Deciphering the functional organization of molecular networks via graphlets-based methods and network embedding techniques

Sergio Doria Belenguer

UNIVERSITAT DE BARCELONA

FACULTAD DE BIOLOGÍA

DOCTORAL PROGRAMME IN BIOMEDICINE

DECIPHERING THE FUNCTIONAL ORGANIZATION OF MOLECULAR NETWORKS VIA GRAPHLETS-BASED METHODS AND NETWORK EMBEDDING TECHNIQUES

Dissertation of:
Sergio Doria Belenguer

Supervisor:
Prof. Nataša Pržulj

Tutor:
Prof. Josep Lluis Gelpi

Deciphering the functional organization of molecular networks via graphlets-based methods and network embedding techniques

Sergio Doria Belenguer

Natasa

Nataša Pržulj
Thesis supervisor

Josep Lluis Gelpi
Thesis tutor

Universitat de Barcelona, Faculty of Biology

PhD Program: Biomedicine
Developed at: Barcelona Supercomputing Center

UNIVERSITAT DE BARCELONA

Submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Biomedicine
of the Universitat de Barcelona

Abstract

Advances in capturing technologies have yielded a massive production of large-scale molecular data that describe different aspects of cellular functioning. These data are often modeled as networks, in which nodes are molecular entities, and the edges connecting them represent their relationships. These networks are a valuable source of biological information, but they need to be untangled by new algorithms to reveal the information hidden in their wiring patterns. State-of-the-art approaches for deciphering these complex networks are based on graphlets and network embeddings. This thesis focuses on the development of novel algorithms to overcome the limitations of the current graphlet and network embedding methodologies in the field of biology.

Graphlets are a powerful tool for characterizing the local wiring patterns of molecular networks. However, current graphlet-based methods are mostly applicable to unweighted networks, whereas real-world molecular networks may have weighted edges that represent the probability of an interaction occurring in the cell. This probabilistic information is commonly discarded when applying thresholds to generate unweighted networks, which may lead to information loss. To address this challenge, we introduce probabilistic graphlets, a novel approach that can capture the local wiring patterns of weighted networks and uncover hidden probabilistic relationships between molecular entities. We use probabilistic graphlets to generalize the graphlet methods and apply these to the probabilistic representation of real-world molecular interactions. We show that probabilistic graphlets robustly uncover relevant biological information from the molecular networks. Furthermore, we demonstrate that probabilistic graphlets exhibit a higher sensitivity to identifying condition-specific functions compared to their unweighted counterparts.

Network embedding algorithms learn a low-dimensional vectorial representation for each gene in the network while preserving the structural information of the molecular network. Current, available embedding approaches strictly focus on clustering the genes' embedding vectors and interpreting such clusters to reveal the hidden information of the biological networks. Thus, we investigate new perspectives and methods that go beyond gene-centric approaches. First, we shift the exploration of the embedding space's functional organization from the genes to their functions. We introduce the Functional Mapping Matrix and apply it to investigate the changes in the organization of cancer and control network embedding spaces from a functional perspective. We demonstrate that our methodology identifies novel cancer-related functions and genes that the currently available methods for gene-centric analyses cannot identify. Finally, we go even further and switch the perspective from the organization of the embedded entities (genes and functions) in the embedding space to the space itself. We annotate axes of the network embedding spaces of six species with both, functional annotations and genes. We demonstrate that the embedding

space axes represent coherent cellular functions and offer a functional fingerprint of the cell's functional organization. Moreover, we show that the analysis of the axes reveals new functional evolutionary connections between species.

Dedication

To my family, for those that are still with us, and for those who left.

Declaration

I, Sergio Doria Belenguer, 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.

Acknowledgements

I would like to take this space to express my sincere gratitude to the individuals who have contributed to the completion of this thesis.

First, I would like to thank my research group. To Nataša Pržulj, thank you for trusting me and giving me the opportunity to study in such an exciting and impactful field. You always had kind words when I needed them, as well as tough love when I needed a reality check. To Noël Malod-Dognin, thank you for transforming my articles from a white and colorless page to a piece of art full of yellow. To my lab crew -Daniel Tello, Alexandros Xenos, Misha Rotkevich, and Carme Zambrana- thank you for sharing this experience from the beginning. Without you, this adventure would not have been the same. To the "traitors", Markus Kirolos and Gaia Ceddia, thank you for the time you spent with me (never forget never forgive).

Second, to my friends and family, thank you for your support and for being there for me throughout this process.

Finally, to myself, thank you for never giving up and demonstrating that hard work always pays off.

Contents

1 Introduction 10

1.1 Motivation . . . 10

1.2 Objectives . . . 12

1.3 Contribution . . . 12

1.3.1 Resulting papers . . . 13

1.3.2 Posters and Talks . . . 14

1.4 Thesis outline . . . 14

2 Background 16

2.1 Biological networks . . . 16

2.2 Concepts on Networks . . . 22

2.2.1 Network Representations . . . 24

2.2.2 Network Properties . . . 25

2.3 Network Models . . . 29

2.3.1 Erdös-Renyi Random networks . . . 29

2.3.2 Scale-free Networks . . . 30

2.3.3 Geometric Networks . . . 31

2.3.4 Stickiness Index Based Networks . . . 31

2.4 Network Comparison . . . 32

2.5 Network Clustering and Functional Enrichment Analysis . . . 33

2.5.1 Network Clustering . . . 33

2.5.2 Functional Enrichment Analysis . . . 35

2.5.2.1 GO terms Summary . . . 36

2.6 Graphlets and Graphlet-based methods . . . 37

2.6.1 Graphlets . . . 38

2.6.2 Graphlets-based Measures of Nodes Topology . . . 39

2.6.3 Graphlets-based Measures of Entire Network Topology . . . 41

2.7 Network Embeddings-based Methods . . . 41

2.7.1 NLP-inspired Embedding Approaches . . . 43

2.7.2 Matrix factorization-based approaches . . . 44

3 Probabilistic graphlets capture biological function in probabilistic molecular networks 46

3.1 Motivation . . . 46

3.2 Materials and Methods . . . 47

3.2.1 Definition of probabilistic graphlets . . . 47

3.2.2 Probabilistic graphlet count implementation . . . 49

3.2.3 Definition of topological distance measures . . . 49

3.2.4 Biological datasets 50
3.2.5 Synthetic datasets 51
3.2.6 pGCM visualization and measuring the utility of pGCD for clustering of model networks 51
3.2.7 Measurement of functional enrichments in clusters 52
3.2.8 Assessment of GO annotation similarity 52
3.3 Results and Discussion 53
3.3.1 Probabilistic graphlets are superior for network comparisons . 53
3.3.2 Probabilistic graphlets have similar or better performances than their unweighted counterpart 56
3.3.3 Probabilistic graphlets capture different biological information 57
3.3.4 Probabilistic graphlets uncover more specific biological information 59
3.3.5 Probabilistic and unweighted graphlets find different functions of a gene 60
3.3.6 Robustness analysis 61
3.4 Conclusions 63

4 A functional analysis of omic network embedding spaces reveals key altered functions in cancer 64
4.1 Motivation 64
4.2 Materials and Methods 65
4.2.1 Biological datasets 65
4.2.2 Definition of cancer-related biological annotations 66
4.2.3 Embedding the PPI networks 67
4.2.4 Definition of the Functional Mapping Matrix 67
4.2.5 Measuring the similarity of functional organization of the embedding spaces by using their FMMs 68
4.2.6 Evaluating the functional organization of an embedding space with its FMM 68
4.2.7 Quantifying the “movement” of the annotation embedding vectors in cancer and control embedding spaces 70
4.2.8 Distances between the embedded entities in the embedding space 71
4.3 Results and Discussion 71
4.3.1 Cancer alters the functional organization of the healthy cell embedding space 71
4.3.2 The “movement” of the annotations in the embedding space predicts cancer-related functions 73
4.3.3 The “movement” of cancer-related annotations in the embedding spaces predicts oncogenic genes 76
4.4 Conclusion 78

5 The axes of biology: a novel axes-based network embedding approach to decipher the fundamental mechanisms of the cell 80
5.1 Motivation 80
5.2 Materials and Methods 82
5.2.1 Biological datasets 82
5.2.2 Embedding the PPI networks 83

5.2.3 Annotating the axes of the gene embedding space with GO BP terms 84
5.2.4 Quantifying the evolutionary conservation of biological functions 85
5.2.5 Evaluating our axes-based methodology 86
5.2.6 Generating the Axes-Specific Functional Annotations 87
5.2.7 Analyzing the connection between the ASFAs and evolution . 88
5.3 Results and Discussion 88
5.3.1 The axes of the embedding spaces capture the cell's functional organization 88
5.3.2 Orthonormality and positive constraints improve the functional organization of the gene embedding space 90
5.3.3 Specific biological functions are disentangled by the axes of the embedding space with the increment of dimensions 91
5.3.4 The axes of the embedding space represent the fundamental mechanisms of the cell 92
5.3.5 The axes of the embedding space uncover the human evolutionary history 95
5.3.6 Non-Annotated Axes also capture the functional mechanisms of the cell 97
5.4 Conclusions 99

6 Conclusion 100
6.1 Summary of thesis achievements 100
6.2 Conclusions 102
6.3 Future work 103
6.3.1 Identifying Pan-cancer functions with our FMM 103
6.3.2 Measuring the evolutionary closeness between species using the FMM 104
6.3.3 Uncovering complementary information from network embeddings using different perspectives 105
6.3.4 Generalizing probabilistic graphlets to network embeddings . . 106

Bibliography 108

Appendices 124

A Supplementary Information for Chapter 3 125
A.1 Supplementary Results and Discussion for Chapter 3 125
A.1.1 pGDV captures different functional information that is not seen by GDV 125
A.2 Supplementary Figures for Chapter 3 126

B Supplementary Information for Chapter 4 136
B.1 Supplementary Materials and Methods for Chapter 4 136
B.1.1 Multiplicative update rules 136
B.2 Supplementary Results and Discussion for Chapter 4 137
B.2.1 Impact of the PPI network matrix representation to the functional organization of the embedding space 137

B.2.2 Our FMM-based methodology captures more biological information from the embedding space compared to the actual gene-centric approaches 139
B.2.3 The FMMs reveal the higher-order functional organizations of the GO BP terms in the network embedding spaces 140
B.2.4 FMM discriminates between functionally and not functionally organized embedding spaces 141
B.2.5 FMMs identify novel cancer-related functions 141
B.3 Supplementary Figures for Chapter 4 143
B.4 Supplementary Tables for Chapter 4 152

C Supplementary Information for Chapter 5 158
C.1 Supplementary Results and Discussion for Chapter 5 158
C.1.1 The Axes of the embedding space synthesize the core functions of different species cells 158
C.1.2 The Axes of the embedding space give insights into the evolutionary story of species 159
C.2 Supplementary Figures for Chapter 5 160
C.3 Supplementary Tables for Chapter 5 165

Chapter 1

Introduction

1.1 Motivation

Current technology is producing high-throughput biological data at an ever-growing rate. These data are often modeled as networks, in which nodes are molecular entities, and edges define their relationships, e.g., in protein-protein interaction networks (PPIs), nodes represent proteins and edges indicate physical interactions (binding) between them, as measured by biological experiments. These networks are a valuable source of biological information, but they need to be untangled by new algorithms to reveal the information hidden in their wiring patterns [1].

A powerful approach for uncovering information from a biological network is to measure the local wiring patterns of its nodes. Nodes with similar local wiring patterns share biological functions regardless of their proximity in the biological network, e.g., in a PPI network, functions are not only shared among proteins that physically interact but also among proteins with similar wiring patterns [2]. The best measures to quantify the local wiring patterns rely on graphlets, small, connected, non-isomorphic, induced subgraphs of a larger network [3]. Graphlets and their statistics have been used in network biology to compare biological networks [4], to uncover their functional organization [3, 2, 4] as a base for network alignment algorithms (GRAAL family [3, 5]), or to relate genes in these networks with their biological functions [2, 4, 6]. Moreover, to account for the complex information in molecular systems, graphlets have been generalized in different ways, including directed networks [7], in which the directions of the interactions play an essential role; hypergraphs and simplets, that capture the multi-scale organization of biological systems [8, 9].

However, graphlet-based measures are mostly applicable to unweighted networks, while some molecular interactions, such as gene coexpression [10] or PPIs [11], can be modeled as probabilistic networks, in which edges have an associated probability to reflect the level of confidence in the existence of the molecular interaction. This probabilistic information is usually discretized by applying thresholds to decide which interactions are considered to be real. While this approach permits the removal of noise from the data, it can also eliminate crucial information [10]. In Chapter 3, we propose solutions to extract the additional information hidden in the wiring patterns of probabilistic networks by extending the graphlet-based methodology to probabilistic networks.

Graphlet-based approaches have been successfully applied in network biology, but they also present several limitations. The number of graphlets containing k nodes increases super-exponentially with k. The computation complexity of counting graphlets on k nodes in a network with n nodes is $O(n^k)$, but in practice, it is much lower since biological networks are sparse. This makes the methods inefficient when dealing with large-scale data limiting their application in increasingly complex biological networks. In addition, even when graphlets are computed, their biological interpretation can be challenging due to their large number. In particular, for some graphlet generalizations, such as hypergraphlets, this number goes up to 450 [12]. These limitations highlight the need for alternative approaches that can decipher these notoriously complex biological networks.

An alternative to graphlet-based approaches is *network embedding technique*, also called *graph representation learning*. Unlike graphlet-based methods, which operate on biological networks, network embedding techniques employ dimensionality reduction techniques, such as matrix factorization, autoencoders, or graph neural networks, to represent the networks in a low-dimensional space facilitating the analysis of these extremely complex networks [13, 14]. These algorithms aim to map nodes in a network to a low-dimensional embedding space, spanned by a system of coordinates (a.k.a., embedding axes), in which the closeness of nodes in the original network is preserved [14, 15]. Defining an optimal number of dimensions of the embedding space is key to properly representing the closeness between the nodes. However, there is no gold-standard approach to finding the optimal dimensionality of the embedding space. Thus, researchers have to rely on grid search, domain knowledge, or heuristics [16], e.g., the cophenetic correlation coefficient [17] and rule of thumb [18]. In Chapters 4 and 5, we propose solutions to find the optimal dimensionality of the embedding space.

Network embedding includes a variety of algorithms, such as Natural Language Processing (NLP)-inspired methods [14], e.g., DeepWalk [19], and node2vec [20], and matrix factorization-based approaches, e.g, the Non-negative Matrix Tri-Factorization (NMTF), and the Non-negative Matrix Factorization [21]. These algorithms have been widely applied to uncover new biological knowledge from biological networks, such as novel associations of miRNA with diseases [22], new protein functions [14], or new drug target associations [23]. However, as shown with graphlet-based methods, current embedding approaches also have several drawbacks. These gene-centric methods use the learned embedding vectors of the genes as input to machine learning algorithms to perform downstream tasks [13], such as node clustering and classification [13]. Hence, other possible information sources, such as the embedding axes, remain unexplored. Moreover, by only taking as input the genes' embedding vectors and not the functions of these genes, these gene-centric approaches offer an incomplete analysis of biomedical data. In Chapters 4 and 5, we propose solutions to these limitations by introducing two new network embedding approaches.

In the rest of this Chapter, we first present the objectives of this dissertation. Following this, we introduce the contributions of this dissertation including the resulting scientific papers, posters, and talks. We conclude this section with the dissertation outline.

1.2 Objectives

The goal of this thesis is the development of new computational methods to extract biological information from biological networks. The state-of-the-art approaches to deciphering these complex data are based on graphlets and network embeddings. Thus, in this thesis, we focus on the development of novel algorithms to overcome the limitations of the current graphlet and network embedding methodologies in the field of biology. In this context, our first objective is the generalization of the graphlet-based methodology to probabilistic networks. Our second objective is the development of alternative approaches to the current gene-centric embedding techniques for mining new biological information from network embeddings. The third objective, ideally, is to combine the graphlets' generalization with the new network embedding approaches to extend their capabilities.

1.3 Contribution

During this dissertation, we introduce three new computational methods, one graphlet-based method and two network embedding approaches, to mine biological information from biological networks.

First, we introduce *probabilistic graphlets* as a tool for analyzing the local wiring patterns of probabilistic networks. This method extends the use of graphlets and graphlet-based measures to probabilistic networks. In **Chapter 3**, we introduce and apply probabilistic graphlets on the probabilistic representation of real-world molecular interaction networks and synthetic networks (generated by well-known network models). We demonstrate that probabilistic graphlets outperform their unweighted counterparts in capturing the overall topological similarity between synthetic networks. Moreover, our probabilistic graphlets robustly manage low signal topology information without sacrificing their ability to recover relevant biological information from molecular networks. In contrast, the original unweighted graphlets applied in thresholded networks are highly sensitive to both, the noise and the chosen threshold. Finally, we show that probabilistic graphlets robustly capture condition-specific cellular processes such as stress response mechanisms, which in turn benefit from the use of probabilistic models. Hence, the use of our probabilistic graphlets complements the biological information uncovered by their original unweighted counterparts.

Then, to overcome the limitations of the current gene-centric embedding approaches, we propose a new, function-centric perspective and approach to explore the functional organization of the network embedding space from a functional perspective. We introduce the *Functional Mapping Matrix* (FMM), this matrix captures the mutual positions of functional annotations that we embed in the gene embedding space. In **Chapter 4**, we introduce and apply our FMM-based method to capture the functional organization of different tissue-specific PPI embedding spaces generated by the NMTF algorithm. We show that our FMM can efficiently be applied to address different problems, e.g., to find the optimal dimensionality of the embedding space, to analyze the similarities between the functional organization of different embedding spaces (in Chapter 4, those corresponding to cancer and control), and to find the functional changes produced by cancer. Moreover, we demonstrate that our FMM-based method predicts new cancer-related genes that could not have been identified by currently available methods for gene-centric analysis. We validate these predicted genes by literature curation and retrospective analyses of patient survival data. To conclude, we show that our FMM-based method can be easily extended to other bioinformatics tasks, such as patient and tissue stratification, or to uncover evolutionary similarities between species.

While our FMM-based methodology changes the exploration of the embedding space from the genes' embedding vectors to the vectorial representation of their functional annotations, both approaches still focus on embedded entities (genes and functional annotations) to decipher information from embed networks. In Chapter 5, we propose to change the perspective from the organization of the entities in the embedding space to the space itself. In particular, we introduce a new method that uses the axes of the embedding space where the entities are embedded to capture the cell's functional organization from molecular networks. In **Chapter 5**, we apply our axes-based method to uncover the fundamental mechanisms of the functioning of a cell. To this end, we use it on different species-specific PPI embedding spaces generated by the NMTF and Deepwalk algorithms. We demonstrate that the axes of the embedding space disentangle biological information, with functionally similar gene annotations being associated with the same axis. Moreover, we demonstrate that the embedding axes better disentangle biological information than the classic gene-centric approach. Furthermore, we show that the embedding in orthonormal spaces, which only NMTF-based frameworks allow for, leads to the embedding spaces that best capture the cell's functional organization from the biological networks, i.e., to better disentangle functional information about the cell's mechanisms. In addition, we demonstrate that our axes-based methodology can be applied to find the optimal dimensionality of the embedding space. Finally, we introduce the Axes-Specific Functional Annotations (ASFA for short) for summarizing the functional annotations associated with the axes. We demonstrate that ASFAs are not only functionally coherent, but also can be used to get insights into the evolutionary history of humans.

In the rest of this section, we first list the scientific papers contained in this thesis and their current publication status. Then, we list the posters and talks that were presented at international scientific congresses as a result of this thesis.

1.3.1 Resulting papers

The research conducted as part of this thesis resulted in three scientific papers:

- Doria-Belenguer Sergio, Kirolos Markus, Böttcher René, Malod-Dognin Noël and Pržulj Nataša. "Probabilistic graphlets capture biological function in probabilistic molecular networks". Bioinformatics 36. Supplement_2 (2020), pp. i804–i812. (DOI: https://doi.org/10.1093/bioinformatics/btaa812).

- Doria-Belenguer Sergio, Xenos Alexandros, Ceddia Gaia, Malod-Dognin Noël and Pržulj Nataša. "A functional analysis of omic network embedding spaces reveals key altered functions in cancer". Bioinformatics (2023).
(DOI: https://doi.org/10.1093/bioinformatics/btad281).

- Doria-Belenguer Sergio, Xenos Alexandros, Ceddia Gaia, Malod-Dognin Noël and Pržulj Nataša. "The axes of biology: a novel axes-based network embedding approach to decipher the fundamental mechanisms of the cell". (under submission).

1.3.2 Posters and Talks

The research carried out during the development of this thesis was presented at two scientific conferences through one poster and one oral presentation:

- 19^{th} European Conference on Computational Biology (ECCB2020) virtually held from September 31^{st} to 8^{th}, 2020.
  - Doria-Belenguer Sergio, Kirolos Markus, Böttcher René, Malod-Dognin Noël and Pržulj Nataša. "Probabilistic graphlets capture biological function in probabilistic molecular networks." (Proceedings Systems Track - Speaker).
- 21^{st} European Conference on Computational Biology (ECCB2022) held from 18^{th} to the 21^{st} of September, 2022, in Sitges, Spain.
  - Doria-Belenguer Sergio, Xenos Alexandros, Ceddia Gaia, Malod-Dognin Noël and Pržulj Nataša. "A functional analysis of omic network embedding spaces reveals key altered functions in cancer." (Posters track).

1.4 Thesis outline

The thesis is structured as follows:

In **Chapter 2** we provide an overview of the current state of the knowledge and research in the field of biological networks, highlighting the main issues and questions that the present dissertation aims to address.

In **Chapter 3**, we extend the graphlet-based methodology to probabilistic networks by introducing our probabilistic graphlets. First, we assess their performance compared to unweighted graphlets, by generating synthetic networks based on different well-known random network models and edge probability distributions. We demonstrate that probabilistic graphlets outperform their unweighted counterparts in distinguishing network structures. Following this, we model various real-world molecular interaction networks as weighted graphs with probabilities as weights on edges and we analyze them with our new probabilistic graphlets-based methods. We show that due to their probabilistic nature, probabilistic graphlet-based methods more robustly capture biological information in these data, while simultaneously showing a higher sensitivity to identify condition-specific functions compared to their unweighted graphlet-based method counterparts.

In **Chapter 4**, we introduce the FMM, a new network embedding approach for exploring the functional organization of the embedding space from a functional perspective. We use our FMM-based methodology to explore the functional organization of different tissue-specific PPI embedding spaces generated by the NMTF algorithm. Also, we use our FMM to define the optimal dimensionality of these molecular interaction networks embedding spaces. For this optimal dimensionality, we compare the FMMs of the most prevalent cancers in humans to the FMMs of their corresponding control tissues. We find that cancer alters the positions in the embedding space of cancer-related functions, while it keeps the positions of the non-cancer-related ones. We exploit this spacial "movement" to predict novel cancer-related functions. Finally, we predict novel cancer-related genes that the currently available methods for gene-centric analyses cannot identify; we validate these predictions by literature curation and retrospective analyses of patient survival data.

In **Chapter 5**, instead of using the organization of the embedded entities (e.g., genes or functional annotations) in the embedding space, we propose to use the axes of the space where these entities are embedded to uncover the fundamental mechanisms of the cell, i.e., those mechanisms that summarize the most important functions of the cell. We introduce a new network embedding approach that shifts the exploration of network embeddings from the genes' embedding vectors to the axes of the embedding space. We apply our axes-based method to uncover new biological information from the functional organization of different species-specific PPI embedding spaces generated by the NMTF and Deepwalk algorithms. First, we demonstrate that the embedding axes disentangle coherent biological information from the functional organization of these species-specific PPI embedding spaces. We also show that the properties of the embedding spaces produced by NMTF, such as orthonormality and positive constraint, improve the organization of such embedding spaces resulting in the disentangling of more biological information by their embedding axes. Then, we use our axes-based method to define the optimal dimensionality of the NMTF embedding spaces of each species. For this optimal dimensionality, we perform an in-depth analysis of the biological meaning of the GO BP terms associated with their axes by generating their ASFAs. We investigate the functional coherence of those ASFAs with manual literature curation and analyze them in the context of evolution. Finally, we explore the biological meaning of those axes that do not capture any GO BP terms.

Finally, in **Chapter 6**, we conclude the dissertation by providing a summary of the thesis achievements and future steps.

Chapter 2

Background

In this section, we introduce the background for the research presented in this dissertation. We begin by introducing the main types of real-world biological networks. Then, we introduce the relevant network theoretic analysis of biological networks, including the most simple measures of network topology. We continue by introducing the basic random network models that are used to understand the structural properties of biological networks. After this, we introduce the most frequently used approach to link the topology of a network with biological knowledge, clustering and functional enrichment analysis. Finally, as the main focus of this dissertation is the development of new computational applications to extract biological information from biological networks, we conclude this Chapter by introducing state-of-the-art computational approaches for uncovering the biological information hidden in their wiring patterns: graphlets-based methods and network embedding-based approaches.

2.1 Biological networks

Biological networks are an abstract conceptual model of a definite set of biological entities (e.g., proteins, genes, or metabolic pathways), and the relationships between those entities (i.e., interactions among proteins). These networks appear in different forms and represent different types of information about the cell. Mining these biological networks uncovers valuable insights into the organization and dynamics of biological systems.

Biological networks can be directed or undirected, depending on the presence of a direction among node pairs that form the edges. Similarly, edges can be unweighted or weighted for representing their relative importance in the network (network models are defined in section 2.2). In Chapter 3 of this dissertation, we focus on a special case of undirected and weighted biological networks. In particular, we focus on probabilistic networks, in which the weights represent the probability that an interaction between two nodes occurs. On the other hand, in Chapters 4 and 5, we mainly concentrate on undirected and unweighted biological networks, since the networks of this form still carry valuable amount of information, and the embedding approaches for analyzing these networks are much more scalable to large networks than their weighted counterparts.

The main types of biological networks are PPI networks, genetic interaction networks, gene co-expression networks, metabolic networks, transcriptional regulation networks, and cell signaling networks. In this section, we describe these biological networks and specify in which chapters of this dissertation we analyze them.

PPI networks: Proteins are responsible for the majority of cellular activities and serve as the fundamental components for the structure and function of cells. Most cellular functions need stable and transient PPIs to be formed. PPI networks represent the binding information among all proteins of an organism; nodes represent the proteins and edges represent physical interactions between two proteins. Protein interactions are highly correlated with gene expression. Changes in gene expression alter the number of available proteins in a cell, therefore regulating the type of possible protein interactions in the cell. These changes occur as part of normal cellular processes, such as cell differentiation, or as part of pathological events, such as cancer transformation. Gene expression data is usually used to construct *tissue-specific PPI networks*. This is achieved by isolating the subnetwork of interactions from the global PPI network that involve proteins expressed within the tissue of interest [1, 24]. PPI networks appear as undirected networks. While the interactions in these networks are typically represented as unweighted edges, some studies assign weights to these edges to reflect the level of confidence in the existence of the interaction [25, 11].

PPIs can be detected through various experimental essays, and can also be predicted by computational methods. Experimental PPI detection methods can be characterized by the number of interactions (high vs low-throughput), the types of proteins involved (e.g., membrane, soluble), whether the proteins are modified, the types of interactions detected (e.g., direct and indirect), and the settings where interactions are detected [26]. Specific detection methods identify different numbers of interactions: Low-throughput methods (LT), such as affinity chromatography essays or affinity precipitation [27], identify in the order of tens of interactions. In contrast, high-throughput (HR) methods, such as yeast two-hybrid screening or affinity purification-mass spectrometry experiments [28], detect hundreds to thousands of interactions. The main difference between LT and HR methods is the reliability of their results. Studies based on LT methods usually employ a combination of approaches to identify PPIs [29], and consequently, their results are considered reliable [30, 31]. On the other hand, HR methods bring the possibility of identifying a large fraction of all PPIs in an organism in a relatively short amount of time. However, they have not replaced LT methods due to their higher cost, larger complexity, as well as, concerns about their accuracy [32]. Indeed, the error rates of these high-throughput methods are quite high, and it has been demonstrated that only 35-83% of the PPIs detected by these methods are actually real [33]. Hence, PPIs identified by HR methods have to be validated with LR approaches.

Apart from the number of PPIs that they can detect, experimental PPI detection methods also vary on the type of proteins involved in the interactions they can identify. For instance, some detection methods, such as yeast two-hybrid methods, do not detect interactions between membrane proteins, while other methods, such as mammalian membrane two-hybrid assays, only capture interactions between this type of proteins [34]. Another important distinction between detection methods is the type of interactions that they can capture. PPIs are commonly divided into two classes: direct interactions (when two proteins physically interact) and indirect interactions (when they are part of the same complex but not in direct contact). Some methods, such as affinity purification-mass spectrometry, can identify complexes involving more than two proteins but do not report which pairs of proteins are in direct contact. In contrast, detection methods, such as mammalian membrane two-hybrid, detect direct interaction between proteins but not indirect ones [35]. Hence, the selection of a specific PPI detection method will result in PPI networks with very different properties [26].

Finally, a common limitation of most detection methods is that they do not detect PPIs in an entirely natural setting. Most methods, including yeast two-hybrid, luminescence-based mammalian interactome mapping, and fluorescence resonance energy transfer, require modifications to candidate interacting proteins (e.g., tagging them with a specific binding domain or a fluorescent label), which can potentially interfere with the interactions being studied [36, 37]. Moreover, common methods, including yeast two-hybrid and membrane yeast two-hybrid, are carried out in yeast cells regardless of the organisms the genes are taken from. The genes from different organisms may not behave as in their native environment when they are in the yeast nucleus. Indeed, it is estimated that 50% of the PPIs identified by yeast two-hybrid screening assays are noisy [30].

Regarding the computational methods to predict PPIs, there are numerous approaches that rely on machine learning, statistical, and graph-theoretical approaches. However, interaction predicted by these methods should be used with caution or excluded in most analyses, since PPI networks already contain high levels of experimental noise which will exponentially increase with the inclusion of predicted interactions. Hence, during this dissertation, we build our PPI networks with those PPIs that are experimentally validated and exclude those that are predicted by computational approaches.

Finally, one of the main problems with PPI networks is their incompleteness. For a network of n nodes, there exists $n(n-1)/2$ possible interactions. The estimated size of the human proteome is approximately 70,000 [38], raising the need for testing 2.5 billion PPIs for their existence. In addition to this huge number of possibilities, most of the PPIs identification studies are focused on a certain process or disease, leaving the other parts of the PPI network uncovered. In fact, less than 40% of the human PPI network is known, assuming a total of about 650,000 interactions [39]. Moreover, even for *Saccharomyces Cerevisiae*, the most well-studied organism for PPI networks, only about half of its PPI network has been discovered, assuming an estimated number of interactions ranging from 37,800 to 75,500 [40].

Over the last decades, there has been a great effort to build biological databases and resources on PPI data. The main public databases that contain PPI networks are the Biological General Repository for Interaction Datasets (BioGrid) [41], the Saccharomyces Genome Database (SGD) [42], the Munich Information Center for Protein Sequences (MIPS) [43], the Database of Interacting Proteins (DIP) [44], the Human Protein Reference Database (HPRD) [45], the Molecular Interaction database (MINT) [46], and the Search Tool for the Retrieval of Interacting Genes / Proteins Database (STRING) [47].

In Chapter 3 of this dissertation, we apply our graphlet-based method to the unweighted and probabilistic PPI networks of *Homo sapiens sapiens* and *Saccharomyces Cerevisiae*. We collect these PPI networks from the STRING database. In Chapter 4 of this dissertation, we apply our new network embedding method to analyze the tissue-specific PPI networks of the most prevalent cancers in humans and their corresponding control tissues. The human PPI network used for generating these tissues-specific PPI networks was collected from BioGrid. Finally, in Chapter 5, we use a new network embedding approach to investigate the PPI networks of five species. We obtain these species-specific PPI networks from BioGrid. Details about the construction and acquisition of these PPI networks can be found in their corresponding chapter.

Genetic interaction networks: These networks link genes that genetically interact. Two genes are said to have genetic interaction if simultaneous mutations to both genes result in a phenotype that would have not been expected by the phenotype of both single mutations [48]. The most commonly used phenotype for measuring the effects of gene mutations is the cell fitness; i.e., the ability of a cell to survive and reproduce in the cells' culture. The effects of the mutations on the cell fitness are usually measured as follows. When a single gene is mutated, we quantify the fitness of the mutated cell by measuring its growth rate compared to that of the wild type (non-mutant). Then, we mutate a pair of genes, A and B, in the same cell (double mutant) and measure the fitness of the double mutant. We expect the fitness of the double mutant to be a function of the fitness of the two single mutants, i.e., of the cell with only gene A mutated and the cell with only gene B mutated. When the fitness of the double mutant is significantly different from the expected combined fitness of the single mutants, then we say that genes A and B interact [26]. Genetic interactions include two broad categories: *positive* and *negative* [49]. Positive genetic interactions refer to double mutants with a less severe fitness defect than expected. Conversely, negative genetic interactions refer to a more severe fitness defect than expected, with an extreme case being synthetic lethality [50]. Examples of genetic interactions are synthetic lethality, synthetic sickness, and synthetic growth defect. Genetic interactions are detected by synthetic genetic array (SGA) [51] or synthetic lethal analysis by microarray (SLAM) [52] experiments. In genetic interaction networks, nodes represent the genes and edges connect two genes if the observed phenotypes after the mutation of genes are unexpected. These networks are undirected. Edges can be weighted based on the Z-scores of the observed phenotypes.

Recently, a special case of genetic interaction network has been proposed, the genetic interaction similarity networks (GIS) [53]. These networks link genes with similar genetic interaction profiles. It has been demonstrated that genes belonging to the same pathway or biological process tend to share similar profiles of genetic interactions [54]. Hence, GIS networks highlight genetic relations between diverse biological processes capturing the inherent functional organization of the cell [53]. The similarities between the genetic interaction profiles of the genes are quantified by computing Pearson's correlation coefficient (PCC) between them. In GIS networks, nodes represent the genes and edges connect two genes if the similarity between their genetic interactions profiles is higher than a given threshold. These networks are undirected. Edges can be weighted based on the PCC of the genes' genetic interaction profiles.

The main public databases for obtaining genetic interaction data are BioGRID [41] and Flybase [55]. Contrary to PPIs, genetic interactions have not been deeply studied in all organisms. *Saccharomyces Cerevisiae*, *Schizosaccharomyces Pombe*, *Drosophila Melanogaster*, and *Caenorhabditis Elegans* are the only well-studied organisms for genetics interactions in the last years. Genetic interaction networks are not analyzed in the scope of this dissertation. Regarding GIS data, the only available data source is the study by Costanzo *et al.* [53]. In this study, Costanzo *et al.* examine 5.4 million gene-gene pairs generating gene interaction profiles for approximately 75% of the genome of *Saccharomyces Cerevisiae.*

In Chapter 3 of this dissertation, we apply our graphlet-based method to the unweighted and probabilistic GIS network of *Saccharomyces Cerevisiae.* The GIS data used to construct these networks was obtained from the study by Costanzo *et al.* [53]. The construction and acquisition of these networks are detailed in Chapter 3 of this dissertation.

Gene co-expression (CoEx) networks: These networks represent the correlation of the expression of the genes over time. Two genes are linked if their expression is significantly correlated over time [26]. CoEx networks are built on gene-expression databases. These databases contain different expression profiles that show how gene expression is perturbed by developmental stages, different growth conditions, stress, disease, and specific mutations [56]. The correlation of expression profiles for a set of genes across different experimental conditions suggests that the sets of genes are functionally related. Usually, PCC is used as a measure of gene co-expression. A PCC between two genes of "1", indicates a strong relationship in an aspect of gene expression regulation, and "0" indicates no relationship [57]. There are many genes that only have low PCC co-expressed gene pairs, but are still functionally relevant. Thus, PCC is not usually applied to build co-expression networks. Instead, the PCC is used to compute the mutual rank between the genes. This measure keeps the weak but significant gene co-expression from being discarded [57]. The mutual rank between two genes, A and B, is computed by first ranking gene B in relation to gene A, and then ranking gene A in relation to gene B, using PCCs (this is called the correlation rank). Then, the mutual rank between genes A and B is obtained by geometrically averaging their correlation ranks. Finally, different thresholds on the mutual rank between the genes are applied to choose those genes that are co-expressed [10, 58]. Gene expression profiles are generally obtained from DNA microarray experiments. In gene co-expression networks, nodes represent genes and edges connect two genes if their mutual rank is higher than a given threshold. These networks are undirected. Edges can be weighted based on PCCs or the mutual rank between the expression profiles of the genes.

The main public databases for obtaining gene co-expression data are COXPRESdb [59] and GeneFriends [60]. Specifically, COXPRESdb provides gene co-expression data for 11 species. In Chapter 3 of this dissertation, we apply our graphlet-based method to the unweighted and probabilistic CoEx networks of *Saccharomyces Cerevisiae.* The gene co-expression data used to build these networks was obtained from COXPRESdb.

Metabolic networks: These networks model the metabolism of a cell. The collection of all metabolic reactions that occur in a cell forms a metabolic network [61, 62]. A metabolic reaction transforms one metabolite (substrate) into another (product) and it is catalyzed by an enzyme (protein). Metabolites can be small molecules, such as glucose, or larger molecules, such as polysaccharides. A metabolic reaction involves at least two metabolites (substrate and product) and an enzyme. Metabolic reactions can be divided into two main categories: catabolic reactions, which involve the breakdown of metabolites to release energy (e.g., glucose to pyruvate by cellular respiration), and anabolic reactions, which involve the synthesis of molecules using energy (e.g., synthesis of glycogen from glucose) [63]. These reactions are often represented by directional edges since they represent the chemical conversion of the substrate to a product. However, most metabolic reactions are reversible, e.g., glycogenesis (synthesis of glycogen from glucose) and glycogenolysis (breakdown of glycogen to glucose). Hence, these reactions can also be represented by undirected edges. Since metabolic reactions involve two biological entities (metabolites and enzymes), bipartite networks are a natural choice to represent metabolic reactions [64]. In a bipartite metabolic network, there are two distinct sets of nodes: one set representing metabolites and the other set representing enzymes. Edges exist only between nodes from different sets (i.e., enzymes to metabolites) but not between nodes from the same set (i.e., enzymes to enzymes) [64]. That is, an enzyme is linked to metabolites that are inputs and outputs of the reaction that it catalyzes. Edges in these bipartite networks can be either directed or undirected. Also, there exist metabolite-centric and enzyme-centric network representations. In a metabolite-centric network representation, nodes represent metabolites and edges connect two metabolites if there is an enzyme that mediates the reaction that transforms one metabolite into the other. In an enzyme-centric network representation, nodes represent enzymes and edges connect two enzymes if they catalyze reactions involving the same metabolite. The choice of the metabolic network representation variate depending on the problem being studied.

The main databases for obtaining metabolic data are the Kyoto Encyclopedia of Genes and Genomes (KEGG) [65], GeneDB [66], metaTIGER [67], and ERGO [68]. Analyses of metabolic networks are out of the scope of this dissertation.

Transcriptional regulation networks: These networks model gene expression regulation. The regulation of the gene expression typically involves the binding of proteins (known as transcription factors) to *cis*-regulatory sequences, such as enhancers and promoters, in the genome [69]. Transcriptional regulation networks are a simplified representation of this phenomenon, where nodes represent genes and two nodes, A and B, are connected if the protein product of gene A (transcription factor A) regulates the transcription of B. These networks are directed, since the relation between the nodes is asymmetric, e.g., gene A controls the expression of Y, but Y may not influence the expression of A. One limitation of this representation is that there is no difference between repression and enhancement of gene expression. This is usually remedied by using weighted networks to represent this phenomena [26]. Techniques to detect the gene expression regulation by transcription factors include chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-Seq), chromatin immunoprecipitation followed by microarray analysis (ChIP-chip), and electrophoretic mobility shift assay (EMSA) [70].

The databases that contain transcription regulation information are KEGG [65], EcoCyc [71], GeneNet [72], Reactome [73], RegulonDB [74], JASPAR [75], Phospho.ELM [76], The PHOsphorylation SIte DAtabase (PHOSIDA) [77], TRANSPATH [78], and The Small Molecule Pathway Database (SMPDB) [79]. Analysis of transcription regulatory networks is out of the scope of this dissertation.

Cell signaling networks: These networks model the complex signaling mechanisms inside the cell, or between a cell and the extracellular environment [80]. Regulation of cellular processes allows cells to maintain homeostatic balance and make decisions as to whether to divide, differentiate, or die [81]. In each case, the cell responds to chemical, mechanical, or electrical signals. These signals can be intracellular, when they occur within a single cell, and intercellular, which occurs between cells. Cells respond to these signals through signaling pathways. Signaling pathways transduce the signal from a cell membrane receptor to the nucleus. This transduction is achieved by an ordered sequence of reactions. In short, after the first protein (usually a receptor) in the pathway receives a signal, it activates another protein. This process is repeated through the entire signaling pathway until the signal arrives at the nucleus. As a result, transcription happens in the nucleus enabling the cell to react [81]. A cell signaling network captures this transient sequence of interactions between proteins that transduce a signal from the first protein in the pathway to the nucleus. All signaling pathways of a cell from its signaling network. The nodes of these networks are proteins and the directed edges connecting these proteins represent the signals propagated from one protein to another. These networks are used for modeling the cellular responses to different internal and external stimuli by means of pathways.

The databases that contain cell signaling information are TRANSPATH [78] and the Microbial Signal Transduction (MiST) database [82]. Cell signaling networks are not analyzed in the scope of this dissertation due to their limited availability.

Other biological networks: Biological networks also include *neural networks* modeling the connections of the neurons in the brain [83], *disease-genes networks* modeling the genes that are affected by the same diseases [84], *drug-target networks* modeling the proteins that are targeted by the same drugs [85], and *protein structure networks* modeling the tertiary (3D) structure of a protein [83]. These networks are not analyzed in the scope of this dissertation.

2.2 Concepts on Networks

A graph, or a network G, is a mathematical representation of a set of objects and the relations among them. A network is formally denoted by $G = (V, E)$, where V is the set of nodes that represent the objects, and E is the set of edges that define the relations among the elements of V. An edge $e \in E$ corresponds to a pair of nodes $(u, v) \in V^2$. Edge $e = (u, v)$ is said to join the nodes u and v, which, in turn, are called the *ends* of edge e. An edge is incident to both its ends and the nodes joined by an edge are called *adjacent*. If both ends on an edge correspond to the same vertex, $e = (v, v)$, this edge is called a *loop*. A network is *undirected* if the edges of the network have no orientation; i.e., $\forall(u, v) \in E : (u, v) = (v, u)$. These networks represent a symmetric interaction between the nodes, for instance, a physical contact between two proteins in a PPI network. On the other hand, a network is *directed* if the edges of a network have direction; i.e., $\forall(u, v) \in E : (u, v) \neq (v, u)$. Directed edges are used to represent asymmetric interactions between the nodes, for instance, the regulation of the expression of gene A by the protein product of gene B. Directed edges are represented by arrows and a directed edge originates from at the *source node* and points to the *target node*. If two nodes are joined by two or more edges, then we say that there is a *multi-edge* between those two nodes. Edges in directed and undirected networks can be *weighted* or *unweighted*. A weighted network has a weight, or score, associated with each edge. Formally, a weighted graph is defined by its nodes and edges sets (V and E), as well as a set of possible weights Ω and a function $\omega : E \rightarrow \Omega$ assigning a weight to each edge. Weighted networks can be used to include the probability of the existence of a given interaction between two nodes (in this case $\Omega = [0, 1]$). This special case of weighted networks is called *probabilistic* *networks* and is widely used in biology to represent the uncertainty associated with molecular interactions (we focus on these networks in Chapter 3). As opposed to weighted networks, unweighted networks do not have weights assigned to the edges. Thus, in an unweighted network, each connection between nodes is either present or absent, which is why these networks are also referred to as *binary networks*. Finally, a *simple network* is an undirected and unweighted network that contains no loops ($\forall v \in V : (v, v) \nexists E$), and no multiple edges (we focus on these networks in Chapters 4 and 5). The different networks introduced in this section are illustrated in Figure 2.1.

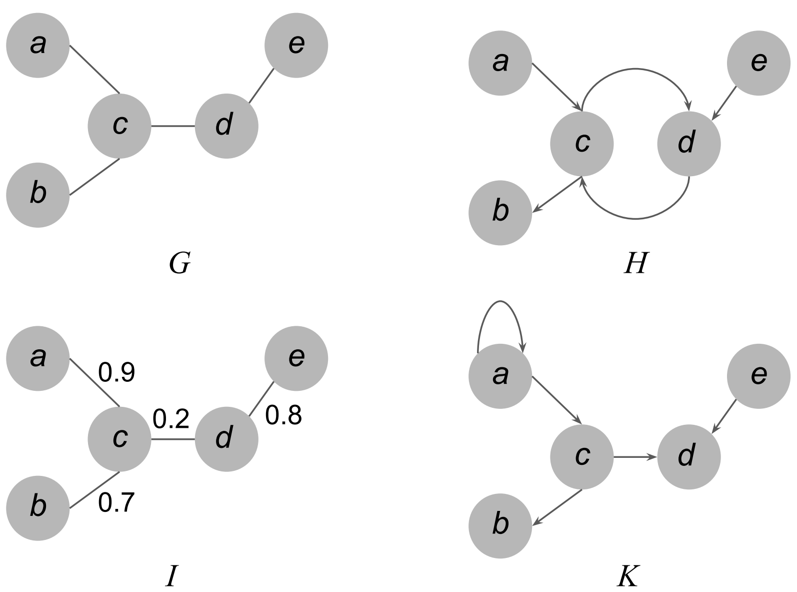


Figure 2.1. Examples illustrating: a simple network G; an unweighted directed network H with a multi edge between nodes c and d; an undirected weighted network I; an unweighted directed network K with a loop in node a.

The *degree* of a node v corresponds to the number of edges incident to v and is denoted by $d(v)$. The *neighborhood* of node v, $N(v)$, is the set of nodes that are adjacent to v. Consider network G of Figure 2.1: node c has degree 3 ($d(b) = 3$) and the nodes in its neighbourhood are a, b, and d. A network $H(V', E')$ is a *subnetwork* of $G(V, E)$ if it contains a subset of nodes of G, $V' \subseteq V$, and a subset of edges connecting those vertices, $E' \subseteq E$. A subgraph $H(V', E')$ of G is *induced* if it contains all the edges in G between the nodes in V'; otherwise, it is a *partial* subgraph. Two networks $G(V, E)$ and $H(V', E')$ are *isomorphic* if there exists a function $f : V \to V'$ such that $vu \in E$ if and only if $f(v)f(u) \in E'$. The function f is a *bijection*, which means that f is a one-to-one mapping of the nodes in G to the nodes of H. In other words, we say that two networks are isomorphic if their nodes and edges can be mapped one-to-one onto each other.

A *walk* in an undirected network $G(V, E)$ is a sequence of nodes of V such that the consecutive vertices are adjacent. A walk $w = v_0 ... v_n \in V^{n+1}$ is *closed* if the starting and ending nodes are identical, i.e., if a walk starts and ends at the same node ($v_0 = v_n$), otherwise it is *open*. The length of a walk w corresponds to the number of edges it contains; here w is of length n [86]. A v-u walk, where $(v, u) \in V^2$, is a walk that starts at node v and ends at node u. The shortest v-u walk is a v-u walk of minimum length. Hence, a path between nodes v and u is an ordered set of edges that need to be traced for reaching node v without visiting any node more than once. Similar to the shortest walk, the shortest path is a path of minimum length. For example, consider network G of Figure 2.1: the shortest path between node a and d is the path the a-c-d. If a walk never visits any node in V more than once, it is called *path*. A network is *connected* if there exists a path from every node to every other node in the network, otherwise, it is *disconnected*. A *random* walk is a walk that starts in a certain node and proceeds from node to node along the edges by randomly choosing the edge to be followed [87]. This process continues until a stopping criterion is met, such as a certain number of steps or reaching a specific node. In an unweighted network, the probability of choosing a certain edge is equal across all available edges. Hence, the transition probability from node v_i to node v_j, denoted by $P(v_i, v_j)$ is given by:

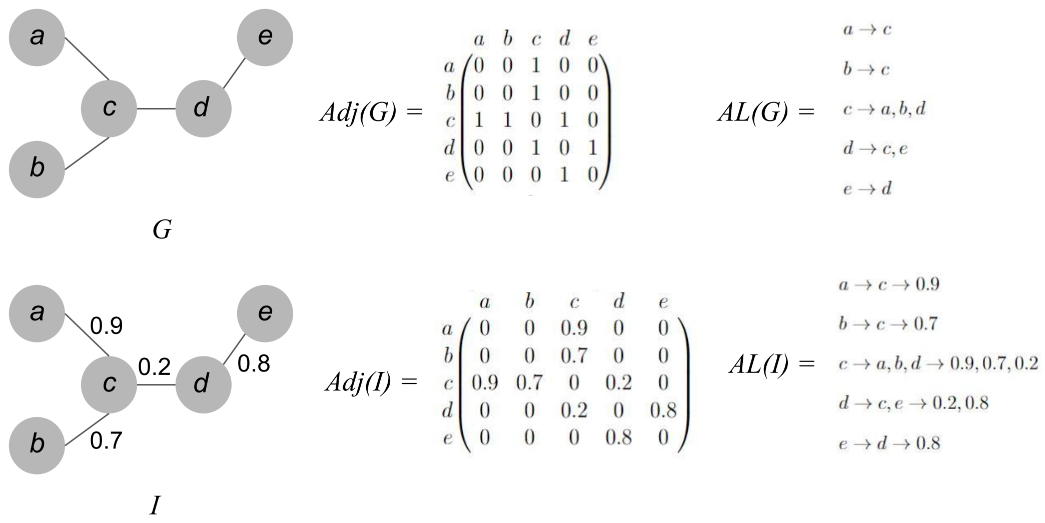
$$P(v_i, v_j) = \begin{cases} \frac{1}{d(v_i)} & \text{if } \ (v_i, v_j) \in E \\ 0 & \text{otherwise} \end{cases} \tag{2.1}$$

Where $d(v_i)$ denotes the degree of node v_i. Similar notions are defined for directed networks with the difference being that the direction of the edges needs to be followed, i.e., a walk $w = v_0 ... v_n \in V^{n+1}$ if defined in a directed network if for any consecutive pairs of nodes, there is a directed edge with source v_{i-1} and target v_i, with $i \in [1, n]$.

2.2.1 Network Representations

There are two standard ways to represent a network $G(V, E)$, with V nodes and E edges: Adjacency list and adjacency matrix [88]. Either way applies to both directed and undirected graphs. The *adjacency list* of $G(V, E)$ consists of an array AL of $|V|$ lists, one for each node in V. For each v in V, the adjacency list $AL[v]$ contains all the nodes such that there is an edge $(v, u) \in E$. That is, $AL[v]$ consists of all nodes adjacent to v in G. For representing weighted networks, an extra list of edge weights should be kept for each node. The *adjacency matrix* of $G(V, E)$ is a $|V| \times |V|$ matrix Adj, where $Ajd[v, u]$ is a non-zero value when nodes u and v are connected, and equal to 0 otherwise. This matrix is symmetric when G is undirected. For representing weighted networks, the edge weights can be encoded in the value of $Adj[u, v]$. Figure 2.2 presents the adjacency lists and adjacency matrices of an unweighted undirected network and a weighted undirected network. The choice of which representation to use depends on the network and the task at hand. If a network is sparse (as most biological networks are), the adjacency list representation is more efficient to store and traverse the network.

Recently, an alternative way of representing a network has been introduced, the *Positive Pointwise Mutual Information* (PPMI) matrix [89]. The PPMI matrix originated in the NLP field to represent the association between words in a lexical corpus (e.g. a passage or a document). In NLP, rows and columns of this matrix represent words and their cells quantify if two words co-occur more frequently than expected at random assuming they are independent. Similarly, the PPMI representation of network $G(V, E)$ is a $|V| \times |V|$ matrix $PPMI$, where rows and columns represent nodes and entry $PPMI[v, u]$ quantify if node v and u co-occur more frequently in



$$Adj(G) = \begin{array}{c|ccccc} & a & b & c & d & e \\ a & 0 & 0 & 1 & 0 & 0 \\ b & 0 & 0 & 1 & 0 & 0 \\ c & 1 & 1 & 0 & 1 & 0 \\ d & 0 & 0 & 1 & 0 & 1 \\ e & 0 & 0 & 0 & 1 & 0 \end{array}$$

$$AL(G) = \begin{array}{l} a \to c \\ b \to c \\ c \to a, b, d \\ d \to c, e \\ e \to d \end{array}$$

$$Adj(I) = \begin{array}{c|ccccc} & a & b & c & d & e \\ a & 0 & 0 & 0.9 & 0 & 0 \\ b & 0 & 0 & 0.7 & 0 & 0 \\ c & 0.9 & 0.7 & 0 & 0.2 & 0 \\ d & 0 & 0 & 0.2 & 0 & 0.8 \\ e & 0 & 0 & 0 & 0.8 & 0 \end{array}$$

$$AL(I) = \begin{array}{l} a \to c \to 0.9 \\ b \to c \to 0.7 \\ c \to a, b, d \to 0.9, 0.7, 0.2 \\ d \to c, e \to 0.2, 0.8 \\ e \to d \to 0.8 \end{array}$$

Figure 2.2. An illustration of networks G and I and their adjacency matrices $Adj(G)$ and $Adj(I)$, and their adjacency lists $AL(G)$ and $AL(I)$. In $Adj(G)$ the 1s represent the existence of an edge from the node in a row to the node in the column. In $Adj(G)$ the non-zero values represent the edge weight between the node in a row to the node in the column.

a random walk than expected by random. In opposite to the previous network representations, the PPMI matrix captures high-order proximities between the nodes in the network; hence, the PPMI matrix is considered a richer representation than the adjacency matrix and the adjacency list (in Chapter 3, we investigate if the extra information encoded in the PPMI can be exploited to capture more biological information from molecular networks). We give more background about the PPMI network matrix representation in section 2.7.1 of this Chapter.

2.2.2 Network Properties

Measures of network structure, also called network properties, are historically and roughly divided into two categories: *local* and *global* ones. In general, if they involve the full network, then they are global; if they involve only one node and its neighborhood, then they are local [26]. Global network properties give an overall view of the network with respect to all nodes and edges, and local network properties evaluate the topology of a network in terms of its subgraphs. In the rest of this section, we describe the most basic measures of global and local network properties, and we illustrate them on the example network, G, that is shown in Figure 2.2. The most complex measures are covered in section 2.6 of this Chapter.

The most elementary local property of a node is its degree (or connectivity). The *degree* of a node is the number of links that the node has to other nodes in the network. For example, the degree of node a in G is 1, and the degree of node c is 3. In molecular networks, it has been shown that the degree of a node is connected to its biological essentially. For instance, proteins with a high degree in the human PPI network are connected with multiple diseases, such as cancer [90]. Moreover, mutations in genes with a high degree in the molecular network cause the death

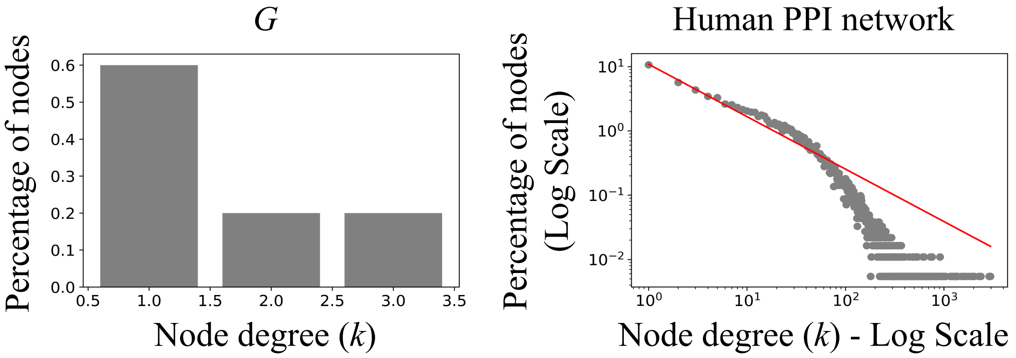


Figure 2.3. Degree distribution of network G and the human PPI network (obtained from BioGrid database v.4.2.191). For the human PPI network, the power-law $\gamma = 0.93$ is also presented.

of the cell [91]. However, these observations may not be universal since they were shown not to hold for more complete PPI networks such as the one of *Saccharomyces Cerevisiae* [92]. This demonstrates that simple topology measures are not enough to capture biological signals, i.e., more sophisticated measures are needed (detailed at the end of this section and in the last two sections of this Chapter).

