# Integration of multi-scale protein interactions for biomedical data analysis

*Thomas Gaudelet*

A dissertation submitted in partial fulfillment

of the requirements for the degree of

**Doctor of Philosophy**

of

**University College London**.

Department of Computer Science

University College London

March 25, 2021

I, Thomas Gaudelet, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

# Abstract

With the advancement of modern technologies, we observe an increasing accumulation of biomedical data about diseases. There is a need for computational methods to sift through and extract knowledge from the diverse data available in order to improve our mechanistic understanding of diseases and improve patient care. Biomedical data come in various forms as exemplified by the various omics data. Existing studies have shown that each form of omics data gives only partial information on cells state and motivated jointly mining multi-omics, multi-modal data to extract integrated system knowledge. The interactome is of particular importance as it enables the modelling of dependencies arising from molecular interactions.

This Thesis takes a special interest in the multi-scale protein interactome and its integration with computational models to extract relevant information from biomedical data. We define multi-scale interactions at different omics scale that involve proteins: pairwise protein-protein interactions, multi-protein complexes, and biological pathways. Using hypergraph representations, we motivate considering higher-order protein interactions, highlighting the complementary biological information contained in the multi-scale interactome. Based on those results, we further investigate how those multi-scale protein interactions can be used as either prior knowledge, or auxiliary data to develop machine learning algorithms. First, we design a neural network using the multi-scale organization of proteins in a cell into biological pathways as prior knowledge and train it to predict a patient's diagnosis based on transcriptomics data. From the trained models, we develop a strategy to extract biomedical knowledge pertaining to the diseases investigated. Second, we propose a general framework based on Non-negative Matrix Factorization to integrate the multi-scale protein interactome with multi-omics data. We show that our approach outperforms the existing methods, provide biomedical insights and relevant hypotheses for specific cancer types.

# Impact Statement

Over the last decades, the need for computational methods to process, understand diverse data has grown along with the exponentially increasing amount of data collected. Machine learning has demonstrated an impressive ability for detecting general patterns within large corpora of data in various field such as natural language and image processing. The application of machine learning to the biomedical field is a promising avenue to improve the practice of medicine, notably through the analysis of rich molecular data. In the last decades, medical research has moved away from the Mendelian paradigm to embrace the concept of Network Medicine, which highlights the intricate molecular interaction underlying diseases. The biomedical field is data-rich, particularly due to the diversity of the data that describes a single organism at different scales.

The contributions of the work presented can be classified in methodological and biological contributions. We introduce methods that can effectively integrate multiple, diverse data sources at the different scales of biology. As such, this Thesis is part of the global effort to bridge the gap between network analysis and machine learning and to harness diverse biomedical data to further our understanding of biology and particularly diseases. We take particular interest in how to extract information from hierarchical interactions between proteins within cells. The molecular organisation is far from random and contains cues that algorithms can exploit to identify biological hypotheses and directions that experimental research can exploit to address major health and social crisis, such as cancer or, in light of recent event, pandemics such as COV-SARS-2. These hypotheses constitute the principal downstream implications of the methods and results presented in this Thesis. New biological hypotheses describing underlying mechanisms of diseases have important implication for patients such as more accurate diagnosis and better care as well as the identification of leads for drug development. In the shorter term, the work presented opens new directions for future academic research and method development. The work was conducted as part of prestigious ERC consolidator grant, and presented in the largest conference of the field (ECCB, acceptance rate $< 20\%$).

# Acknowledgements

This PhD has been a long adventure, and there is a long list of people that have helped me grow along the way, both academically and personally.

First, I would like to thank my research group. To Nataša Pržulj, thank you for giving me the opportunity to study in such an exciting and impactful field. To Noël Malod-Dognin, thanks for helping me along the way and being always available for pints. To Sam Windels, thanks for bearing the recurrent complaining throughout this PhD experience. Second, I would like to thank all the professors that have provided me with guidance or given me opportunities along the way, specifically Iasonnas Kokkinos and the DeepMind team (Thore Graepel, Diana Borsa, and Hado van Hasselt) for giving me the opportunity to be a teaching assistant in their respective courses, I have learnt a lot from these experiences. Last but not least, I would like to thank Jure Leskovec and his group at Stanford, notably Marinka Zitnik, for having me as a visiting student. It was a short stay, but one of the most impactful experiences of my PhD.

Finally, I would like to thank family and friends that have helped me during this period. My parents and siblings who were always there, rooting for me, as well as my grandmother, Yvette Mouyon, without whose help I would not have been here. Emanuela Sala, thank you for your constant support through this adventure and making the global experience easier. The Oxford crew that provided me with distractions throughout the period. In no particular order: James, Beth, Ruben, Fraser, Gala, Kalil, Konrad, Tom, Rebecca, Lexie, and Ben, thanks for being there. And finally, all the different flatmates with whom I have lived in my numerous homes in London and who made life in London more interesting.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Context

The advancement of modern high-throughput capturing and sequencing technologies, and the subsequent decrease of experimental costs, has led to the dramatic increase in the amount and speed at which data is collected in a broad spectrum of biological and clinical experiments. Projects such as the Human Genome project [2] and The Cancer Genome Atlas (TCGA) project[1] are emblematic of this new era in biology and particularly of the joint efforts to increase our knowledge base. The work presented in this Thesis is part of ongoing efforts to develop computational methods to harness the wide range of biological data available to uncover biomedical insights with an eye towards improving medical practice.

The variety of the existing biological data is best exemplified by the different types of **omics** data such as genomics, transcriptomics, and metabolomics. In section 2.1.1, we give a brief overview of some omics data within the context of the cellular machinery. Biomedical applications of computational methods have greatly benefited from access to omics data. We are now able to obtain molecular-level characterisations of both healthy and disease states, allowing researchers to identify molecular alterations that are significantly associated with diseases. However, to be able to understand the implication of any given alteration (e.g. causative or symptomatic), it is essential to get an accurate depiction of its context within the cellular

[1] http://cancergenome.nih.gov/.

machinery. Hardly any entity exists in absolute isolation from others. It is usually a part of a larger system of greater complexity. This remains true at any scale, and notably, molecules in cells interact to carry out complex tasks and functions. The study of single entities in isolation of others ignores the context that shapes their roles as part of the broader systems of interacting components.

Interactomics studies molecular interactions on which rest the different processes required to fulfil a biological roles [3]. Complex traits, such as diseases, are rarely a consequence of a single gene and typically arise from complex molecular interactions [4]. In particular, high-level cellular functions are achieved based on physical bindings between molecules. These include transient bindings which are parts of cascades of chemical reactions that result in specific processes or signals. A protein can also bind to the DNA, effectively preventing (or enabling) the transcription of regions of the DNA. This phenomenon is part of what is referred to as *regulation*. Furthermore, external stimuli can also trigger processes within a cell, for instance, neurons can send "instructions" to muscle cells to contract a muscle. It involves both instruction and coordination messages exchanged between cells in the form of chemical reactions. Understanding the precise mechanisms of a cell is difficult and requires mathematical abstraction to model its rich microcosm. Biological systems are often modelled using graphs (networks) to capture the interactions between pairs of entities while abstracting much of the detail about them. Briefly, a graph is given by a set of vertices (nodes), each corresponding to an entity, and a set of interactions (edges or links) between those entities. We introduce graphs in details in Section 2.2. There are various examples of molecular networks studied in the literature and capturing different types of molecular interactions between pairs of molecular entities; we discuss some examples in Section 2.1.2.

However, as important as molecular networks have been to advancing our understanding of the cellular machinery, they still only capture reductionist pictures of biological mechanisms. One particular issue is that these networks only capture binary (pairwise) interactions, i.e. interactions between pairs of elements. However, molecular interactions occur at different scales. For instance (see Figure 1.1), focusing on proteins, two proteins can physically bind to form a protein-protein interaction (PPI). Then, multiple proteins can physically bind, forming one cohesive unit, called a **protein complex**, with specific functional roles. Additionally, **biological pathways**, which are associated with specific functions or phenotypes, involve multiple proteins and complexes. It is through this **multi-scale interactome** that emerge the complex phenotypes and functions displayed by cells, tissues, and organisms [4].

In this Thesis, we take a specific interest in the multi-scale protein interactome. Notably, we investigate its integration into models designed to study molecular interactions within cells and their pathological states. We demonstrate the importance of considering the different scales of protein interactions for biomedical data analysis. Our motivation is to obtain models that more accurately capture the multi-scale molecular organisation and enable the investigation of higher-order entities such as biological pathways within a pathological context. First, we show that each layer in the multi-scale protein interactome (see Figure 1.1) contains complementary biological information, motivating their associations in analytical models. Utilising this observation, we address some of the challenges in computational biology that we introduce in the next section. Succinctly, we develop new approaches to study cell's pathological states. Our objectives are to get better understandings of diseases through the identification of perturbed cellular entities and mechanisms at the different scales of molecular interactions. Identifying affected biological functions and understanding the molecular basis of diseases are crucial steps for the improvement of medical care and the development, or repurposing, of effective treatments [5].

## 1.2 Challenges in Computational Biology

Computational biology is facing numerous technical and conceptual challenges that need to be addressed to extract valuable information from the rapidly increasing amounts of data. The main objective is to formulate hypotheses and directions for experimental research, which ultimately could lead to both fundamental biological knowledge and improved patient care. An added requirement for models and

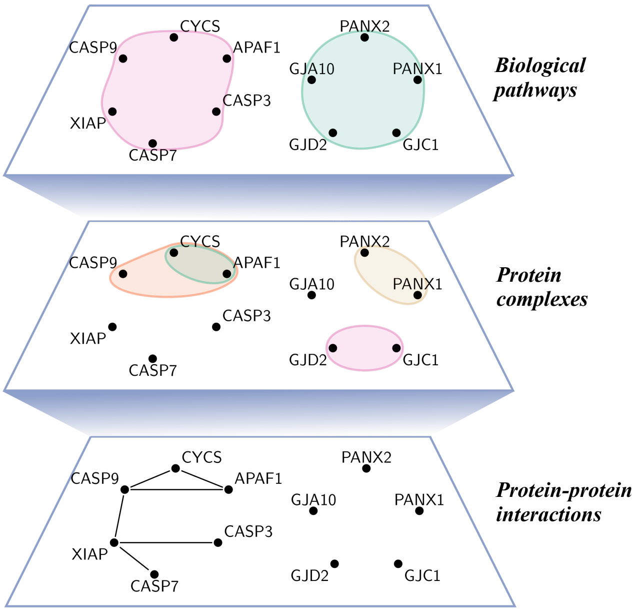


**Figure 1.1:** Representation of different scale of molecular interactions on a set of genes. The bottom panel corresponds to Protein-Protein interaction experimentally validated in BioGrid [6], the middle panel gives protein complexes obtained from Reactome [7], and the top panel gives biological pathways from Reactome [7].

algorithms in biomedical applications, sometimes overlooked in other fields, is interpretability of the results, that is to say, the ability to explain how and why an algorithm is making a "decision". Model interpretability is useful both for experimental design, to test hypotheses, and to understand the underlying mechanisms uncovered by the models. This is especially relevant in biology where we lack intuition that we naturally have in fields such as image or language processing. Furthermore, the European Union passed a resolution in 2016, the General Data Protection Regulation, concerning the use of personal information that includes the "right to an explanation". The law states that when using a patient's data to make a decision, a clinician should be able to detail how the decision has been reached, which includes decisions resulting from algorithm-aided processes.

Naturally, numerous modern challenges are tied to the staggering amount of data available. In genomics, since the completion of the first human genome, thousands of human genomes have been assembled. A recent project[2] in the UK created

[2] https://www.genomicsengland.co.uk/the-100000-genomes-project/

a database of $100,000$ genomes (completed in December 2018). The data takes around 20 petabytes ($10^{15}$ bytes) of storage. A similar project in the USA announced a database of $1,000,000$ genomes. The Gene Expression Omnibus (GEO) website offers access to more than $1,000,000$ samples. Clinical data, including Electronic Health Record (EHR), is also on the rise with the US healthcare system reporting 150 exabytes ($10^{18}$ bytes) of data in 2011. Datasets are often highly heterogeneous, even for the same type of data there are multiple methods and technologies with different coverage, bias and noise [8]. As the quantity of the available data keeps increasing, scalable computational methods need to be devised to uncover biological knowledge of practical value.

We review below two interlinked modern challenges facing computational biology that are at the core of this Thesis: data integration and precision medicine.

### 1.2.1 Data integration

The individual analysis of specific biological data can generate knowledge that is, to some extent, myopic to the broader context. For instance, complex diseases, such as cancer, can be caused by combinations of genetic, molecular, environmental, and lifestyle factors. Hence, no single type of omics data can fully capture such diseases. Thus, a strategy that has been gaining momentum is the collective mining of different data to extract *integrated* knowledge about a system that goes beyond what any single data source can offer [9] (see illustration Figure 1.2). Those approaches have numerous advantages, addressing some limitations and caveats inherent to each technology that generate data.

Integrative approaches have already shown promising improvements in various biomedical applications, such as *protein function prediction* [10, 11, 12, 13] and *biological network inference* [14, 15, 16]. Briefly, protein function prediction is concerned with the characterisation of the role of each protein within the cellular machinery, in both normal and pathological context. There are many ongoing efforts to develop methods to annotate proteins and genes automatically. This is best exemplified by the popularity of the CAFA challenge [17]. The biological network inference task aims to derive the underlying molecular interactions from biological

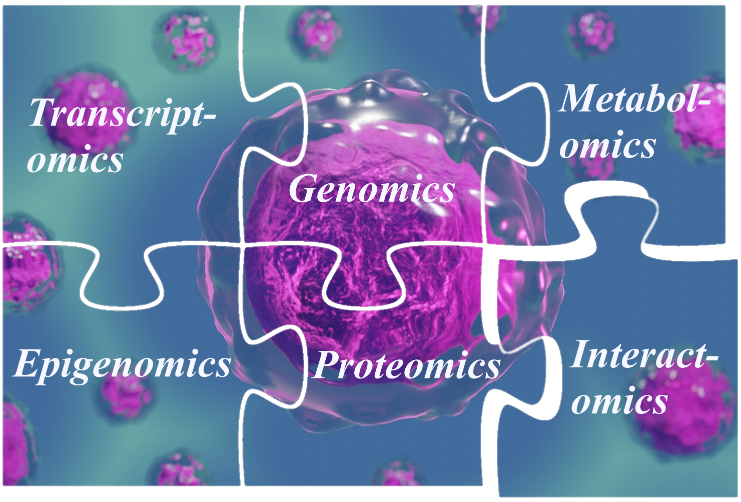


**Figure 1.2:** Illustration of omics data integration exploiting their complementarity.

data. The task boils down to predicting associations between biological entities. Applications include protein–protein interaction predictions [18] and regulatory interaction predictions [14, 19].

Developing efficient computational methodologies to handle and extract information from the various data sources is thus a fundamental challenge of computational biology. Under the *no free lunch theorem* [20], there cannot be one single computational model that solves all problems. As such, each research question requires new tailored approaches. We discuss data integration in more details and review the existing approaches in Section 2.3.

### 1.2.2 Precision Medicine

A modern challenge, linked to the abundance of patient-specific biomedical data, is the personalisation of the practice of medicine based on patient-specific information, such as genetic background or other omics datasets. This field of computational biology is called **precision medicine**, also referred to as personalised, predictive, preventive and participatory (“P4”) medicine. Precision medicine is typically associated with three main tasks: patient stratification, biomarker discovery, and drug repositioning.

**Patient stratification** corresponds to the identification of groups of patients having similar clinical outcome through their molecular data. The aim is to improve risk assessment and prevention, as well as optimise the choice of treatment. The task is essential to characterise complex, heterogeneous diseases that can be caused by different molecular alterations despite similar phenotype manifestation (e.g., cancers [21], Alzheimer's disease [22]). A popular approach to the problem, as highlighted by Wang *et al.* [23], is the molecular-based clustering of patients through the integration of multi-omics data. Kaplan-Meier estimator [24] are typically used to quantify if the subgroups identified in a cohort significantly correlate with prognosis. The estimator gives a p-value of the significance of the correlation based on a log-rank statistical test. A problem closely related to patient stratification is the prediction of a patient's clinical outcome and prognosis based on both clinical and biomedical data. Most approaches addressing this problem follow a two-step framework. First, they start by training machine learning models to predict patients' survivals; such models are often based on Cox proportional hazards [25]. Second, they analyse the models to identify covariates that significantly represents a patient's risk [26, 27, 28]. These covariates can be further used to stratify the cohort in subgroups significantly correlated to prognosis.

**Biomarker discovery** aims to identify biological characteristics (biomarkers) that can be evaluated objectively and are indicators of healthy or pathogenic biological processes. Biomarkers can be used to diagnose a disease early, to evaluate an individual's susceptibility to a disease, to assess the evolution and disease risk, and to predict an individual's response to a treatment [29]. Identifying prognosis biomarkers is a task closely linked to patient stratification. As such, methods are often designed to address both tasks simultaneously. For instance, the clustering-based approaches mentioned above generally include a second step that derives molecular signatures that are representative of each cluster [23]. These signatures can be used both to classify unseen patients and to identify specific, putative biomarkers that characterise the subgroup and are linked to survival. Similarly, the second step used by models that directly predict survival rates can also be used to identify new, putative biomarkers through the identification of covariates significantly linked to patients' survival [26, 27, 28].

**Drug repositioning** aims at finding new applications for existing pharmacotherapies. Drug discovery is a slow and costly process involving many steps from creation to clinical trials. It is estimated that developing a new drug takes over 12 years and cost over $1 billion on average [30, 31]. Repurposing existing drugs to treat new diseases enables circumventing some development steps, reducing cost and time to bring a drug to market. It also reduces risks associated with new drugs such as unknown side effects. A wide range of computational methods has been developed to address the problem as reviewed by Pushpakom *et al.* [32]. Relevant to our work, integrative approaches, including network-based, have been used to mine omics jointly with other biological data to derive new repurposing hypotheses [33, 34, 35].

## 1.3 Thesis contributions and outline

In this Thesis, we explore solutions to the data integration challenge to answer some of the tasks associated with precision medicine. The Thesis breaks down into the following chapters.

**Chapter 2** first describes some omics data types and various molecular networks. The description motivates the view of cells as multi-faceted systems of interacting components that operate at different scales. Secondly, the chapter introduces background notions of graph theory and machine learning that are necessary to frame the context of the Thesis.

**Chapter 3** investigates the information contained in the different scales of protein interactions: PPI, protein complexes, and biological pathways. We utilise hypergraph representations and introduce *hypergraphlets* as a tool for their analysis. We show that each type of interactions contains complementary biological information about proteins. We further demonstrate that new biological insight can be derived from the multi-scale protein interactome by predicting novel protein functional annotations using a simple approach based on hypergraphlet statistics. Our results motivate the integration of the multi-scale protein interactome to machine learning models developed for the analysis of biological data. In the following chapters, we propose approaches to integrate this information into computational models.

**Chapter 4** investigates the use of multi-scale protein interactions as prior knowledge to define and control the structure of a neural network (NN). We train the NN to predict patients diagnosis based on genes' differential expression. We then introduce a methodology, based on the analysis of variation of the function parametrised by the neural network, to derive biomedical knowledge and diseases mechanisms from the trained NN models. Thus, our methodology gives a way to interpret neural network models to derive biological hypotheses such as, in this case, putative diagnostic biomarkers.

**Chapter 5** proposes a framework integrating patient omics and clinical data, the multi-scale protein interactome, drug–protein interactions, and drug similarity data for pan-cancer analysis. The objectives are to uncover novel molecular cancer mechanisms, putative biomarkers, and propose new drug indications for the treatment of cancer subtypes. Our framework can identify links, that have not been studied (or reported) previously, between cancer and higher-order molecular structures. Furthermore, we demonstrate that our approach captures underlying biological mechanisms that govern response to cancer drugs. Our methodology extends to the derivation of patient-specific knowledge, such as personalised drug recommendations.

**Chapter 6** summarizes the contributions, provides a general discussion, and outlines future research directions.

The contributions of the Thesis fall within two categories: methodological and biological.

**Methodological contributions** are the development of novel network analysis tools and machine learning approaches. Specifically,

- Introduction of hypergraphlets, and the associated distance metric, to characterise and compare wiring patterns of nodes in hypergraphs.
- A novel neural network design based on prior biological knowledge and an analytical strategy to derive knowledge from trained models.

- A joint integrative embedding library based on Non-negative Matrix Factorisations.
- A novel framework for association inferences in standard and zero-shot learning settings.

**Biological contribution** is in the novel biomedical knowledge derived from the proposed methodologies. Specifically,

- Prediction of protein function by hypergraph representations of the multi-scale interactome.
- Uncovering of disease molecular mechanisms at the different scales of the multi-scale protein interactome.
- Prediction of drugs to be repurposed for the treatment of specific cancer types.

## 1.4 List of publications

The results presented in this Thesis have been published in peer-reviewed journals and conferences, or are currently under review for publication. We give below the list of original articles written as part of the Thesis work.

1. **Gaudelet, T.**, Malod-Dognin, N. and Pržulj, N., 2018. Higher-order molecular organisation as a source of biological function. Bioinformatics, 34(17), pp.i944-i953.
2. **Gaudelet, T.** and Pržulj, N., 2019. Introduction to Graph and Network Theory, in Pržulj, N. (ed.) *Analysing network data in biology and medicine*. Cambridge University Press, pp. 111-150.
3. **Gaudelet, T.**, Malod-Dognin, N., Sánchez-Valle, J., Pancaldi, V., Valencia, A. and Pržulj, N., 2020. Unveiling new disease, pathway, and gene associations via multi-scale neural network. PLoS One, 15(4), p.e0231059.
4. Lugo-Martinez, J., Zeiberg, D., **Gaudelet, T.**, Malod-Dognin, N., Pržulj, N., Radivojac, P., Classification in biological networks with hypergraphlet kernels, Bioinformatics, btaa768

5. **Gaudelet, T.**, Malod-Dognin, N. and Pržulj, N., 2020. Integrative Data Analytic Framework to Enhance Cancer Precision Medicine. arXiv preprint arXiv:2007.01107.

6. **Gaudelet, T.**, Malod-Dognin, N. and Pržulj, N., MNMFIF: Mixed Non-negative Matrix Factorization Integrative Framework (In preparation)

# Chapter 2

# Background

This chapter introduces the background knowledge necessary to navigate the Thesis. First, we give an overview of the existing types of biological data. Specifically, we introduce various omics data and molecular networks. Second, we introduce notions of graph theory that are essential to understand some of the concepts discussed and used in the Thesis. Finally, we review the literature and discuss methods that have been developed to address data integration problems and that set the stage for the Thesis.

## 2.1 Data

In this section, we introduce different types of biological data within the context of cellular machinery. The resulting description of the cellular mechanisms is far from complete and can only barely scratch the surface of the complex organisations at play. However, the objective is to give an idea of the intricacies of cell mechanisms, to motivate the view of biology as a multi-faceted system of interacting components [9, 36], and to introduce some terminology. This short description also allows us to highlight the different aspects that need to be taken into consideration when studying the inner workings of cells. Note that any biological data typically comes with noise and biases. Noise can be linked to multiple factors such as human error, experimental noise, and variability in underlying system [37, 38, 39]. Biases can also arise at multiple levels such as lack of diversity in patient cohorts [40] or choice of experiments. For instance, Luck *et al.* [41] demonstrated that the publicly available experimental molecular data tend to be biased towards proteins that have been associated with diseases. The use of biased data, notably biased patient data, to improve medical care might jeopardise the fairness of the final system [42]. Thus, addressing these issues is essential for both experimental design and downstream data analysis and, as such, is an important open research problem that, however, falls outside the scope of this Thesis.

### 2.1.1 Omics data

Omics data refer to comprehensive data about molecules of the same type and obtained simultaneously with high-throughput assays. There is a wide-ranging array of omics data. Each captures quantitative information about single entities in the cellular machinery. We introduce some examples below, discussing technologies used to capture the data and landmark findings associated with each type of omics.

**Genomics** data relate to the genetic material in the form of DNA data encapsulated in the nucleus. Genomics is concerned with the analysis of DNA sequences and in particular genes. A *gene* is defined as a region of the DNA that codes for proteins. In this Thesis, we use the term gene to refer indiscriminately to the gene itself (DNA sequence), its associated coding RNA transcripts, and all the proteins it encodes. DNA sequencing technologies have greatly improved over the last decade, both in quality and speed. The latest next-generation sequencers (NGS) can capture up to 16 human genomes over three days. In comparison, the Human Genome Project, that generated the first human genomics data, spanned 12 years and cost $3 billion. The access to full genomes has facilitated the development of Genome-Wide Association Studies (GWAS) and the identification of genetic markers associated with phenotypes, for instance, the identification of cancer driver genetic mutations [43].

**Transcriptomics** measures the presence and relative amount of RNA transcripts transcribed from the DNA at a fixed point in time. The associated data give information on the momentaneous state of cells, highlighting the active components. RNA transcripts are either *coding*, i.e. that translates into proteins, or *non-coding*. The most common approaches to generate transcriptomics data are microarrays [44] and RNA Sequencing (RNA-Seq) [45]. Typically, transcriptomics data is obtained from bulk samples containing multiple cells. However, the ongoing development of single-cell technologies aims to give a more fine-grained depiction of cellular states. Transcriptomics data have enabled comparative studies of gene expression across cohorts, enabling the comparisons of diseases on their molecular expression profiles [46], and across time, allowing for the tracing of tumorigenesis [47].

**Proteomics** aims to catalogue, sequence, and provide quantitative information about proteins, translated from coding RNA in cells, that are responsible for most cellular functions. Due to alternative splicing, post-translational modifications, and mRNA editing, a gene can code for multiple proteins. Thus the pool of human proteins is orders of magnitude larger than the numbers of genes in humans, with an estimated 1.8 million protein [48] compared to around $25,000$ genes. Mass Spectrometry technologies [49] and reverse phase protein lysate microarrays [50] are arguably the most popular approaches to collect proteomics data. The development of proteomics technologies has greatly benefited the study of single proteins with applications toward protein structure prediction [51] and protein function prediction [52].

**Metabolomics** studies the set of *metabolites* present in a system. The term metabolite encompasses any substance that is produced or consumed in chemical reactions occurring in a cell. The set of all chemical reactions is called *metabolism*. MS and nuclear magnetic resonance (NMR) spectrometry are the methods of choice to extract metabolites profile. Research has highlighted that a cell's metabolites vary under different conditions, for instance, diseased state versus healthy state [53], and as such metabolomics data can help with identification and diagnosis of diseased cells, e.g. in cancer [54].

**Epigenomics** is the study of epigenetic changes, which correspond to reversible changes on a cell's DNA, causing perturbation of gene expression. DNA methylation, histone modification, and chromatin structure alteration are examples of epigenetic perturbations [55, 56]. Methods such as ChiP-seq [57] and Bisulfite sequencing [58] can be used to identify epigenetic modifications. Changes in the epigenome have been linked to cellular development and environment [59] as well as to the occurrence of diseases [60].

### 2.1.2 Molecular networks

The omics data described above only give information about single entities without explicitly considering the underlying molecular interactions. Interactomics investigate and catalogues molecular interactions that are often represented as networks (formally introduced in the next section). Below, we describe some of the existing biological networks, highlighting the wide range of molecular interactions on which the cellular machinery relies. Notably, for each molecular network, we detail the type of interactions they represent, how the interactions are captured, and some applications for which the network has been used. Furthermore, we list a few reference databases for molecular networks in Table 2.1.

**Protein-Protein Interaction (PPI) networks** are perhaps the most studied molecular networks. They represent physical interactions, or bindings, between the proteins of a species [61, 62, 63, 64]. Experimentally, PPIs are derived using Yeast Two-Hybrid (Y2H) or affinity capture techniques [65]. A PPI network can be seen as the blueprint of the protein bindings of a species. It does not mean that at any given time, one can observe all those interactions in any given cell (e.g., some proteins might not always be expressed in a cell). As each protein is associated with a specific gene, PPI networks often use genes notation in place of proteins. While this simplification reduces the size of PPI networks dramatically and facilitate the comparison and integration to other molecular networks, it does lead to loss of information and granularity. Research focusing on biological systems has relied extensively on PPI networks in a wide range of applications such as exploring the link and evolution of PPIs across species [66], uncovering functional modules in the network to identify proteins' function [67], linking network properties to protein's essentiality and disease [68, 69], as well as drug discovery [70].

**Transcriptional regulatory networks** model gene expression regulation [71, 72, 73]. Specific genes code for proteins that regulate the expression of other genes; such a protein is called a *transcription factor*. A transcriptional regulation network is a simplified representation of this phenomenon, where gene X is connected to gene Y if the protein product of X controls the expression of Y. Such a relation is *asymmetric*, as X controls the expression of Y, but Y does not necessarily influence the expression of X. Regulatory interactions are inferred experimentally through Yeast One-Hybrid (Y1H) and chromatin immunoprecipitation (ChIP) [65]. Furthermore, numerous computational approaches have been explored to infer regulatory networks from omics data [14, 19]. Properties of regulatory networks have been linked to protein's essentiality and disease [74], and the networks have further helped uncover disease mechanisms at the molecular level [75, 76, 77].

**Metabolic networks** model the metabolism of a cell [78, 79, 80]. The set of all metabolic reactions in a cell forms a metabolic network. A metabolic reaction transforms one metabolite (small molecule) into another and is catalysed by an enzyme (protein). In short, a metabolic reaction involves at least two metabolites and an enzyme. The interactions are usually asymmetric, as the biochemical reactions involved are typically irreversible. Metabolic networks are constructed by compiling experimental results describing metabolic reactions that are reported in the literature [65]. As discussed above, metabolomics studies have exposed the context-dependent composition of the set of metabolites in cells. Under this observation, metabolic networks can help understand the underlying processes that are perturbed, or that arise, under certain conditions, such as cancer [81] or Parkinson's disease [82].

**Co-expression networks** [83] represent the correlation of the expression of transcriptomics data across multiple cell samples [84]. Two genes are connected if their expression is significantly correlated across samples. The underlying assumption is that either one gene controls the expression of the other or a third party simultaneously controls the expression of both genes. Analyses of co-expression networks have uncovered functional modules [85], non-coding RNA functions [86], and system-level properties across cancers [87].

| Databases | URL | Networks | reference |
|---|---|---|---|
| KEGG | www.genome.jp/kegg | M; TR | [88] |
| GeneDB | www.genedb.org | M | [89] |
| BioCyc, EcoCyc, MetaCyc, HumanCyc | biocyc.org | M; TR | [90, 91, 92] |
| MetaTIGER | www.bioinformatics.leeds.ac.uk/metatiger | M | [93] |
| ERGO (previously known as WIT) | ergo.integratedgenomics.com | M | [94] |
| GeneNet | wwwmgs.bionet.nsc.ru/mgs/systems/genenet | TR | [95] |
| Reactome | reactome.org | TR | [7, 96] |
| RegulonDB | regulondb.ccg.unam.mx | TR | [97] |
| JASPAR | jaspar.cgb.ki.se | TR | [98] |
| Phospho.ELM | phospho.elm.eu.org | TR | [99] |
| NetPhorest | netphorest.info | TR | [100] |
| PHOSIDA | 141.61.102.18/phosida/index.aspx | TR | [101] |
| TRANSPATH | www.biobase.de/transpath | TR | [102] |
| SMPDB | smpdb.ca | TR | [103] |
| BioGRID | thebiogrid.org | PPI | [6] |
| HPRD | www.hprd.org | PPI | [104] |
| SGD | www.yeastgenome.org | PPI | [105] |
| IntAct | www.ebi.ac.uk/intact | PPI | [106] |
| HPID | www.hpid.org | PPI | [107] |
| DroID | www.droidb.org | PPI | [108] |
| MIPS | mips.gsf.de | PPI | [109] |
| DIP | dip.doe-mbi.ucla.edu/dip/Main.cgi | PPI | [110] |
| MINT | mint.bio.uniroma2.it | PPI | [111] |
| IID (previously known as I2D/OPHID) | ophid.utoronto.ca/ophidv2.204 | PPI | [112] |
| STRING | string-db.org | PPI | [113] |
| COXPRESdb | coxpresdb.jp | COEX | [114] |
| GeneFriends | genefriends.org | COEX | [115] |

**Table 2.1:** Existing databases for protein-protein interaction (PPI) networks, transcriptional regulatory networks (TR), metabolic networks (M), and gene coexpression networks (COEX).

## 2.2 Basics of Graph and Network Theory

In this section, we formally define graphs and introduce important related notions. We adapt this section from the Chapter *Introduction to Graph and Network Theory* contributed to the book Analysing Network Data in Biology and Medicine, Cambridge University Press, edited by Nataša Pržulj, published February 2019 [116].

### 2.2.1 Definitions

A graph, or network, $G$, is defined by a set of vertices $V$ (also called nodes), a set of edges $E \subseteq V \times V$ (also called links) corresponding to pairs of vertices and representing interactions between vertices, and a function $\omega : E \mapsto \Omega$ associating a label to each edge, where $\Omega$ represents a set of possible labels. A graph is formally denoted by the triplet $G = (V, E, \omega)$. We also use the notations $(V(G), E(G), \omega(G))$ and $(V_G, E_G, \omega_G)$ to denote unambiguously the vertex set, edge set, and labelling function of graph $G$. In this Thesis, we focus mainly on unlabelled graphs, where $\Omega = \{1\}$, for which the function $\omega$ is omitted from the definition, i.e. $G = (V, E)$. Henceforth, to alleviate notations, and unless specified otherwise, a graph is assumed to be unlabelled.

Graphs are used to represent systems of interconnected entities. Each vertex $v \in V$ represents a unique entity, and each edge $e \in E$ traditionally represents a relationship between a pair of vertices $(u, v) \in V \times V$ (also denoted by $uv$ for brevity). The label associated with an edge can represent various properties such as the type of interaction or the confidence level associated with the interaction. An edge can be *undirected* or *directed*. An undirected edge represents a symmetric interaction between vertices, for instance, a physical contact between two proteins. A directed edge represents an asymmetric interaction, for instance, an interaction between a transcription factor and a gene.

A graph is *undirected* if all its edges are undirected. If it also contains no loop (edge connecting a vertex to itself) and no multi-edge (multiple edges connecting the same pair of vertices), the graph is *simple*. There is a wealth of theoretical results for simple graphs that can be found in any textbook dedicated to Graph Theory [117], illustrating their importance. In biology, PPI networks are typically modelled by simple graphs. A graph is *directed* if all of its edges are directed. Directed graphs model transcriptional regulatory networks.

A graph $G = (V, E)$ is *bipartite* if we can divide its vertex set $V$ into two non-overlapping vertex sets $A$ and $B$, such that $V = A \cup B$ and every edge $e$ in $E$ connects a vertex from A to a vertex from B. Bipartite graphs are used to model interactions between two separate sets of entities, for instance, drug–gene interactions.

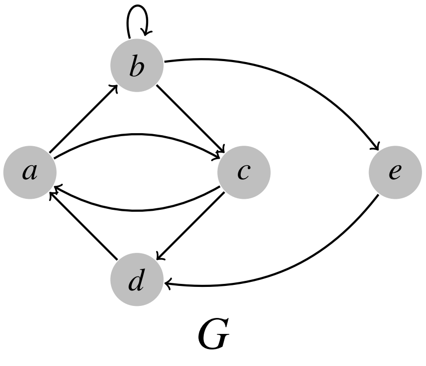
*Multi-layer graphs* [118, 119] and *Hypergraphs* [120, 121] are generalisations of standard graphs that can represent more complex systems and are gaining attention in the literature. The latter is at the core of Chapter 3. For a complete overview of these and additional types of graphs, we refer the reader to the book chapter [116].

#### 2.2.1.1 Graph data structures

Graphs are typically represented as either an adjacency matrix or an adjacency list. For a graph $G = (V, E, \omega)$ containing $n$ vertices, indexed from 0 to $n-1$, i.e. $V = \{0, .., n-1\}$, the adjacency matrix is denoted by $A$, with $A \in \Omega^{n \times n}$. Each row/column index represents a vertex of the graph and the entries of $A$ are such that:

$$A_{i,j} = \begin{cases} \omega_{ij}, & \text{if edge } ij \in E, \\ 0, & \text{otherwise.} \end{cases}$$

Figure 2.1 presents the adjacency matrix of an unweighted directed graph.

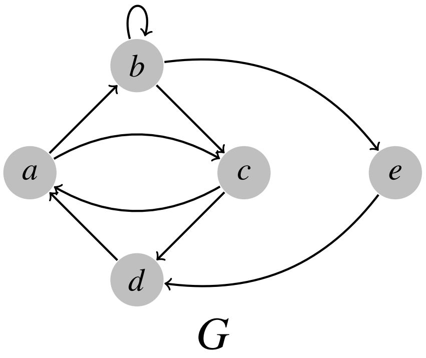


$$A = \begin{array}{c|ccccc|} & a & b & c & d & e \\ a & 0 & 1 & 1 & 0 & 0 \\ b & 0 & 1 & 1 & 0 & 1 \\ c & 1 & 0 & 0 & 1 & 0 \\ d & 1 & 0 & 0 & 0 & 0 \\ e & 0 & 0 & 0 & 1 & 0 \end{array}$$

**Figure 2.1:** An illustration of a graph $G$ and its adjacency matrix A. The 1s represent the existence of a directed edge from the vertex in a row to the vertex in the column, 0s represent the absence of an edge.

For undirected graphs, weighted or not, the adjacency matrix is symmetric, i.e. $A = A^T$. Storing a graph represented by its adjacency matrix requires $O(n^2)$ memory. As a special case, bipartite graphs can be represented by a submatrix of the adjacency matrix where rows are associated to one of the two disjoint sets of the bipartite network, and columns to the other set.

An adjacency matrix structure is preferred if a graph is dense, which means that the number of edges is high (of the order $O(n^2)$) with respect to the total number of possible edges (which is equal to $\binom{n}{2} = \frac{n(n-1)}{2}$ for a simple graph).



$$
\begin{array}{lcc}
L: & & \\
a & \rightarrow & b,c \\
b & \rightarrow & b,c,e \\
c & \rightarrow & a,d \\
d & \rightarrow & a \\
e & \rightarrow & d
\end{array}
$$

**Figure 2.2:** Graph $G$ and its adjacency list, $L$.

An adjacency list is a list of vertices and all of their adjacent vertices, as illustrated in Figure 2.2. To store an adjacency list, one needs $O(m+n)$ memory, where $m$ is the number of edges and $n$ is the number of vertices in the graph. Thus, if the graph is sparse, in the sense that it has $O(n)$ or fewer edges, and the task does not require a matrix representation, then the adjacency list is a more space-efficient structure to store a graph.

#### 2.2.1.2 Degree and neighbourhood

The *degree* of vertex $u$ corresponds to the number of edges incident to $u$ and is denoted by $d(u)$. The *neighbourhood* of vertex $u$ is the set of all vertices adjacent to $u$, formally defined as $N(u) = \{v \in V : uv \in E\}$. Consider graph $G$ of Figure 2.3: vertex $b$ (in red) has degree 3 ($d(b) = 3$) and the vertices in its neighbourhood are circled in red. If the edges are directed, each vertex $u$ has an *indegree* and an *outdegree*. The indegree of $u$ corresponds to the number of edges having $u$ as target vertex and the outdegree of $u$ is the number of edges having source vertex $u$. In graph $H$ of Figure 2.3, these notions are illustrated on vertex $d$: the indegree of $d$ is 2 and the outdegree is 1, illustrated by red and blue edges, respectively.

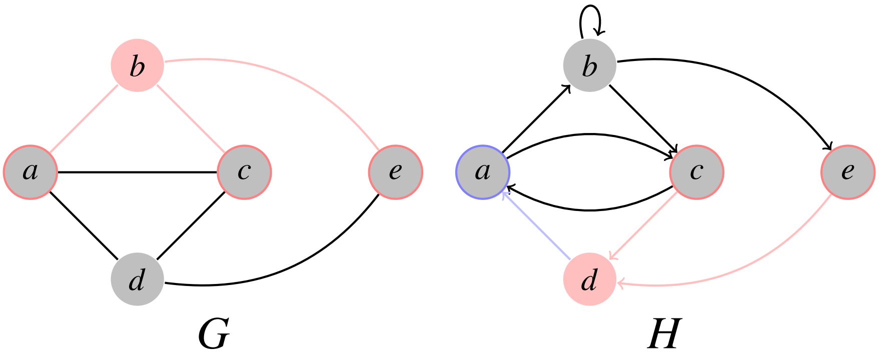


**Figure 2.3:** Examples illustrating neighbourhood and degree notions in undirected graph $G$ and directed graph $H$.

### 2.2.1.3 Subgraphs

A *subgraph* $H$ of graph $G$ contains a subset of vertices of $G$ and a subset of edges connecting those vertices. Formally, if $V(H) \subseteq V(G)$ and $E(H) \subseteq E(G)$, then $H$ is a subgraph of $G$. If $H$ contains all edges from $G$ that connect its vertices, then $H$ is *induced*, or *node-induced*. Consider graph $G = (V,E)$ and set of vertices $A \subseteq V$, then $G[A]$ denotes the subgraph *induced* (or *node-induced*) by $A$ on $G$. Its vertex set is the set $A$ and its edge set includes all the edges in $E$ that have both ends in $A$.

Figure 2.4 illustrates these concepts. In graph $G$, red denotes a non-induced subgraph. $G$ is not induced, because it does not contain *ed* edge. The red subgraph in graph $H$ is induced by the set of vertices $\{a,c,d,e\}$.

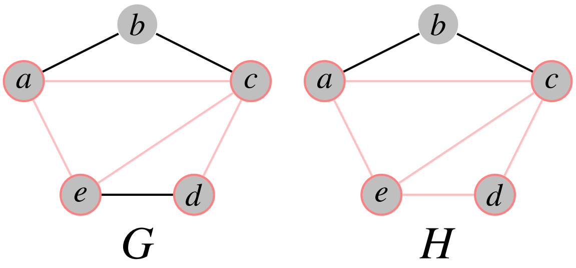


**Figure 2.4:** Examples illustrating in red in graph $G$ a subgraph and in graph $H$ an induced subgraph for the same graph.

#### 2.2.1.4 Graph isomorphism

Two graphs $G$ and $H$ are *isomorphic* if there exists a function $f : V_G \mapsto V_H$ such that $uv \in E(G)$ if and only if $f(u)f(v) \in E(H)$. The function $f$ is a *bijection*, which means that $f$ is a one-to-one mapping of the vertices of $G$ onto the vertices of $H$. Figure 2.5 gives an example of two isomorphic graphs. Finding approximately isomorphic mappings between biological networks is a recurrent problem, such as network alignment tasks [122].

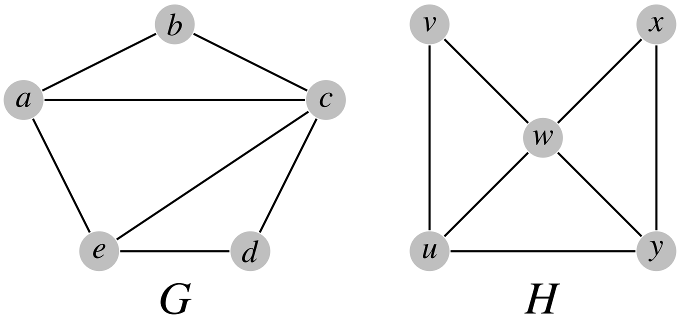


**Figure 2.5:** Examples of isomorphic graphs G and H. The isomorphic function maps the vertices a, b, c, d, and e of graph G to the vertices u, v, w, x, and y of graph H, respectively.

