Graphlet-adjacencies provide complementary views on the functional organisation of the cell and cancer mechanisms

Sam Windels

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Department of Computing

University College London

September 29, 2022

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

Abstract

Recent biotechnological advances have led to a wealth of biological network data. Topological analysis of these networks (i.e., the analysis of their structure) has led to breakthroughs in biology and medicine. The state-of-the-art topological node and network descriptors are based on graphlets, induced connected subgraphs of different shapes (e.g., paths, triangles). However, current graphlet-based methods ignore neighbourhood information (i.e., what nodes are connected). Therefore, to capture topology and connectivity information simultaneously, I introduce graphlet adjacency, which considers two nodes adjacent based on their frequency of co-occurrence on a given graphlet. I use graphlet adjacency to generalise spectral methods and apply these on molecular networks. I show that, depending on the chosen graphlet, graphlet spectral clustering uncovers clusters enriched in different biological functions, and graphlet diffusion of gene mutation scores predicts different sets of cancer driver genes. This demonstrates that graphlet adjacency captures topology-function and topology-disease relationships in molecular networks.

To further detail these relationships, I take a pathway-focused approach. To enable this investigation, I introduce graphlet eigencentrality to compute the importance of a gene in a pathway either from the local pathway perspective or from the global network perspective. I show that pathways are best described by the graphlet adjacencies that capture the importance of their functionally critical genes. I also show that cancer driver genes characteristically perform hub roles between pathways.

Given the latter finding, I hypothesise that cancer pathways should be identified by changes in their pathway-pathway relationships. Within this context, I propose pathway-driven non-negative matrix tri-factorisation (PNMTF), which fuses molecular network data and pathway annotations to learn an embedding space that captures the organisation of a network as a composition of subnetworks. In this space, I measure the functional importance of a pathway or gene in the cell and its functional disruption in cancer. I apply this method to predict genes and the pathways involved in four major cancers. By

using graphlet-adjacency, I can exploit the tendency of cancer-related genes to perform hub roles to improve the prediction accuracy.

Impact statement

Graphlet adjacency is an important work that redefines the neighbourhood of a node in a network based on structural patterns in its connectivity. So far, I used graphlet adjacency as a basis for generalising traditional spectral methods to account for pattern specific interactions (see Chapters 3 and 4, or papers (Windels *et al.*, 2019) and (Windels *et al.*, 2022)). The introduction of graphlet adjacency has led to new lines of research: (1) generalising network embedding algorithms, such as spring embedding, multidimensional scaling and coalescent embedding, to account for connectivity patterns (thesis work of Daniel Tello, 3rd-year PhD Student, University of Barcelona, paper in preparation) and (2) defining new Natural Language Processing (NLP) inspired matrix representations of networks that also account for connectivity patterns (thesis work of Alexandros Xenos, 3rd year PhD student, Polytechnic University of Catalonia, paper in preparation). In the closing notes of this thesis, I list further extensions to the presented methods, including improvements in graphlet counting and defining alternative notions of graphlet based connectivity. Graphlet adjacency also forms the basis of a grant proposal I submitted in Belgium at the Reasearch Foundation Flanders for Post-Doc funding. Lastly, graphlet adjacency is part of the methods and know-how that are being transferred to a graphlet technologies based start-up, via the ERC proof of concept grant 957488 of Professor Nataša Pržulj.

I apply Pathway-driven non-negative matrix factorisation (PNMTF) to identify cancer implicated pathways based on the extent that their interactions with other pathways change in cancer. With this, I effectively introduce a new paradigm in cancer pathway analysis, as conventional approaches identify cancer-related pathways based on their internal rewiring. A current limitation of the PNMTF model is that it only allows to predict genes as cancer related if they are pathway annotated. To overcome this limitation, in the closing notes of the thesis, I redefine PNMTF as a subspace clustering algorithm, which learns the assignment of nodes to subnetworks, i.e., *de novo* pathways.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation and objectives

Biology is flooded with large scale "omic" data. Genomic, proteomic, transcriptomic and other data are typically modelled as *networks* (also called *graphs*). In molecular networks, nodes usually represent genes or proteins and edges represent interactions or relationships between them, such as physical interactions between the proteins (PPI), genetic interactions (GI), or co-expressions (COEX) (Stark *et al.*, 2006; Okamura *et al.*, 2015).

Topological analysis of these networks, i.e. the analysis of their structure, has led to breakthroughs in biology and medicine. At the node level, two main categories of methods exist. *Centrality* based methods quantify the importance of a node either based on its connectivity or based on its frequency of occurrence on shortest paths. Ever since Jeong *et al.* (2001) showed that perturbing highly connected nodes in PPI networks has a higher probability of impacting cell viability, these methods have become a major tool for discovering gene functions and uncovering disease-related genes (e.g., see Wang *et al.* (2011); Guo *et al.* (2016); Asensio *et al.* (2017)). *Graphlet* based approaches describe the local wiring of a network around a node based on its frequency of occurrences on *graphlets*, small, connected, induced subgraphs of different 'shapes' (i.e., paths, triangles, etc.) (Pržulj *et al.*, 2004). Graphlet based methods have been widely applied to molecular networks, for instance to predict protein function (Gaudelet *et al.*, 2018) and to identify age-related genes (Li and Milenković, 2019). Current graphlet approaches all suffer from the same shortcoming: they ignore node-neighbourhood information, i.e., they only capture the structure of the network around a given node, ignoring information based on the identity of its neighbours or its location in the network.

In the case of molecular networks, it is known that they are also structurally organised as a composition of *pathways*, functional subnetworks within the cell that once

activated lead to a certain product or change within the cell (Vogelstein and Kinzler, 2004; DeBerardinis and Chandel, 2016). Pathway-focused network analysis approaches are frequently considered to study diseases, as they provide functional context to the observed gene mutations, which help to identify potential drug targets and determine disease subtypes (Creixell *et al.*, 2015). Current topological approaches prioritise disease implicated pathways based on their internal rewiring, i.e., the rewiring of their nodes. To the best of my knowledge, no approaches exist that take a higher-level view, prioritising disease implicated pathways based on the rewiring of pathway-pathway interactions.

In this thesis, I aim to study the higher-order structure of networks at the node and subnetwork level. In particular, I aim to extend graphlet-based methods to take node-connectivity information into account. Additionally, I take higher-point of view and build methods that capture the organisation of a network as an organisation of large molecular subnetworks. I apply these methods to study the functional organisation of molecular networks, uncover disease mechanisms and disease genes.

1.2 Thesis contributions

In this thesis I introduce new methodologies to study the higher-order topology of networks and apply them to solve problems in biology.