The *average degree* of a network is the arithmetic average of the degrees of all nodes in the network. The average degree of G is equal to $(1 + 1 + 3 + 2 + 1)/5 = 1.6$. The *degree distribution* $P(k)$ of a network gives the probability that a selected node has exactly degree k. $P(k)$ is obtained by counting the number of nodes $N(k)$ with $k = 1, 2...$ and dividing by the total number of nodes in the network. For instance the $P(1)$ of network G is 3 / 5 = 0.6. Figure 2.3 illustrates the degree distribution of G. Most biological networks have a degree distribution that approximates a *power law*, $P(k)$ k^{γ}, where the *degree exponent* $\gamma > 0[100, 101]$ indicates "proportional to." In a log-log plot, a power law distribution corresponds to a straight line with slope $-\gamma$. Networks with a power-law degree distribution are called *scale-free*. The name comes from the fact that power laws have the same forms at all scales. Scale-free networks share an important property: some nodes have a tremendous number of connections to other nodes (high degree), whereas most nodes have just a few [93]. These high-degree nodes are called *hubs*. In a scale-free network, the value of γ determines different properties of the system. The smaller it is, the more important role of the hubs is in the network. In contrast, for $\gamma > 3$ hubs are not relevant. For $2 > \gamma > 3$ there is a hierarchy of hubs, with the most connected hubs being in contact with a small fraction of all nodes [94]. Such networks display an unexpected degree of robustness, the ability of their nodes to communicate being unaffected even by extremely high failure rates. However, high-degree nodes (hubs) are open for targeted attacks which results in the overall failure of the network [95]. Figure 2.3 illustrates the degree distribution of the PPI network of human. The *cumulative degree distribution*, $P_c(k)$, is the probability that a randomly selected node has a degree smaller than k.

The *density* of a network describes the portion of all possible edges between the nodes that are actual edges in the network: *density* $= \frac{2|E|}{|V|(|V|-1)}$. The density of G is equal to $8/20 = 0.4$. If the number of edges is close to the maximal number of possible edges (density close to 1), the network is said to be *dense*. In contrast, a network is sparse if the number of edges is low in comparison to the number of possible edges (density close to 1). Most biological networks are sparse. For instance, the density of the human PPI network is 0.002 (obtained from BioGrid database v.4.2.191). The low sparsity of molecular networks is linked with a process of natural selection and evolution [96]. Sparsity allows the biological systems to have specialized interactions between molecules preventing cross-talk between different processes. Moreover, sparse molecular networks are less costly than dense ones, since they require fewer resources to maintain the interactions between molecules. Finally, although it may seem intuitive that dense networks are more robust, molecular networks have evolved to maintain specific wiring patterns that make them also robust.

To describe how spread the nodes in a network are, we use the *average path length* of the network and the *diameter* of the network. The average path length is measured by computing the shortest paths between all pairs of nodes and averaging them. For instance, the average path length is 1.8 for G. The diameter of the network corresponds to the maximum shortest path distance over all pairs of nodes (e.g., the diameter is 4 for G). Another network measure is the *clustering coefficient*, which is the probability that two neighbors of a node are linked by an edge. The clustering coefficient of a node v, C_v, is computed as:

$$C_v = \frac{2 \times T(v)}{deg(v) \times (deg(v) - 1)} \tag{2.2}$$

where $deg(v)$ is the degree of node v and $T(v)$ is the number of triangles through node v. The clustering coefficient is a measure of the degree to which nodes in a graph form transitive relations. For instance, all the nodes of G have a clustering coefficient of 0, since non of their neighbors are connected. The clustering coefficient can also be thought of as the ratio of the number of *triangles* containing node v to the number of triplets of nodes containing node v. The clustering coefficient takes values between 0 and 1, 0 if no edges connect any pair of nodes in the neighborhood of the node, and 1 if the neighborhood forms a complete network. The *average clustering coefficient* is the arithmetic average of the clustering coefficients of all nodes in the network and characterizes the overall tendency of nodes to form clusters. It is formally defined as: $\bar{C} = \frac{1}{n}\sum_{v \in V} C_v$. A $\bar{C}$. For a given network, the higher the $\bar{C}$ is, the more its nodes tend to form clusters. The $\bar{C}$ of most molecular networks is statistically significantly higher than that of an equivalent random network [97] (random models are introduced in section 2.3). For instance, the average clustering coefficient of the human PPI (from BioGrid database v.4.2.191) is 0.11, while the average clustering coefficient of a random network model with the same number of nodes and edges is 0.04. This indicates that the high clustering is a generic feature of biological networks and suggests a modular organization of such networks, i.e., the organization of its molecular entities into functional sub-units that are sparsely interconnected [98, 99, 99].

Centrality measures the significance of a node within a network. There are different well-known centrality measures. The simplest centrality definition is the *degree centrality.* which is defined as the degree of a node. The degree centrality assumes that important nodes of a network are connected to many other nodes (have a high degree). In network biology, it has been shown the degree of a node in a molecular network is an indicator of its biological relevance. For instance, the dysfunction of proteins with a high degree in the *Saccharomyces Cerevisiae* PPI network is lethal for the cell. In contrast, the mutation of proteins with a low degree does not alter the functioning of the cell [100]. *Closeness centrality*, $C_c(v)$, is another centrality measure that evaluates how closely a node is connected to all other nodes in the network. It is computed as:

$$C_c(v) = \frac{1}{\sum_{v \in V dist(v,u)}} \tag{2.3}$$

where $dist(v, u)$ is the number of edges in a shortest path between nodes v and u. For instance, in G, the closeness centrality of nodes c and e of G are respectively 0.8 and 0.4; higher values representing more central (important) nodes. In network biology, this measure of centrality has been used to identify important metabolites, study the evolution of metabolic networks topology, and compare the metabolic networks of different species [101, 102]. *Betweenness centrality*, $C_b(v)$, is a more detailed centrality measure, for a given node, it evaluates the number of shortest paths in the network that pass through the node. The betweenness centrality is computed as:

$$C_b(v) = \sum_{s \neq t, s \neq v, v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{2.4}$$

where σ_{st} is the total number of shortest paths between nodes s and t and $\sigma_{st}(v)$ is the number of shortest paths between nodes s and t that pass through node v. In G, the betweenness centrality of node a is 0 since none of the shortest paths in the network pass through it. In contrast, the betweenness centrality of node c is 0.83, which highlights its importance in connecting the rest of the nodes in the network. Nodes with high betweenness are known to be bottlenecks in the network, i.e. all the information has to pass through them. In network biology, this measure of centrality is connected with important regulation processes. For instance, proteins with high centrality in metabolic networks act as switchboards for regulating cell metabolism by controlling the trafficking of intracellular metabolites [102].

The *eccentricity* of a node is the maximum of the shortest path distances between the node and all other nodes in the network. Hence, low eccentricity of a node means that the node is in the periphery of the network, while high eccentricity means that the node is “in the middle” of the network. The *eccentricity centrality* is computed as:

$$C_e(v) = \frac{1}{E(v)} \tag{2.5}$$

where $E(v)$ represents the eccentricity of node v. In G, the eccentricities of nodes a c are respectively 3 and 2, and the corresponding eccentricity centralities are 0.33 and 0.5. In network biology, this measure of centrality has not yet been linked with the importance of a molecule in a molecular network. For instance, this centrality measure is not capable of distinguishing between essential and not essential proteins of *Saccharomyces Cerevisiae* from its PPI network [103].

So far, we have introduced the most commonly used centrality measures and we have shown examples of their use in network biology. There exist other centrality measures that are not discussed in this section. They include *eigenvector centrality*, which identifies nodes that are connected to important nodes as measured by a relative score associated with each node, *subgraph centrality*, which measures the participation of a node in all subgraphs in a network, and the *K-Shell decomposition*, which divides the nodes of a network into groups based on their degrees.

Finally, the most sophisticated network measures are *network motifs* and *graphlets*. Network Motifs (or simply called motifs) are small partial subgraphs that occur in a network at frequencies statistically significantly higher than expected by random according to a certain random model [104, 105] (detailed description of the classical random models is provided in section 2.3). Several studies have investigated the motifs of biological networks, e.g., in signaling networks and gene regulatory networks [106, 107, 105], finding that they are functional. Moreover, it has been shown a correlation between specific motifs and biological functions [108]. However, Artzy-Randrup *et al* [104] criticizes the dependence of network motifs on the choice of the random model used in their identification. They claim that most biological networks do not have a random topology, i.e., comparing the frequency of the subgraphs of input with the frequencies in the random network contains some bias as the random network model is not a good model for the real network. Since network comparison is computationally intractable, finding a random network model that accurately mimics a specific network's topology is difficult, i.e., this problem can not be easily solved. On the other hand, motifs are partial subgraphs (they do not include all the nodes and edges of the original network). Because of it, their ability to capture topological similarities is not as strong as that captured by the induced subgraphs. Hence, Przulj*et al* [3] introduced graphlets; that is small, induced, connected, and non-isomorphic subgraphs of a large network. In comparison to network motifs, graphlets are independent of any random network model. Moreover, they are more powerful in capturing the network topology because they are defined as induced subgraphs of a network. Actually, graphlets and graphlet-based methods are considered one of the best approaches to capture the local wiring patterns of the nodes and uncover information from biological networks. In section 2.6 of this Chapter, we focus on graphlet and graphlet-based approaches.

2.3 Network Models

A network model is a collection of rules for generating random networks with specific topological properties. The study of network models aims at understanding the functional mechanisms in the real-world system. Network models have been applied for multiple purposes; e.g., for identifying the overrepresented subgraphs (network motifs) in the network [109, 104, 106], investigating dynamic processes on networks, such as diseases spreading [110, 111], denoising biological networks [112], predicting interaction [112], or guiding interactome detection experiments [113].

In this section, we present some of the most widely used network models in the representation of biological networks namely Erdös-Renyi Random networks, Scale-free networks, geometric networks, and Stickiness Index-based networks.

2.3.1 Erdös-Renyi Random networks

The first random model was introduced by Erdös and Renyi [114]. There are two related ways of computing an Erdös-Renyi (ER) random network of n nodes and m edges. The first is to pick a pair of nodes uniformly at random to form the m

edges. The second is to start with a pair of nodes and, for each possible pair of nodes, add an edge connecting the pairs with a probability p, with $p = \frac{m}{\binom{n}{2}}$. This creates a network with approximately $\frac{pn(n-1)}{2}$ randomly placed edges. ER networks have many proven topological properties [115]. The node degrees follow a Poisson distribution, which indicates that most nodes have approximately the same number of links. The average degree equals $(n1) \times p$. The clustering coefficient is independent of a node's degree and low (since the edges in the network are distributed uniformly at random). The diameter is low, in an order of $log(n)$.

An extension of ER model is the *Generalized Random Model* (ER-DD). In this model, the network is generated as for ER networks, but the degree distribution is forced to match a specified distribution (e.g., the distribution of a molecular network) through the use of the "stubs" method [116]. To generate an ER-DD network of n edges, we first associate to a node a number of stubs that correspond to its degree. We add edges by randomly picking node pairs that have available stubs and connecting them. When an edge is added, the number of available stubs of the connected nodes is decreased by one. Therefore, the degree distribution of these models matches with the given distribution when all stubs are filled. Note that this procedure may not terminate, as it may encounter impossible constraints for adding new edges. In that case, it needs to be restarted until it produces a model network. Similar to ER models, due to the random distribution of the edges, the clustering coefficient of ER-DD models is low [94].

In Chapter 3, we extend the ER model to probabilistic networks and use them to assess the performance of our new graphlet-based method in detecting different network topologies. An illustration of an ER network can be seen in Figure 2.4.

2.3.2 Scale-free Networks

Scale-free networks are characterized by their power-law degree distributions (detailed in section 2.2.2). Barabasi and Albert preferential attachment model (BA) is the most well-known scale-free network model [94]. The BA model of scale-free networks starts from a small seed network $G_0(V_0, E_0)$ and at each time point t a new node v is added to the network, which is connected to an already existing node u with a probability of:

$$\frac{d(u)}{\sum_{w \in V_t} d(w)} \tag{2.6}$$

where $d(u)$ denoting the degree of node u and V_t being the set of nodes of the network at time t. As can be seen, new nodes tend to be connected to those nodes that already have a large number of edges, this process is known as the "rich-gets-richer principle." The clustering coefficient and diameter of BA networks are low. These networks are good models of the World Wide Web [26].

Another scale-free model is the *Scale-free Gene Duplication and Divergence model* (SF-GD) [117]. SF-GD imitates the gene duplication and mutation events for the scale-free network generation. Similarly to the BA model, these networks are generated iteratively from a small seed network (containing at least an edge). Each iteration can be decomposed into two main steps: duplication and divergence. In the duplication step, an existing node v in the network is selected uniformly at random, and a new node u that is connected to all neighbors of v is added to the network. The selected node v and the new node u are also connected with probability p. In the divergence step (also known as the mutation step), for each node connected to v and u, one of the two edges (to v or u) is chosen and deleted with a probability q.

In Chapter 3, we extend the BA model to probabilistic networks and use them to assess the performance of our new graphlet-based method in detecting different network topologies. An illustration of a BA network can be seen in Figure 2.4.

2.3.3 Geometric Networks

Geometric networks incorporate spatial information into the representation of a complex system.

In a *geometric random graph* (GEO) [118], nodes are independently and uniformly distributed in a metric space and they are connected if the distance between them is smaller than or equal to a distance threshold, r (also called the radius). The radius is chosen to adjust the density of the network. GEO model can be altered based on the dimensionality of the metric space and the distance measure used among the nodes (e.g., euclidean or cosine distances). These graphs have Poisson degree distributions, but there exist variants of geometric networks in which this is not the case. Their clustering coefficient is large and their diameters are small. It has been shown that GEO networks model PPI networks much better than other commonly used network models despite the fact that parts of degree distributions of PPI networks often follow a power-law [119, 120].

Inspired by the evolutionary dynamics that are observed in nature, Przulj *et al.* [119] adapt geometric models to imitate the gene duplication and mutation events that occur during the evolution of a biological network defining *Geometric graph with gene duplications and mutations* (GEO-GD) [119]. Instead of randomly distributing the nodes uniformly at random in space, the GEO-GD model, the model starts with a small network and adds nodes to mimic gene duplication and mutations. A node is randomly selected and duplicated. Then, the duplicated node is placed uniformly at random within a distance of $2r$ of the original node, where r is the distance threshold introduced earlier. The duplication process is repeated until a certain number of nodes, specified by the user, is reached. The nodes are then connected if they are within a radius r. GEO-GD networks have been shown to fit real PPI networks better than the traditional GEO networks [119].

In Chapter 3, we extend the GEO model to probabilistic networks and use them to assess the performance of our new graphlet-based method in detecting different network topologies. An illustration of a GEO network can be seen in Figure 2.4.

2.3.4 Stickiness Index Based Networks

As the GEO-GD model, another biologically motivated network model is the *Stickiness Index Based Network Model* (STICKY) [121]. This model is built on two main assumptions: high-degree proteins have many binding domains and these domains are highly involved in physical interactions, and a pair of proteins are more likely to interact if they both have high degrees (many domains). Based on these assumptions, the idea is to assign a "stickiness" index to a protein proportional to its degree. The stickiness index of a node v is defined as:

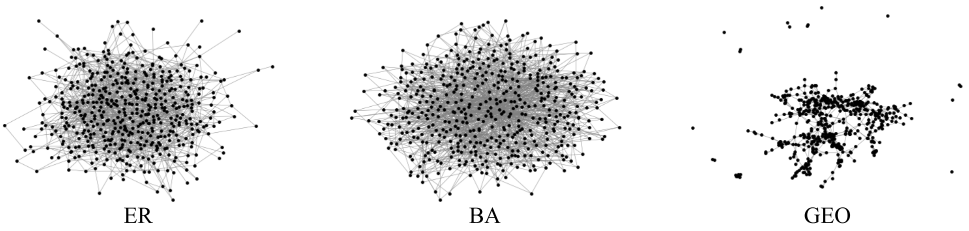


Figure 2.4. Illustration of networks that have 500 nodes and 1% edge density and generated from three network models. The corresponding models are Erdös-Renyi random network (ER), Barabasi and Albert preferential attachment model (BA), and geometric random graph (GEO).

$$\theta_v = \frac{d(v)}{\sqrt{\sum_{u \in V(G)} d(u)}} \tag{2.7}$$

where $V(G)$ is the set of all nodes in network G and $d(v)$ is the degree of node v. Nodes are connected with a probability proportional to the product of their "stickiness" indices. This model is applied to investigate the topological structure of PPI networks [121]. In this dissertation, we do not consider STICKY-based model networks.

2.4 Network Comparison

The network comparison problem includes three sub-problems: Network topology comparison, network alignment, and network querying. Briefly, the network topology comparison problem focuses on defining a distance measure that evaluates the overall topological similarity between two networks. The network alignment problem aims to produce a mapping between the nodes of two networks such that the correspondence between the edges of the two networks is maximized. Finally, the network querying problem searches for a small topological pattern in a large graph. In the third Chapter of this dissertation, we face this problem by introducing a new graphlet-based method to compare synthetic probabilistic networks and real-world molecular networks modeled as probabilistic networks.

The easiest approaches to compare two networks are based on comparing the global topological properties that are introduced in section 2.2.2. Some of these global properties (i.e., average degree, average clustering coefficient, and diameter) can be directly compared by taking their absolute difference. For those global properties that are in form of distributions (e.g., degree distribution), the most direct approach is to compute the Euclidean distance of the two distributions. Given two distributions d_i and d_j, the Euclidean distance $D(d_i, d_j)$ is computed as:

$$D(d_i, d_j) = \sqrt{\sum_{k=0}^{max(d_i,d_j)} (d_i(k) - d_j(k))^2} \tag{2.8}$$

The distributions are usually normalized before computing the Euclidean distance in order to highlight a specific part of the distribution. An alternative approach to compare distributions is by the use of standard statistical tests such as Kolmogorov-Smirnov [122] or Mann-Whitney U [123]. Other approaches to compare networks are based on more complex measures. Among them, the leading edge method to compare the topological similarity between networks is the Graphlet Correlation Distance (GCD) [4]. To compute this distance, we have to obtain the graphlet correlation matrices (GCMs) of the networks that we want to compare. These matrices summarize the topology of a network of any size into an 11×11 symmetric matrix with values between -1 and 1. Then, the GCD between two networks, $G1$, and $G2$, is defined as the Euclidean distance of the upper triangle values of their GCMs. The GCD outperforms other measures when applied to synthetic and real networks. Moreover, it has been broadly applied to several real-world problems [4]. In section 2.6 we give details about this graphlet-based distance (details about GCM and GCD can be found in section 2.6.3 of this Chapter).

2.5 Network Clustering and Functional Enrichment Analysis

One of the most commonly used approaches in network biology is the clustering and enrichment analysis approach. This approach is divided into two different steps: (1) the clustering of the nodes, and (2) the functional enrichment analysis of the clusters. *Network clustering* involves grouping the biological entities in the network into clusters based on a similarity metric (e.g., the local topological similarity of the nodes or distance of their vectorial representations in the embedding space). On the other hand, *functional enrichment analysis* evaluate the biological significance of the clusters by assessing the statistical over-representation of specific biological functions (represented by functional annotations) within the clusters. Together, these approaches are powerful tools in various scenarios, such as finding the link between the local topology of the nodes and their biological functions [2], the prediction of protein functions [124], or the identification of new oncogenes [125]. In this section, we first give an overview of the most used network clustering algorithms. Then, we focus on the functional enrichment analysis of the clusters and provide a description of the most used functional annotations databases. We end this section by describing methods to summarize large lists of functional annotations.

2.5.1 Network Clustering

Network clustering is one of the most popular approaches for analyzing the topological and functional properties of biological networks. The goal of network clustering is to identify clusters of nodes in the molecular network that exhibit significant clustering properties. Network clustering methods usually require a topological measure of the nodes in the network, e.g., their local wiring patterns, to cluster them based on their similarity. These clusters uncover topological and functional modules of the network. Moreover, they enable the prediction of the functions of unknown molecules by assigning to it the function of another molecule that belongs to the same cluster [124]. Depending on the way the biological entities are clustered, network clustering methods can be broadly divided into two categories: hard clustering approaches and soft clustering methods. *Hard clustering* approaches assign each node to only one cluster, i.e., there are no overlapping clusters. In contrast, *soft clustering* methods allow each node to be assigned to more than one cluster at the same time, i.e., there are overlapping clusters.

Network clustering methods include several clustering algorithms. One of the most well-known hard clustering algorithms is *k-means.* This method groups the nodes into k clusters based on their distance in a d-dimensional Euclidean space. The algorithm starts by initializing k cluster centers randomly in the space. Then, for each node, the algorithm computes its Euclidean distance to all the cluster centers and assigns the node to the closest one. Thereafter, the centers are updated based on the nodes assigned to each cluster. This process of assigning nodes to the clusters and updating the cluster centers is iteratively repeated until convergence. In biomedicine, k-means has been widely applied for several tasks, such as clustering gene expression data [126] or clustering proteins based on their topological structure [127]. However, its application to the analysis of biological networks is limited. One reason is that the algorithm requires the d-dimensional Euclidean space [26]., i.e., it does not capture the non-linear relationships between the nodes. In addition, it is not robust against outliers, which limits its use on biological networks where a few highly connected nodes can have a significant impact on the results. A variant of this algorithm is *k-medoids.* K-medoids follow the same concept as k-means by iteratively assigning the nodes to the closest cluster center and updating the center. However, these methods use medoids (objects from the data set) instead of means to define cluster centers, i.e., are less sensitive to outliers. Additionally, k-medoids can handle non-numeric data and do not require a Euclidean space, making them more versatile compared to k-means [128]. Despite these benefits, k-medoids also present the same drawback as k-means, they both require the number of clusters, k, to be specified beforehand.

A totally different clustering algorithm is the *Hierarchical clustering.* This hard clustering method takes as input only a $N \times N$ distance matrix and dynamically builds a hierarchy of clusters based on the similarity of the objects. The process of building this hierarchy can be done in two different ways: *agglomerative* and *divisive.* In agglomerative clustering, the algorithm starts with each node as its own cluster and successively merges smaller clusters into larger ones until all nodes are in the same cluster or a stopping criterion is met. On the other hand, In divisive clustering, the algorithm starts with all nodes in the same cluster and splits it into smaller ones until each node forms its own cluster or a stopping criterion is met. Hierarchical clustering has two main advantages over k-means and k-medoids algorithms. First, it does not require the number of clusters as input. Second, this method not only provides a separation of the objects into clusters but also gives the hierarchical relationship between the clusters.

During this dissertation, we use different clustering algorithms for different tasks. In Chapters 3, 4, and 5, we apply the k-medoids algorithm to cluster genes and proteins based on different similarity measures, such as the similarity of their local wiring patterns in the PPI network or the cosine distances of their embedding vectors in the PPI embedding space. Also, in Chapter 4, we use hierarchical clustering to explore the hierarchical organization of the genes' embedding vectors in different tissues-specific PPI network embedding spaces.

2.5.2 Functional Enrichment Analysis

Functional enrichment analysis is a statistical method used to determine whether certain biological functions, pathways, or disease-related processes are over-represented in a set of genes or proteins. In network biology, this method is usually applied after clustering the nodes (e.g., genes or proteins) of the molecular network to interpret and understand the biological significance of the clusters. Functional enrichment analysis involves computing the probability that a biological function is enriched (over-represented) in a cluster. This probability is computed by using sampling without replacement strategy (also called the hypergeometric test) [129]. The probability that a biological function represented by annotation a is enriched in a cluster of genes C_j is computed as:

$$p(a) = 1 - \sum_{i}^{X-1} \binom{K}{i} \binom{M-k}{N-i} / \binom{M}{N} \tag{2.9}$$

where N is the number of genes in C_j (only annotated genes from the cluster are taken into account), X is the number of genes in C_j that is annotated with a, M is the number of annotated genes in the molecular network, and K is the number of genes in the molecular network that are annotated with a. The biological functions of genes (or, more properly, the proteins and non-coding RNA molecules expressed by the genes) are often represented by functional annotations. *Functional annotations* provide a computational representation of current scientific knowledge about the functions of genes from many different organisms making this information computationally tractable [130]. These annotations are usually organized in ontologies. An *ontology* is a formal representation of knowledge of a body of knowledge within a given domain. Ontologies usually consist of a set of concepts (functional annotations) with relations that operate between them. Over the past few decades, several curated ontology resources such as the Gene Ontology (GO) [130], Reactome [73], and KEGG [65] have been developed and demonstrated to significantly enhance the progress of biological and medical research [131]. Each of these ontologies provides different information about the functioning of the genes, e.g., the pathways in which they are involved or the cellular location in which their protein products perform their functions.

Among the previous resources, GO has the largest concepts and records. GO describes our knowledge of the biological domain with respect to three aspects: (1) Molecular Function, (2) Cellular process, and (3) Biological processes. *GO Molecular Function* (ML) annotations describe activities that occur at the molecular level, such as “catalysis” or “transport” without specifying where, when, or in what context the action takes place. *GO Cellular Component* (CC) annotations give information about locations relative to cellular structures in which a gene product performs a function, either cellular compartments (e.g., mitochondrion) or stable macromolecular complexes of which they are parts (e.g., the ribosome). *GO Biological Process* (BP) annotations provide a higher-level perspective of the larger processes that are accomplished by multiple activities. The structure of GO can be defined in terms of a directed acyclic network, where each GO annotation is a node, and the links between the annotations are edges between the nodes [130]. GO acyclic network is loosely hierarchical, with “child” annotations being more specialized than their “parent” annotations, but unlike a strict hierarchy, an annotation could have more than one parent annotation. Thus, the “level” of a GO annotation is the position of the annotation in the hierarchy, and it indicates the specificity of the function represented by the annotation. The GO annotations are designed to be species-agnostic, i.e., a single GO can annotate the genes of multiple organisms. Finally, each annotation has an evidence code to indicate how the biological function of a particular gene is supported. The evidence can be: experimental, phylogenetic, computational, author, curatorial, and automatically generated. To avoid noise, in this dissertation we only keep those annotations that are experimentally validated.

Finally, a common limitation of all curated ontology resources is their update rate. These ontologies aim to represent the current state of knowledge in biology, hence they are constantly revised and expanded as biological knowledge accumulates. However, the assignment of annotations to genes is a very time-consuming process because there are millions of genes mentioned in biomedical literature, and the database curators need to find evidence passages for each gene.

The main databases for obtaining functional annotations data are GO [130], KEGG [65], Reactome [73], and the National Center for Biotechnology Information (NCBI) [132]. In combination with clustering algorithms, during this dissertation, we use the functional annotations as follows. In Chapter 2, we use GO BP, GO ML, and GO CC annotations to evaluate the ability of our new graphlet-based method to capture biological information from different probabilistic molecular networks. In Chapter 4, we employ GO BP terms to assess the functional organization of different tissues-specific PPI embedding spaces (if they capture the organization of the cell) and explore their functional organization from a functional perspective. These annotations were obtained from the NCBI web server. Finally, in Chapter 5, we use GO BP terms to uncover biological information from the functional organization of different species-specific PPI embedding spaces. These annotations were also obtained from the NCBI web server.

2.5.2.1 GO terms Summary

One problem when interpreting the biological meaning of the clusters is that the list of GO terms that are statistically overrepresented in each cluster is usually large and highly redundant [133]. Redundancy between GO terms can be explained by their functional similarity. For instance, let GO_1 be the parent annotation of GO_2 in the GO hierarchy. Let GO_1 represent a generic function such as “Carbohydrate metabolic process” and let GO_2 represent a more specific function that is connected with the previous, such as “Carbohydrate biosynthetic process.” Then, the probability of having GO_1 and GO_2 enriched in the same cluster is high since they both are likely to annotate the same genes. Thus, the chances of having enriched functionally related annotations in the same cluster are high. The functional relationship and redundancy between functional annotations are usually evaluated by their *semantic similarity*, which is a measure of the similarity of their meaning. These functional relationships are then exploited to summarize large lists of GO terms (as we show below). Semantic similarity between GO annotations can be calculated based on various methods such as lexical and ontological measures.

The most known lexical measure of similarity is based on the TF-IDF (Term Frequency-Inverse Document Frequency). TF-IDF is a statistic that reports how important a word is to a document (e.g., chapters of a book) in a corpus (e.g., a textbook) [134]. The TF-IDF of a word w in a document d is computed as:

$$\text{TF-IDF}_{w,d} = (1 + \log \text{TF}_{w,d}) \cdot \log \frac{N}{\text{DF}_w}, \tag{2.10}$$

where w is a word in the document, d, TF is the number of occurrences of w in d, DF is the number of documents containing w, and N is the total number of documents. For two GO terms, GO_1 and GO_2, their TF-IDF-based similarity is computed as follows. Consider s_1 and s_2 as the description (documents) of GO_1 and GO_2 and $s_1 \cap s_2$ as the vocabulary, v (the set of words that forms the description of the two annotations). For each word i in the vocabulary v, the TF-IDF of i in s_1 and s_2 is computed obtaining the TF-IDF vector of each annotation, $TF-IDF_{GO_1}$ and $TF-IDF_{GO_2}$. These vectors have a size of $|v|$ and each entry i corresponds to the TF-IDF value of the word i in s_1 and s_2, respectively. Then, the similarity between GO_1 and GO_2, $S(GO_1, GO_2)$, is computed as $S(GO_1, GO_2) = 1 - cosine(GO_1, GO_2)$, where $cosine(GO_1, GO_2)$ corresponds to the cosine distance between vectors A_1 and A_2. This similarity ranges from 0 if the GO terms are not similar to 1 if they are similar. Apart from the similarities between annotations, TF-IDF can be also applied to summarize a large list of annotations by finding the most important words in their descriptions (those with a high TF-IDF).

TF-IDF-based measures are a good measure of similarity between two GO terms, however, they do not take into account the information of the ontology. Thus, ontological measures are preferred when analyzing the functional similarity between two functional annotations. These measures compare the meaning of two functional annotations based on their relationships in an ontology. The most used ontological measures of semantic similarity are Resnik's semantic similarity [135] and Lin's semantic similarity [136]. For two GO terms, GO_1 and GO_2, these measures calculate their similarity based on the information content (how specific is the information) of their most informative common ancestor (their closest shared parent annotation) in the ontology. The values of Resnik's semantic similarity range from 0 to infinite. Thus, Lin's semantic similarity is preferred. Lin's semantic similarity is based on Resnik's similarity and ranges from 0 for terms without similarity to 1 for terms with maximum similarity.

Several available methods, such as REVIGO [133], use semantic similarity measures to summarize the GO terms of a large list. These methods summarize long lists of GO terms by finding a representative subset of the terms using clustering algorithms that relies on semantic similarity measures. In Chapter 3, we use the TF-IDF between different GO terms to summarize their meaning. In Chapter 4, we use Lin's semantic similarity between GO BP terms to evaluate the organization of tissues-specific PPI embedding spaces from a functional perspective. Finally, in Chapter 5, we apply Lin's semantic similarity to evaluate the functional coherence of GO BP terms associated with the same embedding axis. In the same Chapter, we develop a new TF-IDF-based measure to uncover the biological meaning of the embedding axes of different species-specific PPI embedding spaces.

2.6 Graphlets and Graphlet-based methods

As introduced in section 2.2.2 of this Chapter, to efficiently capture network topological patterns, many network measures have been proposed. Among them, *graphlets* [3] have been shown to be the most sensitive measures for capturing the topological characteristics of complex biological networks. Thus, graphlets have been used as the basis for sensitive measures of network [3, 137, 4], and nodes (or edges) [2, 138]. At the same time, graphlets have been used to develop state-of-the-art *graphlet-based* algorithms for many computational problems, such as node centrality computation [139], clustering [140], link prediction [141], network comparison [4, 142, 143, 144, 145], and network alignment [146, 129, 147]. In the context of biomedicine, graphlet-based methods have been efficiently applied to various problems, such as the study of human aging [148], cancer [149, 150, 151], and other diseases [152]. Importantly, graphlets have been shown to be superior to many other measures of network topology, such as network motifs [153, 105], and many various centrality measures [139, 148].

In the rest of this section, we first introduce the original graphlets and provide an overview of their extended versions. Then, we describe graphlet-based measures that focus on the topology of individual nodes, including the Graphlet Degree Vector and the Graphlet Degree Distance. Finally, we provide an overview of the graphlet-based measures of the entire network, with a particular emphasis on the Graphlet Correlation Matrix and the Graphlet Correlation Distance.

2.6.1 Graphlets

Formally, graphlets are defined as connected non-isomorphic induced subgraphs of an undirected network [3] (these concepts are defined in section 2.2.2). Figure 2.5, illustrates all possible graphlets containing two to five nodes. The number of graphlets containing k nodes increases exponentially with k. The computational complexity of counting graphlets of k nodes in a network of n nodes is $O(n^k)$. For these reasons, in practice, graphlets up to five nodes have typically been studied. Yet, the statistics of 2- to 5-node graphlets have been proven to be detailed enough to capture most of the network topological properties, as most molecular networks have a short diameter. Moreover, in some cases, it has been shown that the use of 4-node graphlets can be sufficient to extract information from the network topology [141, 140, 154]. Hence, the need of using 5-node graphlets is an open question that needs to be investigated.

The nodes of all 2- to 5-node graphlets are also annotated with automorphism orbits (simply called orbits), where each orbit defines a group of nodes that are topologically symmetrical in a graphlet. See Figure 2.5 for an illustration. There are 73 orbits for 2- to 5-node graphlets [2]. Using the orbits of graphlets, Przulj *et al.* [137] generalize the notion of node degree to graphlet degree: the i^{th} graphlet degree of a node N is the number of graphlets that N touches at orbit i. Note that, with this definition, the 0^{th} graphlet degree corresponds to the standard definition of node degree. The graphlet degree of the nodes is the basis for the majority of the graphlet-based measures that we introduce in the next subsections.

Graphlets were originally designed to analyze undirected unweighted networks. However, in the last decade, graphlets have been generalized to many different network models, such as directed networks [7], dynamic networks [140], hyper networks [12], or heterogeneous networks [155]. That is, graphlets have been adapted to accommodate the increasing complexity of biological data, which demands more sophisticated network models. However, current graphlets have not been extended to weighted networks, which are one of the most widely used network models in

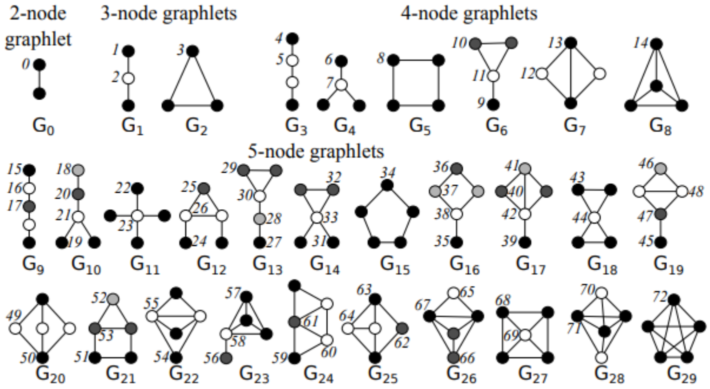


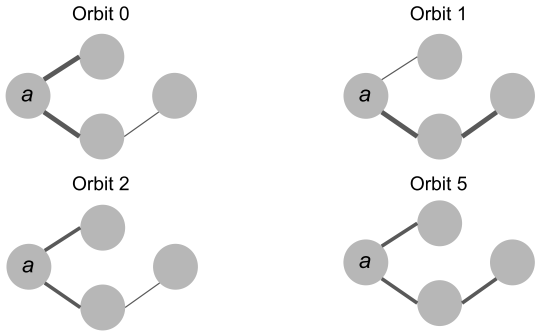
Figure 2.5. All 2- to 5-node graphlets $G0, G1, G2, ..., G29$, and their automorphism orbits 0,1,...,72. Nodes belonging to the same automorphism orbit are of the same shade in each graphlet.

biology (detailed in section 2.1). In Chapter 3 of this dissertation, we address the limitations of the current graphlet methodology by extending it to weighted networks, specifically probabilistic networks. For the next of this section, we will focus on the measures based on the original unweighted graphlets.

2.6.2 Graphlets-based Measures of Nodes Topology

Many network tasks require a way of summarizing the topology around a node in the network. This information is captured by the graphlet degree vector. Assuming the existence of o orbits ($o = 15$ for 2- to 4-node graphlets and $o = 73$ for 2- to 5-graphlets), the *Graphlet Degree Vector* (also known as graphlet degree signature) of a node is a o-dimensional vector where each value represents the graphlet degree of the node for a particular orbit. The Graphlet Degree Vector (GDV) computation for a node a in a toy network of four nodes is illustrated in Figure 2.6. The GDV of all n nodes in a network can be represented by an $n \times o$ matrix (named *GDV matrix*), where matrix entry (i, j) contains the information of how many times node i touches orbit j.

Comparing the GDVs of two nodes give a measure of the topological similarity of the nodes. However, directly comparing their GDVs could not be appropriate since orbits are not completely independent of each other [2]. For instance, the differences in orbit 0 of two nodes will automatically imply the differences in all other orbits of these nodes, since all orbits contain orbit 0. The *GDV-similarity* was designed to remove orbit dependencies [2]. This approach assigns a higher weight w_i to those orbits that are independent, and a lower weight to those orbits that are dependent. Specifically, if an orbit i is affected by o_i number of other orbits, then its corresponding weight is computed as $w_i = 1 - (\frac{log(o_i)}{log(n_o)})$, where n_o represents the total number of different orbits for graphlets of up to size n. For example, for $n = 4$, there are $n_o = 15$ different orbits. Although these weights remove some orbit dependencies, it was shown that for a given node, some of its orbit counts (columns in the GDV) are redundant, meaning that their counts can be derived



Orbit	0	1	2	3	4	5	6	…	14
GDV(*a*)	2	1	1	0	0	1	0	…	0

Figure 2.6. An illustration of the GDV for 2- to 4-node graphlets (i.e., their 15 orbits) of a node a in a toy network of four nodes, and its computation.

from the counts of non-redundant orbits [4]. Hence, the GDV-similarity is usually computed by considering only non-redundant orbits. Given two nodes u and v, their GDV-similarity is computed as follows [2]. Let u_i and v_i denote the i^{th} coordinate of the GDVs of u and v, respectively. Then, the distance $D_i(u, v)$ between the i^{th} orbits of nodes u and v is:

$$D_i(u,v) = w_i \times \frac{|log(u_i+1) - log(v_i+1)|}{log(\max\{u_i, v_i\}+2)} \tag{2.11}$$

Given the $D_i(u, v)$ values for each orbit i (e.g., for all 15 orbits of the 2- to 4-node graphlets), the GDV distance between u and v, $D(u, v)$ is computed as:

$$D(u,v) = \frac{\sum_{i=0}^{14} D_i}{\sum_{i=0}^{14} w_i} \tag{2.12}$$

The value of $D(u, v)$ is 1 if the nodes are topologically different and 0 otherwise. Finally, the GDV-similarity $S(u, v)$ between the nodes u and v is: $S(u, v) = 1 - D(u, v)$. As mentioned above, to avoid redundancies between orbits, one can simply discard from the GDV those columns that contain counts for the redundant orbits and then apply the above formulas. The GDV similarity between nodes is frequently used as input for clustering algorithms to sub-group nodes based on their topological similarity. Afterward, functional enrichment analysis is applied to these clusters to uncover the biological information hidden in the wiring patterns of molecular networks (details about network clustering and functional enrichment analysis can be found in section 2.5).

As graphlets, GDV, and GDV similarity have been generalized to directed network [7], dynamic networks [140], or heterogeneous networks [155]. However, up to date these graphlet-based measures have not been extended to weighted networks. To address this issue, in Chapter 3, we extend the GDV and GDV similarity to weighted networks, precisely to probabilistic networks.

2.6.3 Graphlets-based Measures of Entire Network Topology

The GDV similarity is an efficient measure of the topological similarity between two nodes. However, sometimes we aim to capture the compare the topological similarity between two entire networks. In this case, the leading edge graphlet-based method to compare the topological similarity between two networks is the *Graphlet Correlation Distance* (GCD) [4]. The GCD between two networks, N, and $N1$, is defined as the Euclidean distance of the upper triangle values of their *Graphlet Correlation Matrices* (GCMs). The GCM of each network is computed as follows. Consider network N with n nodes and its GDV matrix of size $n \times o$ ($o = 11$ non-redundant orbits for 2- to 4-node graphlets and $o = 25$ non-redundant orbits for 2- to 5-node graphlets). Recall that each entry (i, j) in the networks' GDV matrix represents the number of times that a node i touches orbit j. The GCM of this network N, GCM_N, is defined as a matrix of size $o \times o$, where the value at position (i, j) in the GCM_N represents the Spearmans' correlation coefficient between columns i and j of the GDV matrix. Then, having the respective GCMs of N and $N1$, GCD is defined as the Euclidean distance of the upper triangle matrix values of GCM_N and GCM_{N1}. Formally the GCD between these two GCMs, $GCD(GCM_N, GCM_{N1})$, for 1- to 4-node graphlets ($o = 15$) is computed as:

$$GCD(GCM_N, GCM_{N1}) = \sqrt{\sum_{i=0}^{14} \sum_{j=i+1}^{14} (GCM_N(i,j) - GCM_{N1}(i,j))^2} \tag{2.13}$$

GCM and GCD are only applicable for undirected unweighted networks. However, in Chapter 3 of this dissertation, we extend them to probabilistic networks.

2.7 Network Embeddings-based Methods

The complexity of biological data has been increasing. As a result, biological networks have gradually incremented their size becoming more complex and complete. In this new scenario, the application of topological measures, such as graphlets and graphlet-based approaches (introduced in section 2.6), that operate directly on the networks is time-consuming and sometimes computationally intractable. Thus, a growing body of work applies network embedding techniques to simplify, visualize, and facilitate the analysis of this large and notoriously complex network data [14]. Network embedding techniques (also called graph representation learning) generate vector representations for network elements such that the learned representations, i.e., embeddings, capture the structure and semantics of networks [156]. This problem is often formalized as follows: Given a network $G(V, E)$, with V nodes and E edges, and its corresponding adjacency matrix (introduced in 2.2.1), the goal is to learn a function $V \rightarrow \mathbb{R}^d$ that maps each node to a d-dimensional ($d < |V|$) vector that captures its structural properties. Figure 2.7 illustrates the d-dimensional embedding space of a toy network G of 14 nodes.

To extract biological information from molecular networks, classic network approaches rely on network properties (e.g., degrees or clustering coefficients, introduced in section 2.2.2 of this Chapter) or on carefully engineered features to measure network structures (e.g., graphlets, introduced in section 2.6 of this Chapter) [156]. In contrast, network embedding techniques automatically learn to encode networks into low-dimensional representations (i.e., embeddings) using transformation techniques based on deep learning and nonlinear dimensionality reduction [156]. The flexibility of learned representations shows in a variety of downstream tasks that representations can be used for. In particular, in the biomedical field, these learned representations are often used for the following tasks:

Node, link, and network property prediction: The objective is to learn representations of network elements, such as nodes, edges, subgraphs, and the entire network. Representations are optimized so that performing algebraic operations in the embedding space reflects the network topology, e.g., topologically similar nodes are embedded close in the embedding space [14]. Optimized representations are usually used as input for models to predict properties of the network elements, such as the function of proteins in a molecular networks [157] (i.e., node classification task), the binding affinity of a chemical compound to a target protein [158] (i.e., link prediction task), and the toxicity profile of a candidate drug [159] (i.e., graph classification task).

Latent network learning: Network embedding techniques exploit relational inductive biases for data that come in the form of networks. However, in many biological problems, the networks are not readily available for learning, e.g., gene regulatory networks are not complete. In this context, latent network learning is concerned with inferring the network from the data [156]. The latent graph can be application-specific and optimized for the downstream task.

Network generation: The objective is to generate a network G representing a biomedical entity that is likely to have a property of interest, such as high drug-likeness [160]. The model is given a set of networks M with such a property and is tasked with learning a non-linear mapping function characterizing the distribution of networks in M. The learned distribution is used to optimize a new network G with the same property as input networks [156].

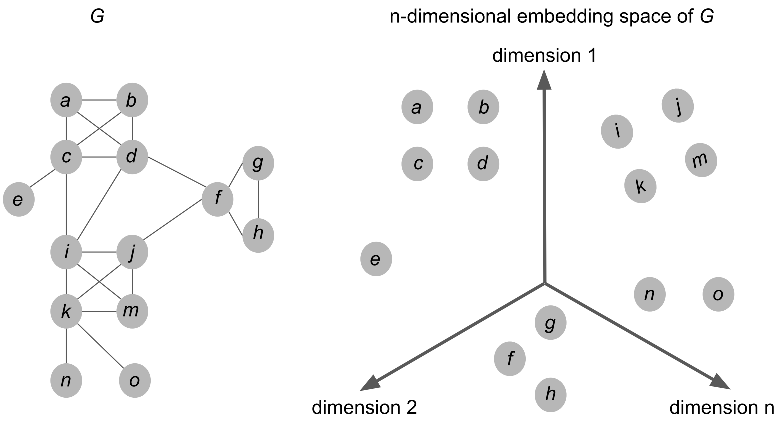


Figure 2.7. An illustration of the d-embedding space of a toy network G of 14 nodes. The node closeness in the original network G is preserved in the embedding space.

Network embedding techniques encompass a wide range of methods, including manifold learning, topological data analysis, graph neural networks and generative graph models [156, 14], that have been widely applied in biomedical research to identify cancer-related genes [125], to subtype cancers [161], to stratify patients [162], or to repurpose drugs [163]. These algorithms include NLP-inspired approaches, such as DeepWalk [19], and node2vec [20], and matrix factorization-based approaches, such as Non-negative Matrix Tri-Factorization (NMTF) [21]. In the rest of this section, we first introduce the most used NLP-inspired embedding approaches, such as node2vec, Deepwalk, and LINE [164]. We show that these algorithms implicitly factorize the PPMI matrix network representation (introduced in section 2.2.1). Following this, we focus on factorization-based embedding approaches, specifically the NMTF.

2.7.1 NLP-inspired Embedding Approaches

The recent advantages in network embeddings have been largely influenced by word2vec, a skip-gram model originally proposed for embeddings [165], whose input is a text corpus composed of sentences in natural language and output is the latent vector representation of each word in the corpus [166]. Inspired by these settings, NLP-inspired approaches consider the node paths traversed by random walks over networks as the sentences and leverage skip-gram for learning the vectorial representation of the nodes. By doing it, diffusion-based methods preserve both the local and the global network structures. In the last decade, many NLP-inspired embedding approaches have been developed, such as DeepWalk, LINE, and node2vec. The main difference between these models lies in the type of random walk that they use to generate the node representations. *DeepWalk* algorithm uses simple random walks over the network to generate the sentences [19]. *Node2vec* extends DeepWalk by introducing a flexible notion of random walks, called biased random walks [20]. These biased random walks allow balancing the exploration-exploitation trade-off and learning node representations that capture diverse information from the graph (details about the exploration-exploitation trade-off can be found in [20]). Finally, *LINE* uses both first- and second-order proximity of the nodes (their neighbors and the nodes connected to their neighbors) to learn the embeddings [164].

Recently, it was shown that skip-gram models used to obtain word embeddings in NLP are implicitly factorizing a word-context matrix, called the shifted Pointwise Mutual Information (PMI) matrix [167]. The cells of this matrix are the PMI values of the respective word and context pairs, shifted by a global constant. However, the PMI values of a word context pair (w, c) that did not occur in the corpus would be $PMI(w, c) = log0 = -\infty$. To address this issue, in the field of NLP, the negative entries of this matrix are replaced with 0s, resulting in a sparse matrix called the shifted positive PMI (PPMI) matrix (introduced in section 2.2.1 of this Chapter). Levy *et al.,* [167] showed that the exact factorization of the shifted PPMI matrix with Singular Values Decomposition (SVD) could achieve solutions that are at least as good skip-gram models for word similarity tasks. Formally the PPMI of two words, w and c, is defined as:

$$PPMI(w, c) = \max \left\{ 0, log \frac{\#(w, c) \times |C|}{\#w\#c} \right\} \tag{2.14}$$

where $|C|$ is the size of the corpus, $\#(w, c)$ is the number of times two words co-occur in the corpus, and $\#w$ and $\#c$ are the number of times the words w and c occur in the corpus, respectively.

Motivated by these observations, Qiu *et al.,* [166] investigated the theoretical mechanism behind DeepWalk, LINE, and node2vec. In their work, Qiu *et al.,* [166] demonstrated that these algorithms are implicitly factorizing a random-walk-based mutual information matrix. This matrix is equivalent to the shifted PMI matrix on networks, as its cells quantify how frequently two nodes, i and j, of the network co-occur in a random walk compared to what would be expected if the co-occurrences of the nodes were independent. Qiu *et al.,* [166] also found the closed formulas to approximate the shifted PMI matrices that DeepWalk, LINE, and node2vec are implicitly factorizing. As in NLP, to ease the use of these shifted PMI matrices of networks in matrix factorization approaches, such as SVD, while preserving the resulting network qualities, these matrices are substituted by the simpler PPMI matrix by setting the shifted parameter to zero and replacing all negative values with zeros [166]. Contrary to the adjacency matrix representation introduced in section 2.2.1 of this Chapter, the PPMI matrix captures high-order proximities between the nodes in the network, hence, the PPMI matrix is a richer representation than the adjacency matrix (we demonstrate this in Chapter 3).

Taking advantage of the previous findings, Xenos *et al.,* [89] extended the NLP-inspired embedding methods to the Non-negative Matrix Tri-Factorization (NMTF), a well-known embedding technique that has successfully been used in the biological context to analyze large-scale omics data types [21] (we introduce NMTF in the next subsection), by applying NMTF to the PPMI matrix representations of molecular networks obtained by the Deepwalk closed formula with its default settings [89] (see equation 2.15). One of the advantages of using NMTF over the previous NLP-inspired embedding approaches is that it requires way fewer parameters to tune, thanks to the careful modeling of the relationships between the data points that it takes as input. Moreover, contrary to these embedding approaches, the embedding space produced by NMTF can have valuable properties, e.g., orthonormality, that may lead to an easier interpretation and deeper scientific insight [168]. In Chapter 5, we demonstrate that these properties have an impact on the ability of the embedding space to capture the functional organization of the cell.

$$Deepwalk : X = \max \left\{ 0, log \left(vol(N) \left(\frac{1}{T} \sum_{r=1}^{T} (D^{-1}Ajd)^r \right) \right) - log\ b \right\} \tag{2.15}$$

where Ajd is the adjacency matrix of the network N, D is the diagonal matrix of Ajd, $vol(N)$ is the volume of G, $T = 10$ is the length of the random walk and $b = 1$ is the number of negative sampling in skip-gram. Note that the second argument corresponds to the Deepwalk closed formula by [19], i.e., to the shift PMI matrix implicitly decomposed by the Deepwalk algorithm.

2.7.2 Matrix factorization-based approaches

The most frequently used approach for matrix factorization is NMTF, an extension of Non-negative Matrix Factorization (NMF) and a well-known machine learning (ML) technique introduced for co-clustering and dimensionality reduction [21]. In an NMTF, a non-negative matrix, $X \in \mathbb{R}^{n_1 \times n_2}$ such as the PPMI matrix or the adjacency matrix is approximated by the product of three lower-dimensional matrix factors, $P \in \mathbb{R}^{n_1 \times k}$, $S \in \mathbb{R}^{k \times k}$, and $G \in \mathbb{R}^{k \times n_2}$, where $k << min(n_1, n_2)$ [169].

Setting the rank parameter $k << min(|v|, |v|)$ provides a dimensionality reduction [170]. NMTF is closely related to k-means clustering [21]. From a clustering point of view, entries in X represent the relationship between two different types of objects, e.g., the relation between n_1 patients (type 1) and n_2 genes (type 2). This matrix is factorized into P, S, and G, where U is the cluster indicator matrix of type 1 object (grouping the n_1 objects of type 1 in k clusters), G is the cluster indicator matrix of objects of type 2 (grouping the $n2$ objects of type 2 into k clusters), and S is the compressed representation of X that related the clusters of P in G.

NMTF is a co-clustering approach, which means that one can extract objects from P clusters that group together objects of type 1 according to their relationships with objects of type 2 while extracting from G objects of type 2 according to their relationship with objects of type 1. Such clusters can be extracted from P and G with a procedure called "hard-clustering" [171], in which each data point i is assigned to a cluster j, $1 \leq j \leq k$, such that $P_{i,j}$, is the maximum value in row i (same for G). Another important property of NMTF is its completion property. Namely, after solving NMTF, the reconstructed matrix $\hat{X} = PSG^T$ has new entries not observed in X but emerging from inferred factors (also called latent factors) that can be used for prediction.

NMTF can be also described from an embedding point of view. To simplify, consider a non-negative matrix X of size $n_1 \times n_1$, where each entry represents an interaction between two objects of the same type, for instance, edges between genes in a molecular network. This matrix is decomposed into three factors, P, S, and G, where the set of the rows of the matrix $P \cdot S$ defines the set of embedding vectors of the genes, E, and the set of the columns of G defines the basis (also called embedding axis), B, of the space in which the genes are embedded. Orthonormality constraint can be applied to the basis-defining matrix ($G^T G = I$). This constraint leads to minimal col-linearities (hence, minimizing the dependencies) between the vectors of the basis, B, of the embedding space [172]. It has been hypothesized that this orthonormality improves the ability of NMTF to reveal the cell's functional organization. However, this hypothesis has not been properly tested, i.e., the real impact of orthonormality remains unclear. In Chapter 5, we assess this hypothesis by comparing orthonormal PPI embedding spaces with non-orthonormal PPI embedding spaces generated by NMTF. The NMTF decompositions with orthonormality constraint, $NMTF_{orthonormal}$, and without this constraint $NMTF$, are done by minimizing the functions:

$$\begin{aligned} NMTF_{orthonormal} &: min_{P,S,G \geq 0} \|X - PSG^T\|_F^2, G^T G = I, \\ NMTF &: min_{P,S,G \geq 0} \|X - PSGT\|_F^2, \end{aligned}$$

where F denotes the Frobenius norm. These optimization problems are NP-hard [21]; thus, they are heuristically solved by using a fixed point method that starts from an initial solution and iteratively uses multiplicative update rules [21]. Such rules guarantee convergence towards a locally optimal solution that verifies the Karush-Kuhn-Tucker (KKT) conditions [21] (these rules can be found in Supplementary section B.1.1). In Chapters 4 and 5, we apply NMTF on different PPI networks to capture the cell's functional organization.

Chapter 3

Probabilistic graphlets capture biological function in probabilistic molecular networks

In this chapter, we extend the graphlets to probabilistic networks by introducing probabilistic graphlets. We use probabilistic graphlets to generalize the state-of-the-art graphlet-based methods to capture the local topology of the nodes in a network (i.e., GDV and GDV distance) and to capture the topology of an entire network (GCM and GCD). First, we evaluate the ability of probabilistic graphlets to capture the overall topological similarity between probabilistic networks based on well-known random network models. Then, we apply our methodology to analyze the probabilistic representations of different molecular interaction networks and investigate the biological information uncovered by our probabilistic methodology. The results of this chapter are published in Doria-Belenguer *et al.,* [173] and were presented at the ECCB'19 conference.

3.1 Motivation

A powerful approach for uncovering information from a biological network is to measure the local wiring patterns of its nodes. The best measures to quantify the local wiring patterns rely on graphlets -small, connected, non-isomorphic, induced subgraphs of a larger network [3]-(Figure 2.5). The descriptive power of graphlets has been widely exploited for comparing network topologies and mining networks for local topological similarities [2, 4]. Moreover, to account for the increasing complexity of biological network data, the original graphlets have been extended to concepts of directed [7], dynamic [140], heterogeneous [155], or node-ordered [143] graphlets.

However, current models do not account for uncertainty associated with molecular interactions, which is an integral part of biological systems. Uncertainty can be defined as the confidence that a specific interaction is occurring and has been linked to the accuracy of the underlying measurements. For network representations, it is common to discretize these measurements by applying thresholds to decide which interactions are considered to be real. While this approach permits the removal of noise from the data, it can also eliminate crucial information. To overcome this limitation and utilize more of the available information about biological interactions, thresholded networks can be replaced by probabilistic networks with probabilities as weights on edges (links). In weighted networks, each edge is assigned a weight that contains additional information about the edge, such as the strength of the interaction. A special case of these models is probabilistic networks, in which the weights represent the probability that an interaction between two nodes occurs. We hypothesize that probabilistic networks can accurately represent the uncertainty about molecular interactions and capture their probabilistic nature. Nevertheless, the currently available network methodology is mostly defined for unweighted networks and is not applicable to these probabilistic networks.

