Deep learning methods for Network Biology

Lorenzo Madeddu[1] and Giovanni Stilo[2]

madeddu@uniroma1.it, giovanni.stilo@univaq.it

Lives on earth are regulated by of complex system of interactions. Modelling those interactions through the network paradigms allows researchers to discover and understand the fundamental molecular mechanisms which drive the biological processes and lead to humans diseases. The advancement made in the development of sequencing technologies has produced a growing amount of biological data. The aforementioned preconditions are at the base of a flourishing production of Deep Learning methods able to cope with the complexity and the data abundance of this domain. For those reasons, this chapter provides a comprehensive overview of the recent advancement in the deep learning network-based approaches focusing on biology, medicine and pharmacological crucial research's problems. At first, the needed biological and network science backgrounds are presented. Secondly, a comprehensive overview of the biologicals' networks and resources are provided. Finally, we will discuss the most recent methods in the field organising them into three broad categories related to the Interactome, the Network Pharmacology, and the frontier biologicals problems.

[1]Sapienza University of Rome

[2]University of L'Aquila. Author's work is partially supported by Territori Aperti a project funded by Fondo Territori Lavoro e Conoscenza CGIL CISL UIL and by SoBigData-PlusPlus H2020-INFRAIA-2019-1 EU project, contract number 871042.

1. Introduction

The last decades are characterised by the advance of high-throughput technologies, such as the yeast two-hybrid screening and the next-generation sequencing. This advancement in technologies has boosted the creation of large 'omics' datasets (i.e. molecular data) which helped to reveals and understand the complex interconnections among all the subparts[a] of an organism.

In this context, researchers need new *"holistic"* tools to cope with the dimension and the complexity of the biological data. The response to this demand, as we will see, is the modelisation offered by the Networks formalism. Biological networks, such as protein-protein interaction network, are new representations of *"omics"* data and combine network science and biology approaches to analyse the interconnection of biological processes. The study of the structures and functions of the biological networks is known as Network Biology or Systems biology. The study of the pathogenic behaviour and drug processes in biological networks is referred to as Network Medicine. Systems biology and network medicine are the keys to: *i)* understand the biological mechanisms and *ii)* address challenges on both diagnostic and therapeutic aspects. The research literature has studied biological networks by using a plethora of graph-mining and classic machine learning approaches. However, diving in the complexity of the molecular interconnections across several levels of the organism's organisation is challenging. To overcome this limitations, researchers are adopting new powerful strategies in network biology and network medicine. Adopt Deep learning techniques in biology allow to explore the latent mechanisms and untangle the intricate molecular interconnections that standard approaches are not able to. In last years, promising deep learning methods in graph-mining, as Graph Convolutional Networks,[1] have been successfully applied to biological networks to solve problems as drug repurposing[2]and identify new disease genes.[3]

This chapter, discuss the deep learning network-based approaches applied to both network biology and network medicine. The chapter is organised as follow. Section 2 provides fundamentals knowledge about Networks' Theory (2.1), *Learning Problems on Networks* (2.2) and *Ground concepts of the System Biology* (2.3). The third section (3 is dedicated to describe in a formal way the most important *Biological Networks* (3.1) and the *Public available resources* (3.2). In section 4,5 and 6 the most recent Deep Learning based methods - organised by the tackled problem and the used data - are presented. Lastly, section 7 discusses the future directions and concluding remarks.

[a]In a modelisation of the organism as a system.

2. Background knowledge

In this section, it is presented a comprehensive summary of the key aspects related to network biology. More precisely, in section 2.1 the fundamental network concepts are briefly introduced. Section 2.2 describes the learning/prediction problems in network science. Finally, in section 2.3 the most important biological concepts are summarised.

2.1. *Networks background and formalisation*

In this sub-section we summarises the network concepts and properties most useful in network biology:

- Network: A simple network or graph is a mathematical formalism that allows describing in an abstract and concise way both the components of a system and their interactions. Formally, a graph $G = (V, E)$ is defined as a tuple containing the two sets V and E. V is the set of objects, called nodes or vertices of the graph. The set of edges $E \subseteq \{(u, v)|u, v \in V\}$ contains the relationships among the objects of the system. Edges - also called links or ties - can be either **directed** or **undirected**.

Directed edges are necessary to describe asymmetrical relationships. Having a relationship among nodes i and j does not imply having the same relationship between nodes j and i. On the other hands, the *undirected* case implies that the relationships in the system are symmetrical. It is not necessary to distinguish the relationship among nodes i and j from the one from node j to i. To collect differences among the *importance* of each relationship is useful to define a weighting function $w(e) : \quad E \to \mathbb{R}$, which returns values of the edges according to the semantics of the system. When a weighting function is defined, we call the graph a **weighted** one. Each simple graph (**binary**) is described by a square matrix A - called the adjacency matrix - of sizes $|V| \times |V|$, where its element A_{ij} contains 1 if exists an edge from vertex i to vertex j, or zero otherwise. When the weighting function $w(e)$ is defined, the elements A_{ij} of the adjacency matrix are equivalent to the values of the weighting function $w(i, j)$, or zero otherwise. A graphical representation of a simple binary undirected graph and of a weighted directed one is depicted in figure 1.

The **complex systems** or the heterogeneous networks - where different information is associated with the system components and several kinds of relationships exist - can be modelled in many ways. They can be modelled as a decorated graph with categorical and/or numerical attributes on both edges and vertices. To encode this information with the graph, several (one for each attribute) mapping functions are defined. Later on, we will refer to this kind of graph as one with **attributes**.

The **bipartite** graph contains two distinct kinds of vertices, and

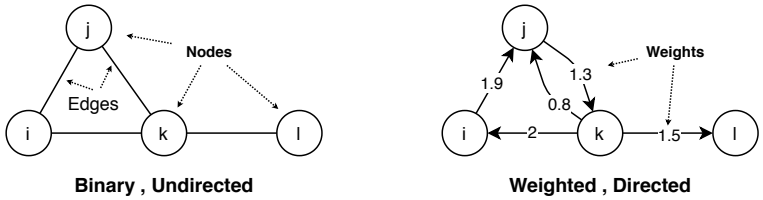

Fig. 1. Examples of several types of simple networks

no two vertices within the same set are adjacent. A bipartite graph $G = (V, U, E)$ is defined as a triple containing the three sets V, U and E. V and U are two disjoint sets of vertices: $V \cap U = \emptyset$. The set of edges $E \subseteq \{(u, v)| \quad (u \in V \cap v \in U) \cup (v \in V \cap u \in U)\}$ contains the relationships among the two different kinds of vertices. The bipartite graph is a special case of a k-partite graph where $k = 2$.

- Other important Networks' concepts and properties:

- **Degree**: the node's degree is the number of edges (incoming and/or outgoing in the directed case) of a node;
- **Path/Diameter**: a path, or walk, is a sequence of nodes in which every node is adjacent to the next one in the sequence. The diameter is the longest path between any pair of nodes;
- **Hub**: is a node which exposes a degree higher than the average degree of the network (normally it is a property analysed in the directed graphs considering only the out-degree);
- **Connected Graph**: a graph is strongly connected if exists a path for any pair of its nodes;
- **Induced Sub-graph**: an induced sub-graph H is a graph containing a node subset of V and all the edges among them. Let be G the initial graph; $H = (V_H \subseteq V, E_H = \{(u, v)|u, v \in V_H \quad and \quad (u, v) \in E\})$;
- **Motif**: a Motif is a recurrent and statistically significant partial sub-graph. Note that the motifs differ from graphlets since they can be partial sub-graphs, whereas motifs are induced sub-graphs;
- **Community**: the definition of *"community"* may change according to the application domain. From a network perspective, usually, a community is a locally dense sub-graph where the edges' density among the inside community's nodes is higher than the edges' density among the inside community's nodes and outside ones;
- **Random Network**: a network generated by a random distribution probability (e.g. Erdős–Rényi model[4]). A node $v \in V$ is connected to another node of the network following a certain probability p.

- **Scale-Free Network**: a network whose degree distribution follows the power-law distribution;[5]
- **Small-World Network**: a network in which any pair of nodes is connected to the other nodes of the network by a *small*, proportionally to $log(|V|)$, path.[6]

2.2. *Learning Problems on Networks*

Networks are mathematical tools used to study and predict properties in the domains' applications related, but not limited, to social behaviours, economic events, biology processes, traffic flow and internet connections. Despite the differences, every application on those domains is reducible to the same general networks' prediction tasks:

- **Node Classification:** Typically a class is used to group a collection of nodes that expose similar characteristics. The classes are typically referred by unique textual descriptions – namely labels or tags. Using nodes which are already classified, it is possible to train a classifier which captures some knowledge. A trained classifier can be used to assign a class to an unknown node. For instance, we would like to learn which proteins are related to a certain disease. The classification tasks can be either binary or multi-class (a given data instance may be assigned to one or many classes, respectively).
- **Link Prediction:** A fundamental problem with networks is that the link information in the graph may be of doubtful quality or not present at all. Thus, inferring the existences of edges between nodes has been referred to as link prediction.[7] Link prediction is a challenging problem that has been studied in various guises in different domains. The prediction of a link can be solely based on structural information (graph associations) or also on the attribute information.
- **Community Detection:** Community detection refers to the procedure of identifying groups (sub-graphs) of tightly interacting vertices (i.e., nodes) in the network.
- **Representation Learning:** Representation learning is about deriving a succinct representation of the input data (graph in our settings[8]). This succinct representation allows to achieve better performances (of the classification/prediction task) or to reduce the computational complexity. Exists many representation learning approaches: in deep learning, the representation is typically achieved by the composition of multiple non-linear transformations (hidden layers) of the input data.

2.3. *Ground concepts of the System Biology*

Hereafter we present the fundamental biological concepts necessary to understand the papers collected in this chapter. Note that in the network biology, the followings biological/molecular concepts are normally

organised in databases which collects this information in one place. To have a complete overview of these databases see section 3.2. The biological concepts are tightly interconnected. We decided to present at first the most important or general ones, further all the others. For this reason, we invite the reader to go through all this section to better understanding all the ground concepts.

- **Cells:** are building blocks of the structure of an organism. A cell is a structure containing DNA, cytoplasm, cellular structures (e.g. ribosomes) and several molecules (e.g. proteins), all surrounded by a membrane. The specialisation of a cell (in term of functionality) is determined by the gene expression. A collection of cells that work together to accomplish the same function is called tissue.

- **Biological Pathway:** is an ordered sequence of actions among molecules (e.g. proteins or complexes), in a cell that leads to the creation of product or change the cellular state. As example, a pathway could lead to the synthesis of a molecule, regulate genes, or spur a cell to move. The most common pathway types are metabolic pathways, gene regulation pathways, and signal transduction pathways. A metabolic pathway is a series of chemical reactions occurring within a cell. Gene regulation pathways turn genes on and off. The signal transduction pathway is a series of intracellular molecular events as a response to the activation of a cell's receptor by an extracellular signal. Enzymes, usually proteins, are involved in almost all metabolic pathways to accelerates chemical reactions.

- **Proteins:** are large molecules composed by amino acids. They are responsible for biological processes in the cell. The sequence of amino acids and the environment determine both the protein's three-dimensional structure and its specific function. The instructions to build the sequence of amino acids of a protein are stored in the genes. As a convention, proteins are classified by family and domain. A **protein family** is a group of proteins sharing a common evolutionary origin reflecting a related set of functions or/and structural similarities. **Domains** are a compact way to describe the three-dimensional structure of a protein in structural units(responsible for a particular function or interaction). Another important aspect related to proteins is their ontological organisation. The **Go-terms** are proteins/genes' ontological concepts organised into their three fundamental facets: molecular activities, cellular structures and biological processes.

- **Protein-Protein Interaction (PPI):** is a physical interaction which takes place in a cell among the structural region of two proteins. A PPI is responsible for biological processes and can be permanent, as in the case of the **protein complexes**[b], or can be transient as in the case of the signals. Protein-Protein Interactions are essential to understand how molecular biological processes are carried.

[b]Protein Complex (e.g. Haemoglobin) is an aggregation of proteins connected by protein-protein interactions.