Methodological contributions. I built methods to describe and capture the higher-order topology of networks:

- **Chapter 3**: To capture the higher-order organisation of nodes in a network, I introduce *graphlet adjacency*, which considers two nodes connected based on their frequency of co-occurrence in a given graphlet (induced connected subgraphs of different 'shapes', such as paths and triangles). I use graphlet adjacency to generalise spectral methods spectral embedding and spectral clustering. This work has been published in (Windels *et al.*, 2019).

- **Chapter 4**: To capture the topological importance of nodes in a network based on their higher-order topology, as captured by graphlet adjacency, I introduce graphlet eigencentrality. For a given graphlet, a node has a high graphlet eigencentrality if it and its neighbours frequently occur on that graphlet. This work has been published in (Windels *et al.*, 2022).

- **Chapter 5**: To capture the organisation of a network as a composition of subnet-

works, I propose pathways-driven non-negative matrix tri-factorisation (PNMTF), which fuses network data and prior domain-specific knowledge assigning nodes to subnetworks in the network. PNMTF allows to measure the topological importance of nodes and subnetworks in a network and the rewiring of their interactions between two different network states.

Biological contributions. I apply the developed methods to study the functional organisation of molecular networks and disease mechanisms. I results show that these methods capture strong topology-function and topology-disease relationships in molecular networks:

- **Chapter 3**: Graphlet adjacency provides complimentary views of the functional organisation of molecular networks. This is shown for multiple types of molecular networks, species and functional annotations. These results have been published in (Windels *et al.*, 2019).

- **Chapter 4**: Graphlet adjacency allows to predict complementary sets of cancer implicated genes. This is shown for multiple types of molecular networks. These results have been published in (Windels *et al.*, 2022).

- **Chapter 5**: PNMTF allows to predict genes and the pathways involved in four major cancers based on their functional importance in the healthy cell and their changing functional relationships in cancer. By combining PNMTF with graphlet adjacency, the tendency of driver genes to perform hub roles can be exploited to increase prediction accuracy. Strong literature support is provided for the top predicted genes, of which six are predicted as potential cancer-specific drug targets.

1.3 Thesis outline

The thesis is outlined as follows:

Chapter 2: In this chapter I present relevant concepts from molecular biology, network science and machine learning. Further, I provide an overview of biological network data, model network data and network analysis methods. In particular, I focus on graphlet based methods and spectral analysis. I also present key machine learning based data-integration approaches that have been used in cancer precision medicine, with the main focus on non-negative matrix factorization (NMF) and non-negative matrix tri-factorization (NMTF). This chapter is based on (Malod-Dognin *et al.*, 2019a).

Chapter 3: In this chapter I introduce graphlet-adjacency, which I use to generalise spectral embedding, spectral clustering and spectral diffusion. Through spectral clustering and enrichment analysis, I show that different graphlet adjacencies capture complementary biological function. I explain the complementarity of biological functions captured by different graphlet adjacencies by showing that they capture different local topologies in model networks. Finally, by diffusing pan-cancer gene mutation scores based on different Graphlet Laplacians, I find complementary sets of cancer related genes. Hence, I demonstrate that Graphlet Laplacians capture topology-function and topology-disease relationships in biological networks. This chapter is based on (Windels *et al.*, 2019).

Chapter 4: In this chapter I build on the previously defined graphlet adjacency to define graphlet eigencentrality. I apply this method to further investigate the relationships between the topological features of genes in molecular networks as captured by graphlet adjacencies and their biological functions, taking a more descriptive pathway-based approach. I show that pathways are best described by the graphlet adjacencies that capture the importance of their functionally critical genes. I also show that cancer driver genes characteristically perform hub roles between pathways. This chapter is based on (Windels *et al.*, 2022).

Chapter 5: In this chapter I introduce pathway-driven non-negative tri-matrix factorisation (PNMTF), which learns the organisations of pathways in the healthy cell. Building on the observation that cancer drivers perform hub-roles between pathways, I apply PNMTF to predict cancer pathways based on their changing pathway interactions in cancer. From the set of genes involved in the prioritised pathways, I prioritise genes implicated in cancer. By combining PNMTF with graphlet adjacency, I can better capture the hub-roles of genes and increase the prediction accuracy.

1.4 List of publications

Major results of this thesis have either been published in peer-reviewed journals. Below is a list of published articles.

Publications:

- Malod-Dognin, N., Windels, S. F. L., and Pržulj, N. (2019). *Machine Learning for Data Integration in Cancer Precision Medicine: Matrix Factorization Approaches*,

a chapter in *Analyzing Network Data in Biology and Medicine*, edited by Nataša Pržulj, Cambridge University Press.

- Malod-Dognin, N., Petschnigg, J., Windels, S. F. L., Povh, J., Hemmingway, H., Ketteler, R., and Pržulj, N. (2019). Towards a data-integrated cell. *Nature Communications*, **10**(1), 805.
- Windels, S. F. L., Malod-Dognin, N., and Pržulj, N. (2019). Graphlet Laplacians for topology-function and topology-disease relationships. *Bioinformatics*, **35**(24), 5226–5234.
- Windels, S. F. L., Malod-Dognin, N., and Pržulj, N. (2021). Graphlet eigencentralities capture novel central roles of genes in pathways. *PLoS One*, **35**(24), 5226–5234.
- Windels, S. F. L., Malod-Dognin, N., and Pržulj, N. (2022). Identifying cellular cancer mechanisms through pathway-driven data integration. *Bioinformatics*, btac493.

Chapter 2

Background

In this thesis, I study the higher-order structure of molecular networks. In this chapter, I introduce key concepts from biological network analysis, spectral theory, data-integration applied machine learning and pathway focused approaches to study cancer disease mechanism, that are extensively used throughout this work.

The machine learning algorithms presented throughout this chapter are exclusively examples of *unsupervised learning* or *semi-supervised learning*, i.e., they uncover patterns in the data without or with incomplete prior knowledge of those patterns. In contrast, *supervised learning* learns to predict an A priori known class label for each data point. In network-biology applications however, such labels are typically not available or incomplete. After all, the nature of network biology is hypothesis generating or explorative, aiming to uncover disease mechanisms, patient subtypes, etc. Hence, supervised learning approaches fall outside of the scope of this chapter.

2.1 Network analysis

Network analysis studies the relationship between different discrete entities, such as genes, patients or drugs, through their representation as networks. Below, I first define the different biological networks subject of this thesis (Section 2.1.1). Then, I define different network centrality measures, which are used to describe different notions of the importance of a node in a network (Section 2.1.2). Next, I define different classical network descriptors that measure different key structural properties of a network (Section 2.1.3). I then go on to define graphlet-based node descriptors and network descriptors, which are the current state-of-the-art (Section 2.1.4). Finally, I define theoretical model networks that are frequently used to study biological networks (Section 2.1.5).

