010}$	0.301 ± 0.005

Table 7.2. Interaction Functions Comparison. Adjusted mean reciprocal rank and adjusted hits@10 scores for evaluating LOGOS with different interaction functions on the test set. Mean $\pm$ standard deviation of 5 replicas.

Surprisingly, the best-performing model, ComplEx, is not the most recent one, RotatE (which is second in line). In light of these results, we selected the ComplEx model with the best AMRR for our subsequent case study predictions.

7.2.3.3 Drug Repurposing for Alzheimer's Disease

Our evaluation of LOGOS's performance involved multiple steps. Initially, we verified its ability to prioritize drugs, which, in this case, is the correct entity type for completing the triple. Remarkably, out of the first 1,000 proposed entities, 997 were indeed drugs, demonstrating LOGOS's strong capability to prioritize them and achieving a ROC-AUC of 0.99.

Subsequently, we focused on specific predictions for Alzheimer's disease (AD) indications. For drugs already linked to AD in PATHOS (Epicriptine, Donepezil, Tacrine, Aducanumab, Galantamine, Ipidacrine, Rivastigmine, Acetylcarnitine), the ROC-AUC reached 0.94, considering only the drugs in the ranking (otherwise the ROC-AUC would have been of 1).

Table 7.3 shows the top 15 highest-scoring novel drug repurposing candidates for AD, including the canonical name of the drug, evidence category and PMID (literature reference supporting the interpretation). Among these top 15 predicted drug candidates, 6 drugs are validated for treating AD based on literature evidence, while 2 candidates exhibit a potential relationship with AD.

Some noteworthy examples include Daratumumab, Clomethiazole, and Fusidic acid. Daratumumab, an FDA-approved human antibody targeting CD38 for mul-

	DrugBankID	Drug Name	Evidence Category	PMID
1	DB01588	Prazepam		
2	DB06470	Clomethiazole	Neuroprotective and proposed	27129593[338], 24116891[339]
3	DB00541	Vincristine		
4	DB00234	Reboxetine	Neuroprotective in animal model	31297718[340]
5	DB09331	Daratumumab	In clinical trial	32144994[341], 33343293[342], 35300725[343], 35516416[344]
6	DB11089	Docusate		
7	DB00625	Efavirenz	Proposed	36581878[345]
8	DB02703	Fusidic acid	Related[a]	34537590[346]
9	DB00514	Dextromethorphan	Neuroprotective and potential protective effect against dementia	36113413[347]
10	DB06282	Levocetirizine		
11	DB03754	Tromethamine	Related[b]	8380642[348]
12	DB01551	Dihydrocodeine		
13	DB06654	Safinamide	Proposed	35001806[349]
14	DB06788	Histrelin		
15	DB08877	Ruxolitinib		

Table 7.3. Prioritized Drugs for AD. 15 highest-scoring drug repurposing candidates for Alzheimer's Disease. It reports: DrugBank ID, canonical name, evidence category and literature reverence.

[a] Aggregation inhibitor and disaggregator of silk fibroin, it was suggested as a treatment for protein aggregation disorders.

[b] Blocks amyloid beta channels, a mechanism that was proposed as a useful strategy for drug discovery for treatment of AD

tiple myeloma, is currently in a phase two clinical trial (NCT04070378) for mild to moderate AD due to its immunomodulatory effects on non-plasma cells and its potential to cross the blood-brain barrier.[343] Clomethiazole, an anticonvulsant with demonstrated neuroprotective properties, is considered a promising candidate

for future combination therapies addressing neuronal injury[338] and could serve as a lead compound for anti-neurodegenerative drug discovery[339]. Fusidic acid is of particular interest as it aligns with one of the key theories about Alzheimer's disease etiology: protein aggregation. Indeed, this compound has been suggested as a therapeutic approach for protein aggregation disorders.[346]

7.2.3.4 Huntington's Disease Phenotype Prediction

The top 5,967 entities (over 174,367 possible entities) in the ranking were phenotypes, reaffirming LOGOS's ability to determine the appropriate entity type for completing the triple.
Focusing solely on phenotypes, the ROC-AUC for those already cataloged in PATHOS reached 0.97.
Among the first 50 phenotypes (see Table C.3), 14 were part of training sets (with none in the validation nor test sets). Impressively, an additional 16 matched descriptions found in relevant literature references: dementia, tremor, muscle spasm, abnormality of extrapyramidal motor function, dysarthria, shuffling gait, ataxia, postural tremor, generalized-onset seizure, drooling[350, 351], abnormal autonomic nervous system physiology[352], abnormality of somatosensory evoked potentials[353, 354], muscle weakness[355]. This outcome not only validates LOGOS's capabilities, but can also suggest potential research avenues for clinical applications, aiming to enhance the anamnesis process.

7.2.3.5 Proteins Related to Multiple Sclerosis

Out of the entire pool of 174,367 entities, the model correctly prioritized proteins, with the top 17,479 entities in the ranking belonging to this category, underscoring LOGOS efficiency in identifying the appropriate entity type for the task.
For validating the gene ranking, we assessed the ROC-AUC for known ones in PATHOS, resulting in a high score of 0.97.
To provide a more detailed assessment, we examined the first 100 proteins (see list in Table C.4) in the ranking, carrying out a gene ontology (GO) enrichment analysis. This analysis revealed strong associations with biological processes (see Table C.5) related to MS: chronic inflammation regulation or response (chronic inflammatory response to antigenic stimulus, regulation of chronic inflammatory

response to antigenic stimulus), toll-like receptors (toll-like receptor TLR6:TLR2 signaling pathway), macrophage (positive regulation of cellular response to macrophage colony-stimulating factor stimulus), vitamin D (positive regulation of vitamin D biosynthetic process, positive regulation of calcidiol 1-monooxygenase activity). Concerning molecular functions related to MS (see Table C.6), the analysis identified relationships with: anandamide (anandamide 11,12 epoxidase activity, anandamide 8,9 epoxidase activity, anandamide 14,15 epoxidase activity)[356, 357], death receptor (death receptor agonist activity)[358], NAD (NAD+ nucleotidase, cyclic ADP-ribose generating, NAD(P)+ nucleosidase activity)[359, 360], MHC class II (MHC class II protein complex binding)[361].

7.2.4 Limitations

The project, while achieving its core objectives, is not without limitations. Some difficulties are common to many data science initiatives and relate primarily to data availability and model development. For instance, information regarding all biological entities remains unavailable, and the data (and consequently the relations) we do possess is notably skewed towards the most extensively studied entities. Additionally, incoherence of entity identifiers among source databases can result in errors during data integration and linkage. Moreover, the stability and maintenance of the source databases used for PATHOS are essential for keeping it up-to-date. Any disruptions or inconsistencies in these source databases may affect the quality of the KG.

While PATHOS is comprehensive, there may still be valuable data types, such as chemical and physical information, side effects or druggability, that are not included. The absence of certain entity and relation types may limit the range of insights and predictions that can be made.

On the model development side, LOGOS was thoughtfully constructed and optimized. However, like any model, there is room for improvement. Potential enhancements include exploring better encoding techniques, increasing embedding size, and optimizing the selection of anchor entities for even more accurate predictions.

In summary, while the project has achieved its primary objectives, it is important to recognize these limitations, and addressing them in the future could further improve the capabilities of PATHOS and LOGOS.

7.2.5 Conclusions

In conclusion, we gathered and integrated data on relevant biological entities from 24 distinct databases, creating the comprehensive knowledge graph PATHOS.
Additionally, we developed LOGOS, a knowledge graph embedding model, and evaluated its capabilities across three critical tasks: drug repurposing for Alzheimer's disease, phenotype selection for Huntington's disease, and the identification of proteins related to multiple sclerosis.
LOGOS succeeded in each of these tasks, demonstrating its potential to help drug research and foster innovative advancements in the field.

8 Data Analysis

8.1 Emergent Adverse Events in Single-pill Combinations

This project was conducted during a research period at Chemotargets in Barcelona, Spain. It was prompted by the Pharmacovigilance Risk Assessment Committee (PRAC) warning (issued on the 30th September 2022) regarding serious renal, gastrointestinal, and metabolic toxicities associated with the combination of ibuprofen and codeine.[†]

Leveraging the capabilities of the ClarityPV[‡] pharmacovigilance platform, which has been developed by Chemotargets to consolidate drug safety data throughout the entire lifetime of a drug (including safety pharmacology, preclinical toxicology, clinical safety, and postmarketing reports) the safety signals[§] related to the ibuprofen and codeine combination were examined. Prominent safety signals, namely renal tubular acidosis and hypokalaemia, align with the PRAC warning, and these signals were not associated with the individual drugs.

Encouraged by these initial results and recognizing the absence of a systematic exploration of safety signals for drug combinations versus individual drugs, this project embarked on a comprehensive analysis of the integrity of medicinal products and safety data within ClarityPV. The objective was to identify and rank all potential safety concerns specifically related to drug combinations.

Details

Authors Luca Menestrina, Ricard Garcia-Serna, Jordi Mestres

Type Research Article

Status In Preparation

[†] https://www.ema.europa.eu/en/news/meeting-highlights-pharmacovigilance-risk-assessment-committee-prac-26-29-september-2022

[‡] https://claritypv.com

[§] A safety signal is the information on a new or known adverse event that is potentially caused by a medicine and that warrants further investigation.

Detection of Safety Signals for Fixed-dose Drug Combinations

8.1.1 Introduction

The identification of adverse events associated with drug therapies is a critical aspect of pharmacovigilance, ensuring the safety and well-being of patients.[362] Fixed-dose combinations, which consist of two or more drugs combined into a single formulation, have gained popularity in medical practice due to their potential benefits such as improved adherence and simplified dosing regimens.[363] However, it is crucial to assess whether fixed-dose combinations pose any additional risks compared to their individual drug components.

In this study, we aimed to investigate adverse events that are more frequently observed in fixed-dose combinations compared to the individual drugs they contain. To achieve this, we leveraged ClarityPV[364], a comprehensive pharmacovigilance platform, and employed a computational algorithm specifically designed for this analysis. This algorithm facilitated the systematic evaluation of large-scale pharmacovigilance data, allowing us to identify and rank relevant triplets of drug combinations and associated adverse events overrepresented when compared to the individual drugs.

8.1.2 Methods

8.1.2.1 Data Collection

Data was collected from ClarityPV, a comprehensive pharmacovigilance platform developed by Chemotargets, integrating safety data across the entire lifetime of drugs, including information on safety pharmacology, preclinical toxicology, clinical safety, and postmarketing reports from various spontaneous reporting systems such as FAERS (FDA Adverse Event Reporting System)[365], VigiBase[366], JADER (Japanese Adverse Drug Event Report), and VAERS (Vaccine Adverse Event Reporting System)[367]. Data collection began by identifying all fixed-dose combinations (here referred to as "single-pill" combinations) available in the ClarityPV database, which represent specific drug formulations where multiple active ingredients are combined into a single dosage form. Then the associated adverse drug reactions

(ADRs) reported in patients who took these single-pill combinations were retrieved.
To ensure the reliability and quality of the data, reports with a high number of molecules or ADRs (those exceeding the 99th percentile of their respective distributions) were excluded, thereby eliminating outliers that could potentially bias the analysis.
Additionally, cases in which the indication was erroneously reported as an ADR were removed to eliminate potential data entry errors.

8.1.3 Metrics Computation

The Proportional Reporting Ratio (PRR) is a widely used pharmacovigilance metric that measures the strength of the association between a specific drug (or drug combination) and an adverse drug reaction.[368]
Given this contingency table:

Number of reports	**Analysed ADR**	**All other ADRs**
Analised drug	a	b
All other drugs	c	d

Table 8.1. PRR Contingency Table. The table shows the number of reports categorized by the presence or absence of the adverse drug reaction (ADR) and the drug in study.

We can define the PRR as:

$$PRR = \frac{\frac{a}{a+b}}{\frac{c}{c+d}} \tag{8.1}$$

PRRs were calculated for five different conditions: i) single-pill combinations (when the fixed-dose drug is reported with the brand name), ii) individual drugs within the combinations (also when only one of them is reported), iii) individual drugs co-subministered (both the components of the fixed-dose are reported together, irrespectively of the brand name), iv) exclusive combinations (when in the report appear only the drugs within the combination and no other substances), and v) exclusive combinations without single-pills (when in the report appear only the drugs within the combination, excluding the cases in which the brand name is reported).

In order to compare the PRR obtained for the different conditions, we computed the relative change:

$$relative\ change = \frac{PRR_{single\_pill} - \frac{\sum_{i=1}^{n} PRR_{drug_n}}{n}}{\frac{\sum_{i=1}^{n} PRR_{drug_n}}{n}} \tag{8.2}$$

This metric provides insights into the variation of PRR between the single-pill and the two individual drugs.

8.1.3.1 Filtering

To focus on the most relevant cases, the results were filtered based on a specific criterion, ensuring that only the most relevant side effects were considered. After analyzing the distribution of PRR values for each side effect across all drugs, all those where the single-pill calculated value was not among the higher 25% were discarded for that single-pill. With this, we kept only those side effects with high overreporting in comparison to the general drug population.

8.1.4 Results and Discussion

The analysis of the pharmacovigilance data contained in ClarityPV yielded a ranking of triplets, each representing a specific fixed-dose combination and its associated adverse events at the highest causality level.
In this section, we present the findings obtained from our computational algorithm, emphasizing its effectiveness in identifying relevant cases.

Within the dataset, a total of 43,208 triplets were identified, encompassing 551 single-pill fixed-dose combinations and 543 individual drugs, forming the basis for our analysis. The most frequently observed individual drugs included hydrochlorothiazide (3,706 instances), amlodipine (1,860 instances), paracetamol (2,493 instances), aspirin (1,635 instances), metformin (1,723 instances), and ethinylestradiol (1,648 instances).
These drugs are commonly used in clinical practice, with hydrochlorothiazide and amlodipine primarily prescribed for hypertension management, paracetamol and aspirin serving as analgesics, metformin being a widely used oral antidiabetic medication, and

ethinylestradiol being a component in many hormonal contraceptives.
Shifting our attention to ADRs, our analysis detected a total of 1,044 distinct ADRs associated with the analyzed triplets. Notably, the three most prevalent ADRs, namely drug reaction with eosinophilia and systemic symptoms syndrome (519 instances), hypersensitivity (494 instances), and anaphylactic reaction (485 instances), represent typical reactions to extraneous substances in the body. However, more serious ADRs, including neuroleptic malignant syndrome (406 instances), acute pancreatitis (388 instances), noninfectious encephalopathy/delirium (375 instances), and cardiomyopathy (373 instances), followed these more generic and classic reactions.

The ranking order for the identified triplets was determined based on the relative change, which served as a measure of the increase of adverse event occurrence for fixed-dose combinations compared to the baseline incidence for the corresponding individual drugs.
Considering a threshold of 1 for the relative change can be convenient as it directs the attention to cases where the adverse event occurrence for a drug combination is at least twice as high as the mean of that for the individual drugs, ensuring that associations with a more pronounced increase in risk are prioritized. By applying this criterion to filter the results, we detected 3402 triplets relating single-pill combinations and over-represented adverse events. These associations involved 828 adverse events for 378 single-pill combinations.

Notably, our findings align with previously reported cases documented in the literature (some relevant examples are presented in Table 8.2), providing validation for the accuracy of the algorithm in identifying known adverse events.

Before delving into the literature-backed triplets, it could be worth it to inspect the first entries of the ranking. In most of these initial cases, the lack of substantial medical evidence to demonstrate that the ADRs are overrepresented in the single-pill combination compared to the individual components can be reasonably explained.
For instance, the first most frequent single-pill combination in this set, which involves ciprofloxacin, dexamethasone, displays a relatively high incidence of ear-related ADRs (Ear infection, Ear swelling, Tympanic membrane perforation, Ear discomfort, Ear pain). However, upon closer examination, these ADRs might be better characterized as indications rather than side effects of the medication. A similar interpretation can be

	Drug1	Drug2	ADR	Drug1 PRR	Drug2 PRR	Single-pill PRR	Relative change
43	ibuprofen	codeine	Renal tubular acidosis	11.43	4.42	211.12	25.64
232	tinidazole	norfloxacin	Fixed eruption	58.86	33.92	414.83	7.95
305	tinidazole	norfloxacin	Drug eruption	15.52	6.80	86.63	6.76
363	ibuprofen	codeine	Metabolic acidosis	3.41	2.31	19.43	5.80
393	olanzapine	fluoxetine	Hyperlipidaemia	8.96	3.18	38.78	5.39
439	naproxen	esomeprazole	Renal failure	3.69	12.02	47.37	5.03
472	ibuprofen	codeine	Hypokalaemia	0.89	1.48	6.91	4.83
534	tinidazole	norfloxacin	Severe cutaneous adverse reactions (SMQ)	4.00	3.94	21.33	4.37
622	olanzapine	fluoxetine	Blood triglycerides increased	18.26	2.76	51.33	3.88
711	olanzapine	fluoxetine	Blood cholesterol increased	8.87	2.73	25.88	3.46
777	fluorouracil	folinate	Conjunctivitis	1.85	1.88	7.99	3.27
1040	fluorouracil	folinate	Conjunctival disorders (SMQ)	0.97	1.13	3.73	2.56
1174	naproxen	esomeprazole	Renal toxicity (DME)	2.32	7.88	16.82	2.30

Table 8.2. Relevant Emergent ADRs. The table shows some relevant examples of fixed-dose combinations presenting an adverse drug reaction that is overrepresented compared to the individual drugs.

applied to the third most common entry, which includes gentamicin and prednisolone, where Conjunctivitis (but also Keratitis and Eye pain) may be inherent to the drug's intended use.

Considering the fourth most frequent entry, a combination of aspirin and omeprazole, some ADRs such as Aspirin-exacerbated respiratory disease, Asthma/bronchospasm (smq), and Nasal polyps, are typical (they have a high PRR) also of aspirin use alone. On the other hand, there are also cases that could be worth a more thorough analysis. A relevant example is the combination of naloxone and hydromorphone, which ap-

pears to be strongly linked to gastrointestinal bleeding (Haematochezia, Gastrointestinal haemorrhage (smq), Ischaemic colitis (smq), Haemorrhage terms (excl laboratory terms) (smq)). This association is not as pronounced when considering the individual drugs separately.

Within the following results, several noteworthy findings have emerged. In particular, we have identified key cases that validate the accuracy of our approach and provide further insights into the safety profiles of fixed-dose combinations.
One of the key outcomes is the presence of cases involving ibuprofen and codeine, which were the focus of scrutiny by the European Medicines Agency's Pharmacovigilance Risk Assessment Committee (EMA PRAC) that motivated this project.*[369] It is interesting to notice that the case of these two drugs, associated with renal tubular acidosis, occupies the position 42 in the ranking, which is the highest among the cases that we found having evidence in the literature.
Furthermore, four other notable examples have emerged from our analysis.
The combination of olanzapine and fluoxetine was found being implicated in the worsening of total cholesterol and triglyceride levels.[370] This is reflected in the identified adverse drug reactions, such as increased blood triglycerides, increased blood cholesterol, and hyperlipidemia.
Another noteworthy finding involves the combination of fluorouracil and folinate, which has been associated with conjunctivitis and conjunctival disorders.[371]
Similarly, the combination of tinidazole and norfloxacin has shown an association with skin hyperpigmentation, drug eruption, and severe cutaneous adverse reactions.[372–374]
However, probably, the most significant and clinically relevant case we have identified is the association of naproxen and esomeprazole with renal failure. Notably, this combination has received attention from the EMA, leading to the issuance of a PRAC assessment report that states that the causal relationship between the esomeprazole/naproxen and tubulointerstitial nephritis (with possible progression to renal failure) is at least a reasonable possibility and instructs the amendment of the

* https://www.ema.europa.eu/en/documents/prac-recommendation/new-product-information-wording-extracts-prac-recommendations-signals-adopted-26-29-september-2022_en.pdf

information associated to products containing these drugs.*
These findings underscore the potential utility of the proposed computational algorithm in aiding the identification of emergent adverse events.