- **Deoxyribonucleic acid (DNA):** *"is a complex and hereditary molecule that contains all of the information necessary to build and maintain an organism"*[c]. All organisms have a full copy of the DNA held within every cell. The DNA is composed of a chain of linked blocks, called nucleotides. Each nucleotide is characterised by one of the four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T)[d]. The ordered sequence - constituted by about 3 billion of base pairs - encodes the information necessary to make the essential molecules such as proteins. The DNA is a string of roughly 3 billion characters where the alphabet is {A, C, G, T}.
- **Gene:** is a section (contiguous subsequence) of the DNA that encodes information of a functional unit. The Human Genome Project[e] estimated that humans have between $20,000$ and $25,000$ genes. Mutations in the DNA cause diverse versions of the genes, namely *alleles*. The information stored in genes - as the rest of the DNA - contribute to regulating the mechanisms behind the proteins' synthesis.
- **Genotype and Phenotype:** the **genotype** is the collection of an organism's genes. The term **phenotype** refers to the observable traits of an organism. The phenotype is determined by the genotype, the gene expression and environmental factors.
- **Gene/Protein Nomenclature:** several committees are responsible for genes and proteins nomenclature. Despite the standardization effort, the multitude of nomenclatures cased ambiguities[f]. The HUGO Human Gene Nomenclature Committee (HGNC)[g] is responsible to assign names and symbols (shorter-form of the names) to the humans' genes. Alternatively, the most popular nomenclatures in bioinformatics are the Entrez ID and the Ensembl ID[h]. UniprotKB, complete the nomenclatures' overview, providing names also for protein isoforms. In bioinformatics, Entrez and UniportKB are the most used nomenclatures for a programmatic access. The literature generally adopt the HGNC names or acronyms.
- **Gene Expression:** is the cellular process of synthesizing proteins or functional RNA. Gene expression is composed of two **main** steps: transcription and translation. The transcription is the process of transferring the information stored in a gene to the cytoplasm through the messenger-RNA (mRNA) which promote the translation process. During the translation process, the readed mRNA sequence in collaboration with the transfer-RNA (tRNA) allows to assembles the protein.
- **Gene Regulation:** is the process of controlling the gene expression profile in a cell. Gene regulation increase, decrease or suppress the expression of

[c] https://www.nature.com/scitable/topicpage/introduction-what-is-dna-6579978/
[d] https://ghr.nlm.nih.gov/primer/basics/dna
[e] https://www.genome.gov/
[f] https://www.uniprot.org/help/different_protein_gene_names
[g] https://www.genenames.org/
[h] https://m.ensembl.org/info/genome/genebuild/gene_names.html

genes. Since only a fraction of genes are expressed ("turned on"), the gene expression profile defines the type and the activities of a cell at a certain time. Mechanisms of gene expression control can occur at any step of the gene expression but most commonly during the transcription process. *Transcription factors and microRNAs are key regulators of the gene expressions.* The **Transcription Factors (TF)** regulate the gene transcription (increase or decrease the amount of gene product of a gene). The **microRNAs (miRNAs)** regulate the gene expression at the post-transcription level by binding to target mRNAs preventing their expression. **Alternative splicing** is a regulation mechanism that enables the coding of multiple proteins (called protein isoforms), instead of one. Lastly, the long non-coding RNAs (lncRNAs), still under investigation, are involved in several molecular functions playing a key role in some cancer mechanisms.

- **Disease:** a general definition states the disease as any harmful deviation from the normal structural or functional state of an organism. The abnormal condition generates disorders in the whole organism or any of its parts. Generally, the disease is associated with certain signs and symptoms and differs in nature from physical injury. The disease process is the result of complex dysfunctional mechanisms. As an example, a disease could alter the normal pathways causing patho-phenotypes[i].

- **Virus:** is a submicroscopic infectious agent that can multiply only within the living cells of an organism. Consequently, a virus uses the chemical machinery of the living cells to continue to live and to reproduce itself. *Host specific* viruses cause many common human infections and are also responsible for several diseases. As examples, the common cold, caused by one of the rhinoviruses, and the AIDS, caused by HIV, are virus. Viruses may have their genetic material (DNA or RNA).

- **Gene-disease association:** (or simply a disease gene) it happens when a gene or a gene product is involved in the diseases process. A mutated gene which changes its expression may consequently change the produced proteins (and/or its quantity), thus, drafting the basis of the dysfunctional process. As example, a mutation in *TP53* lead to several tumoral diseases such as breast cancer, bladder cancer and lung cancer[j].

- **Disease Module:** some diseases are caused by more than one faulty gene product (disease gene or disease protein). Barabási et al.[9] have moved the focus from the disease genes to their complex set of interactions: *"a disease is rarely a consequence of an abnormality in a single gene product but reflects the perturbations of the complex intracellular network"*. The **disease module** is defined as the local neighbourhood of genes which linked to the same disease interact together and alter the normal structure and the biological functions.

[i] https://www.genome.gov/about-genomics/fact-sheets/Biological-Pathways-Fact-Sheet

[j] https://ghr.nlm.nih.gov/gene/TP53#conditions

- **Drug:** in pharmacogenomics, a drug is a chemical compound which by interacting with specific molecules of the organism (e.g. enzymes or receptors proteins) produces a therapeutic effects. A drug which binds to a receptor (a protein family) change its molecular structure and its functionalities. This change induces a signalling pathway within the cell or inhibiting the functions of the receptor itself. The specific drug-induced alterations of the biological processes contrast the disease-induced ones and thus causing a therapeutic *"fixing"* effect.

  We like to remark that in precision medicine[k] is fundamental to select those set of drugs which contrast the disease-altered biological processes and have a minimal or negligible impact on the other patient biological processes.

3. Biological Networks and public available resources

In this section, we present the most common biologicals' networks (3.1) and an exhaustive collection of biologicals' databases (3.2).

3.1. *Biological networks*

Nodes of a molecular network, usually represent genes, gene products or biological functions while edges define the molecular connections between these entities. In the following, we describe the most common types of biological networks providing a full explanation and a tabular description:

3.1.1. *Protein-Protein Interaction Network (PIN)*

Protein-Protein Interaction Network (PIN), often referred as interactome, is an undirected binary graph of Protein-Protein Interactions (PPIs) where proteins are nodes and undirected edges are physical interactions. Contrary to the common convention the PINs present an ambiguity to keep in mind. The absence of an edge may imply that does not exist the interaction between the two proteins or that the related lab test was not performed yet. Additionally, PIN's nodes can be decorated with categorical attributes representing biological knowledge such as GO terms, diseases relationships and other proteins features as domains. From the topological point of view, PINs have characteristics similar to those of scale-free[10] networks[l]. As reported by Jeong et al.,[11] in PINs, the hubs are likely to be proteins which are essential for many fundamental life processes. Another typical property of the PINs is its modular organization.[10]

[k]Precision Medicine is an emergent approach to patient care that allows clinicians to select treatments that are most likely to help patients based on a genetic understanding of their disease.

[l]Scale-free networks are characterised also by the presence of large hubs.

PINs are composed of modules, subgraphs of proteins that tend to collaborate together in order to accomplish a biological function. Those modules play a key role in the life of an organism by enabling biological processes and by constituting protein complexes.

The PINs are mainly used to identify key proteins or modules, essential for the development of biological processes like diseases.[9] Discovering such modules in the human PIN helps to understand how diverse phenotypes might be linked at the molecular level. PIN and modules are the key-concepts which allows clinicians to investigate on co-morbidity and on drug repurposing.

We like to remark that, notwithstanding in the past decade were witnessed systematic efforts to increase its coverage and accuracy, the human PIN remains highly incomplete and noisy.[12] For the aforementioned reasons, it is still an active and open research field.

Nodes:	Protein	**Edges:**	Physical interactions
Type:	Undirected	**Weights:**	Binary
Nodes Attributes:	Pathways, diseases, domains, families, tissues		
Edges Attributes:	Usually not present		

3.1.2. *Drug-Target Network (DTN)*

A drug-target network is a bipartite undirected network where one set of nodes are composed by drugs and the other set contains their targets molecules. An edge/interaction is present when a drug has the tendency to bind to a target,[13] a protein (e.g. GPCR) peptide or nucleic acid. Furthermore, drugs can target several molecules at once causing also adverse side effects to a patient.[14]

Since one of the main use of DTN is in the field of drug repurposing, many networks report also categorical attributes. These attributes constitute additional knowledge as disease, proteins' domains, and drugs' pathway or categories. The DTN networks can be also expanded with protein-protein and drug-drug interactions. The drug repurposing approaches leverage enriched DTNs - with the functional and modular structures of the PINs - to achieve better performances.[2,14,15]

Another kind of network related to the DTN is the *drug-drug interactions (DDIs) network*. In DDIs network, nodes are drugs, and the edges represent a change, often adverse, in the effect that one drug has on the target if combined with the other drug. For this reason, DDIs are usually employed to identify drugs' side effects and drive polypharmacy therapies. Similarly to diseases, it has been observed that several drugs impact on the PIN neighbourhood of their targets. For example, a drug can be engineered to produce beneficial effects by interacting with proteins in the PIN neighbourhood of the diseases proteins.[15]

Finally, the concept of drug-target module, hypothesised by Cheng et. al.,[14] captures the tendency of proteins targeted by the same drug to form a localised neighbourhood as it happens in the case of the diseases modules. An interesting property observed in polypharmacy is: two drugs have a therapeutic effect only if the drug-target modules overlap with the disease module but no among them.

Nodes:	Drugs, Proteins	**Edges:**	Physical/Functional interactions
Type:	Undirected	**Weights:**	Binary
Nodes Attributes:	Diseases, protein features, drug features		
Edges Attributes:	Side effects		

3.1.3. *Gene Expression Network (GEN)*

A gene co-expression network is an undirected, often weighted, correlation network where nodes are genes and edges represent significant correlation in the expression between two genes. Gene nodes can be decorated with categorical attributes adding biological knowledge as diseases and GO-terms. GENs are typically constructed in two steps: in the first step, a correlation measure (Pearson or Spearman coefficients, Mutual Information, Euclidean Distance) is computed between each pair of genes, by using their expression data (Microarrays, RNA-Seq). In the second step, a significance threshold (a threshold cutoff or the Fisher's Z-transformation) is applied over the previously computed correlation values in order to identify the co-expressed genes.

Studying GENs permits to identify functionally related genes or co-expression module with key roles in biological pathways.[16]

We want also to highlight as, conversely to the PPIs networks, the gene co-expression ones, depend on the conditions (typically the temporal context) in which the samples have been collected.

Nodes:	Genes	**Edges:**	Correlations of gene expression
Type:	Undirected	**Weights:**	Real values
Nodes Attributes:	GO-Terms, diseases		
Edges Attributes:	Usually not present		

3.1.4. *Gene Regulatory Network (GRN)*

A gene regulatory network is a weighted, directed, bipartite network of gene regulatory dependencies. The interactions are between a *"regulator molecule"* (Transcription Factor; RNA; miRNA etc.) and a *"regulated molecule"*, usually a gene. Nodes can either be characterised by other categorical attributes. The GRN captures the complex work-flow of the

gene regulation system. Studying the gene regulatory process allows the researchers to understand how molecular mechanisms work, and thus, identifying the key patterns and players which emerge in specific conditions like a disease state.[17]

Nodes:	TFs, RNAs, miRNAs, lncRNA, Genes	**Edges:**	Biological interactions
Type:	Directed	**Weights:**	Real values
Nodes Attributes:	GO-Terms, diseases		
Edges Attributes:	Usually not present		

3.1.5. *Brain Network*

The connectome, also known as the brain network, is used to organise and represent information about functional or physical connectivity among different brain regions.[18] As often was observed in several biological networks, also in the brain one, it is possible to recognise the scale-free topology.[19] Other studies have also found out that the connectome exposes some similarities with the small-world network model[20] and have identified some recurrent structural motifs.[21] Moreover, a brain network exposes modular structures related to cerebral functions[22] as[m] those that typically are present in the PINs' networks.

To understand how the connectome was built, we must consider the different techniques employed, to detect the brain's zones and understand how they interact one to another. The two most popular ones are i) The structural Magnetic Resonance Imaging (sMRI or MRI) which detects anatomical structures and provides a map of physical neural connections and ii) The functional Magnetic Resonance Imaging (fMRI) which allows to tracks the oxygen changes associated with blood flow helping to build the brain's activity map.