2.1.1 Biological data and their network representation

Biology is flooded with large-scale "omic" data. Genomic, transcriptomic, proteomic, metabolomic and other data are typically modelled as *networks* (also called *graphs*). In this thesis I will focus on gene-based interaction networks, illustrated in Figure 2.1, that capture different types of relationships (represented by unweighted and undirected edges) between genes or proteins (represented by nodes).

- A *protein-protein interaction (PPI) network* consists of proteins (nodes) and physical bindings (edges). PPI networks are an example of a *molecular interaction network*, in which nodes represent molecules and edges represent the molecules that physically bind (Gligorijević and Pržulj, 2015). As proteins are encoded by genes, often gene notations are used for proteins. Firstly, this allows for easy integration and comparison with other gene-based interaction networks. Secondly, this allows for the computation of context specific PPI networks, e.g. for a given tissue or disease, by inducing the set of context specific expressed genes on the set of nodes in the generic PPI network (see for example (Kotlyar *et al.*, 2019)). Key PPI databases include BioGRID (Oughtred *et al.*, 2019) and IID (Kotlyar *et al.*, 2019).

- A *co-expression (COEX) network* captures which genes (nodes) are typically co-expressed (edges), i.e. have a very similar expression pattern under different experimental conditions. A COEX network is an example of a *similarity network*, in which nodes are connected if they are functionally or structurally similar. Creating a COEX network is a two-step process. First, a pairwise similarity score is computed between all the genes, based on their expression values under different experimental conditions. Usually, the Pearson Correlation Coefficient (PCC) is used. Next, some thresholding strategy is applied. For instance, one strategy is to connect each gene to the top 1% of genes that have the most correlated expression levels. A different approach would be to connect all gene pairs that show statistically significantly correlated gene expression levels. It should be clear that depending on the similarity metric and thresholding strategy used, COEX networks of vastly different topology could be generated based on the same gene-expression dataset. A key database for gene co-expression data is the COXPRESdb database (Obayashi *et al.*, 2019).

- A *genetic interaction (GI) network* captures the relations between genes (nodes) of which their simultaneous mutation has measurable effect on the cell's phenotype

(edges). If the simultaneous mutation of two genes has a positive (negative) effect on the cell viability, statistically significantly larger than expected from their individual mutation, the genes have a positive (negative) genetic interaction (Costanzo *et al.*, 2010). Note that the genes part of a genetic interaction do not necessarily physically interact. Rather, they are functionally associated and are therefore an example of a *functional association network*. Genetic interactions can be found in the BioGRID database (Oughtred *et al.*, 2019).

I denote a networks as $H = (V, E)$, where V is the set of n nodes in the network, and E is the set of edges (West, 2001). The adjacency of all nodes in network H is represented by an $n \times n$ symmetric *adjacency matrix* A, where the entries A_{uv} are 1 if nodes $\{u, v\} \in V$ are connected by an edge in E (i.e. are 'adjacent') and 0 otherwise. For a given node u, its (direct) neighbourhood, N_u, is defined as the set of nodes in the network it is connected to by an edge: $N_u = \{v : e_{uv} \in E \vee e_{vu} \in E\}$. The degree of a node u, k_u, is equal to the number of nodes in its neighbourhood: $k_u = |N_u|$.

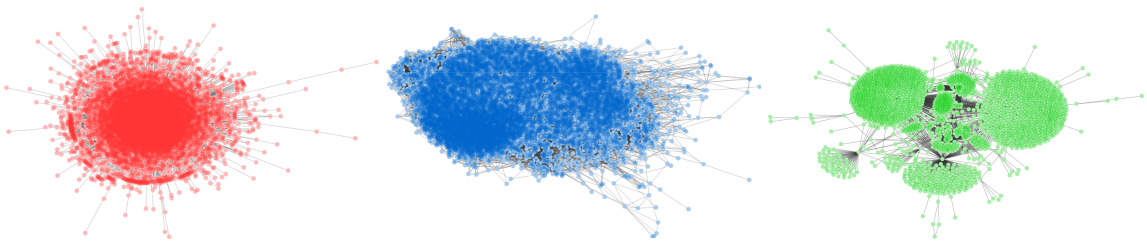


Figure 2.1: Molecular networks in human. From left to right, a breast cancer specific human PPI, COEX and GI network. This Figure is adapted from Figure 1 in (Malod-Dognin *et al.*, 2019b).

2.1.2 Node centrality

Ever since Jeong *et al.* (2001) showed that perturbing highly connected nodes in PPI networks is more likely to impact cell viability, *node centrality* measures, which measure different notions of topological importance of a node in a network, have become a major tool for discovering gene functions and uncovering disease-related genes (e.g. see Wang *et al.* (2011); Isik *et al.* (2015); Guo *et al.* (2016); Asensio *et al.* (2017)). Below I focus on four major node centrality measures. Similar concepts can also be considered from the edge point of view. For more details, I refer the reader to Newman (2010).

Centrality measures can be classified into two categories. The first category of centrality measures quantifies the importance of a node based on its connectivity.

- The *degree centrality* considers highly connected nodes to be the most important

nodes in the network. As such, the degree centrality of a node is synonymous for its degree: it is its number of neighbours in the network, or equivalently, the number of edges in the network including the node. Formally, the degree centrality, of a node, $u \in V$, is:

$$DC(u) = |N_u| = \sum_{v=1}^{n} A_{uv}. \tag{2.1}$$

- The *eigencentrality* is a more sophisticated version of the degree centrality, which considers the most important nodes to be the nodes that are highly connected to other highly connected nodes in the network. Formally, the centrality of a node $u \in N$, EC_u, is defined as the average of the centralities of its m neighbours:

$$EC(u) = \frac{1}{\lambda} \sum_{v=1}^{m} EC(v) A_{uv}, \tag{2.2}$$

  where λ is a constant. More detail into the intuition behind the eigencentrality and its computation is provided in chapter 2.2 on spectral theory.

The second category of centrality measures quantifies the importance of a node based on network paths. A *path* in a network is the sequence of node and edges traversed when following edges from one to another across the network, without revisiting the same nodes or edges. The *length* of a path is equal to the number of edges traversed. The *shortest path* between a pair of nodes is a path such that the number of edges and nodes traversed is minimal. The *graph geodesic distance* between a pair of nodes is the length of the shortest path between them. Note that the shortest path between a pair of nodes in a network is not necessarily unique.