8.1.5 Limitations

Despite our best efforts, this study is not exempt from certain limitations commonly encountered in pharmacovigilance report analyses. The first one is about the quality of the reports analyzed, including factors such as completeness, order, and cleanness. Due to the reliance on spontaneously reported ADRs, the accuracy and consistency of the information within these reports may vary. Furthermore, this study did not consider the dosage of individual drugs, which could potentially influence the occurrence and severity of adverse events in drug combinations. Additionally, in the process of analyzing pharmacovigilance reports, the indication of a drug is sometimes erroneously reported as ADR. Another limitation is the inherent variability in the number of reports available for individual drugs or single-pill formulations, which could impact the robustness of the analysis. Finally, it is worth noting that this study focused on combinations of two molecules, so potential interactions involving more than two molecules were not considered. Despite these limitations, the findings of this study provide valuable insights into detecting emerging adverse events in drug combinations, but further research is warranted to address these limitations and enhance the comprehensiveness and accuracy of such analyses.

8.1.6 Conclusions

In conclusion, this study successfully developed and implemented a computational algorithm aimed at identifying adverse events specific to fixed-dose drug combinations that were highly overrepresented when compared to the individual drugs. The consistency between our findings and prior knowledge indicates the reliability and effectiveness of the computational approach employed in this investigation.

The first entries in the ranking challenge us to distinguish unique ADR profiles

* https://www.ema.europa.eu/en/documents/psusa/esomeprazole/naproxen-cmdh-scientific-conclusions-amendments-product-information-implementation-timetable-psusa/00001270/202204_en.pdf

in single-pill combinations versus individual components, underscoring the need for a meticulous analysis to accurately identify and understand ADRs in the context of combination therapies.
Despite this, at the top of our ranking, we detected 13 associations between single-pill combinations and overrepresented adverse events that are supported by the literature. These associations involved 13 adverse events for 5 single-pill combinations, including top-selling pairs like ibuprofen-codeine, that had already been addressed by regulatory agencies worldwide.

By uncovering specific combinations and associated adverse events, this research provides valuable insights that can inform future studies and facilitate informed clinical decision-making in the realm of drug safety and pharmacovigilance. Moreover, the methodology employed in this study holds promise as a foundation for the development of a tool capable of promptly raising alerts regarding disproportionately reported adverse drug reactions in fixed-dose combinations compared to their individual drug components, greatly enhancing pharmacovigilance efforts. These findings also emphasize the need for continued monitoring and evaluation of fixed-dose drug combinations as they are prescribed to the population.

With this work, we lay the groundwork for future investigations and highlight the necessity of collaborative efforts among researchers, healthcare professionals, and regulatory agencies to ensure the safe and effective use of fixed-dose drug combinations in clinical practice. Through comprehensive pharmacovigilance measures and collaborative research, we can improve patient safety and optimize the use of fixed-dose drug combinations, ultimately enhancing the quality of healthcare delivery.

Part III

Conclusions

9 Outcomes and Future Perspectives

In this thesis, applications of network theory and machine learning to drug research are presented, showcasing their multifaceted potentialities in advancing our understanding of complex biological systems. The research efforts were grounded in the pursuit of knowledge at the intersection of network theory, computational methods, and drug research. As this study comes to a conclusion, the key contributions, findings, and implications derived from this research are distilled.
The initial goal, as stated in chapter 4, was to leverage available biomedical knowledge to explore novel approaches supporting drug research. The studies have yielded practical solutions that address challenges in the field. These solutions have not only offered insights into the complexities inherent to biological systems but have also provided valuable tools to navigate and manipulate these intricacies.

More specifically, the outcomes drawn from each project are:

COVIDrugNet COVIDrugNet, a web application that allows users to capture a holistic view and keep up to date on how the clinical drug research has responded to the SARS-CoV-2 infection was developed.
Careful analyses of the COVID-19 drug-target system, based on COVIDrugNet, can help to understand the biological implications of the proposed drug options, and eventually improve the search for more effective therapies.

DEGA DEGA, a python package for differential gene expression analysis was proposed. It also includes the capability of identifying key regulatory genes (switch genes) which are likely to play a pivotal role in driving significant changes across various biological contexts. The application of this tool revealed critical regulatory genes with potential implications for understanding and treating glioblastoma.

Drug-induced Impulsivity Taking advantage of the FDA Adverse Events Reporting System, pivotal events driving the exacerbation of drug-induced impulse control disorders associated to aripiprazole and pramipexole were identified, primarily encompassing psychiatric, social and metabolic events.

Unsupervised Pipeline for Drug Repurposing An automated pipeline for identifying potential repurposable drugs and combinations was presented. It eventually collected a number of plausible opportunities to treat Huntington's disease and multiple sclerosis.

PATHOS and LOGOS A comprehensive KG about relevant biological data was built, and a GML model able to make predictions based on it was developed. This KGE model has demonstrated its versatility by succeeding in tasks such as drug repurposing for Alzheimer's disease, phenotype selection for Huntington's disease, and the identification of genes related to multiple sclerosis.

Emergent Adverse Events in Single-pill Combinations A computational algorithm aimed at identifying adverse events specific to fixed-dose drug combinations that are highly overrepresented when compared to the individual drugs was developed and implemented.

The study, including each of the projects within it, is not immune to certain limitations common in data analysis, regarding mainly data availability and quality.
The complete characterization of all biologically relevant entities remains a formidable challenge, and this limitation reverberates across various aspects of the study. Partial mitigation of this issue has been achieved through an extensive integration of data from a diverse array of databases. However, it is evident that more data is imperative, and while some may be obtained by organizing unstructured data into more structured machine-readable formats, a substantial portion will only be accessible through conducting further experiments and assays. Moreover, the research, like a significant portion of current endeavors, primarily centers on proteins, and it is essential to broaden the focus to include additional omics, with particular attention to RNA.
It's important to recognize that data in this field is in a perpetual state of evolution. Online resources, while beneficial for their capacity to facilitate updates, can pose limitations when they become unavailable, potentially impeding the execution of pipelines, either partially or entirely.

A pivotal aspect shaping the future of this research is the dynamic evolution of data in this field. The expansion of data sources by adding more data, incorporating new

data types, and correcting inaccuracies and mistakes is a continual process. Staying up-to-date and adapting the research methodologies accordingly is paramount for staying at the forefront of drug research.
Another critical aspect is the quality and standardization of data, ensuring them is an ongoing challenge. Collaborative efforts that focus on the curation and verification of data can lead to more uniform, complete and higher quality data.
In addition, fostering collaboration with other researchers and promoting interdisciplinary studies is essential. Collaboration and knowledge exchange with researchers from different fields would lead to a broader perspective, richer and more innovative insights.
Ultimately, in this game, open science and open data play a pivotal role, they will guide future research toward greater accuracy, comprehensiveness, and applicability.

In conclusion, this research underscores the value of data science and network-based methodologies in addressing the complex effects of drugs on biological systems. Looking ahead, and acknowledging the dynamic nature of scientific exploration, dedication remains to adapt to evolving data and promote open science and data sharing within the specialized domain. These efforts aim to foster a more profound comprehension of biological systems, ultimately facilitating the development of safer and more effective drugs. This, in turn, contributes to the advancement of the field of drug research.

Part IV

References

1. Gentili, P. L. *Untangling Complex Systems: a grand challenge for science* (CRC PRESS, 2019).

2. San Miguel, M. *et al.* Challenges in complex systems science. *The European Physical Journal Special Topics* **214,** 245–271. doi:10.1140/epjst/e2012-01694-y (2012).

3. Recanatini, M. & Menestrina, L. Network modeling helps to tackle the complexity of drug–disease systems. *WIREs Mechanisms of Disease* **15,** e1607. doi:10.1002/wsbm.1607 (2023).

4. Gaudelet, T. *et al.* Utilising Graph Machine Learning within Drug Discovery and Development. *Briefings in Bioinformatics* **2021,** 1–22. doi:10.1093/bib/bbab159 (2021).

5. Vespignani, A. Twenty years of network science. *Nature* **558,** 528–529. doi:10.1038/d41586-018-05444-y (2018).

6. Euler, L. Solutio Problematis ad geometriam situs pertinentis. *Commentarii Academiae Scientiarum Imperialis Petropolitana* **8,** 128–140 (1736).

7. Sylvester, J. J. On an Application of the New Atomic Theory to the Graphical Representation of the Invariants and Covariants of Binary Quantics, with Three Appendices. *American Journal of Mathematics* **1,** 64–104. doi:10.2307/2369436 (1878).

8. Caldarelli, G. & Catanzaro, M. *Networks: A Very Short Introduction* (Oxford University Press, 2012).

9. Caldarelli, G. *Scale-Free Networks: Complex Webs in Nature and Technology* doi:10.1093/acprof:oso/9780199211517.001.0001 (Oxford University Press, 2007).

10. Newman, M. E. J. *Networks: An Introduction* doi:10.1093/acprof:oso/9780199206650.001.0001 (Oxford University Press, 2010).

11. Trudeau, R. J. *Introduction to Graph Theory* (Dover Publications, 1976).

12. Chartrand, G. *Introductory Graph Theory* (Dover Publications, 1977).

13. Recanatini, M. & Cabrelle, C. Drug Research Meets Network Science: Where Are We? *Journal of Medicinal Chemistry* **63,** 8653–8666. doi:10.1021/acs.jmedchem.9b01989 (2020).

14. De Domenico, M. *et al.* Mathematical formulation of multilayer networks. *Physical Review X* **3,** 041022. doi:10.1103/physrevx.3.041022 (2014).

15. Hammoud, Z. & Kramer, F. Multilayer networks: aspects, implementations, and application in biomedicine. *Big Data Analytics* **5,** 2. doi:10.1186/s41044-020-00046-0 (2020).

16. Nicholson, D. N. & Greene, C. S. Constructing knowledge graphs and their biomedical applications. *Computational and Structural Biotechnology Journal* **18,** 1414–1428. doi:10.1016/J.csbj.2020.05.017 (2020).

17. Bonner, S. *et al.* A review of biomedical datasets relating to drug discovery: a knowledge graph perspective. *Briefings in Bioinformatics* **23,** bbac404. doi:10.1093/bib/bbac404 (2022).

18. Schlichtkrull, M. *et al.* Modeling Relational Data with Graph Convolutional Networks. *The Semantic Web. ESWC 2018. Lecture Notes in Computer Science* **10843,** 593–607. doi:10.1007/978-3-319-93417-4_38 (2018).

19. Rossi, A. *et al.* Knowledge graph embedding for link prediction: A comparative analysis. *ACM Transactions on Knowledge Discovery from Data* **15.** doi:10.1145/3424672 (2021).

20. Newman, M. E. J. in *The New Palgrave Dictionary of Economics* 1–8 (Palgrave Macmillan, 2008). doi:10.1057/978-1-349-95121-5_2565-1.

21. Newman, M. E. J. Assortative Mixing in Networks. *Physical Review Letters* **89,** 208701. doi:10.1103/physrevlett.89.208701 (2002).

22. Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature* **393,** 440–442. doi:10.1038/30918 (1998).

23. Boccaletti, S. *et al.* Complex networks: Structure and dynamics. *Physics Reports* **424,** 175–308. doi:10.1016/j.physrep.2005.10.009 (2006).

24. Newman, M. E. J. Fast algorithm for detecting community structure in networks. *Physical Review E* **69,** 066133. doi:10.1103/physreve.69.066133 (2004).

25. Menczer, F., Fortunato, S. & Davis, C. A. *A First Course in Network Science* doi:10.1017/9781108653947 (Cambridge University Press, 2020).

26. Newman, M. E. J. Detecting community structure in networks. *The European Physical Journal B* **38,** 321–330. doi:10.1140/epjb/e2004-00124-y (2004).

27. Von Luxburg, U. A tutorial on spectral clustering. *Statistics and Computing* **17,** 395–416. doi:10.1007/s11222-007-9033-z (2007).

28. MacQueen, J. *Some methods for classification and analysis of multivariate observations* in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* **5.1** (University of California Press, 1967), 281–298.

29. Ester, M. *et al. A density-based algorithm for discovering clusters in large spatial databases with noise* in *KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (1996), 226–231. doi:10.5555/3001460.3001507.

30. Girvan, M. & Newman, M. E. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America* **99,** 7821–7826. doi:10.1073/pnas.122653799 (2002).

31. Clauset, A., Newman, M. E. J. & Moore, C. Finding community structure in very large networks. *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics* **70,** 066111. doi:10.1103/physreve.70.066111 (2004).

32. Nguyen, L. V. *et al.* Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* **2008,** P10008. doi:10.1088/1742-5468/2008/10/p10008 (2008).

33. Erdős, P. & Rényi, A. On Random Graphs I. *Publicationes Mathematicae* **6,** 290–297 (1958).

34. MilGram, S. The Small World Problem. *Psychology Today* **2,** 60–67 (1967).

35. Guare, J. *Six Degrees of Separation: A Play* (Vintage Books, 1990).

36. Newman, M. E. J. The Structure and Function of Complex Networks. *SIAM REVIEW* **45,** 167–256. doi:10.1137/s00361450342480 (2003).

37. Menestrina, L., Cabrelle, C. & Recanatini, M. COVIDrugNet: a network-based web tool to investigate the drugs currently in clinical trial to contrast COVID-19. *Scientific Reports* **11,** 19426. doi:10.1038/s41598-021-98812-0 (2021).

38. Barabási, A. L. & Albert, R. Emergence of Scaling in Random Networks. *Science* **286,** 509–512. doi:10.1126/science.286.5439.509 (1999).

39. Clauset, A., Shalizi, C. R. & Newman, M. E. J. Power-law distributions in empirical data. *SIAM Review* **51,** 661–703. doi:10.1137/070710111 (2009).

40. Pastor-Satorras, R. & Vespignani, A. Epidemic Spreading in Scale-Free Networks. *Physical Review Letters* **86,** 3200. doi:10.1103/PhysRevLett.86.3200 (2001).

41. Lü, L. & Zhou, T. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications* **390,** 1150–1170. doi:10.1016/j.physa.2010.11.027 (2011).

42. Liben-Nowell, D. & Kleinberg, J. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology* **58,** 1019–1031. doi:10.1002/asi.20591 (2007).

43. Csermely, P. *et al.* Structure and dynamics of molecular networks: A novel paradigm of drug discovery A comprehensive review. *Pharmacology and Therapeutics* **138,** 333–408. doi:10.1016/j.pharmthera.2013.01.016 (2013).

44. Abate, C., Decherchi, S. & Cavalli, A. Graph neural networks for conditional de novo drug design. *WIREs Computational Molecular Science* **13,** e1651. doi:10.1002/wcms.1651 (2023).

45. McCarthy, J. & Hayes, P. J. in *Machine Intelligence 4* 463–502 (Edinburgh University Press, 1969).

46. Yang, X. *et al.* Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery. *Chemical Reviews* **119,** 10520–10594. doi:10.1021/acs.chemrev.8b00728 (2019).

47. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* http://www.deeplearningbook.org (The MIT Press, 2016).

48. McCulloch, W. S. & Pitts, W. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics* **5,** 115–133. doi:10.1007/bf02478259 (1943).

49. Russell, S. & Norvig, P. *Artificial Intelligence: A Modern Approach* (Pearson, 2021).

50. Zhu, H. Big data and artificial intelligence modeling for drug discovery. *Annual Review of Pharmacology and Toxicology* **60,** 573–589. doi:10.1146/annurev-pharmtox-010919-023324 (2020).

51. Lecun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521,** 436–444. doi:10.1038/nature14539 (2015).

52. Raymond, J. L. & Medina, J. F. Computational Principles of Supervised Learning in the Cerebellum. *Annual Review of Neuroscience* **41,** 233–253. doi:10.1146/annurev-neuro-080317-061948 (2018).

53. Joachims, T. *Transductive inference for text classification using support vector machines* in *International Conference on Machine Learning* (1999), 200–209.

54. Vapnik, V. N. *Statistical learning theory* (John Wiley & Sons, Ltd, 1998).

55. Sutton, R. S. & Barto, A. G. *Reinforcement Learning, An Introduction* (The MIT Press, 2018).

56. Pan, S. J. & Yang, Q. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering* **22,** 1345–1359. doi:10.1109/tkde.2009.191 (2010).

57. Caruana, R. Multitask Learning. *Machine Learning* **28,** 41–75. doi:10.1023/a:1007379606734 (1997).

58. Voulodimos, A. *et al.* Deep Learning for Computer Vision: A Brief Review. *Computational Intelligence and Neuroscience* **2018,** 7068349. doi:10.1155/2018/7068349 (2018).

59. Young, T. *et al.* Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine* **13,** 55–75. doi:10.1109/mci.2018.2840738 (2018).

60. Chami, I. *et al.* Machine Learning on Graphs: A Model and Comprehensive Taxonomy. *Journal of Machine Learning Research* **23,** 1–64 (2022).

61. West, R. *et al. Knowledge base completion via search-based question answering* in *WWW 2014 - Proceedings of the 23rd International Conference on World Wide Web* (Association for Computing Machinery, 2014), 515–525. doi:10.1145/2566486.2568032.

62. Paulheim, H. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic Web* **8,** 489–508. doi:10.3233/sw-160218 (2017).

63. Nickel, M. *et al. A review of relational machine learning for knowledge graphs* in *Proceedings of the IEEE* **104** (2016), 11–33. doi:10.1109/jproc.2015.2483592.

64. Waddington, C. H. *The strategy of the genes* doi:10.4324/9781315765471 (Taylor & Francis Group, 1957).

65. Han, J. D. J. Understanding biological functions through molecular networks. *Cell Research* **18,** 224–237. doi:10.1038/cr.2008.16 (2008).

66. Redhu, N. & Thakur, Z. in *Bioinformatics: Methods and Applications* 381–407 (Academic Press, 2022). doi:10.1016/b978-0-323-89775-4.00024-9.

67. Barabási, A. L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics* **5,** 101–113. doi:10.1038/nrg1272 (2004).

68. Kaufmann, S. H. Paul Ehrlich: founder of chemotherapy. *Nature Reviews Drug Discovery* **7,** 373. doi:10.1038/nrd2582 (2008).

69. Hopkins, A. L. Network pharmacology: The next paradigm in drug discovery. *Nature Chemical Biology* **4,** 682–690. doi:10.1038/nchembio.118 (2008).

70. Loscalzo, J. in *Network Medicine: Complex Systems in Human Disease and Therapeutics* 137–152 (Harvard University Press, 2017). doi:10.4159/9780674545533-007.