The analysis of brain networks led to great advancement in the research areas related to degenerative brain-related diseases. As an example, works[23] related to the study of Alzheimer's Disease, reports evidence that the functional brain networks have structural properties that are less similar than those of the small-world model. That discovery allows to a better understanding that Alzheimer causes a disconnection among the distant brain zones.

Nodes:	Brain regions	**Edges:**	Physical connections
Type:	Undirected	**Weights:**	Binary
Nodes Attributes:	Usually not present		
Edges Attributes:	Usually not present		

[m] From the structural perspective.

3.2. *Public available resources*

Data collections (databases) constitute a grounding block for deep learning methods. The network biology databases reflect the biological/molecular information of the organism, presented in section 2.3, by centralising them in one place. The databases can be classified in: Primary databases are based on experimental results submitted by researchers. They contain primary information (as, nucleotide or protein) and annotations regarding, bibliographies, function, cross-references to other databases, and so forth. Secondary databases summarise the results from analyses (e.g. computational predictions) of primary databases. These analyses allow finding common features of biological classes, which can be used to classify unknown biological data. Databases that contain biological or medical information, are usually classified as secondary ones. However, many databases have integrated both data from primary and secondary databases over the years and, thus, their classification has grown uncertain.

It should be noted that the same biological information can be collected in several databases maintained by different national or transnational institutions: the National Institutes of Health (NIH) for the U.S. maintains the PubChem database and the European Bioinformatics Institute (EBI) for Europe maintain the corresponding one ChEMBL.

In adjunction to the original article (containing its data description, structure and functionality of the database) exist dedicated journal issues (e.g. The journal Nucleic Acids Research[n]) which yearly reviews all the recently updated and available biological databases.

We present in table 1 an exhaustive list of the publicly available databases. The table reports for each database its name, a description of the contained data and three binary columns which identify the primary databases and classify their content as features (attributes on nodes or edges) or/and as associations (edges). The direct link to the original resource is provided, but for readability propose it is available only in the electronic format of the book. The table is organised by groups of databases [o]: gene/proteins, diseases, drugs, gene expression, gene regulation and pathways related data. Following the authors' expertise and the recognised use in of the databases. The most important databases have been underlined[p].

[n]Freely accessible also on the Web

[o]Note that a database can be present in more than one group.

[p]Due to the several databases' versions released during years keeping track of citations in a precise way is a hard task. Thus, the underline must be considered advice.

Database	Content description	Primary	Features	Associations	Website
Gene/Protein Data					
Uniprot[24]	proteins' additional information as domains, families and sequences	-	yes	-	link
Gene Ontology (GO)[25]	collection of ontological terms (Go-Term) related to gene and gene products	-	yes	-	link
Human Protein Atlas[26]	human proteins' distribution in cells, tissues, organs and cancer types	-	yes	yes	link
BioGRID[27]	large collection of protein protein interactions (PPI)	yes	-	yes	link
IntAct[28]	large collection of protein protein interactions (PPI)	yes	-	yes	link
DIP[29]	collection of protein protein interactions (PPI)	-	-	yes	link
HPRD[30]	collection of human only protein protein interactions (PPI)	yes	-	yes	link
HuRI[31]	database of highly reliable PPIs	yes	-	yes	link
APID[32]	a database which contains PPIs collected from other several primary databases	-	-	yes	link
HIPPIE[33]	collection of protein protein interactions (PPI)	-	-	yes	link
CLAIRE-COVID	protein related data for COVID-19 tasks	-	yes	yes	link
Disease Data					
ICD[34]	international classification of Diseases	-	yes	-	link
SNOMED[35]	modern disease taxonomy	-	yes	-	link
MeSH[36]	taxonomy of medical literature	-	yes	-	link
UMLS[37]	mapping among several medical terminologies	-	yes	-	link
DO[38]	diseases' ontology and vocabularies with biomedical data	-	yes	-	link
HPO[39]	a phenotype ontology: contains symptom-disease associations	-	yes	yes	link
MalaCards[40]	data-mining based human disease knowledge-base	-	yes	yes	link
Orphanet[41]	European source for disease related data (i.e. symptom-disease associations)	-	yes	yes	link
DisGeNET[42]	large GDA collection: integrates multiple expert-curated and computational sources	-	-	yes	link
OMIM[43]	diseases' vocabulary and expert-curated GDAs' collection	yes	-	yes	link
ClinVar[26]	collection of association between human variations and phenotypes	-	-	yes	link
PsyGeNET[44]	collection of GDAs for psychiatric diseases	-	-	yes	link
HuGE Navigator[45]	text-mining based GDA	-	-	yes	link
COSMIC[46]	collection of cancer related GDAs	-	-	yes	link
CTD[47]	literature-curated associations between drugs, genes and diseases	-	-	yes	link
GWAS Catalog[48]	collection of GDA for gene variants from Genome-WAS	yes	-	yes	link
GWASdb[49]	collection of GDA for gene variants from Genome-WAS	yes	-	yes	link
PheWAS[50]	collection of GDA for gene variants from Phenome-WAS	yes	-	yes	link
CLAIRE-COVID	disease related data for COVID-19 tasks	-	yes	yes	link
Drug Data					
DrugBank[51]	pharmaceutical knowledge-base. Useful for DTIs/DIs/DDAs and drug side effects	-	yes	yes	link
SIDER[52]	drug-side effect association collection	-	yes	yes	link
ChEMBL[53]	European manually-curated drug knowledge-base. Useful for DTIs	yes	yes	yes	link
PubChem[54]	American manually-curated drug knowledge-base. Useful for DTIs	yes	yes	yes	link
ChEBI[55]	collection and ontology of small chemical compound (e.g. drugs)	-	yes	-	link
TTD[56]	database of drug targets, targeted disease and pathway information	-	yes	yes	link
OFFSIDES[57]	database of drug-side effects associations	-	yes	yes	link
STITCH[58]	collection of experimental-validated and computational predicted DTIs	-	-	yes	link
SuperPred[59]	database of experimental-validated and computational predicted DTIs	-	-	yes	link
DGIdb[60]	collection of DTIs collected from several sources	-	-	yes	link
TWOSIDES[57]	collection of DTIs	-	-	yes	link
BindingDB[61]	collection of DTIs with affinity measurements	yes	-	yes	link
PharmGKB[62]	collection of DTIs with genetic variants	-	-	yes	link
CTD[47]	database of literature-curated associations between drugs, genes and diseases (i.e. DDAs)	-	-	yes	link
SupertTarget[63]	database of DTIs with information about side-effects, pathways and Gene Ontology terms	-	yes	yes	link
Matador[63]	collection of physical and functional, or indirect, DTIs	-	-	yes	link
TDR targets[64]	collection of drugs-targets resource for neglected human diseases	-	-	yes	link
DCDB[65]	database of drugs combinations (e.g. DIs)	-	-	yes	link
RepoDB[66]	collection of repurposed drugs with indications (e.g. DDAs)	-	-	yes	link
CLAIRE-COVID	drug related data for COVID-19 tasks	-	yes	yes	link
Gene Expression Data					
GEO[67]	American public collection of gene expressions data	yes	yes	-	link
ArrayExpress[68]	European public collection of gene expression data	yes	yes	-	link
TCGA[69]	repository of gene expression profiles associated with cancer	yes	yes	-	link
Gene Regulation Data					
TransFAC[70]	database of regulatory interactions	-	yes	yes	link
GTRD[71]	database of transcription factor binding sites	-	yes	-	link
TRRUST[72]	manually curated collection of human and mouse regulatory networks	-	-	yes	link
miRTarBase[73]	collection of miRNA-gene associations	-	-	yes	link
miRWalk[74]	database of miRNA-gene associations	-	-	yes	link
miR2Disease[75]	collection of miRNA-disease associations	-	-	yes	link
HMDD[76]	collection of miRNA-disease associations	-	-	yes	link
miRCancer[77]	database of miRNA-cancer associations	-	-	yes	link
PhenomiR[78]	miRNA knowledge-base and collection of miRNA-disease associations	-	yes	yes	link
dbDEMC[79]	miRNA-expression profiles in human cancers	-	yes	-	link
TCGA[69]	collection of miRNA data in cancers	-	yes	-	link
lncRNASNP2[80]	database of miRNA-lncRNA associations	-	-	yes	link
LncRNADisease[81]	database of lncRNA-disease associations	-	-	yes	link
LncRNA2Target[82]	collection of lncRNA-protein associations	-	-	yes	link
Pathways Data					
Reactome[83]	European-hosted pathway database	-	yes	-	link
KEGG[84]	Japanese-hosted pathway database	-	yes	-	link
WikiPathways[85]	pathway database based on collaborative platform	-	yes	-	link
MSigDB[86]	database integrating pathways from other resources	-	yes	-	link
Pathway Commons[87]	database integrating pathways from other resources	-	yes	-	link

4. Deep Learning for Interactome (I)

4.1. *PPI prediction (PPIP)*

Proteins own a key role in the biological processes. Having a complete collection of PPIs of an organism is crucial for the many molecular networks' studies (e.g. the Disease Gene Identification or the Drug-Target Association Identification). Notwithstanding the time-consuming and labour-intensive dedicated to the built of PPI detection systems, it is still necessary to put efforts to complete the PPI maps of several organisms. The protein-protein interaction identification task may helps to detect physical interactions which were not previously present among the proteins of an organism. The protein-protein interaction network (see Section 3.1.1), the Proteins, Gene, Gene Expression and Cells are useful concepts presented in section 2.3.

Three main experimental approaches are used to detect human protein-protein interactions: systematic experiments, literature curation, computational predictions. Systematic experiments, as Yeast two-Hybrid (Y2H) and Affinity-Purification with Mass-Spectrometry (AP-MS), are the most reliable approaches which provide diverse types of PPIs: Y2H detects binary interactions and AP-MS detects one-to-many (complexes) interactions.[12]

Notwithstanding their accuracy, systematic approaches are prone to identify false positives and false negatives. As noted by,[88] PPIs collections built on a scientific literature review are richer but exposes a *lower in quality* since the adopted methods are error-prone and can be affected by investigation biases. Computational approaches, on the other hand, are inexpensive - if compared to the laboratory experiments - but due to their synthetic nature, the produced predictions must be validated with biological experiments. Several approaches were proposed in recent years. The oldest, but also one of the most used one as a comparison, is node2vec: a skip-gram based approach - proposed in 2016 by Grover and Leskovec[89] - which learn the nodes network representation (embeddings) leveraging only structural information. Kishan et al.[90] (GNE) and Luo et al.[91] are based on a classic DNN approach. The most recent ones[92–94] are based on GNN. HO-VGAE, proposed by Xiao and Deng,[92] is also based on graph variational auto-encoder (GAE). Here the aim of the authors is to predict PPIs by improving the GCN's aggregation scheme in the GAE to explore higher-order neighbourhoods of each node in the Human PIN. Moreover, they integrate L3 principle, a recently discovered property of the Human PIN.[95] In their comparison evaluation on the Human's PIN, HO-VGAE outperforms node2vec[89] by 3.4% AUPRC. Similarly to the HO-VGAE, SkipGNN[93] improves the aggregation scheme to collect information from direct and second-order neighbours. Has been shown that similarity in second-order PPIs can be highly predictive of PPIs (i.e. L3 principle).[95] In their compara-

tive evaluation on the Human PIN, SkipGNN outperforms node2vec[89] by 14.8% AUPRC and 15.1% AUROC. Similarly to the HO-VGAE and SkipGNN, HOGCN[94] improve the aggregation scheme to collect information from direct k-order neighbours. In their comparative evaluation on the Human PIN, HOGCN outperforms node2vec[89] by 15.7% AUPRC and 15.6% AUROC and SkipGNN by 0.9% AUPRC and 0.5% AUROC.