To extract the additional information hidden in the wiring patterns of probabilistic networks, in this chapter of the dissertation, we generalize the graphlet-based methodology to probabilistic networks and demonstrate its utility in capturing additional information when compared to the original, unweighted graphlet-based methodology applied to thresholded, unweighted networks. We generalize the GCD [2] to probabilistic graphlets (pGCD) and use it for comparing probabilistic networks based on three random network models: Barabasi and Albert preferential attachment [94], Erdös-Renyi [114], and Geometric random graph [118]. We evaluate the ability of pGCD to detect differences between not only unweighted different network topologies (generated by different network models) but, on top of that, edge weight probability distributions for the same, unweighted, topologies. Moreover, we assess its capability to detect different edge weight distributions by randomly applying edge probabilities to each of the network models based on the empirical distributions of real molecular interaction networks, as well as a uniform distribution. We observe that pGCD can differentiate between probabilistic distributions and network topologies, while GCD can only differentiate between network topologies. We conclude that pGCD is an accurate alternative for comparing probabilistic networks.

Lastly, using probabilistic graphlets, we analyze the probabilistic representations of molecular interaction networks for budding yeast and human and compare our methodology against unweighted graphlets. We show that probabilistic graphlets not only deal more efficiently with noise but also capture additional, more specific biological information compared to their unweighted counterparts. We argue that these differences arise from the information contained in lower confidence interactions that are lost when thresholding the data.

3.2 Materials and Methods

3.2.1 Definition of probabilistic graphlets

A *probabilistic network* is a weighted graph G = (V, E, p), where V is a set of nodes, E is a set of edges that connect nodes in V, and where each edge, e $\in$ E, is associated with probability, p(e), between 0 and 1. Implicitly, each edge, e, is also associated with the probability of not being connected, p($\neg$ e) = 1-p(e). Each pair of nodes in the network that is not connected by an edge is associated with probability 0.

Since these probabilities are independent, the outcome probability of a weighted subgraph can be calculated by the product of the probabilities of the edges and non-edges (see example in Figure 3.1). This outcome probability can also be seen as a measure of "how much induced" the chosen subgraph is in a probabilistic sense.

To that end, we define a probabilistic graphlet to be any subgraph, weighted by the probability of its corresponding product of edge weights. For instance, given a triangle with edge-probabilities p_1, p_2, and p_3, the probability that the 3-node path consisting of edge p1 and p2 results in an induced subgraph is given by product, $p_1p_2(1\text{-}p_3)$. Hence, this subgraph is considered a probabilistic graphlet with weight $p_1p_2(1\text{-}p_3)$, even though it is not an induced subgraph of the network when ignoring the probabilities on the edges (see Figure 3.1).

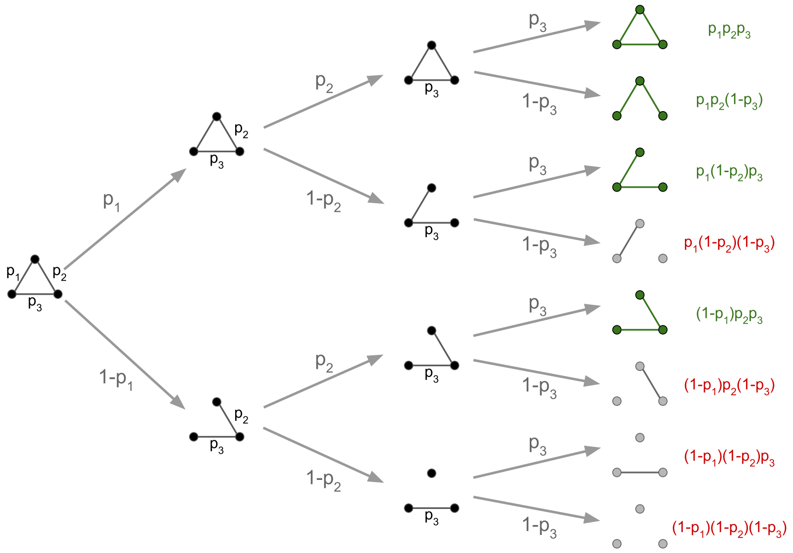


Figure 3.1. Counting probabilistic graphlets for a triangle with edge probabilities p_1, p_2, and p_3. Displayed are all 2^3 possible outcomes of a probabilistic triangle (graphlet G_2 in Figure 2.5). Out of these eight outcomes, only four results in 3-node graphlets (shown in green). Note that 3-node paths can appear as graphlets (second, third, and fifth outcome) even though they are not induced subgraphs in the original probabilistic network (shown on the very left). In case p_1, p_2, and p_3 only take values in 0 and 1, there is only one possible outcome with non-vanishing probability, leading to a coherent definition of a probabilistic and an unweighted graphlet. The fact that certain realizations result in non-connected 3-node subgraphs (shown in red) can be exploited to exclude these outcomes a *priori* when calculating graphlet counts, making the computation of probabilistic graphlet counts feasible.

For a fixed graphlet, G_i (Figure 2.5), its probabilistic graphlet count within a network is now defined to be the weighted sum over all instances of the probabilistic graphlet, G_i. This definition is equivalent to taking the expected value of the "classical" (unweighted) graphlet counts of G_i across all possible realizations of the probabilistic network. If unweighted networks are replaced by probabilistic networks with edge weights being zero (non-existing edge) or one (existing edge), all graphlet counts of the unweighted network coincide with the probabilistic graphlet counts of the probabilistic network. In that sense, probabilistic graphlets form a natural generalization of unweighted graphlets.

We further extend the notion of *graphlet orbits* (or simply *orbits*) to probabilistic orbits and their counts. Orbits are equivalence classes of nodes within graphlets, accounting for "symmetries" within a graphlet, and are usually labeled from 0 to

73 for up to 5-node graphlets [137] (detailed in Chapter 2, section 2.6.1). In short, given orbit, i, the number of times a node is being "touched" by that orbit is called the i^{th} graphlet orbit count of the node. Note that the 0^{th} orbit count coincides with the degree of a node and thus, orbit counts are also referred to as *graphlet degrees*. The vector comprising all graphlet degrees for a given node is referred to as the GDV of the node [2] (see Figure 2.6). Remind that the matrix consisting of the row-wise collection of these node feature vectors is called the *GDV matrix* (introduced in Chapter 2, section 2.6.2). It has 73 columns, describing the global behavior of the orbits over all nodes, and one row for each node describing the local topology of the node via its graphlet degrees.

In analogy to the probabilistic graphlet counts described above, we define the *probabilistic graphlet degrees* to be the expected number of (unweighted) graphlet degrees across all realizations of the probabilistic network. We call the resulting feature vector the *probabilistic graphlet degree vector* (pGDV) and the matrix composed of all pGDVs the *pGDV matrix*. We demonstrate that the pGDV matrix of a probabilistic network captures valuable information, which the GDV matrix of the thresholded network fails to encompass.

Due to the computational complexity of the probabilistic method, we perform our calculations for only up to 4-node graphlets, resulting in 15 orbits labeled from 0 to 14.

3.2.2 Probabilistic graphlet count implementation

We modify the gtrieScanner software [174] to extend its applicability to probabilistic networks. Given a fixed number of nodes in a graphlet, n, the algorithm calculates all instances of graphlets of size n, ignoring calculations of non-connected subgraphs a *priori* to avoid redundant calculations (Figure 3.1, red). The output is the probabilistic graphlet count, as well as the probabilistic orbit count, for each node in the network and each graphlet of size n. This function was written in C++ version 11 and compiled using g++ (version 7.4.0) under Ubuntu 18.04.

3.2.3 Definition of topological distance measures

As introduced in Chapter 2, section 2.6.2, the current state-of-the-art measure of topological similarity between local wiring patterns of the nodes in an unweighted network is the GDV. We generalize the GDV distance formula 2.11 on the *pGDV distance* of nodes in probabilistic networks to measure their wiring pattern similarity.

The leading edge method to compare the topological similarity between unweighted networks is GCD (introduced in Chapter 2, section 2.6.3). We extend the GCD to probabilistic networks by first calculating the GCM from the pGCV matrix of the nodes and then applying the formula 2.13. We define the extension of this distance to probabilistic networks as probabilistic GCD (pGCD). As introduced in Chapter 2, section 2.6.3, Yaveroglu *et al.* [4] showed that the GCD based on eleven non-redundant orbits performs better than the GCD based on all fifteen orbits for up to 4-node graphlets, due to redundancy equations describing relationships between different orbits. However, these equations are not applicable to probabilistic networks, and thus we use all fifteen orbits to calculate the pGCD between the networks.

3.2.4 Biological datasets

We use three types of molecular interaction networks for budding yeast and human (see Table 3.1). From STRING v11.0 [47] we collect the experimental PPIs of human and budding yeast and we use the experimental scores from STRING as the edge probabilities. From the study by Costanzo, *et al.* [53] we collect the GIS and we used the PCC between the genetic interaction profiles of each pair of genes as edge probabilities, all negative PCCs were set to 0. From COXPRESdb (v7.3) [59] we obtain the CoEx network and we define the edge probability as 1 minus the value of the mutual rank of a gene pair divided by the total number of genes (see https://coxpresdb.jp for details). Details about these biological networks can be found in Chapter 2, section 2.1.

Network	Confidence	Threshold	# Nodes	# Edges	Density
PPI Budding yeast	High	>=0.7	2130	15.226	0.75%
	Low	all edges	4980	788,548	0.36%
GIS Budding yeast	High	>=0.4	1555	2912	0.24%
	Medium	>=0.2	4732	30137	0.26%
	Low	>=0.05	5707	1,635,781	10%
CoEx Budding yeast	High	>=0.82	5757	171,384	1%
	Medium	>=0.66	5855	856,905	5%
	Low	>=0.60	5855	1,713,792	10%
PPI Human	High	>=0.7	3903	39,790	0.24%
	Low	all edges	9890	119,116	0.52%

Table 3.1. Statistics for the biological networks in this study. Threshold values (3^{rd} column) depend on network type: for PPI, experimental evidence of the interaction; for GIS, PCC of the genetic interaction profiles; for CoEx, 1 minus Mutual Rank divided by the total number of genes. "# Edges" contains the number of edges of the network. "# Nodes" contain the number of nodes of the network. "Density" contains the edge density of the network.

To compare probabilistic and unweighted networks, we form probabilistic and unweighted counterpart networks by applying different thresholds to the raw data and only retaining edges whose assigned probability passed the corresponding threshold. In particular, we define high and low confidence thresholds for the two PPI networks, as well as high, medium, and low confidence thresholds for the GIS and CoEx networks (see Table 3.1).

For each of these thresholds, we generate a probabilistic model with all edge probabilities satisfying the applied threshold. We also generate unweighted network representations of the probabilistic networks presented above by applying the same thresholds on the edge probabilities, as summarized in Table 3.1.

Details about the applied thresholds can be found in Table 3.1. Briefly, PPI thresholds are applied as recommended by the STRING authors. In GIS, thresholds are chosen following the methodology of Costanzo *et al.* [53]. Finally, the CoEx thresholds for high, medium, and low confidence are selected using the top 1%, 5%, and 10% of genes based on mutual rank.

To investigate functional enrichments as well as to test the amount of biological information that our methodology is capable of extracting from the networks, we decide to collect all the GO terms that are associated with each gene (details in

Chapter 2, section 2.5.2). To do so, from the most specific experimentally validated GO terms [130] of each gene, we perform a back-propagation to the most general term using GOATOOLS [175] keeping all the GO terms in the process. GO annotations are obtained from the GO Consortium database (accession date: 20.09.2019) (Gene Ontology Consortium: going forward, 2015), National Center for Biotechnology Information (accession date: 22.11.2018; https://www.ncbi.nlm.nih.gov/) and Ensembl (accession date 27.02.2020) [176].

3.2.5 Synthetic datasets

To assess the performance of pGCD in detecting different edge probability distributions, as well as different network topologies, we use synthetic networks that we generate from three random network models: BA [94], ER [114], and GEO [118] (these models are introduced in Chapter 2, section 2.1). Given that the budding yeast networks shown in Table 3.1 feature around 3000 nodes with an edge density of 0.3%, we choose these parameters for generating the model networks. Since these models are defined only for unweighted networks, we develop a pipeline to extend these synthetic networks to probabilistic models. This protocol consisted of three steps: first, we generate an unweighted network from one of the three random network models mentioned above. Then, we obtain the empirical probability distribution of weights on edges from the real-world biological network data. Finally, for each edge in the unweighted model network, we randomly sample one probability from the empirical distribution and use it as the probability of the corresponding edge.

3.2.6 pGCM visualization and measuring the utility of pGCD for clustering of model networks

To visualize the pGCM and GCM matrices, we use heatmaps. To visualize GCD and pGCD distances between networks, we use the uniform manifold approximation and projection for dimension reduction (UMAP) [177]. UMAP is a non-linear dimension reduction method similar to tSNE [178]. However, unlike tSNE which focuses only on preserving local structures, UMAP additionally aims to preserve the overall structure of the high-dimensional data cloud. We use the latter technique to embed our data points (networks) in two dimensions and visualize how the networks are clustered according to pGCD or GCD.

To evaluate the performance of the pGCD for clustering probabilistic networks, we test how well our distance measure groups probabilistic networks of the same type by using standard Precision-Recall (PR) curves: for small increments of parameter $\epsilon < 0$, if the distance between two networks is smaller than ϵ, then the pair of networks is retrieved. For each ϵ, precision is the fraction of correctly retrieved pairs (i.e., grouping together two probabilistic networks from the same model and with the same edge probability distribution), while recall is the fraction of the correctly retrieved pairs overall correct ones. The Area Under the Precision-Recall curve (AUPR), also called average precision, standardly measures the quality of the grouping by a given distance measure. We choose Precision-Recall curve analysis, as it is known to be more robust to large numbers of negatives (in our case, negatives would be pairs of networks from different models that are grouped together) than

Receiver Operator Characteristic (ROC) curve analysis [6].

3.2.7 Measurement of functional enrichments in clusters

To compare the performance of unweighted graphlets and probabilistic graphlets for extracting biological information from unweighted and probabilistic molecular networks, we apply a clustering and functional enrichment analysis approach (introduced in Chapter 2, section 2.5). In particular, we apply k-medoids (this algorithm is detailed in Chapter 2, section 2.5.1) based on GDV and pGDV distances (equation 2.11) between genes in networks to generate clusters of genes with similar local wiring patterns. To account for the random initialization of k-medoids, we perform 10 iterations for each chosen k. For the k clusters found in each iteration, we perform a GO term enrichment analysis while controlling the false discovery rate (FDR) [179]. We consider a cluster as enriched when at least one GO term is significantly enriched among its constituent genes ($\mathrm{FDR} < 0.05$). Likewise, we consider a gene in a cluster to be enriched when it is associated with at least one enriched GO term in the cluster. For each k, we calculate the mean and the 95% confidence intervals (based on bootstrapping) [180] of the percentage of enriched clusters, the percentage of GO terms enriched across all clusters, and the percentage of enriched genes per cluster.

To see whether the unweighted and probabilistic networks capture different biological information, we compute the Jaccard Index (JI) [181] of the enriched GO terms found in both methods for each k. Furthermore, we calculate the mean level of the enriched GO terms captured specifically by each method, as well as the intersection of both methods (unweighted and probabilistic). Lastly, we analyze the ability of both methods to capture different and complementary functions of genes. To do so, we first fix k following the rule of thumb ($k \approx \sqrt{\frac{N}{2}}$, where N is the number of nodes) [18] and select the enriched GO terms that are captured by each method. We use these terms to retrieve the genes responsible for their enrichment and we divide them into three sets: unique for probabilistic (s1), unique for unweighted (s3), and the intersection between the two methods (s2) (see Figure, a Venn diagram). To conclude, we test the robustness of our method by counting the number of times that a GO term is captured in all ten repetitions and compare it against the corresponding counts for the unweighted method.

3.2.8 Assessment of GO annotation similarity

To further understand the differences between the biological information captured by each method, we fix a specific k following the rule of thumb mentioned in the previous section. For each method, we keep only those GO terms that are enriched in all ten repetitions. We subdivide these GO terms into three groups: probabilistic-only, unweighted-only, and intersection. Next, analogous to REVIGO [133], we summarize and remove the redundant GO term information using two steps: 1) for every GO term, perform a back-propagation to the parent term of level 2 or 3; 2) generate a TF-IDF [182] matrix and group the description of the terms by their pairwise cosine similarity. Given these simplifications, two descriptions are considered the same if their cosine similarity was equal to or higher than 0.7. The most common approaches to summarize large lists of GO terms can be found in Chapter 2, section 2.5.2.

3.3 Results and Discussion

3.3.1 Probabilistic graphlets are superior for network comparisons

To determine whether probabilistic graphlets are equally sensitive as unweighted ones, we assess their ability to distinguish probabilistic networks by their topology, edge probability distribution, or both. First, we visualize the GCMs and pGCMs of the model networks introduced in section 3.2.5 and the clustering results obtained from applying UMAP on GCD and pGCD distances between these networks. Then, we evaluate the ability of pGCD to capture differences between network models (topologies) and edge probability distributions.

We compute the pGCMs of the 600 synthetic probabilistic networks described in section 3.2.5 and measure the pGCD distances between them. To compare our methodology with the unweighted graphlet-based methodology, we also compute the GCMs and GCD distances between their unweighted synthetic network counterparts (detailed in section 3.2.3). As previously shown by Yaveroglu *et al.,* [4], we observe that unweighted networks generated from different network models are characterized by clearly distinct GCMs, as visualized by differences in their heatmaps (Figure 3.2 A and B). In contrast, the GCMs of unweighted networks generated by thresholding the probabilistic networks (recall that the probabilistic networks were obtained by weighting the edges of unweighted networks by different edge probability distributions) are mostly identical, leading to indistinguishable heatmap representations (in Figure 3.2, heatmaps in panels A and C are almost identical, and so are heatmaps in panels B and D). This behavior is expected, as the unweighted counterparts of the probabilistic networks have the same topology, which is slightly altered by assigning probabilities to edges (from different probability distributions). When we instead visualize the heatmaps of pGCMs of these probabilistic networks, clear differences are visible in the heatmaps that are dependent on the edge probability distribution applied during the generation of each network (Figure 3.2, E-F).

Likewise, when visualizing these model networks by performing UMAP-based embeddings using the pGCD distances between the networks (detailed in section 3.2.6), we find that probabilistic networks belonging to the same random graph model are well-separated by their edge probability distribution (Figure 3.3 A and Supplementary Figure A.1). Curiously, a synthetic uniform distribution cluster with the empirical PPI distribution (Figure 3.4). Upon closer inspection, we observe that the variance and the mean of both distributions are practically identical (see Table in Figure 3.4), which confirms that the two networks are accurately clustered based on pGCD. We investigate this behavior further, by choosing the Beta distribution, a continuous and versatile family of distributions on [0,1], to fit the weight distributions of the real-world networks (PPI, GIS, and CoEx). In particular, we fit the parameters alpha and beta such that the resulting Beta distribution matches the first two moments (mean and variance) of the empirical distributions; this is known as the “Method of moments” [183]. As before, our methodology correctly

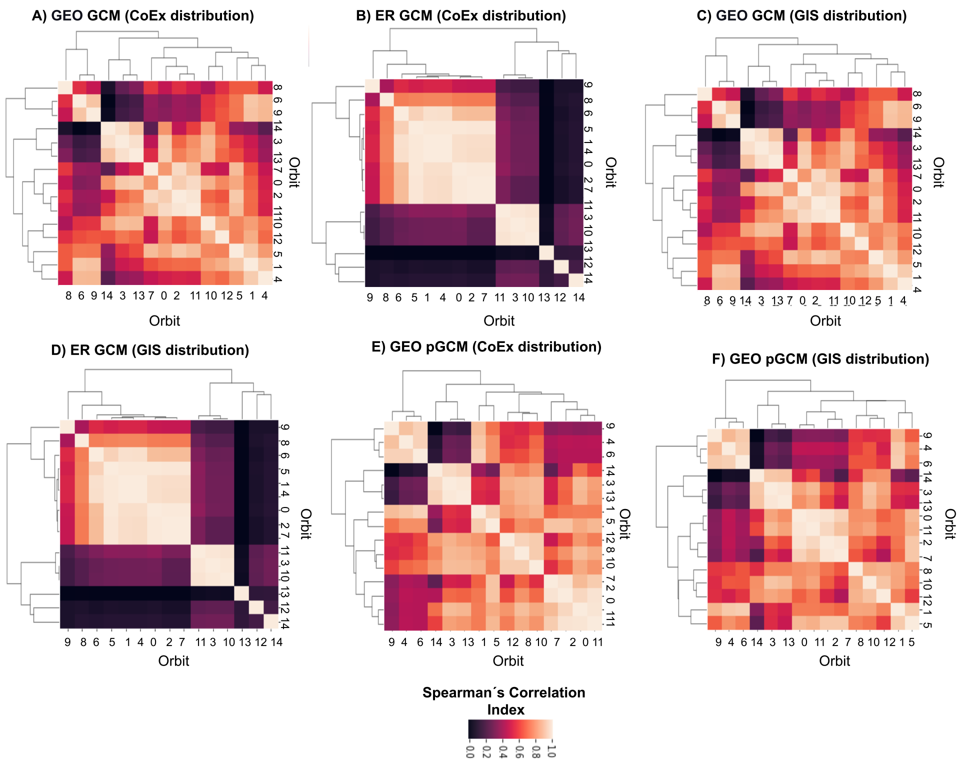


Figure 3.2. Different edge probability distributions of the same network model are characterized by different pGCMs. Heatmaps of the GCMs of (A) a Geometric Random Graph (GEO) unweighted network using the empirical CoEx distribution; (B) an Erdös-Renyi (ER) unweighted network using the empirical CoEx distribution; (C) GEO unweighted network using the empirical GIS distribution; (D) ER unweighted network using the empirical GIS distribution. Note that GCMs struggle to distinguish edge probability distributions and thus all empirical distributions will result near identical heatmaps. Heatmaps of the pGCMs of (E) a probabilistic GEO model using the empirical CoEx edge probability distribution; (F) a probabilistic GEO network using the empirical GIS probability distribution. Columns and rows in each heatmap represent orbits 0 to 14 clustered by their pairwise Spearman correlation coefficient.

clustered distributions with similar moments, including the empirical distributions from the real-world networks (Supplementary Figure A.2). These results again confirm that the pGCD is highly sensitive and able to group networks correctly by their edge probability distributions. Furthermore, when embedding all networks into a single space using UMAP with pGCD distance, we find that the clusters separate the networks by both topology (network model) and probability distribution (Figure 3.3 B), while GCDs only cluster the networks by their underlying network model (Figure 3.3 C). Thus, probabilistic graphlets offer superior sensitivity to distinguish network topologies compared to their unweighted counterparts.

To formally assess the sensitivity of probabilistic graphlets to distinguish between different network models (topologies) and edge probability distributions, we use our pGCD to cluster model networks with different underlying edge probability distributions. We obtain these distributions from different real-world networks. Then, we evaluate this clustering by using AUPR (detailed in section 3.2.6). We find that our method efficiently groups networks with similar topology (generated

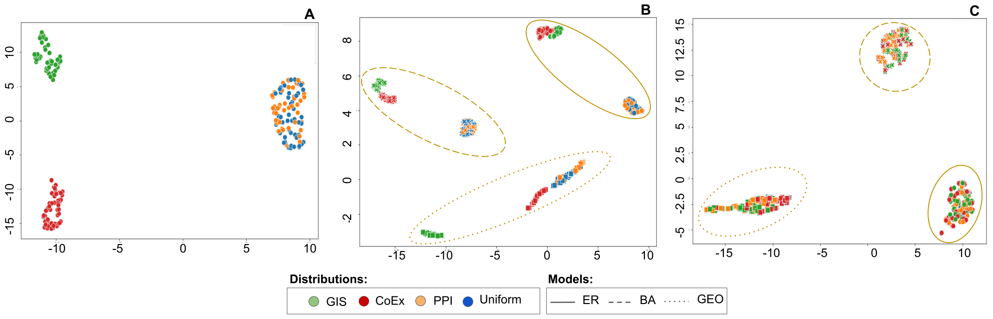
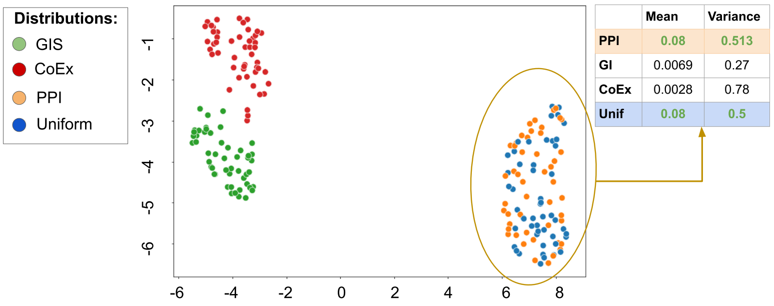


Figure 3.3. pGCMs can separate networks by topology and edge probability distribution. Edge probabilities were sampled based on empirical distributions from GIS, CoEx, and PPI or from a synthetic uniform distribution. Panels represent the UMAP embedding of (A) pGCMs from probabilistic networks based on the Barabasi and Albert preferential attachment model (BA) using the indicated probability distribution; (B) pGCMs from probabilistic networks based on the Erdös-Renyi (ER), BA or Geometric random graph (RG) model using the indicated probability distributions; (C) GCMs of unweighted networks created from thresholding the unweighted networks shown in (B). Colors represent the utilized edge probability distribution, while solid, dotted, and dashed lines indicate the model used for generating the networks.

from the same random graph model) by their edge weight distributions (AUPRs of 0.894, 0.899, and 0.939 for ER, BA, and RG networks, respectively). Moreover, when clustering all model networks simultaneously, pGCD accurately clusters them by random graph model (AUPR = 0.999), as well as by both model and edge distribution (AUPR = 0.980). However, when solely grouping all model networks by their edge probability distribution independent of their network model, pGCD is incapable of doing so (AUPR = 0.53). This behavior is expected, as the model used for generating a network may have a much higher impact on its topology compared to the applied edge probability distribution.



	Mean	Variance
PPI	0.08	0.513
GI	0.0069	0.27
CoEx	0.0028	0.78
Unif	0.08	0.5

Figure 3.4. pGCMs clustering is highly dependent on the variance and mean of the underlying edge probability distributions. Edge probabilities were sampled based on empirical distributions from GIS, CoEx, and PPI or from a synthetic uniform (Unif) distribution. The panel represents the UMAP embedding of pGCMs of the probabilistic network generated by the Barabasi and Albert preferential attachment model (BA) using the indicated probability distribution.

In conclusion, while topological measures based on unweighted graphlets, such as GCM and GCD, are able to differentiate networks by the topology of their underlying network model, probabilistic graphlets are additionally capable of accurately differentiating probabilistic networks of the same model by their edge probability distributions.

3.3.2 Probabilistic graphlets have similar or better performances than their unweighted counterpart

Here, we compare the ability of our probabilistic graphlets to capture biological information from real molecular networks to that of their unweighted counterparts. To this aim, we apply our methodology to real-world examples of probabilistic molecular networks of budding yeast (PPI, GIS, and CoEx) and human (PPI) (detailed in section 3.2.4). To evaluate the amount of information uncovered from these probabilistic networks, we cluster genes based on the distances between their pGDVs and analyze the enrichment of these clusters in GO biological process (BP), cellular component (CC), and molecular function (MF) annotations (detailed in sections 3.2.3 and 3.2.7). Finally, we examine the difference in the GO term enrichments resulting from unweighted and probabilistic graphlet-based techniques.

For budding yeast PPI and GIS networks, we find that probabilistic graphlets yield similar, or higher percentages of enriched clusters, GO terms, and genes, compared to their unweighted counterparts. For instance, we see that probabilistic graphlets consistently capture higher percentages of enriched GO BP, GO MF, and GO CC terms in these networks regardless of being of high, medium, or low confidence (Figure 3.5 A-E and Supplementary Figures A.3- A.5). Similarly, the percentage of enriched clusters in these molecular networks is comparable to, or greater than, the percentage for unweighted graphlets, as shown in Figure 3.5 A-E). Moreover, with the exception of high confidence GIS network (Figure 3.5 M), probabilistic graphlets exhibit higher percentages of enriched genes for all types of GO terms (BP, MF, and CC) across these networks (Figure 3.5 N-O and Supplementary Figures A.3- A.5). We explain the results observed in high confidence GIS network by its low number of genes that results from the stringent threshold that was used to define high-confidence interactions (see Table 3.1).

In addition to the previous results, we also observe that the clustering and enrichment analysis is dependent on the number of clusters, k, as we see variations in the performance of both methods across different values of k, However, as shown in Figure 3.5, our probabilistic approach outperforms its unweighted counterparts over all the tested values of k.

Finally, all of these results also hold for the human PPI network as well, with probabilistic graphlets exhibiting even better performance in the human than in the budding yeast PPI network (Supplementary Figures A.6 and A.7).

Contrary to our earlier findings, we do not find any differences in the percentage of enriched clusters, genes, or GO terms in the budding yeast CoEx network (Supplementary Figures A.8-A.10). We hypothesize that the variation in the performance of probabilistic graphlets is linked to the type of molecular interactions represented in each network and the meaning of the probabilities assigned to each network's edges. In particular, PPIs are physical interactions between proteins that can vary based on environmental changes or tissue specificity, which makes it important to represent

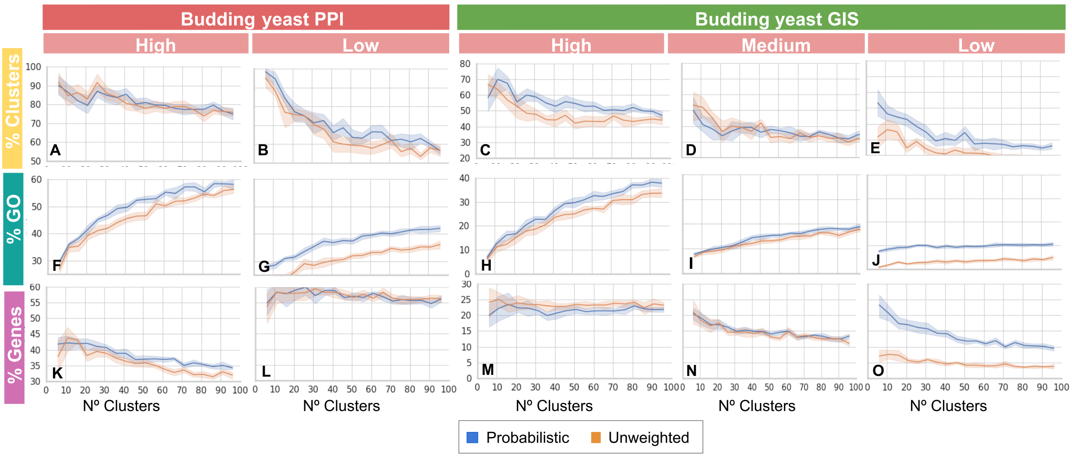


Figure 3.5. Probabilistic graphlets capture as much or more biological information than their unweighted counterparts. Lines represent the mean and the shaded area represents the 95% confidence intervals based on bootstrapping of each enrichment statistic for unweighted and probabilistic graphlets depending on the number of clusters k, across 10 repetitions. High, medium, and low indicate the confidence threshold for the underlying networks.

them using edge probabilities to capture their dynamic nature [47]. Likewise, the connections between genes in GIS networks reflect shared regulatory mechanisms whose activity varies based on conditions [53]. In contrast, CoEx networks present more complex gene interactions as co-expression can result from shared regulatory mechanisms, like transcription factor binding and spatial proximity [184], but can also happen randomly depending on the chosen co-expression method [185]. Therefore, the benefits of representing CoEx networks by probabilistic networks may be limited. Yet, even in that case, our methodology has similar or better performances than its unweighted counterpart.

Collectively, our results demonstrate the better performance of probabilistic graphlets across a range of different enrichment metrics, as well as their capability to utilize the additional network topology resulting from lower measurements of biological signals (captured by lower probabilities on edges) to cluster network nodes and extract biological information. In contrast, unweighted graphlets struggle with increasing noise levels and as a result, their performance deteriorates when adding low-confidence interactions to unweighted networks.

3.3.3 Probabilistic graphlets capture different biological information

To further elucidate whether probabilistic graphlets are capable of capturing additional and different biological information compared to their unweighted counterparts, we compare the GO BP terms that we find enriched in the previous clustering and enrichment analysis (section 3.3.2) using Jaccard index (JI) (introduced in section 3.2.7).

For budding yeast and human PPI networks, we find that the enriched GO

BP terms enriched are different in both, low and high-confidence PPI networks (JI different to 1), i.e., probabilistic graphlets captured distinct biological information in these data. However, we see the GO BP terms enriched are more different in low-confidence PPI networks than in high-confidence ones (see 3.6 A-B). Similarly, we find that our methodology captures different GO BP enriched in both budding yeast GIS and CoEx networks. In these networks, we find the highest differences in the medium confidence ones (see 3.6 C-D). whereas the low confidence networks produce results either similar to the high confidence (for GIS), or the medium confidence ones (for CoEx). In addition, the results show that the number of clusters, k, affects the similarity of the enriched GO BP terms captured by each methodology. However, the trend remains consistent across most values of k tested, meaning that the differences in the biological information between the methods are similar (see 3.6). Most of the results discussed for GO BP annotations also hold for GO MF and CC terms (see Supplementary Figure A.11 and A.12).

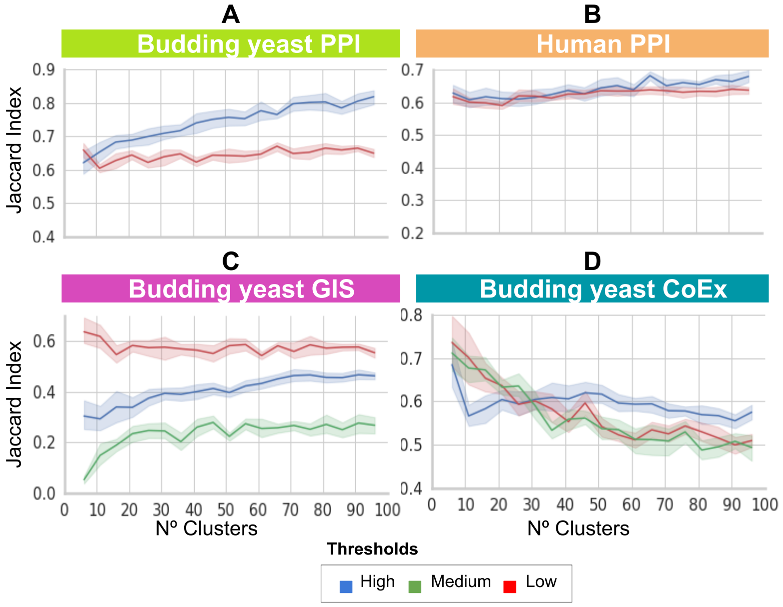


Figure 3.6. Probabilistic graphlets capture distinct information compared to their unweighted counterparts. Jaccard index (JI) of the BP GO terms captured by probabilistic and unweighted networks across different confidence thresholds and for different numbers of clusters. Lines represent the mean and the shaded area represents the 95% confidence intervals based on bootstrapping for ten repetitions.

Taken together, these results suggest that the differences between the biological information captured by each method depend on the threshold chosen and the biological network modeled. As seen in section 3.3.2, this dependence is a consequence of the impact of the chosen threshold on the topology of the resulting unweighted network, which also depends on the molecular data that is modeled. For instance, if we compare the number of edges in high-confidence PPI networks for humans and budding yeast. The resulting high-confidence human network, obtained from a common source and thresholded at ¿=0.7, encompasses 33.4% of all available edges, while the corresponding fraction for the budding yeast network is only 7.63%. Thus, as pointed out earlier, selecting a confidence threshold can strongly influence the network´s local structure and the biological information that graphlets can extract from its wiring patterns. However, this limitation does not affect probabilistic graphlets since all interactions in the network are taken into account and no information is discarded

3.3.4 Probabilistic graphlets uncover more specific biological information

Besides the overlap of the enriched GO terms, we also investigate the specificity of the biological information captured by each methodology. To this aim, we assess the mean level in the GO hierarchy of the enriched GO terms. The level of a term in the GO hierarchy is the distance of the term from the ontology root node, where higher levels describe more specific functions (more information can be found in Chapter 2, section 2.5.2).

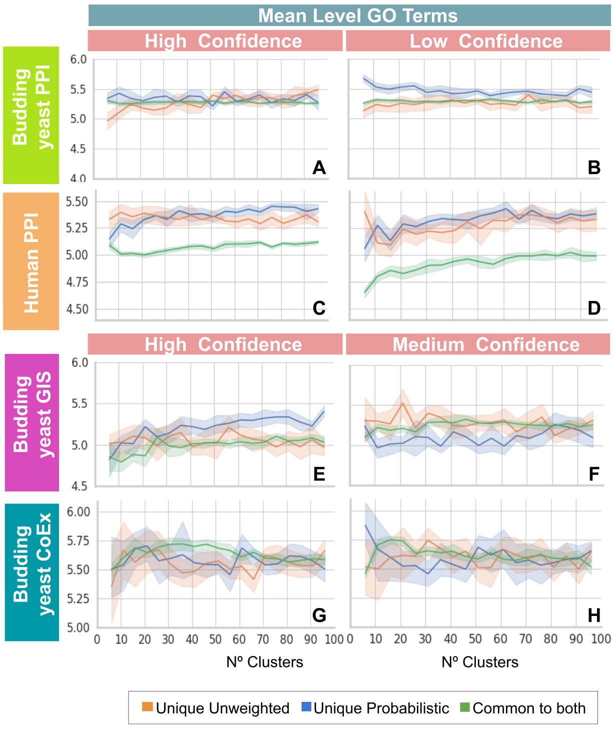


Figure 3.7. Mean level of the BP GO terms captured by probabilistic and unweighted networks in human PPI network and budding yeast PPI, GIS, and CoEx networks. Panels represent the mean level of the GO terms captured by each method or their intersection across a range of cluster sizes k, as well as the shaded area represents the 95% confidence intervals based on bootstrapping of the means across ten repetitions.

We find only small differences in the mean level of enriched GO terms when comparing terms unique to probabilistic graphlets, unique to unweighted graphlets, or their intersection in high confidence budding yeast PPI network. In contrast, probabilistic graphlets show clearly higher mean levels for the low confidence yeast PPI networks (see Figure 3.7 A-B). Thus, probabilistic graphlets capture more specific biological functions in low-confidence yeast PPI networks than their unweighted counterparts. Also, this suggests that our methodology benefits from retaining low-confidence interactions.

Regarding the low and high-confidence human PPI networks, we find that both, probabilistic and unweighted graphlets, individually capture more specific terms than their intersection (see Figure 3.7 C-D). This indicates that the specificity of the biological functions captured by each method is the same, but more interestingly, that they could be capturing complementary information that is not shared in their intersection.

For GIS networks, probabilistic graphlets capture more specific functional information than unweighted graphlets, as do their intersection for high-confidence interactions. However, we do not see this for medium confidence interactions (see Figure 3.7 E). Lastly, as previously seen for the enrichment measures, both probabilistic and unweighted graphlets show only small differences in the mean levels of enriched GO terms for CoEx networks (see Figure 3.7 G-H), which could mean that using the co-expression values (section 3.2.4) as edge-probabilities in CoEx networks or that CoEx networks, in general, do not benefit from a probabilistic model (as also found in section 3.3.3).

In conclusion, these results suggest that probabilistic graphlets capture equally or more specific biological information (GO terms) in some molecular networks when compared to the unweighted graphlet-based methodology. Moreover, as seen for human PPI networks, the two graphlet methodologies might be capturing complementary functional information. In the next section, we further investigate the biological information that is uniquely captured by our probabilistic graphlets.

3.3.5 Probabilistic and unweighted graphlets find different functions of a gene

To further understand the reasons for the differences in the biological information captured by each method, we choose a specific value of k and retrieve all enriched GO terms that are captured by each method. For these terms, we extract the genes responsible for the corresponding functional enrichments and divided them into three sets (section 3.2.7).

First, we observe that both, the total number of genes responsible for the functional enrichments as well as the proportion of uniquely enriched genes, are higher in probabilistic graphlets than in unweighted ones, i.e., our probabilistic methodology extracts biological information from a larger number of genes (see Figure 3.8). Moreover, Moreover, the existence of an intersection between the gene sets (Figure 3.8, yellow) indicates that the methodologies extract different and complementary biological information from the same set of genes.

In summary, these results demonstrate that the two methods identify different functional information from different genes, as well as complementary functional information from the same genes. Moreover, it suggests that the two approaches may complement each other to extract biological information from the molecular interaction datasets. This complementarity is a consequence of the fact that pGDV vectors can capture functional information that cannot be captured by GDV vectors, and vice versa (we investigate these topological differences in Supplementary Section A.1.1).

3.3.6 Robustness analysis

We also test the robustness of our methodology in capturing biological information from different networks by counting how often an enriched GO term is specific to or shared between the ten cluster repetitions performed for each network. Interestingly, across all networks and GO annotation domains, probabilistic graphlets are more robust in finding the same GO terms throughout all ten independent repetitions when compared to unweighted graphlets (Supplementary Figures A.13-A.15). This could once again be a limitation of thresholding probabilistic networks when creating unweighted representations, as specific interactions may or may not be retained, which can result in more divergent clusterings when compared to probabilistic graphlets.

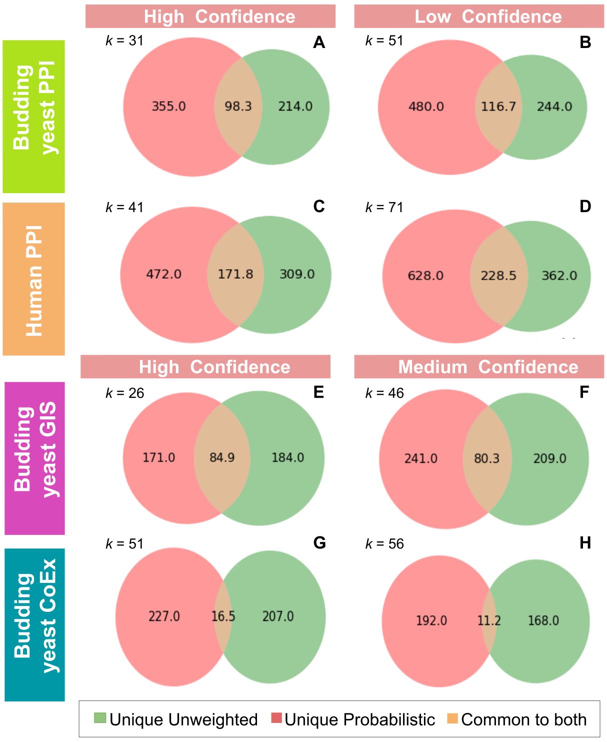


Figure 3.8. Probabilistic graphlets capture different biological information than unweighted graphlets from the same set of genes. Each panel represents a Venn diagram with the mean count of genes responsible for enriched GO terms (BP) uniquely captured by each methodology in different networks (rows) and at different confidence thresholds (columns), across 10 repetitions. The value of k for each network is fixed based on the rule-of-thumb ($k \approx \sqrt{\frac{N}{2}}$, where N represents the number of nodes in the network) and is displayed in each panel.

To further assess the robustness of our method in a biological context, we select all BP GO terms captured by either graphlet methodology across all ten repetitions. We select BP GO terms over MF and CC ones since they represent a higher-level perspective of biological functions (details in Chapter 2, section 2.5.2). Then, we generate simplified GO (sGO) terms using a two-step procedure (see section 3.2.8 for details) and identify all sGO terms that are specific to either graphlet methodology.

In budding yeast and human PPI networks, probabilistic graphlets identify a high number of biological response processes, such as response to oxygen-containing compounds, response to heat, response to hormones, and others. Furthermore, we find that response terms are detected significantly more often using the probabilistic methodology for low-confidence PPI networks (budding yeast PPIs: p-value of 0.012 and human PPIs: p-value of 0.01). These results suggest that our methodology is able to more reliably capture condition-specific biological processes that are only activated when cells are responding to specific stimuli and that therefore might benefit more from a probabilistic nature modeling approach. A possible explanation for the enrichment of these response terms is that the corresponding PPIs are less likely to be experimentally detected due to them occurring only in specific conditions [30], which in turn would result in these interactions being removed from high confidence networks (if the associated scores are too low). Hence, such interactions would only be included in unweighted networks when a low confidence threshold is applied, but would then be indistinguishable from spontaneous interactions that are likewise reported with low confidence.

In the case of GIS networks, when we examine the union of all sGO terms detected regardless of the confidence threshold applied, we find that 79.36% of these sGO terms are uniquely identified by probabilistic graphlets, while the rest are captured by both methodologies or uniquely by the unweighted one (6.34% and 14.28%, respectively). Among these 79.36% sGO terms, we find that functions related to the cytoskeleton are particularly frequent (15.34% of these sGO terms). For instance, biological functions related to microtubule-based movement, spindle localization, and cellular component assembly. Once again, these findings may suggest that processes related to behavior and control of the cytoskeleton can benefit from probabilistic models, perhaps due to influences of the cell cycle on the detectability of interactions [186, 187, 188].

Lastly, for CoEx networks we do not find any particular differences between the sGO terms captured by either methodology, which is in line with our previous observations regarding enriched clusters, GO terms, and genes (Figure 3.5) as well as gene signatures (Figure 3.8 G-H).

We once again observe that the chosen confidence threshold has a large impact on the information captured by each method. For example, in GIS networks, unweighted graphlets do not find any enriched terms for the low-confidence network, nor do they identify any of the terms found by probabilistic graphlets for the high-confidence network. Moreover, in the case of CoEx, we observe that most of the terms captured by unweighted graphlets are also found by probabilistic graphlets when we change from a high to a medium confidence threshold (for instance, "response to drug" or "response to nitrogen compound." As seen in previous sections, these results highlight the importance of thresholding and its influence on the topology of the data as well as the biological information that can be decoded from the network.

We also observe that probabilistic graphlets capture more stable biological information when comparing the same biological network across different confidence levels, while unweighted graphlets rarely find the same or related terms (Supplementary Table A.13). Thus, as seen previously, probabilistic graphlets are less affected by the introduction of noise due to low confidence interactions and are able to consistently identify relevant biological processes, regardless of the chosen confidence threshold.

3.4 Conclusions

Taken together, our results demonstrate that probabilistic graphlets robustly manage low signal topology information without sacrificing their ability to recover relevant biological information from the network, while unweighted graphlets used in thresholded networks are sensitive to noise, as well as the chosen threshold and its impact on network topology. Moreover, we find that probabilistic graphlets can robustly extract condition-specific processes such as stress response mechanisms, which in turn benefit from the use of probabilistic models.

Chapter 4

A functional analysis of omic network embedding spaces reveals key altered functions in cancer

In this Chapter, we propose to explore the functional organization of the gene embedding space from a functional perspective. We introduce a new function-centric approach, the *Functional Mapping Matrix* (FMM), and apply it in the context of cancer research. First, we evaluate the ability of the FMM in capturing the functional organization of the PPI tissue-specific embedding spaces of the most prevalent cancers in humans (breast, prostate, lung, and colorectal cancer) and their corresponding control tissues (breast glandular cells, prostate glandular cells, lung pneumocytes, and colorectal glandular cells). Then, we use our FMM-based method to investigate the changes in the functional organization of the cancer embedding spaces with respect to their control spaces, we exploit these changes to predict new cancer-related functions. Finally, we extend our FMM-based method to predict new cancer-related genes. he results of this chapter is published in Doria-Belenguer *et al.,* [189] and were presented at the ECCB'21 conference.

4.1 Motivation

In cancer research, different network embedding algorithms have been used to identify cancer-related genes [125], to subtype cancers [161], to stratify patients [162] and to repurpose drugs [163]. These algorithms include NLP-inspired methods, e.g., DeepWalk [19] and node2vec [20]; and matrix factorization-based approaches. In particular, NMTF is an extension of NMF and a well-known machine learning (ML) technique introduced for co-clustering and dimensionality reduction [21]. Unlike NMF, which factorizes the matrix representation of a network into two low-dimensional non-negative matrices, NMTF generates the embedding space by decomposing it into the product of three non-negative matrices, providing more degrees of freedom in the data modeling and analysis than NMF does [21]. One of the advantages of NMTF over deep neural network-based ML approaches is that it requires way fewer parameters to tune, thanks to the careful modeling of the relationships between the data points that it takes as input. As shown by Xenos *et al.,* [89], the molecular network embedding space produced by NMTF can have valuable properties, e.g., orthonormality, that may lead to an easier interpretation and deeper scientific insight [168].

However, current approaches for mining embedded biological networks use the genes' embedding vectors as input to machine learning algorithms to perform downstream tasks. These gene-centric approaches have demonstrated their potential in identifying new gene mutations in cancer cells involved in the initiation and progression of the disease [190]. However, they offer incomplete analyses of cancer data, since they do not take as input the functional implications of such genomic variations. Thus, changing the gene-centric paradigm to a functional-based one could be key to revealing additional functional information about cancer.

To improve our understanding of cancer, we generate cancer and control (healthy) gene embedding spaces by applying the NMTF algorithm to the corresponding tissue-specific PPI networks (detailed below). Then, to explore these gene embedding spaces from a functional perspective, we propose to embed biological functions, represented by GO BP annotations [191], into these gene embedding spaces. Finally, we capture the functional organization of a given gene embedding space with our new FMM, which encodes the mutual positions of the biological function embedding vectors in the space. First, we use our FMM-based method to identify the optimal dimensionality of cancer and control gene embedding spaces. Then, we apply the FMM to explore the functional changes in the most prevalent cancers (breast, prostate, lung, and colorectal) compared to their corresponding control tissues. We find that the changes in the distances between the embedding vectors of biological functions in cancer compared to the control embedding space are related to cancer. Indeed, we observe that cancer changes the distances between embedding vectors of cancer-related biological functions, while it preserves the positions of other biological functions. We exploit this observation to predict novel cancer-related functions, e.g., alternative translational mechanisms, or the response to unfolded protein accumulation. Moreover, we find a set of 8 annotations that are altered in all four cancer types. These annotations describe important cellular functions that may be commonly altered in different cancers, e.g., stress-activated MAPK cascade. Also, we demonstrate that our approach is not restricted only to functionally-based analyses of cancer, but also that it can be used to mine for new genomic knowledge from the embedding space. For instance, we use it to identify novel cancer-related genes, i.e., PRDM11, C9orf72, MINDY3, and H4C6, that could have an important role in the studied cancer types.

4.2 Materials and Methods

4.2.1 Biological datasets

Tissue-Specific PPI networks. We analyze cancer and control tissue-specific PPI networks that we generate by using the same methodology as in [1]. To this end, we collect the experimentally validated PPIs of human from BioGRID v.4.2.191 [41]. We model this human PPI data as a PPI network, in which nodes represent genes (or equivalently in this study, their protein products) and edges connect the nodes (genes) whose corresponding proteins physically bind. We use this generic human PPI network to generate our tissue-specific PPI networks. Following [1], we collect the tissue-specific gene expression data for breast, prostate, lung, and colorectal cancer tissues, as well as their corresponding control tissues of origins (breast glandular cells, prostate glandular cells, lung pneumocytes, and colorectal glandular cells, respectively) from the Human Protein Atlas (HPA) database v.20.0 [192]. For each tissue, we only consider the genes whose expression value is available in the HPA and that have at least one PPI in the generic human PPI network. We generate our eight tissue-specific PPI networks, in which nodes are genes that are expressed in the corresponding tissue, and two nodes are connected by an edge if they interact in the generic human PPI network (details about PPI tissues-specific networks can be found in Chapter 2, section 2.1). The network statistics of the tissue-specific networks are presented in Supplementary Table B.1.

Network Representation. We represent the tissue-specific PPI networks with their PPMI matrices, X, where each entry in the matrix contains information about how frequently two nodes co-occur in a random walk in the corresponding PPI network (details about PPMI matrices can be found in Chapter 2, section 2.7.1). Following Xenos *et al.,* [89], we use the DeepWalk closed formula by Perozzi *et al.,* [19] (see formula 2.15) with its default settings to compute the PPMI matrix. This formula can be interpreted as a diffusion process that captures high-order proximities between the nodes in the network; hence, PPMI is a richer representation than the adjacency matrix [89]. As a result of the extra information encoded in the PPMI, its corresponding embedding spaces better capture the functional organization of the cell than the ones generated by using the adjacency matrix (the details of this comparison are presented in Supplementary section B.2.1).

Biological Annotations. We use the GO BP annotations of genes' biological functions in a cell [191] (functional annotations are introduced in Chapter 2, section 2.5.2). We collected the experimentally validated GO BP annotations of genes from NCBI's web server (collected on 28 September 2021).

4.2.2 Definition of cancer-related biological annotations

Computational cancer research is usually based on computationally processing information about genes and not their annotations. Although a standard definition of a cancer driver (oncogene) exists [193], there does not exist a standard definition of a cancer-related GO BP term. Oncogenes are a functionally heterogeneous group of genes whose products regulate multiple cellular processes [194]. Despite this heterogeneity, oncogenes also participate in common molecular mechanisms that are known to be cancer-related, e.g., cell proliferation [195]. Thus, we propose to consider as cancer-related the most representative biological functions in which the oncogenes participate (detailed below).

We download the set of all 725 genes considered to be oncogenes in COSMIC [196] (collected on 01 December 2021). We find the most representative biological functions of these oncogenes by performing an enrichment analysis of our oncogenes set in GO BP functions (based on the hypergeometric test [197], see Chapter 2, section 2.5.2). A GO BP annotation is considered to be significantly enriched in our set of oncogenes, compared to all other genes if its enrichment p-value is lower than or equal to 5% after correction for multiple hypothesis testing [198]. We find 104 significantly enriched GO BP annotations in our set of oncogenes: these are our *cancer-related annotations*. To validate our set of cancer-related annotations, we calculate the Lin's semantic similarity [136] between our set of cancer-related functions and the set of 135 "cancer hallmark" annotations defined by [199]. With an average Lin's semantic similarity between the sets of 0.67, (see Supplementary Figure B.1) we conclude that the two sets are highly functionally related, i.e., our set of cancer-related annotations is related to the cancer hallmarks.

4.2.3 Embedding the PPI networks

To embed genes according to the PPMI matrix representation of a molecular network, X, we use NMTF to decompose X as the product of three non-negative factors, $X \approx P \cdot S \cdot G^T$, where the set of the rows of the matrix $P \cdot S$ defines the set of embedding vectors of the genes, E, and the set of the columns of G defines the basis, B, of the space in which the genes are embedded [200] (Figure 4.1 A illustrates the NMTF factorization on two different PPMI matrices, cancer, and control). Importantly, we apply the orthonormality constraint to the basis-defining matrix ($G^TG = I$), since it leads to minimal co-linearities (hence, minimizing the dependencies) between the vectors of the basis, B, of the embedding space [172]. The decomposition of the orthonormal NMTF is done by minimizing the function 2.7.2. To generate initial P, S, and G matrices, we use the SVD-based strategy [201]. This strategy makes the solver deterministic and also reduces the number of iterations that are needed to achieve convergence [201]. The quality of the factorization is usually evaluated by the Relative Square Error (RSE) between the input matrix, X, and its corresponding decomposition, PSG^T, as $RSE = \frac{\|X - PSG^T\|_F^2}{||X||_F^2}$. The iterative solver is normally stopped when the RSE is not increasing anymore.