71. Silverman, E. K. & Loscalzo, J. Developing New Drug Treatments in the Era of Network Medicine. *Clinical Pharmacology & Therapeutics* **93,** 26–28. doi:10.1038/clpt.2012.207 (2013).

72. Stumpf, M. P. *et al.* Estimating the size of the human interactome. *Proceedings of the National Academy of Sciences of the United States of America* **105,** 6959–6964. doi:10.1073/pnas.0708078105 (2008).

73. Tsai, C. J., Ma, B. & Nussinov, R. Protein-protein interaction networks: how can a hub protein bind so many different partners? *Trends in Biochemical Sciences* **34,** 594–600. doi:10.1016/j.tibs.2009.07.007 (2009).

74. Yu, H. *et al.* TopNet: a tool for comparing biological sub-networks, correlating protein properties with topological statistics. *Nucleic Acids Research* **32,** 328–337. doi:10.1093/nar/gkh164 (2004).

75. Palla, G. *et al.* Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **435,** 814–818. doi:10.1038/nature03607 (2005).

76. Goh, K. I. *et al.* The human disease network. *Proceedings of the National Academy of Sciences of the United States of America* **104,** 8685–8690. doi:10.1073/pnas.0701361104 (2007).

77. Barabási, A. L., Gulbahce, N. & Loscalzo, J. Network medicine: A network-based approach to human disease. *Nature Reviews Genetics* **12,** 56–68. doi:10.1038/nrg2918 (2011).

78. Feldman, I., Rzhetsky, A. & Vitkup, D. Network properties of genes harboring inherited disease mutations. *Proceedings of the National Academy of Sciences of the United States of America* **105,** 4323–4328. doi:10.1073/pnas.0701722105 (2008).

79. Di Paola, L. *et al.* Protein Contact Networks: An Emerging Paradigm in Chemistry. *Chemical Reviews* **113,** 1598–1613. doi:10.1021/cr3002356 (2013).

80. Vendruscolo, M. *et al.* Small-world view of the amino acids that play a key role in protein folding. *Physical Review E* **65,** 061910. doi:10.1103/physreve.65.061910 (2002).

81. Atilgan, A. R., Akan, P. & Baysal, C. Small-World Communication of Residues and Significance for Protein Dynamics. *Biophysical Journal* **86,** 85–91. doi:10.1016/s0006-3495(04)74086-2 (2004).

82. Greene, L. H. & Higman, V. A. Uncovering Network Systems Within Protein Structures. *Journal of Molecular Biology* **334,** 781–791. doi:10.1016/j.jmb.2003.08.061 (2003).

83. Del Sol, A. *et al.* Residues crucial for maintaining short paths in network communication mediate signaling in proteins. *Molecular Systems Biology* **2,** 2006.0019. doi:10.1038/msb4100063 (2006).

84. Yildirim, M. A. *et al.* Drug-target network. *Nature Biotechnology* **25,** 1119–1126. doi:10.1038/nbt1338 (2007).

85. Takarabe, M. *et al.* Network-based analysis and characterization of adverse drug-drug interactions. *Journal of Chemical Information and Modeling* **51,** 2977–2985. doi:10.1021/cI200367w (2011).

86. Dobson, C. M. Chemical space and biology. *Nature* **432,** 824–828. doi:10.1038/nature03192 (2004).

87. Maggiora, G. M. & Bajorath, J. Chemical space networks: A powerful new paradigm for the description of chemical space. *Journal of Computer-Aided Molecular Design* **28,** 795–802. doi:10.1007/s10822-014-9760-0 (2014).

88. Bellman, R. E. *Adaptive Control Processes: A Guided Tour* (Princeton University Press, 1961).

89. Lowe, D. Chemical space is big. Really big. *MedChemComm* **6,** 12–12. doi:10.1039/c4md90045f (2015).

90. Scalfani, V. F., Patel, V. D. & Fernandez, A. M. Visualizing chemical space networks with RDKit and NetworkX. *Journal of Cheminformatics* **14,** 1–13. doi:10.1186/S13321-022-00664-x (2022).

91. Zwierzyna, M. *et al.* Design and characterization of chemical space networks for different compound data sets. *Journal of Computer-Aided Molecular Design* **29,** 113–125. doi:10.1007/s10822-014-9821-4 (2015).

92. Zhang, B. *et al.* Comparison of bioactive chemical space networks generated using substructure- and fingerprint-based measures of molecular similarity. *Journal of Computer-Aided Molecular Design* **29,** 595–608. doi:10.1007/s10822-015-9852-5 (2015).

93. Wu, M. *et al.* Design of chemical space networks on the basis of Tversky similarity. *Journal of Computer-Aided Molecular Design* **30,** 1–12. doi:10.1007/s10822-015-9891-y (2016).

94. Kenny, P. W. & Sadowski, J. in *Chemoinformatics in Drug Discovery* 271–285 (John Wiley & Sons, Ltd, 2005). doi:10.1002/3527603743.ch11.

95. Vogt, M. *et al.* Lessons learned from the design of chemical space networks and opportunities for new applications. *Journal of Computer-Aided Molecular Design* **30,** 191–208. doi:10.1007/s10822-016-9906-3 (2016).

96. Zheng, S. *et al.* PharmKG: a dedicated knowledge graph benchmark for bomedical data mining. *Briefings in Bioinformatics,* 1–15. doi:10.1093/bib/bbaa344 (2020).

97. Himmelstein, D. S. *et al.* Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife* **6,** e26726. doi:10.7554/eLife.26726 (2017).

98. Hughes, J. P. *et al.* Principles of early drug discovery. *British Journal of Pharmacology* **162,** 1239–1249. doi:10.1111/j.1476-5381.2010.01127.x (2011).

99. Bleicher, K. H. *et al.* Hit and lead generation: beyond high-throughput screening. *Nature Reviews Drug Discovery 2003 2:5* **2,** 369–378. doi:10.1038/nrd1086 (2003).

100. Yang, K. *et al.* Analyzing Learned Molecular Representations for Property Prediction. *Journal of Chemical Information and Modeling* **59,** 3370–3388. doi:10.1021/acs.jcim.9b00237 (2019).

101. Xue, D. *et al.* Advances and challenges in deep generative models for de novo molecule generation. *WIREs Computational Molecular Science* **9,** e1395. doi:10.1002/wcms.1395 (2019).

102. Weininger, D. SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules. *Journal of Chemical Information and Computer Sciences* **28,** 31–36. doi:10.1021/cI00057a005 (1988).

103. Cheng, F., Kovács, I. A. & Barabási, A. L. Network-based prediction of drug combinations. *Nature Communications* **10,** 1197. doi:10.1038/s41467-019-09186-x (2019).

104. Chen, Q., Allot, A. & Lu, Z. Keep up with the latest coronavirus research. *Nature* **579,** 193. doi:10.1038/d41586-020-00694-1 (2020).

105. Ahamad, S. *et al.* Primed for global coronavirus pandemic: Emerging research and clinical outcome. *European Journal of Medicinal Chemistry* **209,** 112862. doi:10.1016/j.ejmech.2020.112862 (2021).

106. Chai, K. L. *et al.* Convalescent plasma or hyperimmune immunoglobulin for people with COVID-19: a living systematic review. *Cochrane Database of Systematic Reviews* **2020,** CD013600. doi:10.1002/14651858.cd013600.pub3 (2020).

107. Pushpakom, S. *et al.* Drug repurposing: Progress, challenges and recommendations. *Nature Reviews Drug Discovery* **18,** 41–58. doi:10.1038/nrd.2018.168 (2018).

108. Xu, J. *et al.* Drug repurposing approach to combating coronavirus: Potential drugs and drug targets. *Medicinal Research Reviews* **41,** 1375–1426. doi:10.1002/med.21763 (2021).

109. Ng, Y. L., Salim, C. K. & Chu, J. J. H. Drug repurposing for COVID-19: Approaches, challenges and promising candidates. *Pharmacology and Therapeutics* **228,** 107930. doi:10.1016/j.pharmthera.2021.107930 (2021).

110. Alexander, S. P. *et al.* A rational roadmap for SARS-CoV-2/COVID-19 pharmacotherapeutic research and development: IUPHAR Review 29. *British Journal of Pharmacology* **177,** 4942–4966. doi:10.1111/bph.15094 (2020).

111. Mucke, H. COVID-19 and the drug repurposing tsunami. *Assay and Drug Development Technologies* **18,** 211–214. doi:10.1089/adt.2020.996 (2020).

112. Gaulton, A. *et al.* The ChEMBL database in 2017. *Nucleic Acids Research* **45,** D945–D954. doi:10.1093/nar/gkw1074 (2017).

113. Kim, S. *et al.* PubChem in 2021: new data content and improved web interfaces. *Nucleic acids research* **49,** D1388–D1395. doi:10.1093/nar/gkaa971 (2021).

114. Wishart, D. S. *et al.* DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Research* **46,** D1074–D1082. doi:10.1093/nar/gkx1037 (2018).

115. Martin, R. *et al.* CORDITE: The curated CORona Drug InTERactions database for SARS-CoV-2. *iScience* **23,** 101297. doi:10.1016/j.isci.2020.101297 (2020).

116. Kuleshov, M. V. *et al.* The COVID-19 drug and gene set library. *Patterns* **1,** 100090. doi:10.1016/j.patter.2020.100090 (2020).

117. Mercatelli, D. *et al.* coronapp: A web application to annotate and monitor SARS-CoV-2 mutations. *bioRxiv,* 2020.05.31.124966. doi:10.1101/2020.05.31.124966 (2020).

118. Loscalzo, J., Barabási, A.-L. & Silverman, E. K. *Network Medicine* (eds Loscalzo, J., Barabási, A.-L. & Silverman, E. K.) 414–458 (Harvard University Press, 2017).

119. Korn, D. *et al.* COVID-KOP: Integrating emerging COVID-19 data with the ROBOKOP database. *Bioinformatics* **37,** 586–587. doi:10.1093/bioinformatics/btaa718 (2021).

120. Sadegh, S. *et al.* Exploring the SARS-CoV-2 virus-host-drug interactome for drug repurposing. *Nature Communications* **11,** 1–9. doi:10.1038/s41467-020-17189-2 (2020).

121. Verstraete, N. *et al.* CovMulNet19, integrating proteins, diseases, drugs, and symptoms: A network medicine approach to COVID-19. *Network and Systems Medicine* **3,** 130–141. doi:10.1089/nsm.2020.0011 (2020).

122. Mitsopoulos, C. *et al.* Coronavirus canSAR - A data-driven, AI-enabled, drug discovery resource for the research community. *ChemRxiv* (2020).

123. Gysi, D. M. *et al.* Network medicine framework for identifying drug-repurposing opportunities for COVID-19. *Proceedings of the National Academy of Sciences of the United States of America* **118.** doi:10.1073/pnas.2025581118 (2021).

124. Gordon, D. E. *et al.* A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* **583,** 459–468. doi:10.1038/s41586-020-2286-9 (2020).

125. Zhou, Y. *et al.* Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. *Cell Discovery* **6,** 14. doi:10.1038/s41421-020-0153-3 (2020).

126. Pavlopoulos, G. A. *et al.* Bipartite graphs in systems biology and medicine: A survey of methods and applications. *GigaScience* **7,** 1–31. doi:10.1093/gigascience/giy014 (2018).

127. Hagberg, A. A., Schult, D. A. & Swart, P. J. *Exploring network structure, dynamics, and function using NetworkX* in *Proceedings of the 7th Python in Science Conference (SciPy)* (eds Varoquaux, G., Vaught, T. & Millman, J.) (2008), 11–15.

128. Broido, A. D. & Clauset, A. Scale-free networks are rare. *Nature Communications* **10,** 1–10. doi:10.1038/s41467-019-08746-5 (2019).

129. Albert, R., Jeong, H. & Barabási, A. L. Error and attack tolerance of complex networks. *Nature* **406,** 378–382. doi:10.1038/35019019 (2000).

130. Bohn, M. K. *et al.* Pathophysiology of COVID-19: Mechanisms underlying disease severity and progression. *Physiology* **35,** 288–301. doi:10.1152/physiol.00019.2020 (2020).

131. Gupta, A. *et al.* Extrapulmonary manifestations of COVID-19. *Nature Medicine* **26,** 1017–1032. doi:10.1038/s41591-020-0968-3 (2020).

132. Bassetti, M., Kollef, M. H. & Timsit, J. F. Bacterial and fungal superinfections in critically ill patients with COVID-19. *Intensive Care Medicine* **46,** 2071–2074. doi:10.1007/s00134-020-06219-8 (2020).

133. Beovic, B. *et al.* Antibiotic use in patients with COVID-19: A 'snapshot' Infectious Diseases International Research Initiative (ID-IRI) survey. *Journal of Antimicrobial Chemotherapy* **75,** 3386–3390. doi:10.1093/jac/dkaa326 (2020).

134. Hoffmann, M. *et al.* SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* **181,** 271–280. doi:10.1016/j.cell.2020.02.052 (2020).

135. Parmar, M. S. TMPRSS2: An Equally Important Protease as ACE2 in the Pathogenicity of SARS-CoV-2 Infection. *Mayo Clinic Proceedings.* doi:10.1016/j.mayocp.2021.07.005 (2021).

136. Yang, N. & Shen, H. M. Targeting the endocytic pathway and autophagy process as a novel therapeutic strategy in COVID-19. *International Journal of Biological Sciences* **16,** 1724–1731. doi:10.7150/ijbs.45498 (2020).

137. Chen, Z. *et al.* Comprehensive analysis of the host-virus interactome of SARS-CoV-2. *bioRxiv.* doi:10.1101/2020.12.31.424961 (2021).

138. Shannon, P. *et al.* Cytoscape: A software Environment for integrated models of biomolecular interaction networks. *Genome Research* **13,** 2498–2504. doi:10.1101/gr.1239303 (2003).

139. Avram, S. *et al.* DrugCentral 2021 supports drug discovery and repositioning. *Nucleic acids research* **49,** D1160–D1169. doi:10.1093/nar/gkaa997 (2021).

140. Szklarczyk, D. *et al.* STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research* **47,** D607–D613. doi:10.1093/nar/gky1131 (2019).

141. Piñero, J. *et al.* The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research* **48,** D845–D855. doi:10.1093/nar/gkz1021 (2020).

142. Bienert, S. *et al.* The SWISS-MODEL Repository-new features and functionality. *Nucleic Acids Research* **45,** D313–D319. doi:10.1093/nar/gkw1132 (2017).

143. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Research* **28,** 235–242. doi:10.1093/nar/28.1.235 (2000).

144. The UniProt Consortium. UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Research* **47,** D506–D515. doi:10.1093/nar/gky1049 (2019).

145. Santos, R. *et al.* A comprehensive map of molecular drug targets. *Nature Reviews Drug Discovery* **16,** 19–34. doi:10.1038/nrd.2016.230 (2017).

146. Zhang, J. X. *et al.* Identifying a set of influential spreaders in complex networks. *Scientific Reports* **6,** 1–10. doi:10.1038/srep27823 (2016).

147. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12,** 2825–2830 (2011).

148. Alstott, J., Bullmore, E. & Plenz, D. Powerlaw: A Python package for analysis of heavy-tailed distributions. *PLoS ONE* **9,** 85777. doi:10.1371/journal.pone.0085777 (2014).

149. Plotly Technologies Inc. *Collaborative data science* Montréal, QC, 2015. https://plot.ly.

150. The Apache Software Foundation. *Apache HTTP Server* 1995. http://httpd.apache.org.

151. Seyednasrollah, F., Laiho, A. & Elo, L. L. Comparison of software packages for detecting differential expression in RNA-seq studies. *Briefings in Bioinformatics* **16,** 59–70. doi:10.1093/bib/bbt086 (2015).

152. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140. doi:10.1093/bioinformatics/btp616 (2010).

153. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* **43,** e47. doi:10.1093/nar/gkv007 (2015).

154. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15,** 550. doi:10.1186/s13059-014-0550-8 (2014).

155. Ostrom, Q. T. *et al.* CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2013–2017. *Neuro-Oncology* **22,** iv1–iv96. doi:10.1093/neuonc/noaa200 (2020).

156. Alifieris, C. & Trafalis, D. T. Glioblastoma multiforme: Pathogenesis and treatment. *Pharmacology & Therapeutics* **152,** 63–82. doi:10.1016/j.pharmthera.2015.05.005 (2015).

157. Rominiyi, O. & Collis, S. J. DDRugging glioblastoma: understanding and targeting the DNA damage response to improve future therapies. *Molecular Oncology* **16,** 11–41. doi:10.1002/1878-0261.13020 (2022).

158. Paci, P. *et al.* SWIM: a computational tool to unveiling crucial nodes in complex biological networks. *Scientific Reports* **7,** 44797. doi:10.1038/srep44797 (2017).

159. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57,** 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x (1995).

160. Cook, R. D. Detection of Influential Observation in Linear Regression. *Technometrics* **19,** 18 (1977).

161. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455,** 1061–1068. doi:10.1038/nature07385 (2008).

162. The Cancer Genome Atlas Research Network *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics* **45,** 1113–1120. doi:10.1038/ng.2764 (2013).

163. Paci, P. & Fiscon, G. SWIMmeR: an R-based software to unveiling crucial nodes in complex biological networks. *Bioinformatics* **38,** 586–588. doi:10.1093/bioinformatics/btab657 (2022).

164. Pearson, K. VII. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London* **58,** 240–242. doi:10.1098/rspl.1895.0041 (1895).

165. Fiscon, G. *et al.* Gene network analysis using SWIM reveals interplay between the transcription factor-encoding genes HMGA1, FOXM1, and MYBL2 in triple-negative breast cancer. *FEBS Letters* **595,** 1569–1586. doi:10.1002/1873-3468.14085 (2021).

166. Bottomly, D. *et al.* Evaluating Gene Expression in C57BL/6J and DBA/2J Mouse Striatum Using RNA-Seq and Microarrays. *PLOS ONE* **6,** e17820. doi:10.1371/journal.pone.0017820 (2011).

167. Wang, K. *et al.* PTBP1 knockdown promotes neural differentiation of glioblastoma cells through UNC5B receptor. *Theranostics* **12,** 3847–3861. doi:10.7150/thno.71100 (2022).

168. Stegh, A. H. *et al.* Glioma oncoprotein Bcl2L12 inhibits the p53 tumor suppressor. *Genes & Development* **24,** 2194–2204. doi:10.1101/gad.1924710 (2010).

169. Kumthekar, P. *et al.* A phase 0 first-in-human study using NU-0129: A gold base spherical nucleic acid (SNA) nanoconjugate targeting BCL2L12 in recurrent glioblastoma patients. *Journal of Clinical Oncology* **37,** 3012–3012. doi:10.1200/jco.2019.37.15_suppl.3012 (2019).

170. Paulino, V. M. *et al.* TROY (TNFRSF19) is overexpressed in advanced glial tumors and promotes glioblastoma cell invasion via Pyk2-Rac1 signaling. *Molecular Cancer Research* **8,** 1558–1567. doi:10.1158/1541-7786.mcr-10-0334 (2010).

171. Ding, Z. *et al.* TROY signals through JAK1-STAT3 to promote glioblastoma cell migration and resistance. *Neoplasia* **22,** 352–364. doi:10.1016/j.neo.2020.06.005 (2020).