- The *closeness centrality* considers a node to be central in the network if it is nearby to all other nodes in the network. Formally, the closeness centrality a node, $u \in V$, is equal to the reciprocal of the average distance of u to every other node in the network:

$$CC(u) = \frac{1}{\sum_{v=1}^{n} d(u,v)/n}, \tag{2.3}$$

  were $d(u,v)$ is the distance between u and v.

- The *betweenness centrality* measures the amount of control u has on the flow of information in the network. Formally, the betweenness centrality of a node u is the fraction of shortest paths between all nodes in the network on which u occurs over

all shortest paths in the network:

$$BC(u) = \sum^{s,t \in V} \frac{\sigma(s,t|u)}{\sigma(s,t)}, \tag{2.4}$$

where $\sigma(s,t)$ is the number of shortest paths in the network from node s to node t, and $\sigma(s,t|v)$ is the number of those paths that include node u.

Per definition, these different centrality measures capture different but related notions of centrality. For instance, Figure 2.2 shows that the different centrality measures presented are statistically significantly correlated in an Erdős-Rènyi (ER) random network (see Section 2.1.5, (Erdős Paul and Rényi Alfréd, 1959)). The strength of the correlations between these different centrality measures depends on the topology or wiring pattern of the global network on which these centrality measures are applied (Oldham *et al.*, 2019).

	DC	**EC**	**CC**	**BC**
DC		0.93	0.94	0.95
EC			0.92	0.76
CC				0.84
BC				

Figure 2.2: Different centrality measures are related. I present the pairwise Spearman Rank Correlation between the different presented centrality measures (columns and rows), applied on an Erdős-Rènyi random network of $5,000$ nodes and $625,000$ edges. All centrality measures are highly correlated, as the minimum correlation found is 0.76 and all correlations are statistically significant at the 1% significance level.

2.1.3 Classical network descriptors

Network descriptors summarize the global structure of a network. Here I list global network descriptors frequently used to describe and compare networks.

- The *density* of a network is the ratio of the total number of edges in the network over the potential number of edges in the network: $\frac{|E|}{(n(n-1))/2}$. Equivalently, it is the probability that a randomly selected pair of nodes in the network is connected.

- At the node level, the *local clustering coefficient* is defined as the number of edges between its neighbours over the possible number of edges between them: $CC(U) = \frac{|e_{vw}:v,w \in N_u, e_{vw} \in E|}{(d_u(d_u-1)/2)}$. It quantifies how close the neighbours of a given node are to being a clique. The *global clustering coefficient* is defined as the average of the local clustering coefficient over all nodes. It is a measure of the extent to which nodes in a network tend to cluster together. Equivalently, it reflects the likelihood that a pair of nodes interact given that they share a common neighbour.

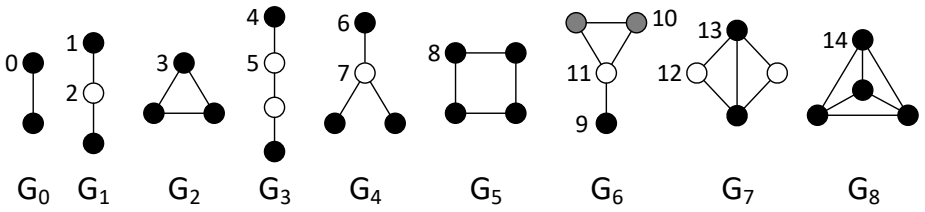
- The *average path length* is the mean geodesic distance between all pairs of nodes in the network. If a network has a small average path length and high clustering coefficient, it is considered to be *small-world*. In this case, nodes are expected to share neighbours so that most nodes can be reached from every other node by a traversing only small number edges. A related measure is the network *diameter*, which is equal to the length of the longest shortest path in the network.

- The *degree distribution* is the distribution of node degrees over all nodes. It is summarized as a vector of counts, i.e. the k^{th} value is the number of nodes that have degree k. Many networks, including PPI networks, display a *scale-free* degree distribution, where some nodes, called *hubs*, have many more connections than other nodes in the network. Formally, a network is scale-free if its degree distribution follows a power law: $P(k) \backsim k^{-\gamma}$, with the constant γ typically ranging from 2 to 3 in real-world scale-free networks (Jeong *et al.*, 2001; Broido and Clauset, 2019).

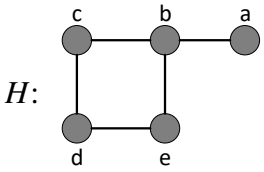
2.1.4 Capturing and quantifying network topology with graphlets

Current state-of-the-art node descriptors and network descriptors are based on *graphlets*: small, connected, non-isomorphic, induced sub-graphs of a large network (Pržulj *et al.*, 2004). Graphlet based methods have been applied to predict protein function (Gaudelet *et al.*, 2018) and identify age-related genes (Li and Milenković, 2019) based on their interaction patterns in PPI networks.

(A)



(B)



(C)

Graphlet	G_0	G_1		G_2	G_3		G_4		G_5	G_6			G_7		G_8
Orbit	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
GDV(a)	1	2	0	0	2	0	1	0	0	0	0	0	0	0	0

Graphlet Count(a)	1	2	0	2	1	0	0	0	0

Figure 2.3: Illustration of graphlets and graphlet degree vector. A: All graphlets with up to 4 nodes, labeled G_0 to G_8. **B:** A dummy network H. The 15 automorphism orbits are differently shaded and labeled (from 0 to 14) within each graphlet. **C:** The graphlet degree vector of node 'a' in the example network, H, and its relationship to graphlet counts. Node 'a' touches graphlet G_0 once on orbit 0, via edge a-b. Node 'a' touches graphlet G_1 twice, each time at orbit 1, via paths a-b-c and a-b-e. This figure is taken from (Windels *et al.*, 2019).

I illustrate graphlets in Figure 2.3-A. Within each graphlet, automorphism orbits are groups of nodes such that node-permutations within the orbit (i.e. swapping them) preserves the structure of the graphlet (Pržulj, 2007), as illustrated by differently shaded nodes in Figure 2.3-A. The *graphlet degree vector* (GDV) quantifies the local topology of a node as a vector of counts of how many times it touches each graphlet at a particular automorphism orbit (Milenković and Pržulj, 2008), as illustrated in Figure 2.3-C. *Representational learning* is an alternative approach for computing vectors that capture the local topology nodes in a network. For instance, node2vec learns a mapping of nodes to a low-dimensional space of features that maximizes the likelihood of preserving network neighborhoods of nodes (Grover and Leskovec, 2016). However, unlike GDV's, which describe the local topology in terms of orbit counts, these learned features are not readily interpretable.