All the aforementioned approaches use solely the PPIs as input data with the exception of the method proposed by Kishan et al.[90] (GNE) which also use gene expressions to build the genes' representation by integrating both the topological structure of the PIN and the gene expression. In their comparison evaluation on the yeast's PIN, GNE outperforms node2vec by 6% in AUROC and 12% in AUPRC. Unfortunately, these studies lack a common benchmark which permits a straight comparison. Each study, even if aligned from the point of view of the adopted evaluation measures, uses a PPI dataset often different from the others works. The most used evaluation measures are the AUROC and the AUPRC, an exception is made by Grover and Leskovec[89] which uses solely the F-1 measure. For an additional overview of this subfield of studies please see the works of Lü and Zhou,[7] Kovács et al.[95]

4.2. *Essential Protein prediction (EPP)*

Essential genes or proteins are molecular components performing key biological processes for the growth and survival of an organism. Furthermore, essential genes tend to be highly conserved in the evolutionary path of common species. For this reason, essential proteins are employed in the gene-disease associations' discovery, drug development of antibiotics, synthetic biology and to assess the minimum set of the essential genes needed for the survival of an organism (i.e. essentialome).[96–98] The traditional biological discovery path used to find essential genes relies on time-consuming experiments as gene knockout, RNA interference, antisense RNA (asRNA) and transposon mutagenesis.[96,98] From a biological network perspective, the essential genes tend to be hub nodes of the PIN network.[11] Since the PIN tends to be a scale-free network, the deletion of essential genes leads to the disruption of its connectivity and, consequently, produces instability in the organism.[99] However, if on the one hand essential proteins are strongly related to network topology, on the other hand, the noise and the incompleteness of the PINs limit the performance of classic network-based prediction methods.[96,97] Several computational approaches, based on deep neural network, were proposed in recent years. The majority of them are based on the node2vec in conjunction with other techniques as LSTM, CNN and GNN. M. Zeng et al.[100] propose an end-to-end model, based on node2vec and LSTM to identify essential proteins by jointly integrate *PIN*, *gene expression* and information on *subcellular localization*. The produced classification of the essential proteins was tested on Yeast data using the F-measure,

AUC, and AP measures. A similar approach was explored by M. Zeng et al.[101] which uses a CNN instead of the LSTM and by X. Zhang et al.[102] which concatenate node2vec proteins' vectors and sequence data to feed a fully connected multi-layer neural network (DNN). We like to note that X. Zhang et al. improves the overall performance of the EPP by exploiting protein sequence and topological features and by addressing the imbalanced learning problem by using a cost-sensitive training function. EPGAT, proposed by J. Schapke et al.,[97] is the most different one both in terms of method and used data. EPGAT construct an attributed *PIN* network to apply a method based on a graph attention neural networks where they used a weighted binary cross-entropy (CE) function. The nodes of the PIN network are decorated with a feature vector based on *gene expression* profiles, *orthology information*, and *subcellular localization*. Even if the aforementioned methods are exhaustively evaluated using several measures (e.g. Accuracy, Recall, Precision, AUROC and AUPRC) they do not present a direct comparison with deep learning techniques. The main baseline methods used for comparison are the degree centrality and the Support Vector Machine but due to the use of different PIN data, it is not possible to highlight a common benchmark. For an additional overview of this subfield of studies please look at the works of Li et al.[96] and Zhang et al.[98]

4.3. *Protein Function Prediction (PFP)*

As already stated, proteins carrying out critical functions of an organism. However, the functions of almost all the proteins are largely unknown. According to Shehu et al.,[103] less than 1% of the proteins have reliable and detailed annotations in the Universal Protein (UniProt) database. Moreover, *"fundamental information is currently missing for 40% of the protein sequences deposited in the National Center for Biotechnology Information (NCBI) database"*. For the aforementioned reason and due to the growing gap between the number of proteins being discovered and their functional characterisation (in particular as a result of experimental limitations), completing the collection of the proteins' function has become a fundamental research problem to address. The first effort made by researchers to overcome this problem was to organise the proteins' functions in a structured and standardised way. The Gene Ontology (GO)[q] is a collection of three protein's functions hierarchical ontologies distinct by the biological aspect: Cellular Component, Biological Processes and Molecular Functions. Secondly, proposing a reliable prediction of protein function through computational methods has become crucial. In this direction, GraphSAGE, proposed by W. Hamilton et al.,[104] presents an inductive framework that leverages node feature information to learn node embeddings. In this work, the authors use a PIN network only to prove the efficacy of their work. In 2018, M. Kulmanov

[q]See section 3.2 for additional details.

et al. propose DeepGO[105] a complex end-to-end deep learning method which exploits protein sequence features, topological cross-species PIN features and dependencies among GO classes to predict the proteins' function. The method is based on the concatenation of two Deep Neural Networks which combines the proteins' sequences and network's representation respectively. DeepNF, by Gligorijević et al.,[106] propose a network fusion method based on Multimodal Deep Autoencoder (MDA) to extract high-level features of genes from multiple heterogeneous interaction networks (six different Yeast protein-protein networks). The method uses the Random Walk with Restart (RWR) to build a high dimensional node embedding which, once it is reduced using a Multimodal Autoencoder (MDA), it is used into an SVM classifier to predict the proteins' function. Lastly, K. Fan et al. proposes Graph2GO,[107] a model to predict protein functions (GO-terms) exploiting a *protein sequence similarity network* and a PIN with node attributes (amino acid sequence, subcellular location, and protein domains). The model uses two variational graph auto-encoders (vGAEs) to learn latent representations for each protein and successively predict proteins' functions with a DNN. Graph2GO improves by 10.5% and 3.33% the micro-AUPRC and macro-AUPRC respectively when compared to deepNF. All works in the area use a combination of PPIs with attributes and normally prefer to use the precision, the recall and the F-measure as comparison metrics. To complete the overview of this subfield of studies it is also possible to look at the works of Shehu et al.,[103] R. Bonetta and G. Valentino.[108] The protein-protein interaction network (see Section 3.1.1) and the Proteins, PPI, Gene and Cells concepts (see Section 2.3) can be helpful concepts to better understand this section.

4.4. *Gene-Disease association Prediction (GDAP)*

The disease-gene identification consists of finding the gene or gene product involved in the origin of a genetic disease (i.e. disease gene). Traditional ways to assess the role of genes in diseases involve time-consuming and expensive[r] analysis such as Linkage Analysis and Genome-Wide association studies (GWAS).[49]

Genome-Wide association studies (GWAS) have led to large collections of disease-gene associations available in public databases like OMIM[43] and DisGeNet.[42] The identification of disease genes provides useful insights to understand disease mechanisms, design new therapies, improve disease prevention approaches and make an accurate risk factors evaluation. For this reason, computational methods in this area of research have become more and more prominent and widely proposed.

The computational approaches that laying in this research area are mainly based on GNN and Skip-Gram using PPI and GDA data with en-

[r] https://www.genome.gov/27541954/dna-sequencing-costs-data/

riched features. Agrawal et al.[109] uses node2vec and a logistic-regression to study how latent network structures of the Human PIN are correlated with the disease modules predictability. In HerGePred[110] (HDGN) a heterogeneous network is built by combining four types of relationships: *disease-gene*, *disease-symptom*, *gene-GO terms*, gene-gene (*PPI*). Then it is applied node2vec to learn the nodes embeddings. Finally, two distinct disease-gene rankings are produced leveraging the cosine-similarity and an RWR based approach. Similarly to DeepWalk and node2vec, SmuDGE[111] perform random walks on heterogeneous networks (composed by PPI, disease-phenotypes and gene-phenotypes associations) and to then apply the Skip-Gram model to learn the diseases' and genes' representations. SmuDGE use two independent methods: i) an unsupervised method based on the cosine-similarity between embedding vectors and the query disease vector; ii) a method based on a deep neural network which uses the genes' and diseases' embeddings to perform the prediction. Another study based on node2vec is the work proposed by Ata et al.[112] where they enrich the generated embeddings with around 500 *gene features* collected from Uniprot. The enriched representation is used to train several binary classification models: SVM, generalized linear model (GLM), Random Forest (RF) and kNN. According to the authors, N2VKO outperforms plain node2vec by roughly 2.6% AUROC on six diseases. In HNEEM,[113]the authors construct a heterogeneous network based on *gene-disease* associations, *gene-chemical* associations and *disease-chemical* associations. Several models are then applied to learn different node embeddings: node2vec, DeepWalk, Higher-Order Proximity Preserved Embedding (HOPE), semi-supervised depth model Structural Deep Network Embedding (SDNE),[114] Graph Factorization (GF)[115] and Laplacian Eigenmaps (LE).[116] Finally, they predict the disease-gene association using a random forest based approach. The method presented by Zhu et al.[117] uses a cascade of deep methods (DeepWalk with a graph convolution layer and a three-layers fully connected network) to predict the disease-gene associations. The presented approach use a heterogeneous network that integrates a gene-gene network, disease-disease network and gene-disease network to outperforms the prediction made by HerGePred by 11.7% AUPRC. HeteWalk[118] performs the Skip-Gram model on meta-path controlled random walks which explore a weighted heterogeneous network of PPIs, miRNA-miRNA, disease-disease, gene-disease, gene-miRNA, miRNA-disease associations. Given a query disease, HeteWalk ranks genes according to the cosine similarity between their representations and the given disease vector. GCN-MF[119] extracts disease and genes' features using the principal component analysis (PCA) and matrix factorisation methods on prior biological knowledge of diseases and genes. These features are then used to build a similarity network which in conjunction with the genes' features are used to train a GCN model. A slightly different approach dgMDL is presented by Luo et al.[120] The method, based on a

multi-modal deep belief net (DBN) first constructs a gene-gene similarity network and a disease similarity network based on PPIs and (GO-)terms data respectively. Then, node2vec is applied to those networks to generate their latent representations. The generated embeddings are finally jointly used into a two-level DBNs. One of the most recent approach is the Random Watcher Walker RW^2.[121] Madeddu et al. proposes an unsupervised deep learning method that exploits both functional and connectivity patterns to predict the gene-disease associations. To do so, RW^2 collects second-order random-walk made over the gene attributes of the humans' PIN to learn, a Skip-Gram based attributes representations. The final prediction is made using attributes' representation which encompasses both structural and functional connectivity. The proposed framework allows to integrate multiple sources of information without manipulating the original network topology producing promising results. CIPHER-SC[122] uses an approach based on a Graph Convolution on a Context-Aware Network with Single-Cell Data. The heterogeneous network is composed by *ontological associations* and *biological relationships* from both a public databases and single-cell data. The node2vec nodes' representations are used into a deep neural network made with one GCN and one DistMult[123] layer. Authors compared CIPHER-SC compared to methods present in literature (HerGePred, SmuDGE and GCN-MF) achieving an increment up to 8.02% in the AUROC and 17.01% in the AUPRC. Finally, HO-VGAE[94] and SkipGNN,[93] described in Section 4.1, improves the GNN's aggregation scheme to predict gene-disease associations. In their comparative evaluation on the GDAP problem, SkipGNN outperforms node2vec[89] by 8.7% AUPRC and 7.8% AUROC. HOGCN outperforms SkipGNN by 2.6% AUPRC and 2.4% AUROC.

It is worth noting that also this prediction problem is characterised by the absence of a sharp benchmark. All the methods can be distinguished mainly by the used data rather than the computational approach. Despite that node2vec, AUROC and AUPRC represent the defacto baseline and measures to compare with. To complete the overview of this subfield of studies it is also possible to look at the works of Luo et al.,[124] Kaushal et al.[125]

5. Deep Learning for Network Pharmacology (NP)

5.1. *Drug-Target Interaction Prediction (DTIP)*

The drug-target association prediction is the task that consists in finding a molecule, usually a protein, which is bounded to a drug. Exist several wet-lab experiments to assess drug-target associations but they are extremely expensive and time-consuming.[126] The identification of drug-target interactions is crucial for drug discovery and drug repurposing and thus to design new therapies and develop Precision Medicine.

The most notable methods which rely on deep neural network tech-

niques are the ones proposed by Zong et al.[127] and Wan et al.[128]

The first method[127] is a similarity-based drug–target one which first constructs a heterogeneous network of *drug–target*, *drug-disease* and *disease–target* associations. Then, the DeepWalk method is applied to the network to build nodes' representation. Authors propose two rule-based inference methods to predict new drug-target associations: a drug-based similarity inference (DBSI) and a target-based similarity inference (TBSI) ones. The former, DBSI, predicts a new drug–target association (d_i, t_l) if the drug d_i is similar to the drug d_j and exist an association among the drug d_j and the target t_l. On the other hand, TBSI predicts a new drug–target association (d_i, t_h) if exist an association among the drug d_i and the target t_l and the target t_l is similar to target t_h. The proposed method mainly take advantage of the topological structure of biological networks to improve the predictions' performances.