172. Liu, X. *et al.* TROY interacts with RKIP to promote glioma development. *Oncogene* **38,** 1544–1559. doi:10.1038/s41388-018-0503-x (2018).

173. Tu, Z. *et al.* Protein disulfide-isomerase A4 confers glioblastoma angiogenesis promotion capacity and resistance to anti-angiogenic therapy. *Journal of Experimental and Clinical Cancer Research* **42,** 77. doi:10.1186/s13046-023-02640-1 (2023).

174. Wang, M. *et al.* PDIA4 promotes glioblastoma progression via the PI3K/AKT/m-TOR pathway. *Biochemical and Biophysical Research Communications* **597,** 83–90. doi:10.1016/j.bbrc.2022.01.115 (2022).

175. Mir, S. E. *et al.* In Silico Analysis of Kinase Expression Identifies WEE1 as a Gatekeeper against Mitotic Catastrophe in Glioblastoma. *Cancer Cell* **18,** 244–257. doi:10.1016/j.ccr.2010.08.011 (2010).

176. Sanai, N. *et al.* Phase 0 trial of azd1775 in first-recurrence glioblastoma patients. *Clinical Cancer Research* **24,** 3820–3828. doi:10.1158/1078-0432.ccr-17-3348 (2018).

177. Engel, G. L. The Need for a New Medical Model: A Challenge for Biomedicine. *Science* **196,** 129–136. doi:10.1126/science.847460 (1977).

178. Kleinman, A., Eisenberg, L. & Good, B. Culture, illness, and care. Clinical lessons from anthropologic and cross-cultural research. *Annals of Internal Medicine* **88,** 251–258. doi:10.7326/0003-4819-88-2-251 (1978).

179. Young, A. The Anthropologies of Illness and Sickness. *Annual Review of Anthropology* **11,** 257–285. doi:10.1146/annurev.an.11.100182.001353 (1982).

180. Lorimer, S., Cox, A. & Langford, N. J. A patient's perspective: the impact of adverse drug reactions on patients and their views on reporting. *Journal of Clinical Pharmacy and Therapeutics* **37,** 148–152. doi:10.1111/j.1365-2710.2011.01258.x (2012).

181. Suh, D. C. *et al.* Clinical and Economic Impact of Adverse Drug Reactions in Hospitalized Patients. *Annals of Pharmacotherapy* **34,** 1373–1379. doi:10.1345/aph.10094 (2000).

182. Rolfes, L. *et al.* The Impact of Experiencing Adverse Drug Reactions on the Patient's Quality of Life: A Retrospective Cross-Sectional Study in the Netherlands. *Drug Safety* **39,** 769–776. doi:10.1007/s40264-016-0422-0 (2016).

183. Calvert, M. *et al.* Guidelines for Inclusion of Patient-Reported Outcomes in Clinical Trial Protocols. *The SPIRIT-PRO Extension* **319,** 483–494. doi:10.1001/jama.2017.21903 (2018).

184. Haag, S. *et al.* Results on patient-reported outcomes are underreported in summaries of product characteristics for new drugs. *Journal of Patient-Reported Outcomes* **5,** 127. doi:10.1186/s41687-021-00402-1 (2021).

185. Weintraub, D. Impulse control disorders in Parkinson's disease: A 20-year odyssey. *Movement Disorders* **34,** 447–452. doi:10.1002/mds.27668 (2019).

186. Fusaroli, M. *et al.* Impulsive conditions in Parkinson's disease: A pharmacosurveillance-supported list. *Parkinsonism & Related Disorders* **90,** 79–83. doi:10.1016/j.parkpeldis.2021.08.006 (2021).

187. Fusaroli, M. *et al.* Behavioral excess and disruptive conduct: A historical and taxonomic approach to the origin of the 'impulse control disorders' diagnostic construct. *Addiction.* doi:10.1111/add.16086 (2022).

188. Grall-Bronnec, M. *et al.* Dopamine Agonists and Impulse Control Disorders: A Complex Association. *Drug Safety* **41,** 19–75. doi:10.1007/s40264-017-0590-6 (2018).

189. Tippmann-Peikert, M. *et al.* Pathologic gambling in patients with restless legs syndrome treated with dopaminergic agonists. *Neurology* **68,** 301–303. doi:10.1212/01.wnl.0000252368.25106.b6 (2007).

190. Cornelius, J. *et al.* Impulse Control Disorders with the use of Dopaminergic Agents in Restless Legs Syndrome: a Case-Control Study. *Sleep* **33,** 81/87 (2010).

191. Fusaroli, M. *et al.* Impulse Control Disorders by Dopamine Partial Agonists: A Pharmacovigilance-Pharmacodynamic Assessment Through the FDA Adverse Event Reporting System. *International Journal of Neuropsychopharmacology* **25,** 727–736. doi:10.1093/ijnp/pyac031 (2022).

192. Béreau, M. *et al.* Hyperdopaminergic behavioral spectrum in Parkinson's disease: A review. *Revue Neurologique* **174,** 653–663. doi:10.1016/j.neurol.2018.07.005 (2018).

193. Jeyadevan, A. *et al.* Quality of life implications for elevated trait impulsivity in people with Parkinson's disease. *Quality of Life Research* **32,** 1143–1150. doi:10.1007/s11136-022-03321-w (2023).

194. Leplow, B. *et al.* Characteristics of behavioural addiction in Parkinson's disease patients with self-reported impulse control disorder and controls matched for levodopa equivalent dose: a matched case–control study. *Journal of Neural Transmission* **130,** 125–133. doi:10.1007/s00702-023-02588-8 (2023).

195. Phu, A. L. *et al.* Effect of impulse control disorders on disability and quality of life in Parkinson's disease patients. *Journal of Clinical Neuroscience* **21,** 63–66. doi:10.1016/j.jocn.2013.02.032 (2014).

196. Skorvanek, M. *et al.* Relationship between the non-motor items of the MDS–UPDRS and Quality of Life in patients with Parkinson's disease. *Journal of the Neurological Sciences* **353,** 87–91. doi:10.1016/j.jns.2015.04.013 (2015).

197. Jesús, S. *et al.* Non-motor symptom burden in patients with Parkinson's disease with impulse control disorders and compulsive behaviours: results from the COPPADIS cohort. *Scientific Reports* **10,** 1–12. doi:10.1038/s41598-020-73756-z (2020).

198. Schrag, A. *et al.* Caregiver-burden in parkinson's disease is closely associated with psychiatric symptoms, falls, and disability. *Parkinsonism & Related Disorders* **12,** 35–41. doi:10.1016/j.parkreldis.2005.06.011 (2006).

199. Antonini, A. *et al.* ICARUS study: prevalence and clinical features of impulse control disorders in Parkinson's disease. *Journal of Neurology, Neurosurgery & Psychiatry* **88,** 317–324. doi:10.1136/jnnp-2016-315277 (2017).

200. Castillon, G., Salvo, F. & Moride, Y. The Social Impact of Suspected Adverse Drug Reactions: An analysis of the Canada Vigilance Spontaneous Reporting Database. *Drug Safety* **42,** 27–34. doi:10.1007/s40264-018-0713-8 (2019).

201. FDA. *FDA Adverse Event Reporting System (FAERS): Latest 605 Quarterly Data Files* 2019. https://www.fda.gov/drugs/questions-and-answers-fdas-adverse-event-reporting-system-faers/fda-adverse-event-reporting-system-faers-latest-quarterly-data-files (2023).

202. Härmark, L., van Hunsel, F. & Grundmark, B. ADR Reporting by the General Public: Lessons Learnt from the Dutch and Swedish Systems. *Drug Safety* **38,** 337–347. doi:10.1007/s40264-015-0264-1 (2015).

203. Härmark, L. *et al.* Patient-Reported Safety Information: A Renaissance of Pharmacovigilance? *Drug Safety* **39,** 883–890. doi:10.1007/s40264-016-0441-x (2016).

204. Inácio, P., Cavaco, A. & Airaksinen, M. The value of patient reporting to the pharmacovigilance system: a systematic review. *British Journal of Clinical Pharmacology* **83,** 227–246. doi:10.1111/bcp.13098 (2017).

205. Rolfes, L. *et al.* Adverse drug reaction reports of patients and healthcare professionals—differences in reported information. *Pharmacoepidemiology and Drug Safety* **24,** 152–158. doi:10.1002/pds.3687 (2015).

206. Watson, S. *et al.* Safety Concerns Reported by Patients Identified in a Collaborative Signal Detection Workshop using VigiBase: Results and Reflections from Lareb and Uppsala Monitoring Centre. *Drug Safety* **41,** 203–212. doi:10.1007/s40264-017-0594-2 (2018).

207. Fusaroli, M. *et al.* Development of a Network-Based Signal Detection Tool: The COVID-19 Adversome in the FDA Adverse Event Reporting System. *Frontiers in Pharmacology* **12,** 740707. doi:10.3389/fphar.2021.740707 (2021).

208. FDA. *FAERS Quarterly Data Extract Files* 2022. https://fis.fda.gov/extensions/FPD-QDE-FAERS/FPD-QDE-FAERS.html (2022).

209. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. *MedDRA* https://www.meddra.org/%7B%5C%%7D0D (2023).

210. Gaimari, A. *et al.* Amyotrophic Lateral Sclerosis as an Adverse Drug Reaction: A Disproportionality Analysis of the Food and Drug Administration Adverse Event Reporting System. *Drug Safety* **45,** 663–673. doi:10.1007/s40264-022-01184-1 (2022).

211. Fusaroli, M. *et al.* Standardization of drug names in the FDA Adverse Event Reporting System: The DiAna dictionary. *medRxiv,* 2023.06.07.23291076. doi:10.1101/2023.06.07.23291076 (2023).

212. Norén, G. N., Hopstadius, J. & Bate, A. Shrinkage observed-to-expected ratios for robust and transparent large-scale pattern discovery. *Statistical Methods in Medical Research* **22,** 57–69. doi:10.1177/0962280211403604 (2013).

213. Church, K. & Hanks, P. Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics* **16,** 22–29 (1990).

214. Recchia, G. & Jones, M. N. More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis. *Behavior Research Methods* **41,** 647–656. doi:10.3758/brm.41.3.647 (2009).

215. Role, F. & Nadif, M. *Handling the Impact of Low Frequency Events on Co-occurrence based Measures of Word Similarity - A Case Study of Pointwise Mutual Information.* in *KDIR* (SciTePress, 2011), 226–231.

216. Fusaroli, M. *et al.* Deliberate Self-Poisoning: Real-Time Characterization of Suicidal Habits and Toxidromes in the Food and Drug Administration Adverse Event Reporting System. *Drug Safety* **46,** 283–295. doi:10.1007/s40264-022-01269-x (2023).

217. Pauline, O. *et al.* Assessment of Reported Adverse Events After Interchanging Between TNF-α Inhibitor Biosimilars in the WHO Pharmacovigilance Database. *BioDrugs* **37,** 699–707. doi:10.1007/s40259-023-00603-8 (2023).

218. Orre, R. *et al.* A Bayesian Recurrent Neural Network for Unsupervised Pattern Recognition in Large Incomplete Data Sets. *International Journal of Neural Systems* **15,** 207–222. doi:10.1142/s0129065705000219 (2005).

219. Chen, S. F. & Goodman, J. *An empirical study of smoothing techniques for language modeling* in *Proceedings of the Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics (ACL), 1996), 310–318. doi:10.3115/981863.981904.

220. Van Borkulo, C. D. *et al.* A new method for constructing networks from binary data. *Scientific Reports* **4,** 1–10. doi:10.1038/srep05918 (2014).

221. Epskamp, S., Borsboom, D. & Fried, E. I. Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods* **50,** 195–212. doi:10.3758/s13428-017-0862-1 (2018).

222. Yule, G. U. On the Methods of Measuring Association Between Two Attributes. *Journal of the Royal Statistical Society* **75,** 579. doi:10.2307/2340126 (1912).

223. Newman, M. E. & Girvan, M. Finding and evaluating community structure in networks. *Physical Review E* **69,** 026113. doi:10.1103/physreve.69.026113 (2004).

224. Jaccard, P. The Distribution of the Flora in the Alpine Zone. *New Phytologist* **11,** 37–50. doi:10.1111/j.1469-8137.1912.tb05611.x (1912).

225. Rezaei, M. & Franti, P. Set matching measures for external cluster validity. *IEEE Transactions on Knowledge and Data Engineering* **28,** 2173–2186. doi:10.1109/tkde.2016.2551240 (2016).

226. Gates, A. J. *et al.* Element-centric clustering comparison unifies overlaps and hierarchy. *Scientific Reports* **9,** 8574. doi:10.1038/s41598-019-44892-y (2019).

227. Meilă, M. *Comparing clusterings by the variation of information* in *Learning Theory and Kernel Machines* **2777** (Springer, 2003), 173–187. doi:10.1007/978-3-540-45167-9_14.

228. Telesford, Q. K. *et al.* The Ubiquity of Small-World Networks. *Brain Connectivity* **1,** 367–375. doi:10.1089/brain.2011.0038 (2011).

229. Okai, D. *et al.* Impulse control disorders and dopamine dysregulation in Parkinson's disease: a broader conceptual framework. *European Journal of Neurology* **18,** 1379–1383. doi:10.1111/j.1468-1331.2011.03432.x (2011).

230. El Otmani, H. *et al.* Othello syndrome in Parkinson's disease: A diagnostic emergency of an underestimated condition. *Revue Neurologique* **177,** 690–693. doi:10.1016/j.neurol.2020.08.006 (2021).

231. Weintraub, D. *et al.* Impulse Control Disorders in Parkinson Disease: A Cross-Sectional Study of 3090 Patients. *Archives of Neurology* **67,** 589–595. doi:10.1001/archneurol.2010.65 (2010).

232. Santangelo, G. *et al.* Pathological gambling in Parkinson's disease. A comprehensive review. *Parkinsonism & Related Disorders* **19,** 645–653. doi:10.1016/j.parkreldis.2013.02.007 (2013).

233. Jiménez-Urbieta, H. *et al.* Pramipexole-induced impulsivity in mildparkinsonian rats: a model of impulse control disorders in Parkinson's disease. *Neurobiology of Aging* **75,** 126–135. doi:10.1016/j.neurobiolaging.2018.11.021 (2019).

234. Riddle, J. L., Rokosik, S. L. & Napier, T. C. Pramipexole- and methamphetamine-induced reward-mediated behavior in a rodent model of Parkinson's disease and controls. *Behavioural Brain Research* **233,** 15–23. doi:10.1016/j.brr.2012.04.027 (2012).

235. Marín-Lahoz, J. *et al.* Depression as a Risk Factor for Impulse Control Disorders in Parkinson Disease. *Annals of Neurology* **86,** 762–769. doi:10.1002/ana.25581 (2019).

236. Blum, A. W., Jacobson, K. & Grant, J. E. Legal settlements and the reporting of adverse drug events: Insights from the aripiprazole products liability litigation. *Psychiatry Research* **309,** 114411. doi:10.1016/j.psychres.2022.114411 (2022).

237. Voon, V. *et al.* Impulse control disorders in parkinson disease: A multicenter case–control study. *Annals of Neurology* **69,** 986–996. doi:10.1002/ana.22356 (2011).

238. Masa, J. F. *et al.* Obesity hypoventilation syndrome. *European Respiratory Review* **28.** doi:10.1183/16000617.0097-2018 (2019).

239. De Jonge, P. *et al.* Associations between DSM-IV mental disorders and diabetes mellitus: A role for impulse control disorders and depression. *Diabetologia* **57,** 699–709. doi:10.1007/s00125-013-3157-9 (2014).

240. Evans, A. H. *et al.* Impulsive and compulsive behaviors in Parkinson's disease. *Movement Disorders* **24,** 1561–1570. doi:10.1002/mds.22505 (2009).

241. De La Casa-Fages, B. & Grandas, F. Dopamine dysregulation syndrome and deep brain stimulation of the subthalamic nucleus in Parkinson's disease. *Neurology Research International* **2011,** 759895. doi:10.1155/2011/759895 (2011).

242. Weintraub, D. *et al.* Clinical spectrum of impulse control disorders in Parkinson's disease. *Movement Disorders* **30,** 121–127. doi:10.1002/mds.26016 (2015).

243. Yu, X. X. & Fernandez, H. H. Dopamine agonist withdrawal syndrome: A comprehensive review. *Journal of the Neurological Sciences* **374,** 53–55. doi:10.1016/j.jns.2016.12.070 (2017).

244. Kataoka, H. & Sugie, K. Delusional jealousy (othello syndrome) in 67 patients with Parkinson's disease. *Frontiers in Neurology* **9,** 129. doi:10.3389/fneur.2018.00129 (2018).

245. Aarsland, D. *et al.* Predictors of Nursing Home Placement in Parkinson's Disease: A Population-Based, Prospective Study. *Journal of the American Geriatrics Society* **48,** 938–942. doi:10.1111/j.1532-5415.2000.tb06891.x (2000).

246. Lopez, A. M., Weintraub, D. & Claassen, D. O. Impulse Control Disorders and Related Complications of Parkinson's Disease Therapy. *Seminars in Neurology* **37,** 186–192. doi:10.1055/s-0037-1601887 (2017).

247. Goerlich-Dobre, K. S. *et al.* Alexithymia - an independent risk factor for impulsive-compulsive disorders in Parkinson's disease. *Movement Disorders* **29,** 214–220. doi:10.1002/mds.25679 (2014).

248. Vriend, C. *et al.* Depression and impulse control disorders in Parkinson's disease: Two sides of the same coin? *Neuroscience & Biobehavioral Reviews* **38,** 60–71. doi:10.1016/j.neubiorev.2013.11.001 (2014).

249. Plummer, N. R. *et al.* Dopamine agonists for traumatic brain injury. *Cochrane Database of Systematic Reviews.* doi:10.1002/14651858.cd013062 (2018).

250. Feltenstein, M. W., Do, P. H. & See, R. E. Repeated aripiprazole administration attenuates cocaine seeking in a rat model of relapse. *Psychopharmacology* **207,** 401–411. doi:10.1007/s00213-009-1671-8 (2009).

251. Feltenstein, M. W., Altar, C. A. & See, R. E. Aripiprazole Blocks Reinstatement of Cocaine Seeking in an Animal Model of Relapse. *Biological Psychiatry* **61,** 582–590. doi:10.1016/j.biopsych.2006.04.010 (2007).

252. Menestrina, L. & Recanatini, M. An unsupervised computational pipeline identifies potential repurposable drugs to treat Huntington's disease and multiple sclerosis. *Artificial Intelligence in the Life Sciences* **2,** 100042. doi:10.1016/j.ailsci.2022.100042 (2022).

253. Talevi, A. & Bellera, C. L. Challenges and opportunities with drug repurposing: finding strategies to find alternative uses of therapeutics. *Expert Opinion on Drug Discovery* **15,** 397–401. doi:10.1080/17460441.2020.1704729 (2020).

254. Choudhury, C., Arul Murugan, N. & Priyakumar, U. D. Structure-based drug repurposing: Traditional and advanced AI/ML-aided methods. *Drug Discovery Today*. doi:10.1016/j.drudis.2022.03.006 (2022).