Two graphlet based network descriptors exist. The *graphlet degree distribution* gen-

eralizes the degree distribution. It is the joint distribution of the graphlet counts over all nodes for each of the different graphlets (Pržulj, 2007). Alternatively, the Graphlet Correlation Matrix (GCM) summarizes network topology as an 11×11 symmetric matrix comprising the pairwise Spearman's correlations between 11 non-redundant orbit counts over all nodes in the network (Yaveroğlu *et al.*, 2014).

Because of biotechnological capturing limitations and human biases, biological networks are globally noisy and locally less noisy, so methods designed to capture the local structure of the data have been shown to generally outperform global methods (Roweis and Saul, 2000; Wu and Scholkopf, 2006). As graphlets are used to capture local wiring patterns, graphlet-based methods have also been shown to be robust to noisy data. For instance, when using GDVs on PPI networks to predict protein functions, similar predictive performance is measured on less complete PPI networks (Milenković and Pržulj, 2008). When clustering different types of molecular networks of different sizes (i.e., completeness) based on their GCD distance, the networks are grouped by their molecular type rather than their size (Yaveroğlu *et al.*, 2014).

2.1.5 Model networks

To provide insight into the organisational principles and evolution of molecular networks, their topology is usually compared to that of theoretical models. Below I list some of the key model networks used today.

- The Erdős-Rènyi random networks (ER) represent uniformly distributed random interactions (Erdős Paul and Rényi Alfréd, 1959). ER networks are generated by fixing the number of nodes to n, and by randomly adding edges between uniformly chosen pairs of nodes (out of the $n(n-1)/2$ possible pairs of nodes) until the density is the same as that of the real network.

- The generalized random model (ER-DD) is an extension of ER model, where the distribution of the degrees of nodes in the generated network mimic the one of an input network (Newman, 2009). ER-DD networks are generate by assigning connection capacities (stubs) to the n nodes of the network, and then adding edges between nodes that have available stubs uniformly at random while reducing the available stubs of the newly connected nodes after each edge addition.

- The random geometric model (GEO) represents proximity relationships between uniformly distributed points in a space (Penrose, 2003). GEO networks are gen-

erated by uniformly distributing n points (nodes) in 3-dimensional space and by connecting nodes by edges if the Euclidean distances between the corresponding points are lower than or equal to threshold r, which is set so to obtain the edge density that similar to that of the real network.

- The Geometric with gene duplication model (GEO-GD) is a geometric model in which the dispersion of nodes is no longer uniformly random, but according to a duplication rule, mimicking the gene duplication process in evolution (Pržulj *et al.*, 2010). GEO-GD network are generated starting from a seed network (i.e. a single edge) to which the duplication process is applied: a randomly chosen parent node is duplicated, and the new node is randomly placed at a distance smaller than or equal to $2r$ (where r is the same distance threshold as in GEO model). This duplication process iterates until the required number of nodes is generated, after which edges are created following the GEO model rules so as to achieve the requested edge density.

- The Barabási-Albert scale-free model network (SF-BA) (Barabási and Albert, 1999) are characterized by a scale-free topology, meaning the networks have a degree distribution that follows a power law (Barabási and Albert, 1999). SF-BA networks are generated starting from small seeds (one edge), to which nodes are added based on the "rich-gets-richer" principle: new nodes are attached to existing nodes of the network with a probability based on that of the nodes in the underlying real network. Scale-free networks are considered to be robust to the random failure of nodes, yet vulnerable to failures of hub-nodes. Firstly, this is because the likelihood that a hub-node is affected is relatively low. Secondly, even when a high degree node is removed, as SF-BA networks are small world, other high degree nodes are expected to be nearby. Therefore, removing a high degree node is not expected to significantly impact the connectedness of the network (Cohen *et al.*, 2000).

- The scale-free model network mimicking gene duplication and mutation events (SF-GD) is a scale-free model that mimics the gene duplication and gene divergence processes from evolution (Vazquez *et al.*, 2001). SF-GD networks are generated starting from a small seed network (one edge), which is grown through iterative duplication and divergence events. In each iteration, a randomly selected existing node, v, is duplicated into a new node, u. This new node is connected to all of the

neighbours of v and may be connected to v itself with probability p. Divergence is achieved by considering all of the shared neighbours of u and v, and removing a connection with probability q (chosen to mimic the edge density of the real network) for either u or v.

- The stickiness-index model network (Sticky) (Pržulj and Higham, 2006) assumes that the higher the degree of two nodes, the higher the probability that they are neighbours. A Sticky network is generated starting from n disconnected nodes, to which randomly stickiness index values are assigned (proportional to the node degrees of the input real network). Then, the probability of connecting two nodes is equal to the product of their stickiness indexes.

- The popularity-similarity-optimization (PSO) model is a geometric model that aims to model how popular nodes tend to gain and lose in popularity over time and how similar nodes are typically more likely to be connected (Papadopoulos *et al.*, 2012). Nodes are generated one by one (simulating time) and placed on a hyperbolic disk at a given angular coordinate (i.e. the direction from the centre of the disk) and radial coordinate (i.e. the distance from the centre of the disk). Newly generated nodes are connected to already existing nodes based on a given probability that is controlled by temperature parameter T. When $T = 0$, newly generated nodes are likely to be connected to nearby nodes, leading to a clustered network structure. When T is 1, newly generated nodes are as likely to connect to nearby nodes as to nodes far away, leading to an unclustered network structure. The radial coordinate models popularity behaviour. This is because nodes are generated progressively more towards the outer rim of the disk. Therefore, by the end of the network generation process, the nodes that have been there the longest and thus had the most opportunities of gaining connections, are near near the centre of the disk. Additionally, the average distance to all other nodes in the disk is the lowest in the centre, so early generated nodes are also more probable to connect because they are more likely to be nearby another node (if $T < 1$). The angular coordinate models node similarity behaviour: similar nodes find themselves on the same direction on the disk and are more likely to be connected.

- The nonuniform popularity-similarity optimization (nPSO) model extends the PSO model to model community structure in a network (Muscoloni and Cannistraci,

2018). To do so, the angular coordinate of a newly generated node is sampled from a Gaussian mixture model with C components (replacing the uniform distribution used in the original PSO model), where C is the number of communities.

Comparing real-world networks to these models provides insight into their evolution and structure, as these model networks have known structural properties. For instance, the clustering coefficient an ER network is equal to its density, e.g. the probability that a pair of nodes in the network is connected. SF-BA networks a scale-free distribution and are robust to the random deletion of nodes. As the clustering coefficient of PPI networks is higher than that of an ER network of similar size and density, one can infer that nodes in PPI networks do not connect randomly. Furthermore, as PPI networks are shown to be scale-free, they are therefore assumed to be similarly robust to random node deletions like SF-BA networks (Vazquez *et al.*, 2004).