NeoDTI[128] is an end-to-end model to predict drug-target interactions from heterogeneous data. The method, first, constructs a network composed of other eight ones: drug-drug structure-based similarity, drug-side effect, drug-target, drug-drug, drug-disease, protein-protein sequence-based similarity, protein-disease and PIN. The proposed deep learning neural network framework takes in input the heterogeneous network to reconstruct the original eight networks adjacency matrices. Finally, NeoDTI predicts drug-target interactions relying on the reconstructed network matrices.

Lastly, SkipGNN[93] and HO-VGAE,[94] described in Section 4.1, improves the GNN's aggregation scheme to predict drug-drug interactions. In their comparative evaluation on the DTIP problem, SkipGNN outperforms node2vec[89] by 15.7% on AUPRC and 20.2% on AUROC. HOGCN outperforms SkipGNN by 0.9% on AUPRC and 1.2% on AUROC.

The main advantages of the proposed methods are the integration of several sources of information to extract non-linear patterns from the data. AUROC and AUPR are the most used metrics in this research field. For an additional overview of this subfield of studies please look at the works of Ezzat et al.,[126] Bagherian et al.[129] and Abbasi et al.[130]

5.2. *Drug-Disease Association Prediction (DDAP)*

A drug-disease association (also named Drug Repurposing) is a synthetic representation of the therapeutic effect which a drug has on a certain disease. It is important to note that a disease is a complex process which has an impact on an organism by modifying its biological processes and thus must be distinguished from the drug-target interaction. Note that the drug discovery process, where a new drug is developed from scratch, is different from the drug repurposing one. Drug discovery is a time-consuming and expensive process while the repurposing process of an existing drug may drastically reduce costs and time of drugs' validation especially if it is addressed by computational approaches.

The first solution to DDAP was proposed by Zeng et al. with DeepDR.[131] DeepDR, is a network-based deep-learning approach for drug repurposing which uses several biological networks: *drug-disease, drug-side-effect, drug–target* and seven *drug-drug* networks. At first, DeepDR generates the low-dimensional representations capturing highly non-linear patterns (drugs' embeddings) by using the DeepNF model (presented in subsection 4.3) on 9 drug-related networks. Then, the learned drugs' embeddings and drug-disease associations are used into a *collective variational autoencoder*[132] to predict novel drug-disease associations. Similarly the approach of Gysi et al.[2] learn diseases' and genes' representations applying the Decagon model (presented in subsection 5.3) on four associations networks (*protein-protein, drug-target, disease-protein, and drug-disease associations*) to predict the drug-COVID19 associations. Lastly, Karimi et al. present a reinforcement learning-based approach, HVGAE.[133] The method learns diseases' and drugs' embeddings using a *hierarchical variational graph auto-encoder* with attentional pooling on PPI, gene-disease and disease-disease networks. The learned representations are used into a reinforcement learning model to identify the drug combinations which maximise the therapeutic efficacy. This work addresses the challenging problem of finding clinical indications for a set of drugs instead of repurposing of a single drug as in the aforementioned works.

The most used metrics in this domain are the AUROC and AUPRC. Unfortunately it is not possible to recognise a common benchmark. An additional overview of this domain is presented in the works of Xue et al.[134] and Jarada et al.[135]

5.3. *Drug-Drug Interaction Prediction (DDIP)*

The combined interactions of two or more drugs with individual biological processes may cause unexpected and critical health complications.[136] Those complications, called adverse drug reactions (ADRs), are dangerous for the patient and expensive for the health system. A significative number of hospital admissions and medical errors are due by DDI.[137]

Notwithstanding the high demanding for improving our understanding of DDIs, the current known side effects of those interactions are less than 1% of the total[138] and their prediction by clinical and wet-lab experiments are extremely expensive and hard to carry out.[139]

For the aforementioned reason, Drug-Drug Interaction detection is a relevant task necessary for the success of patients' treatments. Several computational classic machine learning approaches have been developed,[139,140] the only three methods that can be highlighted as deep based ones are Decagon,[136] SkipGNN[93] and HO-VGAE.[94]

The authors of Decagon presents an *end-to-end multimodal graph auto-encoder* approach for predicting drug-drug associations and their side effects. The input of the model is a multimodal (heterogeneous)

graph composed by *protein-protein interactions, drug-protein target interactions and polypharmacy side effects*. The drugs are the nodes and each side effect is an edge of a different type. The method is based on a novel convolutional neural network for multi-relational link prediction. Decagon addresses the problem of polypharmacy to find side effects of drug combinations. A different task to the one of finding a combined clinical treatment as done by the work of Karimi et al.[133] (presented in details in subsection 5.2).

HO-VGAE[94] and SkipGNN,[93] presented in Section 4.1, improves the GNN's aggregation scheme to predict drug-drug interactions. In their comparative evaluation on the DDIP problem, SkipGNN outperforms node2vec[89] by 6.5% on AUPRC and 7.7% on AUROC. HOGCN outperforms SkipGNN by 3.1% on AUPRC and 2.5% on AUROC.

6. Deep Learning for other biological problems (BIO)

6.1. *miRNA-disease association prediction (MDAP)*

The miRNA-disease association prediction (MDA) is the task of identifying the interactions between microRNA (miRNA) and a disease. miRNAs play a key role in gene regulation with an important impact on biological processes and disease mechanisms. Several studies have shown the usefulness of miRNA-disease associations for personalised diagnosis and drug development.[141] Biological methods to assess miRNA-diseases associations are reverse transcription-polymerase chain reaction, northern blotting and micro-array profiling.[142] However, these experiments as often happen in the biological domain, are expensive and time-consuming. For this reason, applying computational methods, mainly based on CNN and GNN, will benefit the identification of new associations.

Xuang et al. present two works, namely CNNMDA[143] and CNNDMP,[144] both based on a CNN. To overcome the limitation of classic computational methods of the area, the authors embed a higher number of miRNA-diseases associations. The two methods mainly differ from the input data processing step needed by CNN. To extract network representations, CNNMDA uses the non-negative matrix factorization (NMF) and CNNDMP uses RWR. MDA-CNN[145] extracts network-based features by applying an auto-encoder to a three-layer complex network which includes disease similarity network, miRNA similarity network and protein-protein interaction network. MDA-CNN predicts the miRNA-disease associations with a CNN which uses the learned low-dimensional representations as input. The authors of HGCNMDA[141] propose a method based on a heterogeneous network of *human PPIs, miRNA-disease, miRNA-gene, disease-gene* associations. The method first generates the genes' embeddings by applying node2vec to the human's PINs. Then, it uses a *graph convolutional layer* for every network in conjunction with the learned node embeddings. Finally, HGCNMDA

averages the resulting GCN's representations to predict miRNA-disease associations in a link prediction setting. The first step of NIMCGCN[146] is based on the construction of miRNA-disease similarity networks. A graph convolutional networks (GCN) method is used to learn miRNA and disease latent representations. Lastly, the representations are used by a neural inductive matrix completion (NIMC) model to generate an association matrix which allows predict miRNA-disease associations. Another GCN based method (FCGCNMDA) is presented by Li et al.[147] It first constructs a complete network in which nodes represent miRNA-disease pairs. Then the miRNA-disease network along with a feature matrix (i.e. miRNA-disease association scores as node weights) are used by a two-layer graph convolutional networks (GCN) to predict new miRNA-disease associations. Lastly, GAMEDA[142] is a graph auto-encoder based model which first, collect the associations between miRNAs using a bipartite graph. The projected (in the same vector space) miRNAs' and diseases' nodes are then processed by a graph auto-encoder (GAE) to learn dense representations. Finally, GAEMDA uses the miRNAs' and diseases' embeddings with a *bilinear decoder* to reconstruct and predict new miRNA-disease associations.

The reference metrics used by these works are the AUROC and AUPRC. Generally, the works are not going self compare with the other methods in the same field of research. MDAP, as a new area of research, suffers by the absence of common benchmarks and the availability of comprehensive literature reviews.

6.2. *Disease Analysis (DA)*

The discovery of the mechanisms of a complex disease, such as cancer or tumours, is dependent by the underlying interconnected molecular heterogeneous processes. In this context of analysis, GDAP methods, based solely on the analysis of the organism's PIN, generally fail to identify the condition-specific disease drivers. In order to identify the drivers specific of a disease subtype or a patient, several works successfully integrate gene expression profiles with network biology. This integration helps clinicians to make a better diagnosis and select a patient personalised treatment. In this section, we discuss recent works for the analysis of specific disease-related conditions using gene expression data and networks.

Rhee et al.[148] solve the patient classification problem proposing a method based on *relation network (RN)* and *graph convolution neural network (graph CNN)*. The method captures localised patterns of associating genes with graph CNN, and then learn the relationship between these patterns with the RN. The proposed framework is applied to the human PIN and gene expression profiles of patients with breast cancer level 3 to classify the subtype of breast cancer according to the PAM50 scheme. The work proposed by Schulte et al.[149] tackles the problem of

combining multiple omics data types into a single learning model. Their model uses gene expression data for cancer gene prediction by applying a graph convolutional network (GCN). The used attributed PIN is composed of nodes which have features vector extracted from genes' expression profiles in specific cancer condition. DeepDriver[150] uses a Deep CNN applied to a gene-gene similarity network and a gene features vector derived from *gene expression* data. Authors focused their approach on cancer specific genes. iSOM-GSN[151] use the self-organizing maps (SOMs) algorithm to generate a gene-gene similarity network from gene expression data. Then, it enriches the adjacency matrix by integrating each gene with features values of gene expression, DNA methylation and copy number alteration (CNA). Finally, they use the enriched obtained data into CNN to predict the disease states.

Due to the recent introduction of this research field and the limited amount of works present in the area is difficult to clearly identify a common set of evaluation measures and a common benchmark.

6.3. *Brain Analysis (BA)*

The brain network (i.e. connectome see section 3.1.5) analysis is a new emergent field of study. Typically it is used to understand the mechanisms of diseases as schizophrenia, depression, Alzheimer and multiple sclerosis.[152] The connectome is characterised by complex interdependencies between brain regions. The relation between brain network structures and their functional roles is partially known. For this reason, applying computational approaches becomes necessary to improve their current understanding.

Two deep neural network-based methods can be identified in this field. The first one, proposed by Rosenthal et al.,[152] uses node2vec to learn the low-dimensional network representations of brain regions to study their latent relationships. The latter, by Lee et al.,[153] addresses the problem of analysing the natural organisation of the brain networks. The method uses a Graph Auto-Encoder (GAE), with non-negative weight constraints, to a *"structural"* brain network to learn low-dimensional representations of the nodes (brain zones). The non-negative weight constraint in the GAE is the most innovative contribution proposed by this method whom adds interpretability capabilities to the model.

We like to note that this filed of research is characterised by custom qualitative examinations rather than a clear benchmark and a precise set of evaluation measures.

7. Conclusion and future works

This chapter summarised the concepts, datasets, and techniques used by deep learning applications in network biology. A complete overview

of presented works in respect to the used base methods, data and evaluation measures are summarised in table 7. The reviewed works have shown that now it is possible to diving in the biological networks' complexity in a more detailed way. Despite the impressive results achieved by deep learning techniques, future works will need to cope with data and methods issues.

On the one hand, even if recent advances in high-throughput technologies have produced a huge and growing quantity of biological data. This data are affected by quality and reliability problems:
- Incompleteness: the biological networks are characterised by high incompleteness. As shown in table 7, the most used network, with its attributes, is the Human PIN known for only its 20%.[12] This data incompleteness strongly limits deep learning methods' performances.
- Bias: biological knowledge suffers from the study bias. Several PPIs collections are strongly biased towards the most studied genes leading to network structures that are not representative of the topology of the complete Human PIN.[9] Bias in biological networks is a critical issue because it influences the quality of the patterns extracted by network-based methods; hence, their performance. Recently, Luck et al.[31] presented a Human PIN obtained by a systematically and unbiased proteome-wide study.
- Noise: both literature-based data and high-throughput technologies are prone to generate false positives and false negatives.[92,154] For this reason, biological interaction datasets must be completed by an evidence/reliability scores.[28,42]
- Lack of negative knowledge: is common practice in the biomedical literature (due to practical and economical reasons), to do not discuss the biological entities which are not interacting. The absence of this *"negative"* results create uncertainty about the absence of the interaction or the lack of knowledge about it.[121,154] Machine learning methods achieve better performances if they can learn both from positive and negative samples. To solve this issue, negative knowledge is randomly sampled or generated with biological heuristics. However, these generation strategies may force to consider a not known positive interaction as a negative one, leading to lower quality of the model and by overestimating the performances.