255. Guney, E. *et al.* Network-based in silico drug efficacy screening. *Nature Communications* **7,** 10331. doi:10.1038/ncomms10331 (2016).

256. Cheng, F. *et al.* Network-based approach to prediction and population-based validation of in silico drug repurposing. *Nature Communications* **9,** 2691. doi:10.1038/s41467-018-05116-5 (2018).

257. Cheng, F. *et al.* A genome-wide positioning systems network algorithm for in silico drug repurposing. *Nature Communications* **10,** 3476. doi:10.1038/s41467-019-10744-6 (2019).

258. Zhou, Y. *et al.* A network medicine approach to investigation and population-based validation of disease manifestations and drug repurposing for COVID-19. *PLOS Biology* **18,** e3000970. doi:10.1371/journal.pbioO.3000970 (2020).

259. Fang, J. *et al.* Endophenotype-based in silico network medicine discovery combined with insurance record data mining identifies sildenafil as a candidate drug for Alzheimer's disease. *Nature Aging* **1,** 1175–1188. doi:10.1038/s43587-021-00138-z (2021).

260. Peng, Y. *et al.* Screening novel drug candidates for Alzheimer's disease by an integrated network and transcriptome analysis. *Bioinformatics* **36,** 4626–4632. doi:10.1093/bioinformatics/btaa563 (2020).

261. Quan, P. *et al.* Integrated network analysis identifying potential novel drug candidates and targets for Parkinson's disease. *Scientific Reports* **11,** 13154. doi:10.1038/s41598-021-92701-2 (2021).

262. Bates, G. P. *et al.* Huntington disease. *Nature Reviews Disease Primers* **1,** 15005. doi:10.1038/nrdp.2015.5 (2015).

263. Loscalzo, J., Kohane, I. & Barabasi, A. L. Human disease classification in the postgenomic era: A complex systems approach to human pathobiology. *Molecular Systems Biology* **3,** 124. doi:10.1038/msb4100163 (2007).

264. Wright, G. E. *et al.* Interrupting sequence variants and age of onset in Huntington's disease: clinical implications and emerging therapies. *The Lancet Neurology* **19,** 930–939. doi:10.1016/s1474-4422(20)30343-4 (2020).

265. Finkbeiner, S. Huntington's disease. *Cold Spring Harbor Perspectives in Biology* **3,** a007476. doi:10.1101/cshperspect.a007476 (2011).

266. Kobelt, G. *et al.* New insights into the burden and costs of multiple sclerosis in Europe. *Multiple Sclerosis* **23,** 1123–1136. doi:10.1177/1352458517694432 (2017).

267. Leray, E. *et al.* Epidemiology of multiple sclerosis. *Revue Neurologique* **172,** 3–13. doi:10.1016/j.neurol.2015.10.006 (2016).

268. Ramagopalan, S. V. *et al.* Multiple sclerosis: risk factors, prodromes, and potential causal pathways. *The Lancet Neurology* **9,** 727–739. doi:10.1016/s1474-4422(10)70094-6 (2010).

269. Dobson, R. & Giovannoni, G. Multiple sclerosis - a review. *European Journal of Neurology* **26,** 27–40. doi:10.1111/ene.13819 (2019).

270. Duff, K. *et al.* Risperidone and the treatment of psychiatric, motor, and cognitive symptoms in Huntington's disease. *Annals of Clinical Psychiatry* **20,** 1–3. doi:10.1080/10401230701844802 (2008).

271. Hamid, K. M. *et al.* JAK-STAT Lodges in Multiple Sclerosis: Pathophysiology and Therapeutic Approach Overview. *Open Access Library Journal* **4,** e3492. doi:10.4236/oalib.1103492 (2017).

272. Gitler, A. D., Dhillon, P. & Shorter, J. Neurodegenerative disease: Models, mechanisms, and a new hope. *DMM Disease Models and Mechanisms* **10,** 499–502. doi:10.1242/dmm.030205 (2017).

273. Kanehisa, M. *et al.* KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research* **45,** D353–D361. doi:10.1093/nar/gkw1092 (2017).

274. Amberger, J. S. *et al.* OMIM.org: Leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Research* **47,** D1038–D1043. doi:10.1093/nar/gky1151 (2019).

275. Ramos, E. M. *et al.* Phenotype-genotype integrator (PheGenI): Synthesizing genome-wide association study (GWAS) data with existing genomic resources. *European Journal of Human Genetics* **22,** 144–147. doi:10.1038/ejhg.2013.96 (2014).

276. Pletscher-Frankild, S. *et al.* DISEASES: Text mining and data integration of disease-gene associations. *Methods* **74,** 83–89. doi:10.1016/j.ymeth.2014.11.020 (2015).

277. Carbon, S. *et al.* The Gene Ontology resource: Enriching a GOld mine. *Nucleic Acids Research* **49,** D325–D334. doi:10.1093/nar/gkaa1113 (2021).

278. Köhler, S. *et al.* The human phenotype ontology in 2021. *Nucleic Acids Research* **49,** D1207–D1217. doi:10.1093/nar/gkaa1043 (2021).

279. Klopfenstein, D. V. *et al.* GOATOOLS: A Python library for Gene Ontology analyses. *Scientific Reports* **8,** 10872. doi:10.1038/s41598-018-28948-z (2018).

280. Lamb, J. *et al.* The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313,** 1929–1935. doi:10.1126/science.1132939 (2006).

281. Barrett, T. *et al.* NCBI GEO: Archive for functional genomics data sets - Update. *Nucleic Acids Research* **41,** D991–D995. doi:10.1093/nar/gks1193 (2013).

282. Subramanian, A. *et al.* A Next Generation Connectivity Map: L1000 platform and the first 1,000,000 profiles. *Cell* **171,** 1437–1452. doi:10.1016/j.cell.2017.10.049 (2017).

283. Alonso-López, D. *et al.* APID database: Redefining protein-protein interaction experimental evidences and binary interactomes. *Database* **2019,** baz005. doi:10.1093/database/baz005 (2019).

284. Oughtred, R. *et al.* The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Science* **30,** 187–200. doi:10.1002/pro.3978 (2021).

285. Luck, K. *et al.* A reference map of the human binary protein interactome. *Nature* **580,** 402–408. doi:10.1038/s41586-020-2188-x (2020).

286. Breuer, K. *et al.* InnateDB: Systems biology of innate immunity and beyond - Recent updates and continuing curation. *Nucleic Acids Research* **41,** D1228–D1233. doi:10.1093/nar/gks1147 (2013).

287. Meyer, M. J. *et al.* INstruct: A database of high-quality 3D structurally resolved protein interactome networks. *Bioinformatics* **29,** 1577–1579. doi:10.1093/bioinformatics/btt181 (2013).

288. Orchard, S. *et al.* The MIntAct project - IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research* **42,** D358–D363. doi:10.1093/nar/gkt1115 (2014).

289. Csabai, L. *et al.* SignaLink3: A multi-layered resource to uncover tissue-specific signaling networks. *Nucleic Acids Research* **50,** D701–D709. doi:10.1093/nar/gkab909 (2022).

290. Menche, J. *et al.* Uncovering disease-disease relationships through the incomplete interactome. *Science* **347,** 1257601. doi:10.1126/science.1257601 (2015).

291. Jia, J. *et al.* Mechanisms of drug combinations: Interaction and network perspectives. *Nature Reviews Drug Discovery* **8,** 111–128. doi:10.1038/nrd2683 (2009).

292. Süssmuth, S. D. *et al.* An exploratory double-blind, randomized clinical trial with selisistat, a SirT1 inhibitor, in patients with Huntington's disease. *British Journal of Clinical Pharmacology* **79,** 465–476. doi:10.1111/bcp.12512 (2015).

293. Frattola, L. *et al.* Acute treatment of Huntington's chorea with lisuride. *Journal of the Neurological Sciences* **59,** 247–253. doi:10.1016/0022-510x(83)90042-4 (1983).

294. Saft, C. *et al.* Dose-dependent improvement of myoclonic hyperkinesia due to valproic acid in eight Huntington's Disease patients: A case series. *BMC Neurology* **6,** 11. doi:10.1186/1471-2377-6-11 (2006).

295. Mitsonis, C. I. *et al.* Effects of escitalopram on stress-related relapses in women with multiple sclerosis: An open-label, randomized, controlled, one-year follow-up study. *European Neuropsychopharmacology* **20,** 123–131. doi:10.1016/j.euroneuro.2009.10.004 (2010).

296. Karasinska, J. M. & Hayden, M. R. Cholesterol metabolism in Huntington disease. *Nature Reviews Neurology* **7,** 561–572. doi:10.1038/nrneurol.2011.132 (2011).

297. Shukla, S. & Tekwani, B. L. Histone Deacetylases Inhibitors in Neurodegenerative Diseases, Neuroprotection and Neuronal Differentiation. *Frontiers in Pharmacology* **11,** 537. doi:10.3389/fphar.2020.00537 (2020).

298. Hosseini, A. *et al.* Ruxolitinib attenuates experimental autoimmune encephalomyelitis (EAE) development as animal models of multiple sclerosis (MS). *Life Sciences* **276,** 119395. doi:10.1016/j.lfs.2021.119395 (2021).

299. Moscarello, M. A. *et al.* Paclitaxel (Taxol) attenuates clinical disease in a spontaneously demyelinating transgenic mouse and induces remyelination. *Multiple Sclerosis Journal* **8,** 130–138. doi:10.1191/1352458502ms776oa (2002).

300. Gonzalez, G. A. *et al.* Tamoxifen accelerates the repair of demyelinated lesions in the central nervous system. *Scientific Reports* **6,** 1–13. doi:10.1038/srep31599 (2016).

301. Rankin, K. A. *et al.* Selective Estrogen Receptor Modulators Enhance CNS Remyelination Independent of Estrogen Receptors. *Journal of Neuroscience* **39,** 2184–2194. doi:10.1523/jneurosci.1530-18.2019 (2019).

302. Contino-Pépin, C. *et al.* Thalidomide Derivatives for the Treatment of Neuroinflammation. *ChemMedChem* **5,** 2057–2064. doi:10.1002/cmdc.201000326 (2010).

303. Underwood, M., Bonas, S. & Dale, M. Huntington's Disease: Prevalence and Psychological Indicators of Pain. *Movement Disorders Clinical Practice* **4,** 198–204. doi:10.1002/mdc3.12376 (2017).

304. Sprenger, G. P. *et al.* The prevalence of pain in Huntington's disease in a large worldwide cohort. *Parkinsonism and Related Disorders* **89,** 73–78. doi:10.1016/j.parkreldis.2021.06.015 (2021).

305. Pubill, D. *et al.* Orphenadrine prevents 3-nitropropionic acid-induced neurotoxicity in vitro and in vivo. *British Journal of Pharmacology* **132,** 693–702. doi:10.1038/sj.bjp.0703869 (2001).

306. Kolahdouzan, M. & Hamadeh, M. J. The neuroprotective effects of caffeine in neurodegenerative diseases. *CNS Neuroscience and Therapeutics* **23,** 272–290. doi:10.1111/cns.12684 (2017).

307. Ayyadevara, S. *et al.* Aspirin-Mediated Acetylation Protects Against Multiple Neurodegenerative Pathologies by Impeding Protein Aggregation. *Antioxidants and Redox Signaling* **27,** 1383–1396. doi:10.1089/ars.2016.6978 (2017).

308. Günaydın, C. *et al.* Tofacitinib enhances remyelination and improves myelin integrity in cuprizone-induced mice. *Immunopharmacology and Immunotoxicology* **43,** 790–798. doi:10.1080/08923973.2021.1986063 (2021).

309. Benveniste, E. N. *et al.* Involvement of the Janus kinase/signal transducer and activator of transcription signaling pathway in multiple sclerosis and the animal model of experimental autoimmune encephalomyelitis. *Journal of Interferon and Cytokine Research* **34,** 577–588. doi:10.1089/jir.2014.0012 (2014).

310. Königs, C., Friedrichs, M. & Dietrich, T. The heterogeneous pharmacological medical biochemical network PharMeBINet. *Scientific Data* **9,** 393. doi:10.1038/s41597-022-01510-3 (2022).

311. Toutanova, K. & Chen, D. *Observed versus latent features for knowledge base and text inference* in *Proceedings ofthe 3rd Workshop on Continuous Vector Space Models and their Compositionality* (Association for Computational Linguistics, 2015), 57–66. doi:10.18653/v1/w15-4007.

312. Dettmers, T. *et al.* Convolutional 2D knowledge graph embeddings. *32nd AAAI Conference on Artificial Intelligence (AAAI-18),* 221 (2018).

313. Rebele, T. *et al. YAGO: A multilingual knowledge base from wikipedia, wordnet, and geonames* in *The Semantic Web – ISWC 2016. International Semantic Web Conference 2016. Lecture Notes in Computer Science* **9982** (Springer Verlag, 2016), 177–185. doi:10.1007/978-3-319-46547-0_19.

314. Hu, W. *et al. Open Graph Benchmark: Datasets for Machine Learning on Graphs* in *34th Conference on Neural Information Processing Systems (NeurIPS 2020)* (Neural information processing systems foundation, 2020).

315. Galkin, M. *et al. NodePiece: Compositional and Parameter-Efficient Representations of Large Knowledge Graphs* in *International Conference on Learning Representations* (2022).

316. National Center for Biotechnology Information (NCBI). *GENE_INFO/Mammalia* 2023. https://ftp.ncbi.nlm.nih.gov/gene/DATA/GENE%7B%5C_%7DINFO/Mammalia/.

317. Peri, S. *et al.* Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Research* **32,** D497–D501. doi:10.1093/NAR/GKH070 (2004).

318. Cowley, M. J. *et al.* PINA v2.0: mining interactome modules. *Nucleic Acids Research* **40,** D862–D865. doi:10.1093/nar/gkr967 (2012).

319. Bateman, A. *et al.* UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research* **51,** D523–D531. doi:10.1093/nar/gkac1052 (2023).

320. Seal, R. L. *et al.* Genenames.org: the HGNC resources in 2023. *Nucleic Acids Research* **51,** D1003–D1009. doi:10.1093/nar/gkac888 (2023).

321. Natale, D. A. *et al.* Protein Ontology (PRO): enhancing and scaling up the representation of protein entities. *Nucleic Acids Research* **45,** D339–D346. doi:10.1093/nar/gkw1075 (2017).

322. Vasilevsky, N. A. *et al.* Mondo: Unifying diseases for the world, by the world. *medRxiv,* 2022.04.13.22273750. doi:10.1101/2022.04.13.22273750 (2022).

323. Bastian, F. B. *et al.* The Bgee suite: integrated curated expression atlas and comparative transcriptomics in animals. *Nucleic Acids Research* **49,** D831–D847. doi:10.1093/nar/gkaa793 (2021).

324. Mungall, C. J. *et al.* Uberon, an integrative multi-species anatomy ontology. *Genome Biology* **13,** R5. doi:10.1186/gb-2012-13-1-r5 (2012).

325. Rodchenkov, I. *et al.* Pathway Commons 2019 Update: integration, analysis and exploration of pathway data. *Nucleic Acids Research* **48,** D489–D497. doi:10.1093/NAR/GKZ946 (2020).

326. Ursu, O. *et al.* DrugCentral: online drug compendium. *Nucleic Acids Research* **45,** D932–D939. doi:10.1093/nar/gkw993 (2017).

327. Ali, M. *et al.* PyKEEN 1.0: A Python Library for Training and Evaluating Knowledge Graph Embeddings. *Journal of Machine Learning Research* **22,** 1–6 (2021).

328. Sun, Z. *et al. RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space* in *7th International Conference on Learning Representations, ICLR 2019* (International Conference on Learning Representations, ICLR, 2019).

329. Trouillon, T. *et al. Complex Embeddings for Simple Link Prediction* in *33rd International Conference on Machine Learning* (2016).

330. Wang, Z. *et al. Knowledge Graph Embedding by Translating on Hyperplanes* in *Proceedings of the AAAI Conference on Artificial Intelligence* **28** (AI Access Foundation, 2014), 1112–1119. doi:10.1609/aaai.v28i1.8870.

331. Bonner, S. *et al.* Understanding the performance of knowledge graph embeddings in drug discovery. *Artificial Intelligence in the Life Sciences* **2,** 100036. doi:10.1016/j.ailsci.2022.100036 (2022).

332. Bordes, A. *et al. Translating Embeddings for Modeling Multi-relational Data* in *Advances in Neural Information Processing Systems 26* (2013).

333. Yang, B. *et al. Embedding Entities and Relations for Learning and Inference in Knowledge Bases* in *3rd International Conference on Learning Representations* (International Conference on Learning Representations, ICLR, 2015).

334. Trouillon, T. *et al.* Knowledge graph completion via complex tensor factorization. *The Journal of Machine Learning Research* **18,** 4735–4772. doi:10.5555/3122009.3208011 (2017).

335. Ali, M. *et al.* Bringing Light into the Dark: A Large-Scale Evaluation of Knowledge Graph Embedding Models under a Unified Framework. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44,** 8825–8845. doi:10.1109/tpami.2021.3124805 (2022).

336. Berrendorf, M. *et al.* On the Ambiguity of Rank-Based Evaluation of Entity Alignment or Link Prediction Methods. *arXiv* (2020).

337. Thomas, P. D. *et al.* PANTHER: Making genome-scale phylogenetics accessible to all. *Protein Science* **31,** 8–22. doi:10.1002/pro.4218 (2022).

338. Luo, J. *et al.* A multifunctional therapeutic approach to disease modification in multiple familial mouse models and a novel sporadic model of Alzheimer's disease. *Molecular Neurodegeneration* **11,** 35. doi:10.1186/s13024-016-0103-6 (2016).

339. Vandevrede, L. *et al.* Novel analogues of chlormethiazole are neuroprotective in four cellular models of neurodegeneration by a mechanism with variable dependence on GABA A receptor potentiation. *British Journal of Pharmacology* **171,** 389–402. doi:10.1111/bph.12454 (2014).

340. Gutiérrez, I. L. *et al.* Reboxetine Treatment Reduces Neuroinflammation and Neurodegeneration in the 5xFAD Mouse Model of Alzheimer's Disease: Role of CCL2. *Molecular Neurobiology* **56,** 8628–8642. doi:10.1007/s12035-019-0169 5-6 (2019).

341. Ihara, M., Saito, S. & Friedland, R. Drug Repositioning for Alzheimer's Disease: Finding Hidden Clues in Old Drugs. *Journal of Alzheimer's Disease* **74,** 1013–1028. doi:10.3233/jad-200049 (2020).

342. Munafò, A. *et al.* Repositioning of Immunomodulators: A Ray of Hope for Alzheimer's Disease? *Frontiers in Neuroscience* **14,** 614643. doi:10.3389/fnins.2020.614643 (2020).

343. Song, C. *et al.* Immunotherapy for Alzheimer's disease: targeting β-amyloid and beyond. *Translational Neurodegeneration* **11,** 18. doi:10.1186/s40035-022-00292-3 (2022).

344. Cummings, J. *et al.* Alzheimer's disease drug development pipeline: 2022. *Alzheimer's & Dementia: Translational Research & Clinical Interventions* **8,** e12295. doi:10.1002/trc2.12295 (2022).