4.2.4 Definition of the Functional Mapping Matrix

To explore the functional organization of the gene embedding space, obtained as detailed above, we introduce the *Functional Mapping Matrix* (*FMM*). This matrix captures the mutual positions of the functional annotations that we embed in the gene embedding space. In particular, we obtain an FMM by taking as input: the matrix factor, G, which contains the basis, B, of the gene embedding space, and the relation-matrix between the genes and their functional annotations, A, in which entry $A[a_i, g_j]$ is one if annotation a_i annotates gene g_j, and it is zero otherwise. First, we generate the embedding vectors of the functional annotations in the gene embedding space by decomposing the matrix A as the product of two matrix factors, U and G^T, as $A \approx UG^T$, where rows of matrix U (that we call u_i) are the embedding vectors of the annotations, a_i, in the gene embedding space defined by the basis, B (illustrated in Figure 4.1 for two embedding spaces, cancer, and control). Note that, since matrix A is known and matrix factor G is computed as explained in section 4.2.3, we can obtain U by: $U \approx (G^T)^{-1}A$, where $(G^T)^{-1}$ is the Moore-Penrose pseudoinverse of G^T [202]. Finally, the FMM is obtained by computing the pairwise cosine distances between all pairs of the embedding vectors u_i of the annotations a_i (the bottom panel of Figure 4.1 A illustrates two examples of FMMs). In particular, each entry FMM$[i, j] = cos(u_i, u_j)$ corresponds to the cosine distance between the embedding vectors u_i and u_j of the annotations a_i and a_j. Thus, the resulting FMM is a symmetric distance matrix that captures the mutual positions, that henceforth we call *distances*, between the annotation vectors in the embedding space. We choose cosine distance over other distance measures, e.g., the dot-product, since it is a well-known normalized measure [203], which permits direct comparison between different FMMs, i.e., we do not need any normalization step after computing the FMM.

4.2.5 Measuring the similarity of functional organization of the embedding spaces by using their FMMs

For a pair of embedding spaces, we measure the similarity of their functional organization by computing the relative squared error (RSE) between their FMMs. We use the following method to find the smallest number of dimensions, which we call the "optimal dimensionality", after which the functional organization of the gene embedding spaces, as measured by the RSE between the FMMs with increasing numbers of dimensions, does not change anymore. First, we produce the gene embedding space of each cancer and control, tissue-specific PPI networks by using the NMTF algorithm (detailed in section 4.2.3) with different dimensionalities (from 50 to 300 dimensions with a step of 50). Then, we obtain the embedding vectors of each of the GO BP annotations in each of the cancer and the corresponding control gene embedding space and then capture the difference in the position of a GO BP annotation between cancer and control space, measured by our FMM (detailed in section 4.2.4). By tracking the RSEs of the FMMs across dimensions (from 50 to 300 dimensions with a step of 50), we find that the distances of the annotation embedding vectors converge to a stable, i.e., non-changing functional organization, after 200 dimensions for all tissue-specific PPI network embedding spaces (RSE between their FMMs plateaus, i.e., stops decreasing, see Supplementary Figure B.2). In the analysis presented below, we use the optimal dimension of the embedding space that we obtained as described here (for all tissue-specific PPI networks, their optimal dimensions are presented in Supplementary Table B.4).

4.2.6 Evaluating the functional organization of an embedding space with its FMM

From a gene-centric perspective, an embedding space is considered to be functionally organized if genes that participate in similar biological functions are located close in the space [204]. This organization is commonly evaluated by applying various types of clustering methods to the embedding vectors of the genes in the space, followed by functional enrichment analyses of the genes that the clustered vectors correspond to [1] (detailed in Chapter 2, section 2.5). Here, we propose to examine the functional organization of the embedding space from a function-centric perspective. Similar to the gene-centric perspective, we consider an embedding space to be functionally organized if semantically similar annotations, i.e., annotations with high Lin's semantic similarity are embedded close in the space. To evaluate it, we apply our FMM to capture the distances of all pairs of the embedding vectors of the functional annotations in the embedding space (detailed in section 4.2.4).

Then, we analyze the link between the functional similarity of the annotations, measured by their pairwise Lin's semantic similarity, and the distances of their embedding vectors in the embedding space by performing two different experiments. We compute the PCC [205] between the mutual positions of all pairs of annotation vectors in the embedding space, i.e., the cosine distances over all pairs of annotation

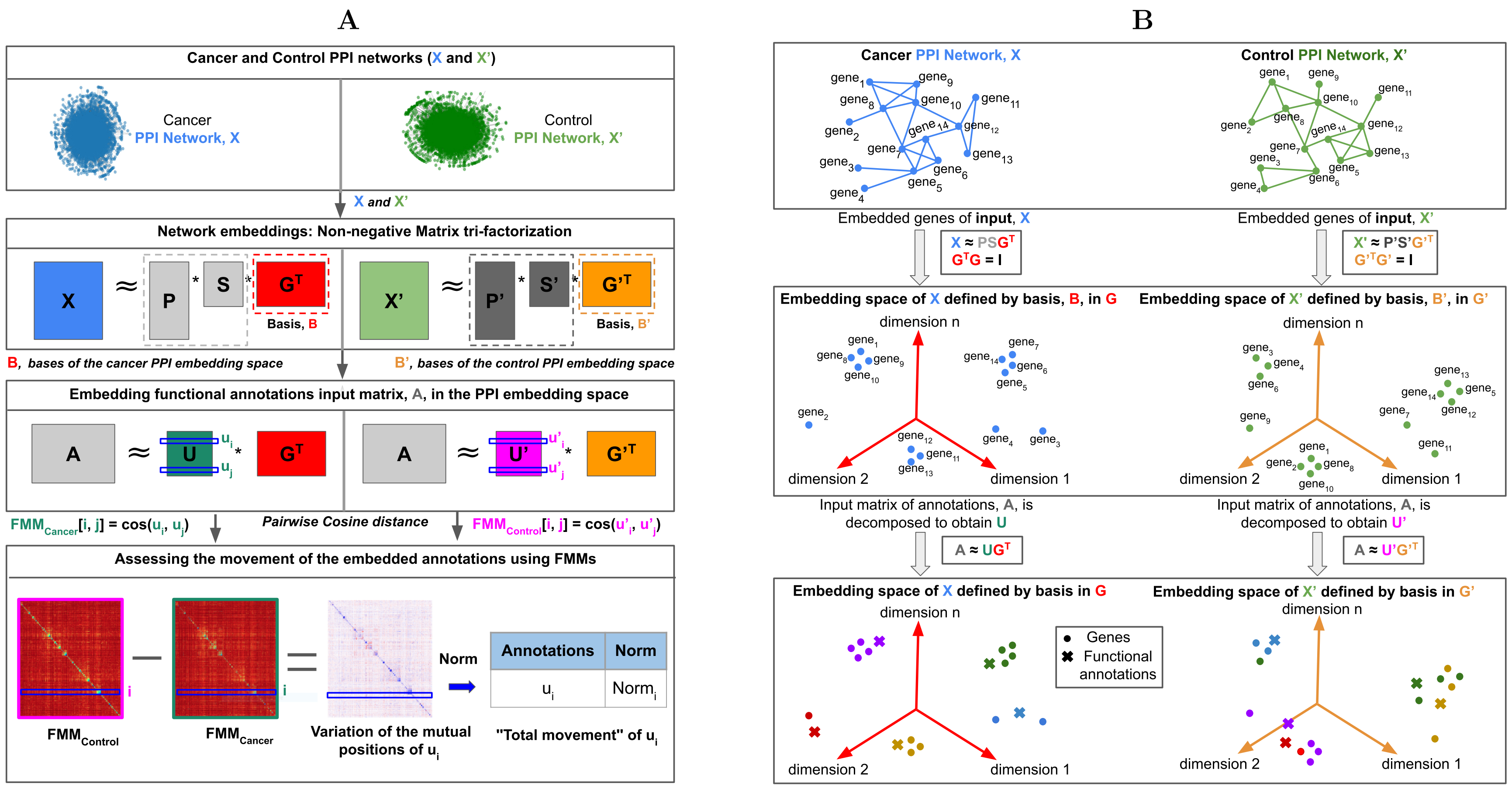

A
Cancer and Control PPI networks (X and X')
Cancer PPI Network, X
Control PPI Network, X'
X and X'
Network embeddings: Non-negative Matrix tri-factorization
X ≈ P * S * G^T
Basis, B
X' ≈ P' * S' * G'^T
Basis, B'
B, bases of the cancer PPI embedding space
B', bases of the control PPI embedding space
Embedding functional annotations input matrix, A, in the PPI embedding space
A ≈ U * G^T
u_i
u_j
A ≈ U' * G'^T
u'_i
u'_j
FMM_Cancer[i, j] = cos(u_i, u_j)
Pairwise Cosine distance
FMM_Control[i, j] = cos(u'_i, u'_j)
Assessing the movement of the embedded annotations using FMMs
FMM_Control
FMM_Cancer
i
Variation of the mutual positions of u_i
Norm
Annotations
Norm
u_i
Norm_i
"Total movement" of u_i
B
Cancer PPI Network, X
Control PPI Network, X'
gene_1
gene_2
gene_3
gene_4
gene_5
gene_6
gene_7
gene_8
gene_9
gene_10
gene_11
gene_12
gene_13
gene_14
Embedded genes of input, X
X ≈ PSG^T
G^TG = I
Embedded genes of input, X'
X' ≈ P'S'G'^T
G'^TG' = I
Embedding space of X defined by basis, B, in G
Embedding space of X' defined by basis, B', in G'
dimension n
dimension 1
dimension 2
Input matrix of annotations, A, is decomposed to obtain U
A ≈ UG^T
Input matrix of annotations, A, is decomposed to obtain U'
A ≈ U'G'^T
Embedding space of X defined by basis in G
Embedding space of X' defined by basis in G'
● Genes
✖ Functional annotations

Figure 4.1. A. Illustration of our new FMM-based method:. For a pair of cancer and control tissues, we construct their tissue-specific PPI interaction networks as explained in section 4.2.1 above (in green and blue for cancer and control, respectively). These networks, represented by their PPMI matrices, X and X', are decomposed as the products of three factors: P, S, and G^T for cancer, and P', S', and G'^T for control, where the set of all rows of G^T and G'^T defines the basis B and B', respectively (illustrated in the second panel from the top). From these matrix factors, we use the bases matrix of the resulting NMTF-based embedding spaces, G^T and G'^T, to generate the matrices U and U', whose i^{th} row are the embedding vectors u_i of annotation a_i in the cancer and control embedding spaces defined by the bases, B and B', respectively (illustrated in the third panel from the top). We capture the distances (cosine distances) between the embedding vectors of all pairs of annotations, in each cancer and control embedding space, by computing FMMs as defined in section 4.2.4 and illustrated at the "Pairwise Cosine distance" line between the two panels at the bottom of the figure. Then, we subtract the cancer and control FMM matrices, FMM$_{Control}$ - FMM$_{Cancer}$, to detect the changes in mutual positions of the embedding vectors u_i of each annotation a_i between cancer and control embedding spaces. Finally, to have the score of "movement" for u_i (illustrated in section 4.2.7), we apply the Euclidean norm to the rows of the matrix identified as "Variation of the mutual positions of u_i" in the bottom panel of the figure. **B. Toy example of our new FMM-based method:** The first panel shows a toy example of cancer and control PPI networks. The second panel shows a three-dimensional (3D) illustration of the embedding spaces of the toy example of cancer and control networks generated by our NMTF framework. The third panel shows the embedding vectors of the biological functions in the aforementioned cancer and control embedding spaces. The colors in the third panel represent the biological functions of the genes.

embedding vectors, and the Lin's semantic similarities over all pairs of annotations. Hence, a negative correlation coefficient indicates that those annotations that are embedded close in the space (lower cosine distance) tend to be functionally similar (high Lin's semantic similarity). Also, we apply the k-medoid algorithm [206] to cluster the annotations based on the distances of their vectors in the embedding space, as captured by our FMM. To define the number of clusters, we use the rule of thumb [18], $k = \sqrt{(}n/2)$, where k corresponds to the number of clusters and n to the number of annotations. Finally, we measure the intra and inter-cluster Lin's semantic similarity for the obtained clusters to assess if the annotations whose embedding vectors cluster in the embedding space are similar in biological function.

4.2.7 Quantifying the "movement" of the annotation embedding vectors in cancer and control embedding spaces

We propose to quantify the changes in the mutual positions (distances), which we call "movement," of the annotation embedding vectors in two different gene embedding spaces defined by bases, B and B'. In this study, we analyze the "movement" of the annotation embedding vectors in cancer and control embedding spaces. To this end, given the pairwise cosine distances of the annotations embedding vectors in the cancer and control embedding spaces, FMM_{Cancer} and $\text{FMM}_{Control}$, we quantify the change in the distance between two embedding vectors of annotations u_i and u_j as $\text{FMM}_{Control}[i, j]$ - $\text{FMM}_{Cancer}[i, j]$. This distance is negative if u_i and u_j are farther in the cancer embedding space than in the control embedding space, positive if they are closer, and zero if there is no change between their positions in the embedding space of cancer and control. By taking all the pairwise distances over all i and j, $\text{FMM}_{Control}[i, j]$ - $\text{FMM}_{Cancer}[i, j]$, we define the distribution of pairwise "movements" (see Supplementary Figure B.3). We define that two annotation embedding vectors, u_i and u_j, are "moving significantly apart" in the embedding space of cancer if their distance is greater than or equal to the 95^{th} percentile of the aforementioned distribution. In contrast, we define that they are "moving significantly closer" in the embedding space of cancer if their distance is smaller than or equal to the distance that corresponds to the 5^{th} percentile of the distribution.

To identify the annotations whose embedding vectors change the most between the cancer and control embedding spaces, first, we calculate the distance between the embedding vectors of each annotation u_i in the control and the cancer embedding spaces, that we call $\text{FMM}_{Control}[i]$ (which is the i^{th} row of matrix $\text{FMM}_{Control}$) and $\text{FMM}_{Cancer}[i]$ (which is the i^{th} row of matrix FMM_{Cancer}, respectively. So the coordinate of vector $\text{FMM}_{Control}[i]$ contains the cosine distances of u_i to all other annotation embedding vectors in the control embedding space. Then, for each annotation embedding vector, u_i, we define the "movement vector" as $D[i] = \text{FMM}_{Control}[i]$ - $\text{FMM}_{Cancer}[i]$. Hence, the "movement vector" contains the differences of the mutual positions in cancer compared to control embedding space (cosine distances) between u_i and all other annotation embedding vectors. Next, we define the "total movement" of annotation, u_i, as the Euclidean norm of its corresponding "movement vector," $D[i]$. In this way, for each annotation, u_i, we define the score of its "total movement" in cancer over control, which is high when its distance to the other annotations changes between the cancer and control embedding spaces (that we call *shifted*) and it is close to zero when it does not change (that we call *stable*). By considering the "total movement" of all annotations, we define the "total movement distribution" (see Figure B.4). We consider as *shifted biological functions* those functional annotations whose embedding vectors' "total movement" is two standard deviations above the mean of the "total movement distribution". In contrast, we define as *stable biological functions* those functional annotations whose embedding vectors' "total movement" is two standard deviations below the mean of the "total movement" distribution.

4.2.8 Distances between the embedded entities in the embedding space

We use the cosine distance to determine the distance between the embedding vectors of two entities (genes or functions in this study) in the same gene embedding space defined by basis B. We recall that in the embedding space defined by B, the embedding vector of gene g_i is the i^{th} row of matrix $P \cdot S$, and that the embedding vector of annotation a_j is the j^{th} row of matrix U (detailed in section 4.2.4 and illustrated in Figure 4.1 B). Before using the cosine distance, we confirm that the embedding vectors of the biological functions (GO BP terms) are significantly closer in space to the embedding vectors in the same space of the genes that they annotate than to the embedding vectors of other genes (Mann-Whitney U *p-value* ≤ 0.05, see Supplementary Table B.5). This confirms that annotations and genes are functionally organized in the embedding space.

4.3 Results and Discussion

Inspired by Malod-Dognin *et al.,* [1] who, in a gene-centric analysis, observed that cancer-related genes are the most rewired between cancer and control embedding spaces and used this property to predict novel cancer-related genes, we use our FMM-based method to confirm that the embedding spaces of both, cancer and control, are functionally organized and that this organization changes between cancer and control. We exploit this observation to predict new cancer-related functions, which we validate by analysis of their enrichment in known cancer-related functions (detailed below), automatic literature search, and manual literature curation for the most promising predictions (section 4.3.2). Moreover, we go beyond and exploit the "movement" of the annotation embedding vectors to predict new cancer-related genes (section 4.3.3), finding four new cancer-related genes, which we validate by literature curation and retrospective analyses of patient survival, but whose role with cancer has yet to be experimentally validated.

4.3.1 Cancer alters the functional organization of the healthy cell embedding space

Here, we focus on applying our FMM-based method to confirm that the embedding spaces of both, cancer and control, are functionally organized (detailed in section 4.2.6). To this end, we generate the embedding spaces of the most prevalent cancers (breast, prostate, lung, and colorectal cancer) and their control tissues (breast glandular cells, prostate glandular cells, lung pneumocytes, and colorectal glandular cells) by applying the NMTF algorithm on the corresponding tissue-specific PPI networks (detailed in sections 4.2.1 and 4.2.3). Then, we use our FMM-based method to embed GO BP terms into these gene embedding spaces and to capture their distances over the cancer and control embedding spaces (detailed in section 4.2.4). By analyzing the FMM of each embedding space, we find that the annotation embedding vectors that cluster together based on their cosine distances in each space have, on average, Lin's semantic similarity 1.32 times larger than those that do not cluster together in space (see column "Fold" in Table 4.1 for the corresponding results for each embedding). Hence, the GO BP terms corresponding to the embedding vectors that cluster together in space are more functionally related than those whose embedding vectors do not cluster in space (see Table 4.1). Thus, both cancer and control embedding spaces are functionally organized. We further confirm this conclusion by comparing these results against a randomized experiment, i.e., when randomly rewiring the PPI networks (detailed in Supplementary section B.2.4). As expected, we find that annotations whose embedding vectors are close in these randomized spaces are not more functionally similar (as measured by the Lin's semantic similarity) than those whose embedding vectors are far in the space, i.e., they are not functionally organized in the randomized space (see Table 4.1 and Supplementary Table B.6).

Having confirmed that both embedding spaces, cancer, and control, for all four cancers, are functionally organized, we investigate if this organization changes between them. To do so, we assess if there are pairs of annotation embedding vectors whose distances in the embedding space are significantly altered in cancers (detailed in section 4.2.7). For the four studied cancers, we find an average of *72,326* (5% of the total number) of pairs that move significantly closer in the cancer space compared to control (see Figure 4.2 for an illustration of this variation). We find that this set of pairs (that are closer) is 1.3 times closer in the cancer space than in the control one. Similarly, we find the same percentage of pairs that move significantly apart in the cancer embedding space compared to the control. Here, we find that this set of pairs (that move apart) is 1.4 times farther in the cancer space in comparison to the control one. In conclusion, these results demonstrate that cancer alters the functional organization of the healthy (control) cell.

We have shown above that cancer alters the functional organization of the control PPI network embedding space by changing the distances of the annotation embedding vectors in the space. Now, we investigate how this change is related to cancer (and if it can be used to predict novel cancer-related functions). We use our FMM-based methodology to identify the annotation embedding vectors that change their distances (that we call "movement") between cancer and control embedding spaces. Then, we compare the "movement" of our set of cancer-related functions and the rest of the annotations. Interestingly, we observe that the embedding vectors of cancer-related functions move the most between cancer and control embedding spaces compared to those of other annotations. Indeed, these annotation vectors move on average 2.4 times more than the rest of the annotation embedding vectors in all four cancers (Mann-Whitney U test with p-value < 0.05). This suggests that the "movement" of the annotation vectors is related to cancer, i.e., it could be exploited to find new cancer-related functions (presented in the next section).

Embedding	**Intra**	**Inter**	**Fold**	**p-value**
Control breast	0.22	0.17	1.29	2.12×10^{-6}
Cancer breast	0.23	0.16	1.43	2.68×10^{-5}
Control prostate	0.24	0.17	1.41	2.24×10^{-6}
Cancer prostate	0.21	0.15	1.40	1.04×10^{-6}
Control colorectal	0.19	0.16	1.18	4.04×10^{-3}
Cancer colorectal	0.21	0.16	1.31	1.68×10^{-5}
Control lung	0.19	0.17	1.11	2.17×10^{-4}
Cancer lung	0.22	0.15	1.46	5.32×10^{-6}
Random Example	0.17	0.17	1.00	0.14

Table 4.1. The embedding spaces of the most prevalent cancers (breast, prostate, lung, and colorectal cancer) and their control tissues (breast glandular cells, prostate glandular cells, lung pneumocytes, and colorectal glandular cells) are functionally organized. The first column, “Embedding,” lists the tissues. The second column, “Intra,” shows the average Lin’s semantic similarity of those annotations whose embedding vectors cluster together based on their cosine distances in the embedding space. The third column, “Inter,” shows the average Lin’s semantic similarity of those annotations whose embedding vectors do not cluster together based on their cosine distances in the embedding space. The fourth column, “Fold,” displays how many times the average Lin’s semantic similarity of those annotations whose embedding vectors cluster together based on their cosine distances in the embedding space is higher than of those annotations whose embedding vectors do not cluster together. The fifth column, “p-value,” shows the p-value from a one-sided Mann-Whitney U test comparing the Lin’s semantic similarity between annotations whose embedding vectors cluster together and those with non-clustered embedding vectors. This table also includes an example of a randomly rewired PPI network (Random Example). The complete information with all the random tissue-specific PPI networks can be found in Supplementary Table B.6.

4.3.2 The “movement” of the annotations in the embedding space predicts cancer-related functions

Here, we exploit the “movement” of the annotations’ vectors to predict novel cancer-related functions. Following the approach detailed in section 4.2.7, we find two groups of annotations based on their “movement:” *shifted* and *stable* group of annotations (the numbers of GO BP annotations in the two sets for each of the four cancers are presented in Supplementary Table B.7). For these sets of annotations, we perform the hypergeometric test (with $alpha = 0.05$, [197]) to assess if they have significantly more, or significantly less cancer-related functions than the background set of genes (the background set of genes contains all genes that are in the corresponding tissue-specific PPI network). We observe that for three out of four cancers, the *shifted* annotations are significantly enriched in cancer-related functions (p-value of 0.85, 0.02, 0.02, and 0.04 for breast, colorectal, prostate, and lung, respectively). In contrast, the *stable* annotations are significantly depleted in these functions (p-value of 0.49, 0.88, 0.80, and 0.68, for breast, colorectal, prostate, and lung, respectively),

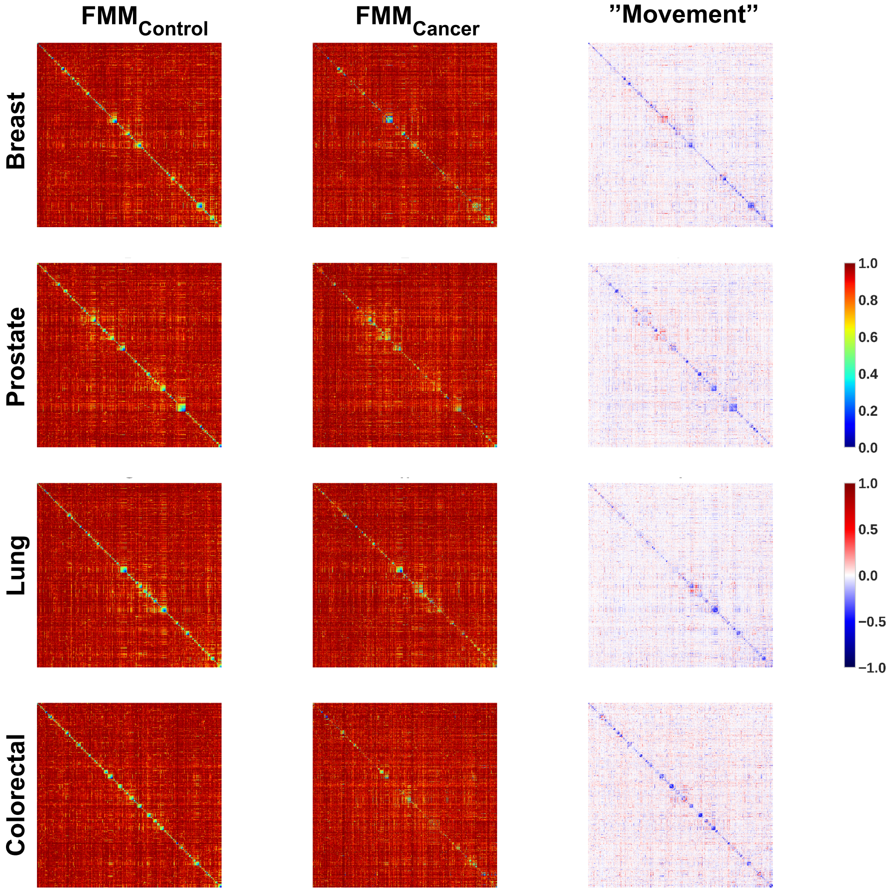


Figure 4.2. The embedding vectors of GO BP terms change their mutual positions in the cancer embedding space with respect to the control embedding space, for each cancer type (breast cancer, prostate cancer, lung cancer, and colorectal cancer) and its corresponding control. Heatmaps in the first and second columns show the cosine distances (mutual positions) between the embedding vectors of the GO BP annotations in control embedding space ($\mathrm{FMM}_{Control}$) and cancer embedding space (FMM_{Cancer}), respectively. Heatmaps in the third column show changes in the mutual positions of the embedding vectors of the functional annotations between cancer embedding space with respect to the control embedding space (computed as $\mathrm{FMM}_{Control}$ - FMM_{Cancer}).

i.e., they have a significant lower percentage of cancer-related functions than the background (see Figure 4.3). This observation does not hold only for the *shifted* annotations of breast cancer (p-value of 0.85). This discrepancy can be attributed to the type of cancer samples used in this analysis and to our definition of cancer-related annotations. While the TCGA's samples of colorectal, lung, and prostate are mostly from adenocarcinomas, over 99% of the TCGA's samples of breast cancer are from neoplasms (see Table 4.2). Indeed, as detailed in section 4.2.2, we use the COSMIC oncogenes to define our cancer-related GO BP terms. These oncogenes are mainly defined from adenocarcinomas samples; in particular, for breast cancer, only 8% of the samples in COSMIC come from neoplasms, while in TCGA, over 99% of the samples come from neoplasms. This highlights the importance of improving the definition of cancer-related functions to include different types of cancer of the same organ.

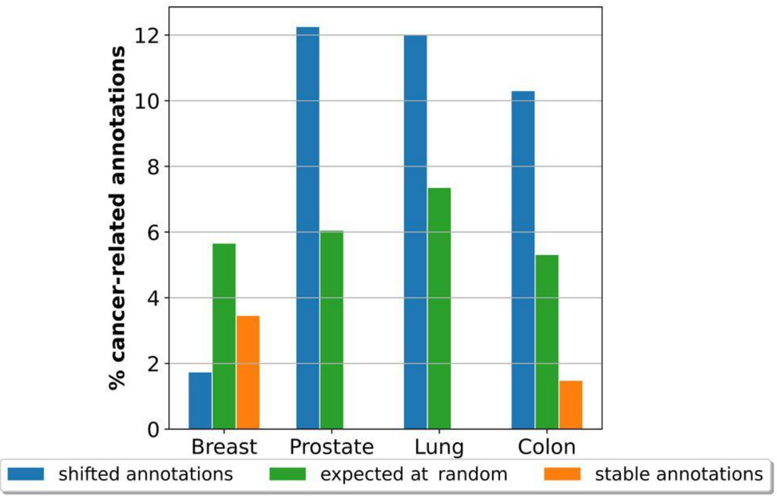


Figure 4.3. "Movement" in the embedding space is related to cancer. The panel contains the percentages of enriched cancer-related GO BP terms out of all GO BP terms (vertical axis) in the *shifted* annotations set (in blue), *stable* annotations set (in orange), and the expected by random (in green), for each cancer type (on the horizontal axis).

Despite the results discussed above, we also find several annotations in the *shifted* set that are not considered to be cancer-related according to our definition. In particular, we find that only 2 (2%), 5 (12%), 5 (10%), and 6 (10%) of the annotations in the *shifted* set are cancer-related for breast, prostate, lung, and colorectal cancer, respectively. Thus, to validate the remaining unknown to be cancer-related annotations, we extend the systematic approach used in the study by Ceddia *et al.,* [163] and conduct a systematic literature search in the PubMed database [207]. We automatically retrieve the number of scientific publications that associate each GO BP term with a specific cancer type. To do so, we search for co-occurrences between the GO BP term and the cancer type in the abstracts of PubMed publications. We find that 33 (58%), 31 (65%), 29 (63%), and 36 (52%) of these annotations have at least one publication demonstrating their role in breast, lung, prostate, and colorectal cancer, respectively. These high percentages of literature validation indicate that the remaining annotations, which we could not validate in the currently available literature, are candidates for novel cancer-related functions.

Finally, we do manual literature curation for the most promising predictions identified above. In particular, we rank the predictions by the magnitude of their "movement" and we investigate the top 10 most "moved" ones. We detect that, although these functions are not reported in the literature to be directly related to cancer, their link with cancer is clear: for instance, we find "the positive regulation of activated T cell proliferation" in breast cancer. This is a well-known regulation process in breast cancer development, and it could be connected to the "cooperation" of breast cancer cells with the immune cells [208]. Other examples include "cleavage furrow formation" and "mitotic spindle midzone assembly" in prostate and colorectal cancers, respectively. The failure of these processes during cell division has been associated with carcinogenesis [209, 210]. Finally, we find "the positive regulation of endodeoxyribonuclease activity" in lung cancer. A deficiency in this process is linked with most of the mutations and genomic alterations that are relevant to

Cancer	TCGA Project	# of patient samples	Disease Type
Breast	BRCA	1,098	1,095 neoplasms 3 adenocarcinomas
Prostate	PRAD	467	459 adenocarcinomas 8 neoplasms
Lung	LUAD, LUSC	1,062	533 neoplasms 529 adenocarcinomas
Colorectal	COAD, READ	456	389 adenocarcinomas 63 neoplasms

Table 4.2. The statistics for the tissue-specific PPI networks in this study. Column one, "Cancer," specifies the type of cancer that we analyzed; column two, "TCGA Project," gives the name of the project from TCGA that produced the data that we used; column three, "# of patient samples" specifies the number of patient samples in the project from column two; column four, "Disease Type," specifies the numbers of patient samples from the corresponding project with a specific cancer type.

cancer [211]. An extended discussion for the rest of the annotations in each cancer type top 10 predictions can be found in Supplementary section B.2.5.

4.3.3 The "movement" of cancer-related annotations in the embedding spaces predicts oncogenic genes

In this section, we investigate if the functions that are shifted in cancer (compared to control) can be used to identify novel cancer-related genes. To this aim, we first demonstrate that the embedding space captures the functions of a given gene by placing its embedding vector close (low cosine distance) to the embedding vectors of those GO BP terms that describe the gene's biological functions (detailed in section 4.2.8). We hypothesize that the alteration in the cosine distance between the gene embedding vector and the GO BP embedding vector may indicate that the gene is losing a function (if the distance increases), or that the gene is gaining a function (if the distance decreases). Hence, we prioritize as cancer-related those genes whose embedding vectors change their distances to the vectorial representations of the *shifted* functions in the embedding space the most.

To evaluate this hypothesis, we first assess if *literature-validated* genes (see the definition below) change significantly more their distances to our *shifted* functions than the background genes in the cancer space compared to control. To this end, similar to the methods explained in section 4.2.7, for each gene, we compute a vector with n positions, where n corresponds to the number of the "shifted" GO terms and in which each entry corresponds to the "movement" (change of mutual positions) of the gene and the GO term. Since this "movement" is bi-directional (getting closer or further), we use the absolute value of the "movement" at each coordinate of this vector, to keep only the magnitude of this "movement" independently of the direction of the "movement". Then, since all the values in the n-dimensional vector are now positive, for each gene we assign as its cancer-related score the maximum value (maximum magnitude of movement) in its corresponding vector. Hence, we define the maximum "movement distribution" of the gene embedding vectors as the set of all aforementioned maximum values of "movement." For each cancer type, we consider as *literature-validated* the genes with at least one publication in PubMed indicating their role in the corresponding cancer type. To do this evaluation, we apply the same systematic approach as the one used to validate the "shifted" annotations in section 4.3.2. In all four cancers, we find that literature-validated genes "move" significantly more towards or away (higher cancer-related score) from our *shifted* functions than the background genes (we compare these two "movement" distributions with Mann-Whitney U test with p-value < 0.05). Thus, we use this property to predict new cancer-related genes. We predict as cancer-related those genes that are above or at the 95^{th} percentile of the maximum "movement" distribution (see Supplementary Figure B.5). In this way, we predict as cancer-related 346, 234, 325, and 379 genes in breast, lung, prostate, and colorectal cancer, respectively, which we call *shifted genes*. In the rest of this section, we validate these predicted cancer-related genes in two ways: systematic literature curation and by retrospective analyses of patient survival curves (detailed below).

Gene name	PubMed Counts	Pan-Cancer Prognostic Marker
C9orf72	0	0
PIK3R2	2	0
TAF13	0	2
MINDY3	0	0
EIF5B	1	3
SSB	7	3
SGSM3	0	1
NKX3-1	314	0
RPS4X	0	2
FAM204A	0	1

Table 4.3. op 10 *shifted* genes (the most *shifted* ones) in prostate cancer. The first column, "Gene name," presents the gene names of the top 10 *shifted* genes. The second column, "PubMed Counts," contains the number of publications in Pubmed that relate the gene to prostate cancer. The third column, "Pan-Cancer Prognostic Marker," indicates how many cancer types the gene is considered to be a prognostic marker based on survival curves collected from the Human Protein Atlas [192].

We validate in the literature that 233 out of 346 (67%), 144 out of 234 (61%), 179 out of 325 (55%), and 187 out of 379 (49%) of our predictions are cancer-related in breast, lung, prostate, and colorectal cancer, respectively. Indeed, among our literature-validated predictions, we find well-known cancer genes, i.e., BRAF in breast cancer (225 publications), CASP8 in lung cancer (123 publications), or MSH6 in colorectal cancer (205 publications). Also, we assess if our cancer gene predictions are prognostic markers of patient survival, which we measure with patient survival curves (we collected the data from the Human Protein Atlas (HPA) [192]). We find that 16 (4.6%), 7 (2.9%), 4 (1.2%), and 17 (4.4%) of these genes are registered in the HPA as breast, lung, prostate, and colorectal cancer prognostic markers, respectively. Since these survival curves are based on differential gene expression analyses [212], we hypothesize that our method prioritizes genes that are not differentially expressed. Indeed, only 38 (11%), 85 (36%), 19 (6%), and 56 (15%) of our predicted cancer-related genes are deferentially expressed in breast, lung, prostate, and colorectal cancer tissues with respect to their corresponding control tissues, respectively (using expression data from TCGA projects, as detailed in Table 4.2). These results align with [1], who demonstrated that there exist important cancer-related genes (validated by wet-lab experiments) that are not differentially expressed in control and cancer. We hypothesize that the role of these genes in cancer could be connected with post-translational modifications (PTM) of their expressed proteins. These modifications modulate the functions and interactions of the proteins after translation [213] and have been reported in several cancer types, e.g., ovarian cancer [214] or skin cancer [215]. In conclusion, our method identifies genes whose transcriptional patterns have not changed and thus is complementary to the traditional differential expression analysis.

Finally, we go beyond the above validation and focus on the top 10 *shifted* genes (the most shifted ones) of each cancer type. We largely validate these top 10 shifted genes, either as cancer biomarkers (of prognosis) or as cancer-related in the literature (see Table 4.3 and Supplementary Tables B.8-B.10). Thus, we conjecture that the remaining four non-validated genes (PRDM11 in lung cancer, C9orf72 and MINDY3 in prostate cancer, and H4C6 in colorectal cancer) are also cancer-related. Indeed, PRDM11 is part of a broad family of transcriptional regulators, several of which are deregulated in cancer [216]. It is highly expressed in the lungs, as well as in peripheral blood immune system cells. Although it has been linked with the enhancement of lymphomagenesis [216], our study is the first one to suggest its role in lung cancer. Another example is MINDY3 in prostate cancer; MINDY3 codes for a protein that contains a caspase-associated recruitment domain and may be involved in apoptosis [217]. Even though it has been identified as a tumor suppressor in lung and gastric cancers [218], our study is the first to link it with prostate cancer. For the same cancer type, prostate cancer, we find C9orf72, a gene that has been associated with several neurodegenerative disorders [219]. Although its role in cancer is unknown, its participation in important cancer-related processes, such as autophagy [216] and inflammation [220], supports our observation that it may be cancer-related. Finally, we predict gene H4C6 as being involved in colorectal cancer, which is a member of the histone H4 family that encodes a replication-dependent histone. Although no publication relates this gene to cancer, its involvement in cellular senescence and mitotic prophase [217] suggests that this gene may have an important role in cancer progression. In conclusion, we introduce a method to predict new cancer-related genes based on their distance to the most *shifted* functional annotations in cancer over control molecular network embedding space. We validate our predictions of new cancer-related genes through literature curation and retrospective analyses of patient survival data. Importantly, these newly predicted cancer-related genes can not be identified by using the traditional differential-expression analysis.

4.4 Conclusion

By introducing our new FMM methodology, we initiate the investigation of the embedding spaces of the tissue- and disease-specific molecular networks from a functional point of view. In the Supplementary section B.2.2, we demonstrate that our FMM methodology better captures the functional interaction between GO BP terms than the traditional gene-centric approach. We show that our FMM can efficiently be applied to address different problems, i.e., to find the optimal dimensionality of the embedding space, to analyze the similarities between the functional organization of different embedding spaces (in this study, those corresponding to cancer and control), or to find the functional changes produced by cancer. Moreover, we use our method to predict four new cancer-related genes for which we found some literature indicating their involvement in cancer, but whose role in cancer has yet to be experimentally validated. Furthermore, our methodology could be easily applied to other bioinformatics tasks, such as patient and tissue stratification, or to uncover evolutionary similarities. Moreover, in the Supplementary section B.2.3, we demonstrate that our FMM captures the hierarchical organizations of the GO BP terms in network embedding spaces. However, extracting novel knowledge from that higher-level organization is left for future study. Finally, our new methodology is generic and can be applied to any discipline that analyzes embedded network data in which the embedded network nodes can be functionally annotated, e.g., social, or economic networks, paving the road to new algorithms for mining the data by utilizing the embedding space from a functional perspective.

Chapter 5

The axes of biology: a novel axes-based network embedding approach to decipher the fundamental mechanisms of the cell

While our FMM-based methodology changes the exploration of the embedding space from the genes' embedding vectors to the vectorial representation of their functional annotations, both approaches still focus on the organization of the embedded entities (genes and functional annotations) in the space to decipher biological information from embed networks. Thus, in this Chapter, we propose to change the perspective from the embedded entities to the space itself. In particular, we introduce a new method that uses the axes of the embedding space where the entities are embedded to capture the cell's functional organization from molecular networks. First, we evaluate if the axes of the embedding space uncover the cell's functional organization from species-specific PPI networks. Then, we analyze if the ability of the embedding axes to uncover the cell's functional organization is affected by the embedding spaces' topological properties, such as orthonormality and non-negativity. Also, we investigate the impact of dimensionality on the ability of the embedding axes to reveal the cell's functional organization. We investigate the functional coherence of the biological information captured by the axes and analyze them in the context of evolution. The content of this chapter is adapted from Doria-Belenguer *et al.,* [221] currently under submission.

5.1 Motivation

Recent approaches for deciphering biological networks are based on network embedding techniques [14]. These algorithms aim to find the vectorial representation of the network nodes in a low-dimensional embedding space spanned by a system of coordinates (a.k.a., embedding axes) while preserving the structural information of the network [14, 15]. Current gene-centric and functional-centric approaches for mining embedded networks uniquely use the organization of the learned vectorial representation of the genes and functions in the space to uncover the cell's functional organization. Hence, other possible information sources, such as the axes of the embedding space where these entities are embedded, remain unexplored.

Revealing the hidden information of a biological network requires not only an embedding algorithm but also methods to integrate their results into biologically interpretable models [222]. Network clustering analyzes the topological and functional properties of the molecular networks by grouping (clustering) together genes whose embedding vectors are in proximity in the embedding space. These clusters represent subgraphs of the original molecular network that exhibit significant clustering properties, i..e., genes within each cluster are more densely connected to each other than to genes outside the cluster. This allows for uncovering topological and functional modules within the molecular network [124]. To functionally interpret these clusters, current approaches rely on several curated ontologies, such as KEGG [223], Reactome [73], and GO [191]. Among these resources, GO has the largest number of concepts and records [131, 191]. GO terms are often used in functional enrichment analysis to evaluate the statistical over-representation of specific biological functions in the genes' clusters [224]. A closely related problem to network clustering is the summarization of such networks, which involves condensing the information contained in large-scale molecular networks by allowing visual interpretation of the clusters [124]. The state-of-the-art summarization techniques, such as spatial analysis of functional enrichment (SAFE) [225], construct a global bird's-eye view of functional organization underlying the molecular network by detecting regions with over-represented functional annotations and providing their visual representation.

Current clustering and summarization network methods have demonstrated their potential in functionally characterizing molecular networks, leading to a better understanding of the cells' machinery and biological processes [225, 124]. However, they present several limitations that hinder the identification and interpretation of the fundamental mechanisms of the cells, i.e., those mechanisms that summarize the most important functions of the cell. First, these methods uncover the cell's functional organization by identifying those functions that are statistically overrepresented in each cluster of genes. Although these clusters attempt to represent the different functional modules of the molecular network, they usually present high redundancy among their enriched functions [124]. Second, the number of functions that can be analyzed is restricted to a pre-defined set of functional annotations. Since the update rate of these annotations by the database curators is slow, it presents a bottleneck for their use [222, 131]. Third, the number of functions that are usually overrepresented in each cluster is large, i.e., their interpretability is not intuitive [133]. Finally, the current methods fail to generate a high-quality functional summary for those clusters of genes that lack functional information [124].

To overcome these limitations, we innovatively propose to use the axes of the embedding space to identify the fundamental mechanisms of a cell. Contrary to the current state-of-the-art approaches that focus on the organization of the genes' embedding vectors or the organization of the genes' functions embedding vectors in the embedding space to find these mechanisms, our method focuses on the axes of the embedding space itself. To identify these mechanisms, we generate the gene embedding spaces of different species by applying the NMTF and Deepwalk algorithms to the corresponding species-specific PPI networks. We apply the NMTF algorithm with and without orthonormality constraints to gain insights into their impact on the functional organization of the embedding space axes. Then, to untangle the biological information hidden in the resulting gene embedding spaces, we embed GO terms in the gene embedding spaces and associate them with the axes of each space.

For the first time, we demonstrate that the axes of the embedding space disentangle biological information from the space, with semantically similar GO terms associated with the same axis, i.e., each axis represents a specific biological function. Moreover, we show that the axes of the orthonormal NMTF gene embedding spaces better untangle biological information from the embedding space than Deepwalk (with orthonormal NMTF associating on average 1.6 times more GO terms to the axes than Deepwalk) and that this information is more coherently stratified across the axes (the GO terms associated to the same axis from the embedding space of orthonormal NMTF have on average 1.2 times higher semantic similarity than the ones associated to the same axis from the embedding space of Deepwalk). We demonstrate that this observation is connected to the properties of the NMTF embedding spaces, such as orthonormality and positive constraint, which improve the organization of such embedding space.

Furthermore, we use our novel axes-based method to define the optimal dimensionality of different species-specific PPI embedding spaces. For this optimal dimensionality, we explore the meaning of the GO terms associated with their axes. To this aim, we apply an NLP-based approach to summarize all the GO terms that are associated with a given axis into a higher-level functional annotation that we term ASFA. We find that these ASFAs define the cell's fundamental mechanisms, and we evaluate their coherence by literature curation. Interestingly, the ASFAs not only define coherent biological processes, such as the cellular response to the accumulation of misfolded proteins or the sensory perception of light but they can also be exploited to find new evolutionary connections between species. For instance, some ASFAs suggest that complex human functions, such as synapses between neurons, are inherited from prokaryotic organisms.

Finally, due to the scarcity of GO annotations, we find that not all axes have associated GO terms, i.e., the biological meaning of the non-annotated axes can not be discovered using the current functional annotations. Thus, we go beyond this limitation and propose to use the description of the genes that are associated with the axes to define their ASFAs. We demonstrate that the corresponding ASFAs are also biologically coherent and complement the biological information obtained from the biological annotations.

5.2 Materials and Methods

5.2.1 Biological datasets

Species-Specific PPI Networks We collect the experimentally validated PPIs of *Homo sapiens sapiens* (human) and of five frequently used model organisms *Saccharomyces cerevisiae* (budding yeast), *Schizosaccharomyces pombe* (fission yeast), *Rattus norvegicus* (rat), *Drosophila melanogaster* (fruit fly) and *Mus musculus* (mouse) from BioGRID v.4.2.191 [41] (details about PPI networks be found in Chapter 2, section 2.1). We model these species-specific PPI data as PPI networks in which nodes represent genes (or in this study, protein products), and edges connect nodes (genes) whose corresponding proteins physically bind. The network statistics of these species-specific PPI networks are described in Supplementary Table C.1.

Network Representation. We represent the species-specific PPI networks with their PPMI matrices, X (details about PPMI matrices can be found in Chapter 2, section 2.7.1). These matrices measure the associations between any two nodes in the corresponding PPI network by comparing the probability that the two nodes co-occur in a random walk to what this probability would be if the occurrences of the nodes in random walks were independent. Following Xenos *et al.,* [89], we use the Deepwalk closed formula by Quiu *et al.,* [166] with its default settings to compute the PPMI matrix (see equation 2.15). This formula can be interpreted as a diffusion process that captures high-order proximities between the nodes in the network. As demonstrated in Chapter 4 section B.2.1, as a result of the extra information encoded in the PPMI, its corresponding embedding space better captures the cell's functional organization than the ones generated by using the adjacency matrix.

Biological Annotations. We use the GO Biological Process (BP) terms to represent the biological functions in a cell [191] (this ontology is introduced in Chapter 2, section 2.5.2). We collect the experimentally validated genes to GO BP terms from NCBI's FTP (gene2go file, collected on 28 September 2021). To better capture the higher level functional organization of the cell, we not only annotate the genes with the GO BP terms that they are associated with in the gene2go file, but also with the ancestors of these terms in the GO ontology. To uncover these ancestor terms, we use GOATOOLS [175] and follow the 'is_a' and 'part_of' links between the GO terms in the ontology's directed acyclic graph (go-basic.obo file, collected on 04 November 2021 from the GO website). Thus, for a given gene, we annotate it with the ancestors of the GO BP terms that annotate it. Supplementary Table C.2 shows the total number of GO BP terms that annotate genes in each species-specific PPI network. From the same gene2go file, we also keep the information about in which species (taxons) each annotation appears after considering extension with ancestor terms (out of the 20 taxons included in the file).

5.2.2 Embedding the PPI networks

To obtain the species-specific PPI embedding spaces, we use two different network embedding algorithms: NMTF [21] and Deepwalk [19] (details about these algorithms can be found in Chapter 2, section 2.7).

NMTF. We use NMTF to decompose the PPMI matrix representation of a molecular network X as the product of three non-negative factors, $X \approx P \cdot S \cdot G^T$, where rows of the matrix $E = P \cdot S$ define the set of embedding vectors of the genes, and the columns of G defines the basis (a.k.a, axes) of the space in which the genes are embedded [200]. We use NMTF with and without applying the orthonormality constraint ("ONMTF" and "NMTF," respectively) to the basis-defining matrix ($G^TG = I$). This constraint leads to minimal co-linearities (hence, dependencies) between the vectors of the axes of the embedding space [172]. The ONMTF and NMTF decompositions are done by minimizing functions in 2.7.2. To generate initial P, S, and G matrices, we use the SVD-based strategy [201]. This strategy makes the solver deterministic and also reduces the number of iterations that are needed to achieve convergence [201]. The quality of the factorization is usually evaluated by the Relative Square Error (RSE) between the input matrix, X, and its corresponding decomposition, PSG^T, as $RSE = \dfrac{\|X - PSG^T\|_F^2}{||X||_F^2}$. The iterative solver is normally stopped when the RSE is not increasing anymore.

Deepwalk. We use Deepwalk with its default settings [19] to learn the embedding vectors of the genes. Similarly to other NLP-based network embedding algorithms, Deepwalk learns these vectors by considering the node paths traversed by random walks over the PPI network as word sentences and leveraging a skip-gram neural network for learning the embedding vectors of the nodes (which are genes in this case) [19]. It has been demonstrated that Deepwalk, like other NLP-based network embedding algorithms, is implicitly performing a matrix factorization [166] (detailed in Chapter 2, section 2.7.1). Thus, we take advantage of the fact that Deepwalk implicitly decomposes the shifted PMI to obtain the basis from Deepwalk embeddings (see Chapter 2, section 2.7.1) as follows. Following Quiu *et al.,* [166], we obtain the shifted PMI matrix by setting the shifted parameter to 0 (number of negative sampling in skip-gram to one, $b = 1$) and substituting all the negative values with zeros (see equation 2.15). Note that with these parameters, the matrix obtained from the Deepwalk closed formula corresponds to the same PPMI, X, that we use as input for NMTF. Then, we obtain the basis from Deepwalk embeddings, as follows: $X \approx E \times G^T \rightarrow G \approx X^T(E^T)^-1$, where -1 denotes the Moore-Penrose pseudoinverse. Importantly, the implicit deposition from Deepwalk has fundamental differences with those from NMTF. First, it is not constrained to be non-negative, i.e., the coordinates of the embedding vectors of the genes, E, can be either positive or negative. Second, the basis, G, can not be a constraint to be orthonormal, i.e., leading to more correlated axes. In other words, Deepwalk decomposition has more degrees of freedom than the NMTF one, which may affect the topology of the gene embedding space.

5.2.3 Annotating the axes of the gene embedding space with GO BP terms

In Chapter 4, we introduced the FMM to explore the functional organization of the gene embedding space from a functional perspective. The FMM uncovers the functional organization of the embedding space by capturing the interactions between all gene functions (in our previous study represented by GO BP terms) based on their mutual positions in the embedding space (details about our FMM-based methodology can be found in Chapter 4, section 4.2.4). While the FMM captures all the pairwise interactions between functions, it does not allow for identifying the most important functional interactions. To overcome this limitation, we propose to use the axes of the embedding space in which the genes are embedded to capture the most relevant interactions between the functional annotations that we embed in the gene embedding space. Our method takes as input: the matrix factor, G, which contains the axes of the gene embedding space, and the relation-matrix between the genes and their functional annotations, A, in which entry $A[i, j]$ is one if annotation a_i annotate gene g_j, and it is zero otherwise.

First, we generate the embedding vectors of the functional annotations in the gene embedding space. To this aim, we decompose the matrix A as the product of two matrix factors, U and G^T, $A \approx UG^T$, where rows of the matrix U (that we call u_i) are the embedding vectors of the annotations, a_i, in the gene embedding space defined by the axes G, i.e., the entry $u_i[j]$ corresponds to the coordinate of the vector u_i in respect to the axis j in G. Since matrix A and G are known, we obtain U by: $U \approx (G^T)^{-1}A$, where $(G^T)^{-1}$ is the Moore-Penrose pseudoinverse of G^T [202]. Then, we associate annotation a_i to axis j if the value of the projection of a_i on j, $u_i[j]$, is statistically significantly larger than expected by random. We assess this statistical significance by performing the following bootstrapping-based method with 100,000 iterations. In each iteration, we randomly shuffle the relation-matrix A and use it as input to obtain the random vectorial representations of the annotations. After all the iterations, the p-value of entry $u_i[j]$ is computed as $p - value = \frac{c+1}{p+1}$ where c corresponds to the number of times that the observed value of $u_i[j]$ is lower or equal to that value of the corresponding random vectorial representation. For each annotation, we correct the resulting p-value for multiple hypothesis testing by applying the False Discovery Rate (FDR [179]) method overall axes. We consider the project of annotation a_i on axis j to be statistically significant if its corrected p-value is lower than or equal to 5%. Finally, following the hard clustering procedure introduced by [17], we consider that annotation a_i is associated with axis j if $u_i[j]$ is statistically significant and is the entry with the maximum value in vector u_i.

5.2.4 Quantifying the evolutionary conservation of biological functions

The extant species are the culmination of billions of years of evolution. During this process, several cellular functions have been conserved, lost, or gained by the species (taxons) via natural selection. To quantify the evolutionary conservation of a given biological function (represented by a GO BP annotation in this study), we introduce the "conservation degree," which we define as the number of different taxons in which the annotation appears (out of the 20 taxons available in the gene2go file obtained from NCBI's FTP, detailed in section 5.2.1). Intuitively, the higher the conservation degree of a function is, the more evolutionary conserved it is (from 1 to 20).

We evaluate if the conservation degree also carries information about the specificity of the function represented by the GO BP term (if it is a high-level or a specialized cellular function) by computing the Pearson's correlation coefficient [205] between our conservation degree and two known measures of functional specificity: the number of genes that are annotated by a particular GO BP term (number of genes for short) and the level of the GO BP terms in the GO hierarchy (level for short). GO BP terms that represent generic cellular functions annotate a large number of genes and have low levels in the GO hierarchy. In contrast, GO BP terms that annotate a low number of genes and have high levels in the GO hierarchy represent more specialized cellular functions. We find that the conservation degree is positively correlated with the number of genes (Pearson correlation coefficient of 0.44 with a p-value lower than 0.05) and negatively correlated with the level (Pearson correlation coefficient of -0.27 with a p-value lower than 0.05). Thus, a high conservation degree relates to generic functions that annotate larger sets of genes, while low conservation degrees relate to more specific functions that annotate smaller sets of genes.

We also investigate if the conservation degrees of the GO BP terms relate to their embedding vector positions in the embedding space. To this aim, we embed GO BP terms into the species-specific PPI embedding spaces (detailed in sections 5.2.2 and 5.2.3) and we study the correlation between the mutual positions of their embedding vectors in the embedding space (measured by their pairwise euclidean distances) and their conservation degree. We find that the higher the conservation degree of a GO BP term, the most likely its vectorial representation is spread (higher average pairwise euclidean distance) in the embedding space (Spearman correlation coefficient of 0.72 with p-value lower than 0.05). Interestingly, we find that after a specific conservation degree (17 in human ONMTF embedding space), the average pairwise euclidean distance drastically increases (from 1.20 to 16.48 in Human orthonormal embedding spaces, see Supplementary Figure C.1). Thus, we use this observation to divide the GO BP terms into three categories: "specific" (conservation degree between 17 to 20), "generic" (conservation degree between 1 to 4), and "background" (GO terms that are neither generic nor specific) for human ONMTF embedding spaces.

5.2.5 Evaluating our axes-based methodology

Current embedding approaches rely on the organization of the genes in the embedding space to uncover the cell's functional organization from molecular networks. These approaches apply functional enrichment analysis to identify those GO BP terms that are statistically overrepresented in the cluster of genes. The GO BP terms that are statistically enriched in each cluster are then summarized in order to represent the cell's functional organization (details about the gene clustering and functional enrichment analysis can be found in Chapter 2, section 2.5). The ability of the gene-centric approach to uncover the cell's functional organization is usually quantified by the number of gene clusters enriched in GO BP terms ("enriched clusters"), the number of GO BP terms enriched across these gene clusters ("enriched GO BP terms"), and the semantical similarity of GO BP terms enriched in the same cluster.

Instead of using the organization of the embedded entities (genes and genes' functions) in the embedding space, in this Chapter, we propose using the axes of the embedding space where the entities are embedded to uncover the cell's functional organization from biological networks. Similar to the standard gene-centric approach, we propose to evaluate the ability of our axes-based method to capture the cell's functional organization by analyzing the amount of GO BP terms that are associated with them. We report both: the percentage of the total GO BP terms that are associated with the axes and the percentage of axes with at least one associated GO BP term. We also investigate whether this captured biological knowledge is coherently stratified across the axes, i.e., if GO BP terms associated with the same axis are more functionally similar than those associated with different axes. To this aim, we compute Lin's semantic pairwise semantic similarity [136] between any two GO BP terms. This measure captures the similarity in the biological concepts represented by the GO BP terms, i.e., a high semantic similarity indicates that two GO BP terms are functionally related. We term "intra-semantic similarity" the average semantic similarity of the pairs of GO BP terms that are associated with the same axis, and "inter-semantic similarity" the average semantic similarity of the pairs of GO BP terms that are associated with different axes. We report how many times the "intra-semantic similarity" and "inter-semantic similarity" are in comparison to the expected random semantic similarity and the p-value of the corresponding one-tailed Mann-Whitney U test. Alternatively, we apply the same methodology but replace Lin's semantic similarity with the shortest path distance in the ontology-directed acyclic graph (DAG) as a measure of functional similarity between the GO BP terms. The lower the shortest path distance between two GO BP terms is, the more functionally related they are.