As a consequence, one of the first steps in the analysis of a real-world network is often to identify the model network it is the most similar in structure to, a process referred to as *model fitting*. For a given real-world network, the best fitting model network is the one to which it has the lowest topological distance. The current state-of-the-art network distance measure is *GCD11* (Yaveroğlu *et al.*, 2015), which measures the distance between a pair of networks as the euclidean distance between their GCM's (see Section 2.1.4) (Yaveroğlu *et al.*, 2014). Of key importance is the fact that GCD11 does not require the assumption of an underlying model network (as for real networks a proper model is often not known) and does not require the alignment of the nodes of the networks whose distance is being measured (which is computationally expensive) (Yaveroğlu *et al.*, 2017).

2.2 Spectral graph theory

I recall that adjacency of all nodes in network $H(V,E)$ is represented by an $n \times n$ symmetric adjacency matrix A, where the entries A_{uv} are 1 if nodes $\{u,v\} \in V$ are connected by an edge (i.e. are 'adjacent') and 0 otherwise. A related network representation is the *Laplacian matrix*, which is defined as $\mathscr{L} = D - A$, where D is the diagonal matrix such that D_{uu} is equal to the degree of node u. Both the adjacency matrix and the Laplacian matrix are *diagonalisable*, i.e. for each of them there exists a basis such that when multiplying them by this basis they are diagonal. This follows from the Spectral Theorem:

Theorem 1 (Spectral Theorem)**.** *If M is an $n \times n$ symmetric matrix, then there exists an*

orthogonal matrix U and a diagonal matrix Λ such that:

$$MU = \Lambda U. \tag{2.5}$$

The values on the diagonal of Λ are called the *eigenvalues* of M. The column vectors of U are known as the *eigenvectors* of M. Computing the eigenvectors and eigenvalues of a matrix is known as performing *eigendecomposition*.

It turns out that the eigenvectors and eigenvalues of both the adjacency matrix and Laplacian matrix capture many different topological properties of the underlying network H. The study of relating the eigenvectors or eigenvalues of the Laplacian or adjacency matrix of a given graph to certain topological properties is known as *spectral graph theory* (see (Chung and Graham, 1997) for an extensive overview to the topic). Here, I present different network analysis tools that make use of spectral theory and that have been applied throughout this thesis: spectral clustering (Section 2.2.1), spectral diffusion (Section 2.2.2), spectral embedding (2.2.3) and eigencentrality (2.2.4).

2.2.1 Spectral clustering

Spectral clustering refers to the class of algorithms that identify groups of densely connected nodes in a network based on the eigendecomposition of its Laplacian matrix (Von Luxburg, 2007). This is because the eigenvectors of the Laplacian naturally uncover the clusters present in the network. Here, I will provide an intuition to why this is the case. Additionally, I will explain how spectral clustering relates to *graph-cut* algorithms, which identify groups of densely connected nodes in a network by cutting the network into increasingly smaller densely connected components. For a more rigorous yet accessible introduction to the subject, I refer the reader to (Vidal *et al.*, 2016).

Assume a network H, consisting out of k components. In this case, the network's corresponding Laplacian matrix will have k eigenvectors in its null space (i.e. there exist k eigenvectors corresponding to the eigenvalue 0). These k vectors are *cluster indicator* vectors, assigning each the n nodes of the network to one of the k components. This property is illustrated in Figure 2.4. When small perturbations make the network connected, as illustrated in Figure 2.5, it can be shown that the k eigenvectors corresponding to the smallest non-zero eigenvalues can still uncover the "true" underlying clustering of the network (Davis and Kahan, 1970). In practice, K-means clustering is applied on the first k eigenvectors to extract a clustering of the nodes, see Algorithm 1. This algorithm is

$$H = \quad U = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{matrix} A \\ B \\ C \\ D \\ E \\ F \end{matrix}$$

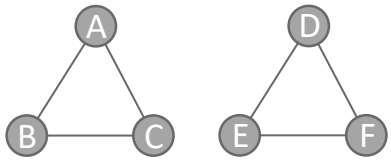


Figure 2.4: Laplacian eigenvectors capture the component structure of a disconnected network. Left: a disconnected dummy network H. Right: the first two eigenvectors of the Laplacian of H. The first (second) eigenvector assigns nodes A, B, C, (D, E, F) to the same cluster, capturing the fact that they form a disconnected component together.

$$H' = \quad U = \begin{bmatrix} 0.5 & 0.3 & 0.2 \\ 0.5 & 0.5 & 0 \\ 0.5 & 0.5 & 0 \\ 0.5 & 0.2 & 0.3 \\ 0.5 & 0 & 0.5 \\ 0.5 & 0 & 0.5 \end{bmatrix} \begin{matrix} A \\ B \\ C \\ D \\ E \\ F \end{matrix}$$

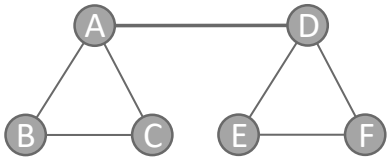


Figure 2.5: Laplacian eigenvectors capture the clustering structure of a connected network. Left: a connected dummy network H'. H' is a lightly perturbed version of the network H in Figure 2.4, with an edge added between nodes A and D. Right: the first three eigenvectors of the Laplacian of H'. The first eigenvector assigns all nodes to the same component, i.e. the fully connected network. The second (third) eigenvector assigns nodes A, B, C, (D, E, F) to the same cluster, capturing the original component structure of H, despite the small perturbation.

what is generally referred to when spectral clustering. Applying spectral clustering using the symmetrically normalized Laplacian is known as *normalized spectral clustering*.

Spectral clustering can also be considered as a graph-cut algorithm. The *minimum-cut problem* considers clustering the nodes of a graph by cutting the graph into d graph partitions whilst minimizing the number of edges being cut. However, in many networks often the cheapest cuts would be to remove the individual nodes that are only connected to the network by a single edge, resulting in clusters containing only a single node. To achieve a more balanced clustering, *ratio-cut problem* normalises each graph cut by the number of nodes in each cut (Hagen and Kahng, 1992). Alternatively, the *normalised-cut* problem normalises each cut by the degree of the nodes in the cut (Ng *et al.*, 2002). Spectral clustering and normalised spectral clustering respectively approximately solve the ratio-cut problem and the normalised cut problem. As such, next to basis vectors and cluster indicator vectors, the eigenvectors of the Laplacian can also be interpreted as graph-cuts.