The generalisation of the graph isomorphism problem is to find if a graph contains a copy of another graph. This problem is called the *subgraph isomorphism problem* and has long been known to be NP-complete [123, 124]. In short, it means that there is no fast and efficient method in general. The clique problem is a particular case of the subgraph isomorphism problem. A *clique* is a fully connected subgraph. The clique problem refers to searching for the largest clique in graph $G$. Detection of cliques in molecular networks has been used to identify groups of consistently co-expressed genes [125, 126] and to match three-dimensional structures of molecules [127, 128]. In the latter example, subgraph isomorphism is used to identify similarities between compounds based on theirs substructures.

An exact comparison of networks is a challenging problem due to the NP-completeness of the underlying subgraph isomorphism problem [129]. Thus, simple heuristics, such as the various centrality measures (see [130, 131, 116] for details and other network statistics), have been used to capture network properties through which networks can be analysed and compared. We next introduce *graphlets* that are used to extract topological statistics from graphs and to define heuristics for network analysis.

### 2.2.2 Graphlets

Graphlets are small, non-isomorphic, connected, and induced subgraphs of a simple graph, which are used to characterise and quantify the local wiring patterns around each node of a network [132, 133].

Within a given graph, automorphic nodes are nodes whose labels can be exchanged without changing adjacency relationships. Formally these nodes can be mapped to each other by an *automorphism*, which is an isomorphism of a graph with itself. Each set of automorphic nodes of a graphlet form what is called an *orbit*. For instance, consider the graphlet corresponding to a path of three nodes (first 3-node graphlets in Figure 2.6). This graphlet has two orbits: the first contains the end nodes of the path and the second contains only the central node. For all 2- to 5-node graphlets, there are a total of 73 different orbits (set of automorphic nodes; see Figure 2.6). In practice, due to the exponential increase in the number of graphlets, existing approaches stop at 5-node graphlets.

Orbits are used to precisely capture the wiring pattern around a node of a simple graph. Specifically, the notion of *graphlet degree* of a node is introduced as an extension to the degree of a node that only captures the number of direct neighbours. The $i^{th}$ *graphlet degree* of a node, $v$, corresponds to the count of graphlets in the network containing $v$ and in which $v$ is in orbit $i$. The $0^{th}$ graphlet degree of a node corresponds to the traditional node's degree. For each node, we define a *graphlet degree vector* (GDV) of size 73 with entry $i$ corresponding to the $i^{th}$ graphlet degree of the node. The GDV of a node gives a purely topological description of the wiring pattern around the node. In practice, for each node, we consider the extended neighbourhood, specifically up to the $4^{th}$ neighbourhood of the node which corresponds to all nodes within the shortest path distance of 4 from the node of interest. Within this set of nodes we consider all *connected* and *induced* sub-

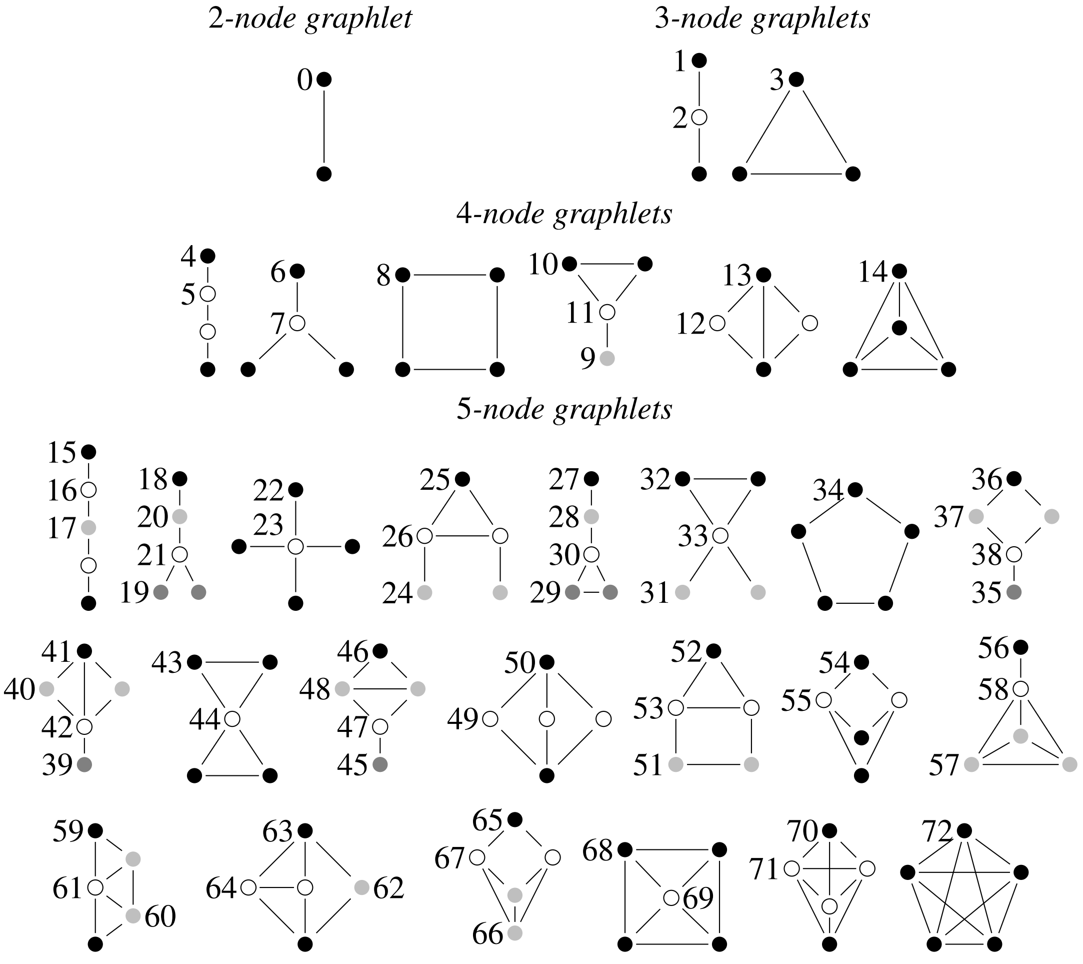
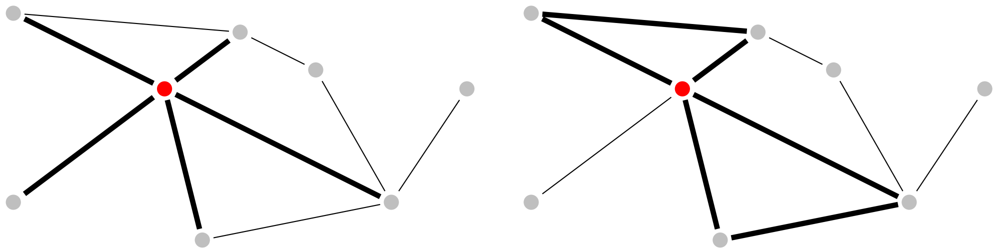


**Figure 2.6:** Representation of the 30 two- to five-node graphlets and their 73 orbits.

graphs that can be formed with 2 to 5 nodes (including the node of interest). For each of these subgraphs, we first identify to which graphlet the subgraph is isomorphic to (i.e. which graphlets it corresponds to) and then identify the orbit the node of interest belongs to and finally we increment the corresponding graphlet degree of the node (see Figure 2.7). This process gives the GDV of the node and is repeated for all nodes of a network.

Based on GDVs of the nodes of a network, one can define a similarity measure to compare the relative wiring patterns of two nodes. This can then be integrated into a kernel, a symmetric matrix capturing similarities (or distance) between entries in an unspecified space (see Section 2.3.2.2 for formal definition), that can further be used to cluster nodes, to relate the wiring patterns of genes in biological networks to their function and to propagate biological annotations to previously uncharacterised



| orbits | 0 | 1 | 2 | 3 | 4 | 5 | 6 | ... | 41 | 42 | 43 | 44 | ... | 72 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| degree | **5** | 3 | 8 | **2** | 0 | 7 | 1 | ... | 0 | 0 | 0 | 1 | ... | 0 |

**Figure 2.7:** Example of the counting process for a node highlighted in red. Thick lines correspond to the bold numbers and highlight specific kind of graphlets. On the left, it corresponds to single edges (orbit 0), giving the degree, 5, of the highlighted node. On the right, we highlight triangle graphlets (orbit 3) containing the highlighted node ($3^{rd}$ orbit degree of the node is 2). Part of the resulting GDV is given in the table below.

genes [134, 135, 136].

The distribution of orbit's degrees can be used to characterise a network. For instance, Yaveroğlu *et al.* [135] proposed the Graphlet Correlation Matrix (GCM) as a network descriptor. Specifically, it computes the correlation between pairs of orbit degree distributions giving a $73 \times 73$ matrix (if all graphlets of size 2 to 5 are considered). This matrix can then be used to compare two networks by defining a matrix distance, such as the Graphlet Correlation Distance (GCD) defined in [135]. Network distances can be used to compare biological networks and uncover their functional organisation [132, 133, 134, 135]. It can also be used to guide network alignment algorithms [137, 138].

A wealth of tools to analyse real-world networks have been built upon graphlets. Note that, the (original) graphlets are defined only for mining simple graphs, which are sometimes not sufficient to capture the complexity of real-world networks. Graphlets have been extended to various kind of networks, among which directed networks [139], ordered graphs (graphs with a global ordering on the vertices) [140], and simplicial complexes [141]. We extend the definition to hypergraphs in Chapter 3, introducing *hypergraphlets*.

Graphlets and extensions offer heuristics to extract node features from complex networks that can then be used as input to machine learning algorithms. Other popular approaches to extract features from graph representations are based on random walks (e.g., DeepWalk [142], node2vec [143]), matrix factorisations discussed in Section 2.3 (e.g., Laplacian eigenmaps [144], HOPE [145]), or graph kernels [146]. For a review of those methods see [147]. In the following section, we discuss the principles of machine learning and the various approaches used to integrate multi-source data.

## 2.3 Machine Learning for Data Integration

With the advent of the Big Data era, the field of *machine learning* has been the focus of growing interests over the last decades due to its ability to detect patterns in large scale datasets. The uncovered pattern can be used, for instance, to predict future data or to highlight fundamental properties of the data source. The underlying mathematical concepts in machine learning are often not new, e.g., neural networks were first used in the 1940s, but their applications were often restricted to small problems and hindered by the limited computing power available at the time. Technological advances have allowed researchers to revisit these ideas and apply them to concrete and data-rich problems leading to some of the most notable achievements, such as the victory of AlphaGO over top human players of the game of GO in 2016 [148]. Designing an algorithm able to master the game of GO was long thought to be too hard or even impossible due to the high complexity of the game. An estimate of the number of legal, possible board settings is of the order of $O\left(10^{170}\right)$. To put it into perspective, researchers estimate that the entire universe contains $O\left(10^{86}\right)$ atoms.

In computational biology, machine learning has been extensively used to address the problem of multi-omics data integration. As discussed in the Introduction, each molecular data type captures different, complementary functional information about cells, tissues, and individuals. Integrating various data type in the same framework, giving more complete representations of the molecular machinery, has proved an efficient strategy in several applications [149, 150, 151]. We give here a brief introduction to data integration, as well as a succinct overview of the existing

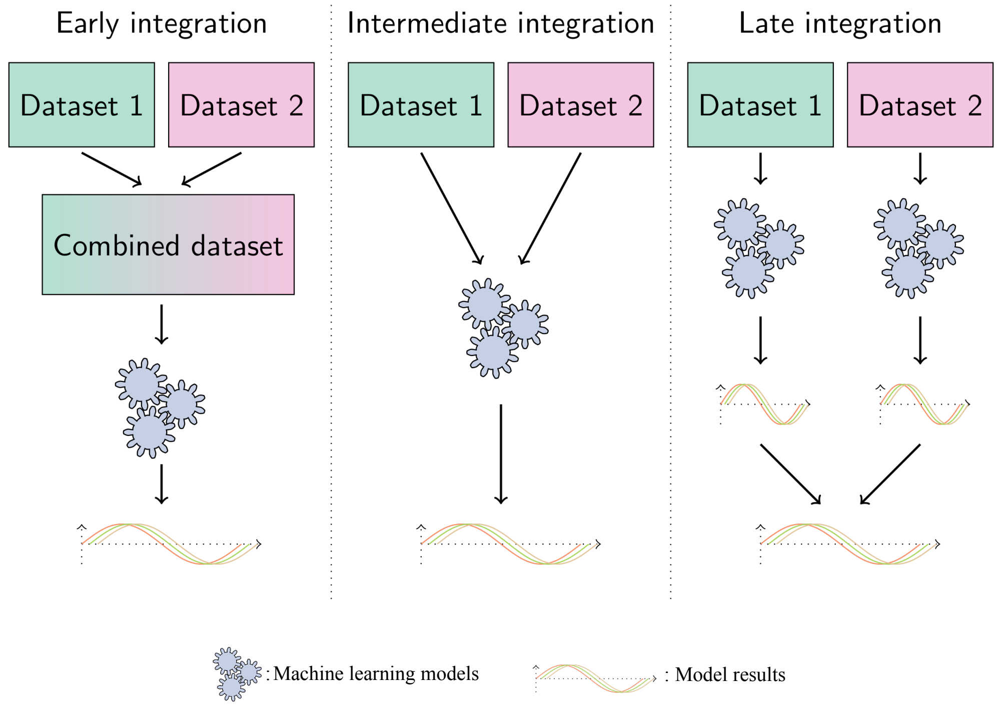


**Figure 2.8:** Illustration of early, intermediate, and late integration approaches.

data integration methods used in computational biology.

### 2.3.1 Definition

Data integration is the process of combining multi-source data into a single, unified view to answer specific questions. Integrative approaches can be decomposed based on the choice of integration process itself. There are three resulting classes: *early*, *intermediate*, and *late* data integration (illustrated in Figure 2.8). Methods that first combine all datasets into one then design a model to analyse the new, integrated dataset fall into the *early* class [152, 153]. Methods in the *late* class start by developing a model for each separate dataset and then use *ensemble learning* approaches to combine the resulting models. The inherent limitations of those approaches are that they do not make use of the datasets' complementarity, which might diminish performance [152, 153]. *Intermediate* integrative approaches aim to infer a joint model over all datasets. Such methods implicitly exploit the complementarity of the data, which can palliate each dataset's intrinsic limitations and seems to lead to better predictive results in practice [154, 152, 155, 153].

Data integration tasks can be further classified as *homogeneous* or *heteroge-*

*neous* depending on the type of data considered. In the former case, datasets capture different information about the same group of entities, obtained, for instance, with biological experiments under different conditions. This is the case when considering multiple networks sharing the same set of nodes, e.g., molecular networks with genes as nodes such as PPI, co-expression networks, and genetic interactions networks [156]. A heterogeneous task involves data relating to different, but interlinked entities. For instance, Gligorijević *et al.* [157] integrate patient genetic mutations data with drug–target interactions as well as molecular networks and drug chemical similarities.

## 2.3.2 Methods for data integration

Here, we review broad (overlapping) classes of machine learning approaches to the data integration task: Bayesian-based methods, Kernel-based methods, Network-based methods, Matrix factorisation-based methods, and neural networks-based methods. We give a brief description of each approach, tying in some of their applications to biomedical data integration.

### 2.3.2.1 Bayesian based integration

A Bayesian network is a directed acyclic graph that captures conditional probabilities between variables (the nodes). More precisely, the network represents a factorisation of the *joint probability distribution* $p(\mathbf{x}|\theta)$, where $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$ are the nodes of the networks and $\theta$ the parameters of the model. The distribution factorises as $p(\mathbf{x}|\theta) = \prod_{i=1}^{n} p(x_i|\mathbf{a_i}, \theta)$, where $\mathbf{a_i}$ is the set of ancestors of $x_i$ in the network, i.e. nodes that are the source of edges connecting to $x_i$. The construction of a Bayesian network requires both learning the network, *structure learning*, and identifying the optimal parameters of the model, *parameter learning*. The variables can be of different types, hence facilitating the integration of heterogeneous data. Bayesian networks have been instrumental in a number of biomedical applications.

For instance, Gevaert *et al.* [158] integrated clinical and microarray data of breast cancer patients in Bayesian networks. The authors then used the resulting network to stratify the cohort into two subgroups that significantly correlate with prognosis. Furthermore, Zhu *et al.* [159] combined genomics, transcriptomics, and interactomics data of yeast using a Bayesian network approach to identify gene regulatory interactions.

Bayesian networks offer a flexible framework for data integration; however, it suffers from a few limitations. First, Bayesian integrative approaches do not scale well. The number of possible Bayesian networks is super-exponential in its number of nodes. This implies that searching for the optimal configuration is an NP-complete problem [160] and inference in a Bayesian network is intractable. Second, Bayesian networks are directed acyclic graphs that are unable to capture feedback loops, which are an essential component of biological systems as they model control mechanisms.

#### 2.3.2.2 Kernel-based integration

Kernel-based approaches first embed a dataset in a higher dimensional feature space. The embedding is given by a function $\phi$ that maps sample $x_i$ of a dataset to a point, $\phi(x_i)$, in the higher dimensional feature space. The embedding is represented by a *kernel matrix*, $K$, capturing the similarity between samples as defined by the kernel function $k(x_i, x_j) = < \phi(x_i), \phi(x_j) >$. The transform, $\phi$, and the feature space need not be specified explicitly. By definition, the kernel is a symmetric, positive semi-definite matrix. The gaussian kernel is arguably the most common kernel that is defined with the kernel function $k(x_i, x_j) = \exp(-\frac{\|x_i - x_j\|_2}{2\sigma^2})$. Another example is the graphlet kernel introduced by Shervashidze *et al.* [161] for graph analysis. Many other graph kernels exist and we refer the reader to the review from Borgwardt *et al.* for more details [146]. Well-known methods that make use of kernels include support vector machines (SVM) [162] and spectral clustering [163].

In an integrative framework, each dataset is represented by a kernel. Each kernel can then be analysed separately before using ensemble learning methods to combine the results (late data integration approach), or the problem can be framed as a multiple kernel learning task (intermediate integration) [164]. Kernel-based approaches enable the integration of heterogeneous data and homogeneous data alike and have been extensively used in biomedical applications.

He *et al.* [165] recently proposed a kernel-based approach to derive disease comorbidities based on the integration of various disease data. The authors first collect disease information, such as associated genes and phenotypes. Then, they define four disease kernels based on 1) interactions of disease genes in a PPI network, 2) overlap of the set of pathways associated to the diseases through their genes, 3) overlap of disease phenotypes, and 4) overlap of the set of functional annotations associated to the diseases through their genes. The authors finally define a rule to integrate those four kernels to derive a final disease–disease similarity score that is interpreted as a comorbidity risk index.

Collier *et al.* [11] introduced LOTUS, a kernel-based integrative method to predict cancer driver genes for both single cancer and pan-cancer settings. In the latter, they first define a gene similarity kernel and a cancer subtype similarity kernel each based on multiple data sources such as PPI network, genes mutation frequency across cohorts, and cancer type-specific features. The two kernels are then combined to create a (cancer type, gene) similarity kernel that is fed to a Support Vector Machine (SVM) model. The SVM is then trained to predict for each (cancer type, gene) pair if, for that cancer types, the gene is either an oncogene, a tumour suppressor, or neither.

Kernels give a practical framework for data integration; however, it has its own limitations. In particular, kernel-based approaches can become quickly expensive when reasoning about multiple groups of entities at a time [8].

#### 2.3.2.3 Network-based integration

Network-based integration methods focus on integrating datasets that correspond to relational data between entities, and that can be represented as networks.

Basic network integration approaches first collapse all networks into one and can be seen as a case of kernel-based integration by interpreting each network adjacency matrix as a kernel. For instance, consider $n$ datasets that can be represented by networks $G_1 = (V_1, E_1), G_2 = (V_2, E_2), \ldots, G_n = (V_n, E_n)$. For homogeneous data, all network representations share a common set of nodes $V (= V_1 = V_2 = \ldots = V_n)$.

Heterogeneous data can be handled by projecting all datasets to a single set of nodes $V$ [166], for instance, the union of all node sets, effectively casting the problem to a homogeneous setting.

The simplest approach to derive one integrated network is to take the union of all networks. Thus, the integrated network is defined by node set $V$ and edge set $E$ containing all the edges from the different networks, $E = E_1 \cup E_2 \cup \ldots \cup E_n$ [167]. However, an issue with this process is that it treats all input networks equally, which might not always be optimal due to the variety of noisiness and completeness levels across datasets. A solution is to take a weighted union of the networks by associating to each network a weight representing its contribution to the optimised, integrated network [168]. This procedure is equivalent to a multi-kernel learning approach applied to the graphs adjacency matrices. Other existing methods rely on more advanced strategies based on message passing theory [169].

Alternatively, methods have been proposed to jointly analyse multiple networks without having to define one unified network. Such approaches typically use diffusion processes that spread information along the edges of the various networks [170]. For instance, Luo *et al.* [171] proposed a network-based approach to repositioning drugs based on bi-directional random walks across two networks representing disease–disease links and drug–drug links. More recently, Ruiz *et al.* [172] introduced a white-box method to link disease and drugs based on random walks through multiple networks: a disease–gene network, a PPI network, a gene–pathway network, a network capturing pathway hierarchy, and a drug–gene network. The authors show that by comparing signatures of random walks starting from drugs to signatures of random walks starting from diseases, they can retrieve known drug indications as well as mechanistic insights into how a drug treats a given disease and how genetic mutations can impact a drug's efficacy.

Network-based integration has been essential in many biomedical tasks; however, those methods are not suited for most omics data that cannot be represented as networks without transformations.

#### 2.3.2.4 Matrix factorisation based integration

Matrix factorisation approaches encompass all methods that decompose a matrix $A$ into the product of $n$ low-dimensional, latent feature matrices $U_i,\ i \in \{1..n\}$, also called latent factors. In practice, this is achieved by finding optimal latent factors $U_i^*,\ i \in \{1..n\}$ such that

$$U_i^* = argmin_{U_i} d(A, \prod_{i=1}^{n} U_i + \varepsilon), \tag{2.1}$$

where $d$ is a function quantifying the distance between its inputs, such as Euclidean distance or Kullback-Leibler divergence, and $\varepsilon$ represents noise. Many models ignore the presence of noise and set $\varepsilon = 0$. Henceforth, to alleviate notation, we will simply represent a matrix factorisation with the notation $A \approx \prod_{i=1}^{n} U_i + \varepsilon$, where the meaning of the symbol $\approx$ implicitly depends on the function $d$ chosen.

Matrix factorisations have been originally developed for applications such as dimensionality reduction, missing data imputation, and clustering. The objective of matrix factorisation methods can effectively be seen as finding embeddings for all entities in a dataset on a low-dimensional, latent manifold under user-defined constraints such that the original data can be best reconstructed.

The most widely known matrix factorisation methods use two factors, i.e. $A \approx UV$, those include principal component analysis (PCA) [173] and non-negative matrix factorisation (NMF) [174]. The former, PCA, solves Equation 2.1 indirectly through the eigendecomposition of the covariance matrix of the dataset, i.e. the covariance matrix of $A$ with our notation. The latter, NMF, solves Equation 2.1 directly through an iterative optimisation procedure, enforcing positivity of the entries of $U$ and $V$ as an additional constraint to facilitate the interpretation of the factors. There exist multiple extensions of NMF that are designed to address different contexts. One example is non-negative matrix tri-factorisation (NMTF) that approximates a matrix by the product of three low-dimensional factors, and that was originally proposed for bi-clustering tasks [175]. NMF-based approaches have gained popularity in the last decade, notably in biomedical applications, due to their successes in various fields and applications. In particular, NMF methods have been used to develop integrative frameworks to mine multi-source datasets jointly.

The underlying mechanism of general matrix factorisation based integration is the collective factorisation of the various datasets through latent factors constraints. These constraints take multiple forms; we discuss two main strategies below, that are sometimes associated together.

The first strategy factorises each dataset in parallel while imposing constraints between factors in different decomposition. For instance, consider two datasets about the same set of entities and represented by matrices $X_1$ and $X_2$. The two datasets could be integrated through the factorisation of both matrices as $X_i \approx U_iV_i,\ i \in \{1,2\}$ with a constraint applied on a function of factors denoted by $f(U_1,U_2)$. The constraint can either be a hard constraint, $f(U_1,U_2) = a$, or an auxiliary objective to optimize in the decomposition, e.g., $\max_{U_1,U_2} f(U_1,U_2)$. The most common formulation sets $U_1 = U_2$, i.e. it enforces the explicit sharing of latent factors across factorisations [176, 177]. In this case, the integration problem can simply be rewritten with a shared latent factor $U$ as $X_i \approx UV_i,\ i \in \{1,2\}$. This effectively means that each entity has a unique embedding across all decompositions. Alternatively, the Partial Least Square (PLS) formulation adds an objective in the joint decomposition: maximising the covariance between pairs of factors [178]. With our notations and our small example, PLS sets $f(U_1,U_2) = \text{covariance}(U_1,U_2)$ and the algorithm aims to find factors $U_1$ and $U_2$ that minimises the factorisation objective function while maximising the covariance between the two latent factors.

The second strategy utilises some data sources to define regularisations for the latent factors of matrix decompositions. Graph regularisation, for instance, has been a popular regularising constraint of latent factors in NMF models. It is used to encourage similar entities, according to data sources typically in the form of a graph (or kernel), to be embedded closer on the latent manifold. For instance, Hofree *et al.* [179] used NMF to factorise a patient–gene matrix representing patients' somatic mutation profiles with graph regularisation based on gene networks that encourage similar embeddings for interacting genes.

Matrix-based data integration has been a popular approach to mine multi-omics datasets. For instance, Malod-Dognin *et al.* [156] derived an integrated cell model, iCell, by integrating three tissue-specific molecular interaction networks in an NMTF-based framework. The authors identified cancer-related genes through the comparative analysis of rewiring in cancer and control iCells. Liu *et al.* [180] proposed a framework based on NMF to integrate heterogeneous physical and functional genomics datasets to predict regulatory chromatin interactions between more than 20,000 promoters and 1.8 million enhancers across 127 human epigenomes.

Matrix factorisations provide a principled framework to analyse multiple homogeneous and heterogeneous datasets collectively. The framework has two main advantages. First, unlike kernel-based approaches, no arbitrary data transformation is required. Note however, that the underlying assumption that the data can be represented in a latent space can lead to some information loss. Second, it enables the joint modelling of all types of relations in the data. Chapter 5 introduces an integrative framework exploiting this feature to jointly model various biomedical entities based on their associations for the analysis of multiple cancers.

#### 2.3.2.5 Deep learning based integration

Deep learning has been one of the main focus of modern Machine Learning owing to its success in numerous applications, such as image processing [181], natural language processing [182], or with powerful algorithms for games such as GO and chess [148]. *Deep learning* is a broad term that used to qualify machine learning models based on artificial neural networks [183]. The epithet “deep” comes from the use of multiple layers in the models. There exist various kinds of layers, such as feed-forward layers, convolutional layers, and recurrent layers [183]. Each layer is typically followed by an activation function that introduces non-linearities. Non-linear activations are essential according to the *universality theorem* [184] that states that any arbitrary function can be parametrised by the superposition of non-linear functions, i.e. by a neural network.

Deep learning has been behind most of the recent progress in Artificial Intelligence. However, there are some specific challenges in its application to biology, as highlighted by the collaborative and thorough review in [185]. The challenges lie in the design and implementation of algorithms that can handle the complexity and heterogeneity of the biological data, and in the interpretability of resulting models and associated results. Model interpretability is a challenge that concerns most methods mentioned in this section. However, it is especially relevant to deep learning due to the high complexity of the models, often presented as black boxes.

However, the deep learning framework is well suited for a variety of data integration problems, and various recent deep learning models have been proposed for the integration of multiple biological data sources. We review some below.

Gligorijevic *et al.* [186] proposed the use of auto-encoders to integrate multiple biological networks for protein function prediction. Auto-encoders models learn both an encoding function that embeds the input into a low-dimensional, latent space, and a decoding function, that reconstructs the input from the embedding. The authors make use of auto-encoders to learn a joint latent space embedding for the networks, then use the learnt features as input to a Support Vector Machine (SVM) model to predict a protein's biological functions.

Deep factorisation models have also been proposed to replace standard matrix factorisation techniques [187, 188]. Those models can be naturally extended to handle data integration problems tackled by matrix factorisation approaches and discussed above. For instance, in the related task of knowledge-graph embedding, Dettmers *et al.* [189] proposed an approach that embeds each entity and each type of relation of the knowledge-graph on a latent manifold, and then uses a model based on convolutional layers to predict (entity,relation,entity)-triplets.

Another approach, baptised Visible Machine Learning, to integrate biological information in the deep learning framework is to use the biological information as prior knowledge to define the structure of neural networks [190]. For instance, Ma *et al.* [191] used the Gene Ontology [192] directed acyclic graph as a template for a feed-forward neural network. The neural network, named DCell, is then trained to predict phenotype related to cellular fitness from genotype data. The trained DCell predicts cellular growth almost as accurately as laboratory observations. We explore a similar approach in Chapter 4.

Additionally, deep learning techniques such as auxiliary losses, model pre-training, and transfer learning, enable researchers to add and combine biological sources in their model. For instance, Yao *et al.* [193] proposed DeepCorrSurv, a method to integrate slide images of cancer patients with omics data for prognosis prediction. The authors used an auxiliary function to maximise the correlation between the projections of each data modality and to extract a common representation used to predict a patient's risk. For the same task, Hao *et al.* [194] recently introduced PAGE-net, a similar model that integrates genomics, slide images, and clinical data from cancer patients, using molecular networks to define part of the network architecture. PAGE-net jointly learns projections for all data modalities, concatenating the resulting features in a final layer that predicts a final patient's risk score.

In this section, we have given a brief description of the various methods used to integrate multi-source datasets, referencing relevant literature in the process. Each of the following chapters will present a more detailed review of the literature relevant to the biomedical problems that it addresses.

# Chapter 3

# The multi-scale protein interactome captures biological function

In this chapter, we propose a new, *multi-scale*, protein interaction *hypernetwork model* that utilises hypergraphs to capture different scales of protein organisation, including PPIs, protein complexes and pathways. In analogy to graphlets, we introduce *hypergraphlets*, small, connected, non-isomorphic, induced sub-hypergraphs of a hypergraph, to quantify the local wiring patterns of these multi-scale molecular hypergraphs and to mine them for new biological information. We apply them to model the multi-scale protein networks of baker's yeast and human and show that the higher-order molecular organisation captured by these hypergraphs is strongly related to the underlying biology. Importantly, we demonstrate that our new models and data mining tools reveal different but complementary biological information compared to classical PPI networks. The content of this chapter is adapted from and extends Gaudelet *et al.* [195], presented at the ECCB'18 conference.

## 3.1 Introduction

In biological systems, molecules do not interact solely in a pairwise fashion. Protein complexes, for instance, are groups of two or more associated proteins. Biological pathways also typically involve multiple molecules. Simple graphs cannot capture the multi-scale organisation of such systems [196, 121]. In the example in Figure 3.1, we observe that the simple graph representation, on the right, of the system on the left blurs the higher-order organisation of the system. Given only the network representation on the right, one might, for instance, falsely assume that the nodes b, c, and d form a complex of three elements, while it is true that b and d form a complex, b and c form a complex, and c, d and e form a complex.

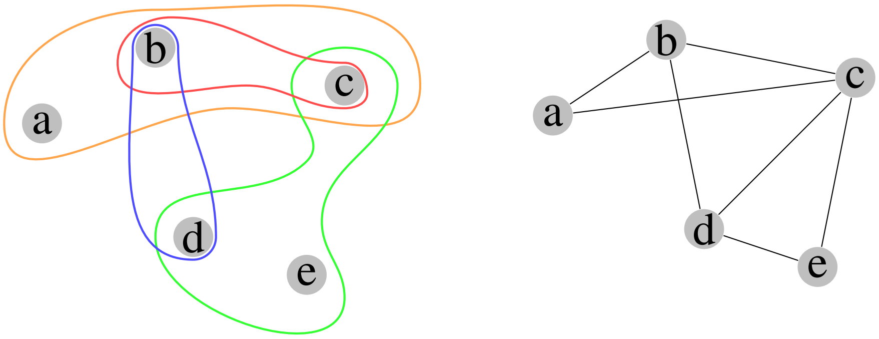


**Figure 3.1:** Illustration of a system with higher order interactions (left) and its simple graph representation (right).

A solution to overcome this limitation is to model a molecular system using hypergraphs [196, 121]. A *hypergraph* is defined by a set of nodes, $V$, and a set of edges, $E$, called *hyperedges*, where each hyperedge corresponds to a set of interacting nodes of any size [120]. This means that a simple graph is a special case of a hypergraph in which all hyperedges are sets of two nodes. The representation of the system in Figure 3.1 (left) is a hypergraph. To analyse data modelled as hypergraphs, it is necessary to develop methods to mine the structure of hypergraphs. A number of simple measures from graph theory have already been extended to hypergraphs, e.g., the clustering coefficient [197], degree distribution [198], and centralities [197, 199]. However, hypergraphs lack more advanced descriptors of local topology. Hence, we introduce hypergraphlets, an extension of graphlets to hypernetworks.

We investigate biological hypernetworks in which nodes are proteins and hyperedges capture PPIs, protein complexes, or biological pathways. The main aim is to check if the topology of these hypernetwork representations of the data carries biological information that goes beyond the information that can be obtained from PPI networks. We use hypergraphlets in this investigation.

## 3.2 Contributions

We motivate studying the higher-order molecular interactions as models that capture additional and different biological information than the widely studied PPI networks. We introduce hypergraphlets as a new tool that unveils the observation of the close link between the multi-scale molecular organisation and biological function. Hypergraphlets can serve as an underlying methodology for many new tools to further study the multi-scale organisation of molecular systems.

We analyse the hypergraph representation of protein interactions of yeast *saccharomyces cerevisiae* and human and show that proteins that are similarly wired in a hypernetwork, independently of their location in the hypernetwork, tend to have similar biological functions. Also, we use the Canonical Correlation Analysis (CCA) [200] to correlate hypergraphlets around proteins in these networks with their biological functions. The results confirm the link between the local wiring patterns of the multi-scale molecular organisation of the cell and biological functions. We use these findings to predict biological functions of uncharacterised proteins from the wiring patterns of the multi-scale molecular organisation. We validate our predictions in the literature.

## 3.3 Materials & Methods

### 3.3.1 Data

We consider six different protein networks across two species, human and baker's yeast. For each species, we consider the protein-protein interaction (PPI) network and two hypernetworks corresponding to protein complexes and biological pathways. Depending on the hypernetwork considered, a hyperedge represents either a protein complex or a biological pathway. These data are used jointly to build hypernetworks capturing multi-scale organisation of proteins in a cell, as detailed in Section 3.3.6 below.

The PPI data is obtained from the BioGRID database [6] (version 3.4.145). Both pathways hypernetworks come from the Reactome database [7] (accessed in April 2017). We only collect the lowest pathways in the Reactome hierarchy. The human protein complexes are downloaded from the CORUM database [201, 202] (all complexes file, accessed in May 2017), while the yeast protein complexes are collected from the CYC2008 database [203] (last updated in 2009). Table 3.1 gives an overview of the sizes of the data sets.

| | Database | # proteins | # (hyper-) interactions |
|---|---|---|---|
| Human | CORUM | 3,145 | 2,138 |
| | Reactome | 9,466 | 1,461 |
| | PPI | 16,008 | 216,865 |
| Yeast | Reactome | 1,465 | 400 |
| | Cyc2008 | 1,607 | 406 |
| | PPI | 5,931 | 87,225 |

**Table 3.1:** Sizes of the data.

To investigate the links between networks and biological functions, we collect gene annotations from the Gene Ontology Consortium (GO) database [192] (downloaded at the end of January 2017). For each protein, we keep only the most specific annotations that are experimentally derived. We separate the annotations based on the three categories: Biological Process (BP), Molecular Function (MF), and Cellular Component (CC). In the analysis, we focus on Biological Process annotations (GO-BP) as they cover a broader set of proteins than the other two categories.

### 3.3.2 Hypergraphlets: the local topology of hypergraphs

We define *hypergraphlets* as small, connected, non-isomorphic, induced sub-hypergraphs of larger hypergraphs. Berge [120] defines an induced sub-hypergraph of a hypergraph $H = (V, E)$ on a set of nodes $A \subset V$ as the hypergraph $H_A$ with set of nodes $A$ and set of unique hyperedges

$$E_{H_A} = \{e \cap A | e \in E, e \cap A \neq \emptyset\}. \tag{3.1}$$

Note that with this definition, hyperedges containing only one node exist for each node. With this definition, an induced hypergraph is simple, i.e. it has no duplicated edges.

Within a given hypergraph, automorphic nodes are nodes whose labels can be exchanged without changing adjacency relationships. Formally these nodes can be mapped to each other by an *automorphism*, which is an isomorphism of a hypergraph with itself. A set of automorphic nodes form what is called an *orbit*. Here, we consider all 1- to 4-node hypergraphlets, which contain a total of $6,369$ different orbits. For 5-node hypergraphlets, we estimate that there are more than a hundred thousands orbits; hence we restrict ourselves to 4-node hypergraphlets. In Figure 3.2, we illustrate all 65 orbits that occur in the 1- to 3-node hypergraphlets.

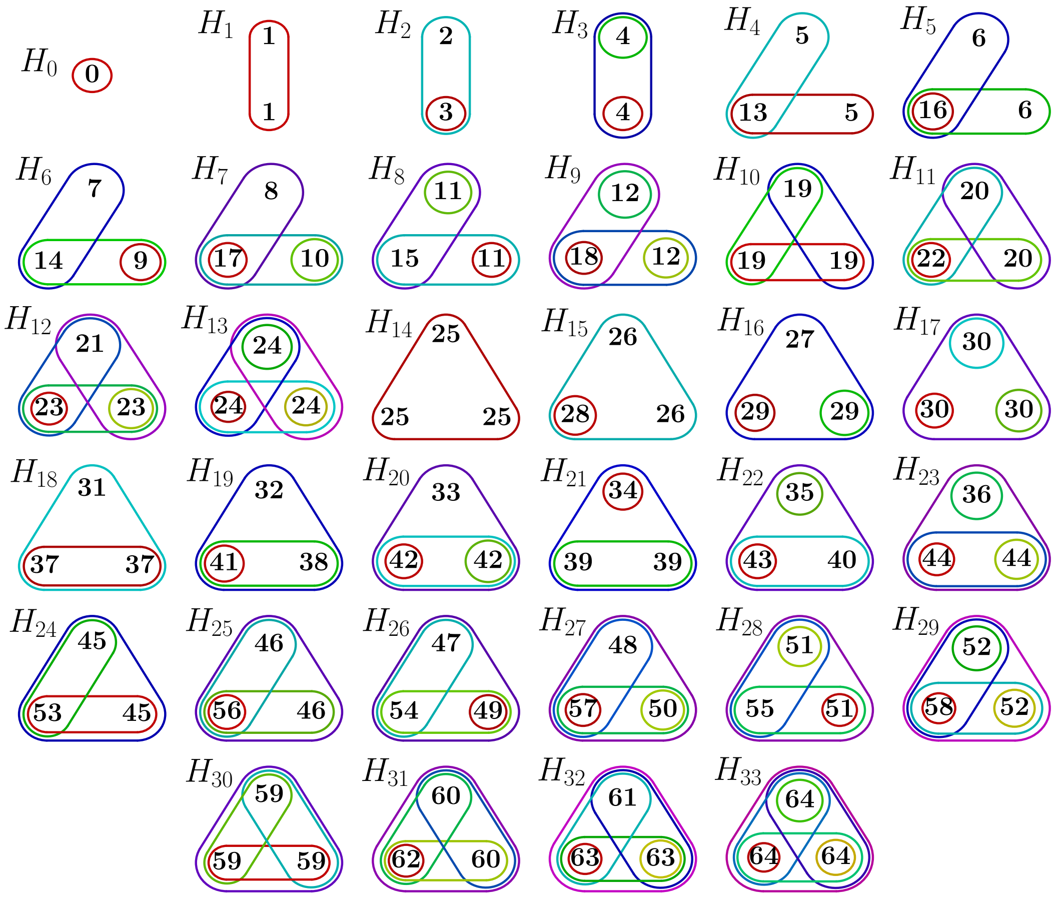


**Figure 3.2:** Illustration of all 1- to 3-node hypergraphlets ($H_0$ to $H_{33}$) and the 65 orbits. Each closed set corresponds to a hyperedge and each node is represented by an integer between 0 and 64 corresponding to the orbit it belongs to.

Analogous to graphlets, we use hypergraphlet orbits to quantify the wiring patterns around each node in a hypergraph. For each orbit $i$ in hypergraphlet $h$, we define the $i^{th}$ *hypergraphlet degree* of a node in the hypergraph $H$ as the number of hypergraphlet orbits $i$ that the node touches.

For each node in a hypergraph, we compute all $6,369$ hypergraphlet degrees resulting in a $6,369$-dimensional vector where entry $i$ corresponds to the $i^{th}$ hypergraphlet degree of the node. We term this vector capturing the local wiring around a node the *Hypergraphlet Degree Vector (HDV).*

Considering a hypergraph with $n$ nodes, with maximal size of hyperedge $l$ and with a maximal degree of a node $d$, where the *degree* of a node corresponds to the number of hyperedges that contain it, an upper bound on the complexity of counting all 1- to $k$-node hypergraphlets is $O(n(ld)^{k-1})$.

Lugo-Martinez *et al.* [204] introduced an alternative definition of hypergraphlets in the context of binary classification problems. They define kernels based on their definition of hypergraphlets and use support vector machines to classify the proteins. The key difference with our definition of hypergraphlets is that they consider the hypergraphlets of a hypergraph as *partial* sub-hypergraphs, thus ignoring some overlaps between hyperedges [204]. A partial sub-hypergraph of a hypergraph $H = (V,E)$, on a set of nodes $A \subset V$ and a set of hyperedges $J \subset E$, is defined as the hypergraph $H_{A,J}$ with set of nodes $A$ and set of unique hyperedges

$$E_{H_{A,J}} = \{e \cap A | e \in J, e \cap A \neq \emptyset\}. \tag{3.2}$$

The counting process in [204] is decomposed in two steps. In the first step, they ignore all hyperedges containing more than four nodes, i.e. $J = \{e \in E | card(e) \leq 4\}$. Hyperedges with more than four nodes are taken into consideration independently in the second step, which decomposes a hyperedge of size $n > 4$ into the $\binom{n}{4}$ subsets of four nodes. Hence, their counting process overlooks an important part of the hypernetwork's topological information. This motivates our redefinition of hypergraphlets as a direct extension of graphlets. However, we could not compare the two approaches, as their implementation is not publicly available, and they recently agreed with us that their definition needed to be changed to alleviate these issues[1].

[1] Personal communication.

### 3.3.3 Related approaches to analyse hypergraphs' topology

A classical approach in the literature when studying a hypergraph $H = (V, E)$ is first to project it to a binary graph. Two often used methods to achieve this are the Clique Expansion (CE) and the Star Expansion (SE) [205].

The Clique Expansion (CE) transforms each hyperedge of a hypergraph in a clique. The weight $\omega$ associated to an edge between nodes $u$ and $v$ is such that it minimises the sum $\sum_{e \in E: u,v \in e} (\omega - \omega_e)^2$, where $\omega_e$ is the weight associated to the hyperedge $e$. As we consider hypergraphs with unweighted hyperedges, from the weight formula, their CE will be unweighted graphs. As a baseline for comparison to evaluate hypergraphlets, we use 2- to 5-node graphlet counts on the CE of a hypergraph.

The Star Expansion (SE) transforms a hypergraph $H = (V, E)$ into a bipartite graph $B$. The sets $V$ and $E$ give the two disjoint node sets of $B$, and the edge set is defined by $\{(v, e) : v \in V, e \in E, v \in e\}$. Furthermore, SE associates a weight to each edge $(v, e)$, corresponding to the ratio of the weight of hyperedge $e$ to its size. We do not use this method for three reasons. First, it has been shown that the hypergraph and bipartite graph representations are not statistically equivalent [197]. Second, by definition, bipartite graphs do not have 3-cycles. This limits the number of features that would be given by graphlets. Lastly, graphlets have not yet been extended to handle weighted graphs.