345. Lerner, A. J. *et al.* CYP46A1 activation by low-dose efavirenz enhances brain cholesterol metabolism in subjects with early Alzheimer's disease. *Alzheimer's Research and Therapy* **14,** 198. doi:10.1186/s13195-022-01151-z (2022).

346. Gasymov, O. K. *et al.* Sodium fusidate prevents protein aggregation of silk fibroin and offers new perspectives for human lens material disaggregation. *Biophysical Chemistry* **279,** 106680. doi:10.1016/j.bpc.2021.106680 (2021).

347. Chen, C. Y. *et al.* The Association Between Dextromethorphan Use and the Risk of Dementia. *American Journal of Alzheimer's Disease and other Dementias* **37,** 1–10. doi:10.1177/15333175221124952 (2022).

348. Arispe, N., Rojas, E. & Pollard, H. B. Alzheimer disease amyloid beta protein forms calcium channels in bilayer membranes: blockade by tromethamine and aluminum. *Proceedings of the National Academy of Sciences* **90,** 567–571. doi:10.1073/pnas.90.2.567 (1993).

349. Gu, X. *et al.* Safinamide protects against amyloid β (Aβ)-induced oxidative stress and cellular senescence in M17 neuronal cells. *Bioengineered* **13,** 1921–1930. doi:10.1080/21655979.2021.2022262 (2022).

350. Rojas, N. G. *et al.* Review of Huntington's Disease: From Basics to Advances in Diagnosis and Treatment. *Journal of Neurology Research* **12,** 93–113. doi:10.14740/jnr.v12I3.721 (2022).

351. Andhale, R. *et al.* Huntington's Disease: A Clinical Review. *Cureus* **14,** e28484. doi:10.7759/cureus.28484 (2022).

352. Andrich, J. *et al.* Autonomic nervous system function in Huntington's disease. *Journal of Neurology, Neurosurgery & Psychiatry* **72,** 726–731. doi:10.1136/jnnp.72.6.726 (2002).

353. Abbruzzese, G. *et al.* Abnormalities of parietal and prerolandic somatosensory evoked potentials in Huntington's disease. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section* **77,** 340–346. doi:10.1016/0168-5597(90)90055-i (1990).

354. Schwarz, M. *et al.* Abnormalities of somatosensory evoked potentials in the quinolinic acid model of Huntington's disease: Evidence that basal ganglia modulate sensory cortical input. *Annals of Neurology* **32,** 358–364. doi:10.1002/ana.410320309 (1992).

355. Bozzi, M. & Sciandra, F. Molecular Mechanisms Underlying Muscle Wasting in Huntington's Disease. *International Journal of Molecular Sciences 2020, Vol. 21, Page 8314* **21,** 8314. doi:10.3390/ijms21218314 (2020).

356. Centonze, D. *et al.* The endocannabinoid system is dysregulated in multiple sclerosis and in experimental autoimmune encephalomyelitis. *Brain* **130,** 2543–2553. doi:10.1093/brain/awm160 (2007).

357. Correa, F. G. *et al.* in *Vitamins and Hormones* 207–230 (Academic Press, 2009). doi:10.1016/s0083-6729(09)81009-1.

358. Cannella, B. *et al.* Multiple sclerosis: Death receptor expression and oligodendrocyte apoptosis in established lesions. *Journal of Neuroimmunology* **188,** 128–137. doi:10.1016/j.neuroim.2007.05.018 (2007).

359. Meyer, T. *et al.* NAD+ metabolism drives astrocyte proinflammatory reprogramming in central nervous system autoimmunity. *Proceedings of the National Academy of Sciences of the United States of America* **119,** e2211310119. doi:10.1073/pnas.2211310119 (2022).

360. Penberthy, W. & Tsunoda, I. The Importance of NAD in Multiple Sclerosis. *Current Pharmaceutical Design* **15,** 64–99. doi:10.2174/138161209787185751 (2009).

361. Ishina, I. A. *et al.* MHC Class II Presentation in Autoimmunity. *Cells* **12,** 314. doi:10.3390/cells12020314 (2023).

362. Pal, S. N. *et al.* WHO strategy for collecting safety data in public health programmes: Complementing spontaneous reporting systems. *Drug Safety* **36,** 75–81. doi:10.1007/s40264-012-0014-6 (2013).

363. Abu Esba, L. C. *et al.* Fixed dose combinations: A formulary decision guide. *Health Policy and Technology* **10,** 100500. doi:10.1016/j.hlpt.2021.02.006 (2021).

364. Chemotargets S.L. *ClarityPV* 2023. https://claritypv.com/.

365. FDA. *FDA Adverse Event Reporting System (FAERS): Latest Quarterly Data Files | FDA* 2023. https://www.fda.gov/drugs/questions-and-answers-fdas-adverse-event-reporting-system-faers/fda-adverse-event-reporting-system-faers-latest-quarterly-data-files.

366. Lindquist, M. VigiBase, the WHO Global ICSR Database System: Basic Facts. *Drug Information Journal* **42,** 409–419. doi:10.1177/009286150804200501 (2008).

367. *VAERS - Data* https://vaers.hhs.gov/data.html.

368. Evans, S. J., Waller, P. C. & Davis, S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiology and Drug Safety* **10,** 483–486. doi:10.1002/pds.677 (2001).

369. EMA. *New product information wording: extracts from PRAC recommendations on signals adopted at the 26-29 September 2022 PRAC meeting* 2022. https://www.ema.europa.eu/en/documents/prac-recommendation/new-product-information-wording-extracts-prac-recommendations-signals-adopted-26-29-september-2022%7B%5C_%7Den.pdf.

370. Croatto, G. *et al.* The impact of pharmacological and non-pharmacological interventions on physical health outcomes in people with mood disorders across the lifespan: An umbrella review of the evidence from randomised controlled trials. *Molecular Psychiatry* **28,** 369–390. doi:10.1038/s41380-022-01770-w (2022).

371. Nobile, M. T. *et al.* Randomised comparison of weekly bolus 5-fluorouracil with or without leucovorin in metastatic colorectal carcinoma. *European Journal of Cancer* **28,** 1823–1827. doi:10.1016/0959-8049(92)90013-r (1992).

372. Kumar, A. *et al.* Norfloxacin Tinidazole induced Steven Johnsons Syndrome: A Case Report. *International Journal of Contemporary Medical Research* **4,** 2336–2337 (2017).

373. Jhaj, R. *et al.* Fixed-drug eruptions: What can we learn from a case series? *Indian Journal of Dermatology* **63,** 332. doi:10.4103/ijd.ijd_481_17 (2018).

374. Ratikanta, T. & Ranjan, P. M. A rare case of recurrent fixed drug eruption and lip edema due to norfloxacin and tinidazole fized dose combination. *International Research Journal of Pharmacy* **6,** 204–205. doi:10.7897/2230-8407.06344 (2015).

Part V

Appendices

List of Figures

List of Tables

A COVIDrugNet

Table A.1. Fittings Evaluation. The table provides the log-likelihood and respective p-value for each function fitted on every analyzed network. For the power-law function, the Kolmogorov-Smirnov distance (D) is provided, and the fitting is considered plausible if the respective p-value is at least 0.1.

	Drug Projection (Entire)		**Drug Projection**	
	D	p-value	D	p-value
Power-Law	0.07	2.4×10^{-1}	0.05	3.3×10^{-1}
	Likelihood-ratio	p-value	Likelihood-ratio	p-value
Truncated Power-Law	-0.58	2.8×10^{-1}	-8.01	6.3×10^{-5}
Exponential	0.91	6.9×10^{-1}	15.71	6.6×10^{-2}
Stretched Exponential	-0.42	5.8×10^{-1}	-6.80	4.8×10^{-3}
Lognormal	-0.33	5.9×10^{-1}	-5.92	7.1×10^{-3}
	Target Projection (Entire)		**Target Projection**	
	D	p-value	D	p-value
Power-Law	0.22	0.0	0.06	9.5×10^{-2}
	Likelihood-ratio	p-value	Likelihood-ratio	p-value
Truncated Power-Law	-329.39	0.0	-9.43	1.4×10^{-5}
Exponential	-349.29	1.0×10^{-21}	-3.57	5.6×10^{-1}
Stretched Exponential	-397.78	1.66×10^{-54}	-9.27	6.6×10^{-4}
Lognormal	-317.17	1.1×10^{-51}	-8.86	1.1×10^{-3}

Figure A.1. Degree Distribution. The degree distributions of both the drug (a) and target (b) projections (red), compared to those of equivalent random graphs (blue). The last ones were generated with the Erdős-Rényi model[33], keeping the same number of nodes and probability of edge creation (ratio between the actual number of edges and the maximum possible edges) of the network they are compared with.

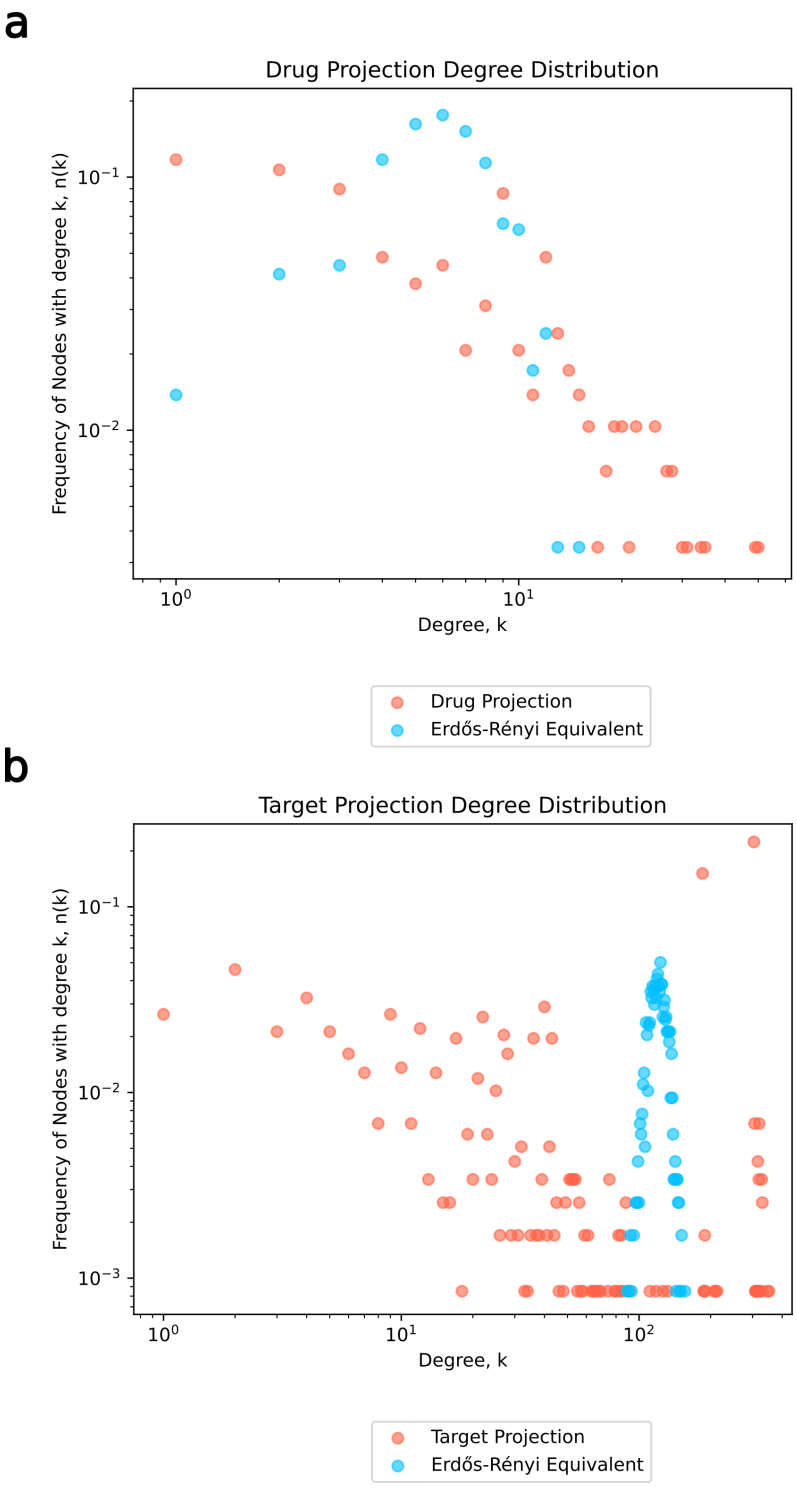


Figure A.2. Degree Distribution Fittings. The degree distributions of the entire drug projection (a) and entire target projection (c) networks, and of the corresponding graphs in which all nodes except Artenimol, Fostamatinib and their exclusive direct neighbors were present (b and d, respectively). On each distribution the following functions are fitted: power-law (orange), truncated power-law (green), exponential (red), stretched exponential (violet) and lognormal (brown). In every chart the probability density function (PDF, blue) is shown too.

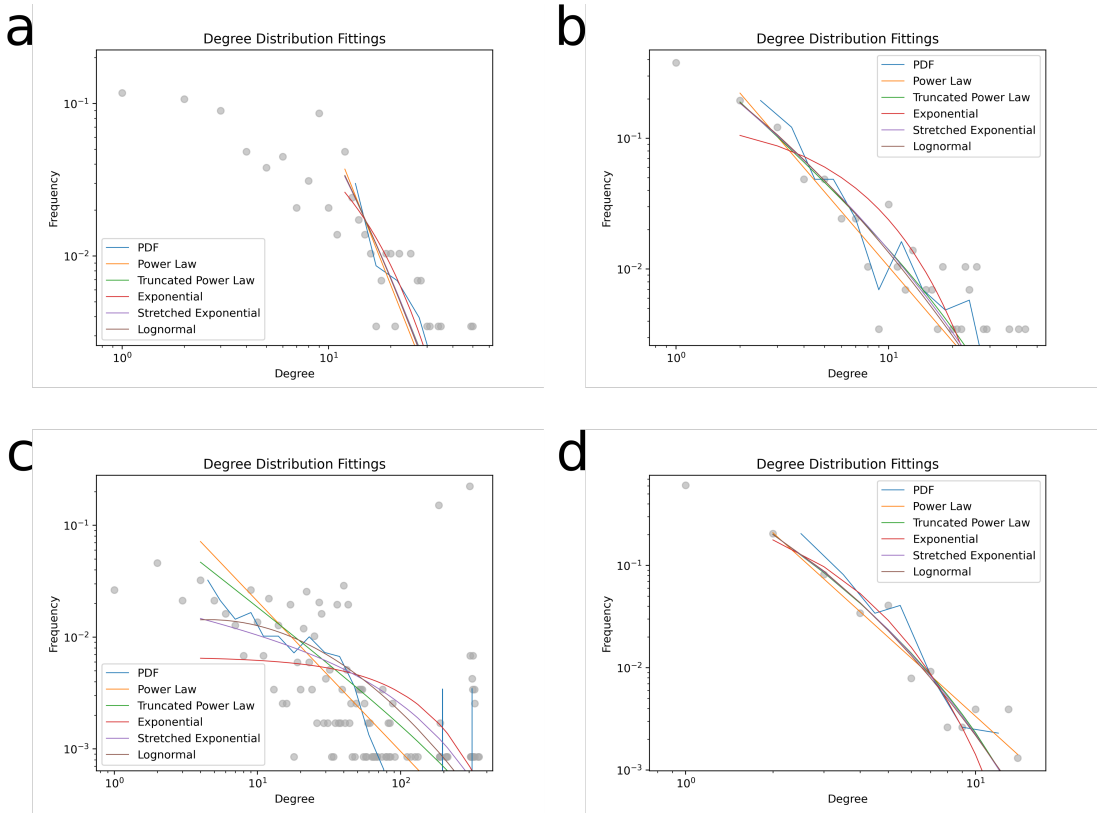


Figure A.3. Clustering Coefficient and Degree Relationship. The relationship between the clustering coefficient and the degree of nodes in the entire drug projection (a), the entire target projection (c) networks, and the corresponding graphs in which all nodes except Artenimol, Fostamatinib and their exclusive direct neighbors were present (b and d, respectively).

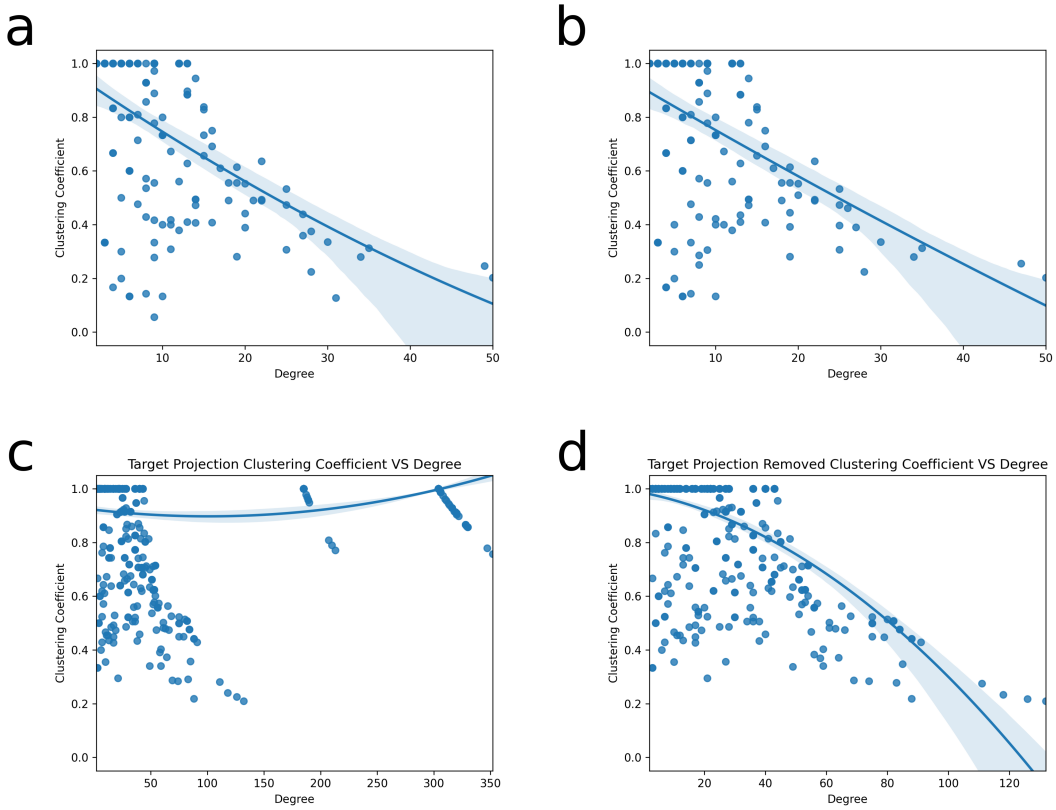
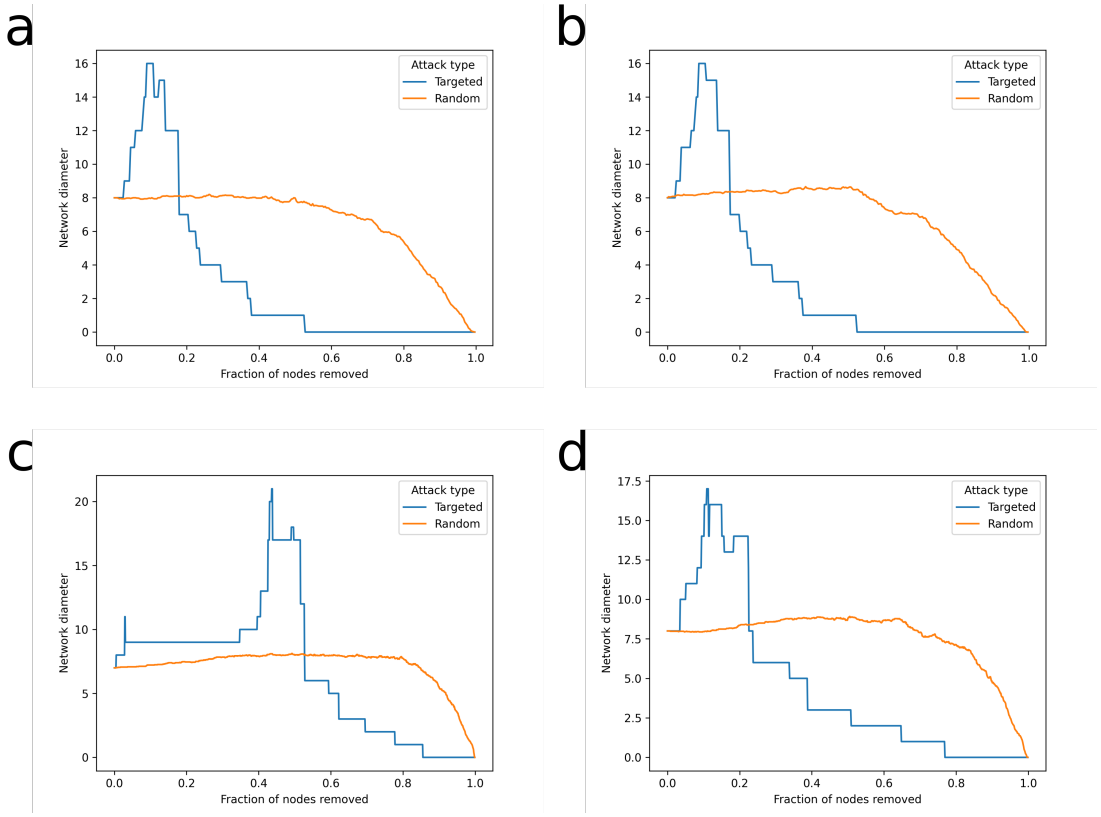


Figure A.4. Network Robustness. The comparison of the network diameter variation in response to targeted attacks (in blue) and random failures (in orange). The investigation was carried out on the entire drug projection (a), the entire target projection (c) networks, and the corresponding graphs in which all nodes except Artenimol, Fostamatinib and their exclusive direct neighbors were present (b and d, respectively).