On the other hand, the future development of the computational methods in network biology must face the following open challenges:
- Heterogeneous data: heterogeneous data in network biology can be both nodes' and edges' features or whole additional networks. In the literature, given the complex and interconnected nature of biomedical entities and the aforementioned lack of knowledge in the biological datasets, there is no agreement on how heterogeneous data must be appropriately handled. As we have seen, several methods can differ solely on the techniques used to tackle this aspect. Modelling and integrating heterogeneous information, even if it is a difficult task, will be the key

strategy to achieve better results.
- Imbalance learning: problems in network biology are usually extremely imbalanced by their nature (e.g. PPIP). If not properly handled, this imbalance will affect the performances by producing overestimated evaluations.
- Biological heuristics: Several graph deep learning methods in network biology are inspired by techniques developed for social networks. However, biological networks rely on different topological patterns. Recent studies[93] are facing this challenge by integrating the L3 principle[95] rather than the social homophily.[89]

To conclude, graph deep learning approaches in network biology represents a powerful tool to unchain the hidden biology, medicine and pharmacology knowledge.

Problems		Deep Learning Methods: Classic: DNN	Classic: CNN	Classic: LSTM	Auto-Encoder (AE): AE	Auto-Encoder (AE): vAE	Auto-Encoder (AE): MDA	Auto-Encoder (AE): GAE	Skip-Gram Based: DeepWalk	Skip-Gram Based: Node2Vec	Others	GNN
I	PPIP	90,91						92		89		92–94
	EPP	102	100		101					100–102		97
	PFP	105	105				106					104
	GDAP	111,117,120							113,117	109,110,112,113,120,122	110,111,113,118,121	93,94,117,119,122
NP	DTIP	128							127			93,94
	DDAP					131	131	2,133				2
	DDIP							136				93,94,136
BIO	MDAP		143–145		145			142		141		141,146,147
	DA	150	150,151									148,149
	BA							153		152		

Problems		Data Types: Gene/Protein: PPI	Gene/Protein: Features	Drug: DTI	Drug: DDI	Drug: DI	Drug: Features	Disease: GDA	Disease: Features	Gene Expression	Gene Regulation	Pathways	Brain
I	PPIP	89–94								90			
	EPP	97,100–102	97,100,102							97,100,101			
	PFP	104–106	104–106										
	GDAP	109–112,117–122	110–112,119,120,122	113	113			93,109–113,117–120,122	110,111,117,119–122		118		
NP	DTIP	128	128	127,128	93,94,127,128	128	128	127,128					
	DDAP	2,133	133	2,131	2,131,133	131	131	2,133	133			133	
	DDIP	136		136		93,94,136	136						
BIO	MDAP	141						141	141–147		141–147		
	DA	148,149						148,150		148–151			
	BA												152,153

Problems		Classic measures: ACC.	Classic measures: RC.	Classic measures: PR.	Classic measures: F-1	Classic measures: MCC	RC@k	PR@k	Ranking/Thresholding measures: F-1@K	Ranking/Thresholding measures: AP	Ranking/Thresholding measures: NDCG	Ranking/Thresholding measures: ROC	PRC
I	PPIP	100–102	100–102	100–102	89			92				90,91,93,94	90–94
	EPP	100–102	100–102	100–102	100,101					101,102		97,100–102	100–102
	PFP	106			104,106	105	105,107		105,107	105		105,107	106
	GDAP	113	113	113	113	113	109,110,117,119,121	110,117,119	110,117	110,117,119	119	93,94,111–113,118–120,122	93,94,113,122
NP	DTIP						127					93,94,127,128	93,94,128
	DDAP						131		133	2,133		2,131,133	2,131
	DDIP									136		93,94,136	93,94,136
BIO	MDAP	141,142,146	142,145	142,145	142,145		143,144					141–147	141,143–145,147
	DA	148,151	151	151	148,151							149–151	149

References

1. T. N. Kipf and M. Welling, Semi-supervised classification with graph convolutional networks, *arXiv:1609.02907* (2016).
2. D. M. Gysi, Í. D. Valle, M. Zitnik, A. Ameli, X. Gan, O. Varol, H. Sanchez, R. M. Baron, D. Ghiassian, J. Loscalzo, et al., Network medicine framework for identifying drug repurposing opportunities for covid-19, *arXiv:2004.07229* (2020).
3. L. Madeddu, G. Stilo, and P. Velardi. Predicting disease genes for complex diseases using random watcher-walker. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing* (2020).
4. P. Erdős and A. Rényi, On the evolution of random graphs, *Publ. Math. Inst. Hung. Acad. Sci.* **5**(1) (1960).
5. A.-L. Barabási and R. Albert, Emergence of scaling in random networks, *science*. **286**(5439) (1999).
6. D. J. Watts and S. H. Strogatz, Collective dynamics of 'small-world'networks, *nature*. **393**(6684) (1998).
7. L. Lü and T. Zhou, Link prediction in complex networks: A survey, *Physica A: statistical mechanics and its applications*. **390**(6) (2011).
8. P. Goyal and E. Ferrara, Graph embedding techniques, applications, and performance: A survey, *Knowledge-Based Systems*. **151** (2018).
9. A.-L. Barabási, N. Gulbahce, and J. Loscalzo, Network medicine: a network-based approach to human disease, *Nature reviews genetics*. **12**(1) (2011).
10. A.-L. Barabasi and Z. N. Oltvai, Network biology: understanding the cell's functional organization, *Nature reviews genetics*. **5**(2) (2004).
11. H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai, Lethality and centrality in protein networks, *Nature*. **411**(6833) (2001).
12. K. Venkatesan, J.-F. Rual, A. Vazquez, U. Stelzl, I. Lemmens, T. Hirozane-Kishikawa, T. Hao, M. Zenkner, X. Xin, K.-I. Goh, et al., An empirical framework for binary interactome mapping, *Nature methods*. **6**(1) (2009).
13. M. AY, K.-I. Goh, M. E. Cusick, A.-L. Barabasi, M. Vidal, et al., Drug–target network, *Nature biotechnology*. **25**(10) (2007).
14. F. Cheng, I. A. Kovács, and A.-L. Barabási, Network-based prediction of drug combinations, *Nature communications*. **10**(1) (2019).
15. F. Cheng, R. J. Desai, D. E. Handy, R. Wang, S. Schneeweiss, A.-L. Barabási, and J. Loscalzo, Network-based approach to prediction and population-based validation of in silico drug repurposing, *Nature communications*. **9**(1) (2018).
16. P. Langfelder and S. Horvath, Eigengene networks for studying the relationships between co-expression modules, *BMC systems biology*. **1**(1) (2007).
17. K. Glass, J. Quackenbush, D. Spentzos, B. Haibe-Kains, and G.-C. Yuan, A network model for angiogenesis in ovarian cancer, *BMC bioinformatics*. **16**(1) (2015).
18. O. Sporns, *Networks of the Brain*. MIT press (2010).
19. V. M. Eguiluz, D. R. Chialvo, G. A. Cecchi, M. Baliki, and A. V. Apkarian, Scale-free brain functional networks, *Physical review letters*. **94**(1) (2005).

20. D. S. Bassett and E. Bullmore, Small-world brain networks, *The neuroscientist.* **12**(6) (2006).
21. O. Sporns and R. Kötter, Motifs in brain networks, *PLoS Biol.* **2**(11) (2004).
22. M. A. Bertolero, B. T. Yeo, and M. D'Esposito, The modular and integrative functional architecture of the human brain, *Proceedings of the National Academy of Sciences.* **112**(49) (2015).
23. E. J. Sanz-Arigita, M. M. Schoonheim, J. S. Damoiseaux, S. A. Rombouts, E. Maris, F. Barkhof, P. Scheltens, and C. J. Stam, Loss of 'small-world'networks in alzheimer's disease: graph analysis of fmri resting-state functional connectivity, *PloS one.* **5**(11) (2010).
24. U. Consortium, Uniprot: a worldwide hub of protein knowledge, *Nucleic acids research.* **47**(D1) (2019).
25. G. O. Consortium, The gene ontology resource: 20 years and still going strong, *Nucleic acids research.* **47**(D1) (2019).
26. M. Uhlén, L. Fagerberg, B. M. Hallström, C. Lindskog, P. Oksvold, A. Mardinoglu, Å. Sivertsson, C. Kampf, E. Sjöstedt, A. Asplund, et al., Tissue-based map of the human proteome, *Science.* **347**(6220) (2015).
27. R. Oughtred, J. Rust, C. Chang, B.-J. Breitkreutz, C. Stark, A. Willems, L. Boucher, G. Leung, N. Kolas, F. Zhang, et al., The biogrid database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions, *Protein Science* (2020).
28. S. Orchard, M. Ammari, B. Aranda, L. Breuza, L. Briganti, F. Broackes-Carter, N. H. Campbell, G. Chavali, C. Chen, N. Del-Toro, et al., The mintact project—intact as a common curation platform for 11 molecular interaction databases, *Nucleic acids research.* **42**(D1) (2014).
29. I. Xenarios, D. W. Rice, L. Salwinski, M. K. Baron, E. M. Marcotte, and D. Eisenberg, Dip: the database of interacting proteins, *Nucleic acids research.* **28**(1) (2000).
30. T. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, et al., Human protein reference database—2009 update, *Nucleic acids research.* **37**(suppl_1) (2009).
31. K. Luck, D.-K. Kim, L. Lambourne, K. Spirohn, B. E. Begg, W. Bian, R. Brignall, T. Cafarelli, F. J. Campos-Laborie, B. Charloteaux, et al., A reference map of the human binary protein interactome, *Nature.* **580**(7803) (2020).
32. D. Alonso-López, F. J. Campos-Laborie, M. A. Gutiérrez, L. Lambourne, M. A. Calderwood, M. Vidal, and J. De Las Rivas, Apid database: redefining protein–protein interaction experimental evidences and binary interactomes, *Database.* **2019** (2019).
33. G. Alanis-Lobato, M. A. Andrade-Navarro, and M. H. Schaefer, Hippie v2. 0: enhancing meaningfulness and reliability of protein–protein interaction networks, *Nucleic acids research* (2016).
34. P. Trott, Int. classification of diseases for oncology, *Journal of clinical pathology.* **30**(8) (1977).
35. K. A. Spackman, K. E. Campbell, and R. A. Côté. Snomed rt: a refer-