Another approach is to view the hypergraph as a simplicial complex from computational geometry [206]. We define simplicial complexes in Appendix A.1. In short, a hypergraph can be seen as a simplicial complex where each hyperedge $e$ is a simplex of dimension $n - 1$, where $n$ is the number of nodes the hyperedge connects. To mine simplicial complex, Malod-Dognin *et al.* [141] recently introduced *simplets*, an extension of graphlets to simplicial complexes. Hence, one can compute for each node of a simplicial complex a Simplets Degree Vector (SDV). As in [141], we limit ourselves to simplets up to four nodes; thus, our SDVs are 32-dimensional. The main difference between the hypergraph and simplicial complex representations is that, with the latter, any subset of the nodes of a simplex is also a simplex, which is not the case with hyperedges and can lead to information loss. Figure 3.3 gives an illustration of this distinction: we observe that in the simplicial complex representation, we lose the information that nodes b and c interact independently of the triangle a, b, and c. Effectively, simplets are a restriction of hypergraphlets, and we can define a mapping from HDV to SDV. We include the simplets statistics in our analysis.

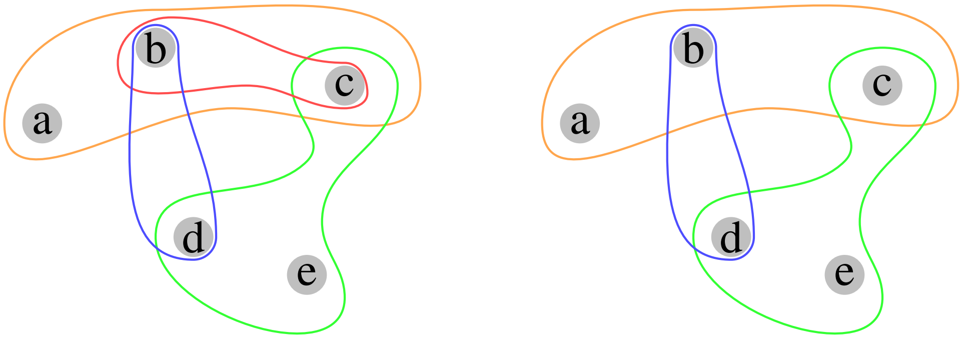


**Figure 3.3:** Hypergraph (left) and simplicial complex (right) representations of the same underlying system. Each closed set corresponds to either a hyperedge of the hypergraph or a facet of the simplicial complex.

### 3.3.4 Topological distance

We define a distance measure to compare the wiring patterns of two vertices in a hypernetwork (or network, depending on the model considered) as follows. Consider a set of proteins $P = \{p_1, p_2, \ldots, p_m\}$ and let $M$ be the matrix representing our data where row $i$ corresponds to the HDV, SDV or GDV of protein $p_i$. Then, we define the distance, $\delta$, between two proteins $p_i$ and $p_j$ as

$$\delta(p_i, p_j) = \left[ \sum_{k \in K} \left( \frac{\log\left(M_{ik}+1\right) - \log\left(M_{jk}+1\right)}{\sigma_k} \right)^2 \right]^{\frac{1}{2}}, \tag{3.3}$$

where $K$ corresponds to the set of orbits considered, $M_{ik}$ denotes the entry of $M$ on the $i^{th}$ row and $k^{th}$ column, and $\sigma_k$ denotes the standard deviation of the distribution of the $k^{th}$ hypergraphlet (or graphlet) orbit degree across our set of data value. We apply to $M$ an element-wise log transformation to reduce the impact of very large orbit counts. Note that this distance is a proper metric as it corresponds to the Euclidean distance between rescaled topological feature vectors (using logarithm and orbit-specific rescaling).

### 3.3.5 Linking local structure to function

We explore two ways to evaluate the link between the local structure of a molecular network and the biological functions of its molecules. First, we cluster the nodes based on the similarity of their wiring patterns defined in Section 3.3.4, and we do the enrichment analysis of the resulting clusters (Section 3.3.5.1). Second, we use CCA to test if biological functions tend to be characterised by specific wiring patterns (Section 3.3.5.2).

#### 3.3.5.1 Cluster enrichment

We cluster proteins that are similarly wired in a hypergraph, a graph or a simplicial complex as measured by distance $\delta$ (see Equation 3.3) between the HDVs, GDVs or SDVs of the nodes and test if the proteins within the same cluster share GO functions.

More precisely, clusters are obtained by using the k-means method [207] based on the distance between pairs of degree vectors of proteins defined in Equation 3.3 (see Appendix A.2). To account for the randomness in the k-means method, we run the clustering algorithm 50 times for each tested value of $k$. For each clustering obtained, we compute the enrichment of its clusters in biological annotations for each GO category. We use the Benjamini-Hochberg procedure to correct for multiple hypothesis testing [208]. We consider a cluster enriched if at least one GO annotation is significantly enriched in the cluster (p-value $< 5\%$). For each value of $k$, we also compute the average of Sum of Squared Error (SSE) and the Normalised Mutual Information (NMI) [209] across all 50 repeats. SSE gives a measure of how close proteins within a cluster are on average according to our similarity measure, and NMI evaluates the stability of the clustering across the 50 runs, i.e. if proteins are consistently clustered together or not. Then, to choose the optimal number of clusters, we use ”the elbow” heuristics on the SSE and NMI plots. For the resulting number of clusters, we select the clustering giving the highest enrichment percentage across the 50 runs of k-means. We test the significance of the enrichment with random permutation tests: we keep the same number and size of clusters and randomly assign proteins to each cluster and measure the enrichments of the resulting clusters. We repeat this process $1,000$ times and compute the significance.

To see whether the different models capture the same or different but complementary biological information, we compare the results across models. First, we quantify the similarity between two models' clusterings with the Adjusted Mutual Information (AMI) [209] score. It measures if any pair of proteins are consistently clustered together or apart in both clusterings, adjusting for chance. Second, we compute the overlap between two models' set of enriched GO annotations using the Jaccard Index [210].

### 3.3.5.2 Canonical correlation analysis

CCA is used to infer correlations between two sets of features, $X$ and $Y$. Consider features $X = (X_1, ..., X_n)$ and $Y = (Y_1, ..., Y_m)$ over the same elements. Then CCA will identify $K$ pairs $(\mathscr{L}_X^k, \mathscr{L}_Y^k)$, called *canonical variates*, of linear combinations of features of $X$ and of features of $Y$, with $K = \min(m,n)$, such that the correlations of $\mathscr{L}_x^k$ and $\mathscr{L}_y^k$ are maximal over all $k$. Each canonical variate is associated a score corresponding to the correlation between its two linear combinations.

In our case, the elements are proteins, the first set of features corresponds to the wiring patterns of proteins in networks or hypernetworks, and the second to the biological functions of proteins from GO. As mentioned above, each protein (node) has a GDV from the PPI network, a HDV from the hypernetwork, a GDV from the CE of the hypernetwork, and a SDV from the simplicial complex derived from the hypernetwork. Hence, we have four matrices of topological features where entries $(i, j)$ correspond to the $j^{th}$ graphlet's, simplet's or hypergraphlet's orbit degree of protein $i$. Also, we associate to each protein a vector of GO–BP annotations. In this vector, an entry is equal to 1 if the gene is annotated with the corresponding GO term, and 0 otherwise. Hence, we form a matrix of biological features, where entries $(i, j)$ correspond to the presence or absence of GO annotation $j$ for protein $i$.

We compute CCA for each combination of topological features and biological annotations to uncover topology-function relationships in the data.

### 3.3.6 Summary of the analysis

As stated above, our main aim is to examine if modelling the higher order of molecular organisation harbours additional biological information and to demonstrate that the wiring patterns of biological hypernetworks are strongly linked to the underlying biology.

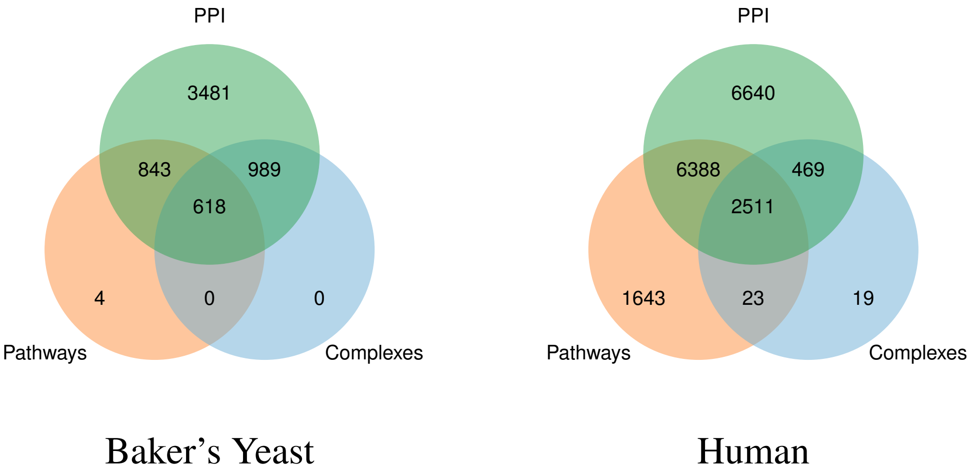


**Figure 3.4:** The overlaps of the protein sets of baker's yeast (left) and human (right). Left: 3,481 proteins participate in PPIs only, 843 in PPIs and pathways, 618 in PPIs, pathways and complexes, 989 in PPIs and complexes, while 4 are in pathways only. Right: 6,640 proteins participate in PPIs only, 6,388 in PPIs and pathways, 2,511 in PPIs, pathways and complexes, 469 in PPIs and complexes, 23 in complexes and pathways, while 1,643 are in pathways only and 19 in complexes only.

We first focus on parts of PPI networks that we know are rich in biological information: protein complexes and pathways. Not all proteins in a PPI network belong to complexes or pathways (see Figure 3.4). Hence to validate our method, we consider four sets of proteins of the PPI networks: those belonging to pathways in human (human-pathways), those belonging to pathways in yeast (yeast-pathways), those belonging to complexes in human (human-complexes), and those belonging to complexes in yeast (yeast-complexes). For each protein in each of these sets, we have four topological signatures: one from the standard graphlets counted on the entire PPI network, one from the hypergraphlet counts in the hypergraph (HG) that we constructed by using only protein complexes (and equivalently pathways), one from the simplet counts on the simplicial complex (SC) derived from the hypergraph, and one from the graphlet counts on the clique expansion (CE) of the hypergraph. For each protein, we also have biological signatures derived from GO annotations. We use these as input into the methods described in Sections 3.3.5.1 and 3.3.5.2. The results of these validations are presented in Sections 3.4.1.1 and 3.4.1.2.

The reason for doing these validations on the sets of data for which we know that they are enriched in biological information (i.e., pathways and complexes) is to demonstrate that our new model and method can correctly identify the biological information. After these validations of the methodology, we use it to perform the analysis of multi-scale protein interaction network data of yeast and human and uncover new biological information. In particular, for each species, we construct a hypergraph that contains all of its PPIs, all of its protein complexes, and all of its pathways; i.e., nodes are proteins, and hyperedges correspond to PPIs, protein complexes, and pathways. The results of analysing these hypergraphs with our methods are presented in Section 3.4.2.

## 3.4 Results & Discussion

### 3.4.1 Validation of our methodology

#### 3.4.1.1 Enrichment Analysis

Having computed the topological vectors from all network models (PPI, HG, CE, and SC) for each protein of each of the four sets of proteins described in Section 3.3.6 (human-pathways, human-complexes, yeast–pathways and yeast–complexes), we apply the methodology detailed in Section 3.3.5.1 to investigate if similarly wired proteins have similar functions. Interestingly, the percentage of enriched clusters is relatively stable as we increase the number of clusters. Hence, any partitioning of the proteins based on the local wiring patterns in a network, quantified by using graphlets, simplets or hypergraphlets, captures the underlying biological information (see Figure 3.5). This result highlights the crucial role played by the way proteins interact in determining protein function without any information about their sequence, or interacting partners. Furthermore, when examining the clusterings obtained at a specific number of clusters, *k* (see Section 3.3.5.1 for details on how *k* is chosen), we observe that the enrichments (top table in Figure 3.6) are all statistically significant, except for the one in grey. Importantly, clusters obtained from models capturing higher-order interactions are more enriched than those obtained from PPI networks. This result shows the importance of higher-order protein interactions, highlighting their links to the underlying biological information. However, based on this enrichment analysis, we can not conclude which representation is best among CE, SC, and HG models.

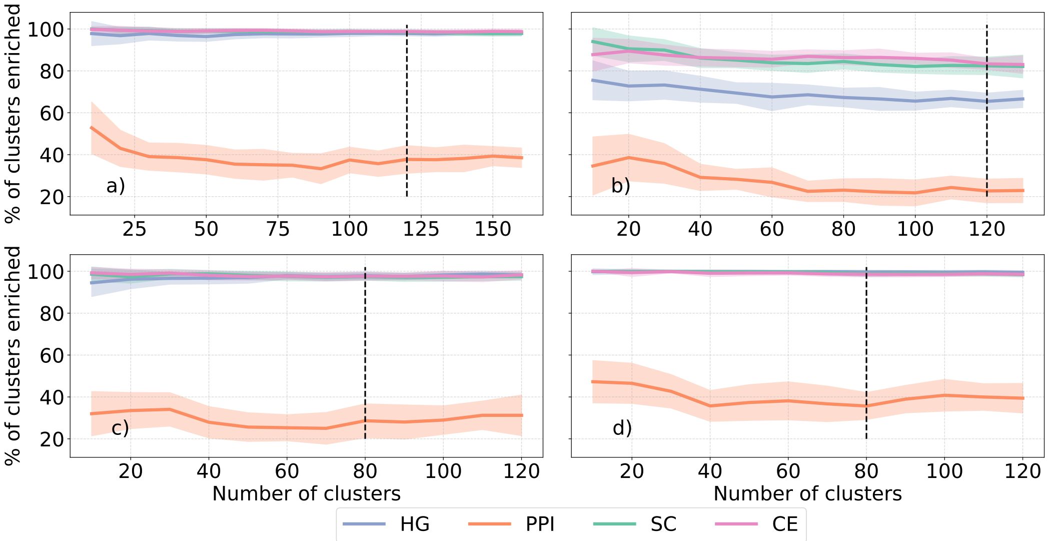


**Figure 3.5:** The panels give the average percentage of clusters enriched with respect to the total number of clusters for a) human-pathways, b) human-complexes, c) yeast-pathways, and d) yeast-complexes. The coloured areas around the lines represent the standard deviation. The colors represent the models from which the clustering is obtained. We only represent the enrichments with respect to GO–BP annotations, the findings are similar for the other type of annotations. The black vertical lines signal the number of clusters selected from the set of NMI and SSE curves according to the procedure described in Section 3.3.5.1.

To further investigate the clusterings, we compute for each the average shortest path distances between pairs of proteins belonging to the same clusters (”within-clusters”) and between pairs of proteins which are in different clusters (”between-clusters”; see the bottom panel in Figure 3.6). We observe a wider gap between

| | HG | PPI | CE | SC |
|---|---|---|---|---|
| Human-pathways | 98.2% (111) | 59.2% (120) | 100% (120) | 100% (117) |
| Human-complexes | 94.3% (105) | 40.3% (119) | 96.4% (111) | 99.1% (112) |
| Yeast-pathways | 100% (71) | 45% (80) | 100% (74) | 100% (76) |
| Yeast-complexes | 100% (51) | 53.75% (80) | 100% (47) | 100% (51) |

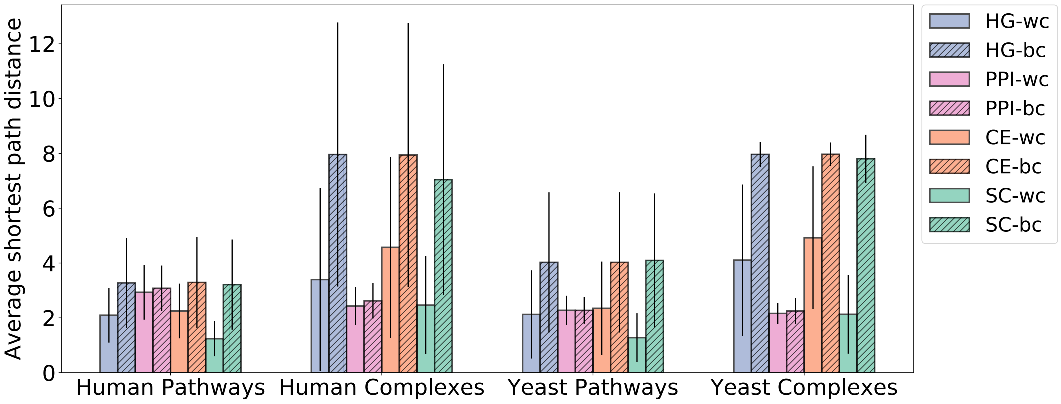


**Figure 3.6:** The top table presents the maximum enrichment measured across clusterings obtained with the "optimal" number of clusters (80 for yeast and 120 for human). The number in parenthesis is the number of non-empty clusters. The color indicates the statistical significance of the maximum enrichment with respect to random permutation tests: black indicates a significant value, grey a non-significant one. The bottom panel gives, for each type of model the average of the shortest path lengths within the clusters (wc) and between clusters (bc) of the best clustering obtained for GO–BP annotations. The results are similar for other GO categories.

within-cluster and between-clusters average shortest path lengths for clustering obtained from the higher-order molecular organisation than from clusterings obtained from PPI networks. Hence, proteins that are topologically similar in the HG, CE, and SC models in addition to sharing biological functions tend to be at shorter distances from each other. This result is consistent with the literature on ”guilt by associations”, which predicts protein functions from their neighbourhoods in molecular networks [211].

Finally, we observe that the clusterings obtained from the PPI model are different from those obtained from the three alternative models both in terms of GO annotations that are enriched and in terms of clustered proteins (see Figures A.1 and A.2 in Appendix). This is because a Jaccard Index close to 0 means that the sets of the enriched GO terms in the PPI and HG clusterings tend not to overlap. Also, AMI scores below 0.1 mean that pairs of proteins belonging to the same clusters in one clustering are typically in different clusters in the other clustering. This demonstrates that higher-order interactomes uncover new biological information that cannot be uncovered from the analysis of PPI networks. Also, it demonstrates the complementarity of the multi-scale representations and that they are capturing different underlying biological information. However, despite observing relatively high AMI scores, greater than 0.5, between the clusterings obtained from HG, CE and SC models (see Figure A.1), the Jaccard Indices measured remain low in comparison, with scores below 0.4 in most cases (see Figure A.2). This result indicates that the different representations of higher-order interactome do not capture the same underlying information.

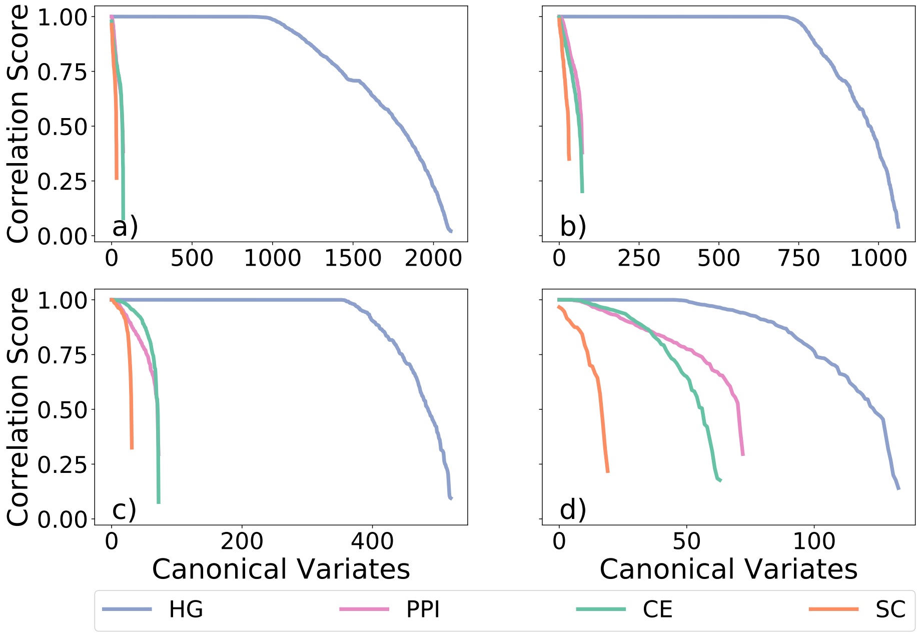


**Figure 3.7:** Canonical correlation score distribution for a) human-pathways, b) human-complexes, c) yeast-pathways, and d) yeast-complexes. The canonical variates represented are all statistically significant (p-value $\leq 5\%$) and are sorted by correlation score. The colours represent the models and the topological signatures from which the canonical variates are obtained.

#### 3.4.1.2 Canonical Correlation Analysis

We investigate the existence of specific topology-function links, i.e. the connection between specific hypergraphlets (or simplets or graphlets) and GO annotations by using CCA described in Section 3.3.5.2. We apply it on the same yeast and human data used in the clustering and enrichment analysis (Section 3.4.1.1): for each set of proteins, we compute the CCA between the topology–containing vectors of each of the associated models (PPI, CE, SC, and HG) and the vector of GO annotations for each category (BP, MF, and CC).

We observe that each model has a number of canonical variates with correlation scores close to 1 (Figure 3.7), which indicates a strong topology-function relationship in these data that was previously highlighted in the context of economic network data [135]. In particular, this means that some functions are strongly linked to specific wiring patterns, and thus, local topology can potentially be used for predicting protein functions. For that purpose, hypergraphlets of HGs have a strong advantage over graphlets of PPI networks or of CEs and the simplets of SCs in the number of canonical variates with a score close to 1, which is 3 to 13 times more variates with HGs.

In Figure A.3, we take a closer look at the most significant CCA variate obtained between HDVs of the proteins of yeast-pathways and their GO–BP annotations. The variate score of 1.0 links a linear combination of GO annotations to a linear combination of hypergraphlets orbits. For instance, this means that a gene annotated with positive regulation of barrier spectrum assembly (GO:0010973) will likely have a relatively large $2644^{th}$ orbit degree in the hypernetwork. Why these specific orbits are linked to these functions is a question that is outside of the scope of this study, and that needs to be further investigated. We find that the GO terms identified here are also biologically coherent: each of the GO–BP terms denoted in blue text in Figure A.3 is annotating at least one protein conjointly with at least one other annotation, that is also denoted in blue text in A.3, according to QuickGO [212]. Furthermore, the only remaining annotation, cell cycle arrest (GO:0007050), has been linked to the MAPK pathway in the literature [213], as have been most of the other terms [214, 215]. Hence, the entire set of GO annotations presented in Figure A.1 is biologically coherent, which validates the relevance of the canonical variate and of our hypergraph-based methodology in capturing functional information.

### 3.4.2 Analysing multi-scale molecular organisation

To explicitly capture the multi-scale organisation of protein interactions, we model them by a hypernetwork containing all PPIs, all protein complexes and all biological pathways as hyperedges (detailed in Section 3.3.6). To assess if the wiring patterns in our new proteins interaction hypernetwork capture the biological functions of its nodes, we do the clustering and enrichment analysis (Section 3.3.5.1), as well as the canonical correlation analysis (Section 3.3.5.2) on these hypernetworks of baker's yeast and human using the same HG, SC, and CE models. We compare the results with those that we obtain by applying the same methodologies to PPI networks.

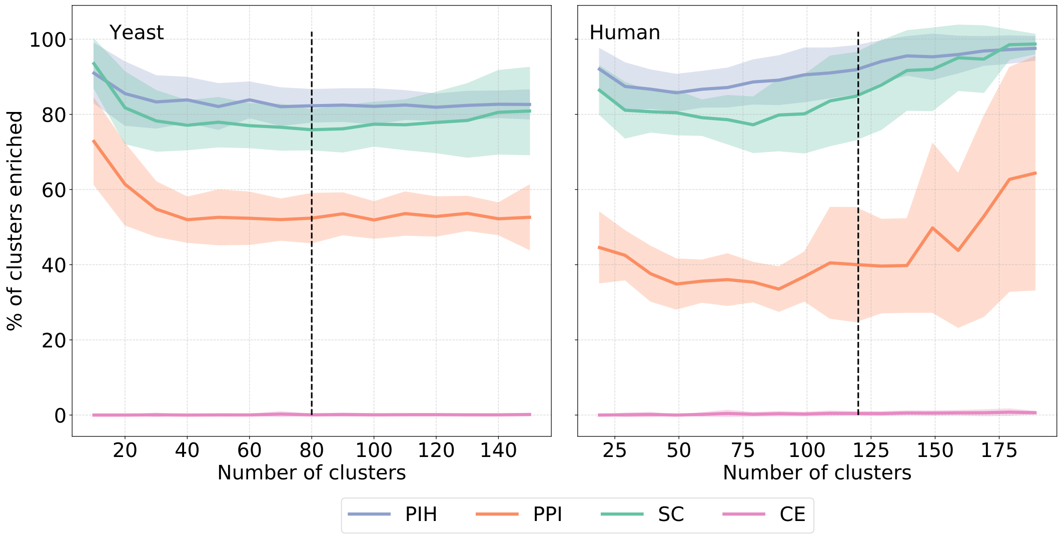


**Figure 3.8:** The panels give the average percentages of clusters enriched with respect to the total number of clusters for a) human and b) yeast. The coloured areas around the lines represent the standard deviation. The colours represent the models from which the clustering is obtained. We only represent the enrichments in terms of GO–BP annotations. Similar results are obtained for the other types. The black vertical lines denote the number of clusters selected from the set of NMI and SSE curves according to the procedure described in Section 3.3.5.1.

In these unifying hypernetwork models of multi-scale molecular organisation, we observe that clusterings of the proteins based on their topological vectors in a network, obtained by using simplets or hypergraphlets, capture the underlying biological information (see Figure 3.8). Furthermore, the clusters obtained from the hypernetwork topology lead to higher enrichments in GO–BP annotations compared to the one obtained from the PPI topology. This shows that our newly proposed

| | HG | PPI | SC |
|---|---|---|---|
| Human | 100.0% (105) | 41.7% (120) | 100.0% (118) |
| Yeast | 91.1% (79) | 68.75% (80) | 90% (80) |

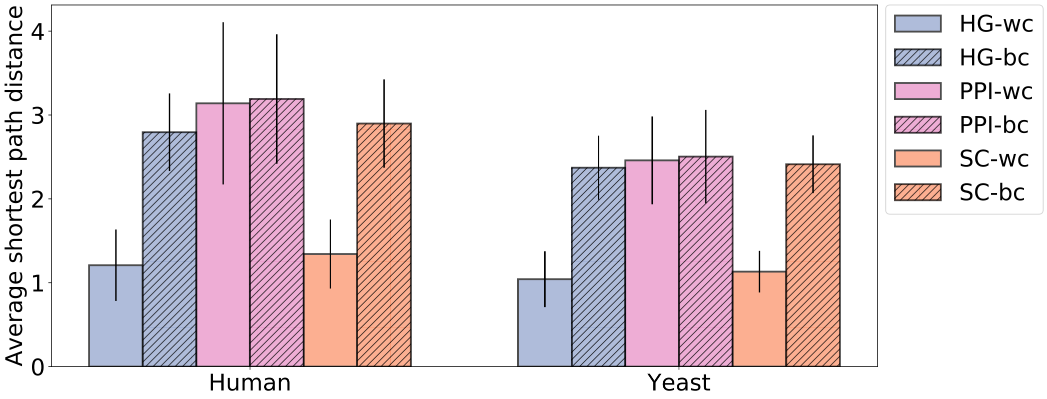


**Figure 3.9:** The top table presents the maximum enrichment measured across clusterings obtained with the ”optimal” number of clusters (denoted by the black vertical lines in the top panels) for GO–BP annotations. The number in parenthesis is the number of non-empty clusters. All enrichments are significant. The bottom panel gives, for each type of model, the average of the shortest path lengths within the clusters (wc) and between clusters (bc) of the best clustering obtained for GO–BP annotations. The results are similar for other GO categories and are not presented here.

model, regardless of the choice of the total number of clusters, $k$, captures more protein biological function in its topology than the standard PPI networks. Interestingly, using graphlets on the CE of the hypergraphs leads to poor performances, with the percentage of clusters enriched consistently close to zero. This result motivates further the need for statistics, such as hypergraphlets and simplets, that can capture the multi-scale organisation of proteins interactions. Due to the poor results, we exclude the CE model from the rest of our analysis. On a different note, we observe that our hypergraphlets perform better than simplets in the sense that we obtain a higher percentage of clusters enriched, on average. This observation suggests that taking into consideration *sub-interactions* (e.g. the interaction between b and c in Figure 3.3 within the hyperedge $\{a,b,c\}$) is necessary to mine our unified hypernetwork.

When choosing the number of clusters, $k$, according to the criteria detailed in Section 3.3.5.1, we observe that all enrichments are statistically significant and that the HG models allow for an increase of over 15% in the number of enriched

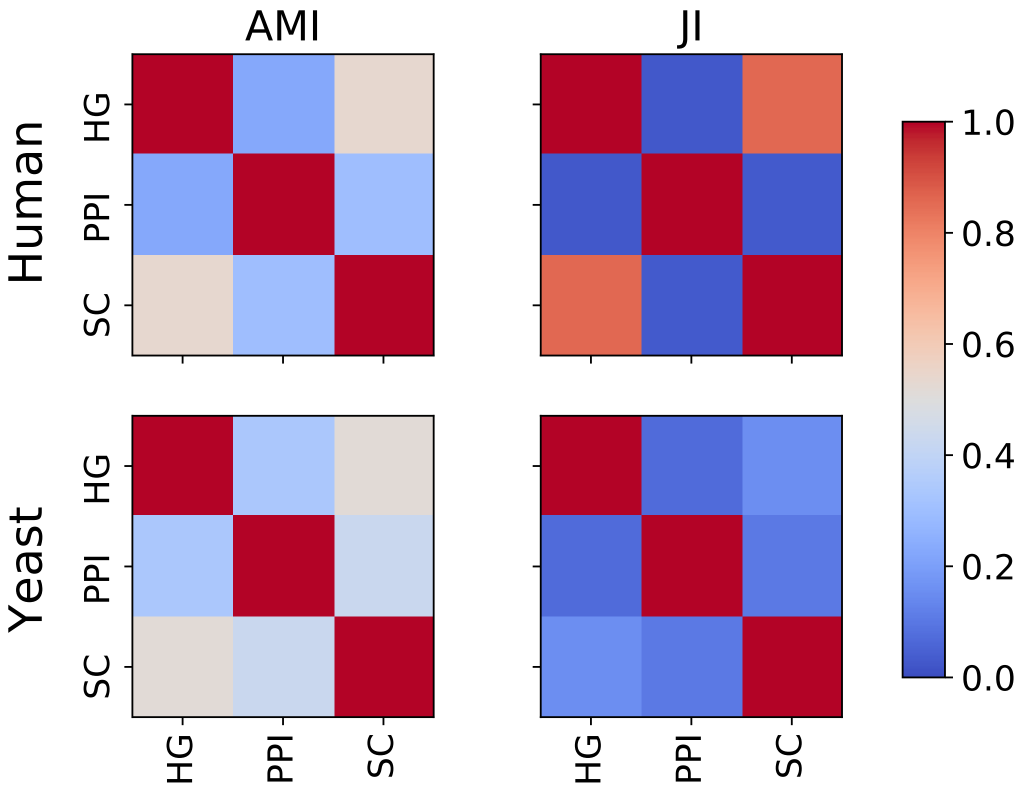


**Figure 3.10:** The panel represents the results of the comparison of the clusterings obtained with the different models. We compute the Adjusted Mutual Information (AMI, left heatmaps) between the clusterings and the Jaccard Index (JI, right heatmaps) between the sets of enriched GO–BP terms for human (top heatmaps) and yeast (bottom heatmaps).

clusters when compared to the PPI networks (see the table in Figure 3.9). This finding underlines the link between multi-scale interaction patterns and biological functions. Interestingly, when investigating the clusters, we observe that a majority of the proteins in the non-enriched clusters only have reported PPIs, but not any associated pathways or complexes. This observation holds for 59% of the proteins in the hypernetwork of yeast and 38% of the proteins in the hypernetwork of human. This might be due to the incompleteness of the pathways and protein complexes data. Our results indicate that when more complete data on complexes and pathways become available, our methodology will be able to extract additional biological information.

We observe that proteins clustered using topological features derived from representations of the multi-scale molecular organisation tend to also be closer in terms of shortest path distances compared to those obtained by clusterings based on the topology of PPI networks (see the bottom panel in Figure 3.9). Interestingly, most proteins clustered together in the HG and SC models are direct neighbours or second neighbours. Hence, the fact that we obtain enriched biological functions in those clusters is consistent with empirical evidence showing that 70-80% of interacting proteins share at least one function. Those evidences were the motivation for the *majority rule* often used in the literature for functional prediction [211].

Finally, we observe that the clusterings obtained from the PPI models are different from those obtained from the HG and SC models both in terms of GO annotations that are enriched, with a Jaccard Index below 0.2, and in terms of similarity of clusters, with an AMI below 0.4 (see Figure 3.10). This confirms that our multiscale model is not equivalent to the standard PPI network and uncover additional biological information complementary to that of the PPI network. When considering the human hypernetwork, we observe that the clusterings obtained from the HG and SC models share a relatively high number of enriched GO–BP annotations with JI score around 0.8. However, this observation does not hold with the yeast hypernetwork. This result confirms further that despite the link between the two models discussed in Section 3.3.3, they are not equivalent.

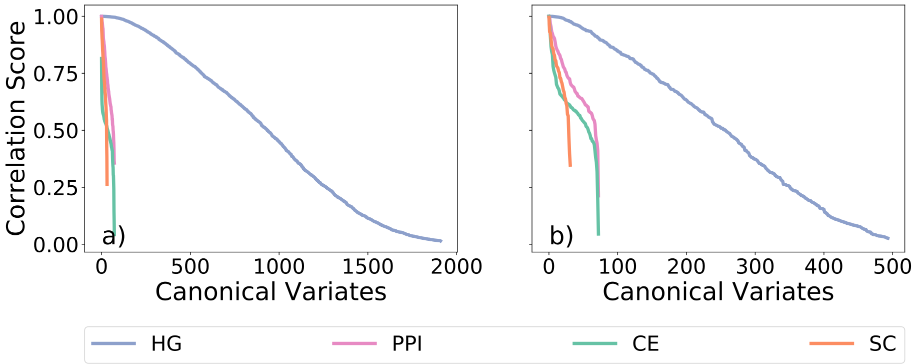


**Figure 3.11:** Canonical correlation score distribution for human hypernetwork. The canonical variates represented are all statistically significant (p-value $\leq 5\%$) and are sorted by correlation score. The colours represent the model and the topological signatures from which the canonical variates are obtained: HG in blue and PPI in orange.

Using CCA (Section 3.3.5.2), we observe that each model has high scoring canonical variates, which indicates that some functions are strongly linked to specific wiring patterns (see Figure 3.11). For that purpose, hypergraphlets of our new HG models have an advantage over graphlets of PPI networks and simplets of the SC models in the number of canonical variates with high correlation score: it has over 300 canonical variates with a score greater than 0.9 compared to only 10 for PPI networks and 4 for SCs. This result indicates that the HG model's local wiring patterns are more correlated with the underlying biology that those of the PPI networks.

Finally, we use the clusterings to investigate the potential of our newly proposed models in conjunction with our hypergraphlets to predict protein functions. As demonstrated above, we identified clusters of proteins with significantly enriched GO annotations. We use these clusters to predict the functions of proteins. For each GO category, we identify two disjoint sets of proteins in each of our hypernetworks: the set of proteins that are experimentally annotated with at least one of the enriched GO terms in their cluster (on which the enrichment computations are based) and the set of proteins that have some predicted annotations in the GO database.

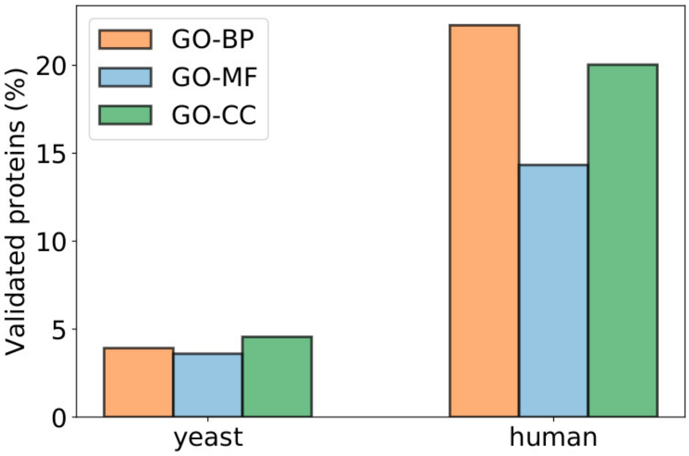


**Figure 3.12:** Percentages of proteins that have at least one of the enriched terms of their clusters in their set of predicted GO annotations (obtained from the GO database [192]). The values correspond to the number of such proteins out of the number of proteins that have at least one putative annotation in the GO database and are not experimentally annotated with any of the enriched terms of their clusters.

| GO ID | Proteins Symbol |
|---|---|
| **GO:0006334** | **HIST1H2AJ** |
| GO:0001580 | LOC107987462; LOC107987425; LOC102725035 |
| **GO:0006364** | **LOC101929876** |
| GO:0035987 | MIR711 |
| GO:0051292 | MIR4260 |
| **GO:0016579** | **MIR6764** |
| **GO:0030199** | **MIR3606** |
| **GO:0030216** | **KRTAP4-7** |
| GO:0006997 | LOC101060521; MIR1181 |
| **GO:0052695** | **UGT2A2**; LOC102724788; GUCY2EP |

**Table 3.2:** Top 10 predictions of GO–BP annotations. The bold font highlight predictions validated by litterature curation.

First, we consider the second set and investigate how many of those proteins have at least one of the enriched terms of their cluster as their predicted GO annotation [192]. For GO–BP, this set contains $11,686$ proteins for human ($4,161$ for yeast). For GO–MF, it contains $7,243$ proteins for human ($3,586$ for yeast). For GO-CC, it contains $6,589$ proteins for human ($3,510$ for yeast). We show that out of these proteins, about 5% for yeast and 15–23% for human have been putatively annotated in GO with at least one of our enriched functions in their clusters (see Figure 3.12), which validates our approach.

Second, we focus on the proteins of the hypernetworks that are unannotated in GO database (this corresponds to 994 proteins for human and 97 proteins for yeast) and investigate the GO–BP annotations we predict for them. We predict function for each of these proteins by associating it with the enriched experimentally obtained GO term that annotates the most proteins in its cluster. We survey the literature to validate our predictions[2] for human (see Table 3.2, the predictions are sorted by statistically significance of the enrichment). We predict that HIST1H2AJ is involved in nucleosome assembly (GO:0006334), which is confirmed in the literature [216].

For microRNA mir–3606, we predict a role in collagen fibril organisation (GO:0030199). Collagen plays a crucial role in cell adhesion, which can involve integrin [217, 218] and mir–3606 has been linked to integrin in the literature as

[2] All predictions are available online at http://www0.cs.ucl.ac.uk/staff/natasa/hypergraphlets/

it has been suggested that mir–3606 can bind to ITGA4 (integrin subunit alpha 4) [219]. We propose that LOC101929876 (40S ribosomal protein S26) is involved in rRNA processing (GO:0006364), which is corroborated by the Reactome database in which the protein is associated with a major pathway of rRNA processing in the nucleolus and cytosol [7]. We also find that microRNA mir–6764 is linked to protein deubiquitination (GO:0016579), which is backed by the Reactome database in which the microRNA is associated to the deubiquitination pathway [7]. We further predict that KRTAP4-7 is linked to keratinocyte differentiation (GO:0030216). Reactome database [7] links KRTAP4-7 to the pathway responsible for keratinisation (GO:0031424) which is a child term of keratinocyte differentiation in the Gene Ontology [192]. We finally find that UGT2A2 is involved in cellular glucuronidation (GO:0052695) which is confirmed by the fact that UGT2A2 is a member of the UDP glucuronosyltransferase family which is responsible for the process of glucuronidation [220].

These results confirm the ability of our hypergraphlets to predict biological functions of proteins from the wiring patterns in our novel model capturing multi-scale organisation of proteins in a cell.

## 3.5 Conclusion

We highlight the importance of considering the higher-order organisation of protein interactions in conjunction with the standard PPI networks. We propose a novel methodology, hypergraphlets, to quantify the local wiring patterns of hypergraphs. We apply it to biological hypernetworks representing protein complexes and pathways of yeast and human and demonstrate a strong link between hypernetwork structure and the function of the proteins. Our novel methodology can mine the biological information hidden in the multi-scale architecture of the molecular organisation. Furthermore, our analysis highlights the superiority, in terms of uncovering the underlying biology, of our multi-scale model when compared to the standard PPI networks. Additionally, we demonstrate that our new hypernetwork model, combined with our hypergraphlets, can be used for functional predictions.

Despite a simple, functional prediction approach, we obtain promising results when using hypergraphlets on our new multi-scale model for functional predictions. It would be interesting to train an advanced machine learning model, such as random forest, using HDVs as features to improve predictions.

Finally, we have demonstrated that the union of networks capturing the multi-scale molecular organisation is strongly linked to the underlying biology of the molecules. However, as discussed in the previous chapter, the simple union of networks give the same weights to each network which, precisely, here blur lines between the different scale of interactions. In the following chapters, we explore integrative approaches that preserve the hierarchical nature of the multi-scale protein interactome. Our methods focus primarily on identifying pathological molecular mechanisms to understand and classify diseases better.

Chapter 4

# Multi-scale protein interactome as prior knowledge to unveil new disease, pathway, and gene associations

In Chapter 3, we have demonstrated that the multi-scale protein interactome captures strong biological signals even with a simple model that does not explicitly use the hierarchical structure of the multi-scale protein interactome.

In this Chapter, we utilise the hierarchical links between protein and biological pathways as prior knowledge to design a Visible Machine Learning model. Specifically, we propose a neural network with a structure based on the multi-scale organisation of proteins in a cell into biological pathways to predict the diagnosis of patients based on differential gene expression. Importantly, through the analysis of our trained model, we uncover disease–disease, disease–gene and disease–pathway associations. The results presented in this chapter are published in Gaudelet *et al.* [221].

## 4.1 Introduction

Symptoms and affected tissues often describe a disease. However, to give a definite diagnosis, physicians often need to analyse patient samples (e.g., blood samples, or biopsies) for typical disease indicators, commonly referred to as disease biomarkers. These may include dysregulated genes, or pathways [222, 223]. Taking into consideration the history of a patient's past and present conditions identifying genetic predispositions, as well as considering associations between diseases, aid in achieving accurate diagnostics and treatments [224]. By also taking advantage of the increasing availability of large scale molecular data, precision medicine aims at improving the understanding of the molecular base of diseases on an individual basis, as well as the relationships between different conditions [225, 226]. The benefits from such work are multiple and include drug re-purposing and identification of new disease biomarkers to improve treatments and diagnoses.

Many studies have investigated disease–gene and disease–pathway associations to improve diagnoses [227, 228, 229, 230]. For instance, Zhao *et al.* [229] propose a ranking of disease genes based on gene expression and protein interactions using Katz-centrality. Hong *et al.* [230] design a tool that identifies significantly disrupted pathways by comparing patient gene expression against controls collected from other experiments. Cogswell *et al.* [231] identify putative gene and pathway biomarkers through change in miRNA in Alzheimer's disease. In specific cancers, Abeel *et al.* [227] use support vector machines and ensemble feature selection methods to select putative gene biomarkers.