We also assess if our axes-based methodology better uncovers the cell's functional organization from biological networks than the gene-centric approach. To this aim, we compare the percentage of enriched GO BP terms and the percentage of enriched clusters to the percentage of GO BP terms associated with the axes and the percentage of axes with at least one associated GO BP term, respectively. We also evaluate if GO BP terms associated with the same axis are on average more semantically similar than GO BP terms enriched in the same gene cluster. In addition, we assess the agreement between our axes-based methodology and the gene-centric by evaluating if GO BP terms associated with the same axis are also enriched in the same gene cluster. We measure this agreement with the adjusted Rand Index [226]. At the time this thesis is submitted, we only have preliminary results for these comparisons and we are still developing a proper method to compare our axes-based methodology to our previous FMM-based approach (see them in Chapter 6, section 6.3.3).

5.2.6 Generating the Axes-Specific Functional Annotations

To obtain annotations that globally summarize the biological functions captured by each axis of the embedding space, we propose to use the GO BP terms captured by them to generate new data-driven functional annotations, which we call *Axes-Specific Functional Annotations* (ASFAs). To this aim, we adapt the Term Frequency Inverse Document Frequency (TF-IDF) used in the NLP field (this methodology is introduced in Chapter 2, section 2.5.2.1). The TF-IDF is a numerical statistic that reports how important a word is to a document (e.g., chapters of a book) in a corpus (e.g., a textbook) [134]. We extend this statistic to our problem by considering all the GO BP terms associated with the axes of the embedding space (all their text descriptions) as the corpus. On the other hand, we consider as a document of this corpus the union of the text descriptions of the GO BP terms that are associated with an axis (ideally having as many documents as axes). Then, we compute the TF-IDF of a word in a document by applying the equation 2.10.

Since not all the words add any semantic meaning to the text (e.g., "where" or "of"), they could add noise to the TF-IDF, i.e., these so-called "stop words" are usually removed before applying the TF-IDF [227]. We take the list of stop words from the NLTK package version 3.6.7 [228]. Similarly, we filter all the "stop words" from the text definitions of the GO BP terms before computing the TF-IDF. Finally, for each axis, we build its ASFA by taking all the words with a TF-IDF higher than 0, i.e., words that are relevant to the document.

Since GO ontology is incomplete, many genes lack GO BP term annotation [227]. Because of this, many genes are left non-annotated (with no associated GO terms). Thus, some axes may not capture any embedded functional annotations from the embedding space. To overcome this issue, we propose to use the description of the genes that are associated with the axes to define their ASFAs. To this aim, we download the gene descriptions file from the Alliance of Genome Resources database v.5.2.1. Then, we associate each gene to the axis for which the projection of the gene's embedding vector has the largest value (in the spirit of the hard clustering procedure of Brunet *et al.,* [17]). Finally, having a set of gene descriptions for each axis, we apply the same TF-IDF-based approach used with the functional annotations to build their ASFAs. In particular, we consider all the genes associated with the axes of the embedding space (all their gene descriptions) as the corpus. On the other hand, we consider as a document of this corpus the union of the gene descriptions of the genes that are associated with an axis (ideally having as many documents as axes). Then, we compute the TF-IDF of a word in a document by applying the equation 2.10.

5.2.7 Analyzing the connection between the ASFAs and evolution

To gain insights into the human evolutionary history, we propose to investigate the link between the ASFAs and evolution. To this aim, we order the ASFAs based on their conservation degree (we introduced the concept of the conservation degree for a GO BP term in section 5.2.4). For an individual ASFA, we obtain its conservation degree by the union of the different taxons in which the GO BP terms associated with its corresponding axis appear. We also search for evolutionary patterns across our ASFAs by identifying those ASFAs that describe biological functions that are conserved from prokaryotic organisms, that appeared for the first time in eukaryotes, or that are unique for vertebrates. To this aim, for a given ASFA, we take all the GO BP terms associated with its corresponding axis and classify the ASFA according to the different taxons in which these GO BP terms appear. Based on these taxons, we consider this ASFA to be related to "prokaryotes" if at least one of the taxons is a prokaryote, "eukaryotes" if all the taxons are eukaryotes, and "vertebrates" if all the taxons are vertebrates.

5.3 Results and Discussion

5.3.1 The axes of the embedding spaces capture the cell's functional organization

In this section, we evaluate if the axes of the embedding space uncover the cell's functional organization from PPI networks. To this end, we generate the embedding spaces of six species (human, budding yeast, fission yeast, rat, fruit fly, and mouse) by applying ONMTF, NMTF, and Deepwalk algorithms on the corresponding species-specific PPI networks (detailed in sections 5.2.1 and 5.2.2). To analyze the impact of dimensionality on the ability of the embedding methods to reveal the cell's functional organization, for each species-specific PPI network and for each embedding method, we generate the embedding spaces with different dimensionalities (from 50 to 1000 dimensions with a step of 50). Then, we embed GO BP terms into these embedding spaces and associate them to the axes of the space (detailed in section 5.2.3). We evaluate the ability of the embedding axes to uncover the cell's functional organization by analyzing the percentage of axes having at least one associated GO BP term, the percentage of the total GO BP terms that are associated with the axes, and the functional similarity of the captured GO BP terms (detailed section 5.2.5). For this section, we focus on human PPI networks indicating whether the results hold for the rest of the species.

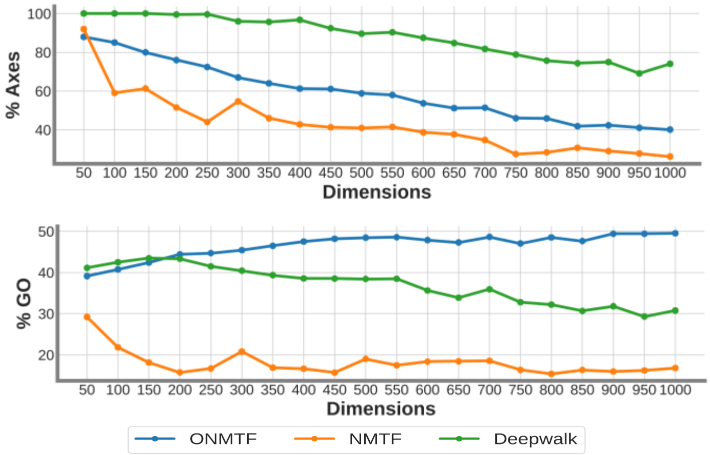


Figure 5.1. The axes of the human ONMTF, NMTF, and Deepwalk embedding spaces uncover the cell's functional organization from the human PPI network. For each PPI embedding space, we use our new axes-based method to capture the GO BP terms that we embed in the embedding space (detailed in section 5.2.3). The top panel shows the percentage of axes that captures at least one embedded GO BP term. The bottom panel shows the percentage of the total embedded GO BP terms that are captured by the axes of the space. For each panel, the horizontal axis displays the number of dimensions of the embedding space. For each panel, the color of the lines corresponds to the three tested embedding algorithms: ONMTF (blue), NMTF (orange), and Deepwalk (green).

We observe that Deepwalk embedding spaces have, on average, the largest number of axes with associated GO BP terms (88.05%) followed by ONMTF (59.25%) and NMTF (42.78%), see Figure 5.1. However, the axes of ONMTF embedding spaces capture a larger number of GO BP terms (37.12%) than the axes of Deepwalk (33.8%), and NMTF embedding spaces (11.95%), see Figure 5.1. These results suggest that Deepwalk embedding spaces capture fewer biological functions but "spreads" them more across the axes (average of 9.7 GO BP terms per axis), while ONMTF spaces capture more biological functions and group them on a smaller number of axes (average of 16 GO BP terms per axis). Furthermore, we investigate whether this captured information is coherently stratified across the axes, i.e., if the GO BP terms that are associated with the same axis are more functionally similar (higher semantic similarity and closer in the ontology DAG) than those associated with different axes (detailed in section 5.2.5). We find that ONMTF embedding spaces not only group more GO BP terms per axis, but the functions that are associated with the same axis are functionally more coherent (3.12 times higher average semantic similarity than expected by random, Mann-Whitney U test with p-value 3.39×10^{-8}) than the ones associated with the same axis in NMTF and Deepwalk embedding spaces (2.6 and 2.1 times larger than expected by random, respectively, Mann-Whitney U test with p-values 3.38×10^{-8} and 2.41×10^{-7}, respectively), see Supplementary Table C.3. Moreover, GO BP terms associated with the axes of the ONMTF spaces are on average closer in the ontology DAG (average shortest path of 4.21), than the ones captured in NMTF and Deepwalk embedding spaces (average shortest path of 4.70, and 5.35, respectively). We find similar results for the rest of the species-specific embedding spaces (see Supplementary Tables C.4 and C.5).

Altogether, these results confirm that the embedding axes capture biological knowledge from the PPI network and that this information is correctly distributed across dimensions, i.e., each axis captures a set of GO terms that are functionally related. Moreover, we demonstrate that the axes of the ONMTF embedding spaces capture more and better-stratified information than the other methods. Hence, the axes of the ONMTF embedding spaces better uncover the cell's functional organization.

5.3.2 Orthonormality and positive constraints improve the functional organization of the gene embedding space

Here we analyze if the ability of ONMTF to produce embedding spaces whose axes capture more, and better stratified functional information can be attributed to the properties of the embedding spaces produced by the ONMTF. ONMTF embedding spaces have two properties, orthonormality, and non-negativity, we assess the effect of these properties in disentangling functional knowledge from the biological networks. Since the embedding space is orthonormal, its axes should represent non-ambiguous and non-dependant directions of the space [89]. We confirm this first property by computing the average pairwise cosine similarity in-between the axes of the ONMTF, NMTF, and Deepwalk embedding spaces. It is important to note that Deepwalk embedding spaces are not constrained to be positive, which means that the cosine similarity is bounded from -1 to 1 instead of from 0 to 1. Thus, to make it comparable to the NMTF and ONMTF spaces, we report the absolute pairwise cosine similarity in-between their axes. A cosine similarity of 1 indicates that two axes are identical (i.e., redundant), and a value of 0 indicates that the axes are orthogonal (i.e., perpendicular).

We observe that the axes of the NMTF embedding spaces have, on average, the largest number of similar axes (average pairwise cosine similarity of 0.014), followed by Deepwalk (average pairwise cosine similarity of 0.10), and ONMTF (average pairwise cosine similarity of 0.76). These results suggest that the majority of the axes in the NMTF embedding space are redundant, i.e., some dimensions do not contribute to disentangling functional knowledge from the biological networks. This high redundancy, in turn, explains the low percentage of GO BP terms associated with the axes of NMTF spaces (11.95%) in comparison to ONMTF (37.12%). We also see that, although the axes of the Deepwalk spaces are not constrained to be orthonormal, their axes have a lower average pairwise cosine similarity (average of 0.10) than the ones of the NMTF. We explain this observation by the degrees of freedom of Deepwalk spaces. In other words, since Deepwalk spaces are not constrained to be positive, the chance that two random vectors are identical is low. This also makes the Deepwalk spaces sparse in comparison to the non-negative ONMTF and NMTF spaces, as can be seen in the percentage of axes with at least one associated GO BP term of the three embedding methods (88.05%, 59.25%, and 42.78% for Deepwalk, ONMTF, and NMTF, respectively). Finally, we observe the absence of non-negativity constraints in Deepwalk embedding spaces decreases its ability to capture the cell's functional organization (GO BP terms less coherently stratified than ONMTF and NMTF, results presented in the previous section). We hypothesize that this observation is connected with the fact that biological processes are often non-negative and additive [169], i.e., positive embedding spaces are more suitable to capture these complex biological mechanisms.

In conclusion, the embedding in positive and orthonormal spaces, which only NMTF-based frameworks allow for, leads to the embedding axes that best capture the cell's functional organization from the biological networks. Thus, in the next sections, we will focus on ONMTF embedding spaces.

5.3.3 Specific biological functions are disentangled by the axes of the embedding space with the increment of dimensions

Having demonstrated that the embedding axes capture GO BP terms from the functional organization of the species-specific PPI embedding spaces, we investigate if the space's dimensionality affects the specificity of the GO BP terms captured by the axes, the amount of GO BP terms captured by the axes, the number of axes with at least one associated GO BP term, and the coherence of the stratification of the GO BP terms across the axes. In particular, to analyze the impact of the dimensionality on the specificity of the GO BP terms captured by the axes, we divide them into three groups: "specific," "generic," and "background" (detailed in section 5.2.4). Then, we take as reference the lowest dimensional embedding space (50 dimensions) and compare the fold increase between the number of "specific," "generic," and "background" GO BP terms associated with its axes and with those captured by the axes of the subsequent species-specific PPI embedding spaces.

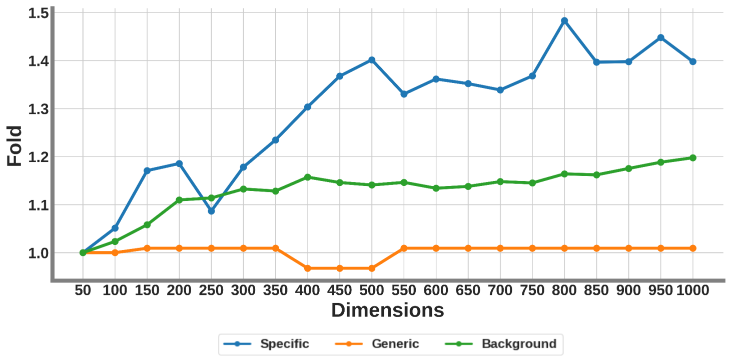


Figure 5.2. Specific biological functions are captured by the axes of the human ONMTF embedding spaces with the increment of dimensions. We take as reference the lowest dimensional embedding space (50 dimensions) and compare the fold increase between the number of "specific," "generic," and "background" GO BP terms associated with its axes and with those captured by the axes of the subsequent species-specific PPI embedding spaces. The horizontal axis displays the number of dimensions of the embedding space.

We find that most of the "generic" functions (average of 90%) are associated with the axes of human lowest dimensional embedding space (50 dimensions). Importantly, we find that increasing the dimensionality of the embedding space does allow us to capture more "generic" functions (fold increase remains close to 1, see Figure 5.2). In contrast, increasing this dimensionality allows for capturing more "background" and "specific" functions, with the specific ones being the ones that most benefit from the increase in the number of dimensions (see Figure 5.2). Moreover, we find that increasing this dimensionality also enhances the stratification of the biological information captured by axes, with more semantic similar GO BP terms associated with the same axis (see Figure 5.3). These results suggest that the embedding space needs more dimensions to disentangle “specific” biological functions encoded in the species-specific PPI networks. Nevertheless, this disentanglement has a limit since after 500 dimensions we observe three signs that indicate no significant benefit in increasing the space’s dimensionality. First, the number of axes capturing at least one GO BP term reduces to less than 50% and the total amount of GO BP terms captured flattens after 500 dimensions (see Figure 5.1). Second, the fold increase of “specific” functions is significantly reduced after 500 dimensions (see Figure 5.2). Third, the semantic similarity of GO BP terms associated with the same axis flatters after 400-500 dimensions. Thus, adding more dimensions does not improve the capture of either more biological information or more specific information from the embedding space. Interestingly, these observations are in line with the results reported in other artificial intelligence fields, such as NLP, where a low dimensionality of the word embedding fails to capture all possible word relations (“specific” relations), and after a certain number of dimensions, the embeddings can not disentangle more word relations [229]. We find similar results for the rest of the studied species-specific ONMTF embedding spaces (see Supplementary Figures C.2 and C.3).

Based on these results, we consider the optimal dimensionality of a given specie-specific PPI embedding space as the one that finds a balance between the three observations introduced above (i.e., amount of information captured, specificity of this information, and the coherence in the stratification of the information captured across the axes). Based on these criteria, we choose 500 dimensions as the optimal dimensionality for the human ONMTF embedding space (the optimal number of dimensions for the rest of species-specific ONMTF embedding spaces can be found in Supplementary Table C.6). This optimal dimensionality is coherent with the number of dimensions usually applied in NLP [230, 231]. In the following sections, we investigate the biological meaning of the axes of the optimal dimensional human ONMTF embedding space in detail.

5.3.4 The axes of the embedding space represent the fundamental mechanisms of the cell

In this section, we perform an in-depth analysis of the biological meaning of the axes of the human PPI embedding space. To this aim, we summarize the set of GO BP terms captured by each axis into ASFAs (as detailed in section 5.2.6). We assess if the ASFAs correctly summarize the set of GO BP terms captured by the axes and evaluate if they describe coherent biological functions by literature curation.

Globally, we find that the ASFAs correctly summarize the biological information captured by the axes and confirm that our ASFAs describe coherent functions of the human cell (see Table 5.1). For instance, axis 12 captures seven GO BP terms (GO:0060354, GO:2000647, GO:190233, GO:1904995, GO:1903121, GO:1903122 and GO:1902034). Individually, these GO BP terms describe the regulation of various cellular processes such as cell adhesion (GO:0060354), leukocyte adhesion (GO:1904995), stem cell proliferation (GO:2000647), hematopoietic stem

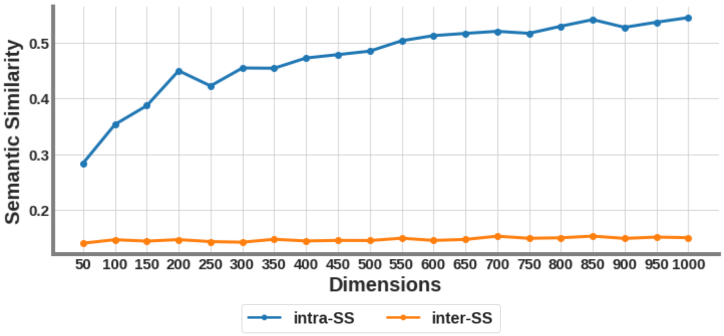


Figure 5.3. Specific biological functions are disentangled by the axes of the human ONMTF embedding spaces with the increment of dimensions. For each dimensional human ONMTF embedding space, we compute Lin's semantic pairwise semantic similarity between any two GO BP terms. The blue line shows the average semantic similarity of the pairs of GO BP terms that are associated with the same axis (intra-SS). The orange line shows the average semantic similarity of the pairs of GO BP terms that are associated with different axis (inter-SS). The horizontal axis displays the number of dimensions of the embedding space.

cell proliferation (GO:190233 and GO:1902034), and TRAIL-dependant apoptotic pathways (GO:1903121 and GO:1903122). As can be seen in Table 5.1, the resulting ASFA summarizes and combines the keywords of this set of GO BP terms. Moreover, it describes a coherent cellular function connected to the induction of apoptosis of tumor and infected cells via TNF-related apoptosis-inducing ligand (TRAIL) [232]. This apoptosis signaling pathway is usually activated by different leukocytes, such as natural killer cells and T cells [233]. TRAIL also coordinates the immune response to tumor cells and infected cells by activating the production of leukocytes by hematopoiesis and controlling the inflammatory processes [233].

Another example is axis 495, which captures five GO BP terms (GO:1900101, GO:1903891, GO:1990440, GO:0036003 and GO:1903893). These GO BP terms describe the response to endoplasmic unfolded protein (GO:1900101, GO:1903891, and GO:1903893) and the regulation of gene expression in response to cellular stress (GO:1990440 and GO:0036003). Their corresponding ASFA correctly summarizes these terms and displays a coherent biological function related to the cellular response against the accumulation of misfolded proteins in the Endoplasmic Reticulum [234] (see Table 5.1). Finally, axis 51 captures twenty-seven GO BP terms that describe multiple cellular processes such as the regulation of telomere, chromosome stability, Cajal body, protein location and stability, and RNA location to the nucleus and gene expression. Hence, the resulting ASFA is connected to the functions of the Cajal bodies that include the biogenesis and modification of different types of ribonucleoprotein, such as Cajal body-specific RNPs (scaRNPs) and telomerase [235] (see Table 5.1). The observed biological coherence of our ASFAs can be explained by the fact that GO BP terms associated with the axes are already functionally coherent (shown in section 5.3.1). We find similar results for the rest of the species (see Supplementary section C.1.1).

In conclusion, by analyzing the biological coherence of the ASFAs, we demonstrate that the axes of the embedding space capture coherent complex cellular func-

Axis	Terms	#GO	Taxons
12	endothelial, negative, regulation, apoptotic, molecule, signaling, cell, stem, activated, leukocyte, vascular, TRAIL, proliferation, adhesion, hematopoietic, production	7	7227, 7955, 9606, 10090, 10116
495	polymerase II, mediated, RNA, unfolded, response, regulation, protein, stress, reticulum, ATF6, promoter, positive, transcription, endoplasmic	5	3702, 4896, 6239, 9606, 10090, 10116, 352472, 559292
51	maintenance, activity, telomere, scaRNA, RNA, telomeric, biosynthetic, Cajal, regulation, protein, stability, nucleus, localization, process, lengthening, establishment, body, DNA, positive, via, stabilization, telomerase, chromosome, organization	27	3702, 4896, 6239, 7227, 7955, 9031, 9606, 9615, 9823, 9913, 10090, 10116, 352472, 511145, 559292
144	activity, anion, negative, aminobutyric, regulation, signaling, inhibitory, assembly, inorganic, chloride, store, pathway, acid, transmembrane, transport, synapse, operated, gamma, calcium, channel	8	4896, 10116, 9606, 9031, 511145, 10090, 9615, 7955, 3702, 352472, 7227, 559292, 6239
61	subunit, spliceosome, processing, nucleobase, RNA, aromatic, heterocycle, snRNP, complex, compound, process, capping, nucleophile, assembly, containing, reactions, spliceosomal, cellular, 3’, mRNA, adenosine, ribonucleoprotein, organization, organic, cyclic, bulged, transesterification, splicing, nucleic, metabolic	20	3702, 4896, 6239, 7227, 7955, 9031, 9606, 9615, 9823, 9913, 10090, 10116, 36329, 39947, 195103, 214684, 227321, 352472, 511145, 559292
492	spliceosome, cis, response, menadione, cellular, via, mRNA, splicing	2	4896, 7227, 9606, 511145
446	vitamin K2, biosynthetic, menaquinone, process	2	511145, 7955, 7227, 9606
120	system, light, visual, nervous, stimulus, process, sensory, perception	4	6239, 7227, 7955, 9606, 10090, 10116
64	neural, crest, cell, stem, specification, fate	3	6239, 7227, 7955, 9031, 9606, 10090
79	heart, thyroid gland, organ, anatomical development	5	10116, 9031, 10090, 9823, 7955, 7227, 6239, 9606, 4896, 214684, 352472, 559292, 227321, 39947, 3702, 352472
473	negative, regulation, activation, cell, proliferation, lymphocyte	3	9031, 9606, 9913, 10090, 10116
370	mediated, natural, killer, leukocyte, activation, cytotoxicity, immunity, lymphocyte, cell, activation	6	7955, 9606, 9615, 9823, 10090, 10116
68	cranial, development, nerve	2	7955, 9031, 9606, 10090, 10116
402	remodeling, regulation, bone, positive, resorption	3	9606, 10090, 10116
36	muscle, skeletal, regeneration, tissue	1	9606, 10090, 10116
406	potential, action, cell, muscle, cardiac	1	7955, 9606, 10090, 10116

Table 5.1. The ASFAs describe coherent functions of the human cell. For the human ONMTF embedding space, we use the GO BP terms associated with its axes to generate the ASFAs (detailed in section 5.2.6). The first column, “Axis,” lists the name of the axes from which each ASFA was obtained. The second column, “Terms,” shows the description of the ASFAs. The third column, “#GO,” displays the number of GO BP terms that are associated with the axis. The fourth column, “Taxons,” shows the Taxonomy ID of the different species for which the associated GO BP terms appear.

tions from the functional organization of the embedding space. These results open a new opportunity for the development of data-driven ontologies using the set of ASFAs to summarize the functional organization of the cell.

5.3.5 The axes of the embedding space uncover the human evolutionary history

Having demonstrated that our ASFAs represent coherent functions of the human cells, in this section, we investigate if they can be used to get insights into the evolutionary history of humans. To this aim, we divide the ASFAs according to their conservation degree into three classes: "prokaryotes," "eukaryotes," and "vertebrates" (detailed in section 5.2.7). We end up with 156 (53%), 101 (35%), and 31 (10%) ASFAs classified as "prokaryotes," "eukaryotes," and "vertebrates" in the human PPI embedding space, respectively. We analyze in detail the meaning of these groups of ASFAs in the context of evolution.

We find that "prokaryotes" ASFAs define functions that are highly conserved in evolution (average conservation degree of 13.7, see Figure 5.4). Interestingly, these functions connect complex human cellular functions to ancient prokaryote ones. For instance, the ASFA of axis 144 has a high conservation degree of 13. Among the taxons that are connected to this ASFA, we find several vertebrates, including rats (taxon id: 10116), mice (taxon id: 10116), and chicken (taxon id: 9031), but also bacteria, such as *E. coli* (taxon id: 511145). This suggests that the biological function represented by this ASFA may be originated in prokaryotes, but is conserved across evolution. Indeed, this function describes the regulation of neuronal synapses in vertebrates by the gamma-Aminobutyric acid (see Table 5.1). Interestingly, the sets of proteins comprising synapse receptors, signaling, and biosynthetic pathways necessary for this regulation arose in prokaryotes to enable prokaryotic organisms to respond and adapt to changing environments [236, 237]. Another example is the ASFA of axis 61, which is extremely conserved across evolution (conservation degree of 20). This ASFA is connected with the RNA preprocessing by the spliceosome (see Table 5.1). Although there is a longstanding debate about the origins of the spliceosome, many researchers agree that it evolved from the group II intron ancestor, which originated within bacteria billions of years ago, during eukaryogenesis [238, 239]. We also find two outliers, i.e., "prokaryotes" ASFAs that are not conserved in evolution. In particular, the ASFAs of axes 492 and 446 show the lowest conservation degree among all prokaryotic ones (average conservation degree of 4 in both cases, see Figure 5.4). Interestingly, both are connected with vitamin K (see Table 5.1). In particular, the ASFA of axis 492 describes the cellular response to vitamin K3. This vitamin is a synthetic form of vitamin K that is often used as a supplement in animals [240]. The response to K3 has been analyzed in some model organisms to verify the safety of its intake, i.e., the conservation degree may not reflect evolutionary conservation in this case. On the other hand, the ASFA of axis 446 describes the synthesis of vitamin K2. In animals, including humans, vitamin K2 is synthesized by bacteria in the gut [241]. Hence, we hypothesize that this ASFA is not describing human functions but the functions of the gut microbiome (e.g., *E. coli* with taxon id: 511145), which explains why this function has not been conserved across evolution and yet is found in humans.

On the other hand, "eukaryotes" ASFAs are newer in evolutionary history since they have an average conservation degree lower than the "prokaryotes" ones (7.3 and 13.7, respectively). We find that these ASFAs reveal evolutionary connections between humans and other eukaryotes. For instance, the ASFA of axis 120 describes a function related to the visual sense (see Table 5.1). Among the taxons that are

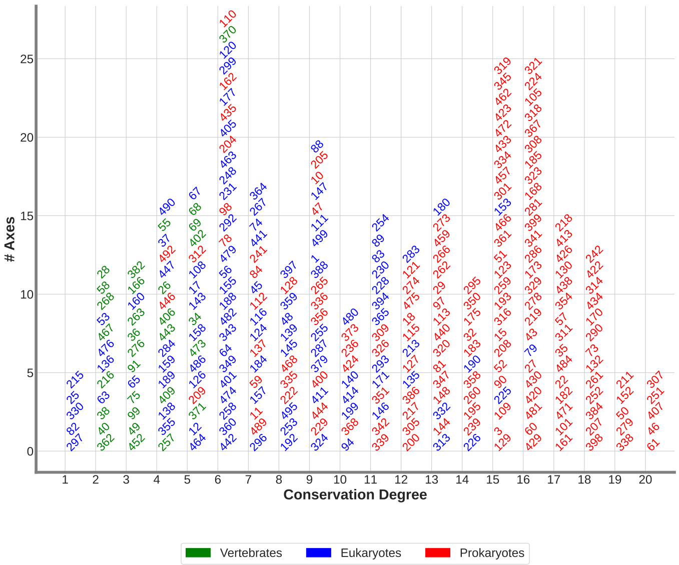


Figure 5.4. The human ASFAs give insights into the evolutionary story of humans. We use the conservation degree of the ASFAs to divide them into three groups: "prokaryotes," "eukaryotes," and "vertebrates" (detailed in section 5.2.7). Then, we order the ASFAs according to their conservation degree. The horizontal displays the conservation degree of the ASFAs. The vertical axis shows the number of ASFAs with a certain conservation degree. Each ASFA is represented in the plot by the number of the axis from which it was obtained.

connected to this ASFA, we find mammals, such as mice (taxon id: 10090) and rats (taxon id: 10116), but also insects, such as the fruit fly (taxon id: 7227). Despite the divergence in the light receptors between these species, this axis further confirms these receptors evolved from a common photoreceptor eukaryotic ancestor [242]. Similarly, the ASFA of axis 64 shed light on the evolutionary divergence in neurogenesis. In particular, this ASFA is connected to the embryonic stem cell differentiation into neural crest (see Table 5.1). Unexpectedly, although this process is considered a functional innovation of vertebrates, we find that the ASFA is connected to two invertebrates, the fruit fly (taxon id: 7227) and *C. elegans* (taxon id: 6239). To understand this observation, we analyze the three GO BP terms associated with axis 64 (GO:0001708, GO:0014036 and GO:0048866). We find that these annotations are connected to stem cell fate differentiation (GO:0001708 and GO:0048866) and neural crest stem cell differentiation (GO:0014036). From these GO terms, the two that appear in fruit fly and *C. elegans* are GO:0001708 and GO:0048866. This supports the hypothesis that regulatory programs involved in neural crest formation evolved from programs already present in the common vertebrate-invertebrate ancestor [243]. Indeed, recently, a group of cells in invertebrates was identified with the characteristics of the neural crest ones [243]. We also focus on the "eukaryote" ASFA that shows the highest conservation degree. With a conservation degree of 16, this ASFA corresponds to axis 79 and describes the molecular mechanisms involved in the development of the human heart and thyroid gland (see Table 5.1). Among the species that are connected to this ASFA, we find a variety of animals that possess these organs, such as rats (taxon id: 10116), chickens (taxon id: 9031), and mice (taxon id: 10090), but also eukaryotes that lack these structures, including budding yeast (taxon id: 559292), fission yeast (taxon id: 4896), and rice (taxon id:

39947). This suggests that the molecular mechanisms involved in the development of these anatomical structures arose early in eukaryotic evolution. Indeed, it has been hypothesized that molecular pathways involved in human organogenesis, such as the Hedgehog proteins, appeared early in the evolution of multicellular organisms through the redeployment of components found in unicellular organisms [244]. The high conservation of such pathways explains the presence of Hedgehog homologs in a variety of evolutionarily distant eukaryotes, including fungi and plants [245, 246].

Finally, the "vertebrate" ASFAs are on average the newest in the evolutionary history of humans (average conservation degree of 3.4, see Figure 5.4). In general, they describe specific traits that are unique to vertebrates. Among them, we find ten ASFAs that describe cellular functions related to the adaptive immune system, which is a system restricted to vertebrates [247, 248]. For instance, lymphocyte proliferation and the activation of natural killer lymphocytes (see axes 473 and 370 in Table 5.1). Moreover, we find eight ASFAs that define functions connected to the development of tissues that are unique to vertebrates, such as cranial development, bone remodeling, skeletal muscle, and cardiac muscle (see axes 68, 402, 36, and 406 in Table 5.1, respectively). Finally, the rest of the "vertebrate" ASFAs are connected to different regulatory processes of the cell and to metabolic processes.

In conclusion, we demonstrate that each axis of the embedding space represents a well-defined function of the human cell. Moreover, by analyzing our new ASFAs, we find evolutionary connections between different species. We find similar results between the rest of the studied species (see Supplementary sections C.1.1 and C.1.2).

5.3.6 Non-Annotated Axes also capture the functional mechanisms of the cell

Finally, in this section, we investigate the biological meaning of those axes without any associated GO BP term (a.k.a. empty axes). To this aim, we recall that genes that form densely connected regions of a PPI network tend to share biological functions [249]. Hence, we investigate if the genes that are associated with the empty axes tend to form such densely connected neighborhoods in the human PPI network. We do this by associating genes to the 206 (41.2%) empty axes of the ONMTF human embedding space. We associate each gene to the axis for which the projection of the gene's embedding vector has the largest value (detailed in 5.2.6). Then, we evaluate the connectivity in the original human PPI network by computing the clustering coefficient between genes associated with the same empty axis.

We see that the average clustering coefficient of those genes associated with the same non-empty axis (axes with associated GO BP terms) is statistically significantly higher than those genes associated with the same empty axis (Mann-Whitney U test *p-value* of 1.76×10^{-63}). However, we find that the average clustering coefficient of those genes associated with the same empty axis is statistically significantly higher than expected by random (Mann-Whitney U test *p-value* of 6.46×10^{-28}), i.e., they form more densely connected sub-networks than randomly chosen genes, which suggests that they are indeed functionally related. Hence, we explain the absence of associated GO terms on these empty axes by the lack of biological functional information (only 48.6% of the human genes in the PPI network are annotated with GO BP terms). In other words, the empty axes capture parts of the human PPI network that have not been yet annotated. We find similar results for the rest of the studied species (see Supplementary Table C.8).

Axis	Terms	#Genes	Empty
9	neurotransmission, glycinergic, gonadotropin, unsaturated, choline, activating, glycosylation, glycine, adenylate, cyclase	19	Yes
76	chylomicron, brood, thymocyte, folding, transcription, microtubule, polymerase, leukocyte, helper, thymus	27	Yes
68	transcription, somitogenesis, polymerase, developmental, skeletal, commitment, midbrain, development, binding, dopaminergic	78	No
370	natural, killer, immunoglobulin, zinc, biosynthesis, transamidation, glutaminyl-tRNAGln, cytotoxicity, eye, adhesion	20	No

Table 5.2. The empty axes of the human ONMTF embedding space capture human cellular functions. For the human ONMTF embedding space, we use the genes associated with its empty axes (axes without associated GO BP terms) and non-empty axes to generate the ASFAs (detailed in section 5.2.6). The first column, "Axis," lists the name of the axes from which each ASFA was obtained. The second column, "Terms," shows the description of the ASFAs. The third column, "#Genes," displays the number of genes that are associated with the axis. The fourth column, "Empty," indicates if the axis is empty ("Yes") or not ("No").

To find the biological meaning of empty axes, we propose to generate their ASFAs from the text description of their associated genes rather than from the text description of their associated GO BP terms (section 5.2.6). Using this approach, we obtain the ASFAs for 97.8% of the axes. We find that the interpretation of these ASFAs is less intuitive (average of 55.47 words) than the ones built using GO BP terms (average of 17.27 words) but are equally coherent. For instance, the ASFA of the empty-axis 461 is connected with the regulation of neural activity (see Table 5.2). Indeed, among the words that define this ASFA, we find glycine (an inhibitory neurotransmitter [250]), choline (regulator of neurological development [251]), and adenylate cyclases (regulator of the energy balance in different parts of the brain [250]). Another example is the ASFA of the empty-axis 76, which is connected to the functions of the thymus (see Table 5.2). This ASFA supports the observation that lipid metabolism ("chylomicron") affects lymphocyte differentiation and survival in the thymus [252].

Finally, we investigate if the ASFAs generated using genes' descriptions (a.k.a, genes' perspective) agree with those generated using functional annotations (GO terms' perspective). Interestingly, we find that the gene perspective ASFAs are not only in agreement with the GO terms perspective ones but also complement them. For instance, from the GO terms perspective, the ASFA of axis 68 is connected to cranial development (see Table 5.1). In this case, the genes' perspective not only agrees with it but also indicates that the ASFA is linked to the neural tube development (see Table 5.2). Similarly, the genes' perspective ASFA of axis 370 complements its GO terms' perspective ASFA. From the GO terms' perspective, this ASFA is connected to the activation of natural killer lymphocytes (see

Tables 5.1). The gene' perspective hallmarks the importance of the "glutaminyl-tRNAGln" and amidotransferase for the correct functioning of their mitochondria, which is connected to the activation of lymphocytes [253] (see Table 5.2).

In conclusion, we demonstrate that all the axes of the embedding space have a coherent biological meaning. For those axes that do not have any GO BP term associated, we propose an approach method that finds the meaning of empty axes. We demonstrate that the ASFAs generated by using it agree with and complement the ones obtained by using the GO BP terms.

5.4 Conclusions

By introducing our new axes-based method, we shift the exploration of the gene embedding spaces' organization from the genes' embedding vectors to the axes of the embedding space. For the first time, we do not discard the axes of the gene embedding space; instead, we demonstrate that they can be used to decipher biological information from the gene embedding space. Moreover, we show that each axis represents a non-redundant cellular function (a.k.a. ASFAs) and their combination offers a summarized functional fingerprint of the cell. This fingerprint can go from a generic overview of the cell to a most specific one depending on the number of dimensions used for generating the gene embedding space. Furthermore, we demonstrate that these ASFAs can be exploited to get insights into the evolutionary history of different species, including humans, i.e., it opens a new way to decipher the functional connections between different species. However, one of the limitations of our approach is the lack of biological information about the genes (only 48.6% of the human genes in the PPI network are annotated with GO BP terms), which results in axes without GO BP associated. We also overcome this issue by associating genes to the embedding axes and using their descriptions to build the corresponding ASFAs. We demonstrate that these ASFAs are also biologically coherent and complement the ASFAs generated using GO BP terms. Finally, our methodology could be easily applied to other bioinformatics tasks, such as the development of data-driven ontology (using the ASFAs as functional annotations and connecting them based on their similarity) or as the bases for network drawing algorithms (using the axes to summarize the functional organization of molecular networks). Finally, our new methodology is generic and can be applied to any discipline that analyzes the organization of networks by using network embeddings, e.g., social, or economic networks, paving the road to new algorithms for mining the data by utilizing the axes of the embedding space.

Chapter 6

Conclusion

In this chapter, we provide a brief summary of our results and contribution presented in this dissertation. We continue by listing the conclusions of the thesis. We finish the dissertation by presenting some future directions to which our new graphlet-based and embedding-based approaches can be applied.

6.1 Summary of thesis achievements

Cells are the basic building blocks of all living organisms. Understanding the complex intracellular processes is crucial not only to identify the fundamental mechanisms of life but also to elucidate the molecular mechanisms of a broad range of diseases. The increasing availability of “omic” data has yielded an unprecedented opportunity to understand the functioning of the cell. This data is often represented as networks. Networks are a valuable source of biological information, but they need to be untangled by new algorithms to reveal the information hidden in their wiring patterns [1]. The state-of-the-art approaches to deciphering these complex data are based on graphlets and network embeddings. In this thesis, we focus on the development of novel algorithms to overcome the limitations of the current graphlet and network embedding methodologies in the field of biology.

In Chapter 3, we propose the use of probabilistic networks to represent the uncertainty about molecular interactions. To extract the biological information hidden in the wiring patterns of probabilistic networks, we generalize the state-of-the-art graphlet-based methods to capture the local topology of the nodes in a network (i.e., GDV and GDV distance) and to capture the topology of an entire network (GCM and GCD) to probabilistic networks. By applying probabilistic graphlets to the probabilistic synthetic networks, we demonstrate that probabilistic graphlet outperforms their unweighted counterparts in capturing the overall topological similarity between synthetic networks. Moreover, we model different molecular interactions as probabilistic networks and show that probabilistic graphlets robustly manage low signal topology information without sacrificing their ability to recover relevant biological information from molecular networks. In contrast, the original unweighted graphlets applied in thresholded networks are highly sensitive to both, the noise and the chosen threshold. Thus, probabilistic graphlets allow for the use of all available data avoiding the use of thresholds that could lead to the loss of crucial information. Finally, we compare the biological information uncovered by probabilistic graphlets to the information uncovered by original graphlets. Interestingly, while original graphlets capture cellular functions that are usually active in all the cells (e.g., carbohydrate metabolic process or oxidative metabolic process), probabilistic graphlets uncover condition-specific cellular functions (e.g., stress response mechanisms or cellular differentiation). We hypothesize that this difference is connected to the probabilistic nature of condition-specific cellular functions that could benefit from the use of probabilistic models. Hence, probabilistic can be used to complement the biological information uncovered by their original unweighted counterparts. However, computing probabilistic graphlets in large networks, such as human molecular networks, is computationally challenging. We leave this issue for future research.

In Chapter 4, we introduce a new, function-centric perspective and approach to explore the functional organization of gene embedding spaces from a functional perspective. Unlike the current gene-centric perspective that focuses on the organization of the vectorial representations of the genes in the embedding space, our new perspective exploits the organization of the genes' functions (represented by functional annotations) in the space to uncover biological information from molecular networks. We introduce the FMM that captures the organization of the annotations' embedding vectors in the gene embedding space by their mutual positions. We develop FMM-based approaches to address fundamental tasks in the network embedding field, e.g., measure the similarity between the functional organization of gene embedding spaces, identify the optimal dimensionality of a gene embedding space, and capture the functional changes between two gene embedding spaces. In this Chapter, we apply our FMM-based methodology to investigate the functional changes produced by the most prevalent cancers in humans (breast, prostate, lung, and colorectal cancer). To this aim, we generate cancer and control (healthy) gene embedding spaces by applying the NMTF algorithm to the corresponding tissue-specific PPI networks. First, we use our FMM to define the optimal dimensionality of these molecular interaction networks embedding spaces. For this optimal dimensionality, we demonstrate that both embedding spaces, cancer, and control, for all four cancers, are functionally organized, i.e., functionally similar annotations are embedded close in the space, and we investigate if this organization changes between them. We show that cancer alters the positions in the embedding space of cancer-related functions, while it keeps the positions of the non-cancer-related ones. We exploit this spacial "movement" to predict novel cancer-related functions, such as alternative translational mechanisms, or the response to unfolded protein accumulation, and we validate them by systematic literature search in the PubMed database. Furthermore, we demonstrate that our FMM-based methodology is not restricted only to functionally-based analyses of cancer, but it can be used to mine for new genomic knowledge from the embedding space. We use our FMM-based methodology to predict cancer-related genes. We show that most of the cancer-related predicted genes are not differentially expressed between cancer and control, i.e., the FMM-based methodology identifies genes whose transcriptional patterns have not changed and this is complementary to the traditional differential expression analysis. Among the most promising cancer-related genes predictions, we identify four genes (PRDM11, C9orf72, MINDY3, and H4C6) for which we found some literature indicating their involvement in cancer, but whose role in cancer has yet to be experimentally validated. In conclusion, our new function-centric approach can complement the knowledge obtained by current gene-centric approaches from omic data by providing a different perspective and additional insights.

In Chapter 5, we propose to change the analysis of the network embeddings from the embedded entities (genes for the gene-centric approach and functional annotations for our functional-centric perspective) to the space itself. We introduce a new approach that uses the axes of the embedding spaces, in which the genes and functions are embedded, to capture the cell's fundamental mechanisms from molecular networks. We apply our axes-based approach to different species-specific PPI embedding spaces that we generate by NMTF and Deepwalk algorithms. To untangle the biological information hidden in the resulting gene embedding spaces, we embed GO terms and genes in the spaces and associate them with the embedding axes. For the first time, we demonstrate that the axes of the embedding space disentangle biological information from the space, with semantically similar GO BP terms associated with the same axis, i.e., each axis captures a coherent cellular function. Moreover, we demonstrate that the embedding in positive and orthonormal spaces, which only NMTF-based frameworks allow for, leads to the embedding axes that best capture the cell's functional organization from the biological networks. We also investigate the impact of dimensionality on the ability of the axes to reveal the cell's functional organization. We demonstrate that, with the increment of dimensions, the axes disentangle specific cellular functions from the molecular networks. However, we find that after reaching a certain number of dimensions, the disentanglement of specific functions stops. We use this observation to define the optimal dimensionality of the embedding spaces and explore the biological meaning of the axes in detail. To this aim, we apply an NLP-based approach to summarize all functional annotations associated with a given axis into a higher-level functional annotation that we term ASFAs. We show that each ASFA represents a coherent cellular function, and we confirm their coherence by literature curation. We demonstrate that ASFAs not only define coherent biological processes, such as the sensory perception of light, but they can also be exploited to find new evolutionary functional connections between the species. Finally, due to the scarcity of GO annotations, we find that not all axes have associated GO terms, i.e., the biological meaning of the non-annotated axes can not be discovered using the current functional annotations. Thus, we propose associating genes to the axes and using their descriptions to define their ASFAs. We demonstrate that the corresponding ASFAs are also biologically coherent and complement the biological information obtained from the biological annotations.

6.2 Conclusions

In this section, we present the general conclusions of the Thesis:

- **Methodological:**

    1. We generalize graphlets to probabilistic networks by introducing probabilistic graphlets.
    2. We introduce a new function-centric methodology to explore network embeddings from a function perspective.
    3. We introduce a new axes-based approach that changes the exploration of network embeddings from the embedded entities (genes and genes' functions) to the embedding space itself.

- **Applications:**

  1. Probabilistic graphlets do not require the use of thresholds and prevent the loss of relevant information.
  2. Probabilistic graphlets manage low signal topology information uncovering more biological information from the network than their unweighted counterparts.
  3. Probabilistic graphlets extract condition-specific processes, which in turn benefit from the use of probabilistic models.
  4. Our function-centric methodology offers a functional map of the embedding space's topology by capturing the functional organization of the genes' functions in the embedding space.
  5. Our function-centric methodology can be applied to address fundamental problems in network embeddings (e.g., to find the optimal dimensionality of the embedding space or to assess the similarity in the topology of different embedding spaces).
  6. Applied to cancer research, our function-centric perspective can be used to predict new cancer-related genes and functions that cannot be detected by using the current gene-centric approaches.
  7. We demonstrate that the embedding axes decipher coherent biological information from the gene embedding space.
  8. Our axes-based methodology captures new interactions between pairs of GO BP terms that are not described in the gene ontology but are still biologically coherent.
  9. We use these newly captured interactions to define new data-driven functional annotations (ASFAs).
  10. We applied our ASFAs to get insights into the biological history of humans.

6.3 Future work

In this section, we propose extensions of our approaches to various biological applications as well as future methodological directions that are relevant to the work presented in this Thesis.

6.3.1 Identifying Pan-cancer functions with our FMM

In this thesis, we apply our FMM-based methodology in the context of cancer research to identify the functional changes produced by the most prevalent cancers in humans (breast, prostate, lung, and colorectal cancer). We demonstrate that our FMM-based methodology can be used to identify cellular functions that are relevant to each cancer type. However, one open question is if the FMM can uncover common cellular functions that are important for all cancer types (a.k.a., pan-cancer functions). We investigate this potential application by analyzing the intersection between functions that are *shifted* (predicted to be cancer-related) in breast, prostate, lung, and colorectal cancer (de definition of *shifted* function can be found in Chapter 4, section 4.2.7).

We find a statistically significant intersection of eight annotations between the *shifted* functions in each cancer type (permutation test with p-value < 0.05). We see that these annotations are connected with cellular response to a chemokine (GO:1990869 and GO:0008543), histone phosphorylation (GO:0016572), positive regulation of the RNA export from the nucleus (GO:0046833), response to radiation (GO:0009314 and GO:0006970), and stress-activated MAPK cascade (GO:0051403 and GO:0007254). We find that these functions are normally deregulated in all types of cancer. For instance, the MAPK signaling cascades are known to be involved in the progression of various human cancers [254]. On the other hand, the cellular response to radiation involves several cellular processes, such as the arrest of cell cycle progression, repair of DNA lesions, or apoptosis, that are known to be linked to cancer [255]. Regarding the histone phosphorylation and the RNA export from the nucleus, these processes could be related to the epigenetic alterations, and the dysregulation of nuclear trafficking observed in cancer [254, 256]. Finally, the response to chemokines has been identified to play an important role in the tumor microenvironment [257].

These preliminary results suggest that there exist common functions between cancer types that could be identified with our FMM-based methodology. This hypothesis could be validated by extending this analysis to the 20 cancer types for which expression data is available in the Human Protein Atlas (HPA) [192]. In addition, as the FMM allows for the identification of cancer genes based on the spatial movement of the functions (see Chapter 4, section 4.3.3), these pan-cancer functions could be used to potentially identify those genes that are important for all cancer types.

6.3.2 Measuring the evolutionary closeness between species using the FMM

In this thesis, we apply our FMM-based methodology in the context of cancer. However, this methodology could be easily extended to other biological research areas, such as evolutionary biology or patient stratification. To demonstrate it, we generate the species-specific PPI embedding spaces by applying NMTF on the PPI network of *Homo sapiens sapiens* (denoted by human), *Saccharomyces cerevisiae* (denoted by budding yeast), *Schizosaccharomyces pombe* (denoted by fission yeast), *Drosophila melanogaster* (denoted by fly), and *Mus musculus* (denoted by mouse). Then, we use our FMM-based method to embed GO BP terms into these embedding spaces and to capture their distances over the species-specific embedding spaces (see Chapter 4, section 4.2.4 for details about the FMM methodology). We measure the similarity between the FMMs of these five species by computing their pairwise RSE (details about this methodology can be found in Chapter 4, section 4.2.5).

We find that the FMMs of evolutionarily related species are more similar (lower RSE between their FMMs) than the FMMs of evolutionarily distant species. For instance, the RSE between human and mouse FMMs is 0.15, while it is 0.20 between human and budding yeast (see Tables 6.1 and 6.2). Hence, this preliminary results demonstrate that our FMM-based methodology can be used to identify the evolutionary closeness between the species but more work is needed on optimizing

	Human	Budding yeast	Fission yeast	Fruit fly	Mouse
Human	0.000	0.204	0.228	0.182	0.159
Budding yeast	0.204	0.000	0.157	0.178	0.217
Fission yeast	0.228	0.157	0.000	0.195	0.242
Fruit fly	0.182	0.178	0.195	0.000	0.180
Mouse	0.159	0.217	0.242	0.180	0.000

Table 6.1. Pairwise RSE between the species-specific FMMs. For the five species: *Homo sapiens sapiens* (denoted by "Human"), *Saccharomyces cerevisiae* (denoted by "Budding yeast"), *Schizosaccharomyces pombe* (denoted by "Fission yeast"), *Drosophila melanogaster* (denoted by "Fruit fly") and *Mus musculus* (denoted by "Mouse"). The table specifies the relative error between their FMMs.

the FMM to enhance the applicability of our approach for evolutionary studies.

	Human	Budding yeast	Fission yeast	Fruit fly	Mouse
Human	0	529	1,017	736	89
Budding yeast	529	0	529	1,017	1,017
Fission yeast	1,017	529	0	1,017	1,017
Fruit fly	736	1,017	1,017	0	736
Mouse	89	1,017	11,017	736	0

Table 6.2. Common ancestor time, Million Yeats Ago (MYA) [258]. For the five species: *Homo sapiens sapiens* (denoted by "Human"), *Saccharomyces cerevisiae* (denoted by "Budding yeast"), *Schizosaccharomyces pombe* (denoted by "Fission yeast"), *Drosophila melanogaster* (denoted by "Fruit fly") and *Mus musculus* (denoted by "Mouse"). The table shows the million years from the common ancestor between the species.

6.3.3 Uncovering complementary information from network embeddings using different perspectives

In this thesis, we describe three perspectives to uncover biological information from network embeddings. The classic gene-centric perspective focuses on the organization of the genes in the embedding space to uncover this information. Our new function-centric perspective (introduced in Chapter 4), uses the organization of the genes' functions (represented by functional annotations) in the embedding space to capture this biological information. Finally, our axes-centric perspective (introduced in Chapter 5) directly uses the embedding space itself to uncover this information. As of the submission of this thesis, we are still working on our axes-based approach, i.e., there are several open questions that need to be solved.

One question is the benefit of using the axes of the embedding space instead of the entities embedded in it (genes and functional annotations) to uncover biological information from network embeddings. One direction to solve this question is to compare the ability of each perspective (gene-, function- and axes-based approaches) to capture the cell's functional organization from network embeddings. We assess it by analyzing the functional coherence in the stratification (grouping) of the GO BP terms captured by each method. First, we use the three methodologies to obtain the clusters of GO BP terms. For the gene-centric approach, GO BP terms are clustered together if they are statistically significantly overrepresented in gene clusters. Similarly, for our axes-centric method, GO BP terms are clustered together if they are statistically significantly associated with the same axis. Finally, for our function-centric approach, GO BP terms are clustered together if their embedding vectors are in proximity in the embedding space. Then, we evaluate the functional coherence in the stratification of the GO BP terms as follows. We compute the pairwise Lin's semantic similarity (SS) [136] between all the GO BP terms. A SS between a pair of GO BP terms is one if they represent an identical biological process and zero if they represent a totally unrelated biological process. We report the mean pairwise SS between GO BP terms that cluster together (Intra SS) and the mean pairwise SS between GO BP terms that do not cluster together (Inter SS). Intuitively, the higher the mean Intra SS and the lower the mean Inter SS the most functional coherence in the stratification of the GO BP terms, i.e., better captures the cell's functional organization.

To conduct the preliminary comparison between methods, we generate a gene embedding space by applying NMTF to the human PPI network. We choose to generate this embedding space with 300 dimensions, as it was found to be optimal based on our FMM-based approach and is close to being optimal based on our axes-based method (see Chapter 4, section 4.2.5 and Chapter 5, section 5.3.3). Then, we follow the criteria detailed above to cluster the GO BP terms based on the gene-, functional-, and axes-centric approaches. Then, we report the fold change between the mean Intra SS and the mean Inter SS.

We find that the axes-based approach stratifies the GO BP terms the best (a fold of 2.8), followed by the FMM-based approach (a fold of 1.6) and the gene-centric approach (a fold of 1.35). Hence, this preliminary result suggests that using the embedding space itself, rather than the spatial organization of the entities (genes and functions) embedded in it, leads to better stratification of biological information uncovered, i.e., the axes of the embedding space are the most effective at capturing the cell's functional organization from molecular networks. We speculate this observation may be connected to the importance of the axes in the embedding space's topology. As introduced in Chapter 2, section 2.7, the embedding spaces are spanned by their axes. Hence, the embedding axes may be essential for the organization of the embedded entities (genes and functions) in the space. To have a proper representation of the cell's functional organization in the embedding space, the axes should capture the most essential biological information hidden in the wiring patterns of molecular networks. While this observation is promising, we need to further validate it.

In conclusion, there are still some open questions that need to be assessed. Currently, we are still working on finding the fairest methods to solve them.

6.3.4 Generalizing probabilistic graphlets to network embeddings

Finally, a future direction is to generalize our probabilistic graphlets to network embeddings. Currently, one of the students in the group is following this research line. In particular, Xenos *et al.,* [89] extended the original graphlets to network embeddings by introducing the GDV PPMI matrix. In their work, Xenos *et al.,* demonstrated that the gene embeddings generated by the NMTF-based decomposition of the GDV PPMI matrix representation of the human PPI network lead to the highest prediction accuracy of cancer genes (at least 89%) [89]. The GDV PPMI matrix is generated by applying the Deepwalk closed formula (introduced in Chapter 2 section 2.7.1) on the GDV matrix. Thus, a simple approach for extending probabilistic graphlets to network embedding would be to directly apply the Deepwalk closed formula on our pGDV (introduced in Chapter 3, section 3.2.7). On the other hand, Xenos *et al.,* are currently working on improving their GDV PPMI matrix as follows. Recall that cells in a PPMI matrix quantify how frequently two nodes, i and j, of the network co-occur in a random walk compared to what would be expected if the co-occurrences of the nodes were independent (see details in Chapter 2, section 2.7.1). Instead of using random walks, the idea behind their new graphlet-based PPMI is to use the graphlet-based (i.e., graphlet adjacency) representation of the network and then quantify how frequently the nodes appear as part of a given graphlet (e.g., triangle). Similarly, we could generalize our probabilistic graphlets to network embeddings by quantifying how frequently the nodes appear as part of a given probabilistic graphlet.

In conclusion, the work of Xenos *et al.,* opens a new research line for extending probabilistic graphlets to network embeddings.

Bibliography

[1] Noël Malod-Dognin et al. “Towards a data-integrated cell”. In: *Nature communications* 10.1 (2019), pp. 1–13.

[2] Tijana Milenković and Nataša Pržulj. “Uncovering biological network function via graphlet degree signatures”. In: *Cancer informatics* 6 (2008), CIN–S680.