B Unsupervised Pipeline for Drug Repurposing

Table B.1. Source Databases for the Human Protein-protein Interactome Construction.

Database	Nodes	Edges	Filters	Notes
APID	15,313	131,509	Quality level 1. Interactions proven by 2 or more experimental evidences.	
BioGRID	19,498	680,439		
HuRI	8,185	51,941		Ensembl IDs mapped to official gene symbols using NCBI database.
InnateDB	15,495	166,882	Confidence score: NP $\geq$ 1. There is at least one publication supporting the interaction that has never been used to support any other interaction (http://wodaklab.org/iRefWeb/faq).	
INstruct	3,616	6,569		
IntAct	4,136	7,853	Confidence value: intact-miscore $\geq$ 0.6. Threshold for high confidence (https://doi.org/10.1093/database/bau131).	
SignaLink	16,484	324,688	Only if source is: SignaLink, ACSN, InnateDB, Signor, PhosphoSite, TheBiogrid, ComPPI, HPRD, IntAct, OmniPath.	
STRING	16,916	415,645	Combined score $\geq$ 700. Threshold for high confidence https://string-db.org/help/faq/#how-to-extract-high-confidence-07-interactions-from-information-on-combined-score-in-proteinlinkstxtgz	

Figure B.1. Huntington's Disease Enriched GO and HPO Terms. The bubbleplots display the top most enriched Gene Ontology (A, molecular functions; B, cellular components) and Human Phenotype Ontology (C) terms for Huntington's disease. On the horizontal axis, the fold enrichment is shown. The color encodes the negative of the false discovery rate logarithm, and the size represents the gene ratio (computed as the ratio of the percentage of genes in the study set related to a specific term, divided by the corresponding percentage in the background, i.e., the entire human proteome).

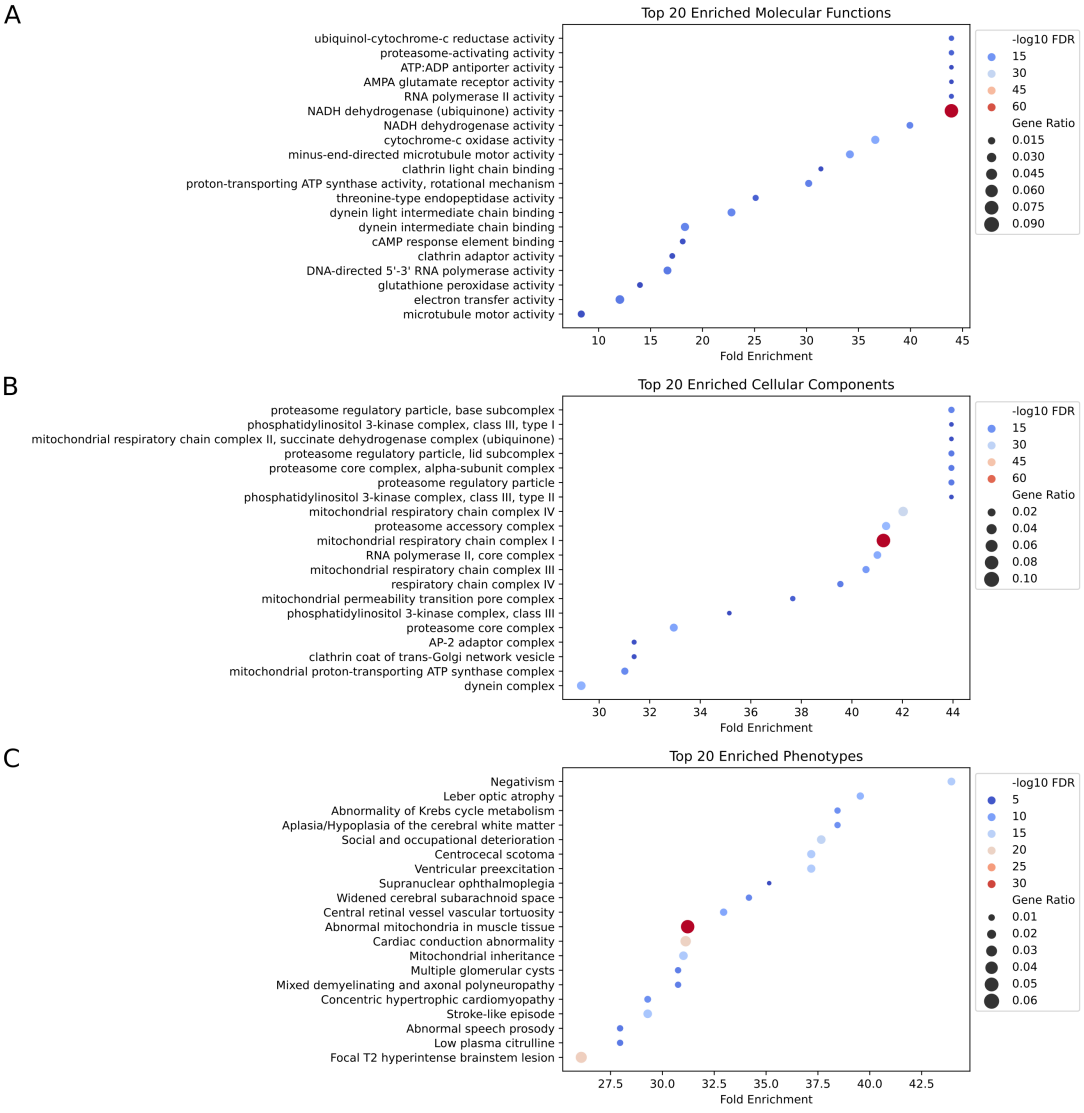


Figure B.2. Multiple Sclerosis Enriched GO and HPO Terms. The bubbleplots display the top most enriched Gene Ontology (A, molecular functions; B, cellular components) and Human Phenotype Ontology (C) terms for multiple sclerosis. On the horizontal axis, the fold enrichment is shown. The color encodes the negative of the false discovery rate logarithm, and the size represents the gene ratio (computed as the ratio of the percentage of genes in the study set related to a specific term, divided by the corresponding percentage in the background, i.e., the entire human proteome).

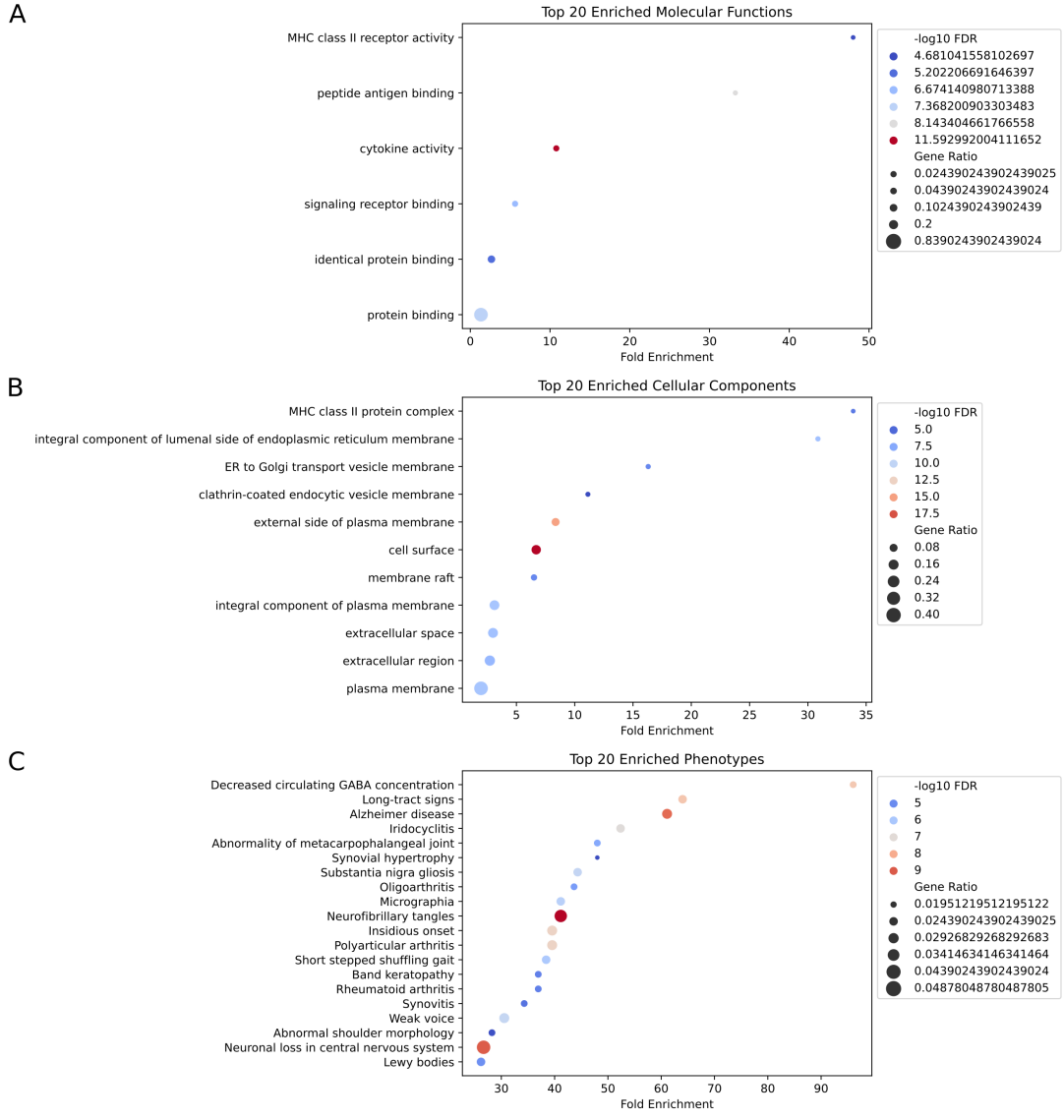


Figure B.3. Huntington's Disease Gene-Target-Drug Network. The Sankey diagram illustrates the interconnections between disease-related genes, drug targets, and drugs. Each drug (right column) is connected to its reported targets (middle column), which, in turn, are proximal on the human interactome to some of the disease-associated proteins (left column). Drugs are colored by the respective ATC code, and the FDR of the IGSEA analysis (see Section 7.1.2) is reported in the label.

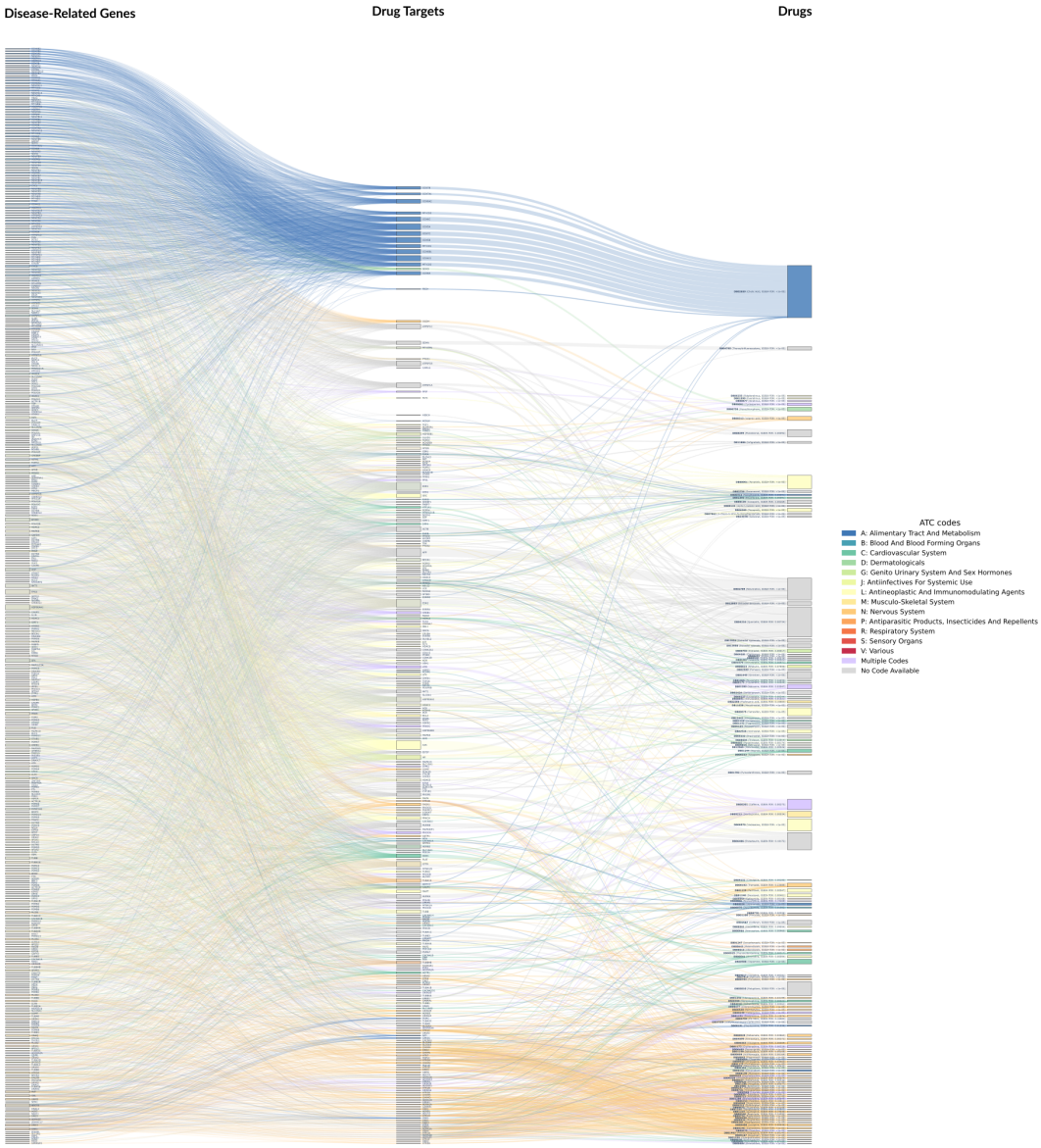
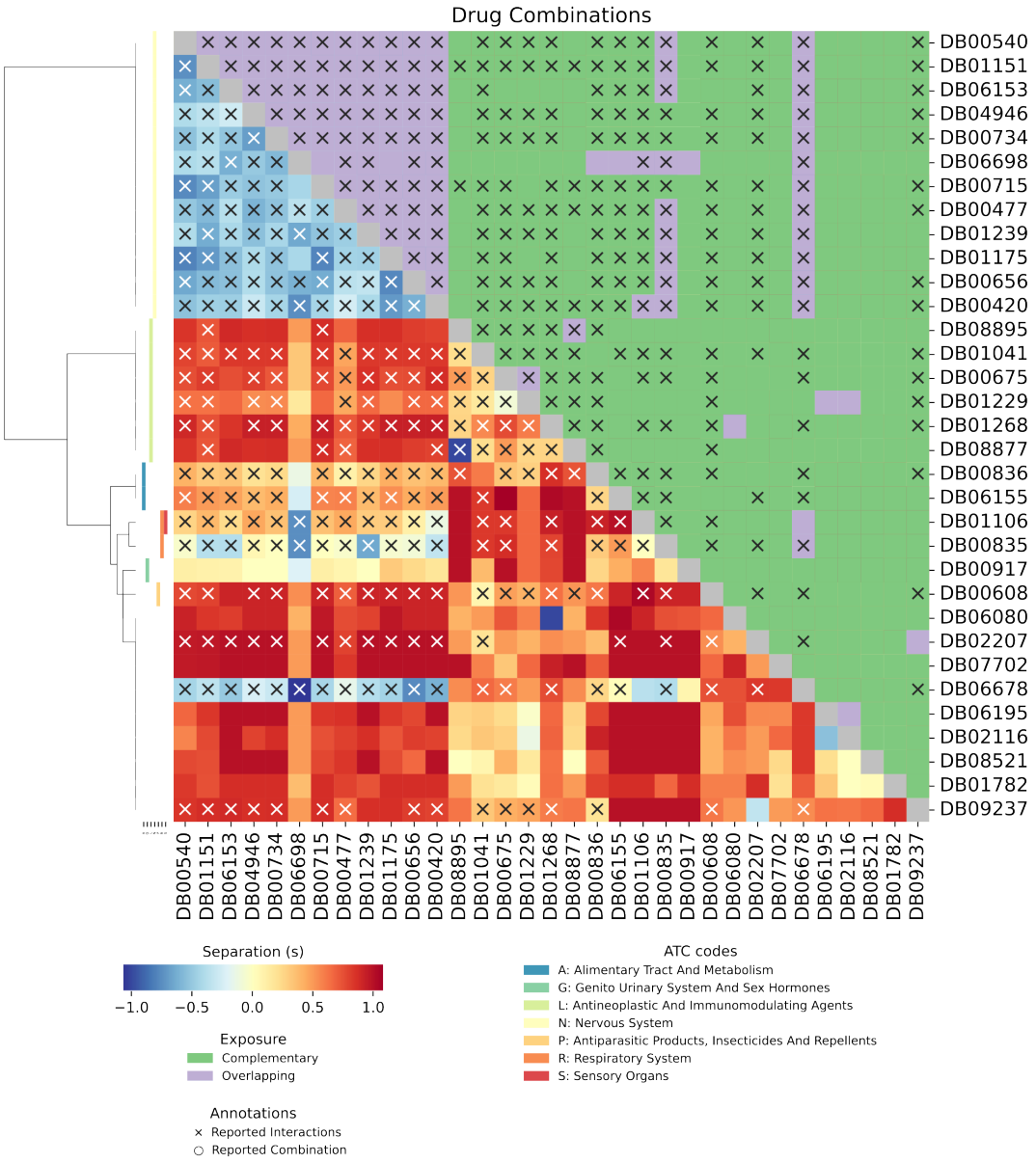


Figure B.4. Multiple Sclerosis Drug Combinations. The annotated heatmap provides info about possible combinations of the selected drugs. A combination is marked with × if an interaction is reported in DrugBank, and with ○ if it is present in an approved formulation. The lower-left part of the heatmap shows the separation of the inspected drugs, color coded from blue (no separation) to red (strongly separated). The upper-right portion, instead, displays the kind of exposure: violet if overlapping and green if complementary. At the leftmost part, the ATC codes of the drugs are reported along with a dendrogram of their hierarchical clustering.