A key issue is that most of these studies consider diseases in isolation, i.e. comparing patients having a disease of interest to healthy individuals; thus, the predicted biomarkers could be shared between various diseases. This limits the discriminative potential of such studies for accurate diagnoses. Indeed, network medicine has shown that diseases can share significant molecular background, as evidenced by numerous studies based on patient historical records [1, 224, 232], biological knowledge of the diseases [225, 226, 165], or patient gene expression profiles [46]. For instance, Goh *et al.* [225] build a disease network, which connects diseases that share at least one gene which, when mutated is linked to both conditions. Lee *et al.* [226] construct a disease network of metabolic diseases, connecting pairs of diseases for which associated mutated enzymes catalyse adjacent metabolic reactions.

Hidalgo *et al.* [1] take a different approach by building a disease network based on disease comorbidities, i.e. two diseases are connected if they tend to co-occur significantly in the patient populations. They used Medicare records of elderly patients to build the network. He *et al.* [165] propose PCID (Predicting Comorbidity by Integrating Data), an approach to predict disease comorbidities by aggregating disease similarity scores derived from different data including protein–protein interactions (PPIs), pathways, and functional annotations.

Sánchez-Valle *et al.* [46] define a disease network, named the Disease Molecular Similarity Network (DMSN) based on patient's differential expression profiles. In their study, the DMSN is generated using positive and negative relative molecular similarities (RR) to measure disease similarity and dissimilarity, respectively, that is then interpreted as an estimate of risk. First, a patient-patient similarity network is generated based on patients' differential expression profiles similarities. Next, using the relative similarity score, diseases are related to each other. The resulting network is directed, and each edge is associated with a positive or negative label indicating either an increased or decreased risk of developing the target disease if the patient has the source disease. The underlying assumption is that having a given disease can increase the risk of developing a disease characterised by a similar gene expression profile.

In these various approaches, a key issue is that either a single data source is used, such as disease–gene mutational data [225], or no new biological knowledge about a specific disease could be derived from the results (e.g., PCID [165]).

## 4.2 Contributions

In this work, we propose an integrative framework based on artificial neural networks (NN) to predict disease–disease links, as well as disease–pathway and disease–gene associations. We train the model to predict patients' diagnoses based on differential gene expression. The NN's structure is designed to mimic the cellular multi-scale functional organisation by integrating gene–pathway annotations. This approach follows on from the Visible Machine Learning body of work introduced in Section 2.3.

We show that our framework achieves good predicting performances on our dataset. By analysing the trained NN, i.e. the underlying weight matrices, we show that we can extract biological knowledge relevant for each disease. Specifically, we use the trained NN to predict novel disease–pathway and disease–gene links and from those predictions we extract disease similarity score used to identify putative comorbidities. We show that our predictions are biologically relevant against established ground-truths and verify the top predictions through manual literature curation ensuring that the sources do not use the same data, to mitigate the risk of argument circularity.

## 4.3 Material & Methods

### 4.3.1 Datasets

The base data used in this project was provided by our collaborators Sanchez-Valle *et al.* [46]. It consists of multiple datasets of gene expressions captured by micro-array technology [233]. The datasets are downloaded from Gene Expression Omnibus (GEO, [234]) and ArrayExpress [235] databases. Each dataset contains measurements from healthy (controls) and affected (patients) subjects. For a given dataset, the measurements originate from bulk samples extracted from the same tissue in each subject. Not all datasets use the same tissue for measurements, as diseases do not necessarily affect the same tissue. Each patient is diagnosed with a single disease. For comparisons, the data is normalised by using the frozen robust multiarray procedure [236] to remove experimental bias. Furthermore, to remove tissue effects, each patient sample is normalised against all the control samples of its original dataset using the Limma method [237]. Up to this point, the data is identical to those used to derive the DMSN network [46]. Then we use the corrected p-values output by Limma to define, for each patient, a vector with size corresponding to the number of genes and in which the $i^{th}$ entry is equal to 1, $-1$, or 0 depending on whether the $i^{th}$ gene is significantly (with 5% cutoff) over-, under-, or normally expressed, respectively, for that patient. Additionally, we exclude patients that have no significantly dysregulated genes, as we cannot learn anything from them.

The set of diseases is curated by hand for associations with Disease Ontology codes [238], standard ICD9 and ICD10 codes, MeSH terms, and OMIM codes [239]. Some of the datasets come from studies investigating subtypes of diseases that are studied by projects linked to other datasets. Based on the number of patients in each study, either these datasets were merged with the more global disease or the patients associated with the more global disease were dropped from the study. Specifically, we drop the global disease if the subtype has many more patients associated with it and merge otherwise. Finally, we exclude diseases that have less than 10 associated patients to capture disease heterogeneity in the final dataset and to have sufficient data for each disease to split in a training and testing set.

Pathway annotations were collected from Reactome database [240] (accessed December 2018). Only the lowest pathways in the hierarchy are considered to avoid dealing with pathway interactions (i.e. pathways containing other pathways). The hierarchy of pathways defined in Reactome could be used to define additional layers in the network. This was not tested as we expect the low number of pathways on the highest level to constrain the overall architecture too much for our number of target classes. Of those pathways, only the ones that have a Traceable Author Statement (TAS) are kept. In total, we consider 1,708 pathway annotations.

The final dataset contains 4,788 samples (patients) diagnosed by one of 83 diseases (see Appendix Table B.1). In total, 20,525 genes have their expressions measured. However, only a subset is used as input to our method described in the following Section, as we restrict ourselves to genes associated with at least one pathway, which leaves 9,247 genes.

### 4.3.2 Neural network based data–integration framework

We propose a neural network (NN) predicting a patient diagnosis based on differential gene expression. The structure of the neural network is based on molecular organisation, more specifically gene–pathway annotations downloaded from Reactome (see Figure 4.1). We integrate molecular organisation data into our model to reflect the idea that complex diseases, such as cancer, can be the results of the perturbations of groups of genes, as opposed to a single gene. Using Reactome data allows us to incorporate prior knowledge into our model in the form of biologically meaningful groupings of genes.

A feed-forward neural network can be expressed as a series of matrix multiplications interleaved with non-linear functions, formally the output $\mathbf{Y}$ of a neural network with $n-1$ hidden layers can be written as

$$\mathbf{Y} = f_n\left(\mathbf{W}_n f_{n-1}\left(\dots f_1((\mathbf{W}_1\mathbf{X}))\right)\right)$$

where $\mathbf{X}$ represents the input data, $\mathbf{W}_i$ the weights of layer $i$, and $f_i(\cdot)$ the non-linear function applied to the output of the $i^{th}$ layer. The optimization problem can be written as the minimization of the loss function $\mathscr{L} = g\left(\hat{\mathbf{Y}}, \mathbf{Y}\right)$, where $\hat{\mathbf{Y}}$ is the objective, or ground truth, and $g(\cdot)$ is a predefined function.

Here, we use the softmax function[183] as the last non-linear function of the NN (common choice for multiclass classification problems) and the hyperbolic tangent non-linear function for hidden layers to allow a hidden unit to have values lying in $[-1,1]$ to represent up- and down-regulations. We use the classical cross-entropy function to define the loss function. Our neural network architecture has only one hidden layer capturing gene–pathway links. Hereafter, we refer to this model as GPD (for Gene–Pathway–Disease). Multinomial logistic regression (MLR) and the proposed GPD architecture can be written as

$$\mathbf{Y}^1 = s\left(\mathbf{W}^1\mathbf{X}\right) \tag{4.1}$$

$$\mathbf{Y}^2 = s\left(\mathbf{W}_2^2 \tanh\left(\mathbf{W}_1^2\mathbf{X}\right)\right) \tag{4.2}$$

where $s$ is the softmax function and tanh denotes the hyperbolic tangent. Matrices $\mathbf{X}$ and $\hat{\mathbf{Y}}$ represent our data. Each column of $\mathbf{X}$ corresponds to the differential gene expression of a patient, and each column of $\hat{\mathbf{Y}}$ corresponds to a patient's diagnosis (the prediction of which is the objective of the framework). $\mathbf{W}^1 \in \mathbb{R}^{n_d \times n_g}$ and $\mathbf{W}_2^2 \in \mathbb{R}^{n_d \times n_p}$ correspond to fully-connected layers. The layer corresponding to $\mathbf{W}_1^2 \in \mathbb{R}^{n_c \times n_g}$ represents biological pathway membership of the genes, i.e. the trainable weights of the matrix correspond only to entries $(i, j)$ where gene $j$ is part of the $i^{th}$ pathways.

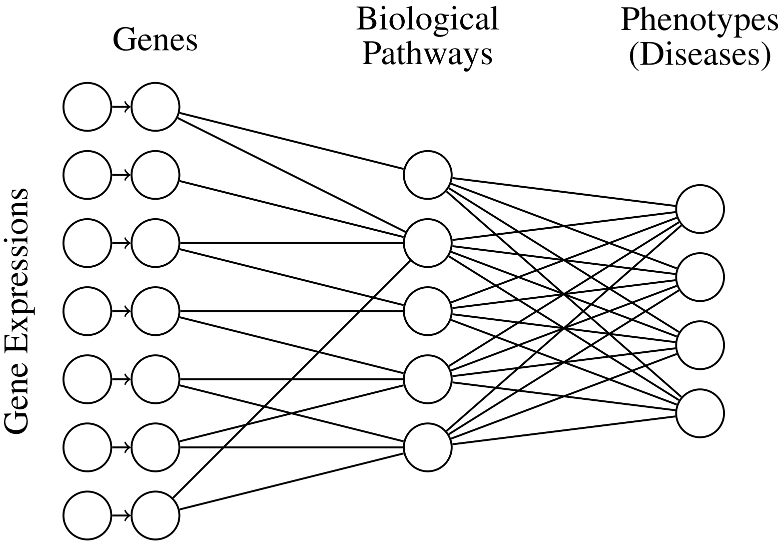


**Figure 4.1:** Example of neural network architecture. For the first layer, the connections are defined by biological information, i.e. a unit representing a gene is connected to all the biological pathways that the gene is involved in. We do not add any prior knowledge on the last layer, thus it is fully connected.

MLR and GPD have a different number of parameters (or free weights): $(573,314)$ for MLR and $(137,838)$ for GPD. Note that, due to this imbalance, we do not expect GPD to outperform MLR in the diagnosis prediction task.

To train both GPD and MLR, we first perform a 10-fold cross-validation to fix the number of training epochs. To fix the number of training epochs, we compute the average number of epochs at which the test loss is the smallest across the runs (see Figure 4.2). As the dataset is imbalanced, we use stratification to split the data, ensuring that at least one patient per disease is in the test set. Using this number of epochs, we perform another 10-fold cross-validation to evaluate the performance of our models. We use the Adam optimizer [241] with learning rate 0.01 and the layer weights are initialized to small values using the initialization process proposed by He *et al.* [242]. We investigated the addition of classical regularisation techniques – L1, L2, and dropout regularisation – as a mean to reduce the capacity of the model to overfit. We found that, given our setting, the performances were better without any regularisations (see Table 4.1).

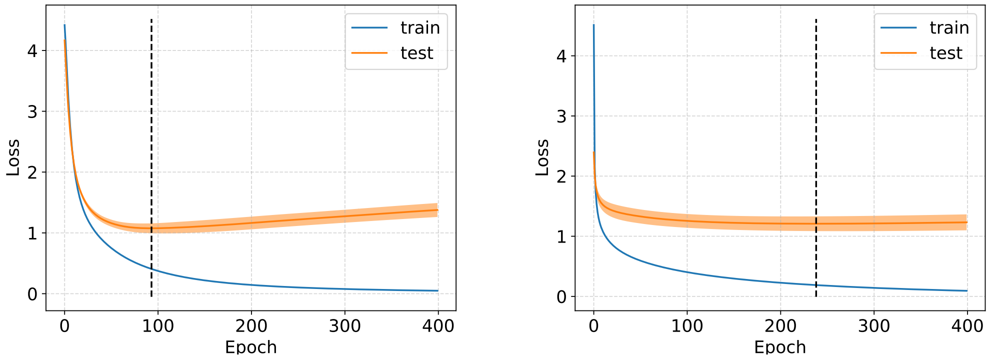


**Figure 4.2:** Train and test loss curve for the MLR model (left) and GDP model (right) with respect to the number of epochs during the cross-validation. The vertical black line indicates the number of epochs which give the lowest loss.

| hyperparameter | 10 | 0.1 | 0.01 | 0.001 | 0 |
|---|---|---|---|---|---|
| L1-regularization | $2026 \pm 4.3$ | $24.5 \pm 0.046$ | $6.39 \pm 0.003$ | $2.71 \pm 0.047$ | $\mathbf{1.09 \pm 0.066}$ |
| L2-regularization | $4.42 \pm 4.8e^{-7}$ | $4.38 \pm 0.012$ | $3.48 \pm 0.032$ | $2.02 \pm 0.026$ | / |
| dropout ratio | 0.25 | 0.5 | 0.75 | 0.9 | 0 |
| dropout | $1.14 \pm 0.121$ | $1.10 \pm 0.130$ | $1.12 \pm 0.086$ | $1.10 \pm 0.081$ | / |

**Table 4.1:** Results of cross-validation to fix regularisation hyperparameters (L1-, L2-, or dropout regularisations). The scores correspond to cross-entropy loss. The best results are obtained with no regularisation (score in bold).

The neural networks are implemented with Tensorflow [243].

### 4.3.3 Predicting disease–disease, disease–pathway, and disease–gene relationships

To identify associations between diseases and genes or pathways, we perform sensitivity analysis [244].

Formally, the local variations $\delta f$ of a single-argument function $f$ due to a change $\delta x = x - x_0$ in input can be approximated with the first order Taylor expansion as

$$\delta f(x) = \frac{df}{dx}(x_0)\delta x + O(x^2).$$

Thus, the magnitude of the local variations of $f$ with respect to perturbation $\delta x$ from $x_0$ is given by $|\frac{df}{dx}(x_0)|$. Based on this approximation, we extract from each neural network a score between an entity represented by a unit of the neural network (e.g., a pathway, or a gene) and each disease (output unit). Specifically, for a neural network NN, we denote $\mathrm{nn}^i : [0,1]^{n_i} \mapsto \mathbb{R}^{n_o}$ the function corresponding to the operation of a neural network NN from the $n_i$ outputs of layer $i$ to the final $n_o$ logits of the neural network, i.e. scores before softmax. E.g., for GPD, we have $\mathrm{nn}_2^1(\mathbf{x}) = \mathbf{W}_2^2 \tanh\left(\mathbf{W}_1^2\mathbf{x}\right)$. Then, the association score $s_{i,j,k}$ between the $j^{th}$ output unit of layer $i$, denoted $u_j^i$, of neural network NN, and disease $k$ is given by

$$s_{i,j,k} = \left| \left[ \frac{\partial \mathrm{nn}^i}{\partial u_j^i}(\mathbf{x_0}) \right]_k \right|,$$

where the reference point is chosen as the null vector, $\mathbf{x_0} = \mathbf{0}$, which corresponds implicitly to a healthy state in our formulation.

The association score between disease $d$ and unit $u$ (representing a gene, or a pathway) is thus measured by the intensity of the local variation of the output unit associated with $d$ with respect to perturbation of $u$. Intuitively, this score measures how the prediction score of disease $d$ is affected by dysregulation of a gene/pathway: we quantify the change in one disease score induced by a dysregulation in the gene expression or pathway activation. We test this scoring approach for the prediction of disease–gene and disease–pathway associations. In particular, we rank disease–gene and disease–pathway pairs based on this score and test if the score correlates with known associations through a Precision-Recall and Receiver Operating Characteristic (ROC) analysis. We focus on manual validation of the top-scoring associations.

Based on this score, we represent each disease by a set formed by the $k_{genes}$ highest scoring genes and a set containing the $k_{pathways}$ highest scoring pathways. We then score disease–disease associations using the Jaccard Index of their sets. The Jaccard Index of two sets $S_1$ and $S_2$ is defined as $\frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}$, where $|\cdot|$ represents the cardinality of a set. Following on from a similar approach used in DisGeNET [245], we interpret those associations as comorbidities. The number of highest scoring pathways and genes considered is set to 150 and 300, respectively, as those numbers gave the best results.

## 4.4 Results & Discussion

### 4.4.1 Classification performances

To validate the relevance of our model, we verify that the classification performances are at least on par with competing methods: MLR, Random Forest (RF), Bernoulli Naive Bayes (nB), and Support Vector Machine (SVM) algorithms (we use the implementation available through the scikit-learning python package [246]). We perform 10-fold cross-validation for the algorithms to fix the hyperparameters (numbers of trees 100, smoothing parameter 0.001 and penalty parameter 100, respectively) and retain the best performing models in terms of cross-entropy loss (the objective function of the neural networks).

We evaluate performances by computing 3 different scores: cross-entropy loss (CEL), micro-average and macro-average precision ($\mathrm{Pre}_\mu$ and $\mathrm{Pre}_M$). Details of each score are given in Appendix B.1.1.

| Algorithm | CEL | $\mathrm{Pre}_\mu$ | $\mathrm{Pre}_M$ |
|---|---|---|---|
| GPD | $1.09 \pm 0.06$ | $0.80 \pm 0.01$ | $0.71 \pm 0.02$ |
| MLR | $\mathbf{1.01 \pm 0.07}$ | $\mathbf{0.84 \pm 0.01}$ | $\mathbf{0.76 \pm 0.01}$ |
| RF | $1.56 \pm 0.24$ | $0.80 \pm 0.01$ | $0.70 \pm 0.03$ |
| nB | $10.63 \pm 0.55$ | $0.66 \pm 0.01$ | $0.60 \pm 0.02$ |
| SVM | $1.42 \pm 0.04$ | $0.72 \pm 0.02$ | $0.59 \pm 0.02$ |

**Table 4.2:** Performances of different classifiers in terms of cross-entropy loss (CEL), micro- and macro-average precisions ($\mathrm{Pre}_\mu$ and $\mathrm{Pre}_M$, respectively). Each score is computed across the 10-fold cross-validation and we provide the standard deviation. Bold scores highlight the best scores for each metric.

We observe that the neural networks; MLR and GPD, give better, or at least on-par, performances when compared to RF, nB, and SVM classifiers as measured by our three metrics (see Table 4.2). This observation justifies the relevance of our GPD model. The best model appears to be the multinomial logistic regression (MLR), which corresponds to the most complex neural network model in terms of the number of parameters (or degree of freedom) since MLR has $\sim 4$ times more parameters than GPD. This analysis shows that using biological knowledge to guide the structure of neural networks, in the limit of the models proposed, does not improve classification performance compared to the multiclass logistic regression (MLR) and only offers slight improvement when compared to a RF classifier (see Table 4.2). However, we show, in the following Sections, that the trained GPD models can be more successfully exploited than MLR to extract biological information. Note as well that the gene–pathway information on which GPD relies is both noisy and incomplete, as biological data often is, and those performances should improve as knowledge improves.

Hereafter, we consider for each model (GPD and MLR) the trained NN that gave the lowest cross-entropy loss during the 10-fold cross-validation.

### 4.4.2 Our GPD model uncovers molecular mechanisms of diseases

To uncover molecular mechanisms of disease, i.e., genes and pathways that are associated with specific diseases, we extract predictions from MLR and GPD using the approach described in Section 4.3.3. We test the performance of our disease–pathway and disease–gene associations predictions by comparing against established databases. We investigate the top predictions of the GPD model through a manual search of the literature.

#### 4.4.2.1 Predicting disease–gene associations

For each model, we compute disease–gene association scores as described in Section 4.3.3, and we test the validity of our predictions against DisGeNET database [245]. We compare the entire set of predictions against two baselines (see Appendix B.1.2 for details): the Frequency of Differential Expression (FDE) and the approach introduced by Zhao *et al.* [229] for *de novo* disease–gene association prediction (Katz).

We use precision–recall and ROC curves to evaluate the performance of our approach and compute the areas under the curves (see Figure 4.3). Interestingly, we observe that the FDE score is a poor predictor of disease–gene associations. We further observe that GPD is the best performing models for this task with Katz coming second. The relatively low overall performances can be partially attributed to the incompleteness of the reported disease–gene associations in DisGeNET. To corroborate this hypothesis, we search the literature for support for the top 10 predicted disease–gene associations by the best performing model, GPD (see Table 4.4). Note that none of those associations is reported in DisGeNET.

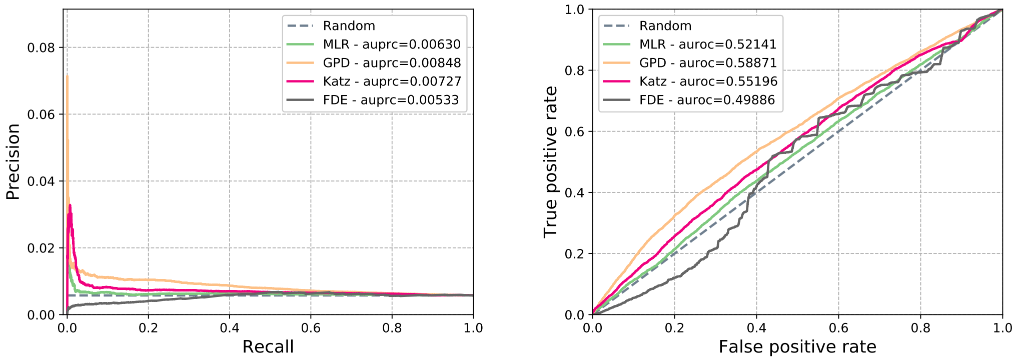


**Figure 4.3:** Precision-recall curve and ROC curve of our predictions for disease–genes associations.

| Disease | Gene | Literature support |
|---|---|---|
| Asthma | UBB | |
| Schizophrenia | RHOA | PMID:16402129 |
| Alzheimer’s disease | FGF23 | PMID:26674092 |
| Autistic disorder | FGF20 | PMID:19204725 |
| Prostate cancer | RPS27A | PMID:15647830 |
| Amyotrophic lateral sclerosis | PSMD13 | |
| Amyotrophic lateral sclerosis | CASP3 | PMID:11715057 |
| Chronic obstructive pulmonary disease | SKP1 | PMID:23713962 |
| Autistic disorder | PSMB2 | |
| Irritable bowel syndrome | PSMA1 | PMID:28717845 |

**Table 4.3:** Top 10 disease–gene predicted by GPD.

We are able to find literature support for 70% of the top 10 predicted disease–gene associations (see Table 4.3). Furthermore, we find indications that some of our top-scoring, non-validated predictions could be relevant, such as the associations of asthma with UBB and amyotrophic lateral sclerosis (ALS) with PSMD13. Ubuquitin B (UBB) belongs to the ubiquitin-proteasome (UPS) and it is known that aberration in the UPS is responsible for inflammatory and autoimmune diseases such as asthma[247]. Moreover, ALS onsets occur typically after age 50 and manifest partially through muscle weakness. PSMD13 is linked to aging [248] and high expression of the gene has been found in skeletal muscle of athletes [249], suggesting that under-expression could be a sign of muscle weakness.

These results validate the relevance of our framework for de novo disease–gene association prediction and confirm the incompleteness of DisGeNET.

#### 4.4.2.2 Predicting disease–pathway associations

For our GPD model, we compute disease–pathway association scores as described in Section 4.3.3, and we test the validity of our predictions by comparison with CTD database [250]. As a baseline, we consider disease–pathway scores corresponding to the average FDE (AFDE) of genes within the pathway for patients having the disease.

We evaluate the results as done previously for disease–gene associations (see Figure 4.4). We observe that GPD convincingly outperforms AFDE. The seemingly poor performances of both approaches can partially be attributed to the incompleteness of CTD database. To test this hypothesis, we search the literature for support for the top 10 disease–pathway associations predicted with our GPD (see Table 4.4). Note that none of these predicted associations is reported in CTD database.

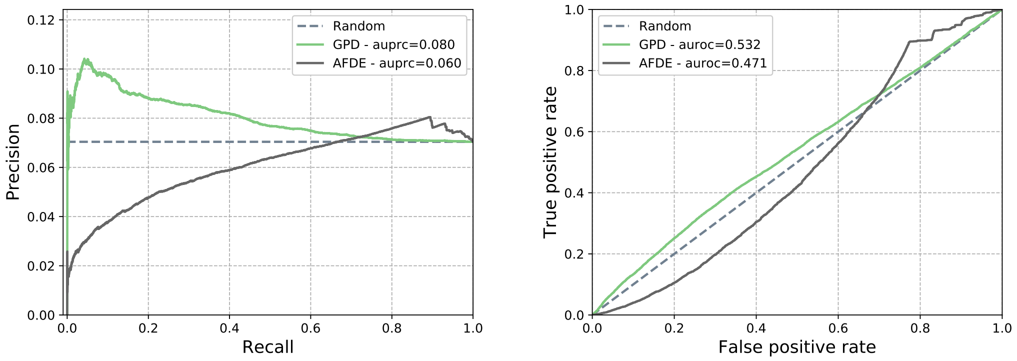


**Figure 4.4:** Precision-recall curve and ROC curve of our disease–pathways associations predictions.

We find literature support for 7 out of the top 10 predicted disease–pathway associations (see Table 4.4). Furthermore, we find indications that some of our

| Disease | Pathway R-HSA- | Literature support |
|---|---|---|
| Autistic disorder | 5653890 | |
| Irritable bowel syndrome | 532668 | PMID:20338921 |
| Irritable bowel syndrome | 391906 | PMID:16835707 |
| Type 2 diabetes mellitus | 499943 | doi:10.2337/diabe-tes.51.2007.S363 |
| Asthma | 391906 | PMID:8603274 |
| Schizophrenia | 71288 | PMID:22465051 |
| Major depressive disorder | 8934903 | PMID:27063986 |
| Type 2 diabetes mellitus | 8939245 | PMID:19667185 |
| Schizophrenia | 5683371 | |
| Sjogren's syndrome | 389661 | |

**Table 4.4:** Top 10 disease–pathway predictions derived from GPD.

top-scoring, non-validated predictions could be relevant, such as the association of autistic disorder with the lactose synthesis pathway (R-HSA-5653890) and the association of schizophrenia with pathway R-HSA-5683371 linked to microphthalmia. The lactose synthesis pathway (R-HSA-5653890) contains three genes: LALBA, SLC2A1, and B4GALT1. All of those genes might be associated with autistic disorders. One patented method to detect autistic disorder (US20140349977A1) includes LALBA as one of the genes of interest. SLC2A1 mutation has been reported in patients diagnosed with autism [251]. Finally, B4GALT1 has been linked with developmental disorders [252]. The pathway R-HSA-5683371 is linked to the eye disease microphthalmia. It is known that schizophrenia is linked to eye abnormalities [253]. Among the 28 genes involved in that pathway, 12 have been linked to the disease in the literature (GOT2, PDHA1, DLD, GCSH, DLAT, PDHB, DAO, OGDH, DHTKD1, GNMT, DDO, PRODH2).

These results show the relevance of our framework for de novo disease–pathway associations prediction despite relatively low retrieval scores against the ground–truth.

### 4.4.3 Our GPD model predicts disease–disease relationships

We rank disease–disease pairs based on the score described in Section 4.3.3 and test our results against a high confidence comorbidity disease network obtained from a large cohort study by Hidalgo *et al.* [1]. We compare our method against DMSN network [46], restricted to our set of diseases, and three alternatives baselines. For the first alternative, we compute disease–disease association score using our approach defined in Section 4.3.3 on the trained MLR network, representing each disease by the top 300 highest scoring genes (which gave the best results based on grid search). The last two baselines associate to each disease–disease pair a Jaccard Index score based on 1) the set of genes associated to each disease in DisGeNET [245] and 2) the set of pathways associated to each disease in CTD database[250]. The results of the comparison are presented using a precision–recall curve (see Figure 4.5).

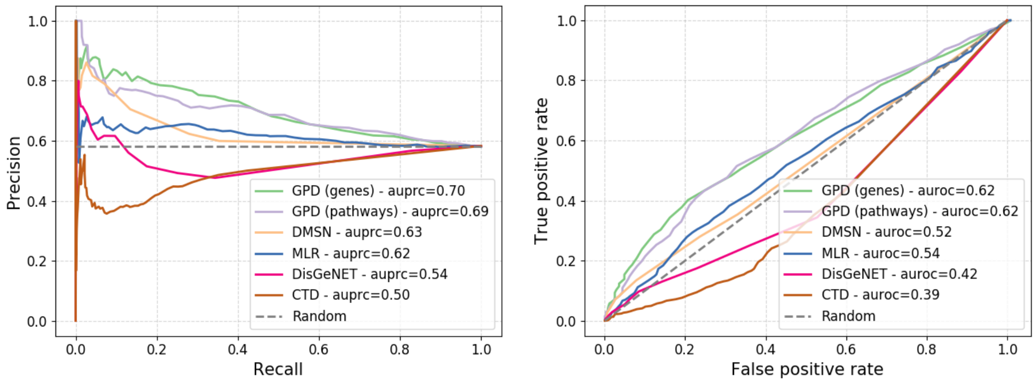


**Figure 4.5:** Precision–recall (top) and ROC (bottom) curves of the test against the disease co-morbidity network built by Hidalgo *et al.* [1].

We observe that our approach outperform convincingly the other approaches in the task of retrieving existing comorbidity links between diseases with over 10% increase compared to DMSN and 30% improvement over DisGeNET in terms of area under the precision–recall curve (auprc). These results strongly support our methodology. The scoring based on disease–gene is performing better than disease–pathway, hence we investigate the top 10 scoring disease–disease associations derived from it (see Table 4.5).

We present and discuss below literature support for the predicted associations between the diseases.

Atrial fibrillation has been linked in the literature to thyroid disease [254] which is known to be comorbid with vitiligo [255]. Atrial fibrillation and peripheral vascular disease are well known comorbid conditions [256]. Alcoholism

| Disease 1 | Disease 2 |
|---|---|
| Atrial fibrillation | Vitiligo |
| Atrial fibrillation | Peripheral vascular disease |
| Alcoholic hepatitis | Osteosarcoma |
| Rhabdoid cancer | Medulloblastoma |
| Cornelia de Lange syndrome | Vitiligo |
| Peripheral vascular disease | Vitiligo |
| Atrial fibrillation | Osteosarcoma |
| Leishmaniasis | Alcoholic hepatitis |
| Sotos syndrome | Vitiligo |
| Follicular lymphoma | Osteosarcoma |

**Table 4.5:** Top 10 disease–disease links predicted using our approach based on the trained GPD.

has been linked to the onset of some cancers notably implicating the transcription factor Nanog which itself has been linked to osteosarcoma [257]. Additionally, a drug used to treat alcoholism, Disulfiram, has recently been proposed as a potential treatment for osteosarcoma [258]. These observations together suggest a shared molecular background for the two conditions. Rhabdoid cancer is a rare form of aggressive cancer affecting young children and with very poor prognostic, which makes it challenging to evaluate comorbid conditions. However, rhabdoid cancer is frequently mistaken for medulloblastoma, indicating some similarity [259]. We found no evidence in the literature supporting a connection between the rare Cornelia de Lange syndrome and vitiligo. Some studies have observed significant comorbidity between vitiligo and psoriasis, and the combination of the two has been linked to cardiovascular diseases, which include peripheral vascular disease [260, 255]. Atrial fibrillation and cardiac complications have been observed as the result of osteosarcoma [261, 262]. Leishmaniasis and alcoholic hepatitis are an unlikely comorbid connection since it would require a patient both to have been infected by parasites of the Leishmania type and have had excessive alcohol intake. This suggests that the interpretation of links based on similarity as co-morbidity can be hasty. However, both disease affects the liver and leishmaniasis has sometimes been misdiagnosed for cirrhosis [263], which suggests that the two diseases might share some similar molecular processes that we would be capturing here. A case of co-occurrence of Sotos syndrome and vitiligo has been reported in the medical literature [264]. It has been postulated that non-Hodgkin's lymphoma (which include follicular lymphoma) and osteosarcoma share underlying mechanisms [265]. Additionally, miR-202 has been identified as a potential tumour suppressor for both conditions [266].

Through this analysis, we have shown that most predicted pairs have either been observed co-occurring or can be connected through underlying mechanisms, thus validating our approach.

## 4.5 Conclusions

In this study, we propose a multi-scale neural network based framework that integrates gene expression data associated with diseases and gene–pathway information. Our integrative framework allows for simultaneously uncovering novel disease-disease associations and molecular disease mechanisms from patient gene expression profiles through the analysis of trained neural networks. We show that GPD achieves good diagnosis prediction on our dataset showing the validity of our integrative process. Furthermore, we show that the associations predicted from the trained models are biologically meaningful and supported by the literature, thus validating our approach and motivating the use of the multi-scale protein interactome as prior knowledge for Visible Machine Learning models.

While the current knowledge about these diseases supports our uncovered molecular disease mechanisms, a next step would be to identify among the predicted genes and pathways suitable biomarkers and drug targets that could be used to improve diagnosis, prognosis, and treatment. We leave this for future work. Also, while our multi-scale NN framework integrates the hierarchical functional organisation of a cell (from genes to biological pathways), our methodology can be extended to include any dataset about diseases of interest, e.g., uncovering molecular mechanisms of cancer from patient somatic mutation profiles or linking diseases to non-coding RNA.

Finally, while we focus on patient data with application to diseases, our methodology can be extended to integrate additional omics data to get more biologically accurate models for the analyses of patients, tissues, and cells. Some further applications include studies of diseases linked to a specific tissue, studies of cell's specialisation, and any study that can benefit from the integration of the hierarchical functional organisation of cells.

## Chapter 5

# Multi-scale protein interactome as auxiliary data to enhance cancer precision medicine

In Chapter 4, we proposed a model using the multi-scale protein interactome as prior knowledge to identify links between diseases, pathways, and genes. In this Chapter, we take a different approach: we integrate the multi-scale protein interactome to a general framework as auxiliary data. Specifically, we use joint matrix factorisation to integrate PPIs, protein complexes, and biological pathways with patients' omics and clinical data, as well as drugs target and similarities. By this process, we derive for each biomedical entity an embedding within the multi-source data context. This enables us to identify molecular mechanisms and drug indications for specific cancer types. The results presented in this chapter have been submitted and are currently under review for publication, a preprint is available on arXiv [267].

## 5.1 Introduction

Over 18 million new cases of cancer and 9 million deaths were recorded worldwide in 2018 [268]. This makes cancer one of the leading causes of death. Cancer is a multi-faceted, complex disease arising from an accumulation of somatic mutations within the genome of normal cells that eventually leads to loss of normal cellular functioning and appearance of tumours that can spread across the body. Technological advances have enabled measurements from patients' tumour biopsies, including gene expression levels, DNA methylations, and somatic mutations. The research into cancer causes, and treatments, has greatly benefited from this wealth of patient data [269, 270].

Cancer projects, including The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC), have made publicly available wide-ranging, multi-modal, multi-omics cancer data, such as patient whole slide images, genome alterations, transcriptome, and epigenome [271, 272]. Free access to these large-scale, diverse databases has dramatically facilitated studies of the biological mechanisms of specific cancer types [271, 273, 274]. The available data have also enabled pan-cancer analyses that study cancer in general to identify common mechanisms and differences across cancer types [275, 274]. Recently, the Pan-Cancer Analysis of Whole Genome (PCAWG) project [274] has informed that our knowledge about cancer is far from complete, as 5% of their cohort was without any known cancer driver mutations. Importantly, these large databases have paved the way for the field of Precision Medicine, whose overarching aim is to improve medical care for patients by tailoring treatment to their individual molecular profiles [276]. Precision medicine has diverse intermediary objectives, for instance, uncovering diagnostic and prognostic biomarkers. This is especially relevant to a heterogeneous disease, such as cancer, which manifests uniquely in every patient.

Cancer can be caused by combinations of genetic, molecular, environmental, and lifestyle factors. Any single type of biological data cannot fully capture such diseases. As such, collective mining of different data has been gaining momentum as a means to extract integrated system knowledge that goes beyond what any single data source can offer [9]. This principle applied to the study of cancer has enabled the discovery of cancer-related genes, or group of genes [277, 157, 156] and the identification of cancer sub-types significantly correlated with patient prognoses [179, 157].

Biological data often have a small number of samples relative to the number of available features. For instance, a typical dataset in TCGA contains a few hundred patients that are each characterised by tens of thousands of features (e.g., expression levels of around 20,000 genes). However, biological features are often redundant due to underlying molecular interactions among biological entities [278]. This has been a motivation for the use of dimensionality reduction and embedding algorithms that are pervasive in bioinformatics [279]. Additionally, due to the low sample to features ratio, dimensionality reduction techniques are often necessary as data pre-processing for machine learning models [279], at the cost of interpretability.

Non-negative matrix factorisation (NMF) approaches are unsupervised algorithms that have extensively been used both as a means to integrate heterogeneous data and to reduce data dimensionality (see Section 2.3). They encompass all methods that decompose a matrix, representing relational links between two sets of entities, into the product of low-dimensional, latent, non-negative matrices, or factors, whose sizes control the degree of dimensionality reduction [280]. Importantly, they can be used to derive an embedding in an unspecified latent space for each entity. Matrix factorization approaches have had numerous applications, including collaborative filtering [281] and biological data integration for cancer analysis [179, 157]. Reconstructing a matrix based on a factorisation has often been used to make predictions and infer new knowledge [157]. NMF approaches have been successfully applied as pre-processing steps for downstream machine learning classifiers [282]

## 5.2 Contributions

We propose a pan-cancer framework to uncover cancer type-specific molecular mechanisms and identify drugs that could be re-purposed (see Figure 5.1). Our framework relies on the simultaneous integration and dimensionality reduction of various data using a joint non-negative matrix factorisation model. Our framework includes more data than the previous studies, integrating patient-specific diagnosis, gene expression, and single nucleotide variants as well as generic network data on human: protein–protein interactions, protein complex associations, biological pathways, drug–target interactions, and drug chemical similarities. To integrates the wealth of data in one framework, we rely on three types of matrix factorisations: non-negative matrix factorisation (NMF), non-negative matrix tri-factorisation (NMTF), and symmetric non-negative matrix tri-factorisation (SNMTF). Data integration is achieved by jointly optimising for multiple factorisation objectives with shared factors. We obtain a *context-aware* embedding of each entity (cancer type, patient, gene, complex, pathway, and drug) that takes into account all the input data. Using boosted decision trees, we predict biologically relevant associations between cancer types and genes, drugs, pathways, and complexes based on the context-rich embeddings of our entities. We choose boosted decision trees due to their simplicity and high performances in a number of competitions [283]. One key insight is that the integration step, by construction, embeds the entities into three latent spaces, each associated with a different family of entities: 1) patient-related entities (i.e., patients and cancer types), 2) gene-related entities (i.e., genes, complexes, and pathways), and 3) drugs. This means that the entities in a given latent space can be substituted with each other when using a model trained to predict associations between one of these classes of entities and cancer types. In this respect, our approach is similar to zero-shot learning [284], which aims to accurately classify at test time samples that belong to classes unseen at training time. In our case, we aim to predict the association of cancer types to unseen classes of entities at training time. Finally, our approach can predict a patient's response to drugs, implying that our framework captures important biology that governs response to cancer drugs.

## 5.3 Material & Methods

### 5.3.1 Data source and processing

We download protein-protein interactions (PPI) data from BioGRID (version 3.5.176). We only keep interactions that have been validated experimentally using yeast-to-hybrid or affinity capture techniques. We obtain protein complexes data from CORUM and Reactome databases (both accessed in April 2019). Reactome is also used to collect all existing pathways of which we only keep pathways that have a traceable author statement (TAS). We further remove disease pathways which are only relevant in the associated disease context.

Patients data are obtained from the DCC Data release of the International Cancer Genome Consortium. We collected patients from 21 cancer cohorts from TCGA studies (see Table 5.1) and kept patients that have RNA-sequencing data, which adds up to 7,998 patients. We consider all Single Nucleotide Variations (SNV) reported in the data releases.

| Cancer | Cohort size | Abbreviation | Cancer | Cohort Size | Abbreviation |
|---|---|---|---|---|---|
| Acute Myeloid Leukemia | 173 | LAML | Bladder Urothelial Carcinoma | 295 | BLCA |
| Brain Lower Grade Glioma | 439 | LGG | Breast invasive carcinoma | 1,041 | BRCA |
| Cervical squamous cell carcinoma and endocervical adenocarcinoma | 259 | CESC | Colon adenocarcinoma | 428 | COAD |
| Glioblastoma multiforme | 159 | GBM | Head and Neck squamous cell carcinoma | 480 | HNSC |
| Kidney renal clear cell carcinoma | 518 | KIRC | Kidney renal papillary cell carcinoma | 222 | KIRP |
| Liver hepatocellular carcinoma | 173 | LIHC | Lung adenocarcinoma | 477 | LUAD |
| Lung squamous cell carcinoma | 428 | LUSC | Ovarian serous cystadenocarcinoma | 262 | OV |
| Pancreatic adenocarcinoma | 142 | PAAD | Prostate adenocarcinoma | 375 | PRAD |
| Rectum adenocarcinoma | 153 | READ | Skin Cutaneous Melanoma | 430 | SKCM |
| Stomach adenocarcinoma | 415 | STAD | Thyroid carcinoma | 500 | THCA |
| Uterine Corpus Endometrial Carcinoma | 508 | UCEC | | | |

**Table 5.1:** List of cancer types considered in this study with associated abbreviations from TCGA.

We consider the set of 15,224 genes whose transcripts are measured for by the RNA-sequencing technology across all datasets and that have at least one PPI with another selected gene according to BioGRID data. We derive a gene expression vector from RNA-seq measurement for each patient that is normalised to Transcripts Per Million (TPM) and rescaled using logarithm in base 2, i.e. the expression score of a gene is given by $\log_2(\text{TPM}+1)$. Note that 559 patients do not have any mutation on any of the 15,224 genes considered.

Drug–target and chemical data is obtained from DrugBank (version 5.1.3) [285]. We consider all drugs that are approved, experimental, or investigational. Drugs chemical similarity is computed using the Tanimoto similarity [286] between circular fingerprints of drugs. The details of the data used can be found in Table 5.2.

| Data | Size | Density | Symbol |
|---|---|---|---|
| Gene expression | $n_{patients} \times n_{genes}$ | n.a | $X_{exp}$ |
| Gene SNV | $n_{patients} \times n_{genes}$ | $1.7\%$ | $X_{snv}$ |
| Patient cancer type | $n_{cancers} \times n_{patients}$ | n.a. | $X_{pct}$ |
| PPI | $n_{genes} \times n_{genes}$ | $0.21\%$ | $X_{ppi}$ |
| Protein complexes | $n_{pc} \times n_{genes}$ | $0.05\%$ | $X_{pc}$ |
| Biological pathways | $n_{bp} \times n_{genes}$ | $0.32\%$ | $X_{bp}$ |
| Drug–target | $n_{drugs} \times n_{genes}$ | $0.013\%$ | $X_{dt}$ |
| Drug tanimoto similarity | $n_{drugs} \times n_{drugs}$ | n.a. | $X_{ts}$ |

**Table 5.2:** Details of the data used. The columns correspond to: 1) the type of data, 2) the size of the matrix representing the data, 3) the density, where applicable, indicates the percentage of the existing links between the entities out of all possible links, and 4) the symbol used in the document to refer to the matrix containing the corresponding data.