[3] Natasa Pržulj, Derek G Corneil, and Igor Jurisica. “Modeling interactome: scale-free or geometric?” In: *Bioinformatics* 20.18 (2004), pp. 3508–3515.

[4] Ömer Nebil Yaveroğlu et al. “Revealing the hidden language of complex networks”. In: *Scientific reports* 4.1 (2014), pp. 1–9.

[5] Noël Malod-Dognin and Nataša Pržulj. “L-GRAAL: Lagrangian graphlet-based network aligner”. In: *Bioinformatics* 31.13 (2015), pp. 2182–2189.

[6] Darren Davis et al. “Topology-function conservation in protein–protein interaction networks”. In: *Bioinformatics* 31.10 (2015), pp. 1632–1639.

[7] Anida Sarajlić et al. “Graphlet-based characterization of directed networks”. In: *Scientific reports* 6.1 (2016), pp. 1–14.

[8] Thomas Gaudelet, Noel Malod-Dognin, and Nataša Pržulj. “Higher-order molecular organization as a source of biological function”. In: *Bioinformatics* 34.17 (2018), pp. i944–i953.

[9] Noël Malod-Dognin and Nataša Pržulj. “Functional geometry of protein interactomes”. In: *Bioinformatics* 35.19 (2019), pp. 3727–3734.

[10] Bin Zhang and Steve Horvath. “A general framework for weighted gene co-expression network analysis”. In: *Statistical applications in genetics and molecular biology* 4.1 (2005).

[11] Steve Horvath. *Weighted network analysis: applications in genomics and systems biology*. Springer Science & Business Media, 2011.

[12] Jose Lugo-Martinez et al. “Classification in biological networks with hypergraphlet kernels”. In: *Bioinformatics* 37.7 (2021), pp. 1000–1007.

[13] Daokun Zhang et al. “Network representation learning: A survey”. In: *IEEE transactions on Big Data* 6.1 (2018), pp. 3–28.

[14] Walter Nelson et al. “To embed or not: network embedding as a paradigm in computational biology”. In: *Frontiers in Genetics* 10 (2019), p. 381.

[15] Sam FL Windels, Noël Malod-Dognin, and Nataša Pržulj. “Identifying cellular cancer mechanisms through pathway-driven data integration”. In: *Bioinformatics* 38.18 (2022), pp. 4344–4351.

[16] Gongxu Luo et al. “Graph entropy guided node embedding dimension selection for graph neural networks”. In: *arXiv preprint arXiv:2105.03178* (2021).

[17] Jean-Philippe Brunet et al. “Metagenes and molecular pattern discovery using matrix factorization”. In: *Proceedings of the National Academy of Sciences* 101.12 (2004), pp. 4164–4169.

[18] Trupti M Kodinariya, Prashant R Makwana, et al. “Review on determining number of Cluster in K-Means Clustering”. In: *International Journal* 1.6 (2013), pp. 90–95.

[19] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. “DeepWalk: Online Learning of Social Representations”. In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’14. New York, New York, USA: ACM, 2014, pp. 701–710. ISBN: 978-1-4503-2956-9. DOI: 10.1145/2623330.2623732. URL: http://doi.acm.org/10.1145/2623330.2623732.

[20] Aditya Grover and Jure Leskovec. “node2vec: Scalable feature learning for networks”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, pp. 855–864.

[21] Chris Ding et al. “Orthogonal nonnegative matrix t-factorizations for clustering”. In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2006, pp. 126–135.

[22] Guanghui Li et al. “Predicting MicroRNA-disease associations using network topological similarity based on DeepWalk”. In: *Ieee Access* 5 (2017), pp. 24032–24039.

[23] Nansu Zong et al. “Deep mining heterogeneous networks of biomedical linked data to predict novel drug–target associations”. In: *Bioinformatics* 33.15 (2017), pp. 2337–2344.

[24] Jure Leskovec and Andrej Krevl. *SNAP Datasets: Stanford large network dataset collection*. 2014.

[25] Amit Surana et al. “Supply-chain networks: a complex adaptive systems perspective”. In: *International Journal of Production Research* 43.20 (2005), pp. 4235–4265.

[26] Nataša Pržulj. *Analyzing Network Data in Biology and Medicine: An Interdisciplinary Textbook for Biological, Medical and Computational Scientists*. Cambridge University Press, 2019.

[27] Bobby-Joe Breitkreutz, Chris Stark, and Mike Tyers. “The GRID: the general repository for interaction datasets”. In: *Genome biology* 4.3 (2003), pp. 1–3.

[28] Aidong Zhang. *Protein interaction networks: computational analysis*. Cambridge University Press, 2009.

[29] Charlotte M Deane et al. “Protein interactions: two methods for assessment of the reliability of high throughput observations”. In: *Molecular & Cellular Proteomics* 1.5 (2002), pp. 349–356.

[30] Christian Von Mering et al. “Comparative assessment of large-scale data sets of protein–protein interactions”. In: *Nature* 417.6887 (2002), pp. 399–403.

[31] Michael E Cusick et al. “Literature-curated protein interaction datasets”. In: *Nature methods* 6.1 (2009), pp. 39–46.

[32] V Srinivasa Rao et al. “Protein-protein interaction detection: methods and analysis”. In: *International journal of proteomics* 2014 (2014).

[33] G Traver Hart, Arun K Ramani, and Edward M Marcotte. “How complete are current yeast and human protein-interaction networks?” In: *Genome biology* 7.11 (2006), pp. 1–9.

[34] Jamie Snider et al. “Fundamentals of protein interaction network mapping”. In: *Molecular systems biology* 11.12 (2015), p. 848.

[35] Stanley Fields and Ok-kyu Song. “A novel genetic system to detect protein–protein interactions”. In: *Nature* 340.6230 (1989), pp. 245–246.

[36] Anne-Claude Gavin et al. “Functional organization of the yeast proteome by systematic analysis of protein complexes”. In: *Nature* 415.6868 (2002), pp. 141–147.

[37] Sean R Collins et al. “Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae”. In: *Molecular & Cellular Proteomics* 6.3 (2007), pp. 439–450.

[38] Ruedi Aebersold et al. “How many human proteoforms are there?” In: *Nature chemical biology* 14.3 (2018), pp. 206–214.

[39] Michael PH Stumpf et al. “Estimating the size of the human interactome”. In: *Proceedings of the National Academy of Sciences* 105.19 (2008), pp. 6959–6964.

[40] Cecilia Blikstad and Ylva Ivarsson. “High-throughput methods for identification of protein-protein interactions involving short linear motifs”. In: *Cell Communication and Signaling* 13.1 (2015), pp. 1–9.

[41] Rose Oughtred et al. “The BioGRID interaction database: 2019 update”. In: *Nucleic acids research* 47.D1 (2019), pp. D529–D541.

[42] J Michael Cherry et al. “SGD: Saccharomyces genome database”. In: *Nucleic acids research* 26.1 (1998), pp. 73–79.

[43] Hans-Werner Mewes et al. “MIPS: a database for genomes and protein sequences”. In: *Nucleic acids research* 28.1 (2000), pp. 37–40.

[44] Ioannis Xenarios et al. “DIP: the database of interacting proteins”. In: *Nucleic acids research* 28.1 (2000), pp. 289–291.

[45] TS Keshava Prasad et al. “Human protein reference database—2009 update”. In: *Nucleic acids research* 37.suppl_1 (2009), pp. D767–D772.

[46] Luana Licata et al. “MINT, the molecular interaction database: 2012 update”. In: *Nucleic acids research* 40.D1 (2012), pp. D857–D861.

[47] Christian von Mering et al. “STRING: a database of predicted functional associations between proteins”. In: *Nucleic acids research* 31.1 (2003), pp. 258–261.

[48] J Bellay and Chad L Myers. “Genetic Redundancy”. In: *Brenner’s Encyclopedia of Genetics: Second Edition.* Elsevier Inc., 2013, pp. 281–283.

[49] Scott J Dixon et al. "Systematic mapping of genetic interaction networks". In: *Annual review of genetics* 43 (2009), pp. 601–625.

[50] Anastasia Baryshnikova et al. "Quantitative analysis of fitness and genetic interactions in yeast on a genome scale". In: *Nature methods* 7.12 (2010), pp. 1017–1024.

[51] Amy Hin Yan Tong et al. "Systematic genetic analysis with ordered arrays of yeast deletion mutants". In: *Science* 294.5550 (2001), pp. 2364–2368.

[52] Siew Loon Ooi, Daniel D Shoemaker, and Jef D Boeke. "DNA helicase gene interaction network defined using synthetic lethality analyzed by microarray". In: *Nature genetics* 35.3 (2003), pp. 277–286.

[53] Michael Costanzo et al. "A global genetic interaction network maps a wiring diagram of cellular function". In: *Science* 353.6306 (2016), aaf1420.

[54] Amy Hin Yan Tong et al. "Global mapping of the yeast genetic interaction network". In: *science* 303.5659 (2004), pp. 808–813.

[55] Jim Thurmond et al. "FlyBase 2.0: the next generation". In: *Nucleic acids research* 47.D1 (2019), pp. D759–D765.

[56] Joshua M Stuart et al. "A gene-coexpression network for global discovery of conserved genetic modules". In: *science* 302.5643 (2003), pp. 249–255.

[57] Takeshi Obayashi et al. "COXPRESdb: a database of coexpressed gene networks in mammals". In: *Nucleic acids research* 36.suppl_1 (2007), pp. D77–D82.

[58] Ramon Xulvi-Brunet and Hongzhe Li. "Co-expression networks: graph properties and topological comparisons". In: *Bioinformatics* 26.2 (2010), pp. 205–214.

[59] Takeshi Obayashi et al. "COXPRESdb v8: an animal gene coexpression database navigating from a global view to detailed investigations". In: *Nucleic Acids Research* 51.D1 (2023), pp. D80–D87.

[60] Sipko van Dam, Thomas Craig, and João Pedro de Magalhães. "GeneFriends: a human RNA-seq-based gene and transcript co-expression database". In: *Nucleic acids research* 43.D1 (2015), pp. D1124–D1132.

[61] Hawoong Jeong et al. "The large-scale organization of metabolic networks". In: *Nature* 407.6804 (2000), pp. 651–654.

[62] Hongwu Ma and An-Ping Zeng. "Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms". In: *Bioinformatics* 19.2 (2003), pp. 270–277.

[63] Jane B Reece et al. *Campbell biology.* Vol. 9. Pearson Boston, 2011.

[64] Balaji Veeramani and Joel S Bader. "Predicting functional associations from metabolism using bi-partite network algorithms". In: *BMC Systems Biology* 4.1 (2010), pp. 1–15.

[65] Minoru Kanehisa et al. "KEGG for linking genomes to life and the environment". In: *Nucleic acids research* 36.suppl_1 (2007), pp. D480–D484.

[66] Christiane Hertz-Fowler, Christopher S Peacock, et al. "Introducing GeneDB: a generic database". In: *TRENDS in Parasitology* 18.10 (2002), pp. 465–467.

[67] John W Whitaker et al. “metaTIGER: a metabolic evolution resource”. In: *Nucleic Acids Research* 37.suppl_1 (2009), pp. D531–D538.

[68] Ross Overbeek et al. “The ERGO TM genome analysis and discovery system”. In: *Nucleic acids research* 31.1 (2003), pp. 164–171.

[69] Patricia J Wittkopp and Gizem Kalay. “Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence”. In: *Nature Reviews Genetics* 13.1 (2012), pp. 59–69.

[70] Rasika Mundade et al. “Role of ChIP-seq in the discovery of transcription factor binding sites, differential gene regulation mechanism, epigenetic marks and beyond”. In: *Cell Cycle* 13.18 (2014), pp. 2847–2852.

[71] Peter D Karp et al. “The ecocyc database”. In: *EcoSal Plus* 8.1 (2018).

[72] Fedor A Kolpakov et al. “GeneNet: a gene network database and its automated visualization.” In: *Bioinformatics (Oxford, England)* 14.6 (1998), pp. 529–537.

[73] Antonio Fabregat et al. “The reactome pathway knowledgebase”. In: *Nucleic Acids Research* 44.D1 (2016), pp. D481–D487.

[74] Araceli M Huerta et al. “RegulonDB: a database on transcriptional regulation in Escherichia coli”. In: *Nucleic acids research* 26.1 (1998), pp. 55–59.

[75] Aziz Khan et al. “JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework”. In: *Nucleic acids research* 46.D1 (2018), pp. D260–D266.

[76] Holger Dinkel et al. “Phospho. ELM: a database of phosphorylation sites—update 2011”. In: *Nucleic acids research* 39.suppl_1 (2010), pp. D261–D267.

[77] Florian Gnad, Jeremy Gunawardena, and Matthias Mann. “PHOSIDA 2011: the posttranslational modification database”. In: *Nucleic acids research* 39.suppl_1 (2010), pp. D253–D260.

[78] Mathias Krull et al. “TRANSPATH®: an information resource for storing and visualizing signaling pathways and their pathological aberrations”. In: *Nucleic acids research* 34.suppl_1 (2006), pp. D546–D551.

[79] Timothy Jewison et al. “SMPDB 2.0: big improvements to the Small Molecule Pathway Database”. In: *Nucleic acids research* 42.D1 (2014), pp. D478–D484.

[80] Narat J Eungdamrong and Ravi Iyengar. “Modeling cell signaling networks”. In: *Biology of the Cell* 96.5 (2004), pp. 355–362.

[81] Evren U Azeloglu and Ravi Iyengar. “Signaling networks: information flow, computation, and decision making”. In: *Cold Spring Harbor perspectives in biology* 7.4 (2015), a005934.

[82] Luke E Ulrich and Igor B Zhulin. “MiST: a microbial signal transduction database”. In: *Nucleic acids research* 35.suppl_1 (2007), pp. D386–D390.

[83] Jerome Y Lettvin et al. “What the frog’s eye tells the frog’s brain”. In: *Proceedings of the IRE* 47.11 (1959), pp. 1940–1951.

[84] Kwang-Il Goh et al. “The human disease network”. In: *Proceedings of the National Academy of Sciences* 104.21 (2007), pp. 8685–8690.

[85] Muhammed AY et al. “Drug–target network”. In: *Nature biotechnology* 25.10 (2007), pp. 1119–1127.

[86] Wolfgang Huber et al. “Graphs in molecular biology”. In: *BMC bioinformatics* 8.6 (2007), pp. 1–14.

[87] Naoki Masuda, Mason A Porter, and Renaud Lambiotte. “Random walks and diffusion on networks”. In: *Physics reports* 716 (2017), pp. 1–58.

[88] Thomas H Cormen et al. *Introduction to algorithms*. MIT press, 2022.

[89] Alexandros Xenos et al. “Linear functional organization of the omic embedding space”. In: *Bioinformatics* 37.21 (2021), pp. 3839–3847.

[90] Pall F Jonsson and Paul A Bates. “Global topological features of cancer proteins in the human interactome”. In: *Bioinformatics* 22.18 (2006), pp. 2291–2297.

[91] Hawoong Jeong et al. “Lethality and centrality in protein networks”. In: *Nature* 411.6833 (2001), pp. 41–42.

[92] Stéphane Coulomb et al. “Gene essentiality and the topology of protein interaction networks”. In: *Proceedings of the Royal Society B: Biological Sciences* 272.1573 (2005), pp. 1721–1725.

[93] Albert-László Barabási and Eric Bonabeau. “Scale-free networks”. In: *Scientific american* 288.5 (2003), pp. 60–69.

[94] Albert-Laszlo Barabasi and Zoltan N Oltvai. “Network biology: understanding the cell’s functional organization”. In: *Nature reviews genetics* 5.2 (2004), pp. 101–113.

[95] Réka Albert, Hawoong Jeong, and Albert-László Barabási. “Error and attack tolerance of complex networks”. In: *nature* 406.6794 (2000), pp. 378–382.

[96] Robert D Leclerc. “Survival of the sparsest: robust gene networks are parsimonious”. In: *Molecular systems biology* 4.1 (2008), p. 213.

[97] Duncan J Watts and Steven H Strogatz. “Collective dynamics of ‘small-world’networks”. In: *nature* 393.6684 (1998), pp. 440–442.

[98] Erzsébet Ravasz et al. “Hierarchical organization of modularity in metabolic networks”. In: *science* 297.5586 (2002), pp. 1551–1555.

[99] Jeff Clune, Jean-Baptiste Mouret, and Hod Lipson. “The evolutionary origins of modularity”. In: *Proceedings of the Royal Society b: Biological sciences* 280.1755 (2013), p. 20122863.

[100] Elena Zotenko et al. “Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality”. In: *PLoS computational biology* 4.8 (2008), e1000140.

[101] Hong-Wu Ma and An-Ping Zeng. “The connectivity structure, giant strong component and centrality of metabolic networks”. In: *Bioinformatics* 19.11 (2003), pp. 1423–1430.

[102] Aurélien Mazurie et al. “Evolution of metabolic network organization”. In: *BMC Systems Biology* 4.1 (2010), pp. 1–10.

[103] Stefan Wuchty and Peter F Stadler. “Centers of complex networks”. In: *Journal of theoretical biology* 223.1 (2003), pp. 45–53.

[104] Ron Milo et al. “Network motifs: simple building blocks of complex networks”. In: *Science* 298.5594 (2002), pp. 824–827.

[105] Uri Alon. “Network motifs: theory and experimental approaches”. In: *Nature Reviews Genetics* 8.6 (2007), pp. 450–461.

[106] Shai S Shen-Orr et al. “Network motifs in the transcriptional regulation network of Escherichia coli”. In: *Nature genetics* 31.1 (2002), pp. 64–68.

[107] Esti Yeger-Lotem et al. “Network motifs in integrated cellular networks of transcription–regulation and protein–protein interaction”. In: *Proceedings of the National Academy of Sciences* 101.16 (2004), pp. 5934–5939.

[108] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. “Network medicine: a network-based approach to human disease”. In: *Nature reviews genetics* 12.1 (2011), pp. 56–68.

[109] Ron Milo et al. “Superfamilies of evolved and designed networks”. In: *Science* 303.5663 (2004), pp. 1538–1542.

[110] Dirk Brockmann and Dirk Helbing. “The hidden geometry of complex, network-driven contagion phenomena”. In: *science* 342.6164 (2013), pp. 1337–1342.

[111] Dane Taylor et al. “Topological data analysis of contagion maps for examining spreading processes on networks”. In: *Nature communications* 6.1 (2015), p. 7723.

[112] Oleksii Kuchaiev et al. “Geometric de-noising of protein-protein interaction networks”. In: *PLoS computational biology* 5.8 (2009), e1000454.

[113] Michael Lappe and Liisa Holm. “Unraveling protein interaction networks with near-optimal efficiency”. In: *Nature biotechnology* 22.1 (2004), pp. 98–103.

[114] Paul Erdős and Alfréd Rényi. “On the strength of connectedness of a random graph”. In: *Acta Mathematica Hungarica* 12.1 (1961), pp. 261–267.

[115] Béla Bollobás and Oliver Riordan. “Random graphs and branching processes”. In: *Handbook of large-scale random networks* (2008), pp. 15–115.

[116] Mark EJ Newman. “The structure and function of networks”. In: *Computer Physics Communications* 147.1-2 (2002), pp. 40–45.

[117] Alexei Vázquez et al. “Modeling of protein interaction networks”. In: *Complexus* 1.1 (2003), pp. 38–44.

[118] Mathew Penrose. *Random geometric graphs*. Vol. 5. OUP Oxford, 2003.

[119] Nataša Pržulj et al. “Geometric evolutionary dynamics of protein interaction networks”. In: *Biocomputing 2010*. World Scientific, 2010, pp. 178–189.

[120] Desmond J Higham, Marija Rašajski, and Nataša Pržulj. “Fitting a geometric graph to a protein–protein interaction network”. In: *Bioinformatics* 24.8 (2008), pp. 1093–1099.

[121] Nataša Pržulj and Desmond J Higham. “Modelling protein–protein interaction networks via a stickiness index”. In: *Journal of the Royal Society Interface* 3.10 (2006), pp. 711–716.

[122] Nickolay Smirnov. “Table for estimating the goodness of fit of empirical distributions”. In: *The annals of mathematical statistics* 19.2 (1948), pp. 279–281.

[123] Henry B Mann and Donald R Whitney. "On a test of whether one of two random variables is stochastically larger than the other". In: *The annals of mathematical statistics* (1947), pp. 50–60.

[124] Sourav S Bhowmick and Boon Siew Seah. "Clustering and summarizing protein-protein interaction networks: A survey". In: *IEEE Transactions on Knowledge and Data Engineering* 28.3 (2015), pp. 638–658.

[125] Lei Chen et al. "Inferring novel genes related to oral cancer with a network embedding method and one-class learning algorithms". In: *Gene Therapy* 26.12 (2019), pp. 465–478.

[126] Heng Lian. "MOST: detecting cancer differential gene expression". In: *Biostatistics* 9.3 (2008), pp. 411–418.

[127] Wei Zhong et al. "Improved K-means clustering algorithm for exploring local protein sequence motifs representing common structural property". In: *IEEE transactions on Nanobioscience* 4.3 (2005), pp. 255–265.

[128] Preeti Arora, Shipra Varshney, et al. "Analysis of k-means and k-medoids algorithm for big data". In: *Procedia Computer Science* 78 (2016), pp. 507–512.

[129] Oleksii Kuchaiev et al. "Topological network alignment uncovers biological function and phylogeny". In: *Journal of the Royal Society Interface* 7.50 (2010), pp. 1341–1354.

[130] Gene Ontology Consortium. "The Gene Ontology (GO) database and informatics resource". In: *Nucleic acids research* 32.suppl_1 (2004), pp. D258–D261.

[131] Yanpeng Li and Hong Yu. "A robust data-driven approach for gene ontology annotation". In: *Database* 2014 (2014).

[132] David L Wheeler et al. "Database resources of the national center for biotechnology information". In: *Nucleic acids research* 35.suppl_1 (2007), pp. D5–D12.

[133] Fran Supek et al. "REVIGO summarizes and visualizes long lists of gene ontology terms". In: *PloS one* 6.7 (2011), e21800.

[134] Anand Rajaraman and Jeffrey David Ullman. *Mining of massive datasets*. Cambridge University Press, 2011.

[135] Philip Resnik and David Yarowsky. "Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation". In: *Natural Language Engineering* 5.2 (1999), pp. 113–133.

[136] Dekang Lin et al. "An information-theoretic definition of similarity." In: *Icml*. Vol. 98. 1998, pp. 296–304.

[137] Nataša Pržulj. "Biological network comparison using graphlet degree distribution". In: *Bioinformatics* 23.2 (2007), e177–e183.

[138] Ryan W Solava, Ryan P Michaels, and Tijana Milenković. "Graphlet-based edge clustering reveals pathogen-interacting proteins". In: *Bioinformatics* 28.18 (2012), pp. i480–i486.

[139] Tijana Milenković et al. “Dominating biological networks”. In: *PloS one* 6.8 (2011), e23016.

[140] Yuriy Hulovatyy, Huili Chen, and Tijana Milenković. “Exploring the structure and function of temporal networks with dynamic graphlets”. In: *Bioinformatics* 31.12 (2015), pp. i171–i180.

[141] Yuriy Hulovatyy, Ryan W Solava, and Tijana Milenković. “Revealing missing parts of the interactome via link prediction”. In: *PloS one* 9.3 (2014), e90073.

[142] Vladimir Vacic et al. “Graphlet kernels for prediction of functional residues in protein structures”. In: *Journal of Computational Biology* 17.1 (2010), pp. 55–72.

[143] Noël Malod-Dognin and Nataša Pržulj. “GR-Align: fast and flexible alignment of protein 3D structures using graphlet degree similarity”. In: *Bioinformatics* 30.9 (2014), pp. 1259–1265.

[144] Jose Lugo-Martinez and Predrag Radivojac. “Generalized graphlet kernels for probabilistic inference in sparse graphs”. In: *Network Science* 2.2 (2014), pp. 254–276.

[145] Serene WH Wong, Nick Cercone, and Igor Jurisica. “Comparative network analysis via differential graphlet communities”. In: *Proteomics* 15.2-3 (2015), pp. 608–617.

[146] Tijana Milenković et al. “Optimal network alignment with graphlet degree vectors”. In: *Cancer informatics* 9 (2010), CIN–S4744.

[147] Mu-Fen Hsieh and Sing-Hoi Sze. “Finding alignments of conserved graphlets in protein interaction networks”. In: *Journal of Computational Biology* 21.3 (2014), pp. 234–246.

[148] Fazle E Faisal and Tijana Milenković. “Dynamic networks reveal key players in aging”. In: *Bioinformatics* 30.12 (2014), pp. 1721–1729.

[149] Tijana Milenković et al. “Systems-level cancer gene identification from protein interaction network topology applied to melanogenesis-related functional genomics data”. In: *Journal of the Royal Society Interface* 7.44 (2010), pp. 423–437.

[150] Hsiang Ho et al. “Protein interaction network topology uncovers melanogenesis regulatory network components within functional genomics datasets”. In: *BMC systems biology* 4 (2010), pp. 1–13.

[151] Sam FL Windels, Noël Malod-Dognin, and Nataša Pržulj. “Graphlet Laplacians: graphlet-based neighbourhoods highlight topology-function and topology-disease relationships”. In: *bioRxiv* (2018), p. 460964.

[152] Xiao-Dong Wang et al. “Identification of human disease genes from interactome network using graphlet interaction”. In: *PloS one* 9.1 (2014), e86142.

[153] Nataša Pržulj. “Protein-protein interactions: making sense of networks via graph-theoretic modeling”. In: *Bioessays* 33.2 (2011), pp. 115–123.

[154] Rob Patro and Carl Kingsford. “Global network alignment using multiscale spectral signatures”. In: *Bioinformatics* 28.23 (2012), pp. 3105–3114.

[155] Shawn Gu et al. "From homogeneous to heterogeneous network alignment via colored graphlets". In: *Scientific reports* 8.1 (2018), p. 12524.

[156] Michelle M Li, Kexin Huang, and Marinka Zitnik. "Graph representation learning in biomedicine and healthcare". In: *Nature Biomedical Engineering* (2022), pp. 1–17.

[157] Hao Zhou, Juan Felipe Beltrán, and Ilana Lauren Brito. "Functions predict horizontal gene transfer and the emergence of antibiotic resistance". In: *Science Advances* 7.43 (2021), eabj5056.

[158] Kexin Huang et al. "SkipGNN: predicting molecular interactions with skip-graph networks". In: *Scientific reports* 10.1 (2020), pp. 1–16.

[159] Aparna Rai, Pramod Shinde, and Sarika Jalan. "Network spectra for drug-target identification in complex diseases: new guns against old foes". In: *Applied Network Science* 3 (2018), pp. 1–18.

[160] Jiaxuan You et al. "Graph convolutional policy network for goal-directed molecular graph generation". In: *Advances in neural information processing systems* 31 (2018).

[161] Han Xu et al. "A network embedding based method for partial multi-omics integration in cancer subtyping". In: *Methods* 192 (2021), pp. 67–76.

[162] Vladimir Gligorijević, Noël Malod-Dognin, and Nataša Pržulj. "Patient-specific data fusion for cancer stratification and personalised treatment". In: *Biocomputing 2016: Proceedings of the Pacific Symposium.* World Scientific. 2016, pp. 321–332.

[163] Gaia Ceddia et al. "Matrix factorization-based technique for drug repurposing predictions". In: *IEEE Journal of Biomedical and Health Informatics* 24.11 (2020), pp. 3162–3172.

[164] Jian Tang et al. "Line: Large-scale information network embedding". In: *Proceedings of the 24th international conference on world wide web.* 2015, pp. 1067–1077.

[165] Tomas Mikolov et al. "Distributed representations of words and phrases and their compositionality". In: *Advances in neural information processing systems* 26 (2013).

[166] Jiezhong Qiu et al. "Network embedding as matrix factorization: Unifying deepwalk, line, pte, and node2vec". In: *Proceedings of the eleventh ACM international conference on web search and data mining.* 2018, pp. 459–467.

[167] Omer Levy and Yoav Goldberg. "Neural word embedding as implicit matrix factorization". In: *Advances in neural information processing systems* 27 (2014).

[168] Sini Isokääntä et al. "Comparison of dimension reduction techniques in the analysis of mass spectrometry data". In: *Atmospheric Measurement Techniques* 13.6 (2020), pp. 2995–3022.

[169] Daniel D Lee and H Sebastian Seung. "Learning the parts of objects by non-negative matrix factorization". In: *Nature* 401.6755 (1999), pp. 788–791.

[170] RZdunek ACichocki et al. *NonnegativeMatrixandTensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation.* 2009.

[171] Ron Zass and Amnon Shashua. "A unifying approach to hard and probabilistic clustering". In: *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1.* Vol. 1. IEEE. 2005, pp. 294–301.

[172] Gilbert Strang. "Linear algebra and its applications. 4th". In: *Brooks Cole* (2006).

[173] Sergio Doria-Belenguer et al. "Probabilistic graphlets capture biological function in probabilistic molecular networks". In: *Bioinformatics* 36.Supplement_2 (2020), pp. i804–i812.

[174] Pedro Ribeiro and Fernando Silva. "Querying subgraph sets with g-tries". In: *Proceedings of the 2nd ACM SIGMOD Workshop on Databases and Social Networks.* 2012, pp. 25–30.

[175] DV Klopfenstein et al. "GOATOOLS: A Python library for Gene Ontology analyses". In: *Scientific reports* 8.1 (2018), pp. 1–17.

[176] Tim Hubbard et al. "The Ensembl genome database project". In: *Nucleic acids research* 30.1 (2002), pp. 38–41.

[177] Leland McInnes, John Healy, and James Melville. "Umap: Uniform manifold approximation and projection for dimension reduction". In: *arXiv preprint arXiv:1802.03426* (2018).

[178] Laurens Van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE." In: *Journal of machine learning research* 9.11 (2008).

[179] Yoav Benjamini and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the Royal statistical society: series B (Methodological)* 57.1 (1995), pp. 289–300.

[180] Tim Hesterberg. "Bootstrap". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 3.6 (2011), pp. 497–526.

[181] Paul Jaccard. "The distribution of the flora in the alpine zone. 1". In: *New phytologist* 11.2 (1912), pp. 37–50.

[182] Stephen Robertson. "Understanding inverse document frequency: on theoretical arguments for IDF". In: *Journal of documentation* (2004).

[183] Jeffrey M Wooldridge. "Applications of generalized method of moments estimation". In: *Journal of Economic perspectives* 15.4 (2001), pp. 87–100.

[184] Ruipeng Lu and Peter K Rogan. "Transcription factor binding site clusters identify target genes with similar tissue-wide expression and buffer against mutations". In: *F1000Research* 7 (2018).

[185] Lin Song, Peter Langfelder, and Steve Horvath. "Comparison of co-expression measures: mutual information, correlation, and model based indices". In: *BMC bioinformatics* 13.1 (2012), pp. 1–21.

[186] Kate Campbell et al. "Building blocks are synthesized on demand during the yeast cell cycle". In: *Proceedings of the National Academy of Sciences* 117.14 (2020), pp. 7575–7583.

[187] Paul S Maddox, Kerry S Bloom, and ED Salmon. "The polarity and dynamics of microtubule assembly in the budding yeast Saccharomyces cerevisiae". In: *Nature cell biology* 2.1 (2000), pp. 36–41.

[188] Marisa Segal and Kerry Bloom. "Control of spindle polarity and orientation in Saccharomyces cerevisiae". In: *Trends in cell biology* 11.4 (2001), pp. 160–166.

[189] Sergio Doria-Belenguer et al. "A functional analysis of omic network embedding spaces reveals key altered functions in cancer". In: *Bioinformatics* (2023).

[190] Jing Jin et al. "Identification of genetic mutations in cancer: challenge and opportunity in the new era of targeted therapy". In: *Frontiers in Oncology* 9 (2019), p. 263.

[191] Alex Bateman et al. "The gene ontology resource: 20 years and still GOing strong". In: *Nucleic Acids Research* 47.D1 (2019), pp. D330–D338.

[192] Fredrik Pontén, Karin Jirström, and Matthias Uhlen. "The Human Protein Atlas—a tool for pathology". In: *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland* 216.4 (2008), pp. 387–393.

[193] Eva YHP Lee and William J Muller. "Oncogenes and tumor suppressor genes". In: *Cold Spring Harbor perspectives in biology* 2.10 (2010), a003236.

[194] Emmanouil P Pappou and Nita Ahuja. "The role of oncogenes in gastrointestinal cancer". In: *Gastrointestinal Cancer Research: GCR* 1.Suppl (2010), S2.

[195] Carolina Vicente-Dueñas et al. "Function of oncogenes in cancer development: a changing paradigm". In: *The EMBO journal* 32.11 (2013), pp. 1502–1513.

[196] Simon A Forbes et al. "COSMIC: somatic cancer genetics at high-resolution". In: *Nucleic Acids Research* 45.D1 (2017), pp. D777–D783.

[197] John A Rice. *Mathematical statistics and data analysis*. Cengage Learning, 2006.

[198] James Dean Brown. "The Bonferroni adjustment". In: *Statistics* 12.1 (2008).

[199] Yi Chen, Fons Verbeek, Katherine Wolstencroft, et al. "Establishing a consensus for the hallmarks of cancer based on gene ontology and pathway annotationsF". In: *BMC Bioinformatics* 22.1 (2021), pp. 1–20.

[200] Liang Hu et al. "Nonnegative matrix tri-factorization with user similarity for clustering in point-of-interest". In: *Neurocomputing* 363 (2019), pp. 58–65. ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2019.07.040. URL: https://www.sciencedirect.com/science/article/pii/S0925231219310057.

[201] Hanli Qiao. "New SVD based initialization strategy for non-negative matrix factorization". In: *Pattern Recognition Letters* 63 (2015), pp. 71–77.

[202] João Carlos Alves Barata and Mahir Saleh Hussein. "The Moore–Penrose pseudoinverse: A tutorial review of the theory". In: *Brazilian Journal of Physics* 42.1 (2012), pp. 146–165.

[203] Amit Singhal et al. “Modern information retrieval: A brief overview”. In: *IEEE Data Eng. Bull.* 24.4 (2001), pp. 35–43.

[204] Thomas Gaudelet, Noël Malod-Dognin, and Nataša Pržulj. “Integrative data analytic framework to enhance cancer precision medicine”. In: *Network and systems medicine* 4.1 (2021), pp. 60–73.

[205] Jacob Benesty et al. “Pearson correlation coefficient”. In: *Noise Reduction in Speech Processing.* Springer, 2009, pp. 1–4.

[206] Hae-Sang Park and Chi-Hyuck Jun. “A simple and fast algorithm for K-medoids clustering”. In: *Expert systems with Applications* 36.2 (2009), pp. 3336–3341.

[207] Lewis Y Geer et al. “The NCBI biosystems database”. In: *Nucleic Acids Research* 38.suppl_1 (2010), pp. D492–D496.

[208] Renee B Chang and Gregory L Beatty. “The interplay between innate and adaptive immunity in cancer shapes the productivity of cancer immunosurveillance”. In: *Journal of Leukocyte Biology* 108.1 (2020), pp. 363–376.

[209] Antonia P Sagona and Harald Stenmark. “Cytokinesis and cancer”. In: *FEBS Letters* 584.12 (2010), pp. 2652–2661.

[210] Neil J Ganem, Zuzana Storchova, and David Pellman. “Tetraploidy, aneuploidy and cancer”. In: *Current Opinion in Genetics & Development* 17.2 (2007), pp. 157–162.

[211] Jan HJ Hoeijmakers. “DNA damage, aging, and cancer”. In: *New England Journal of Medicine* 361.15 (2009), pp. 1475–1485.

[212] Ga Eon Kim et al. “Differentially expressed genes in matched normal, cancer, and lymph node metastases predict clinical outcomes in patients with breast cancer”. In: *Applied Immunohistochemistry & Molecular Morphology* 28.2 (2020), pp. 111–122.

[213] Camilla Thygesen et al. “Characterizing disease-associated changes in post-translational modifications by mass spectrometry”. In: *Expert review of proteomics* 15.3 (2018), pp. 245–258.

[214] Vivekananda Shetty et al. “Investigation of ovarian cancer associated sialylation changes in N-linked glycopeptides by quantitative proteomics”. In: *Clinical proteomics* 9.1 (2012), pp. 1–19.

[215] Lou K Povlsen et al. “Systems-wide analysis of ubiquitylation dynamics reveals a key role for PAF15 ubiquitylation in DNA-damage bypass”. In: *Nature cell biology* 14.10 (2012), pp. 1089–1098.

[216] Cathrine Kolster Fog et al. “Loss of PRDM11 promotes MYC-driven lymphomagenesis”. In: *Blood, The Journal of the American Society of Hematology* 125.8 (2015), pp. 1272–1281.

[217] Marilyn Safran et al. “GeneCards Version 3: the human gene integrator”. In: *Database* 2010 (2010).

[218] Fang Lu et al. “CARP is a potential tumor suppressor in gastric carcinoma and a single-nucleotide polymorphism in CARP gene might increase the risk of gastric carcinoma”. In: *PLoS One* 9.5 (2014), e97743.

[219] Madelyn E McCauley et al. “C9orf72 in myeloid cells suppresses STING-induced inflammation”. In: *Nature* 585.7823 (2020), pp. 96–101.

[220] The UniProt Consortium. “UniProt: a hub for protein information”. In: *Nucleic Acids Research* 43.D1 (2015), pp. D204–D212.

[221] Sergio Doria-Belenguer et al. “The axes of biology: a novel axes-based network embedding approach to decipher the fundamental mechanisms of the cell”. In: (2023).

[222] Michael Ku Yu et al. “DDOT: a Swiss army knife for investigating data-driven biological ontologies”. In: *Cell systems* 8.3 (2019), pp. 267–273.

[223] Minoru Kanehisa and Susumu Goto. “KEGG: kyoto encyclopedia of genes and genomes”. In: *Nucleic acids research* 28.1 (2000), pp. 27–30.

[224] Qi Zheng and Xiu-Jie Wang. “GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis”. In: *Nucleic acids research* 36.suppl_2 (2008), W358–W363.

[225] Anastasia Baryshnikova. “Spatial analysis of functional enrichment (SAFE) in large biological networks”. In: *Computational Cell Biology*. Springer, 2018, pp. 249–268.

[226] William M Rand. “Objective criteria for the evaluation of clustering methods”. In: *Journal of the American Statistical association* 66.336 (1971), pp. 846–850.

[227] Shahzad Qaiser and Ramsha Ali. “Text mining: use of TF-IDF to examine the relevance of words to documents”. In: *International Journal of Computer Applications* 181.1 (2018), pp. 25–29.

[228] Nitin Hardeniya. *NLTK essentials*. Packt Publishing, 2015.

[229] Zi Yin and Yuanyuan Shen. “On the dimensionality of word embedding”. In: *Advances in neural information processing systems* 31 (2018).

[230] Jeffrey Pennington, Richard Socher, and Christopher D Manning. “Glove: Global vectors for word representation”. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 1532–1543.

[231] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *Proceedings of naacL-HLT*. 2019, pp. 4171–4186.

[232] Hassan Dianat-Moghadam et al. “TRAIL in oncology: From recombinant TRAIL to nano-and self-targeted TRAIL-based therapies”. In: *Pharmacological research* 155 (2020), p. 104716.

[233] Delphine Mérino et al. “TRAIL in cancer therapy: present and future challenges”. In: *Expert opinion on therapeutic targets* 11.10 (2007), pp. 1299–1314.

[234] Jin Ye et al. “ER stress induces cleavage of membrane-bound ATF6 by the same proteases that process SREBPs”. In: *Molecular cell* 6.6 (2000), pp. 1355–1364.

[235] U Thomas Meier. “RNA modification in Cajal bodies”. In: *RNA biology* 14.6 (2017), pp. 693–700.

[236] Richard D Emes and Seth GN Grant. “Evolution of synapse complexity and diversity”. In: *Annual review of neuroscience* 35 (2012), pp. 111–131.

[237] Saak V Ovsepian, Valerie B O’Leary, and Nikolai P Vesselkin. “Evolutionary origins of chemical synapses”. In: *Vitamins and Hormones* 114 (2020), pp. 1–21.

[238] Daniel B Haack and Navtej Toor. “Retroelement origins of pre-mRNA splicing”. In: *Wiley Interdisciplinary Reviews: RNA* 11.4 (2020), e1589.

[239] Julian Vosseberg et al. “The spread of the first introns in proto-eukaryotic paralogs”. In: *Communications biology* 5.1 (2022), pp. 1–9.

[240] Hanako Bai et al. “Effects of dietary vitamin K3 supplementation on vitamin K1 and K2 (menaquinone) dynamics in dairy cows”. In: *Animal science journal* 93.1 (2022), e13680.

[241] Jessie L Ellis et al. “Dietary vitamin K is remodeled by gut microbiota and influences community composition”. In: *Gut Microbes* 13.1 (2021), p. 1887721.

[242] David L Williams. “Light and the evolution of vision”. In: *Eye* 30.2 (2016), pp. 173–178.

[243] Yun Kee et al. “Evolutionary conservation of cell migration genes: from nematode neurons to vertebrate neural crest”. In: *Genes & development* 21.4 (2007), pp. 391–396.

[244] Philip W Ingham, Yoshiro Nakano, and Claudia Seger. “Mechanisms and functions of Hedgehog signalling across the metazoa”. In: *Nature reviews genetics* 12.6 (2011), pp. 393–406.

[245] Ying He et al. “OsHIPL1, a hedgehog-interacting protein-like 1 protein, increases seed vigour in rice”. In: *Plant Biotechnology Journal* 20.7 (2022), p. 1346.

[246] Cândida Lucas et al. “Yeast Gup1 (2) proteins are homologues of the Hedgehog morphogens acyltransferases HHAT (L): facts and implications”. In: *Journal of Developmental Biology* 4.4 (2016), p. 33.

[247] Jens Rolff. “Why did the acquired immune system of vertebrates evolve?” In: *Developmental & Comparative Immunology* 31.5 (2007), pp. 476–482.

[248] Inês Trancoso, Ryo Morimoto, and Thomas Boehm. “Co-evolution of mutagenic genome editors and vertebrate adaptive immunity”. In: *Current Opinion in Immunology* 65 (2020), pp. 32–41.

[249] Bolin Chen et al. “Identifying protein complexes and functional modules—from static PPI networks to dynamic PPI networks”. In: *Briefings in bioinformatics* 15.2 (2014), pp. 177–194.

[250] P Legendre. “The glycinergic inhibitory synapse”. In: *Cellular and Molecular Life Sciences CMLS* 58.5 (2001), pp. 760–793.

[251] Emma Derbyshire and Rima Obeid. “Choline, neurological development and brain function: a systematic review focusing on the first 1000 days”. In: *Nutrients* 12.6 (2020), p. 1731.

[252] Duncan Howie et al. “The role of lipid metabolism in T lymphocyte differentiation and survival”. In: *Frontiers in immunology* 8 (2018), p. 1949.

[253] Marisa W Friederich et al. “Pathogenic variants in glutamyl-tRNAGln amidotransferase subunits cause a lethal mitochondrial cardiomyopathy disorder”. In: *Nature communications* 9.1 (2018), pp. 1–14.

[254] Setareh Rezatabar et al. “RAS/MAPK signaling functions in oxidative stress, DNA damage response and cancer progression”. In: *Journal of Cellular Physiology* 234.9 (2019), pp. 14951–14965.

[255] Lei Li, Michael Story, and Randy J Legerski. “Cellular responses to ionizing radiation damage”. In: *International Journal of Radiation Oncology* Biology* Physics* 49.4 (2001), pp. 1157–1162.

[256] Katherine LB Borden. “The nuclear pore complex and mRNA export in cancer”. In: *Cancers* 13.1 (2020), p. 42.

[257] Anna E Vilgelm and Ann Richmond. “Chemokines modulate immune surveillance in tumorigenesis, metastasis, and response to immunotherapy”. In: *Frontiers in Immunology* 10 (2019), p. 333.

[258] Nuala A O’Leary et al. “Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation”. In: *Nucleic Acids Research* 44.D1 (2016), pp. D733–D745.

[259] Chris HQ Ding, Tao Li, and Michael I Jordan. “Convex and semi-nonnegative matrix factorizations”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32.1 (2008), pp. 45–55.

[260] Andrew P Bradley. “The use of the area under the ROC curve in the evaluation of machine learning algorithms”. In: *Pattern recognition* 30.7 (1997), pp. 1145–1159.

[261] Ryota Suzuki and Hidetoshi Shimodaira. “Pvclust: an R package for assessing the uncertainty in hierarchical clustering”. In: *Bioinformatics* 22.12 (2006), pp. 1540–1542.

[262] Christos Vaklavas, Scott W Blume, and William E Grizzle. “Translational dysregulation in cancer: molecular insights and potential clinical applications in biomarker development”. In: *Frontiers in Oncology* 7 (2017), p. 158.

[263] Kanchan Kumari, Paula Groza, and Francesca Aguilo. “Regulatory roles of RNA modifications in breast cancer”. In: *NAR Cancer* 3.3 (2021), zcab036.

[264] Alex Yick-Lun So et al. “The unfolded protein response during prostate cancer development”. In: *Cancer and Metastasis Reviews* 28.1 (2009), pp. 219–223.

[265] K Malik and KW Brown. “Epigenetic gene deregulation in cancer”. In: *British journal of cancer* 83.12 (2000), pp. 1583–1588.

[266] Kamilla ME Laidlaw and Chris MacDonald. “Endosomal trafficking of yeast membrane proteins”. In: *Biochemical Society Transactions* 46.6 (2018), pp. 1551–1558.

[267] Chyi-Ying A Chen and Ann-Bin Shyu. “Mechanisms of deadenylation-dependent decay”. In: *Wiley Interdisciplinary Reviews: RNA* 2.2 (2011), pp. 167–183.

[268] Jonathan P Staley and John L Woolford Jr. “Assembly of ribosomes and spliceosomes: complex ribonucleoprotein machines”. In: *Current opinion in cell biology* 21.1 (2009), pp. 109–118.

[269] Arron Sullivan et al. “Multilayered regulation of TORC1-body formation in budding yeast”. In: *Molecular biology of the cell* 30.3 (2019), pp. 400–410.

[270] Ross Cagan. “Principles of Drosophila eye differentiation”. In: *Current topics in developmental biology* 89 (2009), pp. 115–135.

[271] Bipin Kumar Tripathi and Kenneth D Irvine. “The wing imaginal disc”. In: *Genetics* 220.4 (2022), iyac020.

[272] Yinuo Gu et al. “Role of the innate cytokine storm induced by the influenza A virus”. In: *Viral Immunology* 32.6 (2019), pp. 244–251.

Appendix A

Supplementary Information for Chapter 3

A.1 Supplementary Results and Discussion for Chapter 3

A.1.1 pGDV captures different functional information that is not seen by GDV

To show that the probabilistic and unweighted graphlets extract different biological information from the same set of genes, we test the ability of pGDV to capture different functional information that is not identified by GDV. Following section 3.2.7 of Chapter 2, we fix the number of clusters, k, following the rule-of-thumb, and we select the enriched GO terms that are captured uniquely by probabilistic or unweighted graphlets. Then, we retrieve those genes that are responsible for the enrichments of the previous GO terms. Following this, we divide the genes into three sets: unique in probabilistic (s1), unique in unweighted (s3), and the intersection between the two methods (s2). Then, for each of the clusters obtained by our probabilistic method (c1), we computed the pGDV and GDV distances between s1 and s2. Similarly, for each of the clusters obtained by the unweighted method (c2), we compute the pGDV and GDV distances between s3 and s2. Finally, for each cluster, $c1_i$, we compute the pGDV and GDV distances between s1 genes in $c1_i$ and the s2 genes clustered in c2. We repeat this method for all k clusters and calculate the mean between all the clusters across the 10 clustering runs.

As expected, the similarity between the gene signatures (GDVs and pGDVs) of the unique probabilistic genes (s1) and genes in the intersection (s2) is higher when using the pGDV. On the other hand, the similarity between unique unweighted genes (s3) and genes in the intersection (s2) is higher when using the GDVs. In other words, the different clustering of the s2 genes by each method results from their ability to identify different topological similarities.

For instance, in low confidence human PPI, we observe a mean distance of 0.15 between the unique to probabilistic method genes (s1) and the intersection genes (s2) when using pGDV, while this distance is 0.20 when we using GDV. Similarly, we observe that these differences between methods decrease when higher confidence networks are analyzed. For example, in the high-confidence budding yeast PPI network, the mean distance between the unique to unweighted method genes (s3) and the shared genes (s2) is 0.20 and 0.21 when using GDV and pGDV, respectively. These results are expected since when a high confidence threshold is applied, the edge probability distribution in the probabilistic network is higher (see Table 3.1 in Chapter 2), being closer to one and, consequently closer to a binary network.

All these results support our hypothesis that pGDV captures functional information that differs from the information captured by GDV. This suggests that the two approaches may complement each other in extracting functional information from the molecular interaction networks.

A.2 Supplementary Figures for Chapter 3

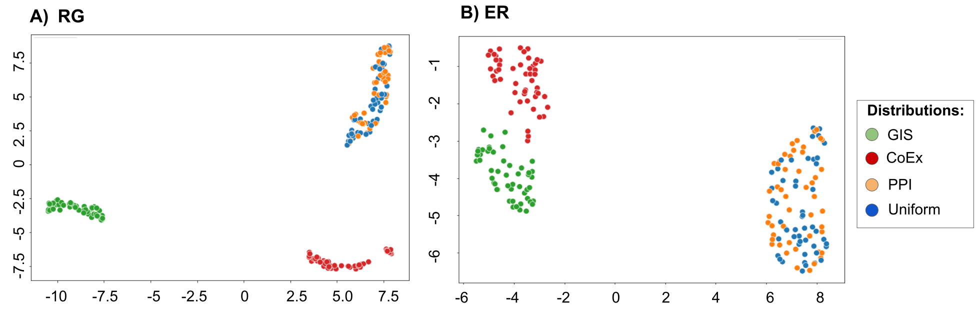


Figure A.1. pGCMs can separate networks by edge probability distribution. Edge probabilities were sampled based on empirical distributions from GIS, CoEx, and PPI or from a synthetic uniform distribution. Panels represent the UMAP embedding of (A) pGCMs of probabilistic geometric random graph (RG) networks using the indicated probability distribution; (B) pGCMs of probabilistic Erdös-Renyi (ER) networks using the indicated probability distribution.

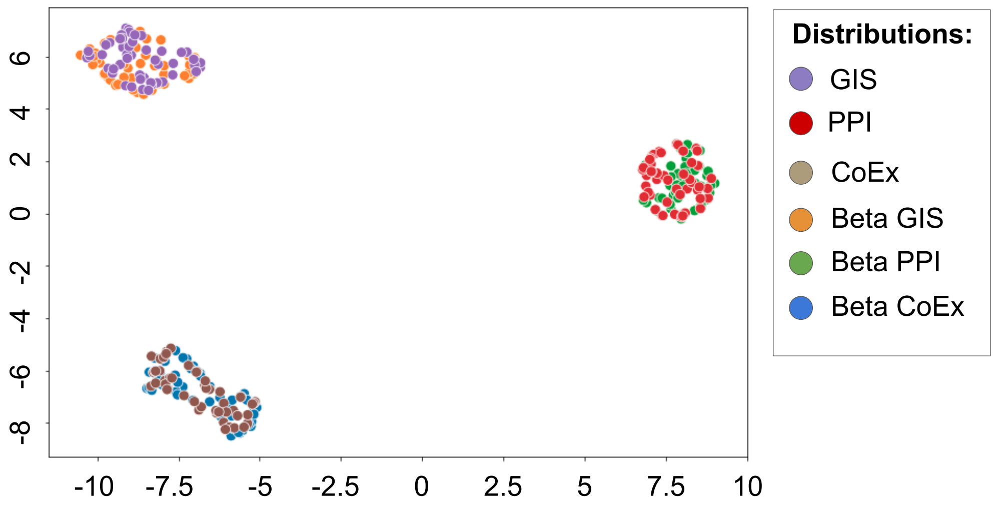


Figure A.2. pGCMs of networks with edge probability distribution sampled based on beta distributions with mean and variance from empirical distributions cluster according to these parameters. Edge probabilities were sampled based on empirical distributions from GIS, CoEx, and PPI or from three different synthetic beta distributions with mean and variance corresponding to each of the empirical distributions. The panel represents the UMAP embedding of probabilistic networks based on the Barabasi and Albert preferential attachment model (BA) using the indicated probability distribution.

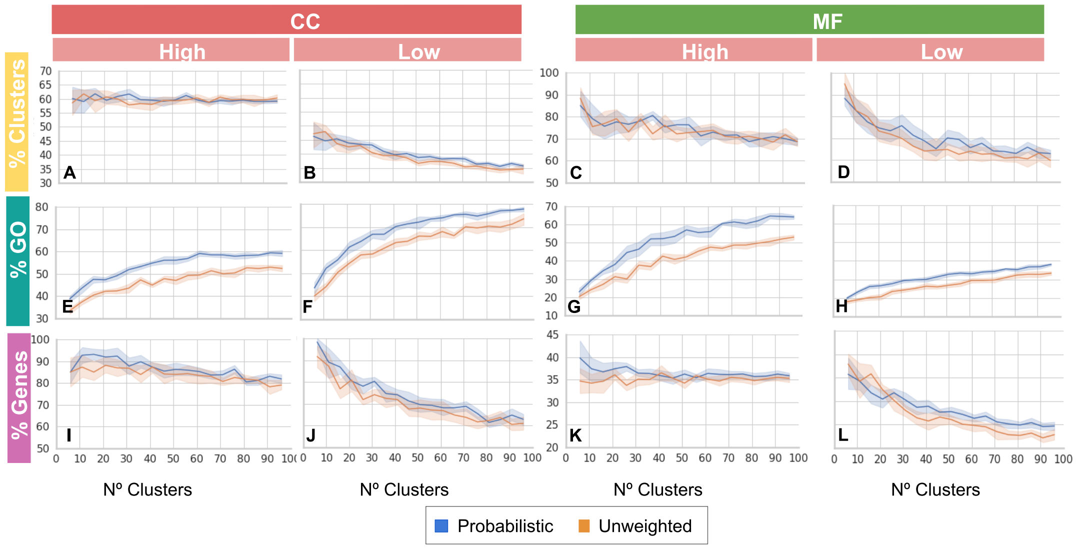


Figure A.3. GO-CC and GO-MF enrichments in the clusters obtained on budding yeast PPI network. Lines represent the mean and the shaded area represents the 95% confidence intervals based on bootstrapping of each enrichment statistic for unweighted and probabilistic graphlets depending on the number of clusters k, across 10 repetitions. High, medium, and low indicate the confidence threshold for the underlying networks.

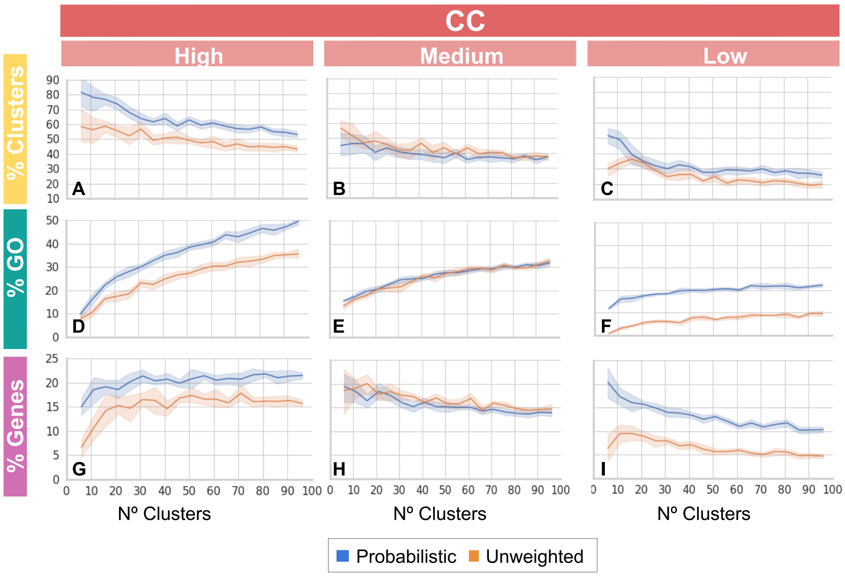


Figure A.4. GO-CC enrichments in the clusters obtained on budding yeast GIS network. Lines represent the mean and the shaded area represents the 95% confidence intervals based on bootstrapping of each enrichment statistic for unweighted and probabilistic graphlets depending on the number of clusters k, across 10 repetitions. High, medium, and low indicate the confidence threshold for the underlying networks.