Algorithm 1 Spectral clustering

Input Number of clusters, d, the Laplacian matrix, $\mathscr{L}$, of an network, H, with n nodes.

Output d clusters of the n nodes of H.

1: Compute the d eigenvectors of $\mathscr{L}$ associated with its d smallest eigenvalues: $U = [\mathbf{u}_1, \ldots, \mathbf{u}_d] \in \mathbb{R}^{n \times d}$.

2: Normalize U so that each column has unit norm.

3: Cluster the n points $\{\mathbf{u}\}_{u=1}^{n}$ into d groups using K-means.

2.2.2 Network diffusion

Network diffusion refers to a family of related techniques, which propagate information on nodes through the network. Here I focus on the diffusion kernel. The diffusion kernel is often called the 'heat kernel', as it can be viewed as describing the flow of heat originating from the nodes across the edges of a graph with time. In network biology nodes typically represent genes and 'heat' on a node represents experimental measurements. For a set of n nodes these measurements are encoded in vector $P_0 \in \mathbb{R}^n$. Information is diffused as follows: $P = HP_0$, where H is a diffusion kernel. The diffusion kernel, H_α^k, is defined as the matrix exponential of the Laplacian matrix (Kondor and Lafferty, 2002):

$$H_\alpha = e^{-\alpha \mathscr{L}}, \tag{2.6}$$

where the parameter $\alpha \in \mathbb{R}$ controls the level of diffusion. H_α can be computed as follows:

$$H_\alpha = \sum_{i=1}^{n} u_i e^{\alpha \lambda_i} u_i^T, \tag{2.7}$$

where u_i is the I-th normalized eigenvector of the Laplacian and λ_i its corresponding eigenvalue.

Many popular bioinformatics software packages are based on network diffusion. For instance, given a molecular network and a set of genes differentially expressed in disease, HotNet identifies significantly altered subnetworks in disease (Reyna *et al.*, 2018). Given a molecular network and a set of genes of interest, GeneMANIA identifies genes in the network that are related (i.e., intertwined) with the genes of interest (Franz *et al.*, 2018).

2.2.3 Spectral embedding

Spectral Embedding projects a network in a low-dimensional space, placing nodes closely together in that latent space if they are part of the same network neighbourhood. Here, I present the Laplacian Eigenmap embedding algorithm (Belkin and Niyogi, 2003) so that two nodes are embedded close in space if they frequently share the same neighbour. Given an unweighted network G with n nodes, a low dimensional embedding, $Y = [\mathbf{y}_1, ..., \mathbf{y}_n] \in \mathbb{R}^{d \times n}$, is computed so that if nodes u and v are adjacent, then $\mathbf{y}(u)$ and $\mathbf{y}(v)$ are close in the d-dimensional space by solving:

$$\begin{aligned} \underset{Y}{\text{minimize}} \quad & \sum_{u=1}^{n} \sum_{v=1}^{n} A(u,v) \left\| \mathbf{y}_u - \mathbf{y}_v \right\|^2 \\ \text{subject to :} \quad & YD\mathbf{1} = \mathbf{0} \text{ and } YDY^T = I, \end{aligned} \tag{2.8}$$

where A is the adjacency matrix of G and D is the degree matrix of G. The columns of Y are found as the generalized eigenvectors associated with the second [1] to (d+1) smallest generalized eigenvalues solving $Y\mathscr{L} = \Lambda YD$, where Λ is the diagonal matrix with the generalized eigenvalues along its diagonal.

2.2.4 Eigencentrality

I recall that the *eigencentrality* is a more sophisticated version of the degree centrality, which considers the most important nodes to be the nodes that are highly connected to other highly connected nodes in the network. Formally, the eigencentrality of a node $u \in V$, $EC(u)$, is defined as the average of the eigencentralities of its n neighbours:

$$EC(u) = \frac{1}{\lambda} \sum_{v=1}^{n} A_{uv} EC(v), \tag{2.9}$$

where λ is a constant. This equation in matrix form:

$$A\mathbf{c} = \lambda \mathbf{c}, \tag{2.10}$$

where $\mathbf{c}$ is the vector of centralities, $\mathbf{c} = (c_1, c_2, \dots)$. From this it is clear that $\mathbf{c}$ is an eigenvector of A. Additionally, for c to be interpretable as a centrality measure, it is necessary for all entries to be non-negative. Per the 'Perron–Frobenius theorem', it can be proven that the eigenvector corresponding to the largest eigenvalue contains only non-negative entries.

[1] Assuming the network is one large connected component.

It should be noted that the eigencentrality is not a useful measure on direct networks. The underlying cause is the fact that eigencentrality of nodes having only outgoing connections would be zero. This is problematic as nodes with (potentially many) incoming edges coming solely from nodes with a centrality of zero, would also have a centrality of zero. To solve this issue, the Katz-centrality (KC) adds some 'free' centrality, β_u, to each one of the nodes: $KC(u) = \alpha \sum_{v=1}^{n} A_{uv} KC(V) + \beta_u$. Another related extension for directed networks is the Page-rank-centrality (PR). To take into account how important a given node is from the perspective of its neighbours, the centrality of its neighbours is weighted by that neighbours number of outgoing edges: $PR(u) = \alpha \sum_{v=1}^{n} A_{uv} PR(v)/d_v^{out} + \beta_u$.

2.2.5 Laplacian alternatives: k-path Laplacian and Vicus

The standard Laplacian captures direct connectivity between nodes in a network. To capture the influence of long-range interactions between nodes, Estrada (2012) proposed the *k-path Laplacian* by generalizing the concepts of adjacency and degree. The k-path Laplacian defines a pair of nodes u and v to be *k-adjacent* if the shortest path distance between them is equal to k. Analogously, *k-path degree*, $\deg_k(u)$, generalizes the concept of the degree to the number of length k shortest paths that have node u as an endpoint. The k-path Laplacian, $\mathscr{L}_k^P$, is defined as:

$$\mathscr{L}_k^P(u,v) = \begin{cases} -1 & \text{if } d(u,v) = k \\ \deg_k(u) & \text{if } u = v \\ 0 & \text{otherwise.} \end{cases} \tag{2.11}$$

Vicus is an alternative to the Laplacian that captures the intricacies of a network's local structure (Wang *et al.*, 2017) based on network label diffusion. Label diffusion is defined as $P = BQ$, where the $n \times d$ matrix Q assigns the n nodes of network G to one of d possible labels (for labeled nodes), B is an $n \times n$ diffusion matrix, and the reconstructed matrix P is an $n \times d$ matrix used for predicting labels for unlabeled nodes. To give Vicus its 'local' interpretation, the label diffusion process determining B is constrained to diffuse information of each node only to its direct neighbourhood (see next paragraph). Under given assumptions and defining Vicus as $\mathscr{L}^V = (I - B^T)(I - B)$, it was shown that Q can be learned as the eigenvectors of $\mathscr{L}^V$. As Q captures the local connectivity between nodes that is implied by the 'localized' diffusion matrix B and can be computed as the eigenvectors of $\mathscr{L}^V$, Vicus is interpreted as a Laplacian matrix.