Genes' annotations used in enrichment analyses are obtained from Gene Ontology (GO) [287] (release 10/06/2019). We keep annotations that have an experimental evidence code (one of EXP, IDA, IPI, IMP, IGI, and IEP). We consider all three GO annotation subtypes separately: Biological Processes (GO–BP), Molecular Function (GO–MF) and Cellular Component (GO–CC). For each, we build a directed acyclic graph (DAG) that connects annotations based on “is a” relationships (we use the go-basic.obo file giving annotations relationships available on GO's website). Then, we propagate the annotations for each gene up the corresponding DAG, which means that we add to the set of annotations of a gene the union of ancestors of the annotations. We remove annotations that annotate less than $0.1\%$, or more than 10% of the 15,224 genes considered, i.e. we prune annotations that are either too rare, or too common. These thresholds were picked by inspecting the distribution of number of genes per annotation. We give the statistics of annotations in Table 5.3

| GO subtype | GO–BP | GO–MF | GO–CC |
|---|---|---|---|
| Number of annotations | 2,322 | 538 | 366 |
| Percentage of genes annotated | 50% | 43% | 45% |

**Table 5.3:** GO annotations statistics for all GO subtypes.

### 5.3.2 Non-negative Matrix Factorizations

Matrix factorizations approaches aim to approximate a matrix $X$ by the product of $n$ smaller matrices $F_i, i \in \{1..n\}$, called factors, i.e. $X \approx \prod_i F_i$. Mathematically, this amounts to finding factors $F_i$, under user-defined dimensional constraints, that minimize the equation $||X - \prod_i F_i||_F^2$, where $|| \cdot ||_F$ represents the Frobenius norm of a matrix. Non-negative matrix factorizations techniques add a non-negativity constraint on the factors, i.e. $\forall i, F_i \geq 0$.

The objective is to obtain lower dimensional representation that captures the essence of the data and can be used to identify missing entries through the matrix completion property. In this work, we use three variants of non-negative matrix factorisations approaches.

**NMF** decomposes a rectangular matrix $X \in \mathbb{R}^{m \times n}$ in the product of two factors $F \in \mathbb{R}_+^{m \times k}$ and $G \in \mathbb{R}_+^{n \times k}$, with $k \leq \min(m,n)$, such that $||X - FG^T||_F^2$ is minimized. With NMF, the embeddings, given by $F$ and $G$, of the two groups of entities whose relational data is given by $X$, are in the same latent space.

**NMTF** decomposes a rectangular matrix $X \in \mathbb{R}^{m \times n}$ in the product of three factors $F \in \mathbb{R}_+^{m \times k_1}$, $S \in \mathbb{R}_+^{k_1 \times k_2}$ and $G \in \mathbb{R}_+^{n \times k_2}$, with $k_1, k_2 \leq \min(m,n)$, such that $||X - FSG^T||_F^2$ is minimized.

**SNMTF** decomposes a symmetric matrix $X \in \mathbb{R}^{n \times n}$ in the product of two factors $G \in \mathbb{R}_{+}^{n \times k}$ and $S \in \mathbb{R}_{+}^{k \times k}$, with $k \leq n$, such that $\|X - GSG^T\|_F^2$ is minimized.

### 5.3.3 Our framework for context-aware embeddings

The core of the framework is gene information (see Figure 5.1.a.). We integrate three types of data about genes. We obtain RNA sequencing (RNA-seq) data and single nucleotide variants (SNV) data for 7,998 patients from ICGC across 21 cancers. Henceforth, we refer to each cancer by its abbreviation given in Table 5.1. We obtain data on gene interactions including protein-protein interaction (PPI) network from BioGRID, protein complexes (PC) from Reactome and CORUM, and biological pathways (BP) from Reactome. These data capture physical and functional relationships between genes and are used to anchor our framework within the context of molecular interactions. The last type of gene data corresponds to drug–target interactions from DrugBank, connecting drugs to proteins that they target. We further add drug chemical similarity information to push similar drugs closer in the latent space. We also add patient diagnosis information through which we embed cancer types and patients in a joint latent space to both push patients closer if they have the same cancer and push molecularly similar cancer types closer. This could help tailor treatments to patients by placing them within a cancer "space" since cancer is a heterogeneous disease and a given cancer type might manifest differently in different people. This may aid characterising cancer of each patient as accurately as possible to personalise treatment options.

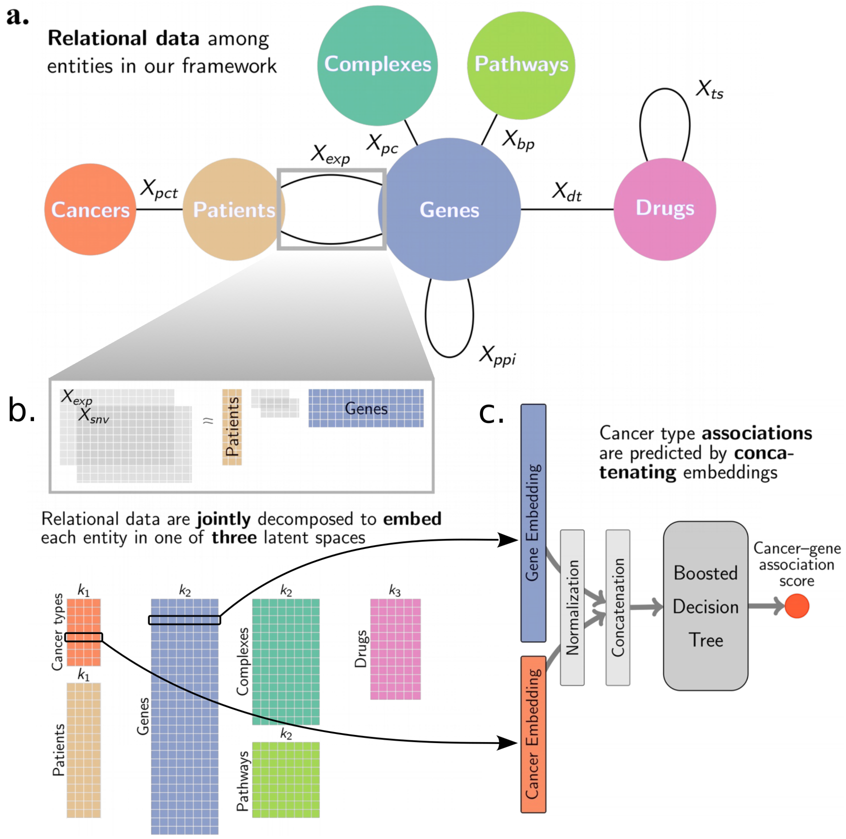


**Figure 5.1:** **a.** Input to our matrix factorisation embedding model: relational data between entities. Each edge corresponds to a type of link and a sub-objective of our joint factorisation model (see Methods Data Processing Table 5.2 for notation). The squares group entities that are embedded in the same joint latent space. **b.** Illustration of the NMTF factorisation sub-objectives corresponding to the edges in the grey box in panel a. Each group of entities is associated, in the decomposition, with a factor that is shared across all sub-objectives involving that group of entities. Through the joint decomposition of all relational data, we derive embeddings for each entity in three latent spaces with dimensions $k_1$, $k_2$, and $k_3$. **c.** We predict associations relevant to cancer types with boosted decision tree classifiers taking as input, for instance, the concatenation of the embeddings of a cancer type and a gene.

Because of the heterogeneity of our input data, our integration framework is based on joint optimisation of different variants of non-negative matrix factorisation: classical Non-negative Matrix Factorisation (NMF), Non-negative Matrix Tri-Factorisation (NMTF), and Symmetric Non-negative Matrix Tri-Factorization (SNMTF). Each variant is best fitted for the decomposition of a different type of relational data. In particular, we use SNMTF to factorise the PPI network and the drug similarities matrix, NMTF to factorise patient molecular data and drug–target data, and NMF for the remaining data. Each edge in Figure 5.1.a corresponds to a sub-objective of our embedding framework, i.e. a specific NMF decomposition. In the joint decomposition, each group of entities is associated with a factor that is shared across all sub-objectives involving that group of entities. For instance, the patient factor is shared by all sub-objectives that involves patient-specific data (diagnoses, gene expressions, and somatic mutations). An entity's embedding is obtained from the factor of the associated group of entities. In practice, we minimize the following general objective function $\mathscr{L}$ over the factors $G_p \in \mathbb{R}^{n_{patients} \times k_1}$, $G_g \in \mathbb{R}^{n_{genes} \times k_2}$, $G_{pc} \in \mathbb{R}^{n_{pc} \times k_2}$, $G_{bp} \in \mathbb{R}^{n_{bp} \times k_2}$, $G_d \in \mathbb{R}^{n_{drugs} \times k_3}$, $S_{exp}$, $S_{snv} \in \mathbb{R}^{k_1 \times k_2}$, $S_{ppi} \in \mathbb{R}^{k_2 \times k_2}$, and $S_{dt} \in \mathbb{R}^{k_3 \times k_2}$ factors:

$$\begin{aligned} \mathscr{L} = & \sum_{x \in \{exp,snv\}} \|X_x - G_p S_x G_g^T\|_F^2 + \sum_{x \in \{pc,bp\}} \|X_x - G_x G_g^T\|_F^2 \\ & + \|X_{ppi} - G_g S_{ppi} G_g^T\|_F^2 + \|X_{pct} - G_{ct} G_p^T\|_F^2 \\ & + \|X_{ts} - G_d S_{ts} G_d^T\|_F^2 + \|X_{dt} - G_d S_{dt} G_g^T\|_F^2, \end{aligned} \tag{5.1}$$

where, henceforth, each $X_D$ represents the matrix associated to data type $D$, see nomenclature in Table 5.2, each $G_E$ factor gives the embeddings of the entities of type $E$, with subscripts $g$, $p$, $ct$, $d$, $bp$, and $pc$ corresponding, respectively, to genes, patients, cancer types, drugs, pathways, and complexes.

The integration of the various data sources is achieved by sharing factors across the NMF sub-objectives that constitute our global objective function $\mathscr{L}$. For instance, the factor $G_g$, corresponding to the genes embeddings, is shared by all decompositions that involve genes which corresponds to the factorisation of PPI data ($X_{ppi}$), the factorisations of patients molecular data ($X_{exp}$ and $X_{snv}$), the factorisation of drug–target data ($X_{dt}$), and the factorisations of higher-order biological entities ($X_{pc}$ and $X_{bp}$). Through this factor sharing and joint optimisation, the framework can harness the relevant information contained across the data sources to derive meaningful embeddings.

Through our integrative framework, we derive embeddings for all entities (cancer types, patients, genes, pathways, complexes, and drugs) that best fit the full context of the framework, i.e. the input relational data. Each entity's embedding, in one of the three latent spaces learnt by our framework, encapsulates the information from the input data that is relevant to that entity; thus we say that this representation is *context-aware*. Our framework has three hyperparameters, denoted by $k_1$, $k_2$ and $k_3$, which correspond to the dimensionalities of the latent spaces. To find suitable values for these hyperparameters, we perform a grid search with $k_1 \in \{2,5,10,15,21\}$, $k_2 \in \{70,80,90,100,110\}$, and $k_3 \in \{40,50,60,70,80\}$. The former is a coarse grid over the range of possible values. For the latter two, due to the large range of possible values, the intervals are restricted around the value $\sqrt{n/2}$, where $n$ is either the number of genes, or the number of drugs. $\sqrt{n/2}$ corresponds to a heuristic commonly used to set the number of clusters [288]. As the selection criterion, we measure if each patient tends to be embedded in the latent space closer to their diagnosis than to other cancer types. We quantify this with the macro-F1 score of the classifier that associates to each patient the closest cancer type in the latent space in terms of cosine distance. We found that the following hyperparameters values maximize this metric: $k_1 = 21$, $k_2 = 70$, and $k_3 = 40$. Appendix Figure C.1 shows the sensitivity of different metrics to the choice of the hyperparameters, which we discuss in the rest of the article.

### 5.3.4 Optimization

The minimisation of the objective function given in Equation 5.1 is achieved through an iterative optimisation process. We use in our framework multiplicative update rules [289] designed to maintain non-negativity of all the factors in the decomposition.

We use an initialization strategy based on the truncated singular value decomposition (SVD) for all factors that has shown better performances than random initialization [290, 156] and has the advantage of giving deterministic solutions. Specifically, consider a factor $G \in \mathbb{R}^{n \times k}$ involved in the decomposition of $l$ data matrices $X_i,\ i \in \{1..l\}$. Without loss of generality, we assume that $G$ is the right hand side factor in the decompositions, i.e. $\forall i, X_i \approx GF_i$. We denote by $U_i$ the right hand side term in the SVD decomposition of $X_i$ ( $X_i = U_i S_i V_i^T$ ). We introduce $U_i^+ = \max(U_i, 0)$ and $U_i^- = \max(-U_i, 0)$. We then denote by $\tilde{U_i} \in \mathbb{R}^{n\times k}$ the matrix where each column is defined by

$$\tilde{U_i}(j) = \begin{cases} \sqrt{s_i(j)}U_i^+(j), & \text{if } \|U_i^+(j)\| \geq \|U_i^-(j)\| \\ \sqrt{s_i(j)}U_i^-(j), & \text{otherwise} \end{cases},$$

where $M(j)$ denotes the $j^{th}$ column of the matrix $M$ and $s_i(j)$ is the $j^{th}$ largest singular value of $X_i$. Factor $G$ is then initialised as $\frac{1}{l}\sum_{i=1}^{l}\tilde{U_i} + \varepsilon$. We initialize the central matrix in NMTF decompositions to $I + \varepsilon$, where $I$ denotes the identity matrix. Under the multiplicative update rules, any entries initialized to zero would stay null. Hence, we add a small $\varepsilon$ everywhere to allow all entries to vary.

The iterative optimisation is ran either for 200 epochs or until the relative variation of the objective function between two consecutive epochs is lower than $10^{-4}$, i.e. when $\frac{|\mathscr{L}_{t+1} - \mathscr{L}_t|}{\mathscr{L}_t} \leq 10^{-4}$ where $\mathscr{L}_t$ corresponds to the value of the objective function at iteration $t$.

### 5.3.5 Boosted decision tree

We use boosted decision trees to predict cancer type associations with entities that are part of our framework from the embeddings derived from the Joint NMF optimisation step (Figure 5.1.c). The ground truths used to train our classifications models are detailed in the relevant sections of the Results & Discussion section. A decision tree partitions the input data iteratively based on features. Boosting signifies deriving a strong classifier from the serial associations of weak classifiers. In our case, the base classifiers are decision trees. The boosted algorithm iteratively adds decision trees to the classifier with the aim of reducing the error of the previous classifier [291].

We discuss here the implementation details that we used. First, we use boosted decision trees from the xgboost package [283]. Boosted decision trees have different hyperparameters that control various aspect of the algorithm: $\eta$ controls the learning rate, $\gamma$ corresponds to a threshold under which a leaf node of the decision tree is not split anymore, the maximal depth of a decision tree, and $\lambda$ controls the L2-regularization (for more details see [283]). We perform a 10-fold cross-validation to fix those hyperparameters with $\eta \in \{0.25, 0.5, 0.75\}$, $\gamma \in \{0, 10\}$, max depth$\in \{6, 12\}$, and $\lambda \in \{1, 10, 100\}$. The best set of parameters is chosen as the one that leads to the classifier with the highest AUROC score in the associated task. Note that we also use early-stopping during training with an 80%/10%/10% train/validation/test split of our data. We use all 10 classifiers trained during the cross-validation process to derive an association score for each possible pair. To ensure that the scoring of the 10 classifiers is comparable, we rescale the output scores to have 0 mean and unit variance. The average of all classifier scores then gives the final association score of an entity pair.

## 5.4 Results & Discussion

### 5.4.1 Patient and cancer embeddings are medically relevant

To evaluate the biomedical relevance of our joint patient and cancer embeddings, we observe that the macro-F1 score is close to 0.8 for our optimal set of hyperparameters (see Figure 5.2.a.) indicating that the majority of patients are embedded closer to their diagnoses than to other cancer types. In addition, we evaluate if patients group in the latent space with respect to either cancer type, or a sampled tissue. To this end, we use hierarchical clustering with cosine distance to group patients in $k$ groups (where $k$ is either the number of cancers, or the number of tissues) and compute the Adjusted Rand Index (ARI) to measure the link between the clustering and the ground truth labelling (either cancer types, or sampled tissues; see Figure 5.2.b.). We observe that patients do not cluster well with respect to sampled tissues, having ARI below 0.2. However, we observe ARI 0.7 with respect to cancer type, indicating that our clusterings resemble diagnostic labelling with some discrepancies. These results are expected, as the inclusion of patient diagnosis data in the framework implies a constraint that aims to embed each patient close to their diagnosis and subsequently, to other patients having the same disease. Note, however, that some patients do not fit well with the rest of their cohorts. This is an impor-

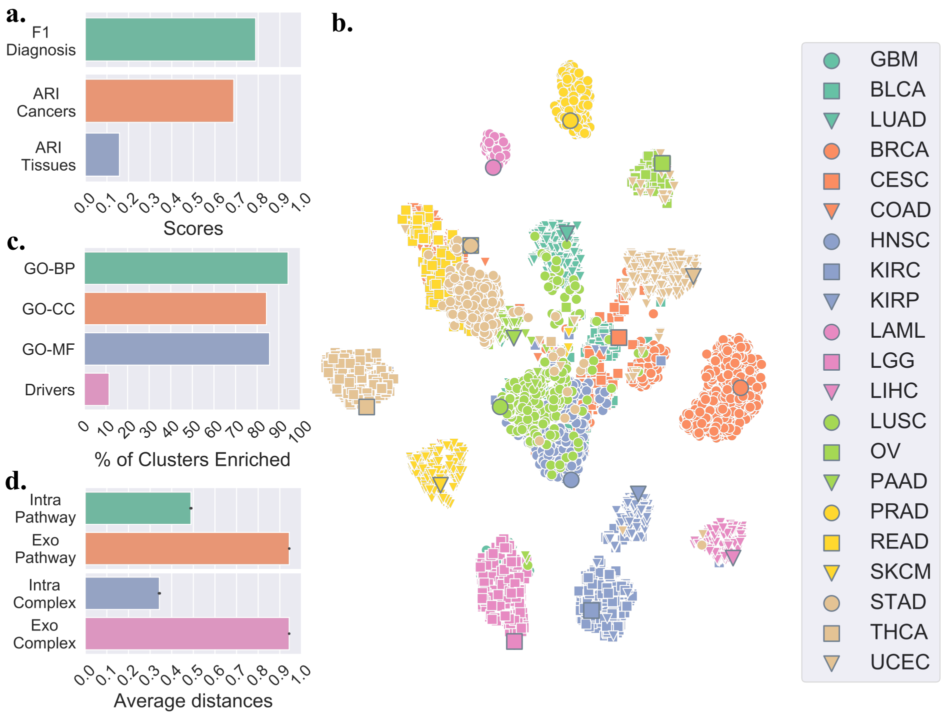


**Figure 5.2:** **a.** Macro-F1 score quantifying the relation between a patient and its cancer type (green) and adjusted random index (ARI) measuring the link between patient clustering for each model and cancer type labelling (orange) and tissue sampled labelling (blue). **b.** t-SNE plot representing the embedding of patients and cancer types in the latent space. The larger circled markers correspond to the embeddings of cancer types, and the smaller ones represent the embeddings of patients. Colours and shapes indicate cancer types (see Methods Table 5.1 for abbreviations meanings). **c.** Percentages of gene clusters enriched in GO-BP, GO-CC, GO-MF, and driver annotations. **d.** Average cosine distance between genes and associated pathways and complexes (intra-pathway and intra-complex) and non-associated pathways and complexes (exo-pathway and exo-complex).

tant observation, as it suggests that those patients might need different care options from the majority of their cohort and further motivates personalising treatments to individual patients.

As an illustration, we visualise our latent space embeddings using T-distributed Stochastic Neighbor Embedding (t-SNE). t-SNE is a machine learning algorithm for nonlinear dimensionality reduction, well-suited for visualisation in a two-dimensional space of high-dimensional data [292]. We observe as expected that patients tend to cluster according to cancer type, with the cancer type itself being also embedded nearby (see Figure 5.2.c.). Additionally, we observe that some cancers are grouped in a meaningful way. For instance, both brain cancers, GBM and LGG, form one group. The cluster in the centre contains mostly squamous cell carcinomas, HNSC, CESC, and LUSC. Both cancers that affect kidneys, KIRP and KIRC, are also grouped. Moreover, READ, COAD, and STAD - which are cancers affecting rectum, colon, and stomach, respectively - form another cluster. We also observe that some patients having a specific type of cancer do not group with the majority of the cohort.

We further investigate if our framework learns a meaningful latent space that translates into actionable representations for new unseen patients. We perform a 10-fold cross-validation in which 90% of patients of each cancer type are used to derive the embeddings of all entities in the framework. We then project the remaining 10% of patients in the derived cancer/patient latent space (see Appendix C.1.1). First, we test if these patients are placed close in space to their diagnosis (quantified as above, see Appendix Table C.1). This gives a macro-F1 score of $0.77 \pm 0.03$, which is close to the score obtained with all patients included in the framework ($\sim 0.8$). This shows that new patients are placed in the latent space according to their diagnoses with accuracy similar to that for patients included in the framework. We also test if the unseen patients tend to be embedded closer to patients having the same diagnoses. For this, we use a k-nearest neighbours classifier with $k = 10$ (see Methods) and measure its macro-F1 score (see Appendix Table C.1). We observe a score of $0.88 \pm 0.02$, which shows that the large majority of new patients are embedded in the latent space closer to patients diagnosed with the same type of cancer. Both results show that our latent space is robust in the sense that we can derive an embedding for new patients that is consistent with that of known patients and cancer types. We also observe that the k-nearest neighbour algorithm gives a more robust diagnosis classifier than finding the closest cancer in the latent space. This means that the local neighbourhood of a patient in the latent space is a better diagnosis indicator than a global predictor derived from cancer types' embedding. This suggests the presence of patient subgroups within a cancer type that display substantially different molecular behaviour.

Overall, our analysis shows that the patients/cancers latent space is consistent with known biology. Furthermore, our framework has the advantage of relaxing the hard clustering derived from patients diagnoses through patient's molecular similarity, highlighting that a patient's molecular profile can be more similar to the profiles of patients with different cancers than to the profiles of patients with the same diagnosis. This observation motivates further the need for pan-cancer perspectives in precision medicine.

### 5.4.2 Our gene latent space is biologically relevant

To evaluate the biological relevance of our genes embeddings, we cluster them in $k_2$ group using hierarchical clustering with cosine distance as well, and measure the enrichments of the clusters in terms of Gene Ontology (GO) annotations and in terms of cancer driver genes (see Methods). We consider all three subtypes of GO annotations: Biological Processes (GO-BP), Cellular Component (GO-CC), and Molecular Function (GO-MF) separately. The significance of the enrichments is computed with a hypergeometric test with Benjamini-Hochberg correction for multiple hypothesis testing and a significance threshold of 0.05. We observe that, regardless of the GO subtype, above 80% of clusters are significantly enriched in at least one annotation (see Figure 5.2.c.). These results show that genes with similar function are embedded closer in the latent space and thus that our genes' embeddings capture known biology. Interestingly, we also observe that around 10% of the clusters are enriched in cancer driver genes, indicating that cancer drivers are embedded closely, i.e. clustered, in the latent space. This highlights the link between the gene latent space and the cancer context that we made a part of our framework. Furthermore, it underlines the relevance of our embeddings for the identification of putative cancer-related genes, discussed in the following section.

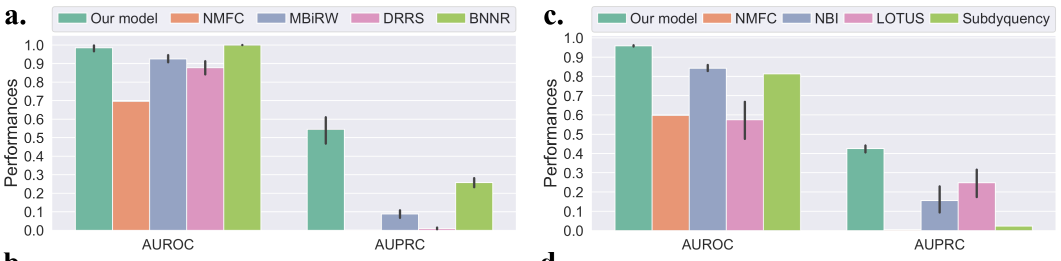
Additionally, we perform an ablation study on the gene interaction data input to investigate the effect that each dataset has on enrichment scores (see Appendix Figure C.2). For each model, all hyperparameters are selected following the same procedure outlined above. First, we observe that adding any gene data is better than not adding them from the point of biological annotation enrichment. For instance, the model without any PPI, complex, or pathway data has 40% of gene clusters enriched in GO-BP annotations, while every model with at least one data source has above 80% of gene clusters enriched for the same annotations. However, there is no clear best model among the ones with diverse combinations of the data, each scoring similar enrichment values with different models performing slightly better for different annotations. Thus, different combinations of data do not seem to lead to significantly different performances but keeping all data in the model enables analysis of each class of entities.

Finally, as pathways and complexes are embedded in the same latent space, we investigate their positioning with respect to genes. In particular, we evaluate whether a gene is embedded closer to its associated higher-order entities, i.e. pathways and complexes, than to those to which it has not been associated yet. To this end, we compute the cosine distances in the latent space between a gene and its associated pathways and complexes, termed "intra–pathway" and "intra-complex" distances, as well as the distances between the gene and all non-associated pathways and complexes, termed "exo–pathway" and "exo-complex" distances. We observe that genes are embedded closer, on average, to their associated higher-order entities than they are to those that they are not associated to (see Figure 5.2.d.), with average distance below 0.5 between a gene and associated entities and above 0.9 between a gene and non-associated entities. These results are significant according to a Mann–Whitney U statistical test (p-value $\sim 0$ in both cases) and underline the relevance of the joint embedding of genes with related higher-order molecular structures in the same latent space. This also suggests that our framework could be used for identifying new genes that are involved in or interact with pathways and protein complexes, which we leave for future work.

### 5.4.3 Predicting cancer type associations

To extract new knowledge for each cancer type, we use our context-aware embeddings to suggest cancer–drug and cancer–gene associations. We cast the problem as a link prediction task for which we train boosted decision trees to predict known associations from our entities' embeddings. After our training step, we use the trained classifiers to predict new associations (see Methods). As pre-processing, we normalise all embeddings to have a unit norm. The normalisation step is crucial for the transfer of a link predictor from one type of entity to another that we discuss in the next section. For each possible cancer–drug pair (or cancer–gene pair), we define the pair's representation by the concatenation of the embeddings of its components, i.e. the concatenation of the cancer's embedding vector with the drug's embedding vector defines the feature vector of the pair. Finally, we use boosted decision trees for link prediction, taking as input a pair's representation and output the association's scores of its component (see Figure 5.1.c.). We choose boosted decision trees due to their simplicity and high performances in several competitions [283].

In the first validation step, we systematically evaluate the performance of our approach with a 10-fold cross-validation using both the Area Under the Receiver Operating Curve (AUROC) and the Area Under the Precision recall Curve (AUPRC) and compare our results to state-of-the-art methods for links prediction. The splits used for the 10-fold cross-validation are performed on the set of known links and considering all non-reported links as part of the negative set of links. Furthermore, we perform an ablation study on the patient–gene data (see Appendix Figure C.5), i.e., we compare the results obtained with those obtained with the framework using less patient data to demonstrate the interest of considering both expression and mutation data jointly. In the second step, we investigate the top 10 drugs and genes associated with cancer types by our methodology. Each pair is scored based on the average of the standardised scores given by 10 classifiers trained for the cross-validation (see Methods). In this step, we only consider drugs and genes that were thus far not associated with any cancer type in the ground-truth data (introduced in each subsection) to avoid trivial cases of information transfer from one cancer to another, which typically happens when one drug or one gene is associated with a majority of cancers. We perform a manual literature curation to validate the top results.



**b.**

| DrugBank ID | Predicted for | Drug Status |
|---|---|---|
| **DB05916** | LUAD;LUSC | I |
| **DB14707** | LUAD;LUSC | A,I |
| **DB05101** | LUAD;LUSC | I |
| **DB12202** | LUAD;LUSC | I |
| **DB05374** | LUAD;LUSC | I |
| **DB01269** | LUAD;LUSC | A,I |
| **DB05931** | LUAD;LUSC | I |
| **DB06186** | LUAD;LUSC | A |
| DB00011 | SKCM | A,I |
| DB00018 | SKCM | A,I |

A:Approved, I:Investigational, E:Experimental

**d.**

| Gene | Predicted for | Support |
|---|---|---|
| **KAT2B** | BRCA | CCGD |
| **MDM2** | LIHC;BRCA | CCGD;CGC |
| **SP1** | BRCA | CCGD |
| **HERC1** | BRCA;COAD;READ | CCGD |
| **SMC3** | LAML | CCGD |
| **NCOA3** | BRCA | CCGD |
| **CHD6** | BRCA | CCGD |
| **SIN3A** | BRCA | CCGD |
| **SRC** | BRCA | CCGD;CGC |
| **PPARG** | BRCA | CCGD;CGC |

CGC:Cancer Gene Census, CCGD:Canditate Cancer Gene Database

**Figure 5.3:** Performances of our cancer–drug association predictor (left column) and cancer–gene associations predictor (right column). The bar charts give the performances of our classifiers measured with 10-fold cross-validation in terms of Area Under the Receiver Operating Characteristic (AUROC) and Area Under the Precision-Recall Curve (AUPRC). The tables give the top ten associations between cancers and drugs (panel b.) and genes (panel d.) that are not associated with any cancers in our data. Drugs or genes highlighted in bold font have been associated with cancer. The support column in the bottom right table indicates which database lists a link between the gene and cancer.

### 5.4.3.1 Our model predicts relevant treatment

To predict cancer–drug associations, we train boosted decision trees to identify known associations that we collect from DrugCentral [293] (last updated October 2018). DrugCentral contains 93 associations in total between our sets of cancer types and drugs. We define our positive set with DrugCentral treatment options and consider all non-reported associations for our negative set. Our classifier takes as input the concatenation of the normalised embeddings of a drug and a cancer type and outputs their association score. We compare our results to four baseline methods: Non-negative Matrix Factorization Reconstruction (NMFR), Measure-based Bi-directional Random Walks (MBiRW) [171], Drug Repositioning Recommendation System (DRRS) [294], and Bounded Nuclear Norm Regularization (BNNR) [35] (see Appendix C.1.2 for implementation details).

We observe that our approach significantly outperforms the competing methods (see Figure 5.3.a.). BNNR achieves slightly better AUROC scores ($\sim 0.99$ compared to our $\sim 0.97$), but it scores significantly lower than our framework in terms of AUPRC ($\sim 0.25$ compared to our $\sim 0.5$). These results show the relevance of our method when compared to the state-of-the-art drug re-purposing approaches. We analyse further the results of our approach through literature curation for the top-scoring drugs that are not associated with any cancer types in DrugCentral.

Among the top 10 drugs that are the most associated to cancer types by our classifiers (see Figure 5.3.b.), a majority is recorded in DrugBank as investigational or approved for the treatment of some cancers. The approved predicted drugs either are not present in DrugCentral, as their approval postdates the DrugCentral release, or target a cancer type not considered in this study. We discuss below supporting information for our top 3 predicted drugs. We provide validations of all of our predictions in Appendix C.2.1.

DB05916 (CT-011) targets gene PDCD1, which has immunomodulating and antitumor activities. CT-011 is currently being investigated for the treatment of tumours and unspecified cancers [285]. DB14707 (Cemiplimab) is an FDA approved drug for the treatment of advanced cutaneous squamous cell carcinoma [285]. Our classifier suggests that it could be used to treat lung cancer and notably lung squamous cell carcinoma (LUSC). DB05101 (Matuzumab) is an investigational drug that targets the EGFR gene, which is often associated with cancers, including lung cancers [295].

The manual literature curation highlights that our predicted drugs are often investigated, or approved for the treatment of forms of cancer and that their targets, or mechanisms of actions, can be linked to the specific cancer types we predict. Overall, the analysis strongly supports our methodology.

#### 5.4.3.2 Our framework identifies genes relevant to cancer types

Based on known cancer genes from IntOGen [43], we train classifiers to identify associations between genes and cancer types. In total, IntOGen reports 1,129 associations between our sets of cancer types and genes. Our positive set is a subset of these cancer–driver associations All non-reported associations are considered as part of our negative set. As above, a classification model takes as input the concatenation of the normalised embeddings of a gene and a cancer type and outputs their association score. We compare the performance of our method with the following state-of-the-art methods: Non-negative Matrix Factorisation Reconstruction (NMFR), Network-Based Integration (NBI) [296], LOTUS [11], and Subdyquency [12]. The methods were developed to predict cancer-related genes in slightly different contexts and are adapted to our problem here (see Appendix C.1.2).

We observe that our approach outperforms the competing methods (see Figure 5.3.c.) in terms of AUROC, which is over 0.9 for our method compared to below 0.8 for the other approaches, and in terms of AUPRC which are around 0.4 for our approach compared to below 0.25 for the other methods. We further evaluate if our approach accurately captures known cancer-related genes that are reported in CCGD but not in IntOGen. We perform this analysis both globally and on a per cancer basis (note that 14 cancer types have data for this test). We observe that our method ranks associations between genes and cancer types in CCGD highly, notably giving AUROC scores above 0.59 (p-values $< 10^{-7}$) for all cases (see Appendix Figure C.3). Below we look at the top 10 genes that are identified by our method (see Figure 5.3.d.). We use the Cancer Gene Census (CGS) [297] and Candidate Cancer Gene Database (CCGD) [298] to find known associations, as well as literature curation (both databases were accessed in August 2019).

We observe that our top 10 scoring genes are listed in either the Candidate Cancer Gene Database (CCGD), or the Cancer Gene Census (CGC) as linked to at least one form of cancer. Furthermore, the pairs MDM2–LIHC, HERC1–COAD, HERC1–READ, SMC3–LAML, NCOA3–BRCA, and CHD6–BRCA are associated in CCGD. Additionally, KAT2B (PCAF) activity has been linked to cancers, and in particular, to breast cancers, in the literature [299, 300, 301]. MDM2 has also been associated with breast cancer [302]. SP1 expression has been linked to breast cancer in multiple prior studies [303, 304, 305]. For each of these three genes, we stratify our BRCA cohort into two groups: patients having higher than average expression of the gene and patients having lower than average expression of the gene.

We compute a logrank statistical test (with 0.05 cut-off) and observe that for each of the three genes, the patient groups have significantly different survival rates with p-values 0.002 for KAT2B, 0.026 for MDM2, and 0.039 for SP1 (see Appendix Figure C.4 for Kaplan-Meir plots). For each of the three genes, higher expressions are associated with lower survival rates. We provide validations for the remaining predictions in Appendix C.2.2.

The literature curation highlights that each gene identified through our approach is relevant to the associated cancer type, with support through existing research and databases, as well as statistical evidence for a connection between gene expression level and patient prognosis. Thus, this supports our methodology.

### 5.4.4 Re-purposing classifiers

Links between types of entities are not always known or available (e.g., associations between cancer type and protein complexes or associations between patients and drugs), which prevents us from using the same methodology to derive new knowledge. However, our framework allows for extrapolating those links from known associations between other types of entities. Our approach relies on the previously observed fact that, by design, some entities are embedded in the same latent spaces (e.g., genes, pathways, and complexes or cancer types and diseases). We have further shown in the first section, that the relative location in the latent space of related entities was biologically consistent, i.e. related entities are closer to each other than non-related entities. Based on these observations, we postulate that a classifier trained from the embeddings of a given type of entities can be re-purposed to predict from the embeddings of another type of entities. For instance, boosted decision trees that learnt to associate genes to cancer types can be used to predict which biological pathways or protein complexes could be associated with which cancer types. This could effectively provide insights into the impact of cancers onto cells by identifying affected higher-order cellular structures and functions. We focus below on the analysis of top associations between higher-order cellular structures and cancer types.

#### 5.4.4.1 Our re-purposed classifiers identify cancer-related protein complexes

To obtain the association score between a cancer type and a protein complex, we simply feed the concatenation of the normalised embeddings of both entities to the 10 boosted decision trees trained to predict cancer–gene associations. The average of the standardised scores across all boosted decision trees gives the final association score.

To the best of our knowledge, there are no comprehensive database reporting associations between cancers and protein complexes. Thus, we are unable to provide global validation scores for our predictions. We proceed by validating the top 3 scoring protein complexes manually (see Table 5.4) below. We provide validations of all of our predictions in Appendix C.2.3.

| Protein Complex | Predicted for |
|---|---|
| IL6:sIL6R:IL6RB:JAKs | all cancers |
| p-7Y-RUNX1:PTPN11 | BRCA;BLCA;PAAD;LUSC;GBM |
| R-HSA-1112759 | BRCA;LAML;BLCA |
| Integrin alpha2bbeta3:SRC | BRCA;BLCA |
| R-HSA-1112753 | BRCA;LAML;BLCA;LGG;GBM |
| R-HSA-1112563 | BRCA |
| SAM68:p120GAP | BRCA |
| JAKs:OSMR | BRCA |
| IL6ST:JAKs | BRCA |
| R-HSA-9632399 | BRCA |

**Table 5.4:** Top 10 protein complexes associated to cancer types.

IL6:sIL6R:IL6RB:JAKs complex plays a role in interleukin 6 signalling, which is linked to cancer [306]. The complex is associated with JAK family of kinases; themselves tied to cancer [307]. p-7Y-RUNX1:PTPN11 complex is involved in the regulation of RUNX1 expression and activity. RUNX1 has been linked to various cancer, sometimes with opposite effects [298]. However, regardless of its precise role in a given cancer, the regulation of RUNX1 appear to be of critical importance as over- or under-expression can have an important impact on the development of cancer [308]. Tyrosine phosphorylated IL6 receptor hexamer:Activated JAKs:Tyrosine/serine phosphorylated STAT1/3 complex (R-HSA-1112759) is involved with interleukin 6 signalling and more specifically serine phosphorylation of STAT family of transcription factors. We have seen previously that interleukin 6 signalling has been linked to cancer. Furthermore, STAT has been linked to various cancers, including breast cancer (BRCA) that our results associate to the protein complex [309].

The literature curation highlights that the existing literature supports our predicted associations between protein complexes and cancer types. Thus, the analysis demonstrates the validity of our boosted decision trees re-purposing approach, as well as the ability of our framework to extract cancer mechanisms at the level of protein complexes.

#### 5.4.4.2 Our re-purposed classifiers identify cancer-related biological pathways

Similarly, we can predict associations between cancer types and biological pathways. The association score between a cancer type and a biological pathway is obtained by simply feeding the concatenation of the normalised embeddings of both entities to the 10 boosted decision trees trained to predict cancer–gene associations. The average of the standardised scores across all boosted decision trees gives the final association score.

CTD database [250] gives associations between diseases and pathways based on shared associated genes and can be used for global validation. We achieve an AUROC score of $0.65 \pm 0.01$ and an AUPRC score of $0.66 \pm 0.01$, which indicates predictions significantly better than random (p-value $\sim 0$) for our re-purposed classifiers. However, note that 52% of all possible associations between our set of cancer types and our set of pathways are reported in the database. This indicates that the condition for association used by CTD might not be sufficiently stringent. This motivates the following manual literature curation to validate our top 10 scoring biological pathways. We discuss the first 3 predicted pathways below and provide validations of all remaining predictions in Appendix C.2.4.

MAPK1 (ERK2) activation pathway (R-HSA-112411) and MAPK3 (ERK1) activation pathway (R-HSA-110056) have been linked to numerous cancers, such

| Pathway | Predicted for |
|---|---|
| R-HSA-112411 | all cancers |
| R-HSA-2262752 | BRCA |
| R-HSA-5654688 | BRCA;GBM;BLCA;LGG;PAAD;LAML;<br>LUSC;STAD;SKCM;LUAD;UCEC |
| R-HSA-5654699 | BRCA |
| R-HSA-8953897 | BRCA |
| R-HSA-110056 | BRCA |
| R-HSA-389357 | BRCA |
| R-HSA-5654719 | BRCA |
| R-HSA-9603381 | BRCA |
| R-HSA-8866910 | BRCA |

**Table 5.5:** Top 10 biological pathways associated to cancer types.

as breast cancer, as discussed in the previous section, and colorectal cancer [310]. The ERK MAPK pathway is critical for cell proliferation and thus is naturally often connected to cancers. Cellular responses to stress pathway (R-HSA-2262752) is a subpathway of the cellular responses to external stimuli pathway (R-HSA-8953897) [240]. Anticancer treatments are often successful when able to induce apoptosis through external stimuli that induce cellular stress [311]. For instance, tumour suppressor gene P53 can be stimulated via cellular stress [312]. Thus, perturbation to those pathways might lead to cancer onset and resilience to treatment.

The literature review highlights the ability of our classifier re-purposing approach to identify associations between biological pathways and cancer types that are supported by the existing literature. Thus, this analysis underlines the ability of our framework to extract cancer mechanisms at the level of biological pathways.

### 5.4.4.3 Predicting patients' responses to cancer drugs

We collect data on patient responses to cancer drugs from TCGA [271]. We only consider patients and drugs that are present in our dataset. The task corresponds to a binary classification where we predict if a patient's response to a drug is positive or negative. A response is considered positive if TCGA reports a complete response of the patient, and negative otherwise. We further discard entries corresponding to combinations of drugs as our model is not suitable for the analysis of these data. From the remaining data, we only consider drugs that have both positive and negative response. After processing, we have 2,589 patient–drug pairs. We split this data in train, validation, and test sets with a $70\%/10\%/20\%$ partition, repeating the experiment 10 times and measuring AUROC and AUPRC scores.

We train a boosted decision tree model to predict patients' responses to drugs. As above, the input to the model is the concatenation of the normalised embedding of a patient and a drug. The output can be interpreted as the success probability of the treatment. Our approach performs well, achieving AUROC score of $0.869 \pm 0.013$ and AUPRC of $0.855 \pm 0.014$. This result suggests that our model can capture some common biological mechanisms that govern response to cancer drugs.

To analyse this claim further, we investigate which features the models use most to predict response. First, we compute the gain, i.e. the relative importance, associated with each feature in each one of the 10 models trained. For each feature, we take the average gain across models as a final feature importance score. Note that we have both patient and drug features; thus, we have two vectors of feature importance scores, $\mathbf{i}_{patients}$ and $\mathbf{i}_{drugs}$. In a second step, we use the central matrices from the NMTF decompositions in our objective to link each feature to both genes and pathways. Specifically, we compute the projection of entity $x$ in either drug or patient space with $G_x^p = G_x S_t$, where $S_t$ is either $S_{dt}$, for the drug space, or $S_{exp} + S_{mut}$, for the patient space. We can then rank the importance of genes, or pathways, by taking the product of the projected embeddings with the feature importance vectors. Interestingly, the two rankings of genes that we obtained retrieve driver genes in IntOGen. We consider all driver genes regardless of cancer type and compute the AUROC and AUPRC scores of the two rankings. We obtain AUROC 0.72 and 0.63 (p-values $< 10^{-20}$) and AUPRC 0.08 and 0.04 in drug and patient space, respectively, which indicate significant correlations between the set of driver genes and the rankings. We take a closer look at the highest-ranked genes and pathways (see Table 5.6).

Interestingly, 8 of the genes identified in Table 5.6 have been linked to cancer response to general, or specific treatments (HSP90AA1 [313], PIK3CA [314], EGFR [315], PTEN [316], PRKACA [317], KRT19 [318], CLDN4 [319] , AGR2

| | Drug space | Patient space |
|---|---|---|
| Genes | HSP90AA1 | KRT19 |
| | PIK3CA | KRT8 |
| | EGFR | KRT18 |
| | PTEN | CLDN4 |
| | PRKACA | AGR2 |
| Pathways | Signal Transduction | Extracellular matrix organisation |
| | Signaling by GPCR | Transport of small molecules |
| | GPCR downstream signalling | Degradation of the extracellular matrix |
| | Metabolism | Signal Transduction |
| | G alpha (i) signalling events | Response to elevated platelet cytosolic Ca2+ |

**Table 5.6:** Top 5 genes and biological pathways associated to drug response prediction based feature importance in both drug and patient spaces.