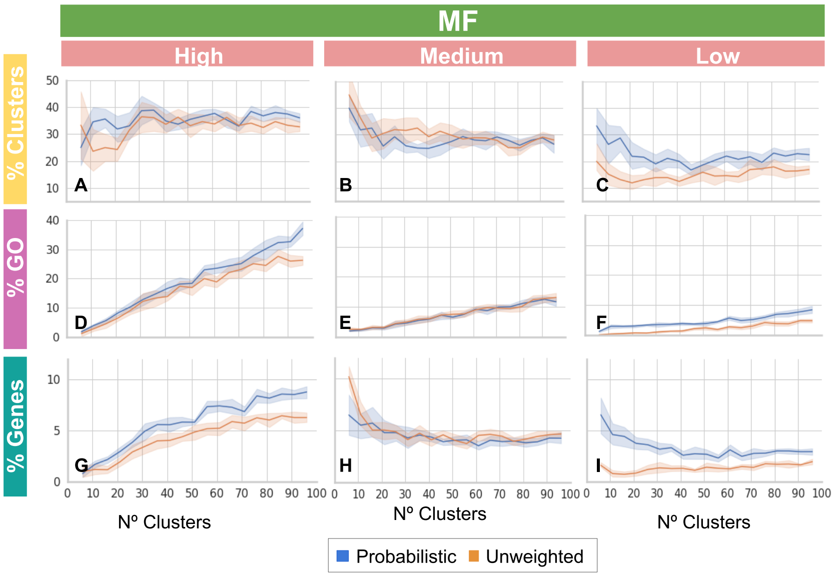


Figure A.5. GO-MF enrichments in the clusters obtained on budding yeast GIS network. Lines represent the mean and the shaded area represents the 95% confidence intervals based on bootstrapping of each enrichment statistic for unweighted and probabilistic graphlets depending on the number of clusters k, across 10 repetitions. High, medium, and low indicate the confidence threshold for the underlying networks.

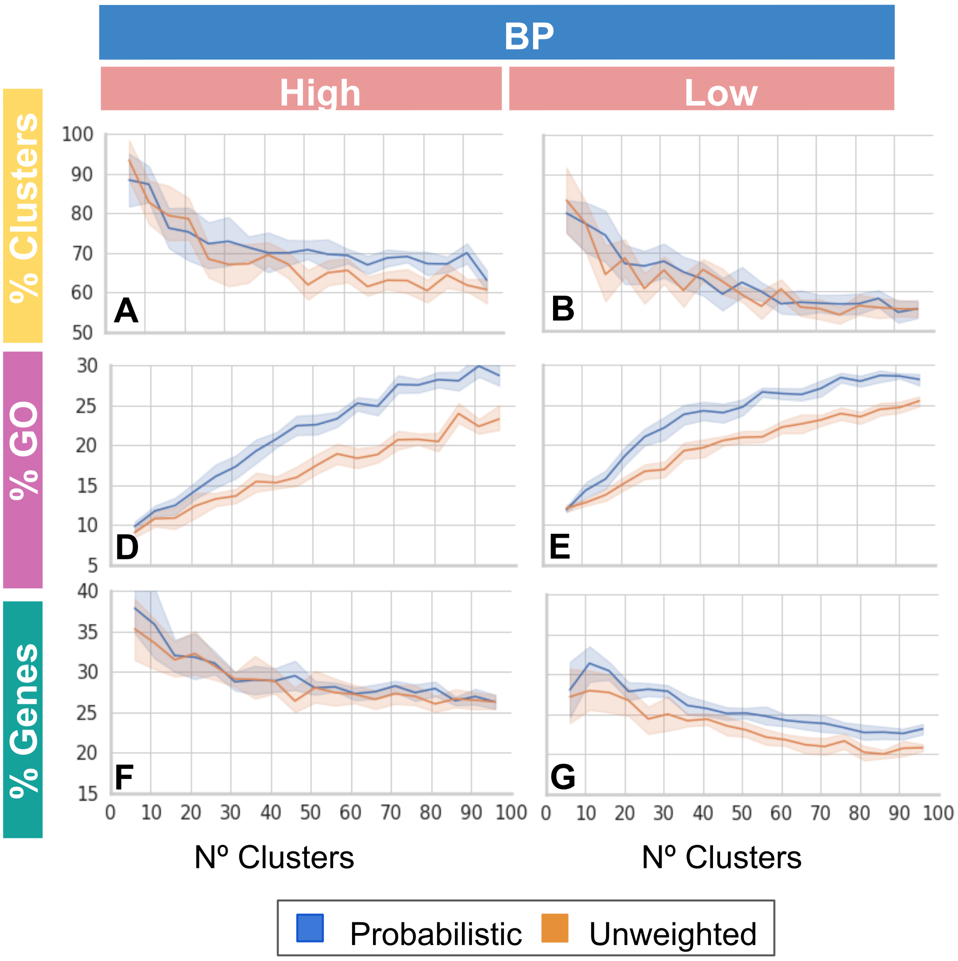


Figure A.6. GO-BP enrichments in the clusters obtained on human PPI network. Lines represent the mean and the shaded area represents the 95% confidence intervals based on bootstrapping of each enrichment statistic for unweighted and probabilistic graphlets depending on the number of clusters k, across 10 repetitions. High, medium, and low indicate the confidence threshold for the underlying networks.

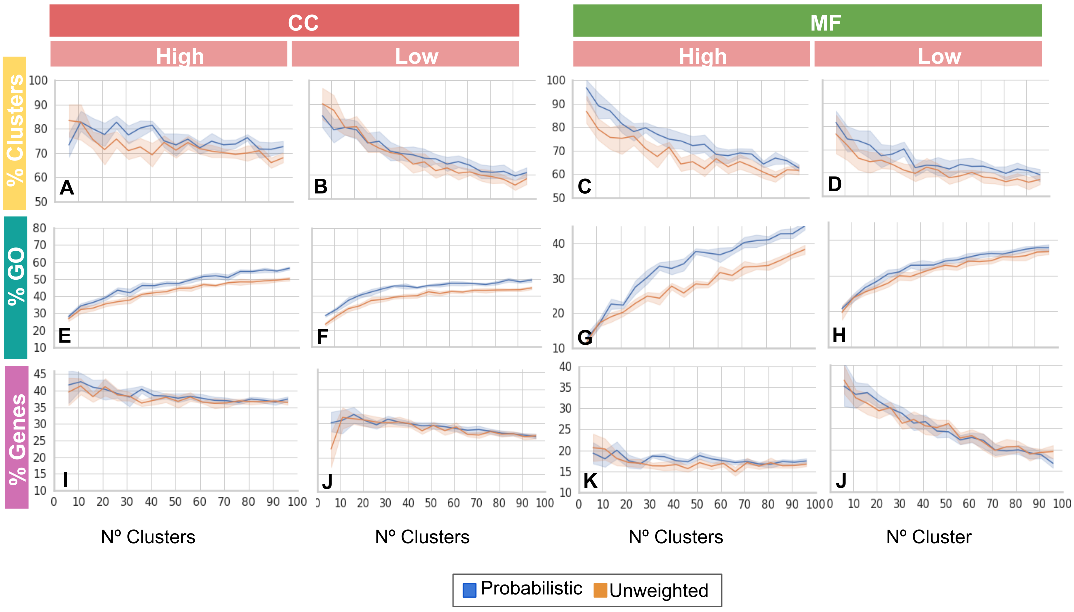


Figure A.7. GO-CC and GO-MF enrichments in the clusters obtained on human PPI network Lines represent the mean and the shaded area represents the 95% confidence intervals based on bootstrapping of each enrichment statistic for unweighted and probabilistic graphlets depending on the number of clusters k, across 10 repetitions. High, medium, and low indicate the confidence threshold for the underlying networks.

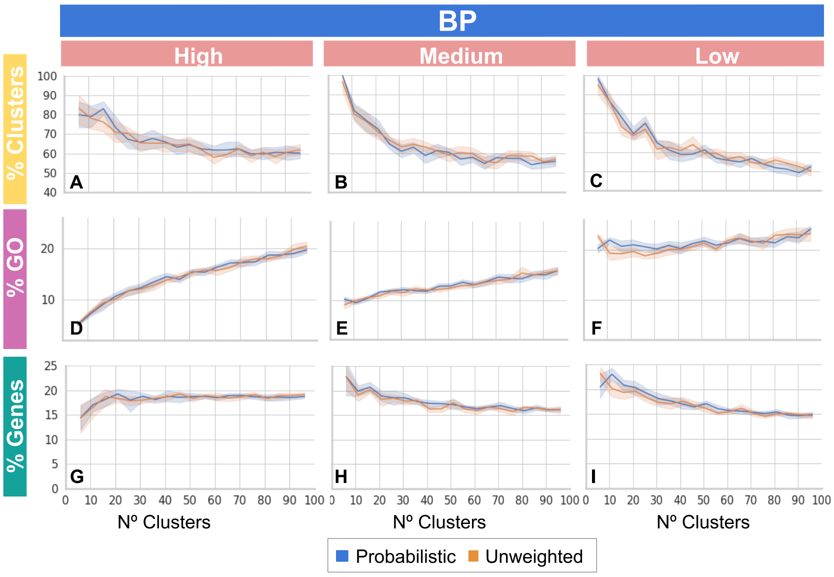


Figure A.8. GO-BP enrichments in the clusters obtained on budding yeast CoEx network. Lines represent the mean and the shaded area represents the 95% confidence intervals based on bootstrapping of each enrichment statistic for unweighted and probabilistic graphlets depending on the number of clusters k, across 10 repetitions. High, medium, and low indicate the confidence threshold for the underlying networks.

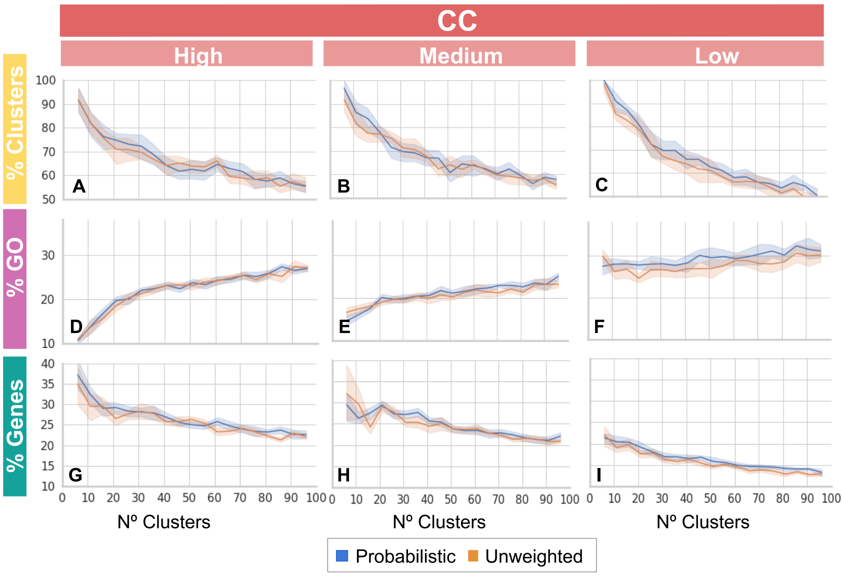


Figure A.9. GO-CC enrichments in the clusters obtained on budding yeast CoEx network. Lines represent the mean and the shaded area represents the 95% confidence intervals based on bootstrapping of each enrichment statistic for unweighted and probabilistic graphlets depending on the number of clusters k, across 10 repetitions. High, medium, and low indicate the confidence threshold for the underlying networks.

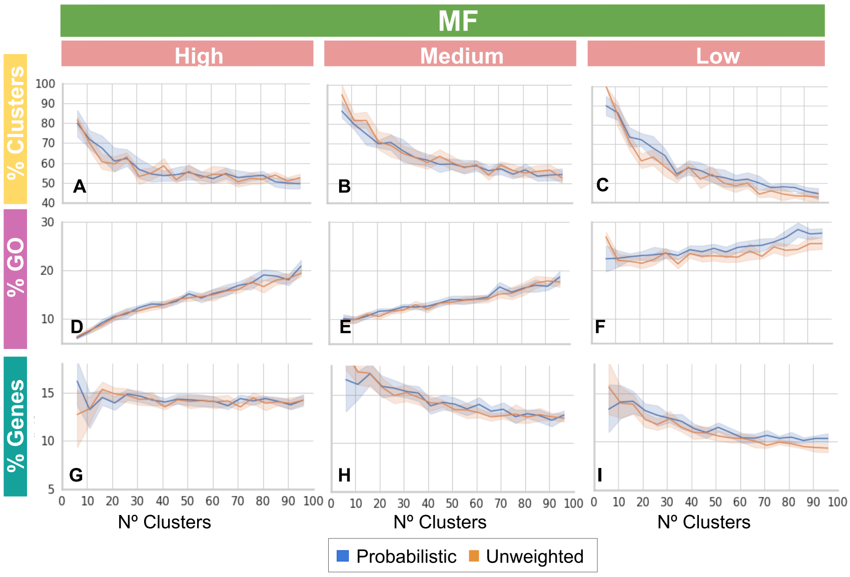


Figure A.10. GO-MF enrichments in the clusters obtained on budding yeast CoEx network. Lines represent the mean and the shaded area represents the 95% confidence intervals based on bootstrapping of each enrichment statistic for unweighted and probabilistic graphlets depending on the number of clusters k, across 10 repetitions. High, medium, and low indicate the confidence threshold for the underlying networks.

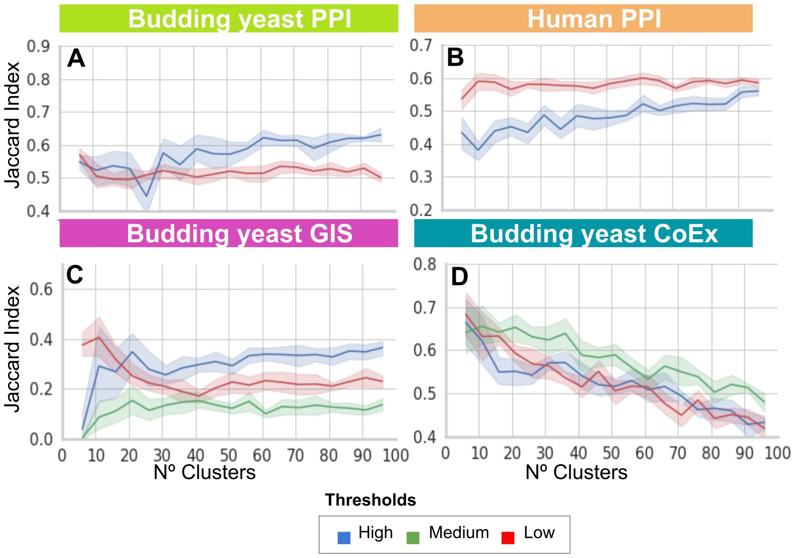


Figure A.11. Probabilistic graphlets capture distinct information compared to their unweighted counterparts. Jaccard index (JI) of the MF GO terms captured by probabilistic and unweighted networks across different confidence thresholds and for different numbers of clusters. Lines represent the mean and the shaded area represents the 95% confidence intervals based on bootstrapping for ten repetitions.

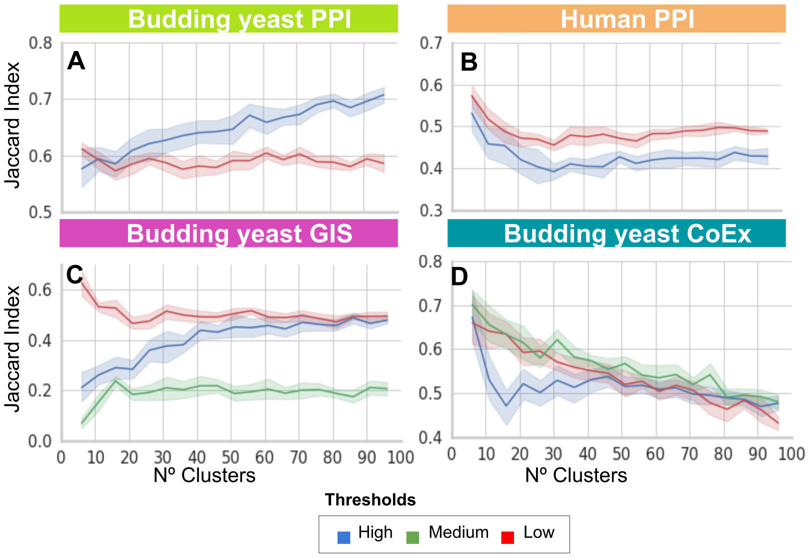


Figure A.12. Probabilistic graphlets capture distinct information compared to their unweighted counterparts. Jaccard index (JI) of the CC GO terms captured by probabilistic and unweighted networks across different confidence thresholds and for different numbers of clusters. Lines represent the mean and the shaded area represents the 95% confidence intervals based on bootstrapping for ten repetitions.

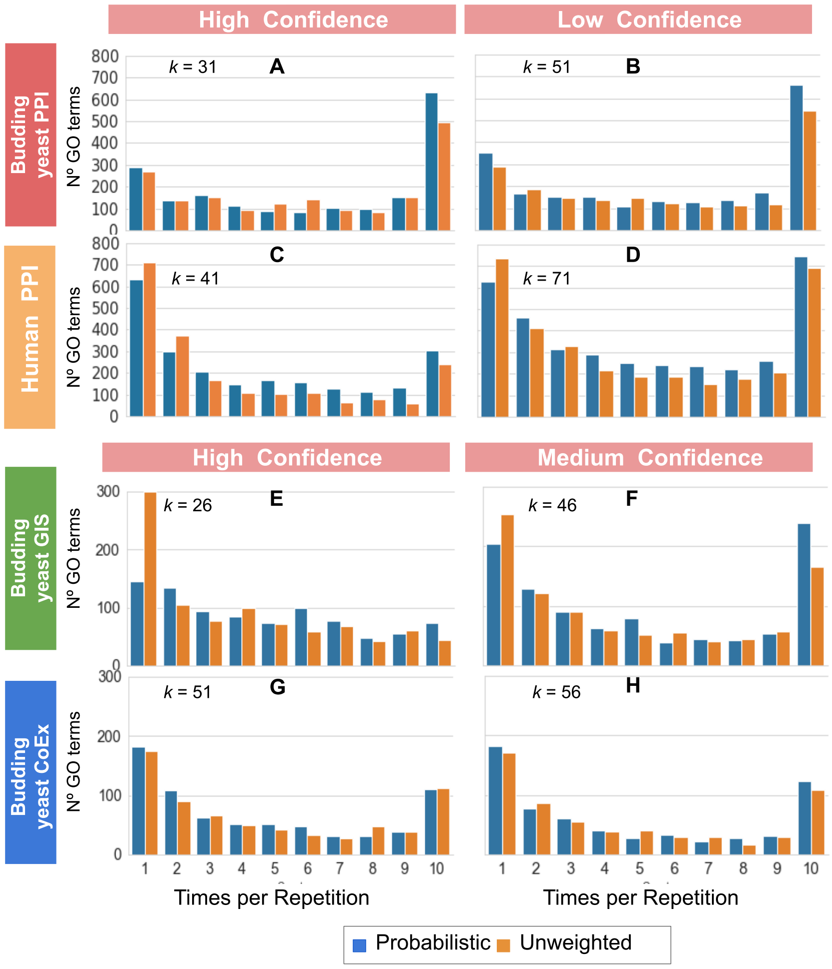


Figure A.13. Probabilistic graphlets are more robust in capturing BP GO terms than unweighted graphlets. The panels show the total count of GO terms that appeared enriched in one to ten repetitions. The value of k was fixed for each network as indicated. The value of k was fixed for each network based on the rule-of-thumb ($k \approx \sqrt{\frac{N}{2}}$, where N represents the number of nodes in the network) and is displayed in each panel.

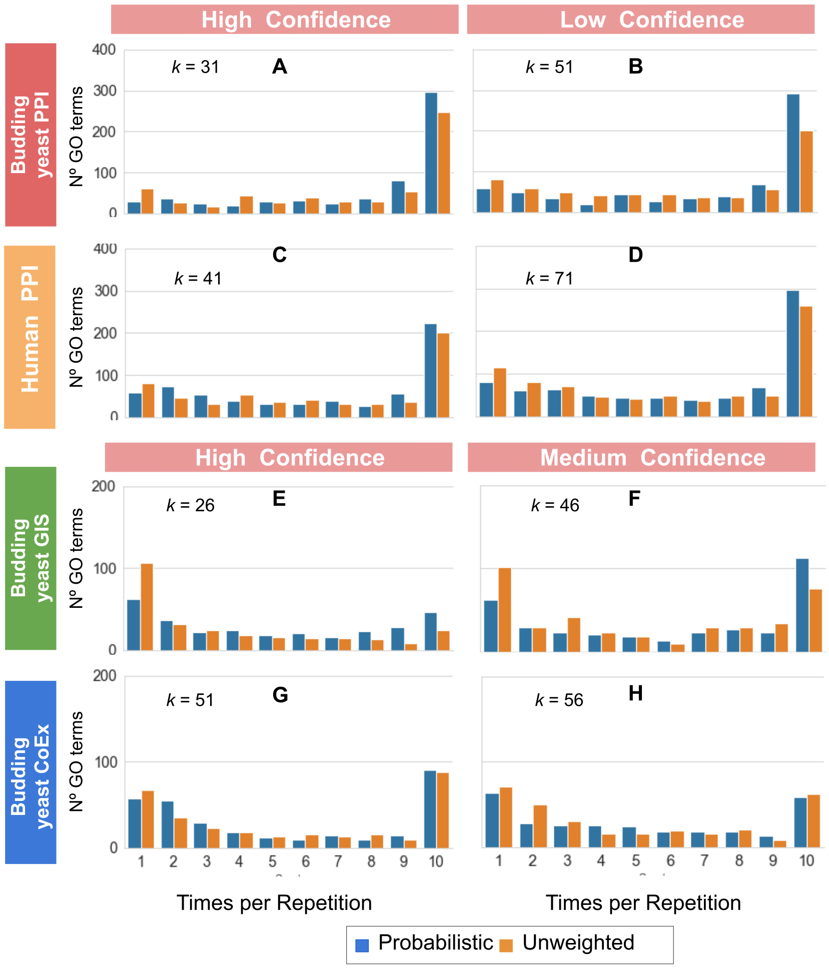


Figure A.14. Probabilistic graphlets are more robust in capturing CC GO terms than unweighted graphlets. The panels show the total count of GO terms that appeared enriched in one to ten repetitions. The value of k was fixed for each network as indicated. The value of k was fixed for each network based on the rule-of-thumb ($k \approx \sqrt{\frac{N}{2}}$, where N represents the number of nodes in the network) and is displayed in each panel.

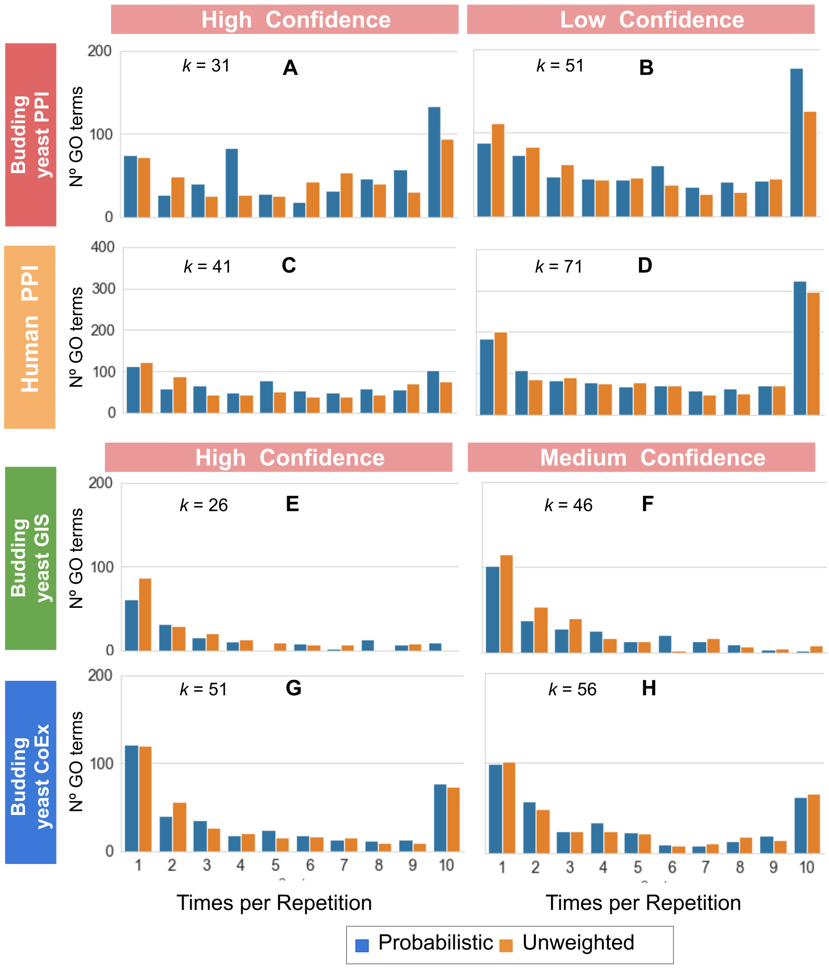


Figure A.15. Probabilistic graphlets are more robust in capturing MF GO terms than unweighted graphlets. The panels show the total count of GO terms that appeared enriched in one to ten repetitions. The value of k was fixed for each network as indicated. The value of k was fixed for each network based on the rule-of-thumb ($k \approx \sqrt{\frac{N}{2}}$, where N represents the number of nodes in the network) and is displayed in each panel.

Appendix B

Supplementary Information for Chapter 4

B.1 Supplementary Materials and Methods for Chapter 4

B.1.1 Multiplicative update rules

As presented in section 4.2.3 of Chapter 4, the Non-negative Matrix Tri-Factorization, NMTF, can be formulated as the following minimization problem:

$$\min_{P,S,G\geq 0} f(P,S,G) = min_{P,S,G\geq 0} \|X - PSG^T\|_F^2, G^TG = I,$$

where F denotes the Frobenius norm, X is the PPMI matrix representation of a molecular network (whose nodes are genes), rows in matrix $P \cdot S$ are the embedding vectors of the genes, and columns in G^T are the axis of the basis describing the space in which the genes are embedded.

Following the semi-NMTF simplification [259] for a more computationally tractable solution, we remove the non-negativity constraint on $S \geq 0$. To solve the optimization problem, we derive the Karush-Kuhn-Tucker (KKT) conditions for our NMTF as follows:

$$\begin{aligned}
\frac{\partial f}{\partial G} = -X^TPS + GS^TP^TPS - \eta_1 = 0, \\
\frac{\partial f}{\partial S} = -P^TXG + P^TPSG^TG = 0, \\
\frac{\partial f}{\partial P} = -XGS^T + PSG^TGS^T - \eta_2, \\
\eta_1, G \geq 0, \\
\eta_1 \odot G = 0, \\
\eta_2, P \geq 0, \\
\eta_2 \odot P = 0,
\end{aligned}$$

where $\odot$ is the Hadamard (element wise) product and matrices η_1 and η_2 are the dual variables for the primal constraint $G, P \geq 0$. For S, we have the following closed formula:

$$S = (P^T P)^{-1}(P^T M G)(G^T G)^{-1}$$

As explained in [26], we derive the following multiplicative update rule to solve the KKT conditions above:

$$G_{ij} \leftarrow G_{ij}\sqrt{\frac{(X^T PS)^+_{ij} + G(S^T P^T PS)^-_{ij}}{(X^T PS)^-_{ij} + G(S^T P^T PS)^+_{ij}}}$$

$$P_{ij} \leftarrow P_{ij}\sqrt{\frac{(XGS^T)^+_{ij} + P(SG^T GS^T)^-_{ij}}{(XGS^T)^-_{ij} + P(SG^T GS^T)^+_{ij}}}.$$

We start from initial solutions, G_{init}, P_{init}, S_{init}, and iteratively use Equations (1) and (2) to compute new matrix factors G, P and S until convergence. To generate initial G_{init}, P_{init} and S_{init}, we use the Singular Value Decomposition based strategy [201]. However, SVD matrix factors can contain negative entries; thus, we use only their positive entries and replace the negative entries with 0, to account for the non-negativity constraint of the NMTF. This strategy makes the solver deterministic and also reduces the number of iterations that are needed to achieve convergence [201].

We measure the quality of the factorization by the sum of the relative square errors (RSE) between the decomposed matrices and the corresponding decompositions:

$$RSE = \frac{||X - PSG^T||^2_F}{||X||^2_F}.$$

In our implementation, the iterative solver stops after 1000 iterations, the value for which the RSE of the decomposition is not decreasing any more.

B.2 Supplementary Results and Discussion for Chapter 4

B.2.1 Impact of the PPI network matrix representation to the functional organization of the embedding space

In this section, we compare the ability of the adjacency and PPMI matrix representations of the tissues-specific PPI networks (detailed in sections 4.2.1 of Chapter 4) to produce functionally coherent network embedding spaces. To this aim, we embed each tissue-specific PPI network by applying our NMTF-based methodology (see section 4.2.3 of Chapter 4) on either its adjacency matrix representation or on its PPMI matrix representation. We generate these embedding spaces with 200 dimensions since this dimensionality corresponds to the optimal dimensionality of such spaces (as detailed in section 4.2.5 of Chapter 4).

In a first step, as standardly done in the literature, we compare the ability of the adjacency and PPMI matrix representations to produce functionally coherent embedding spaces from the gene-centric point of view. For each embedding space, we cluster together genes that are embedded close in space by applying the k-medoid algorithm [206] on the genes' embedding vectors. For the number of clusters, we use the heuristic rule of thumb ($k = \sqrt{\frac{n}{2}}$, where n is the number of nodes in the tissue-specific network) [18]. We end up with 65, 45, 44, 44, 42, 38, 47, and 47 clusters for breast cancer, breast glandular cells, prostate cancer, prostate glandular cells, lung cancer, lung pneumocytes, colorectal cancer, and colorectal glandular cells, respectively. After clustering, we measure the enrichment of those clusters in GO BP annotations by using the sampling without replacement strategy (hypergeometric test) and we consider a GO BP term to be significantly enriched in a gene cluster if the corresponding enrichment p-value, after Benjamini Hochberg correction for multiple hypothesis testing [179], is smaller than or equal to 5%. For each embedding space, we report the percentage of enriched clusters (clusters with at least one enriched GO BP term), the percentage of enriched genes (genes that are annotated with at least one GO BP term that is enriched in their clusters), and the percentage of enriched GO BP terms. As detailed in Supplementary Table B.2, we find that the embedding spaces obtained from the PPMI matrix representations are functionally more coherent, with 74.80% of enriched clusters, 22.87% of enriched genes and 51.10% of enriched GO BP terms (on average over the eight tissues-specific PPI networks), compared to the embedding spaces that are obtained from the adjacency matrix representations (with 71.33% of enriched clusters, 16.23% of enriched genes and 37.56% of enriched GO BP terms on average).

In a second step, we compare the ability of the adjacency and PPMI matrix representations to produce functionally coherent network embedding spaces from our new function-centric point of view. To this aim, we use our FMM-based method to embed and capture the relative positions of the GO BP terms in the eight tissues-specific PPI network embedding spaces described above (detailed in section 4.2.4 of Chapter 4). We evaluate the functional organization of these embedding spaces by assessing if functionally similar GO BP terms (with high Lin's semantic similarity) are located close in the embedding space, and thus have low values in the corresponding FMM. To this aim, we first compute the pairwise Lin's semantic similarity [136] between any two GO BP terms. Then, we cluster GO BP terms based on their proximity in the embedding space (detailed in section 4.2.6 of Chapter 4) and report both the average semantic similarity of the pairs of GO BP terms that are in the same cluster ("intra-SS") and the average semantic similarity of the pairs of GO BP terms that are not clustered together ("inter-SS"). Intuitively, the higher the intra-SS and the lower the inter-SS, the better functionally organized the embedding space is. As detailed in Table 4.1 and Supplementary Table B.3, we find that the embedding spaces obtained from the PPMI matrix representations are more functionally coherent, with an intra-SS of 0.21 and an inter-SS of 0.161 (on average over the eight tissues-specific PPI networks), compared to the embedding spaces obtained from the adjacency matrix representations (with an intra-SS of 0.18 and an inter-SS of 0.165, on average).

Furthermore, for each tissues-specific PPI network, the pairs of GO BP terms that are clustered together in the PPMI-based network embedding spaces have statistically significantly higher Lin's semantic similarity than the pairs of GO BP terms that are clustered together in the adjacency-based network embedding spaces (with all one-sided Mann-Whitney U test p-values being smaller than or equal to 4×10^{-3}, as detailed in Supplementary Table B.3).

To conclude, both the gene-centric and our FMM approach show that the embedding spaces obtained from the PPMI matrix representations of our tissues-specific

PPI networks better capture the cell's functional organization than the embedding spaces obtained from the adjacency matrix representations of these networks. These results further demonstrate that the PPMI matrix is not only a richer representation compared to the adjacency matrix [89], but also that the extra information that it contains is useful for producing a more functionally organized embedding space.

B.2.2 Our FMM-based methodology captures more biological information from the embedding space compared to the actual gene-centric approaches

In this section, we compare the ability of our FMM-based method to uncover functional interactions between GO BP terms from the PPI network embedding spaces to that of the standard gene-centric approach. To this aim, we consider the eight cancer and control tissues-specific PPI networks described in section 4.2.1 of Chapter 4, which we embed by applying our NMTF-based methodology on their PPMI matrix representations (see section 4.2.3 of Chapter 4). We generate these embedding spaces with 200 dimensions since this dimensionality corresponds to the optimal dimensionality of such spaces (as detailed in section 4.2.5 of Chapter 4).

For a given tissues-specific PPI network embedding space, our FMM directly quantifies all the functional interactions between any two GO BP terms that annotate genes in the PPI network by measuring the cosine distance between the GO BP terms' embedding vectors (see section 4.2.4 of Chapter 4). On the other hand, the gene-centric approach does not directly uncover such functional interactions between GO BP terms. Instead, we indirectly uncover them by performing the following gene clustering and enrichment analysis. For each embedding space, we cluster together genes that are embedded close in space by applying the k-medoid algorithm [206] on the genes' embedding vectors. For the number of clusters, we use the heuristic rule of thumb ($k = \sqrt{\frac{n}{2}}$, where n is the number of nodes in the tissue-specific network) [18]. We end up with 65, 45, 44, 44, 42, 38, 47, and 47 clusters for breast cancer, breast glandular cells, prostate cancer, prostate glandular cells, lung cancer, lung pneumocytes, colorectal cancer and colorectal glandular cells, respectively. Then, we measure the enrichment of the resulting gene clusters in GO BP terms by using the sampling without replacement strategy (hypergeometric test) and we consider a GO BP term to be significantly enriched in a gene cluster if the corresponding enrichment p-value, after Benjamini and Hochberg correction for multiple hypothesis testing [179], is smaller than or equal to 5%. Then, we consider that two GO BP terms functionally interact if they are both significantly enriched in the same gene cluster. Finally, for the GO BP terms that are significantly enriched in at least one gene cluster, we measure the agreement between the functional interactions uncovered by the gene-centric approach and the functional interactions that are captured by our FMM methodology by using the following receiver operating characteristic (ROC) curve analysis. In particular, for each GO BP pair, we consider the result of the gene-centric approach as the ground truth, i.e., a pair of GO BP terms is considered as "true" if the two terms are enriched in the same cluster, or as "false" otherwise. Also, for each GO BP pair, we consider as the prediction score their cosine similarity in the embedding space (1 minus their associated value in the FMM). Then, we compute the area under the ROC curve (AUROC) [260] between the ground truth and the prediction score over all the considered GO BP pairs. Note that an AUROC score of 0.5 corresponds to a random classification and a score of 1 to a perfect one. Hence, the closer to one the AUROC score is, the higher the agreement between our FMM-based method and the gene-centric approach.

On average over our eight tissues-specific PPI networks, we find that only 51.1% of the GO BP terms that annotate genes in a network are found to be significantly enriched in at least one gene cluster, leaving about one-half of the functional space unexplored (see Supplementary Table B.2). For the significantly enriched GO BP terms, the functional interactions uncovered by the gene-centric and the FMM approaches are in significant agreement, with an average AUROC of 88% and all p-values $\leq 1 \times 10^{-323}$ (see Supplementary Figures B.6 and B.7). These results confirm that the GO BP terms that are enriched in the same gene cluster tend to be located close in the embedding space and thus, tend to have small association values in the FMM.

In conclusion, our FMM-based method is not only able to uncover the functional organization of biological functions that are identified by the gene-centric approach, but it goes beyond and characterizes the functional organization of all available GO BP terms.

B.2.3 The FMMs reveal the higher-order functional organizations of the GO BP terms in the network embedding spaces

In the previous section, we showed that our FMM better capture the pairwise functional interactions between GO BP terms than the traditional gene-centric approach. Here, we ask if the FMM can uncover the higher-order functional organization of the GO BP terms in a network embedding space. To this aim, we embed all tissue-specific PPI networks by applying our NMTF-based methodology on the PPMI matrix representations of the networks (detailed in sections 4.2.1 and 4.2.3 of Chapter 4). We generate these embedding spaces with 200 dimensions since this dimensionality corresponds to the optimal dimensionality of such spaces (as detailed in section 4.2.5 of Chapter 4). Then, we apply our FMM-based method to embed and capture the relative positions of the GO BP terms in the resulting network embedding spaces (detailed in section 4.2.4 of Chapter 4). To reveal the higher-order functional organization of the GO BP terms in the network embedding spaces, we apply the hierarchical clustering method Pvclust [261] to the rows and columns (representing GO BP terms) of the FMMs. Pvclust evaluates the statistical significance of each cluster in the hierarchy by computing its Approximately Unbiased p-value (AU) [261]. Clusters with an AU value greater than or equal to 95% are considered to be strongly supported by the data, i.e., they are not expected by random.

On average over our eight tissues-specific PPI network embedding spaces, we find that about 53.62% of the clusters in the hierarchies are statistically significant with AUs greater than or equal to 95%. In detail, we find that 54%, 54%, 55%, 52%, 53%, 54%, 53%, and 54% of the clusters in the hierarchy are statistically significant with AUs greater than or equal to 95% for breast cancer, breast glandular cells, prostate cancer, prostate glandular cells, lung cancer, lung pneumocytes, colorectal cancer and colorectal glandular cells tissue-specific PPI embedding space, respectively. Importantly, these significant clusters cover all the GO BP terms that annotate the tissues-specific PPI networks. Furthermore, by reordering the rows and columns of the FMMs according to their corresponding hierarchical clusterings, we observe evident hierarchical organizations of the GO BP embedding vectors in the different network embedding spaces (see Supplementary Figures B.8 and B.9)

In conclusion, these results demonstrate that our FMM methodology captures the higher-order organization of the GO BP terms in the network embedding space. While these results motivate us to compare FMMs across different conditions to uncover condition-related changes in the functional organization of GO BP terms in the network embedding spaces, the extraction of novel knowledge from the hierarchical organization of the GO BP terms is a subject of future study.

B.2.4 FMM discriminates between functionally and not functionally organized embedding spaces

In section 4.3.1 of Chapter 4, we use our novel FMM-based method to confirm that the embedding spaces of both, cancer and control, are functionally organized. Here, we compare these results against a randomized experiment, i.e., when rewiring the previous PPI networks. In particular, for each tissue-specific PPI network, we randomly rewire the corresponding adjacency matrix and compute its corresponding PPMI matrix (detailed in section 4.2.1, of Chapter 4). We follow the same protocol as used for the real tissue-specific networks to generate the corresponding "random" embedding space (detailed in section 4.2.3, of Chapter 4). Next, we apply our FMM-based methodology to obtain the embedding vectors of each of the GO BP annotations and the mutual positions of these vectors, which we call "distances", in the "random" embedding spaces (detailed in section 4.2.4, of Chapter 4). We evaluate the functional organization of these "random" embedding spaces by using the same clustering method as we use with the real PPI networks (detailed in section 4.2.6, of Chapter 4). For each tissue-specific PPI network, we repeat this procedure 100 times. In each repetition, we statistically test if those annotations whose embedding vectors cluster together based on their mutual positions in the space, have a statistically significant higher Lin's semantic similarity than those annotations whose embedding vectors do not cluster. For this test, we use the Mann-Whitney U test (keeping the corresponding p-value in each repetition). After all the repetitions are finished, we correct the p-values for multiple tests by using the Bonferroni correction [198]. As expected, we do not find a statistically significant difference in the Lin's semantic similarity between the annotation whose embedding vectors cluster and the annotations whose embedding vectors do not cluster in the space. Hence, we conclude that the "random" embedding spaces are not functionally organized (see Supplementary Table B.6). These results demonstrate that our methodology correctly discriminates between functionally and not functionally organized embedding spaces.

B.2.5 FMMs identify novel cancer-related functions

In section 4.3.2, of Chapter 4, we use our novel FMM-based methodology to predict new cancer-related functions and we verify the importance of one of our cancer-related predictions (the first annotation in our top 10 annotations predicted to be cancer-related, that we could not validate in the currently available literature). In this section, we extend this discussion to the remaining top 10 predicted cancer-related annotations. Starting with breast cancer, first, we discuss the viral translational termination reinitiation. This function could be connected with the alternative transcriptional regulation pathways described in cancer [262]. In the same cancer, we also find as predicted to be cancer-related the RNA phosphodiester bond hydrolysis, endonucleolytic. This function could be connected with the regulatory roles of RNA modifications reported in this cancer type [263]. Following with prostate cancer, we find the positive regulation of endoplasmic reticulum unfolded protein response. The accumulation of unfolded protein in the ER induces this unfolded protein response as our predicted cancer-related function. It has been shown that the upregulation of this response could provide a growth advantage to tumor cells [264]. Regarding lung cancer, we find the viral translational termination reinitiation as predicted cancer-related function. As discussed for breast cancer (see section 4.3.2 of Chapter 4), this process could also be connected with the alternative transcriptional regulation pathways described in cancer [262]. In lung cancer, we also find the positive regulation of transcription regulatory region DNA binding as predicted cancer-related function. These processes could be connected with the well-known deregulation of the gene expression observed in different cancers [265].

In conclusion, we demonstrate that our predicted cancer-related functions are indeed cancer-related. Thus, our novel FMM-based methodology can be used to identify new cancer-related functions.

B.3 Supplementary Figures for Chapter 4

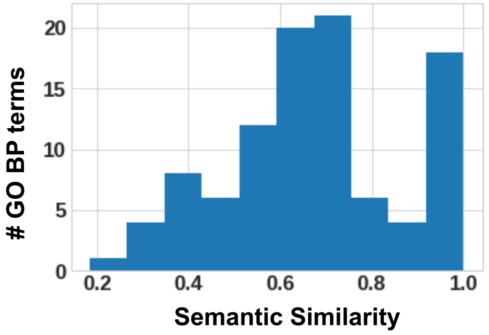


Figure B.1. Lin's semantic similarity between our set of cancer-related GO BP terms (104 annotations) and the set of GO BP terms classified as the set of GO BP cancer hallmark defined by [199] (135 annotations). For each GO BP term in our set, we show its maximum Lin's semantic similarity to one annotation in the cancer hallmarks set.

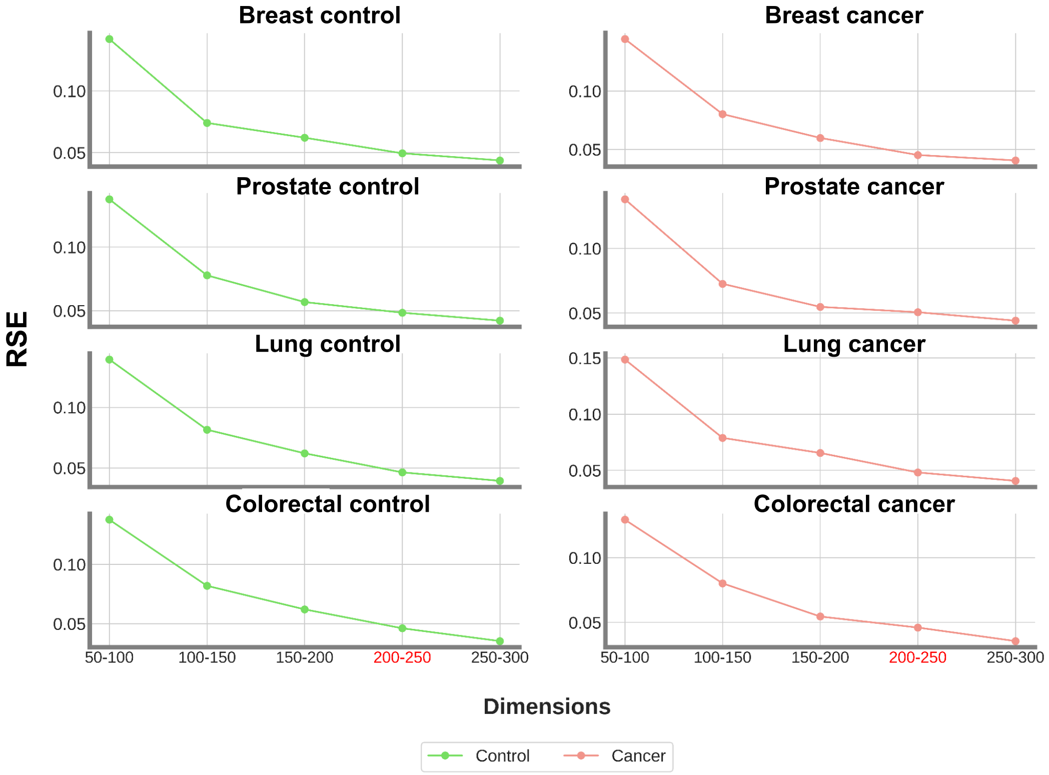


Figure B.2. For each cancer type (breast cancer, prostate cancer, lung cancer, and colorectal cancer) and its corresponding control. Each panel shows the Relative Square Error (RSE) of FMMs corresponding to the cancer and control tissues-specific embedding spaces of increasing dimensions (dimension increasing by 50 starting from 50 and ending with 300).

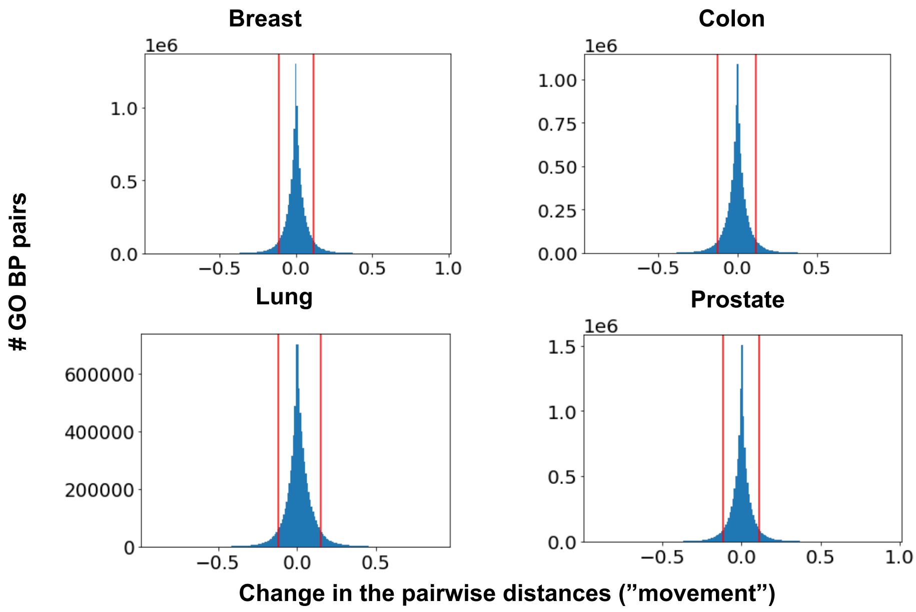


Figure B.3. Change in the pairwise distances (cosine distances), that we call "movement", of the functional annotation embedding vectors between breast, cancer, and control embedding spaces. For a pair of annotation embedding vectors, its "movement" is the difference between the cosine distance between the two embedding vectors in one embedding space (control) and the corresponding cosine distance in the other space (cancer) (defined in section 4.2.7, of Chapter 4). Thus, positive "movement" means that the two annotation embedding vectors got closer in the cancer embedding space, and negative "movement" means that the two annotation embedding vectors got further apart in the cancer embedding space. The red lines represent the 95^{th} and 5^{th} percentiles of the distributions. We use these thresholds to define when two annotation embedding vectors are "moving significantly apart" in the embedding space of cancer (95^{th} percentile) or are "moving significantly closer" in the embedding space of cancer (5^{th} percentile). The panels are for breast, lung, colorectal, and prostate cancers versus controls.

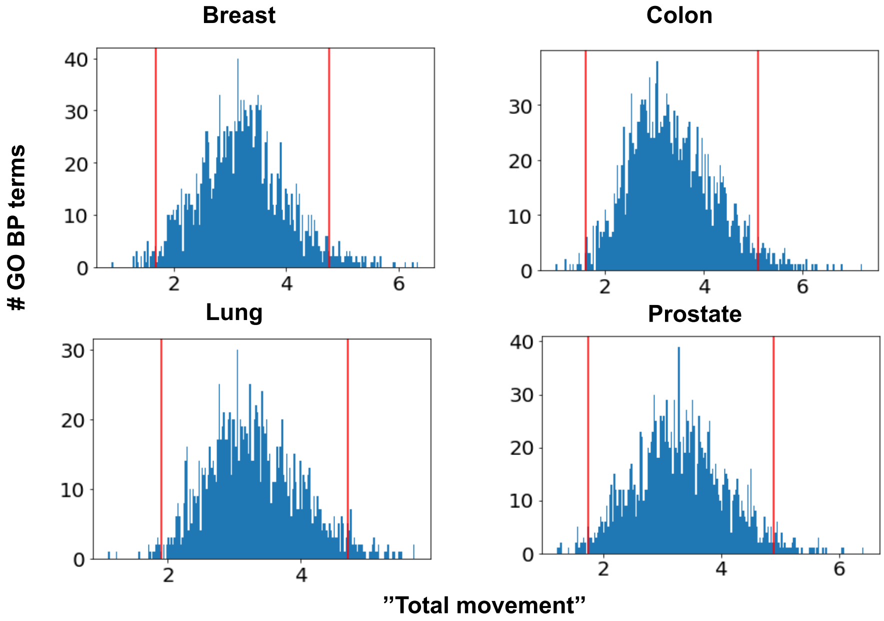


Figure B.4. "Total movement distribution" of the functional annotation embedding vectors. For each annotation embedding vector, we compute its "total movement" (defined in section 4.2.7, of Chapter 4). Thus, those annotation embedding vectors that change their mutual positions, "movement", the most between control embedding space and cancer embedding space have higher "total movement" than those annotation embedding vectors that do not change their "movement". The red lines represent two standard deviations above and below the mean of the distribution. We use these thresholds to define as *shifted biological functions* those functional annotations whose embedding vectors' "total movement" is two standard deviations above the mean of the "total movement distribution." In contrast, we define as *stable biological functions* those functional annotations whose embedding vectors' "total movement" is two standard deviations below the mean of the "total movement" distribution. The distributions are for breast, lung, colorectal, and prostate cancers.

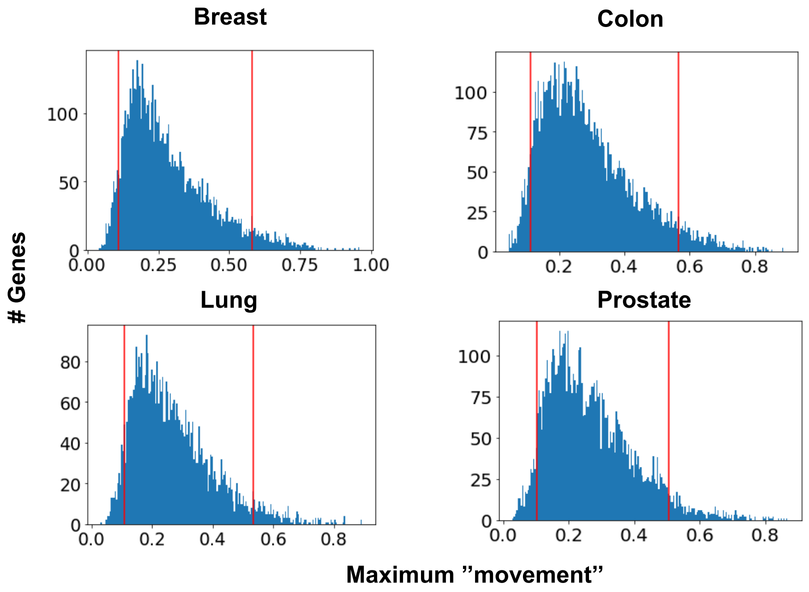


Figure B.5. Gene maximum "movement" distribution. For each gene, we have a vector with n positions, where n corresponds to the number of the "shifted" GO terms. Each entry of this n-dimensional vector corresponds to the "movement" (change of mutual positional) of the gene and the GO term. This "movement" can either be positive (a gene is going closer to the GO term in the cancer space), or negative (a gene is going further from the GO term in the cancer space). Since this "movement" is bi-directional (getting closer or further), we use the absolute value of the "movement" at each coordinate of this vector, to keep only the magnitude of this movement independently of the direction of the "movement". Then, since all the values in the n-dimensional vector are now positive, for each gene we assign as its cancer-related score the maximum value (maximum magnitude of movement) in its corresponding vector. The red lines represent the 95^{th} and 5^{th} percentiles of the distributions. Based on these thresholds, we consider cancer-related gene predictions whose genes that are above the 95^{th} percentile of the maximum "movement" distribution. The distributions are for breast, lung, colorectal, and prostate cancers.

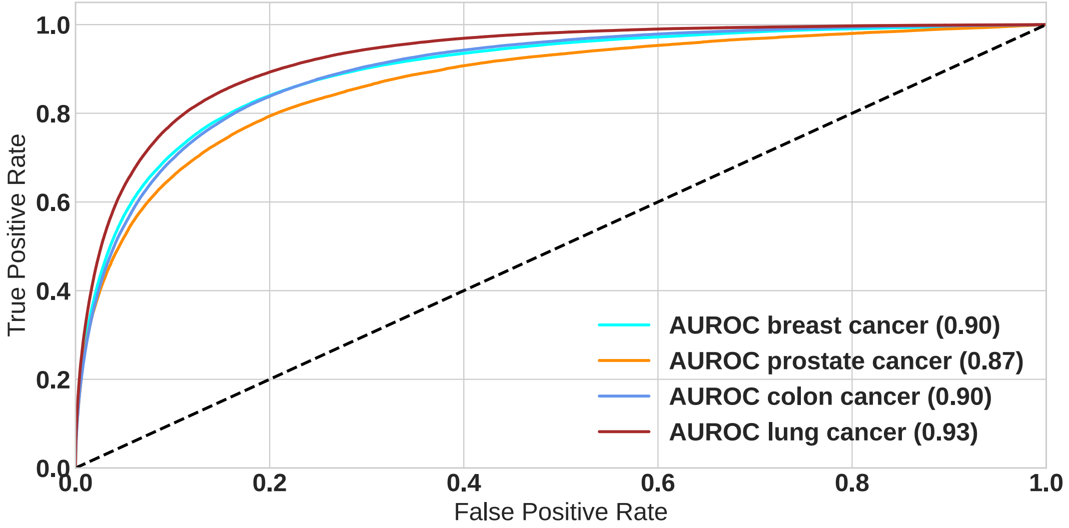


Figure B.6. Our FMM-based method uncovers the functional interactions between GO BP terms that are identified by the standard gene-centric approach (based on clustering and functional enrichment analyses) in four cancer tissue-specific PPI embedding spaces (breast, lung, colorectal, and prostate). For each cancer tissue-specific PPI embedding space, we take the subset of GO BP terms that are statistically enriched based on the gene-centric approach (detailed in Supplementary section B.2.2). Then, for a pair of GO BP terms, we set the ground truth as one if they are enriched in the same cluster (zero otherwise). For the same pair, we set the prediction score as the value of their embedding vectors' cosine distance in the embedding space, as captured by the FMM. Finally, we compute the area under the receiver operating characteristic curve (AUROC) [260] between the ground truth and the prediction score. Each panel shows the corresponding ROC curves with its AUROC.