ence terminology for health care. In *Proceedings of the AMIA annual fall symposium* (1997).
36. C. E. Lipscomb, Medical subject headings (mesh), *Bulletin of the Medical Library Association*. **88**(3) (2000).
37. O. Bodenreider, The unified medical language system (umls): integrating biomedical terminology, *Nucleic acids research*. **32**(suppl_1) (2004).
38. L. M. Schriml, C. Arze, S. Nadendla, Y.-W. W. Chang, M. Mazaitis, V. Felix, G. Feng, and W. A. Kibbe, Disease ontology: a backbone for disease semantic integration, *Nucleic acids research*. **40**(D1) (2012).
39. S. Köhler, N. A. Vasilevsky, M. Engelstad, E. Foster, J. McMurry, S. Aymé, G. Baynam, S. M. Bello, C. F. Boerkoel, K. M. Boycott, et al., The human phenotype ontology in 2017, *Nucleic acids research*. **45**(D1) (2017).
40. N. Rappaport, N. Nativ, G. Stelzer, M. Twik, Y. Guan-Golan, T. Iny Stein, I. Bahir, F. Belinky, C. P. Morrey, M. Safran, et al., Malacards: an integrated compendium for diseases and their annotation, *Database*. **2013** (2013).
41. S. S. Weinreich, R. Mangon, J. Sikkens, M. Teeuw, and M. Cornel, Orphanet: a european database for rare diseases, *Nederlands tijdschrift voor geneeskunde*. **152**(9) (2008).
42. J. Piñero, À. Bravo, N. Queralt-Rosinach, A. Gutiérrez-Sacristán, J. Deu-Pons, E. Centeno, J. García-García, F. Sanz, and L. I. Furlong, Disgenet: a comprehensive platform integrating information on human disease-associated genes and variants, *Nucleic acids research* (2016).
43. A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick, Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders, *Nucleic acids research*. **33**(suppl_1) (2005).
44. A. Gutiérrez-Sacristán, S. Grosdidier, O. Valverde, M. Torrens, À. Bravo, J. Piñero, F. Sanz, and L. I. Furlong, Psygenet: a knowledge platform on psychiatric disorders and their genes, *Bioinformatics*. **31**(18) (2015).
45. W. Yu, M. Gwinn, M. Clyne, A. Yesupriya, and M. J. Khoury, A navigator for human genome epidemiology, *Nature genetics*. **40**(2) (2008).
46. J. G. Tate, S. Bamford, H. C. Jubb, Z. Sondka, D. M. Beare, N. Bindal, H. Boutselakis, C. G. Cole, C. Creatore, E. Dawson, et al., Cosmic: the catalogue of somatic mutations in cancer, *Nucleic acids research*. **47**(D1) (2019).
47. A. P. Davis, C. J. Grondin, R. J. Johnson, D. Sciaky, R. McMorran, J. Wiegers, T. C. Wiegers, and C. J. Mattingly, The comparative toxicogenomics database: update 2019, *Nucleic acids research*. **47**(D1) (2019).
48. D. Welter, J. MacArthur, J. Morales, T. Burdett, P. Hall, H. Junkins, A. Klemm, P. Flicek, T. Manolio, L. Hindorff, et al., The nhgri gwas catalog, a curated resource of snp-trait associations, *Nucleic acids research*. **42** (D1) (2014).
49. M. J. Li, Z. Liu, P. Wang, M. P. Wong, M. R. Nelson, J.-P. A. Kocher, M. Yeager, P. C. Sham, S. J. Chanock, Z. Xia, et al., Gwasdb v2: an update database for human genetic variants identified by genome-wide association studies, *Nucleic acids research*. **44**(D1) (2016).

50. J. C. Denny, M. D. Ritchie, M. A. Basford, J. M. Pulley, L. Bastarache, K. Brown-Gentry, D. Wang, D. R. Masys, D. M. Roden, and D. C. Crawford, Phewas: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations, *Bioinformatics*. **26**(9) (2010).
51. D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, et al., Drugbank 5.0: a major update to the drugbank database for 2018, *Nucleic acids research*. **46**(D1) (2018).
52. M. Kuhn, I. Letunic, L. J. Jensen, and P. Bork, The sider database of drugs and side effects, *Nucleic acids research*. **44**(D1) (2016).
53. A. Gaulton, A. Hersey, M. Nowotka, A. P. Bento, J. Chambers, D. Mendez, P. Mutowo, F. Atkinson, L. J. Bellis, E. Cibrián-Uhalte, et al., The chembl database in 2017, *Nucleic acids research*. **45**(D1) (2017).
54. S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker, et al., Pubchem substance and compound databases, *Nucleic acids research*. **44**(D1) (2016).
55. K. Degtyarenko, P. De Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcántara, M. Darsow, M. Guedj, and M. Ashburner, Chebi: a database and ontology for chemical entities of biological interest, *Nucleic acids research*. **36**(suppl_1) (2007).
56. Y. H. Li, C. Y. Yu, X. X. Li, P. Zhang, J. Tang, Q. Yang, T. Fu, X. Zhang, X. Cui, G. Tu, et al., Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics, *Nucleic acids research*. **46**(D1) (2018).
57. N. P. Tatonetti, P. Y. Patrick, R. Daneshjou, and R. B. Altman, Data-driven prediction of drug effects and interactions, *Science translational medicine*. **4**(125) (2012).
58. M. Kuhn, C. von Mering, M. Campillos, L. J. Jensen, and P. Bork, Stitch: interaction networks of chemicals and proteins, *Nucleic acids research*. **36** (suppl_1) (2007).
59. M. Dunkel, S. Günther, J. Ahmed, B. Wittig, and R. Preissner, Superpred: drug classification and target prediction, *Nucleic acids research*. **36**(suppl_2) (2008).
60. K. C. Cotto, A. H. Wagner, Y.-Y. Feng, S. Kiwala, A. C. Coffman, G. Spies, A. Wollam, N. C. Spies, O. L. Griffith, and M. Griffith, Dgidb 3.0: a redesign and expansion of the drug–gene interaction database, *Nucleic acids research*. **46**(D1) (2018).
61. M. K. Gilson, T. Liu, M. Baitaluk, G. Nicola, L. Hwang, and J. Chong, Bindingdb in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology, *Nucleic acids research*. **44**(D1) (2016).
62. M. Hewett, D. E. Oliver, D. L. Rubin, K. L. Easton, J. M. Stuart, R. B. Altman, and T. E. Klein, Pharmgkb: the pharmacogenetics knowledge base, *Nucleic acids research*. **30**(1) (2002).
63. S. Günther, M. Kuhn, M. Dunkel, M. Campillos, C. Senger, E. Petsalaki, J. Ahmed, E. G. Urdiales, A. Gewiess, L. J. Jensen, et al., Supertarget and

matador: resources for exploring drug-target relationships, *Nucleic acids research.* **36**(suppl_1) (2007).
64. F. Agüero, B. Al-Lazikani, M. Aslett, M. Berriman, F. S. Buckner, R. K. Campbell, S. Carmona, I. M. Carruthers, A. E. Chan, F. Chen, et al., Genomic-scale prioritization of drug targets: the tdr targets database, *Nature reviews Drug discovery.* **7**(11) (2008).
65. Y. Liu, Q. Wei, G. Yu, W. Gai, Y. Li, and X. Chen, Dcdb 2.0: a major update of the drug combination database, *Database.* **2014** (2014).
66. A. S. Brown and C. J. Patel, A standard database for drug repositioning, *Scientific data.* **4**(1) (2017).
67. T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, et al., Ncbi geo: archive for functional genomics data sets—update, *Nucleic acids research.* **41**(D1) (2012).
68. N. Kolesnikov, E. Hastings, M. Keays, O. Melnichuk, Y. A. Tang, E. Williams, M. Dylag, N. Kurbatova, M. Brandizi, T. Burdett, et al., Arrayexpress update—simplifying data submissions, *Nucleic acids research.* **43**(D1) (2015).
69. K. Tomczak, P. Czerwińska, and M. Wiznerowicz, The cancer genome atlas (tcga): an immeasurable source of knowledge, *Contemporary oncology.* **19** (1A) (2015).
70. V. Matys, E. Fricke, R. Geffers, E. Gößling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, et al., Transfac®: transcriptional regulation, from patterns to profiles, *Nucleic acids research.* **31** (1) (2003).
71. I. Yevshin, R. Sharipov, S. Kolmykov, Y. Kondrakhin, and F. Kolpakov, Gtrd: a database on gene transcription regulation—2019 update, *Nucleic acids research.* **47**(D1) (2019).
72. H. Han, J.-W. Cho, S. Lee, A. Yun, H. Kim, D. Bae, S. Yang, C. Y. Kim, M. Lee, E. Kim, et al., Trrust v2: an expanded reference database of human and mouse transcriptional regulatory interactions, *Nucleic acids research.* **46**(D1) (2018).
73. H.-Y. Huang, Y.-C.-D. Lin, J. Li, K.-Y. Huang, S. Shrestha, H.-C. Hong, Y. Tang, Y.-G. Chen, C.-N. Jin, Y. Yu, et al., mirtarbase 2020: updates to the experimentally validated microrna–target interaction database, *Nucleic acids research.* **48**(D1) (2020).
74. C. Sticht, C. De La Torre, A. Parveen, and N. Gretz, mirwalk: An online resource for prediction of microrna binding sites, *PloS one.* **13**(10) (2018).
75. Q. Jiang, Y. Wang, Y. Hao, L. Juan, M. Teng, X. Zhang, M. Li, G. Wang, and Y. Liu, mir2disease: a manually curated database for microrna deregulation in human disease, *Nucleic acids research.* **37**(suppl_1) (2009).
76. Z. Huang, J. Shi, Y. Gao, C. Cui, S. Zhang, J. Li, Y. Zhou, and Q. Cui, Hmdd v3. 0: a database for experimentally supported human microrna–disease associations, *Nucleic acids research.* **47**(D1) (2019).
77. B. Xie, Q. Ding, H. Han, and D. Wu, mircancer: a microrna–cancer association database constructed by text mining on literature, *Bioinformatics.*

29(5) (2013).
78. A. Ruepp, A. Kowarsch, and F. Theis. Phenomir: micrornas in human diseases and biological processes. In *Next-Generation MicroRNA Expression Profiling Technology*. Springer (2012).
79. Z. Yang, L. Wu, A. Wang, W. Tang, Y. Zhao, H. Zhao, and A. E. Teschendorff, dbdemc 2.0: updated database of differentially expressed mirnas in human cancers, *Nucleic acids research*. **45**(D1) (2017).
80. Y.-R. Miao, W. Liu, Q. Zhang, and A.-Y. Guo, lncrnasnp2: an updated database of functional snps and mutations in human and mouse lncrnas, *Nucleic acids research*. **46**(D1) (2018).
81. G. Chen, Z. Wang, D. Wang, C. Qiu, M. Liu, X. Chen, Q. Zhang, G. Yan, and Q. Cui, Lncrnadisease: a database for long-non-coding rna-associated diseases, *Nucleic acids research*. **41**(D1) (2012).
82. L. Cheng, P. Wang, R. Tian, S. Wang, Q. Guo, M. Luo, W. Zhou, G. Liu, H. Jiang, and Q. Jiang, Lncrna2target v2. 0: a comprehensive database for target genes of lncrnas in human and mouse, *Nucleic acids research*. **47**(D1) (2019).
83. A. Fabregat, S. Jupe, L. Matthews, K. Sidiropoulos, M. Gillespie, P. Garapati, R. Haw, B. Jassal, F. Korninger, B. May, et al., The reactome pathway knowledgebase, *Nucleic acids research*. **46**(D1) (2018).
84. M. Kanehisa and S. Goto, Kegg: kyoto encyclopedia of genes and genomes, *Nucleic acids research*. **28**(1) (2000).
85. D. N. Slenter, M. Kutmon, K. Hanspers, A. Riutta, J. Windsor, N. Nunes, J. Mélius, E. Cirillo, S. L. Coort, D. Digles, et al., Wikipathways: a multifaceted pathway database bridging metabolomics to other omics research, *Nucleic acids research*. **46**(D1) (2018).
86. A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo, and J. P. Mesirov, Molecular signatures database (msigdb) 3.0, *Bioinformatics*. **27**(12) (2011).
87. I. Rodchenkov, O. Babur, A. Luna, B. A. Aksoy, J. V. Wong, D. Fong, M. Franz, M. C. Siper, M. Cheung, M. Wrana, et al., Pathway commons 2019 update: integration, analysis and exploration of pathway data, *Nucleic acids research*. **48**(D1) (2020).
88. M. E. Cusick, H. Yu, A. Smolyar, K. Venkatesan, A.-R. Carvunis, N. Simonis, J.-F. Rual, H. Borick, P. Braun, M. Dreze, et al., Literature-curated protein interaction datasets, *Nature methods*. **6**(1) (2009).
89. A. Grover and J. Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD Int. conference on Knowledge discovery and data mining* (2016).
90. K. Kishan, R. Li, F. Cui, Q. Yu, and A. R. Haake, Gne: a deep learning framework for gene network inference by aggregating biological information, *BMC systems biology*. **13**(2) (2019).
91. H. Luo, D. Wang, J. Liu, Y. Ju, and Z. Jin, A framework integrating heterogeneous databases for the completion of gene networks, *IEEE Access*. **7** (2019).
92. Z. Xiao and Y. Deng, Graph embedding-based novel protein interaction