[320]) and the remaining two have been associated to cancer prognosis (KRT8 [321], KRT18 [322]). Put together, the results indicate that our model assigns meaningful importance to features. This is further corroborated when investigating the predicted pathways. We observe that most pathways are linked to external signalling, notably G protein-coupled receptors (GCPRs) signalling. Signal transduction is naturally critical to drug response, as it is the route through which drugs interact with a cell [323, 324]. The state of the extracellular matrix also plays an important role, as it can prevent the penetration of small molecules into the cell, thus impairing pharmacologic treatments [325]. Thus, our model learns to weigh meaningful features that relate to biological processes involved in drug mechanisms of actions.

## 5.5 Conclusions

We introduce a two-step framework to perform data integration, feature reduction, and classification to uncover cancer-related knowledge. First, we develop an integrative non-negative matrix factorisation model to jointly embed entities in multiple connected latent spaces based on heterogeneous, diverse relational data between those entities. Note, that due to the wide range of data incorporated in our framework and the different levels of noise present in each, it might be worthwhile to investigate balancing strategies of the diverse objective functions to improve the results further. Our model can easily be modified to accommodate such approaches. We show that relative positions of entities in our latent spaces are consistent with what we know about them. For instance, we show that genes group in functional domains and are close to associated higher-order molecular structures (pathways and complexes) embedded in the same latent space. Patients tend to be closer to other patients having the same diagnosis and to the diagnosis itself. By taking a pan-cancer approach, we can identify groups of patients with similar molecular manifestations spanning various cancers, confirming that cancer classification may need to be rethought on a global scale and the need for initiatives such as PCAWG [274]. Based on known drug indications for the treatment of each cancer type and known cancer type driver genes, we train boosted decision trees through which we can predict relevant new associations for each cancer type. Due to the joint embedding of different entities in the same latent space, we hypothesised that boosted decision trees trained to identify associations with one type of entities could be repurposed to derive associations with other, less-studied, entities. In this way, we can uncover biological mechanisms affected by each cancer type.

Interestingly, our work opens the door for actionable precision medicine. Through the joint embedding of cancers and patients, boosted decision trees trained on high-level knowledge about cancer types can be re-purposed to help identify patient-specific information, such as potential drug treatment. Furthermore, our model can capture the underlying information relevant to the characterisation of patients' response to drug treatment. However, as the biological validation of such predictions is difficult, requiring cell-line experiments or clinical trials, we leave it for future work.

Our framework is general and flexible and can accommodate additional and different data. While we focus on cancer here, our work paves the way for general cross-diseases analysis that could be useful to identify treatment re-purposing based on molecular similarities among medical conditions.

# Chapter 6

# Conclusions

## 6.1 Thesis summary

The increasing amount of multi-modal biomedical data has enabled researchers to gain insights into biological systems and helped understand and treat pathological states. Integrative approaches (introduced in Section 2.3) harnessing the diverse, wide-ranging data sources, have been instrumental in this process. Notably, the inclusion of biological networks, modelling molecular dependencies as graphs (see Section 2.2), has given rise to integrated system-level representations that are essential to further our comprehension of biological states. Recently, the covid-19 pandemic has highlighted both how far we have progressed in our understanding of pathologies and how much there is still to discover. To answer the global health challenge, the community has been able to leverage and combine the available data to generate relevant hypothesis to understand and identify putative treatments [326, 327, 328]. Hence, among other crucial points, the crisis has highlighted the need for efficient algorithms able to jointly mine multi-source datasets to produce actionable knowledge for the practice of medicine.

This Thesis is part of the ongoing efforts to develop models and algorithms to improve patient care through the joint analysis of diverse biomedical data. We focus on the multi-scale protein interactome and its integration into machine learning models as a way to include the hierarchical structure of biological systems. Importantly, this enables the identification and investigations of pathological perturbation at the different scales of biological systems. Thus, it leads to a better comprehension of the mechanisms of a disease which ultimately enables more efficient diagnosis, prognosis, and the identification of best-suited treatments.

In Chapter 3, we propose the use of hypergraphs to represent protein complexes and biological pathways that constitute higher-order protein interactions in cells. We introduce hypergraphlets as connected, non-isomorphic, induced sub-hypergraph to characterise wiring patterns around nodes of a hypergraph. Using hypergraphlets statistics, we demonstrate that each level in the multi-scale protein interactome captures complementary biological information. This result motivates the importance of considering the multi-scale protein interactome to develop model capturing biological information. Based on this observation, we propose a joint hypergraph model of the multi-scale protein interactome. Using hypergraphlets, we highlight that proteins wired similarly in the hypergraph have similar biological functions. Furthermore, these proteins tend to be at a short distance from each other in the hypergraph in terms of shortest path distance. This result highlights that neighbourhoods in the multi-scale protein interactome are characteristic of biological functions. Based on this result, we propose to predict a protein's biological function based on the functions of proteins with similar wiring patterns. We show that this simple procedure can uncover the biological functions of uncharacterised proteins.

In Chapter 4, we introduce a visible machine learning neural network model based on the multi-scale organisation of proteins into biological pathways in cells. We use our model to predict patient's diagnosis based on differential gene expression. We show that the sparse model diagnosis performances are on par with competitive approaches. More importantly, we can identify links between diseases, pathways, and proteins by investigating the weights of the trained model. Identifying disease co-morbidities is a crucial task that can ultimately help to evaluate a patient's risk of developing a disease based on their medical history. This is especially useful for prevention and preemptive care. The identification of proteins and pathways associated to disease is essential for many applications in precision medicine (discussed in the introduction), such as finding biomarkers to help diagnosis and prognosis or uncovering putative drug targets paving the way for new treatment discovery. We show that our approach outperforms competing methods and heuristics and that our top-scoring associations have strong support in the literature despite not being reported in existing databases.

In Chapter 5, we develop a framework resting on the collective embedding of a wide array of biomedical entities based on their inter-relational links in a pan-cancer context. Specifically, we integrate diagnosis, gene expression and somatic mutation information of patients in multiple cohorts with the multi-scale protein interactome, drug–target and drug similarities data. By using jointly multiple datasets about each type of entities, our framework learns context-rich embeddings that can be used as input to machine learning models for downstream analysis and link predictions. Notably, by integrating the multi-scale protein interactome in the framework, we learn genes' embeddings that take into consideration the multi-scale cellular organisation. This enables us to identify biological function perturbed by cancer types and potential prognosis biomarker genes. Furthermore, by including data about drugs, we predict that existing drugs could be repurposed for the treatment of specific cancer types. We show that our approach outperforms the competing approaches and that the existing scientific literature strongly supports our results.

## 6.2 Future work

Extensions of our approaches to various biological applications are discussed in each chapter's conclusion. Hence, we focus here on future methodological directions that are relevant to the work presented in this Thesis.

### 6.2.1 Scaling up topological analysis of the multi-scale protein interactome

One issue with hypergraph (or simplicial complex) representations, and particularly with the associated topological descriptors (hypergraphlets and simplets) is that they do not scale well. Counting the substructures can be prohibitively time-consuming in the presence of large hyperedges. For instance, counting all hypergraphlets for each node in the hypergraph derived from the complete pathways set of Reactome, containing approximately $10,000$ nodes (proteins) and $1,700$ hyperedges (pathways), requires more than a week of computations.

An approach to address this issue could be to choose a different representation than hypergraph. For instance, an option is to use instead a multi-layer graph representation. Multi-layer graphs are used to represent systems with heterogeneous entities, or heterogeneous types of interactions [118, 119]. A multi-layer graph is a collection of networks, called *layers*, with the potential addition of inter-layer edges, i.e. edges connecting a node from a layer to a node from another layer. A multi-layer graph $M$ is defined by a pair $M = (\mathscr{G}, \mathscr{C})$, where $\mathscr{G}$ denotes the set of layers of $M$ and $\mathscr{C}$ corresponds to its set of inter-layer edges. If $M$ has $n$ layers, we denote them by $\mathscr{G} = \{G_1, \dots, G_n\}$, where $G_i$ corresponds to layer $i$ of $M$ and $G_i = (V_i, E_i)$ (recall that a layer is a graph). An inter-layer edge can connect any pair of nodes in any two layers (i.e. $\mathscr{C} \subseteq \cup_{i,j \in [1,n]: i \neq j} V_i \times V_j$). Note that the node sets of the layers are not necessarily disjoint, thus, to avoid confusion, a superscript denoting the index of the layer is used, such that $V_i = \{v_1^i, v_2^i, \dots, v_{p_i}^i\}$, where $p_i = |V_i|$. For instance in Figure 6.1, the multi-layer graph has four layers, $G_1$, $G_2$, $G_3$, and $G_4$.

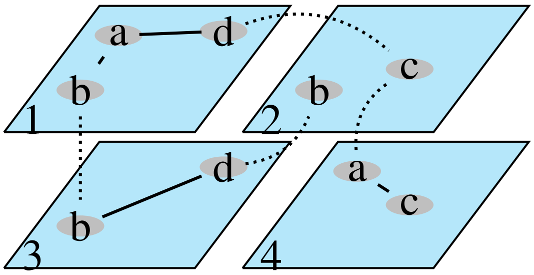


**Figure 6.1:** Examples illustrating a multi-layer graph. The multilayer graph has four layers, $G_1$, $G_2$, $G_3$, and $G_4$. Each layer correspond to a graph: $G_1 = (V_1 = \{a^1, b^1, d^1\}, E_1 = \{a^1b^1, a^1d^1\})$, $G_2 = (V_2 = \{b^2, c^2\}, E_2 = \emptyset)$, $G_3 = (V_3 = \{b^3, d^3\}, E_3 = \{b^3d^3\})$, and $G_4 = (V_4 = \{a^4, c^4\}, E_4 = \{a^4c^4\})$. Finally, the set of inter-layer edges corresponds to $\{b^1b^3, d^1c^2, b^2d^3, c^2a^4\}$, where the superscripts denote the layers. The figure is taken from Gaudelet and Przulj [116].

To represent the multi-scale protein interactome as a multi-layer graph, one can think of different modelling approaches, such as assigning a layer for each type of links (PPI, protein complex, and pathways) or assigning a layer to each higher-order entity (protein complexes and pathways) defining links based on PPIs. Both approaches have the merit of explicitly conserving the hierarchical structure of the multi-scale interactome. Multi-layer graph isomorphism problems can be cast as a graph isomorphism problems [329]. As such, graphlets can be easily extended to multi-layer graphs. For instance, Dimitrova *et al.* [330] recently extended graphlets to a subtype of multi-layer graphs: multiplex networks, which are networks with multi types of edges. However, while simple in theory, there are a few pitfalls. First, as the number of layers increases, so does the combinatorial possibilities and the number of potential substructures. Second, with few layers, nodes in the graph can have a very high degree which drastically increases the time needed for substructure counting. One question in this research direction will be whether a balance between those two extremes can be found to enable a scalable analysis of the multi-scale interactome.

A different research axis would be to develop machine learning models that can either directly approximate substructure counting or implicitly consider the topology in an end-to-end fashion, which we discuss in Section 6.2.3. Approximating graphlet counts has been the focus of multiple studies to scale up the use of graphlets to massive graphs with billions of edges [331, 332, 333]. These methods can serve as a basis to develop accurate estimators of subgraph counts for hypergraphs (or multi-layer graphs).

### 6.2.2 Visible machine learning

Visible Machine Learning (VML; discussed in Chapters 2 and 4) is a promising approach to develop powerful, interpretable, compact models by designing them based on prior biological knowledge.

An exciting research axis is how to handle the inherent noise and incompleteness of biological data in VML models. In the framework proposed in Chapter 4, for instance, we have noise coming from micro-array measurements and data processing, as well as noise and incompleteness from the gene–pathway associations. While we only partially addressed these issues in Chapter 4 with regularisation (eventually discarded), a more comprehensive model should account for them in a more principled way. For instance, the addition of bias terms and dropout on layers input can help handle noise and missing entries in the input data. Furthermore, to handle gene–pathway associations noise and incompleteness, a strategy could be *weight dropout* and, conversely, *weight drop-in*. Weight dropout has been introduced as an extension of standard dropout regularisation by Wan *et al.* [334]. In our case, weight dropout will ensure that the model does not rely on each specific gene–pathway links too heavily, thus accounting for noisy connections. In contrast, drop-in can be defined as the addition of some gene–pathway link with some low probabilities to model missing associations. The adverse effect of drop-in in its general form is that it would transform the layer in a fully connected layer, effectively removing sparsity from the model. One can limit this to an extent by only considering plausible gene–pathway links for drop-in, for instance, predicted gene–pathway associations with no experimental support. By this device, we would maintain sparsity while modelling missing associations.

Additionally, another direction is to leverage more data in similar models. For instance, Reactome gives information on the roles of proteins within pathways (input/output or catalyst). This information can be leveraged to define Pathway Capsule Neural Networks where each pathway is associated with a small neural network based on the proteins within the pathways and following their relative roles. The pathway hierarchy could also be used here in a similar fashion than Ma *et al.* [191] used the Gene Ontology directed acyclic graph. This model could be adapted for diverse tasks, in particular, to study time series data. Notably, it would be interesting to investigate if the model can capture the biological mechanisms governing the evolution of gene expression in time, particularly the cell cycle.

### 6.2.3 Knowledge graph embedding

Our framework in Chapter 5 can be seen as a knowledge-graph embedding approach for link predictions. Tensor and matrix factorisation constitute the state-of-the-art approaches for link predictions [335]. However, such methods cannot handle the addition of features describing entities, for instance, drugs' SMILES signatures or genes' DNA sequences. This issue is often resolved by computing a kernel between entities based on their signature and using those to constrain the embedding objective. In Chapter 5, we used such an approach to encourage the close embedding of chemically similar drugs according to the Tanimoto similarity measure. Although often efficient in practice, this approach implies the choice of a similarity measure that can lead to some loss of information. Furthermore, most recent machine learning practice advocate for end-to-end learning for which input data should be minimally processed, letting algorithms learn to extract features from the raw data. Graph Neural Networks (GNN) are an elegant alternative to analyse networked data that naturally take into consideration node features. GNNs have been first proposed at the beginning of the new millennium; however, they have only gained popularity and attention over the past couple of years owing to their successes in applications such as node classification and link prediction [336].

Various types of GNN layers have been introduced [337, 338, 339, 340, 341], most rest on the same principles. As an illustration, consider a graph $G = (V,E)$ with $n$ nodes, $V = \{1..n\}$, and feature matrix $X \in \mathbb{R}$ such that the $i^{th}$ row, denoted $X_i$, gives the features associated to the $i^{th}$ node. A typical, vanilla GNN layer corresponds to the composition of three functions. 1) A message-passing function MSG defined on the edges of the graph. The MSG function can take multiple forms such as the simple transfer of features from a source node to target node [338] or a more complex attention mechanism [339]. 2) A function AGG that aggregates for each node the messages from all incident edges. The AGG function is often simply the sum, the average, or concatenation of its inputs. 3) A function UPDATE that applies a transformation on the last node features. The UPDATE function is generally a multi-layer perceptron layer. Mathematically, this gives the following equation

$$H_i^{n+1} = \textsc{Update}\left(H_i^n, \textsc{Agg}\left(\left\{\textsc{Msg}\left(H_i^n, H_j^n\right), j \in \mathcal{N}_i\right\}\right)\right),$$

where $H_i^n$ corresponds to the features of node i after n layers, and $\mathcal{N}_i$ denotes the neighbourhood of node i. Thus, from a high-level perspective, a GNN layer updates a node's features based on the features of its direct neighbours in the graph. By stacking two GNN layers, one effectively ensures that a node's final features will depend as well on its second neighbours. With three layers, third neighbours will also have an influence, and so on. The deeper the model, the more extended the perceptive field. Conceptually, GNNs are related to graph diffusion and can be seen as learning weighted diffusion processes on a graph [342]. Importantly, Ying *et al.* [343] introduced GNNExplainer, an approach that enables the investigation of trained GNNs. GNNExplainer sheds light on how the model extract information, identifying important features and edges in the underlying graph. Thus, GNNExplainer enables the interpretation of GNN models which is essential in biological applications.

Three main tasks can be addressed with GNNs: node classification, link prediction, and graph classification. Node classification corresponds to the prediction of a node characteristics based on its features and the features of neighbouring nodes in the graph. For instance, protein biological function prediction based on PPI network [339]. Link prediction aims to uncover missing links in a graph. For instance, Zitnik *et al.* [344] introduced DECAGON to model polypharmacy side effects. Specifically, they construct a multi-layer network capturing PPI, drug–target, and polypharmacy side effects as (drug,side effect,drug) triplets. The authors proposed DECAGON, a GNN model that takes into consideration all types of edges jointly and is trained to predict polypharmacy side effects. The underlying class of problem is knowledge graph embedding that has been addressed with GNN [345] with reported improvements on existing matrix factorisation models. The final task, graph classification, aims to classify entire graphs based on their topology and node features [346]. This task is tied to the graph isomorphism problem. It requires the definition of a pooling function that extracts graph-level features from its nodes [347]. Examples of applications include molecule classification based on their graph representation [348].

Several recent studies have focused on the theoretical analysis of GNN in terms of representation power and generalisation. Vanilla GNNs have been shown to be at most as expressive as the Weisfeiler-Lehman algorithm that was introduced to test if two graphs are isomorphic [341]. Multiple recent publications have put into question the expressiveness of vanilla GNN and notably if GNNs can implicitly, or explicitly, capture wiring patterns. Recently, Chen *et al.* [349] highlighted that vanilla GNNs cannot count substructures in graphs. Furthermore, Dehmamy *et al.* [350] established minimal depth conditions on GCN [337], a type of GNN, to capture graph moments. However, several issues arise as GNNs go deeper such as the bottleneck [351] and over-smoothing [352]. As GNN models improve, it would be of interest to test if the new models can be used to approximate substructure counting effectively and thus if they implicitly take graph topologies into account.

Current thinking in the graph machine learning community to overcome these limitations is directed towards the development of GNN models that go beyond vanilla GNNs, incorporating global and relative information between nodes [353, 354]. Constructing better, task-oriented message passing strategies is also a critical axis of research [355]. These would benefit biomedical applications as it would help identify the most relevant, task-specific information contained within general biological networks. A simple illustration of this idea in biology comes from Kovacs *et al.* [356] whose work highlighted that similar proteins rather than being direct neighbours in a PPI network, instead tend to be two hops neighbours. This idea was recently used by Huang *et al.* [357] when proposing SkipGNN, a GNN that use two hops neighbourhood graph derived from the PPI network to augment its computational graph. The authors demonstrate higher performances on protein function predictions tasks when compared to a model using PPI networks out-of-the-box. This result raises the critical point that the most suited computational graph for a downstream task does not necessarily correspond to the input graph. This observation motivates developing methods able to identify, or infer, the best-suited graph for a specific task. Following this direction, researchers at Google have recently introduced Grale [358], a general-purpose framework that design computational graphs from multi-modal data in order to best address machine learning tasks.

Following this train of thoughts, it would be interesting to evaluate the use of the multi-scale interactome to design GNN models. A first approach would be using higher-order interactions to define meaningful long-range message passing between molecular entities. Vanilla GNNs have been extended to hypergraphs by Bai *et al.* [359]; however, the model suffers from the limitations of vanilla GNNs. Notably, with the multi-scale protein interactome, the high number of connections between proteins would naturally lead to over-smoothing and bottleneck issues. The problem likely requires novel approaches. One idea would be to use the information of protein roles in pathways, as given in Reactome, to define biological paths on a graph of proteins along which message-passing could be advantageously used. A different research direction would be to use the hierarchical molecular organisation to define biologically inspired pooling operations. Graph neural networks have been often presented as an extension of Convolutional Neural Network (CNN) operators used for image processing to irregular structures represented as graphs. However, a clear counterpart to the pooling layers used in image processing is yet to be defined for GNN models. Pooling layers have been instrumental in improving the performances of machine learning models processing images, with the state-of-the-art models using CNN and pooling layers in alternation. A few approaches have been proposed for graphs relying on features, structure, or both [347, 360, 361, 362, 363], however, all have important limitations, and none has been established as gold-standard. While defining a general-purpose graph pooling layer is a difficult problem, in the context of biology, we can take advantage of higher-order interaction between entities to define biologically meaningful pooling. For instance, a specific task could be to predict a patient's survival from its graph representation. Using patients' omics data, one can define a graph representation for each patient, for instance, using generic molecular graphs as a common basis (e.g. [156]). These graphs can be then used as inputs to GNN models in a graph classification setting to predict patients' characteristics such as survival or diagnosis. As graph classification tasks require pooling operations to extract a graph embedding from the nodes' embeddings, one could leverage the hierarchical biological organisation to define these operations. For instance, one could use biological knowledge to pool the subset of proteins involved in task-associated molecular processes, reducing noise associated with global pooling operations. Furthermore, the interpretability of GNN models with tools such as GNNExplainer could further help identify putative biomarkers.

# Appendix A

# Higher order molecular organisation as a source of biological function

## A.1 Simplicial complex

In essence, a simplicial complex is a set of simplices. A simplex is a geometrical figure defined by its number of dimensions (or nodes). A 0-dimensional simplex corresponds to a single node, a 1-dimensional simplex corresponds to a line (2 nodes linked by an edge), a 2-dimensional simplex is a triangle, a 3-dimensional simplex corresponds to a tetrahedron, and so on. Any subset of size $n$ of the $n+1$ nodes of a $n$-dimensional simplex forms a $(n-1)$-dimensional simplex called a *face*. Thus, a $n$-dimensional simplex has $n$ faces (e.g. a triangle has 3 faces corresponding to its edges). A simplicial complex $\mathscr{K}$ satisfies the following two conditions:

1. Every face of a simplex in $\mathscr{K}$ is also a simplex in $\mathscr{K}$.
2. The intersection of any pair of simplices of $\mathscr{K}$ is either empty or a face of both simplices.

A *facet* of a simplicial complex, is a simplex that is not the face of any higher-dimensional simplex. Its set of facets can thus summarise a simplicial complex.

## A.2 K-means

The k-means method aims to find a partition, or *clustering*, $\mathscr{P}$ of $n$ observations $v_1, \ldots, v_n$ (in our case, rescaled degree vectors of the proteins), into $k$ disjoints sets, or *clusters*, i.e. $\mathscr{P} = \{c_1, \dots, c_k\}$ ($k$ is a hyperparameter chosen by the user). Each cluster $c_i$ is defined by a vector called its *centroid*, $\mu_{c_i}$, which corresponds to the average of the rescaled degree vectors of the proteins it contains, i.e. $\mu_{c_i} = \frac{1}{n_i} \sum_{v \in c_i} v$ where $n_i$ corresponds to the number of proteins in cluster $i$. A k-means algorithm starts from random centroids and searches for the clustering that minimises the objective function

$$\operatorname{argmin}_{\mathscr{P}} \sum_{c \in \mathscr{P}} \sum_{v \in c} \|v - \mu_c\|^2. \tag{A.1}$$

Algorithmic details can be found in [207].

## A.3 Supplementary Figures

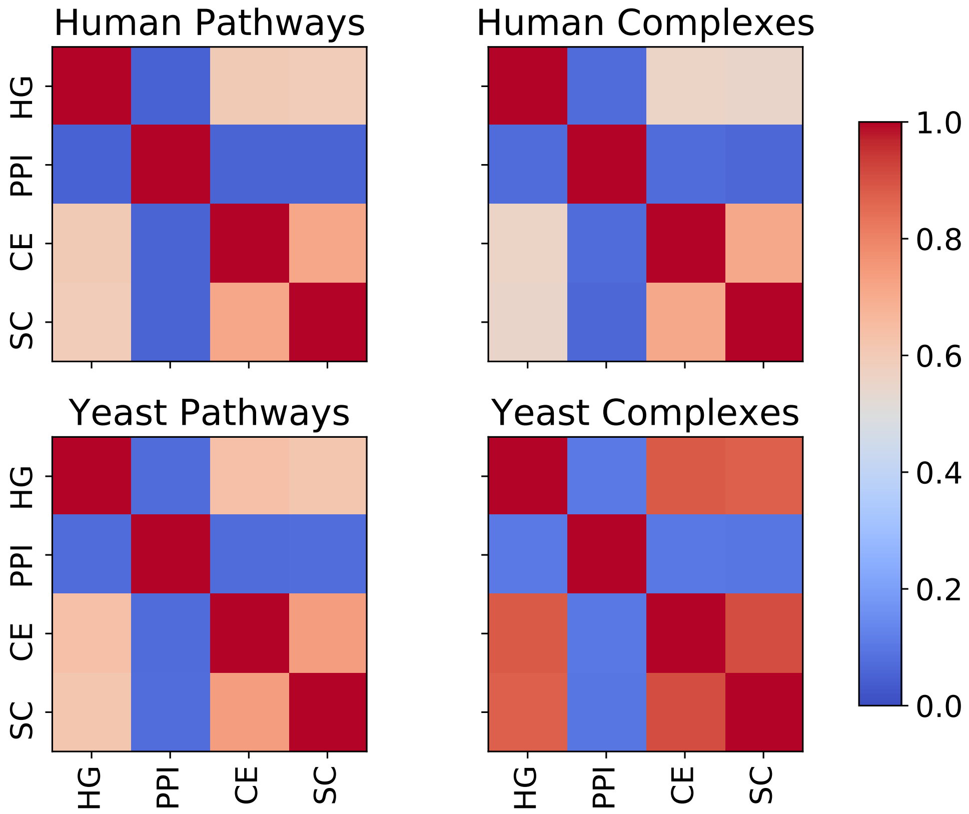


**Figure A.1:** Adjusted Mutual Information scores between pairs of clusterings obtained with the different topological representations for GO–BP.

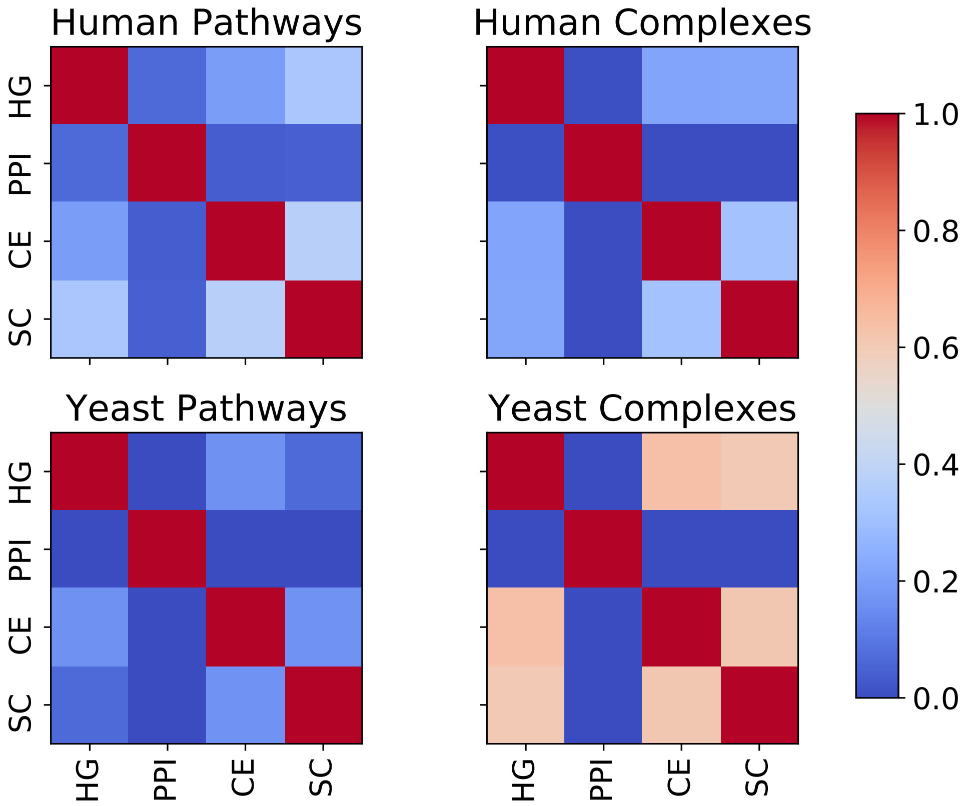
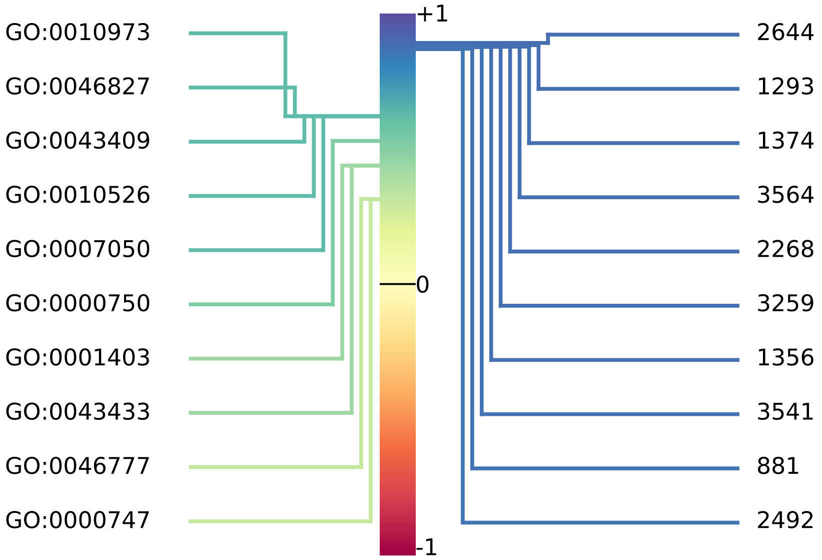


**Figure A.2:** Jaccard Index scores between pairs of clusteringsobtained with the different topological representations for GO–BP.



GO:0010973: positive regulation of barrier septum assembly

GO:0046827: positive regulation of protein export from nucleus

GO:0043409: negative regulation of MAPK cascade

GO:0010526: negative regulation of transposition, RNA-mediated

GO:0046777: protein autophosphorylation

GO:0007050: cell cycle arrest

GO:0000750: pheromone-dependent signal transduction involved in conjugation with cellular fusion

GO:0001403: invasive growth in response to glucose limitation

GO:0043433: negative regulation of sequence-specific DNA binding transcription factor activity

GO:0000747: conjugation with cellular fusion

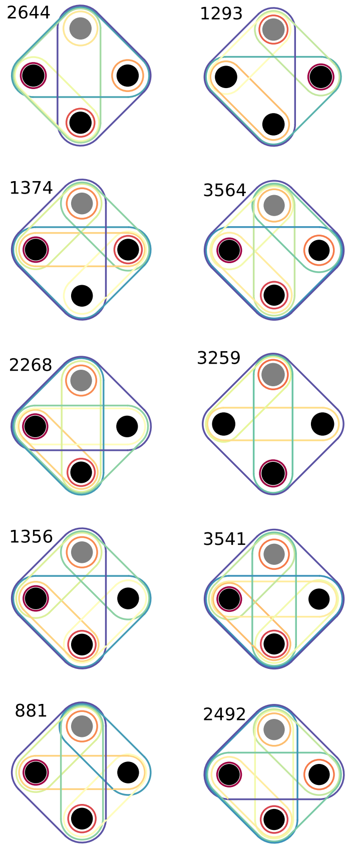


**Figure A.3:** The most significant CCA variate between HDVs of the proteins of yeast-pathways and their GO–BP annotations. The correlation score between the linear combination of annotations and the linear combination of hypergraphlet orbits is 1. The annotations (orbits) illustrated above correspond to the 10 that have the highest Pearson's correlation scores with respect to the linear combinations of annotations (orbits). Each GO term in blue font is annotating at least one protein conjointly with at least one other annotation that is also denoted in blue font, according to QuickGO ontology search engine [212].

## Appendix B

# Unveiling new disease, pathway, and gene associations via multi-scale neural networks

## B.1 Supplementary methods

### B.1.1 Metrics to evaluate classifier

The cross-entropy loss (CEL) of a classifier is defined as

$$\mathrm{CEL} = \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{n} -y_{ij} \log(\hat{y}_{ij}),$$

where $m$ represents the number of samples (patients), $n$ the number of classes (diseases), $y_{ij}$ indicates if patient $i$ is diagnosed with disease $j$ (1 if true 0 otherwise), and $\hat{y}_{ij}$ is the $j^{th}$ output value of the classifier for patient $i$. A relatively small CEL means that the output probability distribution of a classifier is closer to the deterministic one-hot encoding of the true labelling, i.e. the classifier gives a high probability to the true class and very small probabilities to the other classes.

The micro-averaged precision ($\mathrm{Pre}_\mu$) of a classifier gives a measure of the overall precision of the classifier. It is defined as

$$\mathrm{Pre}_\mu = \frac{tp}{m},$$

where $tp$ corresponds to the number of accurately classified patients and $m$ represents the number of patients.

The macro-averaged precision ($\text{Pre}_M$) of a classifier gives an average of the precision across the different classes (diseases) and is defined as

$$\text{Pre}_M = \frac{1}{n}\sum_{i=1}^{n}\frac{tp_i}{m_i},$$

where $tp_i$ corresponds to the number of accurately classified patients for disease $i$ and $m_i$ represents the number of patients diagnosed with the same disease. $\text{Pre}_M$ can be more informative than $\text{Pre}_\mu$ when considering the problem with class imbalance.

### B.1.2 Baselines

The Frequency of Differential Expression (FDE) score of a disease–gene association corresponds to how frequently that gene is consistently differentially expressed in patients having the disease, i.e., for disease $d$ and gene $g$, the association score, $s_{dg}$, is given by

$$s_{dg} = \left| \frac{1}{|\mathscr{P}_d|} \sum_{p \in \mathscr{P}_d} \mathbf{X}_{gp} \right|, \tag{B.1}$$

where we amalgamate entities (disease, gene, and patient) with their indices, $\mathscr{P}_d$ denotes the set of patients having disease $d$, and $\mathbf{X}$ corresponds to the data matrix introduced in Methods.

The Katz method uses disease-specific Protein–Protein Interaction (PPI) network, where each node of a standard PPI network is associated to a score (here the FDE of each gene for the disease considered). The authors then use Katz-centrality on each disease PPI network to extract a final score for each disease–gene association (here we use the absolute value). The higher the score, the higher the association is expected to be true. We download the PPI data from BioGRID [364] and IID [365] and create our PPI network from the union of both databases restricted to our set of genes. Finally, we perform a grid-search to identify the best parameters for the model by trying to maximise the area under the precision-recall curve metric.

## B.2 Supplementary figures

| Disease Name | Patients Count | Disease Name | Patients Count |
|---|---|---|---|
| non-small cell lung carcinoma | 490 | amyotrophic lateral sclerosis | 36 |
| oral cavity cancer | 248 | juvenile myelomonocytic leukemia | 34 |
| psoriasis | 223 | nasopharynx carcinoma | 31 |
| myelodysplastic syndrome | 187 | sarcoidosis | 30 |
| bacterial sepsis | 181 | dermatomyositis | 29 |
| colorectal cancer | 154 | myositis | 29 |
| asthma | 138 | cervical cancer | 28 |
| mature T-cell and NK-cell lymphoma | 131 | multiple sclerosis | 27 |
| alzheimers disease | 128 | turner syndrome | 26 |
| kidney cancer | 121 | interstitial lung disease | 25 |
| schizophrenia | 114 | multiple myeloma | 22 |
| chronic obstructive pulmonary disease | 89 | type 2 diabetes mellitus | 20 |
| pilocytic astrocytoma | 79 | essential thrombocythemia | 19 |
| thyroid cancer | 79 | sjogrens syndrome | 19 |
| bladder carcinoma | 79 | jobs syndrome | 18 |
| cerebrovascular disease | 78 | sotos syndrome | 18 |
| adrenocortical carcinoma | 77 | oral mucosa leukoplakia | 17 |
| uremia | 75 | rhabdoid cancer | 17 |
| endometriosis | 74 | dengue disease | 17 |
| major depressive disorder | 67 | esophagus squamous cell carcinoma | 17 |
| irritable bowel syndrome | 65 | ulcerative colitis | 17 |
| stomach cancer | 65 | anogenital venereal wart | 16 |

| | | | |
|---|---|---|---|
| oligodendroglioma | 64 | alcoholic hepatitis | 15 |
| systemic lupus erythematosus | 61 | campylobacteriosis | 14 |
| hepatocellular carcinoma | 59 | spondylosis | 14 |
| myocardial infarction | 57 | vitiligo | 14 |
| breast cancer | 57 | mitochondrial metabolism disease | 14 |
| malignant pleural mesothelioma | 55 | osteosarcoma | 14 |
| glioblastoma multiforme | 53 | cornelia de lange syndrome | 14 |
| acute myeloid leukemia | 52 | aphthous stomatitis | 13 |
| autistic disorder | 51 | sinusitis | 13 |
| hcv infection | 49 | sickle cell anemia | 13 |
| hepatoblastoma | 49 | atrial fibrillation | 13 |
| pancreatic ductal adenocarcinoma | 46 | hepatitis b | 12 |
| prostate cancer | 46 | peripheral vascular disease | 12 |
| ovarian cancer | 43 | acne | 12 |
| monoclonal gammopathy of undetermined significance | 43 | crohns disease | 11 |
| medulloblastoma | 41 | leishmaniasis | 11 |
| polycythemia vera | 41 | follicular lymphoma | 10 |
| atopic dermatitis | 40 | myelofibrosis | 10 |
| trachoma | 39 | leigh disease | 10 |
| rosacea | 38 | | |

**Table B.1:** Cohort size for each disease in the dataset.

# Appendix C

# Integrative Data Analytic Framework to Enhance Cancer Precision Medicine

## C.1 Supplementary Methods

### C.1.1 Projecting new patients

Projecting new, unseen patients in our framework latent space can be achieved by solving an objective function derived from the framework. Specifically, to find embeddings in our latent space for patients that were not seen by the model during the optimisation process, we minimise the objective function

$$\mathscr{L} = \min_{G.,S.} \sum_{x \in \{exp,snv\}} \|X_x - G_p S_x^* G_g^{*T}\|_F + \|X_{pct} - G_{ct}^* G_p^T\|_F,$$

where $X_x, x \in \{exp, snv\}$ represent the molecular data of the patients, $X_{pct}$ gives the patient diagnosis. The star superscripts $\cdot^*$ denote factors that are fixed in the original framework decomposition. Once this objective is minimised, $G_p$ gives the embeddings of the new patients in the latent space. We measure the quality of the embeddings of the new patients by quantifying if the patient is embedded close to its diagnosis and close to other patients having the same cancer in the original dataset.

The first aspect is quantified with macro-F1 scores of a classifier that associate to each patient the closest cancer in the latent space. The second aspect is quantified with macro-F1 scores of a classifier that associates to each patient the diagnosis that is most represented among the 10 nearest patients in the latent space.

### C.1.2 Baselines

We contrast the performances of our trained boosted decision tree with those of the state-of-the-art methods for the prediction of cancer type associations with genes and drugs. We chose baselines based on the availability of source code (or detailed implementation description), the quality of reported performances, and the concordance of input data with ours. Our implementation of each method is available in the Supplementary Files. When a method requires hyperparameters tuning, the criterion used to identify the best set of hyperparameters is always the AUROC score of the classifier in the associated task.

**Non-negative Matrix Factorization Reconstruction (NMFR)** is based on the reconstruction of the data after factorizations and is the simplest approach based on our framework. The idea is based on the matrix completion property observed in matrix factorizations methods[281]. Here, we propose a simple method that makes use of the link between factors to extract entities' association scores. For instance, cancer–gene association scores, $CG$, are given by

$$CG = G_{ct} G_p^T G_p \frac{(S_{exp} + S_{snv})}{2} G_g^T,$$

where entry $(i, j)$ of matrix $CG$ gives the association score between cancer type $i$ and gene $j$. Cancer–drug association scores, $CD$, are given by

$$CD = G_{ct} G_p^T G_p \frac{(S_{exp} + S_{snv})}{2} G_g^T G_g S_{dt} G_d^T,$$

where entry $(i, j)$ of matrix $CD$ gives the association score between cancer type $i$ and drug $j$.

Performances are measured by how well those association scores correlate to

IntOGen cancer–driver data and DrugCentral cancer–drug data using AUROC and AUPRC.

**MBiRW** [171] was proposed to identify potential new indications for the existing drugs. The method is based on a bi-directional random walk using a drug similarity network, a disease similarity network, and a bipartite network connecting diseases to drugs. The authors report good performances against known ground-truth relative to the competing methods and manually validate de novo predictions.

Here, the drug similarity network adjacency matrix is given by the drug Tanimoto similarities matrix $X_{ts}$. We define the cancer similarity network $X_{cc}$ based on a molecular similarity between cancers. We first associate to each cancer two molecular signatures given by the average of patients gene expression data and SNV data. From each type of data, we define a cancer similarity network that corresponds to the cosine similarity of their molecular signatures. We denote by $X_{cc}$ the final cancer similarity network corresponding to the average of those two similarity networks. Note that the authors use a different disease similarity matrix, the source of which is currently offline. Finally, we use cancer–drug data from DrugCentral to define a bipartite network in which an entry is set to 1 to indicate an association between the corresponding drug and cancer, and 0 otherwise.

The authors propose an iterative method that follows the step given in Algorithm 1.

We perform a 10-fold cross-validation to select hyperparameters $\alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and $M \in \{2, 5, 10, 20\}$ (note that the authors set $M = 2$, and search for $\alpha$). In each run, 10% of known cancer–drug associations are masked in the input to the algorithm and we evaluate how well MBiRW is able to retrieve those.

**DRRS** [294] was proposed to identify potential new indications for the existing drugs as well. The method is based on the matrix completion property of Singular Value Thresholding Algorithm (SVT) using a drug similarity matrix, a disease similarity matrix, and a disease–drug indication matrix. The authors report good performances against known ground-truth relative to the competing methods and

```
Data: cancer–cancer network $X_{cc}$, drug–drug network $X_{ts}$, cancer–drug
      network $CD$, parameter $\alpha$, maximum number of iterations $M$
Result: cancer–drug associations scores $O$
$DD = D_{ts}^{-\frac{1}{2}} X_{ts} D_{ts}^{-\frac{1}{2}}$;where $D_{ts}$ is a diagonal matrix where entry
 $D_{ts}(i,i) = \sum_j X_{ts}(i,j)$
$CC = D_{cc}^{-\frac{1}{2}} X_{cc} D_{cc}^{-\frac{1}{2}}$;where $D_{cc}$ is a diagonal matrix where entry
 $D_{cc}(i,i) = \sum_j X_{cc}(i,j)$
$R_0 = \frac{CD}{\sum CD}$; where $\sum CD$ gives the number of non-zero entries in $CD$
$O = R_0$; $n_iter = 0$;
while $n_{iter} \leq M$ do
    $L = \alpha \cdot O \cdot DD + (1-\alpha) \cdot R_0$;
    $R = \alpha \cdot CC \cdot O + (1-\alpha) \cdot R_0$;
    $O = \frac{L+R}{2}$;
    $n_{iter} = n_{iter} + 1$;
end
```

**Algorithm 1:** MBiRW algorithm.

further use their methods to predict indications for new drugs, validating novel associations.

For our purposes, we use the Tanimoto drug similarity matrix $X_{ts}$, the cancer–cancer similarity matrix derived from ICGC molecular data $X_{cc}$, and cancer–drug associations from DrugCentral $X_{cd}$. The authors define the block matrix $A$

$$A = \begin{pmatrix} X_{ts} & X_{cd}^T \\ X_{cd} & X_{cc} \end{pmatrix},$$

and feed it to the SVT algorithm. The maximum number of epochs is set to the minimum between the number of cancers and the number of drugs. The iteration with the highest AUROC gives the final predictions. We perform a 10-fold cross-validation to evaluate the performance of the algorithm. In each iteration, we mask 10% of known cancer–drug associations and evaluate how well the algorithm retrieves them.