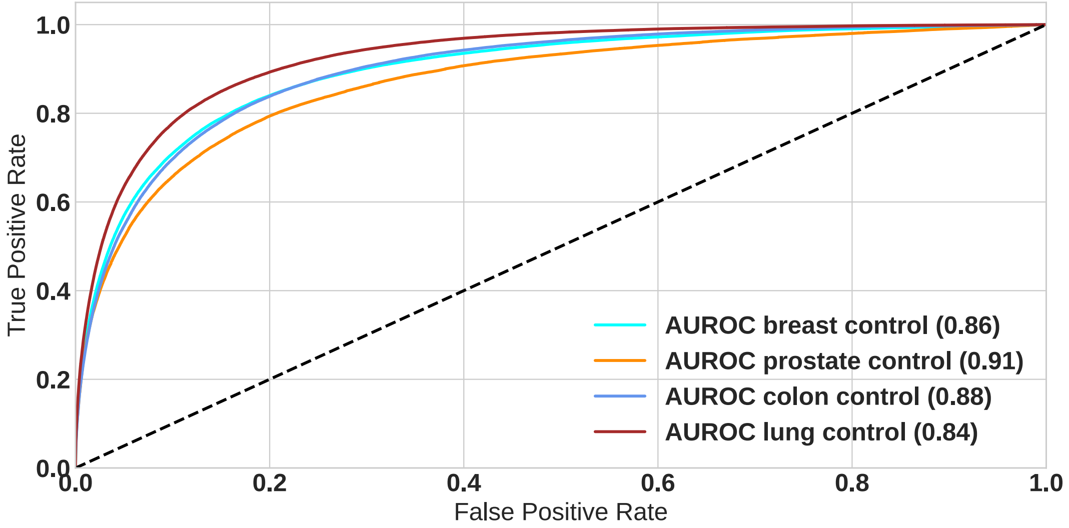


Figure B.7. Our FMM-based method uncovers the functional interactions between GO BP terms that are identified by the standard gene-centric approach (based on clustering and functional enrichment analyses) in the control tissue-specific PPI embedding spaces of four cancer types (breast, lung, colorectal, and prostate). For each control tissue-specific PPI embedding space, we take the subset of GO BP terms that are statistically enriched based on the gene-centric approach (detailed in Supplementary section B.2.2). Then, for a pair of GO BP terms, we set the ground truth as one if they are enriched in the same cluster (zero otherwise). For the same pair, we set the prediction score as the value of their embedding vectors' cosine distance in the embedding space, as captured by the FMM. Finally, we compute the area under the receiver operating characteristic curve (AUROC) [260] between the ground truth and the prediction score. Each panel shows the corresponding ROC curves with its AUROC.

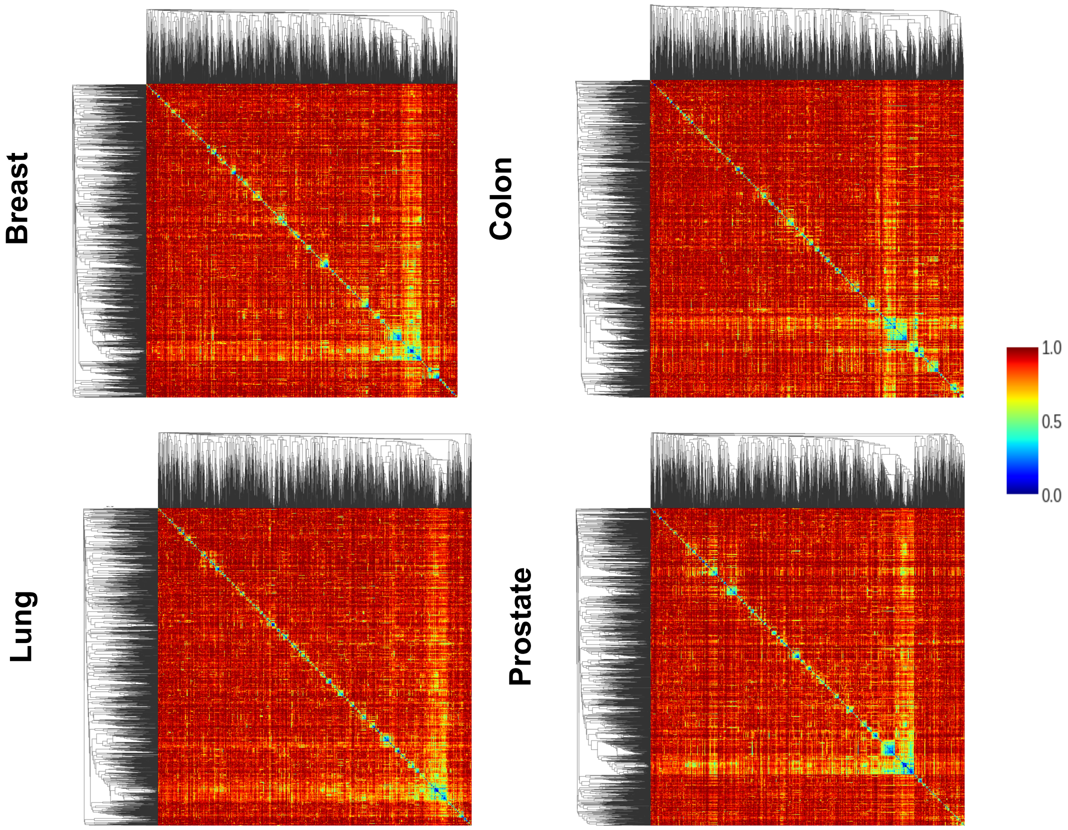


Figure B.8. Heatmaps of the FMMs of breast, lung, colorectal and prostate cancer tissues-specific PPI embedding spaces. For each FMM, we reorder it based on the hierarchical clustering obtained by Pvclust (detailed in Supplementary section B.2.3). For completeness, we plot on the left and the top of each FMM heatmap the dendrogram tree of the corresponding hierarchical clustering.

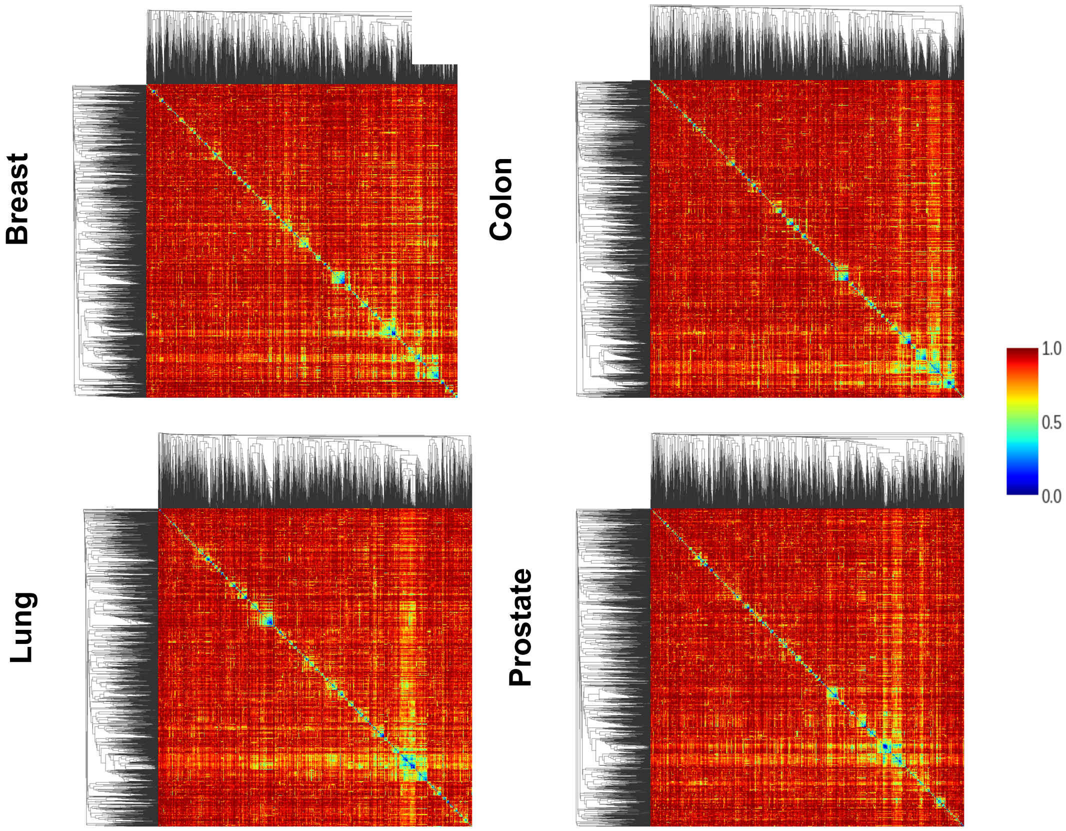


Figure B.9. Heatmaps of the FMMs of the control tissue-specific PPI embedding spaces of four cancer types (breast, lung, colorectal, and prostate). For each FMM, we reorder it based on the hierarchical clustering obtained by Pvclust (detailed in Supplementary section B.2.3). For completeness, we plot on the left and the top of each FMM heatmap the dendrogram tree of the corresponding hierarchical clustering.

B.4 Supplementary Tables for Chapter 4

Network	#Nodes	#Edges	#Density
Breast cancer	8,498	163,893	0.45
Breast control	7,999	160,520	0.50
Prostate cancer	7,885	137,701	0.44
Prostate control	7,837	148,797	0.48
Lung cancer	7,031	126,744	0.51
Lung control	5,912	95,774	0.54
Colorectal cancer	8,941	175,081	0.43
Colorectal control	8,974	185,342	0.46

Table B.1. The statistics for the tissue-specific PPI networks in this study. Column "Network" presents the tissue-specific PPI network that we analyzed column; column, "# Nodes," presents the number of nodes in the PPI network; column, "# Edges," presents the number of edges between the nodes; column, "#Density," presents the edge density of the corresponding PPI network.

Matrix	Data set	%Clusters	%Genes	%GO
PPMI	breast cancer	81.00	23.12	52.44
PPMI	breast control	68.25	22.69	51.73
PPMI	prostate cancer	76.19	23.28	49.36
PPMI	prostate control	80.95	25.37	52.28
PPMI	lung cancer	73.13	25.28	53.01
PPMI	lung control	79.10	24.33	55.97
PPMI	colorectal cancer	77.97	22.05	49.2
PPMI	colorectal control	62.96	16.89	44.86
Adj	breast cancer	70.77	17.87	36.24
Adj	breast control	76.19	18.07	41.84
Adj	prostate cancer	77.78	14.25	40.83
Adj	prostate control	77.19	16.96	38.33
Adj	lung cancer	74.62	20.89	38.76
Adj	lung control	79.10	17.89	41.70
Adj	colorectal cancer	57.63	13.08	30.96
Adj	colorectal control	57.41	10.88	31.87

Table B.2. The embedding spaces of the most prevalent cancers (breast, prostate, lung, and colorectal cancer) and their control tissues (breast glandular cells, prostate glandular cells, lung pneumocytes, and colorectal glandular cells) are functionally organized according to the mutual positions (cosine distances) of the gene embedding vectors in the embedding space (gene perspective). For each tissue-specific PPI embedding space, we cluster genes whose embedding vectors are close in the space based on their cosine distance, and then we measure the enrichment of those clusters in GO BP annotations. The first column, "Matrix," indicates the matrix representation of the tissue-specific PPI network. The second column, "Data set," specifies the tissue-specific PPI network. The third column, "%Clusters," shows the percentage of clusters with at least one GO BP term enriched. The fourth column, "%Genes," presents the percentage of enriched genes in the clusters (out of the total number of genes in the corresponding tissue-specific PPI network). The sixth column, "%GO," shows the percentage of GO BP terms enriched in the clusters (out of the total GO BP terms that annotate the genes of the corresponding tissue-specific PPI network).

Embedding	**Intra-SS**	**Inter-SS**	**Fold**	**p-value Fold**	**p-value (PPMI)**
Control breast	0.18	0.16	1.10	0.0001	0.004
Cancer breast	0.18	0.16	1.10	0.0004	1.31×10^{-8}
Control prostate	0.18	0.17	1.08	0.0074	0.0004
Cancer prostate	0.18	0.17	1.08	0.0002	8.05×10^{-38}
Control colorectal	0.18	0.16	1.11	0.0004	0.0008
Cancer colorectal	0.18	0.16	1.10	0.0004	5.00×10^{-42}
Control lung	0.18	0.17	1.06	0.0020	2.53×10^{-71}
Cancer lung	0.18	0.17	1.09	0.0020	9.73×10^{-57}

Table B.3. The adjacency embedding spaces of the most prevalent cancers (breast, prostate, lung, and colorectal cancer) and their control tissues (breast glandular cells, prostate glandular cells, lung pneumocytes, and colorectal glandular cells) are functionally organized. The first column, "Embedding," lists the tissues. The second column, "Intra-SS," shows the average Lin's semantic similarity of those annotations whose embedding vectors cluster together based on their cosine distances in the embedding space. The third column, "Inter-SS," shows the average Lin's semantic similarity of those annotations whose embedding vectors do not cluster together based on their cosine distances in the embedding space. The fourth column, "Fold," displays how many times the average Lin's semantic similarity of those annotations whose embedding vectors cluster together based on their cosine distances in the embedding space is higher than of those annotations whose embedding vectors do not cluster together. The fifth column, "p-value Fold," shows the p-value from a one-sided Mann-Whitney U test comparing Lin's semantic similarity between annotations whose embedding vectors cluster together and those with non-clustered embedding vectors. The sixth column, "p-value (PPMI)," shows the p-value from a one-sided Mann-Whitney U test comparing Lin's semantic similarity between annotations that cluster together based on their proximity in the PPMI embedding space and those annotations that cluster together based on their proximity in the corresponding adjacency embedding space.

Network	#Optimal Dimensions
Breast cancer	200
Breast control	200
Prostate cancer	200
Prostate control	200
Lung cancer	200
Lung control	200
Colorectal cancer	200
Colorectal control	200

Table B.4. Optimal number of dimensions for each tissue-specific embedding space. Column "Network," specifies the tissue-specific PPI network. Column, "# Optimal Dimensions," contains the optimal number of dimensions that we found experimentally as explained in section 4.2.5 of Chapter 4, which we then used for generating the corresponding embedding space by our NMTF-based procedure explained in Chapter 4.

Sample	Avg Distance Annotate	Avg Distance Not-Annotate
Breast cancer	0.571	0.920
Breast control	0.575	0.921
Prostate cancer	0.598	0.912
Prostate control	0.576	0.926
Colorectal cancer	0.520	0.922
Colorectal control	0.514	0.908
Lung cancer	0.578	0.920
Lung control	0.593	0.922

Table B.5. The embedding vectors of the biological functions (GO BP terms) are significantly closer in space to the embedding vectors in the same space of the genes that they annotate than to the embedding vectors of other genes. Column, “ Sample,” presents the tissues-specific PPI networks. Column, “Avg Distance Annotate,” presents the average cosine distance in the embedding space between the embedding vectors of genes and embedding vectors of those functional annotations that annotate them; column, “Avg Distance Not-Annotate,” presents the average cosine distance in the embedding space between the embedding vectors of genes and embedding vectors of those embedded functional annotations that do not annotate them. In all samples, the difference between these distances is statistically significant (p-value of the Mann-Whitney U test < 0.05).

Embedding	Intra	Inter	Fold	p-value
Random control breast	0.17	0.17	1.00	0.14
Random cancer breast	0.17	0.17	1.00	0.09
Random control prostate	0.17	0.17	1.00	0.06
Random cancer prostate	0.18	0.17	1.05	0.07
Random control colorectal	0.16	0.16	1.00	0.10
Random cancer colorectal	0.17	0.16	1.05	0.08
Random control lung	0.16	0.17	0.94	0.09
Random cancer lung	0.15	0.15	1.00	0.07

Table B.6. Our FMM-based method discriminates between functionally organized embedding spaces and those that are not. For each tissue-specific PPI network, we randomly rewire the networks and generate the random embedding space by using the NMTF algorithm. Then, we use our new FMM-based method to evaluate the functional organization of these tissue-specific PPI embedding spaces (detailed in section 4.2.6 of Chapter 4). The first column, “Embedding,” lists the randomized tissue-specific PPI embedding space. The second column, “Intra,” shows the average Lin’s semantic similarity of those annotations whose embedding vectors cluster together based on their cosine distances in the embedding space. The third column, “Inter,” shows the average Lin’s semantic similarity of those annotations whose embedding vectors do not cluster together based on their cosine distances in the embedding space. The fourth column, “Fold,” displays how many times the average Lin’s semantic similarity of those annotations whose embedding vectors cluster together based on their cosine distances in the embedding space is higher than of those annotations whose embedding vectors do not cluster together. The fifth column, “p-value,” shows the one-sided Mann-Whitney U test p-value.

	Shifted	*Stable*
Breast	58	29
Prostate	49	26
Lung	53	15
Colorectal	68	13

Table B.7. Numbers of GO BP annotations in the *shifted* and *stable* sets in each cancer type. For the four cancer types: breast cancer (denoted by "Breast"), prostate cancer (denoted by "Prostate"), lung cancer (denoted by "Lung"), and colorectal cancer (denoted by "Colorectal"). Column, "*shifted*," presents the number of annotations in the set of *shifted* functions; column, "Stable", presents the number of annotations in the set of stable functions. The details about the definitions of *shifted* and *stable* sets can be found in section 4.2.7 of Chapter 4.

Gene name	PubMed Counts	Prognostic Marker	Pan-Cancer Marker
LDHA	87	-	cervical cancer (unfavorable), liver cancer (unfavorable), lung cancer (unfavorable)
COPG1	1	-	liver cancer (unfavorable)
RPL11	10	yes	breast cancer (favorable), renal cancer (unfavorable)
STK36	0	-	liver cancer (unfavorable)
CD86	94	-	renal cancer (unfavorable)
SMURF1	15	-	-
VRK3	0	-	renal cancer (favorable), urothelial cancer (favorable)
MAPK8IP1	2	-	renal cancer (favorable)
RPL17	1	-	liver cancer (unfavorable)
PIAS4	10	-	endometrial cancer (favorable), pancreatic cancer (favorable)

Table B.8. Top 10 *shifted* genes (the most *shifted* ones) in breast cancer. The first column, "Gene name," presents the gene names of the top 10 *shifted* genes. The second column, "PubMed Counts," presents the number of publications in Pubmed that relate the gene to breast cancer. The third column, "Prognostic Marker," indicates if the gene is a prognostic marker ("yes" if it is a marker, "-" otherwise) in breast cancer (based on survival curves collected from the Human Protein Atlas [192]); the fourth column, "Pan-Cancer Marker," presents whether the gene is a prognostic marker for other cancer types.

Gene name	PubMed Counts	Prognostic Marker	Pan-Cancer Marker
CPSF6	0	-	liver cancer (unfavorable), renal cancer (unfavorable)
PRDM11	0	-	-
SDHB	0	-	renal cancer (favorable)
GLRX2	1	-	renal cancer (unfavorable)
IFITM2	0	-	renal cancer (unfavorable)
C1orf116	0	-	renal cancer (favorable)
H2BC4	0	-	pancreatic cancer (unfavorable), renal cancer (unfavorable)
FUS	13	-	liver cancer (unfavorable)
DDX39B	0	-	renal cancer (unfavorable), urothelial cancer (favorable)
UMAD1	0	-	renal cancer (favorable)

Table B.9. Top 10 *shifted* genes in lung cancer. The first column, "Gene name," presents the gene names of the top 10 *shifted* genes. The second column, "PubMed Counts," presents the number of publications in Pubmed that relate the gene to lung cancer. The third column, "Prognostic Marker," presents if the gene is a prognostic marker ("yes" if it is a marker, "-" otherwise) in lung cancer (based on survival curves collected from the Human Protein Atlas [192]). The fourth column, "Pan-Cancer Marker," presents whether the gene is a prognostic marker for other cancer types.

Gene name	PubMed Counts	Prognostic Marker	Pan-Cancer Marker
H4C6	0	-	-
RPL11	1	-	breast cancer (favorable), renal cancer (unfavorable)
VRK3	0	-	renal cancer (favorable), urothelial cancer (favorable)
RPL17	0	-	liver cancer (unfavorable)
GGA3	0	-	endometrial cancer (unfavorable), liver cancer (unfavorable), renal cancer (unfavorable)
RPS4X	0	-	renal cancer (unfavorable), thyroid cancer (favorable)
C1orf116	0	-	renal cancer (favorable)
NAXE	1	-	endometrial cancer (unfavorable)
RARG	0	-	endometrial cancer (unfavorable), renal cancer (unfavorable)
FUS	1	-	liver cancer (unfavorable)

Table B.10. Top 10 *shifted* genes (the most *shifted* ones) in colorectal cancer. The first column, "Gene name," presents the gene names of the top 10 *shifted* genes. The second column, "PubMed Counts," presents the number of publications in Pubmed that relate the gene to colorectal cancer. The third column, "Prognostic Marker," presents if the gene is a prognostic marker ("yes" if it is a marker, "-" otherwise) in colorectal cancer (based on survival curves collected from the Human Protein Atlas [192]). The fourth column, "Pan-Cancer Marker," presents whether the gene is a prognostic marker for other cancer types.

Appendix C

Supplementary Information for Chapter 5

C.1 Supplementary Results and Discussion for Chapter 5

C.1.1 The Axes of the embedding space synthesize the core functions of different species cells

In section 5.3.4 of Chapter 5, we analyze the biological meaning of the ASFAs obtained from the axes of the human ONMTF embedding space. Here we extend this analysis to the rest of the species-specific embedding spaces obtained by applying ONMTF on the species-specific PPI network of *Homo sapiens sapiens* (denoted by human), *Saccharomyces cerevisiae* (denoted by budding yeast), *Schizosaccharomyces pombe* (denoted by fission yeast), *Rattus norvegicus* (denoted by rat), *Drosophila melanogaster* (denoted by fruit fly), and *Mus musculus* (denoted by mouse) (detailed in sections 5.2.1 and 5.2.2 of Chapter 5). We generate these embedding spaces with different dimensionalities (from 50 to 1000 dimensions with a step of 50). To select the optimal dimensionality of these embedding spaces, we follow the same criteria we did for the human PPI embedding spaces (detailed in section 5.3.3 of Chapter 5). This dimensionality corresponds to 200, 200, 300, 250, and 400 for Budding yeast, Fission yeast, Fruit fly, Rat, and Mouse embedding spaces, respectively. Then, we use the GO BP terms captured by each axis to generate the *Functional AxNotations* of each species (detailed in section 5.2.6 of Chapter 5), and we analyze their biological coherence by literature curation.

Similar to human, we find that all the species-specific ASFAs describe coherent functions of their corresponding species. For instance, the ASFA of axis 79 in budding yeast represents the trafficking of endosomes (see Supplementary Table C.7). Curiously, this yeast is one of the most used models to study this transport process [266]. Another example in budding yeast is the ASFA of axis 82, which is connected to regulating gene expression via mRNA degradation (see Supplementary Table C.7). Precisely with a process that involves the capping of the 7-methylguanosine residue that occurs after the deadenylation of the 3' poly(A) tracts of eukaryotic mRNAs and that serves as a backup mechanism to trigger mRNA decay if initial deadenylation is compromised [267]. Moreover, the ASFA of axis 20 in fission yeast is connected to the generation of large the ribosomal subunit necessary to synthesize proteins [268] (see Supplementary Table C.7). Another example in this yeast is the ASFA of axis 32, which is also related to the synthesis of proteins (see Supplementary Table C.7). However, in this case, the ASFA describes the regulation of protein synthesis via the rapamycin kinase complex I (TORC1) and II (TORC2). In the presence of ample nutrients, TORC1 and TORC2 activate and drive protein, lipid, and nucleotide synthesis by phosphorylating a wide range of proteins [269].

Regarding the fruit fly, we find ASFAs that represent functions that are more complex that the ones observed for the previous yeasts, such as the development of specific tissues. For instance, the ASFA of axis 1 describes the development of the visual nervous system (see Supplementary Table C.7). Briefly, this tissue appears after the differentiation of the neuroectoderm by activating different epidermal growth factor receptors, such as ERBB2 [270]. Another example in the fruit fly is the ASFAs of axis 28, which is related to the wing imaginal disc of this fly (see Supplementary Table C.7). This disc is a tissue of undifferentiated cells that are precursors of the wing and serves as a commonly used model system to study the regulation of growth [271]. Finally, we find that the ASFAs of mouse and rat are also connected to complex cellular functions, such as the immune system or the nervous system. For instance, the ASFA of axis 41 of mouse describes the production of interferon-alpha, interleukins, and cytokines, during the cellular response to a virus infection [272] (see Supplementary Table C.7). On the other hand, the ASFAs of axes 69 and 84 in rat, are connected to the synapsis of neurons and the production of steroids, respectively (see Supplementary Table C.7).

In conclusion, these results demonstrate that the ASFAs describe coherent biological functions. The complete Tables with all the sets of species-specific ASFAs can be found in the Supplementary online data.

C.1.2 The Axes of the embedding space give insights into the evolutionary story of species

In section 5.3.5 of Chapter 5, we show that the human ASFAs give insights into the evolutionary history of humans. In this section, we extend this analysis to the ASFAs of five species, *Saccharomyces cerevisiae* (denoted by budding yeast), *Schizosaccharomyces pombe* (denoted by fission yeast), *Rattus norvegicus* (denoted by rat), *Drosophila melanogaster* (denoted by fruit fly), and *Mus musculus* (denoted by mouse). To this aim, we divide the ASFAs of each species into three classes according to their conservation degree: “prokaryotes,” “eukaryotes,” and “vertebrates” (detailed section 5.2.7 of Chapter 5). Then, We analyze in detail the meaning of these groups of ASFAs in the context of evolution.

We find that 78%, 69%, 59%, 63%, and 40% of all ASFAs in budding yeast, fission yeast, fruit fly, rat, and mouse, respectively are classified as “prokaryotes.” These ASFAs present the lowest conservation degree in all the studied species, i.e., they are conserved in evolution (see Supplementary Figure C.4). We observe that they represent the most basic molecular mechanisms of the cell, such as the translational process in budding yeast, the homeostasis of proteins in fission yeast, the homeostasis of ions in the fruit fly, or the lipid metabolism in mice (see axes 77, 57, 4, and 5, respectively in Supplementary Table C.7).

On the other hand, we find that 22%, 31%, 41%, 33%, and 41% of all ASFAs in budding yeast, fission yeast, fruit fly, rat, and mouse, respectively are classified as "eukaryotes." These ASFAs have on average a lower conservation degree than the "prokaryotes" ones, i.e., they are newer in the evolutionary history. We find that they describe cellular functions that are connected to basic eukaryotic functions, e.g., with Golgi apparatus in budding yeast, signaling transduction in fission yeast, or cytoskeleton (see axes 79, 32, and 51, respectively in Supplementary Table C.7).

Finally, as expected, the only organism that has "vertebrates" ASFAs are rat and mouse. Precisely, we find that 11% and 10% of all ASFAs are classified as "vertebrates" in rats and mice, respectively. These ASFAs have on average the lowest conservation degree, i.e., are the newest in evolution and they describe complex biological functions, such as estrous cycle or odontogenesis in rats, and eyes' lens development or blastocyst development in mice (see axes 81, 19, 86, and 80, respectively in Supplementary Table C.7).

In conclusion, these results demonstrate that the ASFAs of different species can be used to give insights into their evolutionary history.

C.2 Supplementary Figures for Chapter 5

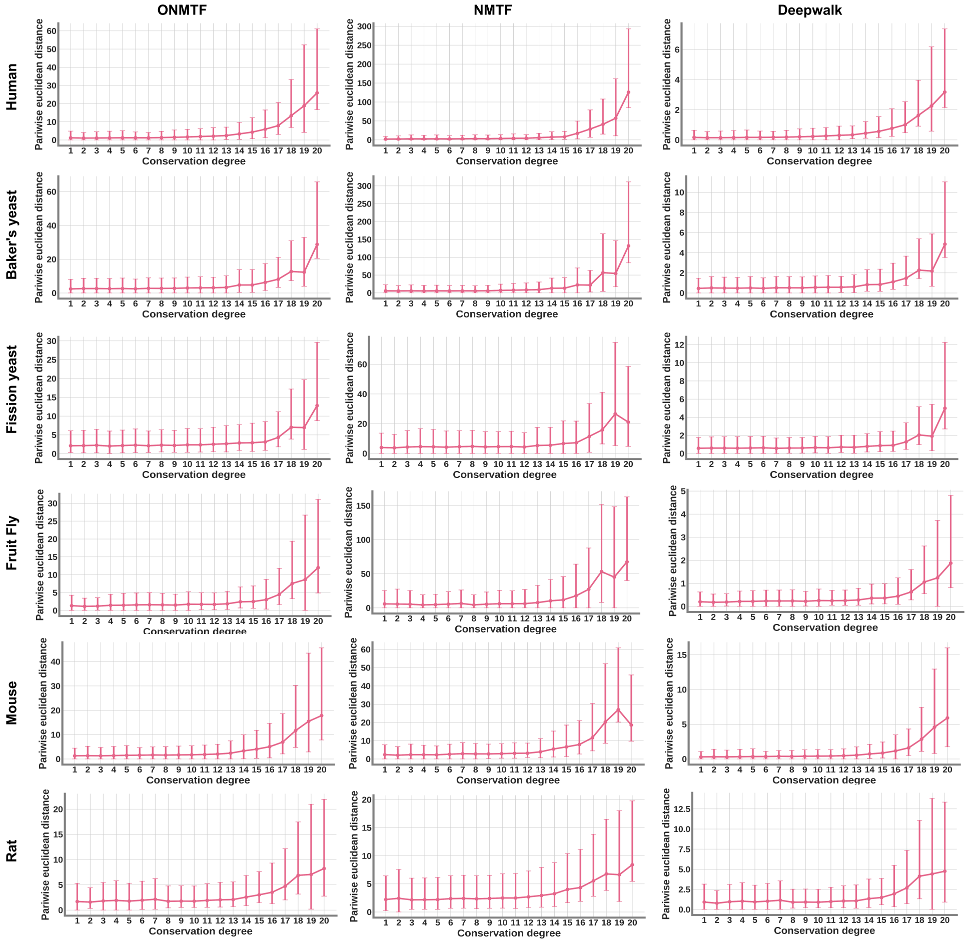


Figure C.1. The conservation degree of the GO BP terms influences the positions of their embedding vector in the species-specific PPI embedding space. We embed GO BP terms into the embedding spaces generated by applying ONMTF, NMTF, and Deepwalk algorithms on the species-specific PPI network of *Homo sapiens sapiens* (denoted by human), *Saccharomyces cerevisiae* (denoted by budding yeast), *Schizosaccharomyces pombe* (denoted by fission yeast), *Rattus norvegicus* (denoted by rat), *Drosophila melanogaster* (denoted by fruit fly), and *Mus musculus* (denoted by mouse) (detailed in sections 5.2.1 and 5.2.2 of Chapter 5). We study the correlation between the mutual positions of their embedding vectors in the space (measured by their pairwise euclidean distances) and their conservation degree (detailed in section 5.2.4 of Chapter 5). In each panel, the horizontal axis displays the conservation degree of the GO BP terms and the vertical axis shows the pairwise euclidean distance distribution of their embedding vectors.

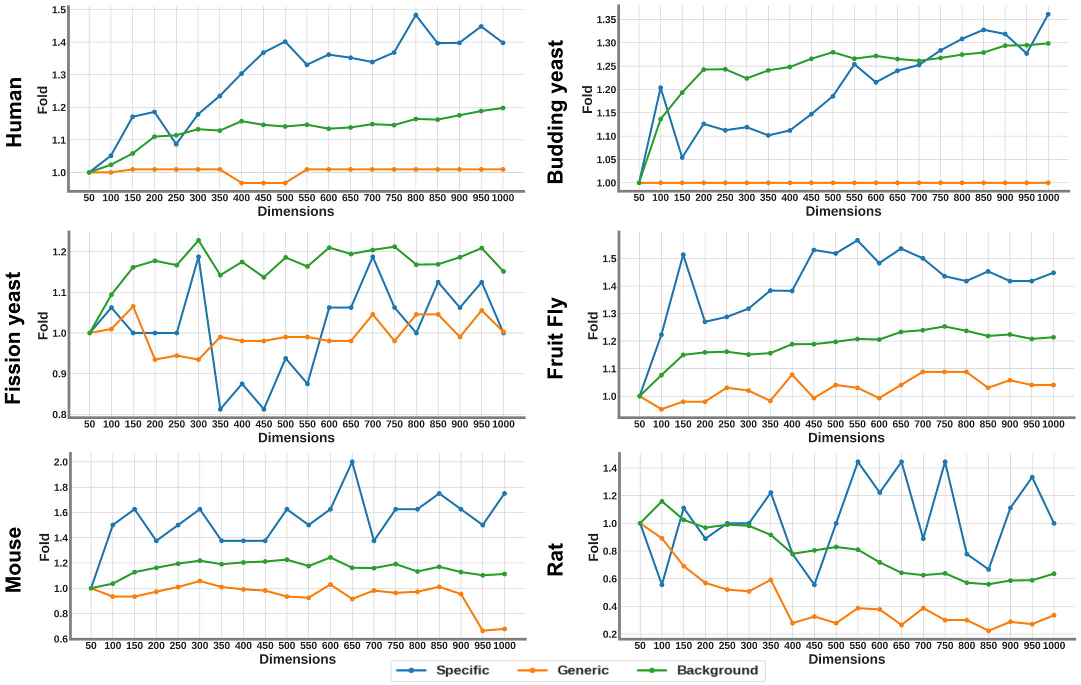


Figure C.2. Specific biological functions are captured by the axes of the species-specific ONMTF embedding spaces with the increment of dimensions. We generate the species-specific PPI embedding spaces by applying ONMTF on the species-specific PPI network of *Homo sapiens sapiens* (denoted by human), *Saccharomyces cerevisiae* (denoted by budding yeast), *Schizosaccharomyces pombe* (denoted by fission yeast), *Rattus norvegicus* (denoted by rat), *Drosophila melanogaster* (denoted by fruit fly), and *Mus musculus* (denoted by mouse) (detailed in sections 5.2.1 and 5.2.2 of Chapter 5). We generate these embedding spaces with different dimensionalities (from 50 to 1000 dimensions with a step of 50). For each species-specific embedding space, we take as a reference the 50-dimensional embedding space and we compute the fold between the number of "specific," "generic," and "background" functional annotations associated with its axes and that of the subsequent species-specific PPI embedding spaces (detailed in sections 5.2.4 and 5.3.3 of Chapter 5). The horizontal axis displays the number of dimensions of the embedding space. The horizontal axis displays the number of dimensions of the embedding space.

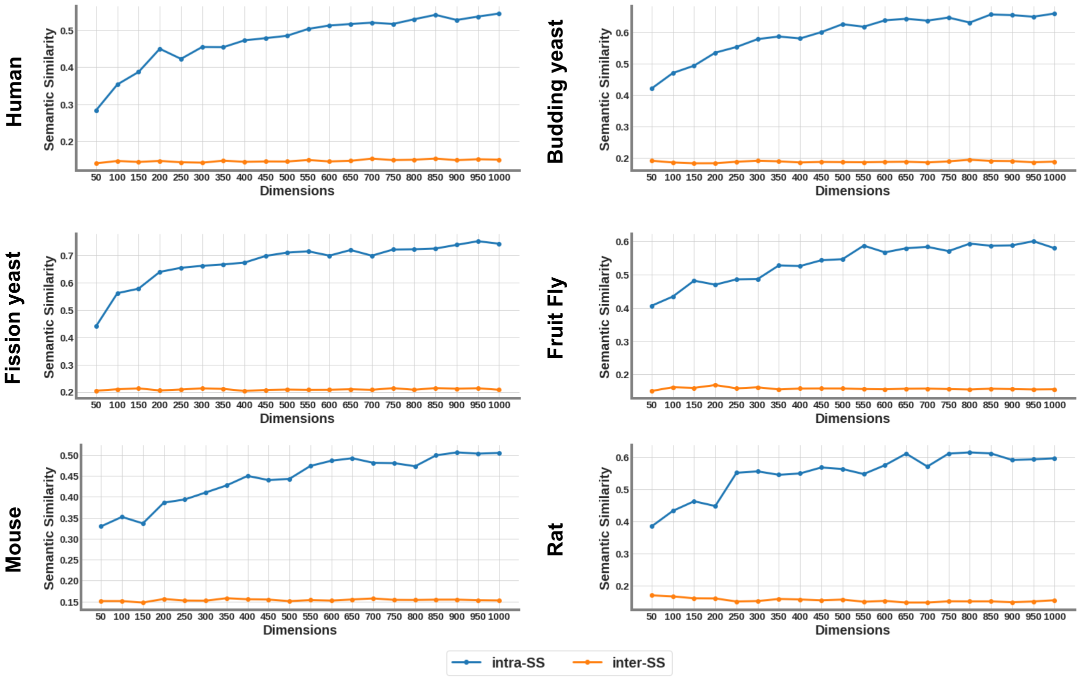


Figure C.3. Specific biological functions are disentangled by the axes of the species-specific ONMTF embedding spaces with the increment of dimensions. We generate the species-specific PPI embedding spaces by applying ONMTF on the species-specific PPI network of *Homo sapiens sapiens* (denoted by human), *Saccharomyces cerevisiae* (denoted by budding yeast), *Schizosaccharomyces pombe* (denoted by fission yeast), *Rattus norvegicus* (denoted by rat), *Drosophila melanogaster* (denoted by fruit fly), and *Mus musculus* (denoted by mouse) (detailed in sections 5.2.1 and 5.2.2 of Chapter 5). We generate these embedding spaces with different dimensionalities (from 50 to 1000 dimensions with a step of 50). For each species-specific embedding space, we compute Lin's semantic pairwise semantic similarity between any two GO BP terms (detailed in section 5.2.4 of Chapter 5). The blue line shows the average semantic similarity of the pairs of GO BP terms that are associated with the same axis (intra-SS). The orange line shows the average semantic similarity of the pairs of GO BP terms that are associated with different axis (inter-SS). The horizontal axis displays the number of dimensions of the embedding space.

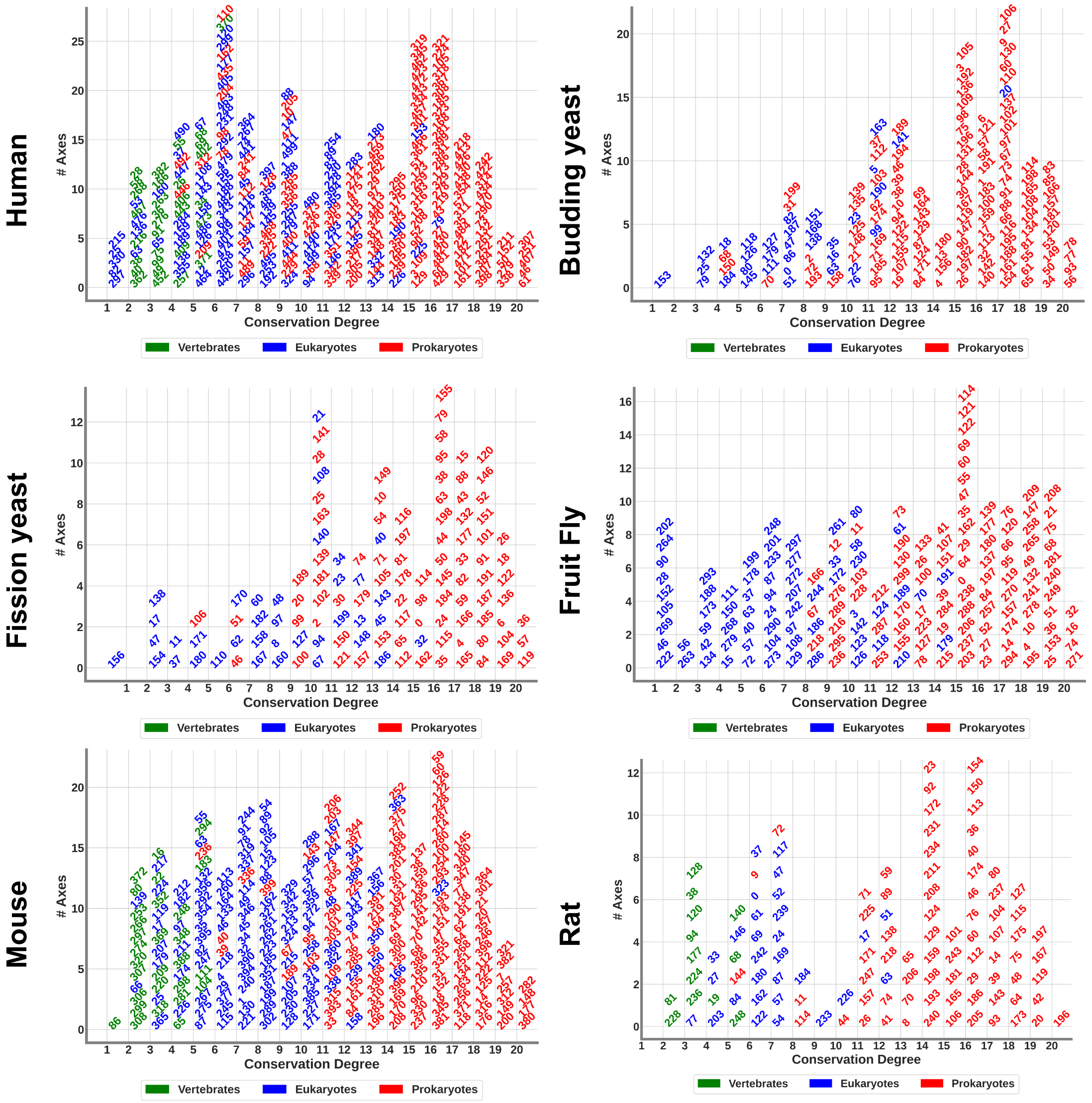


Figure C.4. The ASFAs give insights into the evolutionary story of *Saccharomyces cerevisiae* (denoted by budding yeast), *Schizosaccharomyces pombe* (denoted by fission yeast), *Rattus norvegicus* (denoted by rat), *Drosophila melanogaster* (denoted by fruit fly), and *Mus musculus* (denoted by mouse). For each species, we use the conservation degree of its ASFAs to divide them into three groups: "prokaryotes," "eukaryotes," and "vertebrates" (detailed in section 5.2.7 of Chapter 5). Then, we order the ASFAs according to their conservation degree. In each panel, the horizontal axis displays the conservation degree of the ASFAs and the vertical axis shows the number of ASFAs with a certain conservation degree. Each ASFA is represented in the panels by the number of the axis from which it was obtained.

C.3 Supplementary Tables for Chapter 5

Network	#Nodes	#Edges	#Density
Human	18,290	368,180	0.0022
Budding yeast	5,887	111,307	0.0064
Fission yeast	3,269	10,958	0.0020
Fruit fly	8,917	49,756	0.0012
Mouse	8,043	26,661	0.0008
Rat	2,847	5,252	0.0013

Table C.1. The statistics of the species-specific PPI networks. For the six species: *Homo sapiens sapiens* (denoted by "Human"), *Saccharomyces cerevisiae* (denoted by "Budding yeast"), *Schizosaccharomyces pombe* (denoted by "Fission yeast"), *Drosophila melanogaster* (denoted by "Fruit fly"), *Mus musculus* (denoted by "Mouse") and *Rattus norvegicus* (denoted by "Rat"). The first column, "Network," lists the species. The second column "# Nodes," show the number of nodes in the species-specific PPI network. The third column, "# Edges," contains the number of edges between the nodes. The fourth column, "# Density," specifies the edge density of the corresponding species-specific PPI network.

Species	# GO BP terms
Human	6,864
Budding yeast	3,042
Fission yeast	1,864
Fruit fly	3,712
Rat	2,828
Mouse	6,343

Table C.2. Number of GO BP annotations for each species-specific PPI network. For the six species: *Homo sapiens sapiens* (denoted by "Human"), *Saccharomyces cerevisiae* (denoted by "Budding yeast"), *Schizosaccharomyces pombe* (denoted by "Fission yeast"), *Drosophila melanogaster* (denoted by "Fruit fly"), *Rattus norvegicus* (denoted by "Rat") and *Mus musculus* (denoted by "Mouse"). The first column, "Network," lists the species. The second column, "# GO BP terms," presents the number of GO BP terms that annotates at least one gene in the corresponding species-specific PPI network.

Embedding algorithm	Intra SS	Inter SS	Random SS	Shortest Paths
ONMTF	0.50	0.16	0.16	3.71
NMTF	0.42	0.16	0.16	3.90
Deepwalk	0.35	0.16	0.16	4.31

Table C.3. On average, the GO BP terms captured by the axes of the human PPI embedding spaces generated by the ONMTF embedding algorithm are more coherent and better organized than those of the NMTF and Deepwalk spaces. We generate the human PPI embedding spaces by applying ONMTF, NMTF, and Deepwalk algorithms on the PPI network of *Homo sapiens sapiens* (detailed in sections 5.2.1 and 5.2.2 of Chapter 5). We generate these embedding spaces with different dimensionalities (from 50 to 1000 dimensions with a step of 50). For each human PPI embedding space, we use our new axes-based method to capture the GO BP terms that we embed in the space (detailed in section 5.2.3). Then, we investigate how coherently the captured GO BP terms are distributed across the axes according to the gene ontology (detailed in section 5.2.5 of Chapter 5). The first column, "Embedding algorithm," lists the embedding algorithms used for generating the embedding spaces. The second column, "Intra SS," shows the average Lin's semantic similarity between the GO BP terms that are associated by the same axis averaged across dimensions and species. The third column, "Inter SS," presents the average Lin's semantic similarity between the GO BP terms that are captured by different axes averaged across dimensions and species. The fourth column, "Random SS," shows the global average Lin's semantic similarity between any two GO BP terms. The fifth column, "Shortest Paths," displays the mean shortest paths in the GO ontology-directed acyclic graph between the GO BP terms associated with the same axis averaged across dimensions and species.

Embedding algorithm	% Axes	% GO
ONMTF	53.72	57.40
NMTF	61.80	48.12
Deepwalk	68.00	35.50

Table C.4. On average, the axes of the species-specific PPI embedding spaces generated by the ONMTF embedding algorithm are the best for capturing the cell's functional organization from PPI networks. We generate the species-specific PPI embedding spaces by applying ONMTF, NMTF, and Deepwalk algorithms on the species-specific PPI network of *Homo sapiens sapiens*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Rattus norvegicus*, *Drosophila melanogaster*, and *Mus musculus* (detailed in sections 5.2.1 and 5.2.2 of Chapter 5). We generate these embedding spaces with different dimensionalities (from 50 to 1000 dimensions with a step of 50). For each species-specific PPI embedding space, we use our new axes-based method to capture the GO BP terms that we embed in the space (detailed in Material and Methods, section 5.2.3). The first column, "Embedding algorithm," lists the embedding algorithms used for generating the embedding spaces. The second column, " % Axes," presents the percentage of axes that captures at least one embedded GO BP term averaged across dimensions and species. The third column, " % GO," shows the percentage of the total embedded GO BP terms that are associated with the axes of the space averaged across dimensions and species.

Embedding algorithm	Intra SS	Inter SS	Random SS	Shortest Paths
ONMTF	0.54	0.16	0.16	3.71
NMTF	0.48	0.18	0.16	3.90
Deepwalk	0.46	0.18	0.16	4.31

Table C.5. On average, the GO BP terms captured by the axes of the species-specific PPI embedding spaces generated by the ONMTF embedding algorithm are more coherent and better organized than those of the NMTF and Deepwalk spaces. We generate the species-specific PPI embedding spaces by applying ONMTF, NMTF, and Deepwalk algorithms on the species-specific PPI network of *Homo sapiens sapiens*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Rattus norvegicus*, *Drosophila melanogaster*, and *Mus musculus* (detailed in sections 5.2.1 and 5.2.2 of Chapter 5). We generate these embedding spaces with different dimensionalities (from 50 to 1000 dimensions with a step of 50). For each species-specific PPI embedding space, we use our new axes-based method to capture the GO BP terms that we embed in the space (detailed in section 5.2.3). Then, we investigate how coherently the captured GO BP terms are distributed across the axes according to the gene ontology (detailed in section 5.2.5 of Chapter 5). The first column, "Embedding algorithm," lists the embedding algorithms used for generating the embedding spaces. The second column, "Intra SS," shows the average Lin's semantic similarity between the GO BP terms that are associated by the same axis averaged across dimensions and species. The third column, "Inter SS," presents the average Lin's semantic similarity between the GO BP terms that are captured by different axes averaged across dimensions and species. The fourth column, "Random SS," shows the global average Lin's semantic similarity between any two GO BP terms. The fifth column, "Shortest Paths," displays the mean shortest paths in the GO ontology-directed acyclic graph between the GO BP terms associated with the same axis averaged across dimensions and species.

Species	# Dimensions
Human	500
Budding yeast	200
Fission yeast	200
Fruit fly	300
Rat	250
Mouse	400

Table C.6. The optimal number of dimensions for the six species-specific ONMTF embedding spaces. For the species-specific PPI embedding spaces generated by applying the ONMTF algorithm on the species-specific PPI network of *Homo sapiens sapiens* (denoted by human), *Saccharomyces cerevisiae* (denoted by budding yeast), *Schizosaccharomyces pombe* (denoted by fission yeast), *Rattus norvegicus* (denoted by rat), *Drosophila melanogaster* (denoted by fruit fly), and *Mus musculus* (denoted by mouse), we use our axes-based method to find their optimal dimensionality (detailed in section 5.3.3 of Chapter 5). The first column, "Species," lists the species. The second column, "# Dimensions," shows the optimal dimensionality of the species-specific PPI embedding space according to our axes-based method.

Species	Axis	Terms	#GO	Taxons
Budding yeast	79	endosome, Golgi, early, transport,	1	559292, 9606, 6239
Budding yeast	82	decapping, methylguanosine, RNA, cap, nuclear, deadenylation, mRNA, dependent, transcribed	3	4896, 9606, 10090, 3702, 7227, 559292, 6239
Budding yeast	77	methylation, subunit, benzene, regulation, translation, nucleus, initiation, compound, fidelity, process, gene, export, rRNA, assembly, amide, expression, small, containing, tRNA, positive, transport, post-transcriptional, ribosomal, cellular, translational, metabolic	12	4896, 214684, 10116, 9606, 9031, 511145, 10090, 36329, 39947, 195103, 9615, 7955, 3702, 352472, 9913, 7227, 559292, 227321, 6239, 9823
Fission yeast	20	subunit, large, biogenesis, complex, ribosomal, ribonucleoprotein	2	4896, 9606, 36329, 10090, 511145, 7955, 3702, 7227, 559292
Fission yeast	32	TORC2, regulation, TORC1, reproductive, signaling, process, positive	6	4896, 10116, 9606, 9031, 10090, 9615, 7955, 3702, 352472, 9913, 7227, 559292, 227321, 6239, 9823
Fission yeast	57	catabolic, protein, removal, conjugation, organonitrogen, compound, process, deneddylation, small, cellular, SCF, proteasomal, dependent, proteolysis, ubiquitin, metabolic, modification	9	4896, 214684, 10116, 9606, 9031, 36329, 10090, 511145, 195103, 39947, 9823, 9615, 7955, 3702, 352472, 9913, 7227, 559292, 227321, 6239
Fruit fly	1	synaptic, olfactory, mediated, vesicle, follicular, factor, negative, tyrosine, peptidyl, regulation, photoreceptor, epithelium, clathrin, filament, neuron, dorsal-ventral, transduction, eye, specification, signaling, cell, compound, commitment, learning, epidermal, growth, ERBB2, assembly, pathway, positive, transport, cascade, communication, dependent, organization, phosphorylation, fate, signal, modification, receptor	25	4896, 10116, 9606, 9031, 511145, 10090, 9823, 9615, 7955, 3702, 352472, 9913, 7227, 559292, 6239
Fruit fly	28	negative, vein, regulation, disc, derived, specification, imaginal, wing	1	7227
Fruit fly	4	cation, biosynthetic, metal, divalent, regulation, retinal, aldehyde, ion, compound, lipid, process, olefinic, inorganic, transport, diterpenoid, cellular, retinoid, homeostasis, metabolic	9	4896, 214684, 10116, 9606, 9031, 511145, 10090, 36329, 39947, 9823, 9615, 7955, 3702, 352472, 9913, 7227, 559292, 6239
Mouse	41	immune, type, lipopolysaccharide, negative, alpha, interferon, response, regulation, innate, pattern, signaling, involved, recognition, pathway, virus, dsRNA, interleukin, inflammatory, cytokine, production, receptor	10	10116, 9606, 9031, 511145, 10090, 9823, 9615, 7955, 3702, 352472, 9913, 7227, 559292, 6239
Mouse	7	biosynthetic, estrogen, glycerophospholipid, glycerolipid, lipid, process, phosphatidylcholine, metabolic	5	4896, 214684, 10116, 9606, 9031, 36329, 10090, 511145, 39947, 195103, 9823, 9615, 7955, 3702, 352472, 9913, 7227, 559292, 6239
Mouse	86	type, induction, lens, eye, camera	1	10090
Mouse	80	blastocyst, development	1	10090, 9606
Rat	69	synaptic, signaling, trans, anterograde, transmission, chemical	4	10116, 9606, 10090, 7955, 7227, 6239
Rat	84	mediated, intracellular, signaling, steroid, pathway, hormone, androgen, receptor	3	10116, 9606, 10090, 3702, 7227
Rat	51	negative, regulation, polymerization, ion, microtubule, polymerization, import, calcium	4	4896, 10116, 9606, 9031, 10090, 7955, 3702, 352472, 9913, 7227, 559292, 6239
Rat	81	estrous, cycle, ovulation	2	10090, 10116
Rat	19	odontogenesis	1	10090, 7955, 10116, 9606

Table C.7. The species-specific ASFAs describe coherent functions of six species. For the species-specific PPI embedding spaces generated by applying the ONMTF algorithm on the species-specific PPI network of *Homo sapiens sapiens* (denoted by human), *Saccharomyces cerevisiae* (denoted by budding yeast), *Schizosaccharomyces pombe* (denoted by fission yeast), *Rattus norvegicus* (denoted by rat), *Drosophila melanogaster* (denoted by fruit fly), and *Mus musculus* (denoted by mouse), we use our new axes-based method to capture the GO BP terms that we embed in the space (detailed in sections 5.2.1, 5.2.2, and 5.2.3 of Chapter 5). Then, we use the GO BP terms captured by the axes of the embedding spaces to generate the ASFAs (detailed in section 5.2.6 of Chapter 5). The first column, "Species," lists the species. The second column, "Axis," lists the name of the axes from which each ASFA was obtained. The third column, "Terms," shows the description of the ASFAs. The fourth column, "#GO," displays the number of GO BP terms that are associated with the axis. The fifth column, "Taxons," shows the Taxonomy ID of the different species for which the associated GO BP terms appear.

Species	Empty vs Non-Empty	Empty vs Random
Human	1.76×10^{-63}	6.46×10^{-28}
Budding yeast	1.67×10^{-8}	6.94×10^{-43}
Fission yeast	7.42×10^{-17}	1.81×10^{-62}
Fruit fly	8.85×10^{-15}	1.05×10^{-42}
Rat	0.11	7.89×10^{-6}
Mouse	0.01	9.18×10^{-30}

Table C.8. Genes that are associated with the empty axes tend to form densely connected neighborhoods in the species-specific PPI networks. We generate the species-specific embedding spaces by applying the ONMTF algorithm on the species-specific PPI network of *Homo sapiens sapiens* (denoted by human), *Saccharomyces cerevisiae* (denoted by budding yeast), *Schizosaccharomyces pombe* (denoted by fission yeast), *Rattus norvegicus* (denoted by rat), *Drosophila melanogaster* (denoted by fruit fly), and *Mus musculus* (denoted by mouse). For each species-specific PPI embedding space, we associate genes with their embedding axes. Then, we evaluate the connectivity in the original species-specific PPI network by computing the clustering coefficient between genes associated with the same axis (detailed in section 5.2.6 of Chapter 5). The first column, "Species," lists the species. The second column, "Empty vs Non-Empty," shows the p-value from a one-sided Mann-Whitney U test comparing if the clustering coefficient of the genes associated with non-empty axes (axes with at least one associated GO BP term) is statistically higher than the clustering coefficient of genes associated with empty axes (axes with non-associated GO BP terms). The third column, "Empty vs Random," displays the p-value from a one-sided Mann-Whitney U test comparing if the clustering coefficient of the genes associated with empty axes is statistically higher than expected by random.