C PATHOS and LOGOS

Table C.1. Node Types in PATHOS.

Type	# Nodes
protein	58908
biologicalProcess	27668
disease	23314
anatomicalEntity	14288
molecularFunction	11228
phenotype	8641
drug	8282
sequence	8067
proteinModification	4954
cellularComponent	4054
pathway	3968
proteinFamily	518
cell	226
proteinComplex	221
sequenceGroup	25
peptide	4
entityHavingProteicPart	1

Table C.2. Source Files for PATHOS.

	Source	License	File	Version
1	NCBI	Public Domain	`Homo_sapiens.gene_info.gz`	2023-07-04 (accessed: 2023-07-04)
2	APID	CC-BY-NC	`9606_Q1.txt`	accessed: 2023-07-04
3	BioGRID	MIT	`BIOGRID-ORGANISM-4.4.223.tab.zip`	4.4.223 (accessed: 2023-07-04)
4	HuRI	CC BY 4.0	`HuRI.tsv`	2020-03-09 (accessed: 2023-07-04)
5	InnateDB	DESIGN SCIENCE LICENSE	`all.mitab.gz`	2022-01-29 (accessed: 2023-07-04)
6	INstruct	All rights reserved (Authorization obtained by e-mail contact with Haiyuan Yu <haiyuan.yu@cornell.edu>)	`sapiens.sin`	2020-08-13 (accessed: 2021-10-05)
7	IntAct	CC-BY 4.0	`intact.zip`	2023-06-03 (accessed: 2023-07-04)
8	SignaLink	CC BY-NC-SA 3.0	`slk3db_dump_json.tgz`	2022-03-11 (accessed: 2023-07-04)
9	STRING	CC BY 4.0	`human.name_2_string.tsv.gz`	2019-01-27 (accessed: 2023-07-04)
10	STRING	CC BY 4.0	`9606.protein.links.full.v11.5.txt.gz`	2021-10-30 (accessed: 2023-07-04)
11	HPRD	Freely Available for non-commercial purposes	`HPRD_FLAT_FILES_041310.tar.gz`	2016-08-20 (accessed: 2022-05-30)
12	PINA	Freely Downloable All Rights Reserved (check with the develop team https://omics.bjcancer.org /pina2012/contact.do)	`Homo sapiens-20140521.tsv`	2014-10-27 (accessed: 2022-05-30)
13	DisGeNET	CC BY-NC-SA 4.0	`disease_mappings.tsv.gz`	2020-05-15 (accessed: 2023-07-04)
14	DisGeNET	CC BY-NC-SA 4.0	`curated_gene_disease_associations.tsv.gz`	2020-05-07 (accessed: 2023-07-04)
15	MONDO	CC BY 4.0	`mondo.obo`	2023-07-03 (accessed: 2023-07-04)
16	HPO	Freely Available (with conditions)	`hp.obo`	accessed: 2023-07-04
17	HPO	Freely Available (with conditions)	`phenotype.hpoa`	accessed: 2023-07-04
18	HPO	Freely Available (with conditions)	`genes_to_phenotype.txt`	accessed: 2023-07-04
19	DISEASES	CC BY 4.0	`human_disease_knowledge_filtered.tsv`	2023-07-02 (accessed: 2023-07-04)

20	UniProt	CC BY 4.0	`HUMAN_9606_idmapping.dat.gz`	2023-06-28 (accessed: 2023-07-04)
21	PathwayCommons	Freely Available, under the license terms of each contributing database (www.pathwaycommons .org/pc2/datasources)	`PathwayCommons12.All.uniprot.gmt.gz`	2019-09-18 (accessed: 2023-07-04)
22	HGNC	Freely Available	`gene_with_protein_product.txt`	2023-07-03 (accessed: 2023-07-04)
23	GO	CC BY 4.0	`goa_human.gaf.gz`	accessed: 2023-07-04
24	GO	CC BY 4.0	`go.obo`	accessed: 2023-07-04
25	PRO	CC BY 4.0	`pro_reasoned.obo`	68.0 (accessed: 2023-07-04)
26	Uberon	CC-BY 3.0	`human-view.obo`	accessed: 2023-07-04
27	Bgee	CC0 1.0	`Homo_sapiens_expr_simple.tsv.gz`	2021-02-15 (accessed: 2021-10-05)
28	DrugBank	CC BY-NC 4.0	`all-full-database`	5.1.10 (accessed: 2023-05-23)
29	DrugCentral	CC BY-SA 4.0	`drug2disease.tsv`	2023-05-10 (accessed: 2023-07-04)

LOGOS Hyperparameters

Batch Size : 256

Num Epochs : 100

Training Loop : sLCWA

Optimizer : Adam

- **Learning Rate** : 0.0001

Loss : NSSA

- **Adversarial Temperature** : 0.6868102318671975
- **Margin** : 50

Model : NodePiece

- **Aggregation** : MLP
- **Embedding Dimension** : 128
- **Entity Initializer** : Xavier Uniform
- **Interaction** : ComplEx
- **Number of Tokens** : 20, 5
- **Tokenizers** :
  - **Searcher** : ScipySparse
  - **Max Iter** : 100
  - **Selection** : MixtureAnchorSelection
  - **Number of Anchors** : 10,000
  - **Ratios** : 0.8, 0.2
  - **Selections** : Degree, Random

Negative Sampler : Bernoulli

- **Number of Negatives per Positive** : 100

Table C.3. First 50 Phenotypes Selected for Huntington's Disease.

	ID	Name	Train Set	Val Set	Test Set
1	HP:0030015	Female anorgasmia	✗	✗	✗
2	HP:0002072	Chorea	✓	✗	✗
3	HP:0003324	Generalized muscle weakness	✓	✗	✗
4	HP:0000716	Depression	✓	✗	✗
5	HP:0000741	Apathy	✓	✗	✗
6	HP:0002307	Drooling	✗	✗	✗
7	HP:0002340	Caudate atrophy	✓	✗	✗
8	HP:0002362	Shuffling gait	✗	✗	✗
9	HP:0002174	Postural tremor	✗	✗	✗
10	HP:0002529	Neuronal loss in central nervous system	✓	✗	✗
11	HP:0002460	Distal muscle weakness	✗	✗	✗
12	HP:0002071	Abnormality of extrapyramidal motor function	✗	✗	✗
13	HP:0001283	Bulbar palsy	✗	✗	✗
14	HP:0012332	Abnormal autonomic nervous system physiology	✗	✗	✗
15	HP:0001336	Myoclonus	✓	✗	✗
16	HP:0002151	Increased serum lactate	✗	✗	✗
17	HP:0001332	Dystonia	✓	✗	✗
18	HP:0002921	Abnormal cerebrospinal fluid morphology	✗	✗	✗
19	HP:0001288	Gait disturbance	✓	✗	✗
20	HP:0008652	Autonomic erectile dysfunction	✗	✗	✗
21	HP:0001260	Dysarthria	✗	✗	✗
22	HP:0003387	Decreased number of large peripheral myelinated nerve fibers	✗	✗	✗
23	HP:0006801	Hyperactive deep tendon reflexes	✗	✗	✗
24	HP:0000726	Dementia	✗	✗	✗
25	HP:0002063	Rigidity	✓	✗	✗
26	HP:0002922	Increased CSF protein concentration	✗	✗	✗
27	HP:0002197	Generalized-onset seizure	✗	✗	✗
28	HP:0030319	Weakness of facial musculature	✗	✗	✗

29	HP:0012751	Abnormal basal ganglia MRI signal intensity	✗	✗	✗
30	HP:0001251	Ataxia	✗	✗	✗
31	HP:0012416	Hypercapnia	✗	✗	✗
32	HP:0100021	Cerebral palsy	✗	✗	✗
33	HP:0000738	Hallucinations	✓	✗	✗
34	HP:0003394	Muscle spasm	✗	✗	✗
35	HP:0025331	Upgaze palsy	✗	✗	✗
36	HP:0007377	Abnormality of somatosensory evoked potentials	✗	✗	✗
37	HP:0012670	Orthostatic syncope	✗	✗	✗
38	HP:0001337	Tremor	✗	✗	✗
39	HP:0011289	EEG with temporal sharp slow waves	✗	✗	✗
40	HP:0001324	Muscle weakness	✗	✗	✗
41	HP:0002120	Cerebral cortical atrophy	✗	✗	✗
42	HP:0000737	Irritability	✓	✗	✗
43	HP:0009045	Exercise-induced rhabdomyolysis	✗	✗	✗
44	HP:0000488	Retinopathy	✗	✗	✗
45	HP:0002141	Gait imbalance	✓	✗	✗
46	HP:0000739	Anxiety	✓	✗	✗
47	HP:0000511	Vertical supranuclear gaze palsy	✗	✗	✗
48	HP:0000802	Impotence	✗	✗	✗
49	HP:0410263	Brain imaging abnormality	✗	✗	✗
50	HP:0040141	Tardive dyskinesia	✗	✗	✗

Table C.4. First 100 Proteins Related to Multiple Sclerosis.

	ID	Name	Train Set	Val Set	Test Set
1	TTR	transthyretin	✗	✗	✗
2	ALB	albumin	✗	✗	✗
3	TSNAX-DISC1	TSNAX-DISC1 readthrough (NMD candidate)	✗	✗	✗
4	MIR885	microRNA 885	✗	✗	✗
5	POMC	proopiomelanocortin	✓	✗	✗
6	SHBG	sex hormone binding globulin	✗	✗	✗

7	ADIPOQ	adiponectin, C1Q and collagen domain containing	✗	✗	✗
8	SLC10A2	solute carrier family 10 member 2	✗	✗	✗
9	CYP2D6	cytochrome P450 family 2 subfamily D member 6	✗	✗	✗
10	TF	transferrin	✗	✗	✗
11	MIR99A	microRNA 99a	✗	✗	✗
12	MIR346	microRNA 346	✗	✗	✗
13	CNR2	cannabinoid receptor 2	✗	✗	✗
14	MIR505	microRNA 505	✗	✗	✗
15	CYP2C8	cytochrome P450 family 2 subfamily C member 8	✗	✗	✗
16	TNF	tumor necrosis factor	✗	✗	✗
17	CP	ceruloplasmin	✗	✗	✗
18	CYP2B6	cytochrome P450 family 2 subfamily B member 6	✗	✗	✗
19	VEGFA	vascular endothelial growth factor A	✗	✗	✗
20	ACE2	angiotensin converting enzyme 2	✗	✗	✗
21	GSTM1	glutathione S-transferase mu 1	✗	✗	✗
22	IL6	interleukin 6	✗	✗	✗
23	RBP4	retinol binding protein 4	✗	✗	✗
24	MIR412	microRNA 412	✗	✗	✗
25	CYP2E1	cytochrome P450 family 2 subfamily E member 1	✗	✗	✗
26	TLR4	toll like receptor 4	✗	✗	✗
27	MIR433	microRNA 433	✗	✗	✗
28	SLC30A6	solute carrier family 30 member 6	✗	✗	✗
29	MIR766	microRNA 766	✗	✗	✗
30	MIR192	microRNA 192	✗	✗	✗
31	VKORC1	vitamin K epoxide reductase complex subunit 1	✗	✗	✗
32	MIR218-1	microRNA 218-1	✗	✗	✗
33	HMOX1	heme oxygenase 1	✗	✗	✗
34	DAOA	D-amino acid oxidase activator	✗	✗	✗
35	MTHFR	methylenetetrahydrofolate reductase	✗	✗	✗
36	TYK2	tyrosine kinase 2	✓	✗	✗

37	HLA-DQA2	major histocompatibility complex, class II, DQ alpha 2	X	X	X
38	GJB5	gap junction protein beta 5	X	X	X
39	MIR17	microRNA 17	X	X	X
40	IL10	interleukin 10	X	X	X
41	MIR98	microRNA 98	X	X	X
42	PTGS2	prostaglandin-endoperoxide synthase 2	X	X	X
43	PSCA	prostate stem cell antigen	X	X	X
44	EDN1	endothelin 1	X	X	X
45	IGHG1	immunoglobulin heavy constant gamma 1 (G1m marker)	X	X	X
46	EPHX2	epoxide hydrolase 2	X	X	X
47	FGF2	fibroblast growth factor 2	X	X	X
48	FAXDC2	fatty acid hydroxylase domain containing 2	X	X	X
49	TRH	thyrotropin releasing hormone	X	X	X
50	OXT	oxytocin/neurophysin I prepropeptide	X	X	X
51	SCN10A	sodium voltage-gated channel alpha subunit 10	X	X	X
52	ERGIC3	ERGIC and golgi 3	X	X	X
53	MIR296	microRNA 296	X	X	X
54	TFF2	trefoil factor 2	X	X	X
55	MIR3622B	microRNA 3622b	X	X	X
56	INS	insulin	X	X	X
57	LRP2	LDL receptor related protein 2	X	X	X
58	PLCG2	phospholipase C gamma 2	X	X	X
59	NGF	nerve growth factor	X	X	X
60	THBD	thrombomodulin	X	X	X
61	TLR6	toll like receptor 6	X	X	X
62	MIR30B	microRNA 30b	X	X	X
63	FCGR1A	Fc gamma receptor Ia	X	X	X
64	SCD	stearoyl-CoA desaturase	X	X	X
65	WT1	WT1 transcription factor	X	X	X
66	HLA-DPB1	major histocompatibility complex, class II, DP beta 1	X	X	X

67	MPO	myeloperoxidase	✗	✗	✗
68	GC	GC vitamin D binding protein	✗	✗	✗
69	SH2B3	SH2B adaptor protein 3	✗	✗	✗
70	IGF2	insulin like growth factor 2	✗	✗	✗
71	PRKCQ	protein kinase C theta	✗	✗	✗
72	IFNG	interferon gamma	✗	✗	✗
73	SLC22A1	solute carrier family 22 member 1	✗	✗	✗
74	PROC	protein C, inactivator of coagulation factors Va and VIIIa	✗	✗	✗
75	UGT1A1	UDP glucuronosyltransferase family 1 member A1	✗	✗	✗
76	FXYD6	FXYD domain containing ion transport regulator 6	✗	✗	✗
77	HP	haptoglobin	✗	✗	✗
78	SERPINA1	serpin family A member 1	✗	✗	✗
79	HLA-DRB1	major histocompatibility complex, class II, DR beta 1	✗	✗	✓
80	MIR218-2	microRNA 218-2	✗	✗	✗
81	TLR2	toll like receptor 2	✗	✗	✗
82	PPARG	peroxisome proliferator activated receptor gamma	✗	✗	✗
83	ZAP70	zeta chain of T cell receptor associated protein kinase 70	✗	✗	✗
84	UCN	urocortin	✗	✗	✗
85	CHRNB2	cholinergic receptor nicotinic beta 2 subunit	✗	✗	✗
86	MIR708	microRNA 708	✗	✗	✗
87	MSR1	macrophage scavenger receptor 1	✗	✗	✗
88	ATP1B2	ATPase Na+/K+ transporting subunit beta 2	✗	✗	✗
89	ABCB1	ATP binding cassette subfamily B member 1	✗	✗	✗
90	APOC3	apolipoprotein C3	✗	✗	✗
91	SLCO1B1	solute carrier organic anion transporter family member 1B1	✗	✗	✗
92	HELLPAR	HELLP associated long non-coding RNA	✗	✗	✗

93	MIR629	microRNA 629	X	X	X
94	LRP1	LDL receptor related protein 1	X	X	X
95	MS4A1	membrane spanning 4-domains A1	X	X	X
96	CCR2	C-C motif chemokine receptor 2	X	X	X
97	C2	complement C2	X	X	X
98	IFNB1	interferon beta 1	X	X	X
99	EPO	erythropoietin	X	X	X
100	RNASE3	ribonuclease A family member 3	X	X	X

Table C.5. First 10 Enriched Biological Processes.

	ID	Label	Fold Enrichment	FDR
1	GO:0002439	chronic inflammatory response to antigenic stimulus	242.26	0.0028
2	GO:0038124	toll-like receptor TLR6:TLR2 signaling pathway	242.26	0.0028
3	GO:1990268	response to gold nanoparticle	242.26	0.0028
4	GO:1903974	positive regulation of cellular response to macrophage colony-stimulating factor stimulus	242.26	0.0028
5	GO:0042496	detection of diacyl bacterial lipopeptide	242.26	0.0027
6	GO:0060557	positive regulation of vitamin D biosynthetic process	242.26	0.0027
7	GO:0017187	peptidyl-glutamic acid carboxylation	161.51	0.0042
8	GO:1904466	positive regulation of matrix metallopeptidase secretion	161.51	0.0042
9	GO:0002874	regulation of chronic inflammatory response to antigenic stimulus	161.51	0.0042
10	GO:0060559	positive regulation of calcidiol 1-monooxygenase activity	161.51	0.0042

Table C.6. First 10 Enriched Molecular Functions.

	ID	Label	Fold Enrichment	FDR
1	GO:0062188	anandamide 11,12 epoxidase activity	161.51	0.0239
2	GO:0062187	anandamide 8,9 epoxidase activity	161.51	0.0233
3	GO:0038177	death receptor agonist activity	121.12	0.0321
4	GO:0062189	anandamide 14,15 epoxidase activity	121.12	0.0313
5	GO:0061809	NAD+ nucleotidase, cyclic ADP-ribose generating	45.42	0.0112
6	GO:0050135	NAD(P)+ nucleosidase activity	45.42	0.0104
7	GO:0008392	arachidonic acid epoxygenase activity	45.42	0.0101
8	GO:0023026	MHC class II protein complex binding	35.89	0.0022
9	GO:0070330	aromatase activity	27.95	0.0310
10	GO:0005179	hormone activity	18.93	0.0000