**BNNR** [35] was developed for re-purposing of drugs. The method is also based on the completion property of SVT. The algorithm follows the steps given in Algorithm 2. Compared to DRRS, BNNR incorporates a regularisation term to balance the approximation error and the rank properties and thus can handle the noisy drugdrug and diseasedisease similarities. It also adds a constraint that clips the association scores to the interval $[0,1]$. The authors found that those additions benefited performances compared to DRRS for retrieval of known associations. They further manually validated the top-scoring associations through literature curation.

```
**Data:** block matrix $A$, hyperparameters $\alpha$ and $\beta$
**Result:** cancer–drug associations scores $W$
$tol_1 = 2 \cdot 1e^{-3}$; $tol_2 = 1e^{-5}$;
$M = 300$; $n_{iter} = 0$;
$s_1 = 1$; $s_2 = 1$;
$X = A$; $Y = X$; $W = X$;
$mask = T > 0$;
**while** $n_{iter} \leq M$ *or* $tol_1 < s_1$ *or* $tol_2 < s_2$ **do**
    $F = \frac{1}{b}(Y + \alpha T) + X$;
    $W = F - \frac{\alpha}{\alpha+\beta} F \odot mask$;
    $W[W < 0] = 0$; $W[W > 1] = 1$;
    $X_t = \mathrm{svt}(W - \frac{1}{\beta} Y, \frac{1}{\beta})$;
    $Y = Y + \beta(X_t - W)$;
    $s_t = s_1$; $s_1 = \frac{\|X_t - X\|_F}{\|X\|_F}$;
    $s_2 = \frac{|s_1 - s_t|}{\max(1, |s_t|)}$;
    $X = X_t$; $n_{iter} = n_{iter} + 1$;
**end**
```

**Algorithm 2:** BNNR algorithm.

We perform a 10-fold cross-validation to evaluate the performance of the algorithm and fix the hyperparameters $\alpha \in \{0.1, 1, 10, 100\}$ and $\beta \in \{0.1, 1, 10, 100\}$.

**Network Based Integration (NBI)** [296] was developed to identify cancer-related genes that are not necessarily mutated or differentially expressed. The method is based on network heat diffusion process over a molecular network. The original paper focuses on a single cancer for which they collect differential gene expression data and SNV data. The authors assess performances by first measuring how accurately their method retrieves known cancer driver genes and then validate novel cancer–gene associations.

Network heat diffusion is defined by the iterative update of scores $X^0$ associated to the network's nodes following the equation

$$X^{n+1} = \alpha W X^n + (1 - \alpha) X^0,$$

where $W$ denote the network data and $X^n$ corresponds to the updated scores after $n$ iterations. The iterative process terminates when $\|X^{n+1} - X^n\|_2 < 10^{-6}$.

The authors set $W$ to the PPI normalized adjacency $W = \Delta^{-\frac{1}{2}} X_{ppi} \Delta^{-\frac{1}{2}}$, where $\Delta^{-\frac{1}{2}}$ is a diagonal matrix where entry $i$ corresponds to the degree of gene $i$ in the PPI network. The authors use the diffusion process both on patients differential gene expression and SNVs, obtaining two diffused vector scores per patient. They then handcraft 13 cancer-specific features for each gene based on the results. Those features are then used as input to a logistic regression classifier trained to predict known cancer drivers.

As differential gene expression is not available to us, we use the same gene expression values that are input to our framework. Since our analysis is across cancers, we compute gene features for each cancer types, i.e. each cancer–gene pair is associated with a 13-dimensional feature vector. The method has two hyperparameters: $\alpha$ for the heat diffusion process, and $C$ that controls regularisation of the logistic classifier. We perform a 10-fold cross-validation procedure to pick the best pair of hyperparameters, with $\alpha \in \{0.25, 0.5, 0.75\}$ and $C \in \{0.01, 1, 100\}$, and to evaluate the performance of a logistic regression classifier trained on those features to predict cancer-specific driver genes.

**LOTUS** [11] is a method that achieved the state-of-the-art results for the more specific tasks of identifying oncogenes and tumour-suppressing genes. Each task is tackled separately, and each with 3 different gene features that are not available to us. However, the method can be adapted to the simpler task of retrieving cancer driver genes. To this end, we use gene expression data, SNVs, and gene methylation data, that are available in ICGC, as gene features. A patient's gene methylation is defined as the average beta value of all associated CpG islands.

The authors of LOTUS propose both a cancer-specific framework and a pan-cancer framework; we use the latter here. The method revolves around the Support Vector Machine (SVM) algorithm. The authors first define for each sample–gene pair 3 features that are then averaged across all samples to give the final gene features. In their work, the final features correspond to the number of damaging missense mutations, the total number of missense mutations, and the entropy of the spatial distribution of the missense mutations on each gene, for the prediction of oncogenes. For the prediction of tumour-suppressing genes, the features are the number of frameshift mutations, the number of loss-of-function mutations, and the number of splice site mutations. Note that when defining those features, the authors do not differentiate across cancer types. In our case, the features correspond to the mutation frequency, the average gene expression, and the average gene methylation across all samples. To ensure that those features are comparable, we normalise the distributions to have 0 mean and unit variance.

The authors then define both a gene kernel $K_g$ and a cancer kernel $K_c$. The gene kernel is defined as the average of a kernel corresponding to a gene similarity matrix derived from 3-dimensional features defined above and a kernel derived from the PPI network. We have

$$K_g = \frac{1}{2}\left(\Phi\Phi^T + e^{-L}\right),$$

where $\Phi \in \mathbb{R}^{15,224\times 3}$ represents the gene features and $L$ is the normalized Laplacian of the PPI network, $L = I - D^{-\frac{1}{2}} X_{ppi} D^{-\frac{1}{2}}$, $I$ represents the identity matrix and $D$ the diagonal matrix with entries corresponding to the degree of each node in the PPI network. The cancer kernel is defined as the sum of three kernels

$$K_c = \frac{1}{3}\left(I + J + X_{cc}\right),$$

where $I$ represents the identity matrix, $J$ corresponds to the matrix filled with ones, and $X_{cc}$ is a cancer similarity matrix. As above, we use the cancer similarity matrix defined based on cancers molecular similarities.

The final kernel $K$ for (cancer,gene) pairs used for pan-cancer analysis is defined by

$$K\left((c,g),(c',g')\right) = K_c(c,c') \times K_g(g,g'),$$

where $c$ and $c'$ represent cancers and $g$ and $g'$ represent genes. The hyperparameter of the model corresponds to the regularisation coefficient $C$ of the SVM algorithm. Due to the large size of the full kernel $K$, we use the same strategy as the authors of LOTUS and randomly sample negative (cancer,gene) pairs from all (cancer,gene) pairs that are not reported in IntOGen. This effectively boils down to using a submatrix of kernel $K$ as input that contains as many positive (cancer–driver associations) and negative pairs. As before, we perform cross-validation to pick $C \in \{1, 10, 100, 1000\}$ and evaluate the performance of the method on our task.

**Subdyquency** [12] is a method based on random walks on a network to identify cancer drivers. It achieves the state-of-the-art results for the retrieval of known cancer drivers. The framework is defined for specific cancers, and we extend it to pan-cancer. In their framework, the authors build a network between "outlier" and mutated genes. The outlier genes are genes whose expression is significantly different with respect to the cohort. They correspond to genes with absolute z-score strictly greater than 2. The set of outlier (mutated) genes is defined as all genes being at least outlier (mutated) for one patient. Here, we consider all cancers together to define those sets. Directed interactions between genes are obtained from the Functional Interactions (FI) network [366] derived from Reactome [240]. We downloaded the 2019 version of the FI network. The authors define a bipartite graph between the two sets of genes whose edges correspond to directed links in the FI network. The edge weights are defined based on the localisation of proteins in a cell as given by the COMPARTMENT database [367] (see the original paper for details). As done by the authors, we downloaded all data relating to human regardless of evidence type (obtained in April 2020). We denote the adjacency matrix of this bipartite graph with $W \in \mathbb{R}^{n_m \times n_o}$, where $n_m$ and $n_o$ represent the number of mutated and outlier genes, respectively. Then, for each patient $p$, the authors define a feature vector for the outlier gene set, denoted by $O_p$, and another one for the mutated gene set, denoted by $M_p$. Specifically, consider gene $i$ in the mutated set. If gene $i$ is mutated for patient $p$, then $M_p(i)$ is set to the mutation frequency of gene $i$ in the cohort of patients having the same cancer as $p$, and 0 otherwise. For gene $j$ in the outlier set, if $j$ is not an outlier for patient $p$, then $O_p(j)$ is set to 0, if $j$ is an outlier and is also in the mutated set, then $O_p(j)$ is set to $M_p(j)$, else it is set to the outlier frequency of $j$ across the set of patients having the same cancer as patient $p$. The authors then propose the three steps procedure simulating a random walk on the bipartite graph using both feature vectors

$$\begin{aligned} R_p^m &= \alpha M_p + (1-\alpha) W O_p, \\ R_p^o &= \alpha O_p + (1-\alpha) W^T R_p^m, \\ R_p^m &= \alpha M_p + (1-\alpha) W R_p^o, \end{aligned}$$

where $\alpha \in [0,1]$ is the sole hyperparameter of the model. The final cancer–gene scores are derived by summing the $R_p^m$ vectors across patients. Higher scores indicate a stronger association between a cancer and a gene. We perform a cross-validation to pick $\alpha \in [0,1]$ and evaluate the performance of the method on our task.

## C.2 Supplementary Results

### C.2.1 Cancer–drug associations

Due to space limitations, we discuss here the supporting literature for the remaining predicted drugs in Figure 3.b. of the main article.

DB12202 (Zalutumumab) targets EGFR gene and is investigated for the treatment of Squamous Cell Cancer and Head and Neck Cancer.

DB05374 (Rindopepimut) is a drug investigated for the treatment of brain cancers. It targets the mutant protein EGFRv3, which has recently been identified as a target in lung cancer therapy as well as [368].

DB01269 (Panitumumab) is approved for the treatment of EGFR-expressing colorectal carcinoma. Since EGFR is often also involved in lung cancers, the predicted associations here are relevant.

DB05931 (Pegdinetanib) is an investigational drug for the treatment of unspecified cancers. It binds to gene VEGFR-2 regulating primary tumour angiogenesis pathways, thus blocking ligands from binding to VEGFR-2.

DB06186 (Ipilimumab) is an approved drug for the treatment of multiple cancers, such as renal cell carcinoma, melanoma, and colorectal cancer. It binds CTLA4 to block the T-cell inhibition signal pathway. Erfani *et al.* [369] suggested that therapies targeting CTLA4 might be beneficial to lung cancer patients.

DB00011 and DB00018 (interferon alpha-n1 and interferon alpha-n3) are proteins that both targets interferon alpha/beta receptors 1 and 2. A similar protein, interferon alpha-2b, is among the approved treatments of SKCM. The mechanisms of action of these protein-based treatments are identical according to Drugbank. This supports the prediction of both DB00011 and DB00018 for the treatment of SKCM.

### C.2.2 Cancer–gene associations

Due to space limitations, we discuss here the supporting literature for the remaining predicted genes in Figure 3.d. of the main article.

We predict that HERC1 is associated with BRCA, COAD, and READ. The associations with COAD and READ are already reported in CCGD. Furthermore, HERC1 has been linked to migration and invasion of breast cancer cells [370]. We further observe, with a logrank statistical test (0.05 cut-off), that higher than the average expression of HERC1 in our BRCA cohort leads to significantly lower survival rates (pvalue 0.0498; see Supplementary Figure 2.d). Inversely, lower than the average expression of HERC1 in our READ cohort indicates significantly lower survival rates (p-value 0.008; see Supplementary Figure 2.e).

Both NCOA3 and CHD6 are linked to BRCA in CCGD. We further observe here, with a logrank statistical test, that higher than the average expression of those genes indicates lower survival rates in our BRCA cohort with p-values 0.0014 and 0.0155, respectively (see Supplementary Figure 2.f/g).

We predict that SIN3A gene is linked to BRCA, which is supported by existing literature [371, 372].

SRC is a proto-oncogene linked to colon cancer, according to NCBI. It has notably been connected to breast cancer in the scientific literature [373].

PPARG has been identified as a potential target for cancer treatment and prevention [374]. It has also been associated with the induction of apoptosis in breast cancer cells [375]. This is consistent with our logrank statistical analysis which shows that higher than the average expression of PPARG in our BRCA cohort is associated with significantly higher survival rates (p-value 0.016; see Supplementary Figure 2.g).

### C.2.3 Cancer–complex associations

Due to space limitations, we discuss here the supporting literature for the remaining predicted protein complexes in Table 1 of the main article.

Integrin alpha2bbeta3:SRC complex plays a role in integrin signalling, more specifically in the phosphorylation of SRC kinase. Our prediction is supported by the fact that integrin alpha2bbeta3 is part of the oncogenic MAPK signalling pathways[240] which includes inactive SRC (phosphorylated Y530).

IL6:Tyrosine phosphorylated hexameric IL-6 receptor:Activated JAKs:p-Y546,Y584-PTPN11 complex (R-HSA-1112753) is implicated in MAPK1/MAPK3 signalling. Both MAPK1 and MAPK3 are identified as driver genes in some cancers[298] and phosphorylated PTPN11 (p-Y546,Y584-PTPN11) is linked to PI3K/AKT signalling in cancer[240]. Furthermore, the MAPK pathway, through interplay with PI3K, has been linked to breast cancer [376] and leukaemia [377].

Tyrosine phosphorylated IL6ST:Activated JAKs complex (R-HSA-1112563) is linked to both interleukin 6 signalling, MAPK1/MAPK3 signalling, and phosphorylated PTPN11. Based on the discussion above, this complex is relevant to breast cancer.

SAM68:p120GAP complex has been linked to the insulin receptor signalling pathway. More specifically, Sánchez-Margalet *et al.* [378] reports that Sam68 is associated with p120GAP after insulin stimulation and links GAP to the PI3K pathway. We have already seen that PI3K pathway has been linked to breast cancer in previous work. It is also the case that insulin receptor signalling plays a role in breast cancer [379].

JAKs:OSMR complex is also involved in Interleukin-6 family signalling. Furthermore, it has been suggested that OSMR and JAK/STAT3 signalling can promote breast cancer progression [380].

IL6ST:JAKs complex is a subunit of IL6:sIL6R:IL6RB:JAKs that is involved in signalling by interleukins. The same supporting evidence links it to cancer.

Me260-ESR1:STRN:ESTG:MyrG-pY419 SRC:PI3K alpha complex (R-HSA-9632399) is part of the extra-nuclear estrogen signalling pathway which has been linked to tumour progression and metastasis [381]. Furthermore, protein MyrG-p-Y419-SRC, bound to PI3K in the complex, is associated with various cancer pathways[240], including PI3K/AKT signalling in cancer.

### C.2.4 Cancer–pathway associations

Due to space limitations, we discuss here the supporting literature for the remaining biological pathways in Table 2 of the main article.

SHC-mediated cascade:FGFR1 pathway (R-HSA-5654688), SHC-mediated cascade:FGFR2 pathway (R-HSA-5654699), and SHC-mediated cascade:FGFR4 pathway (R-HSA-5654719) are all three sub-pathways of FGFR signalling pathway which has been connected to cancer [382]. The role of SHC in this is unclear, but observations suggest that it might contribute to the activation of the MAPK pathway [240].

CD28 dependent PI3K/Akt signalling pathway (R-HSA-389357) is associated with cell growth and survival, roles that are critical in the development of cancer. The PI3K/Akt signalling pathway has been identified as a potential therapeutic target for cancers, including breast cancer [383]. Furthermore, mutation of CD28 receptors has recently been linked to increased risk in breast cancer [384].

Our results suggest that activated NTRK3 signals through PI3K pathway (R-HSA-9603381) are associated with breast cancer. The activation of NTRK3 correlates with activating phosphorylation of AKT, the principal mediator of PI3K signalling. Thus, this association is directly related to cancer, as shown in a previous discussion.

TFAP2 (AP-2) family regulates transcription of growth factors, and their receptors pathway (R-HSA-8866910) is associated with breast cancer through ESR1 and ERRB2. The entry for the pathway in Reactome [240] details the link between TFAP2 family and expression of ESR1 in breast cancer.

## C.3 Supplementary figures and tables

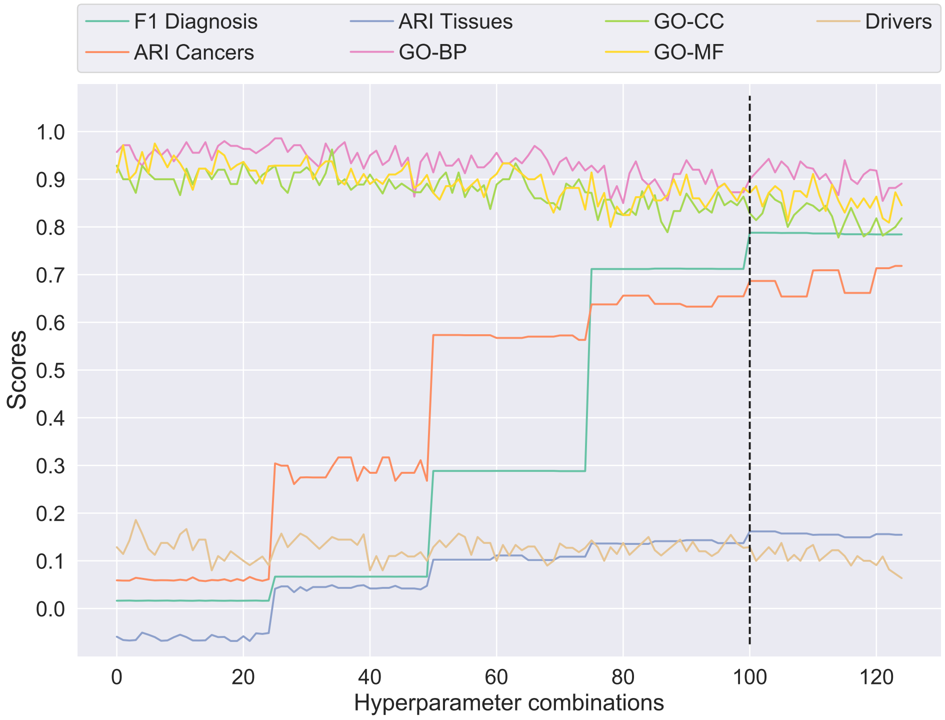


**Figure C.1:** Framework sensitivity to hyperparameters with respect to different scores relating to patients and genes embeddings. Sequence of hyperparameter combinations is defined with set product $(k_1, k_2, k_3) \in \{2, 5, 10, 15, 21\} \times \{70, 80, 90, 100, 110\} \times \{40, 50, 60, 70, 80\}$. Black dashed vertical line indicates the best set of hyperparameters according to the macro-F1 score with respect to patients' diagnosis.

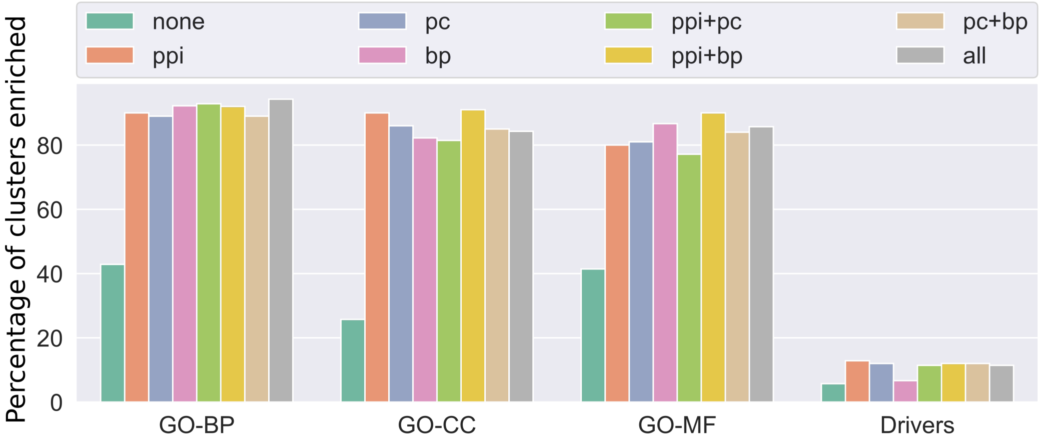


**Figure C.2:** Evolution of enrichment scores based on gene data ablation.

| Closest cancer macro-F1 | Nearest neighbours macro-F1 |
|---|---|
| $0.773 \pm 0.028$ | $0.882 \pm 0.015$ |

**Table C.1:** F1 scores measuring how closely new patients are embedded into the latent space to their cancer type and patients having the same cancer.

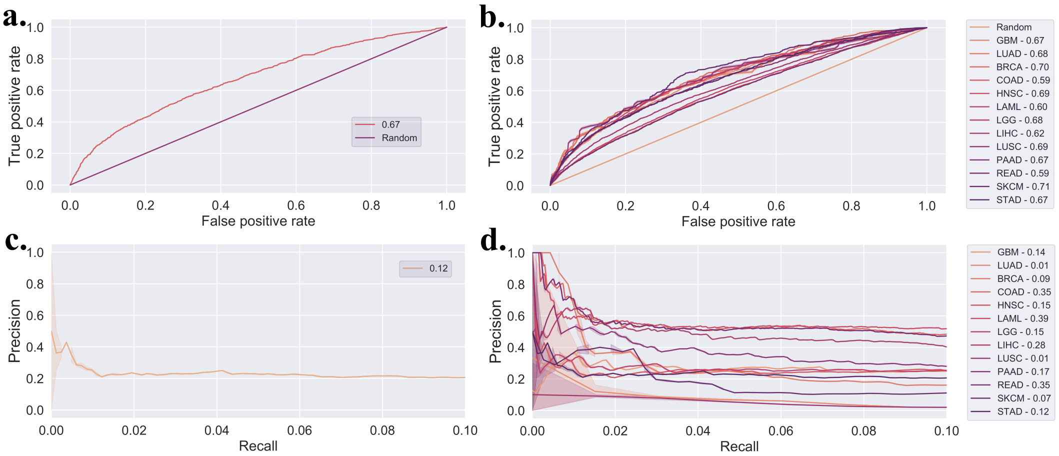


**Figure C.3:** Global validation against CCGD driver genes of our predicted associations between genes and cancer types. Top row gives receiver operator curves for **a.** all predictions and **b.** per cancer type predictions. The bottom row give precision recall curves for **c.** all predictions together and **d.** per cancer type predictions. Precision-recall curve are cut at recall 0.1 to show top ranked precision of predictions in more detail. Each value in the legends corresponds to either AUROC or AUPRC score of the non-restricted associated curve.

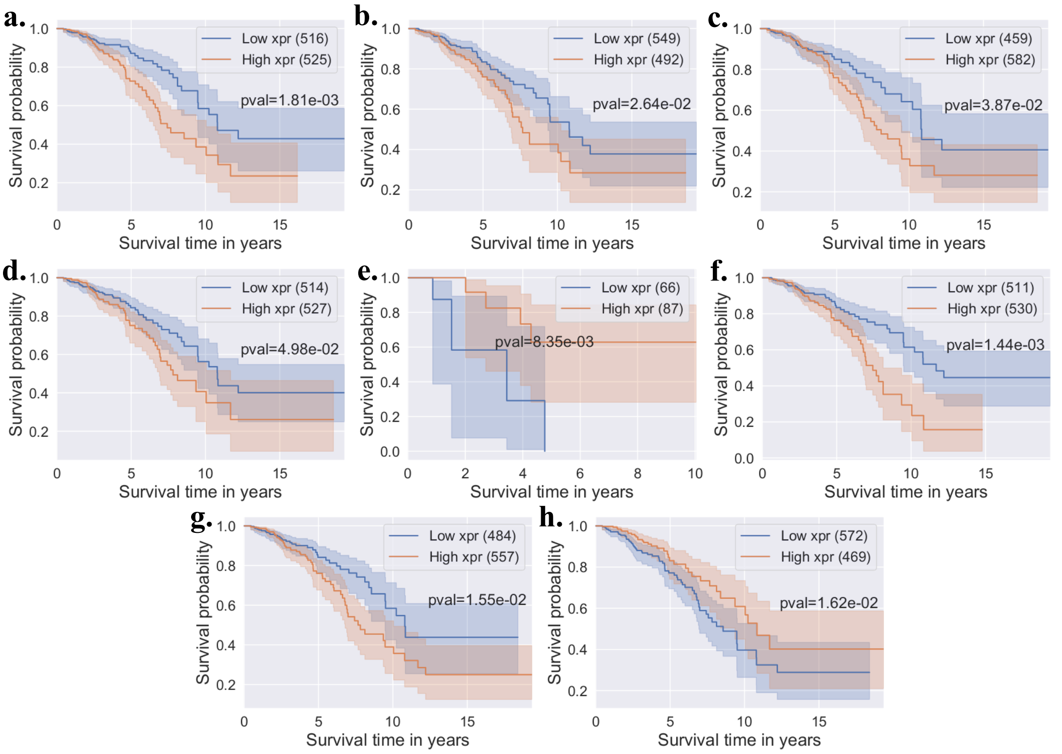


**Figure C.4:** Kaplan-Meir curves comparing patients survival within a given cohort based on the relative expression of a gene. Each panel corresponds to a gene–cancer pair: a. KAT2B–BRCA, b. MDM2–BRCA, c. SP1–BRCA, d. HERC1–BRCA, e. HERC1–READ, f. NCOA3–BRCA, g. CHD6–BRCA, and h. PPARG–BRCA. Numbers in parenthesis in legends correspond to the numbers of patients falling in the associated category.

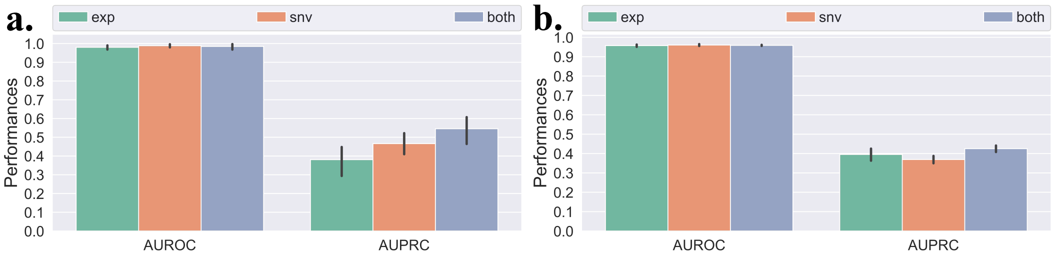


**Figure C.5:** Ablation of patient–gene data effect on performances for **a.** drug–cancer type and **b.** gene–cancer type link prediction tasks.

# Bibliography

[1] C. A. Hidalgo, N. Blumm *et al.* A dynamic network approach for the study of human phenotypes. *PLoS Computational Biology*, 5(4):e1000353, 2009.

[2] I. H. G. S. Consortium *et al.* Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931, 2004.

[3] B. Alberts. *Molecular biology of the cell.* Garland Science, 2017.

[4] A.-L. Barabási, N. Gulbahce, and J. Loscalzo. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68, 2011.

[5] C. J. Vaske, S. C. Benz *et al.* Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using paradigm. *Bioinformatics*, 26(12):i237–i245, 2010.

[6] A. Chatr-Aryamontri, R. Oughtred *et al.* The BioGRID interaction database: 2017 update. *Nucleic Acids Research*, 45(D1):D369–D379, January 2017.

[7] A. Fabregat, K. Sidiropoulos *et al.* The reactome pathway knowledgebase. *Nucleic Acids Research*, 44(D1):D481–D487, January 2016.

[8] V. Gligorijević and N. Pržulj. Methods for biological data integration: perspectives and challenges. *Journal of the Royal Society Interface*, 12(112):20150571, 2015.

[9] H. Ge, A. J. Walhout, and M. Vidal. Integrating “omic” information: a bridge between genomics and systems biology. *Trends in Genetics*, 19(10):551–560, 2003.

[10] X. Ma, T. Chen, and F. Sun. Integrative approaches for predicting protein function and prioritizing genes for complex phenotypes using protein interaction networks. *Briefings in Bioinformatics*, 15(5):685–698, 2014.

[11] O. Collier, V. Stoven, and J.-P. Vert. Lotus: A single-and multitask machine learning algorithm for the prediction of cancer driver genes. *PLoS Computational Biology*, 15(9):e1007381, 2019.

[12] J. Song, W. Peng, and F. Wang. A random walk-based method to identify driver genes by integrating the subcellular localization and variation frequency into bipartite graph. *BMC Bioinformatics*, 20(1):238, 2019.

[13] R. Cao, Z. Zhong, and J. Cheng. Smiss: a protein function prediction server by integrating multiple sources. *International Journal of Computational Intelligence in Bioinformatics and Systems Biology*, 2(1):22–30, 2020.

[14] M. Hecker, S. Lambeck *et al.* Gene regulatory network inference: data integration in dynamic modelsa review. *Biosystems*, 96(1):86–103, 2009.

[15] Y. Guan, D. Gorenshteyn *et al.* Tissue-specific functional networks for prioritizing phenotype and disease genes. *PLoS Computational Biology*, 8(9):e1002694, 2012.

[16] N. Zarayeneh, E. Ko *et al.* Integration of multi-omics data for integrative gene regulatory network inference. *International Journal of Data Mining and Bioinformatics*, 18(3):223, 2017.

[17] N. Zhou, Y. Jiang *et al.* The cafa challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biology*, 20(1):1–23, 2019.

[18] L. J. Lu, Y. Xia *et al.* Assessing the limits of genomic data integration for predicting protein networks. *Genome Research*, 15(7):945–953, 2005.

[19] T. Gross, M. J. Wongchenko *et al.* Robust network inference using response logic. *Bioinformatics*, 35(14):i634–i642, 2019.

[20] D. H. Wolpert and W. G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, 1997.

[21] I. Dagogo-Jack and A. T. Shaw. Tumour heterogeneity and resistance to cancer therapies. *Nature Reviews Clinical Oncology*, 15(2):81, 2018.

[22] R. Au, R. J. Piers, and L. Lancashire. Back to the future: Alzheimer's disease heterogeneity revisited. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 1(3):368, 2015.

[23] D. Wang and J. Gu. Integrative clustering methods of multi-omics data for molecule-based cancer classifications. *Quantitative Biology*, 4(1):58–67, 2016.

[24] E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

[25] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

[26] F. Miao, Y.-P. Cai *et al.* Risk prediction of one-year mortality in patients with cardiac arrhythmias using random survival forest. *Computational and Mathematical Methods in Medicine*, 2015, 2015.

[27] T. Ching, X. Zhu, and L. X. Garmire. Cox-nnet: an artificial neural network method for prognosis prediction of high-throughput omics data. *PLoS Computational Biology*, 14(4):e1006076, 2018.

[28] J. Yao, X. Zhu, and J. Huang. Deep multi-instance learning for survival prediction from whole slide images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 496–504. Springer, 2019.

[29] R. M. Califf. Biomarker definitions and their applications. *Experimental Biology and Medicine*, 243(3):213–221, 2018.

[30] R. C. Mohs and N. H. Greig. Drug discovery and development: Role of basic biological research. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 3(4):651–657, 2017.

[31] O. J. Wouters, M. McKee, and J. Luyten. Estimated research and development investment needed to bring a new medicine to market, 2009-2018. *Jama*, 323(9):844–853, 2020.

[32] S. Pushpakom, F. Iorio *et al.* Drug repurposing: progress, challenges and recommendations. *Nature Reviews Drug Discovery*, 18(1):41–58, 2019.

[33] M. Lotfi Shahreza, N. Ghadiri *et al.* A review of network-based approaches to drug repositioning. *Briefings in Bioinformatics*, 19(5):878–892, 2018.

[34] G. Dissez, G. Ceddia *et al.* Drug repositioning predictions by non-negative matrix tri-factorization of integrated association data. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 25–33, 2019.

[35] M. Yang, H. Luo *et al.* Drug repositioning based on bounded nuclear norm regularization. *Bioinformatics*, 35(14):i455–i463, 2019.

[36] M. Vidal. A unifying view of 21st century systems biology. *FEBS Letters*, 583(24):3891–3894, 2009.

[37] R. Sloutsky, N. Jimenez *et al.* Accounting for noise when clustering biological data. *Briefings in Bioinformatics*, 14(4):423–436, 2013.

[38] L. S. Tsimring. Noise in biology. *Reports on Progress in Physics*, 77(2):026601, 2014.

[39] C. A. Vallejos, J. C. Marioni, and S. Richardson. Basics: Bayesian analysis of single-cell sequencing data. *PLoS Computational Biology*, 11(6):e1004333, 2015.

[40] N. J. Isaac and M. J. Pocock. Bias and information in biological records. *Biological Journal of the Linnean Society*, 115(3):522–531, 2015.

[41] K. Luck, D.-K. Kim *et al.* A reference map of the human binary protein interactome. *Nature*, 580(7803):402–408, 2020.

[42] M. A. Gianfrancesco, S. Tamang *et al.* Potential biases in machine learning algorithms using electronic health record data. *JAMA Internal Medicine*, 178(11):1544–1547, 2018.

[43] A. Gonzalez-Perez, C. Perez-Llamas *et al.* Intogen-mutations identifies cancer drivers across tumor types. *Nature Methods*, 10(11):1081, 2013.

[44] G. Hardiman. Microarray platforms–comparisons and contrasts. *Pharmacogenomics*, 5(5):487–502, 2004.

[45] Z. Wang, M. Gerstein, and M. Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009.

[46] J. Sánchez-Valle, H. Tejero *et al.* Interpreting molecular similarity between patients as a determinant of disease comorbidity relationships. *Nature Communications*, 11(1):1–13, 2020.

[47] S. D. Praktiknjo, B. Obermayer *et al.* Tracing tumorigenesis in a solid tumor model at single-cell resolution. *Nature Communications*, 11(1):1–12, 2020.

[48] O. N. Jensen. Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry. *Current Opinion in Chemical Biology*, 8(1):33–41, 2004.

[49] R. Aebersold and M. Mann. Mass spectrometry-based proteomics. *Nature*, 422(6928):198–207, 2003.

[50] B. Spurrier, S. Ramalingam, and S. Nishizuka. Reverse-phase protein lysate microarrays for cell signaling analysis. *Nature Protocols*, 3(11):1796, 2008.

[51] A. W. Senior, R. Evans *et al.* Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, 2020.

[52] D. Lee, O. Redfern, and C. Orengo. Predicting protein function from sequence and structure. *Nature Reviews Molecular Cell Biology*, 8(12):995–1005, 2007.

[53] N. Vinayavekhin, E. A. Homan, and A. Saghatelian. Exploring disease through metabolomics. *ACS Chemical Biology*, 5(1):91–103, 2010.

[54] A. J. Levine and A. M. Puzio-Kuter. The control of the metabolic switch in cancers by oncogenes and tumor suppressor genes. *Science*, 330(6009):1340–1344, 2010.

[55] M. J. Ziller, H. Gu *et al.* Charting a dynamic dna methylation landscape of the human genome. *Nature*, 500(7463):477, 2013.

[56] C. M. Rivera and B. Ren. Mapping human epigenomes. *Cell*, 155(1):39–55, 2013.

[57] P. J. Park. Chip–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10(10):669, 2009.

[58] M. Frommer, L. E. McDonald *et al.* A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual dna strands. *Proceedings of the National Academy of Sciences*, 89(5):1827–1831, 1992.

[59] J. Zhu, M. Adli *et al.* Genome-wide chromatin state transitions associated with developmental and environmental cues. *Cell*, 152(3):642–654, 2013.

[60] A. Kundaje, W. Meuleman *et al.* Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, 2015.

[61] P. Uetz, L. Giot *et al.* A comprehensive analysis of protein–protein interactions in saccharomyces cerevisiae. *Nature*, 403(6770):623, 2000.

[62] T. Ito, T. Chiba *et al.* A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences*, 98(8):4569–4574, 2001.

[63] U. Stelzl, U. Worm *et al.* A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, 122(6):957–968, 2005.

[64] T. Rolland, M. Taşan *et al.* A proteome-scale map of the human interactome network. *Cell*, 159(5):1212–1226, 2014.

[65] M. Vidal, M. E. Cusick, and A.-L. Barabási. Interactome networks and human disease. *Cell*, 144(6):986–998, 2011.

[66] M. Zitnik, M. W. Feldman *et al.* Evolution of resilience in protein interactomes across the tree of life. *Proceedings of the National Academy of Sciences*, 116(10):4426–4433, 2019.

[67] C. Brun, F. Chevenet *et al.* Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. *Genome Biology*, 5(1):R6, 2003.

[68] M. P. Joy, A. Brock *et al.* High-betweenness proteins in the yeast protein interaction network. *BioMed Research International*, 2005(2):96–103, 2005.

[69] F. Emmert-Streib, S. Tripathi *et al.* The human disease network. *Systems Biomedicine*, 1(1):20–28, 2013.

[70] L. Mabonga and A. P. Kappo. Protein-protein interaction modulators: advances, successes and remaining challenges. *Biophysical Reviews*, pages 1–23, 2019.

[71] A. Blais and B. D. Dynlacht. Constructing transcriptional regulatory networks. *Genes & Development*, 19(13):1499–1511, 2005.

[72] K. Yamaguchi-Shinozaki and K. Shinozaki. Transcriptional regulatory networks in cellular responses and tolerance to dehydration and cold stresses. *Annual Review of Plant Biology*, 57(1):781–803, 2006.

[73] T. I. Lee. Transcriptional regulatory networks in saccharomyces cerevisiae. *Science*, 298(5594):799–804, 2002.

[74] H. Yu, P. M. Kim *et al.* The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Computational Biology*, 3(4):e59, 2007.

[75] A. Aytes, A. Mitrofanova *et al.* Cross-species regulatory network analysis identifies a synergistic interaction between foxm1 and cenpf that drives prostate cancer malignancy. *Cancer Cell*, 25(5):638–651, 2014.

[76] J. Huang, Z. Sun *et al.* Identification of microrna as sepsis biomarker based on mirnas regulatory network analysis. *BioMed Research International*, 2014, 2014.

[77] C. L. Plaisier, S. OBrien *et al.* Causal mechanistic regulatory network for glioblastoma deciphered using systems genetics network analysis. *Cell Systems*, 3(2):172–186, 2016.

[78] H. Jeong, B. Tombor *et al.* The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, October 2000.

[79] H. Ma and A. P. Zeng. Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics*, 19(2):270–277, 2003.

[80] H. W. Ma, X. M. Zhao *et al.* Decomposition of metabolic network into functional modules based on the global connectivity structure of reaction graph. *Bioinformatics*, 20(12):1870–1876, 2004.

[81] K. Hiller and C. M. Metallo. Profiling metabolic networks to study cancer metabolism. *Current Opinion in Biotechnology*, 24(1):60–68, 2013.

[82] T. Eckert, C. Tang, and D. Eidelberg. Assessment of the progression of parkinson's disease: a metabolic network approach. *The Lancet Neurology*, 6(10):926–932, 2007.

[83] J. M. Stuart. A Gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302(5643):249–255, 2003.

[84] T. Obayashi, Y. Kagaya *et al.* Coxpresdb v7: a gene coexpression database for 11 animal species supported by 23 coexpression platforms for technical evaluation and evolutionary inference. *Nucleic Acids Research*, 47(D1):D55–D62, 2018.

[85] L. Mao, J. L. Van Hemert *et al.* Arabidopsis gene co-expression network and its functional modules. *BMC Bioinformatics*, 10(1):346, 2009.

[86] Q. Liao, C. Liu *et al.* Large-scale prediction of long non-coding rna functions in a coding–non-coding gene co-expression network. *Nucleic Acids Research*, 39(9):3864–3878, 2011.

[87] Y. Yang, L. Han *et al.* Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nature Communications*, 5(1):1–9, 2014.

[88] M. Kanehisa, M; Goto, S; Furumichi, M; Tanabe, M; Hirakawa. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Researh*, 38(Database issue):355–360, 2010.

[89] C. Hertz-Fowler. GeneDB: a resource for prokaryotic and eukaryotic organisms. *Nucleic Acids Research*, 32(90001):339D–343, 2004.

[90] D. Weaver, S. Gama-Castro *et al.* The EcoCyc database. *EcoSal Plus*, 6(1), 2014.

[91] P. Romero and P. Karp. PseudoCyc, a pathway-genome database for pseudomonas aeruginosa. *Journal of Molecular Microbiology and Biotechnology*, 5(4):230–239, 2003.

[92] R. Caspi, R. Billington *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Research*, 44(D1):D471–D480, 2016.

[93] J. W. Whitaker, I. Letunic *et al.* metaTIGER: A metabolic evolution resource. *Nucleic Acids Research*, 37(SUPPL. 1):D531–8, 2009.

[94] R. Overbeek, N. Larsen *et al.* The ErgoTM genome analysis and discovery system, 2003.

[95] E. A. Ananko, N. L. Podkolodny *et al.* GeneNet: a database on structure and functional organisation of gene networks. *Nucleic Acids Research*, 30(1):398–401, 2002.

[96] A. Fabregat, K. Sidiropoulos *et al.* The reactome pathway knowledgebase. *Nucleic Acids Research*, 44(D1):D481–D487, 2016.

[97] S. Gama-Castro, H. Salgado *et al.* RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (gensor units). *Nucleic Acids Research*, 39(SUPPL. 1):D98–105, 2011.

[98] A. Sandelin. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research*, 32(90001):91D–94, 2004.

[99] F. Diella, S. Cameron *et al.* Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins. *BMC Bioinformatics*, 5(1):79, 2004.

[100] M. L. Miller, L. J. Jensen *et al.* Linear motif atlas for phosphorylation-dependent signaling. *Science Signaling*, 1(35):ra2–ra2, 2008.

[101] F. Gnad, S. Ren *et al.* PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biology*, 8(11):R250, 2007.

[102] C. Choi, M. Krull *et al.* TRANSPATH – A high quality database focused on signal transduction. In *Comparative and Functional Genomics*, volume 5, pages 163–168, 2004.

[103] A. Frolkis, C. Knox *et al.* SMPDB: the small molecule pathway database. *Nucleic Acids Research*, 38(SUPPL.1):D480–D487, 2009.

[104] T. S. Keshava Prasad, R. Goel *et al.* Human protein reference database - 2009 update. *Nucleic Acids Research*, 37(SUPPL. 1):D767–72, 2009.

[105] J. M. Cherry, E. L. Hong *et al.* Saccharomyces genome database: the genomics resource of budding yeast. *Nucleic Acids Research*, 40(D1):D700–5, 2012.

[106] S. Kerrien, Y. Alam-Faruque *et al.* IntAct – open source resource for molecular interaction data. *Nucleic Acids Research*, 35(SUPPL. 1):D561–D565, 2007.

[107] K. Han, B. Park *et al.* HPID: the human protein interaction database. *Bioinformatics*, 20(15):2466–2470, 2004.

[108] J. Yu, S. Pacifico *et al.* DroID: the drosophila interactions database, a comprehensive resource for annotated gene and protein interactions. *BMC Genomics*, 9(1):461, 2008.

[109] H. W. Mewes, D. Frishman *et al.* MIPS: a database for genomes and protein sequences. *Nucleic Acids Research*, 30(1):31–4, 2002.

[110] I. Xenarios. DIP: the database of interacting proteins. *Nucleic Acids Research*, 28(1):289–291, 2000.

[111] L. Licata, L. Briganti *et al.* MINT, the molecular interaction database: 2012 update. *Nucleic Acids Research*, 40(D1):2006–2008, 2012.

[112] K. R. Brown and I. Jurisica. Online predicted human interaction database. *Bioinformatics*, 21(9):2076–2082, 2005.

[113] D. Szklarczyk, A. Franceschini *et al.* STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, 43(D1):D447–D452, 2015.

[114] T. Obayashi and K. Kinoshita. COXPRESdb: a database to compare gene coexpression in seven model animals. *Nucleic Acids Research*, 39(SUPPL. 1):D1016–D1022, 2011.