prediction via higher-order graph convolutional network, *PloS one*. **15**(9) (2020).
93. K. Huang, C. Xiao, L. Glass, M. Zitnik, and J. Sun, Skipgnn: Predicting molecular interactions with skip-graph networks, *Scientific Reports*. **10** (2020).
94. K. KC, R. Li, F. Cui, and A. Haake, Predicting biomedical interactions with higher-order graph convolutional networks, *arXiv:2010.08516* (2020).
95. I. A. Kovács, K. Luck, K. Spirohn, Y. Wang, C. Pollis, S. Schlabach, W. Bian, D.-K. Kim, N. Kishore, T. Hao, et al., Network-based prediction of protein interactions, *Nature communications*. **10**(1) (2019).
96. X. Li, W. Li, M. Zeng, R. Zheng, and M. Li, Network-based methods for predicting essential genes or proteins: a survey, *Briefings in bioinformatics*. **21**(2) (2020).
97. J. Schapke, A. Tavares, and M. Recamonde-Mendoza, Epgat: Gene essentiality prediction with graph attention networks, *arXiv:2007.09671* (2020).
98. X. Zhang, M. L. Acencio, and N. Lemke, Predicting essential genes and proteins based on machine learning and network topological features: a comprehensive review, *Frontiers in physiology*. **7** (2016).
99. R. Albert, H. Jeong, and A.-L. Barabási, Error and attack tolerance of complex networks, *nature*. **406**(6794) (2000).
100. M. Zeng, M. Li, Z. Fei, F. Wu, Y. Li, Y. Pan, and J. Wang, A deep learning framework for identifying essential proteins by integrating multiple types of biological information, *IEEE/ACM transactions on computational biology and bioinformatics* (2019).
101. M. Zeng, M. Li, F.-X. Wu, Y. Li, and Y. Pan, Deepep: a deep learning framework for identifying essential proteins, *BMC bioinformatics*. **20**(16) (2019).
102. X. Zhang, W. Xiao, and W. Xiao, Deephe: Accurately predicting human essential genes based on deep learning, *PLOS Computational Biology*. **16** (9) (09, 2020).
103. A. Shehu, D. Barbará, and K. Molloy. A survey of computational methods for protein function prediction. In *Big Data Analytics in Genomics*. Springer (2016).
104. W. Hamilton, Z. Ying, and J. Leskovec. Inductive representation learning on large graphs. In *Advances in neural information processing systems* (2017).
105. M. Kulmanov, M. A. Khan, and R. Hoehndorf, Deepgo: predicting protein functions from sequence and interactions using a deep ontology-aware classifier, *Bioinformatics*. **34**(4) (2018).
106. V. Gligorijević, M. Barot, and R. Bonneau, deepnf: deep network fusion for protein function prediction, *Bioinformatics*. **34**(22) (2018).
107. K. Fan, Y. Guan, and Y. Zhang, Graph2go: a multi-modal attributed network embedding method for inferring protein functions, *GigaScience*. **9**(8) (2020).
108. R. Bonetta and G. Valentino, Machine learning techniques for protein function prediction, *Proteins: Structure, Function, and Bioinformatics*. **88**(3) (2020).

109. M. Agrawal, M. Zitnik, J. Leskovec, et al. Large-scale analysis of disease pathways in the human interactome. In *PSB* (2018).
110. K. Yang, R. Wang, G. Liu, Z. Shu, N. Wang, R. Zhang, J. Yu, J. Chen, X. Li, and X. Zhou, Hergepred: heterogeneous network embedding representation for disease gene prediction, *IEEE journal of biomedical and health informatics*. **23**(4) (2018).
111. M. Alshahrani and R. Hoehndorf, Semantic disease gene embeddings (smudge): phenotype-based disease gene prioritization without phenotypes, *Bioinformatics*. **34**(17) (2018).
112. S. K. Ata, L. Ou-Yang, Y. Fang, C.-K. Kwoh, M. Wu, and X.-L. Li, Integrating node embeddings and biological annotations for genes to predict disease-gene associations, *BMC systems biology*. **12**(9) (2018).
113. X. Wang, Y. Gong, J. Yi, and W. Zhang. Predicting gene-disease associations from the heterogeneous network using graph embedding. In *2019 IEEE Int. Conference on Bioinformatics and Biomedicine (BIBM)* (2019).
114. D. Wang, P. Cui, and W. Zhu. Structural deep network embedding. In *Proceedings of the 22nd ACM SIGKDD Int. conference on Knowledge discovery and data mining* (2016).
115. A. Ahmed, N. Shervashidze, S. Narayanamurthy, V. Josifovski, and A. J. Smola. Distributed large-scale natural graph factorization. In *Proceedings of the 22nd Int. conference on World Wide Web* (2013).
116. M. Belkin and P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering, *Advances in neural information processing systems*. **14** (2001).
117. L. Zhu, Z. Hong, and H. Zheng. Predicting gene-disease associations via graph embedding and graph convolutional networks. In *2019 IEEE Int. Conference on Bioinformatics and Biomedicine (BIBM)* (2019).
118. Y. Xiong, M. Guo, L. Ruan, X. Kong, C. Tang, Y. Zhu, and W. Wang, Heterogeneous network embedding enabling accurate disease association predictions, *BMC medical genomics*. **12**(10) (2019).
119. P. Han, P. Yang, P. Zhao, S. Shang, Y. Liu, J. Zhou, X. Gao, and P. Kalnis. Gcn-mf: Disease-gene association identification by graph convolutional networks and matrix factorization. In *Proceedings of the 25th ACM SIGKDD Int. Conference on Knowledge Discovery & Data Mining* (2019).
120. P. Luo, Y. Li, L.-P. Tian, and F.-X. Wu, Enhancing the prediction of disease–gene associations with multimodal deep learning, *Bioinformatics*. **35**(19) (2019).
121. L. Madeddu, G. Stilo, and P. Velardi, A feature-learning-based method for the disease-gene prediction problem, *Int. Journal of Data Mining and Bioinformatics*. **24**(1) (2020).
122. Y. Zhang, L. Chen, and S. Li, Cipher-sc: Disease-gene association inference using graph convolution on a context-aware network with single-cell data, *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2020).
123. B. Yang, S. W.-t. Yih, X. He, J. Gao, and L. Deng. Embedding entities and relations for learning and inference in knowledge bases. In *Proceedings of the*

Int. Conference on Learning Representations (ICLR) 2015 (May, 2015).
124. P. Luo, B. Chen, B. Liao, and F.-X. Wu, Predicting disease-associated genes: Computational methods, databases, and evaluations, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* (2020).
125. P. Kaushal and S. Singh, Network-based disease gene prioritization based on protein–protein interaction networks, *Network Modeling Analysis in Health Informatics and Bioinformatics*. **9**(1) (2020).
126. A. Ezzat, M. Wu, X.-L. Li, and C.-K. Kwoh, Computational prediction of drug–target interactions using chemogenomic approaches: an empirical survey, *Briefings in bioinformatics*. **20**(4) (2019).
127. N. Zong, H. Kim, V. Ngo, and O. Harismendy, Deep mining heterogeneous networks of biomedical linked data to predict novel drug–target associations, *Bioinformatics*. **33**(15) (2017).
128. F. Wan, L. Hong, A. Xiao, T. Jiang, and J. Zeng, Neodti: neural integration of neighbor information from a heterogeneous network for discovering new drug–target interactions, *Bioinformatics*. **35**(1) (2019).
129. M. Bagherian, E. Sabeti, K. Wang, M. A. Sartor, Z. Nikolovska-Coleska, and K. Najarian, Machine learning approaches and databases for prediction of drug–target interaction: a survey paper, *Briefings in bioinformatics* (2020).
130. K. Abbasi, P. Razzaghi, A. Poso, S. Ghanbari-Ara, and A. Masoudi-Nejad, Deep learning in drug target interaction prediction: Current and future perspective, *Current Medicinal Chemistry*. **27** (2020).
131. X. Zeng, S. Zhu, X. Liu, Y. Zhou, R. Nussinov, and F. Cheng, deepdr: a network-based deep learning approach to in silico drug repositioning, *Bioinformatics*. **35**(24) (2019).
132. Y. Chen and M. de Rijke. A collective variational autoencoder for top-n recommendation with side information. In *Proceedings of the 3rd Workshop on Deep Learning for Recommender Systems* (2018).
133. M. Karimi, A. Hasanzadeh, and Y. Shen, Network-principled deep generative models for designing drug combinations as graph sets, *Bioinformatics*. **36**(Supplement_1) (2020).
134. H. Xue, J. Li, H. Xie, and Y. Wang, Review of drug repositioning approaches and resources, *Int. journal of biological sciences*. **14**(10) (2018).
135. T. N. Jarada, J. G. Rokne, and R. Alhajj, A review of computational drug repositioning: strategies, approaches, opportunities, challenges, and directions, *Journal of Cheminformatics*. **12**(1) (2020).
136. M. Zitnik, M. Agrawal, and J. Leskovec, Modeling polypharmacy side effects with graph convolutional networks, *Bioinformatics*. **34**(13) (2018).
137. D. W. Bates, N. Spell, D. J. Cullen, E. Burdick, N. Laird, L. A. Petersen, S. D. Small, B. J. Sweitzer, and L. L. Leape, The costs of adverse drug events in hospitalized patients, *Jama*. **277**(4) (1997).
138. N. Rohani and C. Eslahchi, Drug-drug interaction predicting by neural network using integrated similarity, *Scientific reports*. **9**(1) (2019).
139. F. Cheng and Z. Zhao, Machine learning-based prediction of drug–drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties, *Journal of the American Medical Informatics Association*.

21(e2) (2014).
140. W. Zhang, H. Zou, L. Luo, Q. Liu, W. Wu, and W. Xiao, Predicting potential side effects of drugs by recommender methods and ensemble learning, *Neurocomputing*. **173** (2016).
141. C. Li, H. Liu, Q. Hu, J. Que, and J. Yao, A novel computational model for predicting microrna–disease associations based on heterogeneous graph convolutional networks, *Cells*. **8**(9) (2019).
142. Z. Li, J. Li, R. Nie, Z.-H. You, and W. Bao, A graph auto-encoder model for mirna-disease associations prediction, *Briefings in Bioinformatics* (2020).
143. P. Xuan, H. Sun, X. Wang, T. Zhang, and S. Pan, Inferring the disease-associated mirnas based on network representation learning and convolutional neural networks, *Int. journal of molecular sciences*. **20**(15) (2019).
144. P. Xuan, Y. Dong, Y. Guo, T. Zhang, and Y. Liu, Dual convolutional neural network based method for predicting disease-related mirnas, *Int. journal of molecular sciences*. **19**(12) (2018).
145. J. Peng, W. Hui, Q. Li, B. Chen, J. Hao, Q. Jiang, X. Shang, and Z. Wei, A learning-based framework for mirna-disease association identification using neural networks, *Bioinformatics*. **35**(21) (2019).
146. J. Li, S. Zhang, T. Liu, C. Ning, Z. Zhang, and W. Zhou, Neural inductive matrix completion with graph convolutional networks for mirna-disease association prediction, *Bioinformatics*. **36**(8) (2020).
147. J. Li, Z. Li, R. Nie, Z. You, and W. Bao, Fcgcnmda: predicting mirna-disease associations by applying fully connected graph convolutional networks, *Molecular Genetics and Genomics: MGG* (2020).
148. S. Rhee, S. Seo, and S. Kim. Hybrid approach of relation network and localized graph convolutional filtering for breast cancer subtype classification. In *Proceedings of the Twenty-Seventh Int. Joint Conference on Artificial Intelligence, IJCAI-18* (2018).
149. R. Schulte-Sasse, S. Budach, D. Hnisz, and A. Marsico. Graph convolutional networks improve the prediction of cancer driver genes. In *Int. Conference on Artificial Neural Networks* (2019).
150. P. Luo, Y. Ding, X. Lei, and F.-X. Wu, deepdriver: predicting cancer driver genes based on somatic mutations using deep convolutional neural networks, *Frontiers in Genetics*. **10** (2019).
151. N. Fatima and L. Rueda, isom-gsn: an integrative approach for transforming multi-omic data into gene similarity networks via self-organizing maps, *Bioinformatics*. **36**(15) (2020).
152. G. Rosenthal, F. Váša, A. Griffa, P. Hagmann, E. Amico, J. Goñi, G. Avidan, and O. Sporns, Mapping higher-order relations between brain structure and function with embedded vector representations of connectomes, *Nature communications*. **9**(1) (2018).
153. P. Lee, M. Choi, D. Kim, S. Lee, H.-G. Jeong, and C. E. Han. Deep learning based decomposition of brain networks. In *2019 Int. Conference on Artificial Intelligence in Information and Communication (ICAIIC)* (2019).
154. J. Loscalzo, *Network medicine*. Harvard University Press (2017).