[115] S. van Dam, T. Craig, and J. P. de Magalhães. GeneFriends: a human RNA-seq-based gene and transcript co-expression database. *Nucleic Acids Research*, 43(Database issue):D1124–D1132, 2015.

[116] T. Gaudelet and N. Pržulj. Introduction to graph and network theory. In N. Pržulj, editor, *Analysing network data in biology and medicine*, chapter 3, pages 111–150. Cambridge University Press, 2019.

[117] J. A. Bondy and U. S. R. Murty. *Graph theory with applications*, volume 290. London: Macmillan, 1976.

[118] M. Kivelä, A. Arenas *et al.* Multilayer networks. *Journal of Complex Networks*, 2(3):203–271, 2014.

[119] S. Boccaletti, G. Bianconi *et al.* The structure and dynamics of multilayer networks. *Physics Reports*, 544(1):1–122, 2014.

[120] C. Berge. *Graphs and Hypergraphs*, volume 6. Amsterdam: North-Holland publishing company, 1973.

[121] S. Klamt, U.-U. Haus, and F. Theis. Hypergraphs and cellular networks. *PLoS Computational Biology*, 5(5):e1000385, 2009.

[122] P. H. Guzzi and T. Milenković. Survey of local and global biological network alignment: the need to reconcile the two sides of the same coin. *Briefings in Bioinformatics*, 19(3):472–481, 2018.

[123] R. M. Karp. Reducibility among combinatorial problems. *50 Years of Integer Programming 1958-2008: From the Early Years to the State-of-the-Art*, pages 219–241, 2010.

[124] M. R. Garey and D. S. Johnson. *Computers and intractability*, volume 29 of *A Series of Books in the Mathematical Sciences*. New York: wh freeman, San Francisco, Calif., 2002.

[125] A. Ben-Dor, R. Shamir, and Z. Yakhini. Clustering gene expression patterns. *Journal of Computational Biology*, 6(3-4):281–297, 1999.

[126] Z. Shi, C. K. Derow, and B. Zhang. Co-expression module analysis reveals biological processes, genomic gain, and regulatory mechanisms associated with breast cancer progression. *BMC Systems Biology*, 4(1):74, 2010.

[127] E. J. Gardiner, P. J. Artymiuk, and P. Willett. Clique-detection algorithms for matching three-dimensional molecular structures. *Journal of Molecular Graphics and Modelling*, 15(4):245–253, 1997.

[128] R. Samudrala and J. Moult. A graph-theoretic algorithm for comparative modeling of protein structure. *Journal of Molecular Biology*, 279(1):287–302, 1998.

[129] S. A. Cook. The complexity of theorem-proving procedures. In *Proceedings of the Third Annual ACM Symposium on Theory of Computing*, pages 151–158. ACM, 1971.

[130] S. R. Paladugu, S. Zhao *et al.* Mining protein networks for synthetic genetic interactions. *BMC Bioinformatics*, 9(1):426, 2008.

[131] M. Newman. *Networks: An Introduction.* Oxford University Press Inc., New York, 2010.

[132] N. Pržulj, D. G. Corneil, and I. Jurisica. Modeling interactome: Scale-free or geometric? *Bioinformatics*, 20(18):3508–3515, December 2004.

[133] N. Pržulj. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2):e177–e183, 2007.

[134] T. Milenkovic and N. Pržulj. Uncovering Biological Network Function via Graphlet Degree Signatures. *Cancer Informatics*, pages 257–273, 2008.

[135] Ö. N. Yaveroğlu, N. Malod-Dognin *et al.* Revealing the hidden language of complex networks. *Scientific Reports*, 4:4547, 2014.

[136] D. Davis, Ö. N. Yaveroğlu *et al.* Topology-function conservation in protein–protein interaction networks. *Bioinformatics*, 31(10):1632–1639, 2015.

[137] O. Kuchaiev, T. Milenković *et al.* Topological network alignment uncovers biological function and phylogeny. *Journal of the Royal Society Interface*, page rsif20100063, 2010.

[138] N. Malod-Dognin and N. Pržulj. L-GRAAL: Lagrangian graphlet-based network aligner. *Bioinformatics*, 31(13):2182–2189, July 2015.

[139] A. Sarajlić, N. Malod-Dognin *et al.* Graphlet-based characterization of directed networks. *Scientific Reports*, 6(October):35098, 2016.

[140] N. Malod-Dognin and N. Pržulj. Gr-align: fast and flexible alignment of protein 3d structures using graphlet degree similarity. *Bioinformatics*, 30(9):1259–1265, 2014.

[141] N. Malod-Dognin and N. Pržulj. Functional geometry of protein interactomes. *Bioinformatics*, 35(19):3727–3734, 2019.

[142] B. Perozzi, R. Al-Rfou, and S. Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 701–710. ACM, 2014.

[143] A. Grover and J. Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 855–864. ACM, 2016.

[144] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in Neural Information Processing Systems*, pages 585–591, 2002.

[145] M. Ou, P. Cui *et al.* Asymmetric transitivity preserving graph embedding. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1105–1114. ACM, 2016.

[146] K. Borgwardt, E. Ghisu *et al.* Graph kernels: State-of-the-art and future challenges. *arXiv preprint arXiv:2011.03854*, 2020.

[147] W. L. Hamilton, R. Ying, and J. Leskovec. Representation learning on graphs: Methods and applications. *arXiv preprint arXiv:1709.05584*, 2017.

[148] D. Silver, A. Huang *et al.* Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[149] M. D. Ritchie, E. R. Holzinger *et al.* Methods of integrating data to uncover genotype–phenotype interactions. *Nature Reviews Genetics*, 16(2):85, 2015.

[150] M. Bersanelli, E. Mosca *et al.* Methods for the integration of multi-omics data: mathematical aspects. *BMC Bioinformatics*, 17(2):S15, 2016.

[151] Y. Li, F.-X. Wu, and A. Ngom. A review on machine learning principles for multi-view biological data integration. *Briefings in Bioinformatics*, page bbw113, 2016.

[152] G. R. Lanckriet, T. De Bie *et al.* A statistical framework for genomic data fusion. *Bioinformatics*, 20(16):2626–2635, 2004.

[153] M. Žitnik and B. Zupan. Data fusion by matrix factorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(1):41–53, 2015.

[154] P. Pavlidis, J. Weston *et al.* Learning gene functional classifications from multiple data types. *Journal of Computational Biology*, 9(2):401–411, 2002.

[155] M. H. Van Vliet, H. M. Horlings *et al.* Integration of clinical and gene expression data has a synergetic effect on predicting breast cancer outcome. *PloS One*, 7(7):e40358, 2012.

[156] N. Malod-Dognin, J. Petschnigg *et al.* Towards a data-integrated cell. *Nature Communications*, 10(1):1–13, 2019.

[157] V. Gligorijević, N. Malod-Dognin, and N. Pržulj. Patient-specific data fusion for cancer stratification and personalised treatment. In *Biocomputing 2016: Proceedings of the Pacific Symposium*, pages 321–332. World Scientific, 2016.

[158] O. Gevaert, F. D. Smet *et al.* Predicting the prognosis of breast cancer by integrating clinical and microarray data with bayesian networks. *Bioinformatics*, 22(14):e184–e190, 2006.

[159] J. Zhu, B. Zhang *et al.* Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nature Genetics*, 40(7):854, 2008.

[160] D. M. Chickering. Learning bayesian networks is np-complete. In *Learning from Data*, pages 121–130. Springer, 1996.

[161] N. Shervashidze, S. Vishwanathan *et al.* Efficient graphlet kernels for large graph comparison. In *Artificial Intelligence and Statistics*, pages 488–495, 2009.

[162] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[163] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, pages 849–856, 2002.

[164] M. Gönen and E. Alpaydın. Multiple kernel learning algorithms. *Journal of Machine Learning Research*, 12(Jul):2211–2268, 2011.

[165] F. He, G. Zhu *et al.* Pcid: A novel approach for predicting disease comorbidity by integrating multi-scale data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 14(3):678–686, 2017.

[166] K. Sun, N. Buchan *et al.* The integrated disease network. *Integrative Biology*, 6(11):1069–1079, 2014.

[167] J. Dutkowski, M. Kramer *et al.* A gene ontology inferred from molecular networks. *Nature Biotechnology*, 31(1):38, 2013.

[168] S. Mostafavi, D. Ray *et al.* Genemania: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biology*, 9(1):S4, 2008.

[169] B. Wang, A. M. Mezlini *et al.* Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods*, 11(3):333, 2014.

[170] Y.-F. Huang, H.-Y. Yeh, and V.-W. Soo. Inferring drug-disease associations from integration of chemical, genomic and phenotype data using network propagation. *BMC Medical Genomics*, 6(3):S4, 2013.

[171] H. Luo, J. Wang *et al.* Drug repositioning based on comprehensive similarity measures and bi-random walk algorithm. *Bioinformatics*, 32(17):2664–2671, 2016.

[172] C. Ruiz, M. Zitnik, and J. Leskovec. Discovery of disease treatment mechanisms through the multiscale interactome. *bioRxiv*, 2020.

[173] K. Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[174] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788, 1999.

[175] C. Ding, T. Li *et al.* Orthogonal nonnegative matrix t-factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 126–135. ACM, 2006.

[176] R. Shen, A. B. Olshen, and M. Ladanyi. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, 25(22):2906–2912, 2009.

[177] J. Liu, C. Wang *et al.* Multi-view clustering via joint nonnegative matrix factorization. In *Proceedings of the 2013 SIAM International Conference on Data Mining*, pages 252–260. SIAM, 2013.

[178] W. W. Chin. The partial least squares approach to structural equation modeling. *Modern Methods for Business Research*, 295(2):295–336, 1998.

[179] M. Hofree, J. P. Shen *et al.* Network-based stratification of tumor mutations. *Nature Methods*, 10(11):1108, 2013.

[180] D. Liu, J. Davila-Velderrain *et al.* Integrative construction of regulatory region networks in 127 human reference epigenomes by matrix factorization. *Nucleic Acids Research*, 47(14):7235–7246, 2019.

[181] K. He, X. Zhang *et al.* Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[182] A. Vaswani, N. Shazeer *et al.* Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[183] I. Goodfellow, Y. Bengio *et al.* *Deep learning*, volume 1. MIT press Cambridge, 2016.

[184] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989.

[185] T. Ching, D. S. Himmelstein *et al.* Opportunities and obstacles for deep learning in biology and medicine. *bioRxiv*, page 142760, 2018.

[186] V. Gligorijević, M. Barot, and R. Bonneau. deepnf: Deep network fusion for protein function prediction. *bioRxiv*, page 223339, 2017.

[187] J. Fan and J. Cheng. Matrix completion by deep matrix factorization. *Neural Networks*, 98:34–41, 2018.

[188] J. Schreiber, T. Durham *et al.* Multi-scale deep tensor factorization learns a latent representation of the human epigenome. *bioRxiv*, page 364976, 2019.

[189] T. Dettmers, P. Minervini *et al.* Convolutional 2d knowledge graph embeddings. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[190] K. Y. Michael, J. Ma *et al.* Visible machine learning for biomedicine. *Cell*, 173(7):1562–1565, 2018.

[191] J. Ma, M. K. Yu *et al.* Using deep learning to model the hierarchical structure and function of a cell. *Nature Methods*, 15(4):290, 2018.

[192] J. A. Blake, K. R. Christie *et al.* Gene ontology consortium: Going forward. *Nucleic Acids Research*, 43(D1):D1049–D1056, January 2015.

[193] J. Yao, X. Zhu *et al.* Deep correlational learning for survival prediction from multi-modality data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 406–414. Springer, 2017.

[194] J. Hao, S. C. Kosaraju *et al.* Page-net: Interpretable and integrative deep learning for survival analysis using histopathological images and genomic data. In *Pacific Symposium on Biocomputing*, volume 25, pages 355–366. World Scientific, 2020.

[195] T. Gaudelet, N. Malod-Dognin, and N. Pržulj. Higher-order molecular organization as a source of biological function. *Bioinformatics*, 34(17):i944–i953, 2018.

[196] V. Lacroix, L. Cottret *et al.* An introduction to metabolic networks and their structural analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 5(4):594–617, October 2008.

[197] E. Estrada and J. A. Rodríguez-Velázquez. Subgraph centrality and clustering in complex hyper-networks. *Physica A: Statistical Mechanics and its Applications*, 364:581–594, 2006.

[198] M. Latapy, C. Magnien, and N. D. Vecchio. Basic notions for the analysis of large two-mode networks. *Social Networks*, 30(1):31–48, 2008.

[199] N. Pearcy, J. J. Crofts, and N. Chuzhanova. Hypergraph Models of Metabolism. *International Journal of Biological, Biomolecular, Agricultural, Food and Biotechnological Engineering*, 8(8):19–23, 2014.

[200] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical Correlation Analysis: An Overview with Application to Learning Methods. *Neural Computation*, 16(12):2639–2664, December 2004.

[201] A. Ruepp, B. Brauner *et al.* CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Research*, 36(Database):D646, January 2007.

[202] A. Ruepp, B. Waegele *et al.* CORUM: The comprehensive resource of mammalian protein complexes-2009. *Nucleic Acids Research*, 38(SUPPL.1):D497–D501, January 2009.

[203] S. Pu, J. Wong *et al.* Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Research*, 37(3):825–831, February 2009.

[204] J. Lugo-Martinez and P. Radivojac. Classification in biological networks with hypergraphlet kernels. *arXiv:1703.04823*, 2017.

[205] S. Agarwal, K. Branson, and S. Belongie. Higher order learning with graphs. In *Proceedings of the 23rd International Conference on Machine learning*, pages 17–24. ACM, 2006.

[206] J. R. Munkres. *Elements of algebraic topology*. CRC Press, 2018.

[207] J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

[208] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[209] N. X. Vinh, J. Epps, and J. Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11:2837–2854, 2010.

[210] P. Jaccard. The distribution of the flora in the alpine zone. *New Phytologist*, 11(2):37–50, 1912.

[211] A. Vazquez, A. Flammini *et al.* Global protein function prediction from protein-protein interaction networks. *Nature Biotechnology*, 21(6):697, 2003.

[212] D. Binns, E. Dimmer *et al.* Quickgo: a web-based tool for gene ontology searching. *Bioinformatics*, 25(22):3045–3046, 2009.

[213] K. M. Pumiglia and S. J. Decker. Cell cycle arrest mediated by the mek/mitogen-activated protein kinase pathway. *Proceedings of the National Academy of Sciences*, 94(2):448–452, 1997.

[214] H. D. Madhani and G. R. Fink. The control of filamentous differentiation and virulence in fungi. *Trends in Cell Biology*, 8(9):348–353, 1998.

[215] M. C. Gustin, J. Albertyn *et al.* Map kinase pathways in the yeast-saccharomyces cerevisiae. *Microbiology and Molecular Biology Reviews*, 62(4):1264–1300, 1998.

[216] C. Díaz-Jullien, A. Pérez-Estévez *et al.* Prothymosin $\alpha$ binds histones in vitro and shows activity in nucleosome assembly assay. *Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology*, 1296(2):219–227, 1996.

[217] J. Jokinen, E. Dadu *et al.* Integrin-mediated cell adhesion to type i collagen fibrils. *Journal of Biological Chemistry*, 279(30):31956–31963, 2004.

[218] S. Testaz and J.-L. Duband. Central role of the $\alpha 4\beta 1$ integrin in the coordination of avian truncal neural crest cell adhesion, migration, and survival. *Developmental Dynamics*, 222(2):127–140, 2001.

[219] N. Wong and X. Wang. mirdb: an online resource for microrna target prediction and functional annotations. *Nucleic Acids Research*, 43(D1):D146–D152, 2014.

[220] C. King, G. Rios *et al.* Udp-glucuronosyltransferases. *Current Drug Metabolism*, 1(2):143–161, 2000.

[221] T. Gaudelet, N. Malod-Dognin *et al.* Unveiling new disease, pathway, and gene associations via multi-scale neural network. *PloS One*, 15(4):e0231059, 2020.

[222] V. Emilsson, G. Thorleifsson *et al.* Genetics of gene expression and its effect on disease. *Nature*, 452(7186):423, 2008.

[223] J. T. Dudley, R. Tibshirani *et al.* Disease signatures are robust across tissues and experiments. *Molecular Systems Biology*, 5(1):307, 2009.

[224] A. B. Jensen, P. L. Moseley *et al.* Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nature Communications*, 5:4022, 2014.

[225] K.-I. Goh, M. E. Cusick *et al.* The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690, 2007.

[226] D.-S. Lee, J. Park *et al.* The implications of human metabolic network topology for disease comorbidity. *Proceedings of the National Academy of Sciences*, 2008.

[227] T. Abeel, T. Helleputte *et al.* Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics*, 26(3):392–398, 2009.

[228] Z. He and W. Yu. Stable feature selection for biomarker discovery. *Computational Biology and Chemistry*, 34(4):215–225, 2010.

[229] J. Zhao, T.-H. Yang *et al.* Ranking candidate disease genes from gene expression and protein interaction: a katz-centrality based approach. *PLoS One*, 6(9):e24306, 2011.

[230] G. Hong, H. Li *et al.* Identifying disease-associated pathways in one-phenotype data based on reversal gene expression orderings. *Scientific Reports*, 7(1):1348, 2017.

[231] J. P. Cogswell, J. Ward *et al.* Identification of mirna changes in alzheimer's disease brain and csf yields putative biomarkers and insights into disease pathways. *Journal of Alzheimer's Disease*, 14(1):27–41, 2008.

[232] V. Kannan, F. Swartz *et al.* Conditional disease development extracted from longitudinal health care cohort data using layered network construction. *Scientific Reports*, 6:26170, 2016.

[233] A. Schulze and J. Downward. Navigating gene expression using microarraysa technology review. *Nature Cell Biology*, 3(8):E190, 2001.

[234] T. Barrett, S. E. Wilhite *et al.* Ncbi geo: archive for functional genomics data setsupdate. *Nucleic Acids Research*, 41(D1):D991–D995, 2013.

[235] N. Kolesnikov, E. Hastings *et al.* Arrayexpress updatesimplifying data submissions. *Nucleic Acids Research*, 43(D1):D1113–D1116, 2015.

[236] M. N. McCall, B. M. Bolstad, and R. A. Irizarry. Frozen robust multiarray analysis (frma). *Biostatistics*, 11(2):242–253, 2010.

[237] G. K. Smyth. Limma: linear models for microarray data. In *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, pages 397–420. Springer, 2005.

[238] L. M. Schriml, C. Arze *et al.* Disease ontology: a backbone for disease semantic integration. *Nucleic Acids Research*, 40(D1):D940–D946, 2011.

[239] V. A. McKusick. *Mendelian inheritance in man: a catalog of human genes and genetic disorders*, volume 1. JHU Press, 1998.

[240] A. Fabregat, S. Jupe *et al.* The reactome pathway knowledgebase. *Nucleic Acids Research*, 46(D1):D649–D655, 2018.

[241] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[242] K. He, X. Zhang *et al.* Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1026–1034, 2015.

[243] M. Abadi, P. Barham *et al.* Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pages 265–283, Savannah, GA, 2016. USENIX Association.

[244] J. M. Zurada, A. Malinowski, and I. Cloete. Sensitivity analysis for minimization of input data dimension for feedforward neural network. In *Proceedings of IEEE International Symposium on Circuits and Systems-ISCAS'94*, volume 6, pages 447–450. IEEE, 1994.

[245] J. Piñero, A. Bravo *et al.* Disgenet: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Research*, 45(D1):D833–D839, 2017.

[246] F. Pedregosa, G. Varoquaux *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[247] J. Wang, M. A. Maldonado *et al.* The ubiquitin-proteasome system and its role in inflammatory and autoimmune diseases. *Cellular and Molecular Immunology*, 3(4):255–261, 2006.

[248] D. Bellizzi, S. Dato *et al.* Characterization of a bidirectional promoter shared between two human genes related to aging: Sirt3 and psmd13. *Genomics*, 89(1):143–150, 2007.

[249] E. Koltai, Z. Bori *et al.* Master athletes have higher mir-7, sirt3 and sod2 expression in skeletal muscle than age-matched sedentary controls. *Redox Biology*, 19:46–51, 2018.

[250] A. P. Davis, C. J. Grondin *et al.* The comparative toxicogenomics database: update 2019. *Nucleic Acids Research*, page gky868, 2018.

[251] N. Hashimoto, K. Kagitani-Shimono *et al.* Slc2a1 gene analysis of japanese patients with glucose transporter 1 deficiency syndrome. *Journal of Human Genetics*, 56(12):846, 2011.

[252] B. van der Zwaag, L. Franke *et al.* Gene-network analysis identifies susceptibility genes related to glycobiology in autism. *PloS One*, 4(5):e5324, 2009.

[253] E. F. Torrey and R. H. Yolken. Schizophrenia and infections: the eyes have it. *Schizophrenia Bulletin*, 43(2):247–252, 2017.

[254] C. Selmer, J. B. Olesen *et al.* The spectrum of thyroid disease and risk of new onset atrial fibrillation: a large population cohort study. *British Medical Journal*, 345:e7895, 2012.

[255] A. M. Dahir and S. F. Thomsen. Comorbidities in vitiligo: comprehensive review. *International Journal of Dermatology*, 57(10):1157–1164, 2018.

[256] M. Proietti, V. Raparelli *et al.* Adverse outcomes in patients with atrial fibrillation and peripheral arterial disease: a report from the eurobservational research programme pilot survey on atrial fibrillation. *Ep Europace*, 19(9):1439–1448, 2017.

[257] K. Machida, C.-L. Chen *et al.* Cancer stem cells generated by alcohol, diabetes, and hepatitis c virus. *Journal of Gastroenterology and Hepatology*, 27:19–22, 2012.

[258] J. A. Crasto, M. S. Fourman *et al.* Disulfiram reduces metastatic osteosarcoma tumor burden in an immunocompetent balb/c or-thotopic mouse model. *Oncotarget*, 9(53):30163, 2018.

[259] P. C. Burger, I. Yu *et al.* Atypical teratoid/rhabdoid tumor of the central nervous system: a highly malignant tumor of infancy and childhood frequently mistaken for medulloblastoma: a pediatric oncology group study. *The American Journal of Surgical Pathology*, 22(9):1083–1092, 1998.

[260] M. Arunachalam, F. Dragoni *et al.* Non-segmental vitiligo and psoriasis comorbidity–a case-control study in italian patients. *Journal of the European Academy of Dermatology and Venereology*, 28(4):433–437, 2014.

[261] O. Sogabe and T. Ohya. Right ventricular failure due to primary right ventricle osteosarcoma. *General Thoracic and Cardiovascular Surgery*, 55(1):19–22, 2007.

[262] T. Dohi, H. Ohmura *et al.* Primary right atrial cardiac osteosarcoma with congestive heart failure. *European Journal of Cardio-thoracic Surgery*, 35(3):544–546, 2009.

[263] L. Giannitrapani, M. Soresi *et al.* Progressive visceral leishmaniasis misdiagnosed as cirrhosis of the liver: a case report. *Journal of Medical Case Reports*, 3(1):7265, 2009.

[264] P. Dahlqvist, R. Spencer *et al.* Pseudoacromegaly: a differential diagnostic problem for acromegaly with a genetic solution. *Journal of the Endocrine Society*, 1(8):1104–1109, 2017.

[265] A. Makis, S. Polychronopoulou, and S. Haidas. Osteosarcoma as a second tumor after treatment for primary non-hodgkin's lymphoma in a child with ataxia-telangiectasia: presentation of a case and review of possible pathogenetic mechanisms. *Journal of Pediatric Hematology/Oncology*, 26(7):444–446, 2004.

[266] Z. Sun, T. Zhang *et al.* mir-202 suppresses proliferation and induces apoptosis of osteosarcoma cells by downregulating gli2. *Molecular and Cellular Biochemistry*, 397(1-2):277–283, 2014.

[267] T. Gaudelet, N. Malod-Dognin, and N. Przulj. Integrative data analytic framework to enhance cancer precision medicine. *arXiv preprint arXiv:2007.01107*, 2020.

[268] International Agency for Research on Cancer *et al.* Latest global cancer data: cancer burden rises to 18.1 million new cases and 9.6 million cancer deaths in 2018. press release, 2018.

[269] T. Shen, N. C. Yeat *et al.* Clinical applications of next generation sequencing in cancer: from panels, to exomes, to genomes. *Frontiers in Genetics*, 6:215, 2015.

[270] A. Roos and S. A. Byron. Genomics-enabled precision medicine for cancer. In *Precision Medicine in Cancer Therapy*, pages 137–169. Springer, 2019.

[271] K. Tomczak, P. Czerwińska, and M. Wiznerowicz. The cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary Oncology*, 19(1A):A68, 2015.

[272] International Cancer Genome Consortium. International network of cancer genome projects. *Nature*, 464(7291):993, 2010.

[273] S. Nik-Zainal, H. Davies *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, 534(7605):47–54, 2016.

[274] P. Campbell, G. Getz *et al.* Pan-cancer analysis of whole genomes. *Nature*, 578(7793):82, 2020.

[275] N. Andor, T. A. Graham *et al.* Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nature Medicine*, 22(1):105, 2016.

[276] E. A. Ashley. Towards precision medicine. *Nature Reviews Genetics*, 17(9):507, 2016.

[277] M. D. Leiserson, F. Vandin *et al.* Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature Genetics*, 47(2):106, 2015.

[278] H. Kitano. Computational systems biology. *Nature*, 420(6912):206–210, 2002.

[279] W. Nelson, M. Zitnik *et al.* To embed or not: network embedding as a paradigm in computational biology. *Frontiers in Genetics*, 10, 2019.

[280] Y.-X. Wang and Y.-J. Zhang. Nonnegative matrix factorization: A comprehensive review. *IEEE Transactions on Knowledge and Data Engineering*, 25(6):1336–1353, 2012.

[281] R. Mehta and K. Rana. A review on matrix factorization techniques in recommender systems. In *2017 2nd International Conference on Communication*

*Systems, Computing and IT Applications (CSCITA)*, pages 269–274. IEEE, 2017.

[282] P. Guyot, E.-H. Djermoune, and T. Bastogne. Assessment of non-negative matrix factorization for the preprocessing of long-term ecg. In *Annual Meeting of Safety Pharmacology Society, SPS 2018*, 2018.

[283] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

[284] Y. Xian, B. Schiele, and Z. Akata. Zero-shot learning-the good, the bad and the ugly. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4582–4591, 2017.

[285] D. S. Wishart, Y. D. Feunang *et al.* Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic Acids Research*, 46(D1):D1074–D1082, 2017.

[286] N. Nikolova and J. Jaworska. Approaches to measure chemical similarity–a review. *QSAR & Combinatorial Science*, 22(9-10):1006–1026, 2003.

[287] M. Ashburner, C. A. Ball *et al.* Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25, 2000.

[288] T. M. Kodinariya and P. R. Makwana. Review on determining number of cluster in k-means clustering. *International Journal*, 1(6):90–95, 2013.

[289] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*, pages 556–562, 2001.

[290] C. Boutsidis and E. Gallopoulos. Svd based initialization: A head start for nonnegative matrix factorization. *Pattern Recognition*, 41(4):1350–1362, 2008.

[291] B. P. Roe, H.-J. Yang, and J. Zhu. Boosted decision trees, a powerful event classifier. In *Statistical Problems in Particle Physics, Astrophysics and Cosmology*, pages 139–142. World Scientific, 2006.

[292] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[293] O. Ursu, J. Holmes *et al.* Drugcentral: online drug compendium. *Nucleic Acids Research*, page gkw993, 2016.

[294] H. Luo, M. Li *et al.* Computational drug repositioning using low-rank matrix approximation and randomized algorithms. *Bioinformatics*, 34(11):1904–1912, 2018.

[295] M. L. Sos, M. Koker *et al.* Pten loss contributes to erlotinib resistance in egfr-mutant lung cancer by activation of akt and egfr. *Cancer Research*, 69(8):3256–3261, 2009.

[296] M. Ruffalo, M. Koyutürk, and R. Sharan. Network-based integration of disparate omic data to identify “silent players” in cancer. *PLoS Computational Biology*, 11(12), 2015.

[297] Z. Sondka, S. Bamford *et al.* The cosmic cancer gene census: describing genetic dysfunction across all human cancers. *Nature Reviews Cancer*, 18(11):696–705, 2018.

[298] K. L. Abbott, E. T. Nyre *et al.* The Candidate Cancer Gene Database: a database of cancer driver genes from forward genetic screens in mice. *Nucleic Acids Research*, 43(Database issue):D844–8, January 2015.

[299] D. Zhao, Y. Mo *et al.* Notch-induced aldehyde dehydrogenase 1a1 deacetylation promotes breast cancer stem cells. *The Journal of Clinical Investigation*, 124(12):5453–5465, 2014.

[300] G. Zhang, W. Zhang *et al.* Microrna-200c and microrna-141 are regulated by a foxp3-kat2b axis and associated with tumor metastasis in breast cancer. *Breast Cancer Research*, 19(1):73, 2017.

[301] E. Bondy-Chorney, A. Denoncourt *et al.* Nonhistone targets of kat2a and kat2b implicated in cancer biology. *Biochemistry and Cell Biology*, 97(1):30–45, 2019.

[302] J. Lukas, D.-Q. Gao *et al.* Alternative and aberrant messenger rna splicing of the mdm2 oncogene in invasive breast cancer. *Cancer Research*, 61(7):3212–3219, 2001.

[303] R. Duan, W. Porter, and S. Safe. Estrogen-induced c-fos protooncogene expression in mcf-7 human breast cancer cells: role of estrogen receptor sp1 complex formation. *Endocrinology*, 139(4):1981–1990, 1998.

[304] X. Wang, W. Peng *et al.* Expression and prognostic value of transcriptional factor sp1 in breast cancer. *Chinese Journal of Cancer*, 26(9):996–1000, 2007.

[305] Y. Wang, X. Cai *et al.* Hbxip up-regulates acsl1 through activating transcriptional factor sp1 in breast cancer. *Biochemical and Biophysical Research Communications*, 484(3):565–571, 2017.

[306] N. Kumari, B. Dwarakanath *et al.* Role of interleukin-6 in cancer progression and therapeutic resistance. *Tumor Biology*, 37(9):11553–11572, 2016.

[307] A. Verma, S. Kambhampati *et al.* Jak family of kinases in cancer. *Cancer and Metastasis Reviews*, 22(4):423–434, 2003.

[308] K. A. Janes. Runx1 and its understudied role in breast cancer. *Cell Cycle*, 10(20):3461–3465, 2011.

[309] C. V. Clevenger. Roles and regulation of stat family transcription factors in human breast cancer. *The American Journal of Pathology*, 165(5):1449–1460, 2004.

[310] J. Y. Fang and B. C. Richardson. The mapk signalling pathways and colorectal cancer. *The Lancet Oncology*, 6(5):322–327, 2005.

[311] I. Herr and K.-M. Debatin. Cellular stress response and apoptosis in cancer therapy. *Blood, The Journal of the American Society of Hematology*, 98(9):2603–2614, 2001.

[312] J. Pflaum, S. Schlosser, and M. Müller. p53 family and cellular stress responses in cancer. *Frontiers in Oncology*, 4:285, 2014.

[313] X. Xiao, W. Wang *et al.* Hsp90aa1-mediated autophagy promotes drug resistance in osteosarcoma. *Journal of Experimental & Clinical Cancer Research*, 37(1):201, 2018.

[314] B. Weigelt, P. Warne, and J. Downward. Pik3ca mutation, but not pten loss of function, determines the sensitivity of breast cancer cells to mtor inhibitory drugs. *Oncogene*, 30(29):3222–3233, 2011.

[315] D. Haber, D. Bell *et al.* Molecular targeted therapy of lung cancer: Egfr mutations and response to egfr inhibitors. In *Cold Spring Harbor Symposia on Quantitative Biology*, volume 70, pages 419–426. Cold Spring Harbor Laboratory Press, 2005.

[316] M. Frattini, P. Saletti *et al.* Pten loss of expression predicts cetuximab efficacy in metastatic colorectal cancer patients. *British Journal of Cancer*, 97(8):1139–1145, 2007.

[317] S. E. Moody, A. C. Schinzel *et al.* Prkaca mediates resistance to her2-targeted therapy in breast cancer cells and restores anti-apoptotic signaling. *Oncogene*, 34(16):2061–2071, 2015.

[318] S. K. Saha, K. Kim *et al.* Cytokeratin 19 (krt19) has a role in the reprogramming of cancer stem cell-like cells to less aggressive and more drug-sensitive cells. *International Journal of Molecular Sciences*, 19(5):1423, 2018.

[319] Z. Gao, X. Xu *et al.* C terminus of clostridium perfringens enterotoxin downregulates cldn4 and sensitizes ovarian cancer cells to taxol and carboplatin. *Clinical Cancer Research*, 17(5):1065–1074, 2011.

[320] R. Hrstka, V. Brychtova *et al.* Agr2 predicts tamoxifen resistance in postmenopausal breast cancer patients. *Disease Markers*, 35, 2013.

[321] L. Xie, Y. Dang *et al.* High krt8 expression independently predicts poor prognosis for lung adenocarcinoma patients. *Genes*, 10(1):36, 2019.

[322] H. Zhang, X. Chen *et al.* Egr1 decreases the malignancy of human non-small cell lung carcinoma by regulating krt18 expression. *Scientific Reports*, 4:5416, 2014.

[323] A. Persidis. Signal transduction as a drug-discovery platform. *Nature Biotechnology*, 16(11):1082–1083, 1998.

[324] K. Sriram and P. A. Insel. G protein-coupled receptors as targets for approved drugs: how many targets and how many drugs? *Molecular Pharmacology*, 93(4):251–258, 2018.

[325] P. A. Netti, D. A. Berk *et al.* Role of extracellular matrix assembly in interstitial transport in solid tumors. *Cancer Research*, 60(9):2497–2503, 2000.

[326] D. M. Gysi, Í. D. Valle *et al.* Network medicine framework for identifying drug repurposing opportunities for covid-19. *arXiv preprint arXiv:2004.07229*, 2020.

[327] K. Hsieh, Y. Wang *et al.* Drug repurposing for covid-19 using graph neural network with genetic, mechanistic, and epidemiological validation. *arXiv preprint arXiv:2009.10931*, 2020.

[328] X. Zeng, X. Song *et al.* Repurpose open data to discover therapeutics for covid-19 using deep learning. *Journal of Proteome Research*, 2020.

[329] M. Kivelä and M. A. Porter. Isomorphisms in multilayer networks. *IEEE Transactions on Network Science and Engineering*, 5(3):198–211, 2017.

[330] T. Dimitrova, K. Petrovski, and L. Kocarev. Graphlets in multiplex networks. *Scientific Reports*, 10(1):1–13, 2020.

[331] X. Chen, Y. Li *et al.* A general framework for estimating graphlet statistics via random walk. *arXiv preprint arXiv:1603.07504*, 2016.

[332] N. K. Ahmed, T. L. Willke, and R. A. Rossi. Estimation of local subgraph counts. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 586–595. IEEE, 2016.

[333] R. A. Rossi, R. Zhou, and N. K. Ahmed. Estimation of graphlet counts in massive networks. *IEEE Transactions on Neural Networks and Learning Systems*, 30(1):44–57, 2018.

[334] L. Wan, M. Zeiler *et al.* Regularization of neural networks using dropconnect. In *International Conference on Machine Learning*, pages 1058–1066, 2013.

[335] A. Rossi, D. Firmani *et al.* Knowledge graph embedding for link prediction: A comparative analysis. *arXiv preprint arXiv:2002.00819*, 2020.

[336] Z. Zhang, P. Cui, and W. Zhu. Deep learning on graphs: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[337] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[338] W. Hamilton, Z. Ying, and J. Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, pages 1024–1034, 2017.

[339] P. Veličković, G. Cucurull *et al.* Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[340] K. Xu, W. Hu *et al.* How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

[341] C. Morris, M. Ritzert *et al.* Weisfeiler and leman go neural: Higher-order graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4602–4609, 2019.

[342] K. Xu, C. Li *et al.* Representation learning on graphs with jumping knowledge networks. *arXiv preprint arXiv:1806.03536*, 2018.

[343] Z. Ying, D. Bourgeois *et al.* Gnnexplainer: Generating explanations for graph neural networks. In *Advances in Neural Information Processing Systems*, pages 9244–9255, 2019.

[344] M. Zitnik, M. Agrawal, and J. Leskovec. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, 34(13):i457–i466, 2018.

[345] M. Schlichtkrull, T. N. Kipf *et al.* Modeling relational data with graph convolutional networks. In *European Semantic Web Conference*, pages 593–607. Springer, 2018.

[346] W. Hu, M. Fey *et al.* Open graph benchmark: Datasets for machine learning on graphs. *arXiv preprint arXiv:2005.00687*, 2020.

[347] Z. Ying, J. You *et al.* Hierarchical graph representation learning with differentiable pooling. In *Advances in Neural Information Processing Systems*, pages 4800–4810, 2018.

[348] J. Klicpera, J. Groß, and S. Günnemann. Directional message passing for molecular graphs. *arXiv preprint arXiv:2003.03123*, 2020.

[349] Z. Chen, L. Chen *et al.* Can graph neural networks count substructures? *arXiv preprint arXiv:2002.04025*, 2020.

[350] N. Dehmamy, A.-L. Barabási, and R. Yu. Understanding the representation power of graph neural networks in learning graph topology. In *Advances in Neural Information Processing Systems*, pages 15413–15423, 2019.

[351] U. Alon and E. Yahav. On the bottleneck of graph neural networks and its practical implications. *arXiv preprint arXiv:2006.05205*, 2020.

[352] K. Oono and T. Suzuki. Graph neural networks exponentially lose expressive power for node classification. *arXiv preprint arXiv:1905.10947*, 2019.

[353] P. Velickovic, W. Fedus *et al.* Deep graph infomax. In *ICLR (Poster)*, 2019.

[354] J. You, R. Ying, and J. Leskovec. Position-aware graph neural networks. *arXiv preprint arXiv:1906.04817*, 2019.

[355] D. Flam-Shepherd, T. Wu *et al.* Neural message passing on high order paths. *arXiv preprint arXiv:2002.10413*, 2020.

[356] I. A. Kovács, K. Luck *et al.* Network-based prediction of protein interactions. *Nature Communications*, 10(1):1–8, 2019.

[357] K. Huang, C. Xiao *et al.* Skipgnn: Predicting molecular interactions with skip-graph networks. *arXiv preprint arXiv:2004.14949*, 2020.

[358] J. Halcrow, A. Mosoi *et al.* Grale: Designing networks for graph learning. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2523–2532, 2020.

[359] S. Bai, F. Zhang, and P. H. Torr. Hypergraph convolution and hypergraph attention. *arXiv preprint arXiv:1901.08150*, 2019.

[360] H. Gao and S. Ji. Graph u-nets. *arXiv preprint arXiv:1905.05178*, 2019.

[361] F. Diehl, T. Brunner *et al.* Towards graph pooling by edge contraction. In *Proceedings of the ICML Workshop on Learning and Reasoning with Graph-Structured Data, Los Angeles, CA, USA*, volume 15, 2019.

[362] F. M. Bianchi, D. Grattarola, and C. Alippi. Mincut pooling in graph neural networks. *arXiv preprint arXiv:1907.00481*, 2019.

[363] E. Ranjan, S. Sanyal, and P. P. Talukdar. Asap: Adaptive structure aware pooling for learning hierarchical graph representations. In *AAAI*, pages 5470–5477, 2020.

[364] R. Oughtred, C. Stark *et al.* The biogrid interaction database: 2019 update. *Nucleic acids research*, 47(D1):D529–D541, 2018.

[365] M. Kotlyar, C. Pastrello *et al.* Integrated interactions database: tissue-specific view of the human and model organism interactomes. *Nucleic Acids Research*, 44(D1):D536–D541, 2015.

[366] G. Wu, X. Feng, and L. Stein. A human functional protein interaction network and its application to cancer data analysis. *Genome Biology*, 11(5):R53, 2010.

[367] J. X. Binder, S. Pletscher-Frankild *et al.* Compartments: unification and visualization of protein subcellular localization evidence. *Database*, 2014, 2014.

[368] Z. Zhang, J. Jiang *et al.* Chimeric antigen receptor t cell targeting egfrviii for metastatic lung cancer therapy. *Frontiers of Medicine*, 13(1):57–68, 2019.

[369] N. Erfani, S. M. Mehrabadi *et al.* Increase of regulatory t cells in metastatic stage and ctla-4 over expression in lymphocytes of patients with non-small cell lung cancer (nsclc). *Lung Cancer*, 77(2):306–311, 2012.

[370] N. Goto, H. Hiyoshi *et al.* Estrogen and antiestrogens alter breast cancer invasiveness by modulating the transforming growth factor-$\beta$ signaling pathway. *Cancer Science*, 102(8):1501–1508, 2011.

[371] S. J. Ellison-Zelski and E. T. Alarid. Maximum growth and survival of estrogen receptor-alpha positive breast cancer cells requires the sin3a transcriptional repressor. *Molecular Cancer*, 9(1):263, 2010.

[372] M. J. Lewis, J. Liu *et al.* Sin3a and sin3b differentially regulate breast cancer metastasis. *Oncotarget*, 7(48):78713, 2016.

[373] B. Elsberger. Translational evidence on the role of src kinase and activated src kinase in invasive breast cancer. *Critical Reviews in Oncology/Hematology*, 89(3):343–351, 2014.

[374] C. D. Allred and M. W. Kilgore. Selective activation of ppar$\gamma$ in breast, colon, and lung cancer cell lines. *Molecular and Cellular Endocrinology*, 235(1-2):21–29, 2005.

[375] H.-N. Yu, Y.-R. Lee *et al.* Induction of g1 phase arrest and apoptosis in mda-mb-231 breast cancer cells by troglitazone, a synthetic peroxisome proliferator-activated receptor $\gamma$ (ppar$\gamma$) ligand. *Cell Biology International*, 32(8):906–912, 2008.

[376] H. Hu, A. Goltsov *et al.* Feedforward and feedback regulation of the mapk and pi3k oscillatory circuit in breast cancer. *Cellular Signalling*, 25(1):26–32, 2013.

[377] J.-C. Wu, C.-S. Lai *et al.* Tetrahydrocurcumin, a major metabolite of curcumin, induced autophagic cell death through coordinative modulation of pi3k/akt-mtor and mapk signaling pathways in human leukemia hl-60 cells. *Molecular Nutrition & Food Research*, 55(11):1646–1654, 2011.

[378] V. Sánchez-Margalet and S. Najib. Sam68 is a docking protein linking gap and pi3k in insulin receptor signaling. *Molecular and Cellular Endocrinology*, 183(1-2):113–121, 2001.

[379] A. Belfiore and F. Frasca. Igf and insulin receptor signaling in breast cancer. *Journal of Mammary Gland Biology and Neoplasia*, 13(4):381–406, 2008.

[380] L. Lapeire, A. Hendrix *et al.* Cancer-associated adipose tissue promotes breast cancer progression by paracrine oncostatin m and jak/stat3 signaling. *Cancer Research*, 74(23):6806–6819, 2014.

[381] S. Saha Roy and R. K. Vadlamudi. Role of estrogen receptor signaling in breast cancer metastasis. *International Journal of Breast Cancer*, 2012, 2012.

[382] I. Ahmad, T. Iwata, and H. Y. Leung. Mechanisms of fgfr-mediated carcinogenesis. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1823(4):850–860, 2012.

[383] I. A. Mayer and C. L. Arteaga. The pi3k/akt pathway as a target for cancer treatment. *Annual Review of Medicine*, 67:11–28, 2016.

[384] Y. Zeng and N. Lai. Association between the cd28 c.17 +3 t¿c polymorphism (rs3116496) and cancer risk: An updated meta-analysis. *Medical Science Monitor*, 25:1917, 2019.