The label diffusion matrix, B, is defined as follows. The authors of Vicus modified the diffusion process slightly from the original, in the sense that diffusion is limited to the direct neighborhood of each node to give Vicus its 'local' characteristics (instead of the diffusion process being applied over the entire graph at once). Formally, they define for each node u a local $K \times K$ adjacency matrix W_u, which is the sub-matrix of the adjacency matrix A limited to u and its $K-1$ neighbours, $N(u)$. Then, S_u is defined as the row-normalized transition matrix of W_u:

$$S_u(v,t) = \frac{W_u(v,t)}{\sum_{l=1}^{K} W_u(v,l)}. \tag{2.12}$$

Matrix β_u encodes the label diffusion for u and its direct neighbours $\beta_u = (1-\alpha)(I - \alpha S_u)^{-1}$, where α controls the level of label diffusion. Label diffusion matrix, B, of G is defined as:

$$B(u,v) = \begin{cases} \frac{\beta_u(u,v)}{1-\beta_u(u,u)} & \text{if } v \in N(u) \\ 0 & \text{otherwise.} \end{cases} \tag{2.13}$$

Vicus has been applied to protein module discovery and ranking of genes for cancer subtyping (Wang *et al.*, 2017).

2.3 The different types of data-integration methods

In recent years rapid biological advancements have flooded biology with new large-scale data on different biological entities (genes, patients, drugs, etc.) and the relations between them (Gligorijević *et al.*, 2016a). As these data provide complementary views on the same biology, data integration methods have been proposed in order to simultaneously mine these heterogeneous data and provide system levels insight not achievable from analysis of these network data individually. Here, I briefly describe the underlying principles of key machine learning approaches used to integrate biological data, including: Network-based data integration (Section 2.3.1), Bayesian data integration (Section 2.3.2), Kernel-based data integration (Section 2.3.3), Neural network based approaches (Section **??**). However, the core of this section focusses on non-negative matrix factorization based approaches (Section 2.3.4).

2.3.1 Network-based data integration

Network-based methods offer simple ways to integrate different network data. For homogeneous integration of d different networks $H_1 = (V,E_1)$, $H_2 = (V,E_2)$, $\ldots$, $H_d = (V,E_d)$,

that share the same set of nodes, V, but have different sets of edges ($E_i, 1 \leq i \leq d$) connecting their nodes, the simplest network based data integration consist of taking the union of all networks, i.e. $H_{union} = (V, \bigcup_i E_i)$ (Dutkowski *et al.*, 2013) (illustrated in Figure 2.6). However, this simple approach does not consider varying noisiness and incompleteness of the different datasets, as all input networks contribute equally to the integrated model.

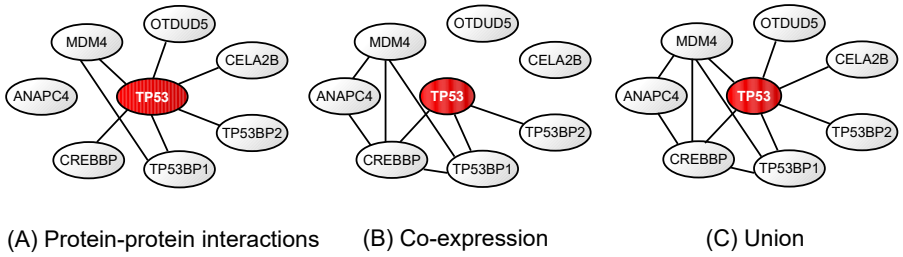


Figure 2.6: Illustration of network-based integration. Part of the local interactions arround protein TP53 (a tumor suppressor protein that regulates cell division by preventing cells from growing and dividing in an uncontrolled way), as taken from STRING database (Szklarczyk *et al.*, 2014). **A**: The protein-protein interactions between the proteins. **B**: The proteins that are co-expressed. **C**: The union of the two networks presented.

To overcome this limitation, another approach consists of considering the adjacency matrices of each network, $A_i, 1 \leq i \leq d$, and combining them using a linear combination: $A_{union} = \sum_i w_i A_i$, where $w_i \geq 0$ is the weight associated to network N_i so that the quality of the integrated model is optimized (Mostafavi *et al.*, 2008; Chen *et al.*, 2013). Finding weights w_i requires solving a system of linear equations, which will assign lower weights to "less contributing" networks. More advanced methods use message passing theory (Pearl, 2014) to iteratively update the input networks, making them more similar to each other after each iteration, until they converge to a single, integrated model (Wang *et al.*, 2014a).

In heterogeneous network integration, nodes and edges of different types are joint into a single large network. Earlier network based approaches for heterogeneous integration (e.g. of networks having different sets of nodes and edges) often requires a preliminary step in which all networks are projected on a set of common nodes (Davis and Chawla, 2011; Sun *et al.*, 2014), resulting in information loss. More advanced approaches, called *network propagation* methods, can directly integrate heterogeneous data using diffusion processes that spread information along the edges of the networks (Guo *et al.*, 2011; Huang *et al.*, 2013).

Recently, heterogeneous network integration has seen a revival as part of deep learn-

ing frameworks. In this context, heterogeneous networks have been rebranded as *knowledge graphs*, so called as they represent all the knowledge of a given domain (Ehrlinger and Wöß, 2016). For instance, Decagon is a graph convolutional neural network trained to predict drug combinations' side effects (Zitnik *et al.*, 2018). It is trained on a knowledge graph (i.e., a single large heterogeneous network) of protein–protein interactions, drug–protein target interactions and drug–drug interactions (i.e., polypharmacy side effects).

2.3.2 Bayesian approaches

A Bayesian network is a probabilistic graphical model that represents a set of random variables and their conditional dependencies. A Bayesian network is based on a directed acyclic graph (DAG, see chapter 3) in which nodes represent variables and in which a directed edge from node y to node x represents the conditional probability between x and y. The conditional probability between x and y is denoted by $p(x|y)$ and is the probability of x given the value of y. A Joint Probability Distribution (JPD) of a Bayesian network having n nodes, $x = \{x_1, x_2, \ldots, x_n\}$, is:

$$p(x|\theta) = \prod_{i=1}^{n} p(x_i|Pa(x_i)), \tag{2.14}$$

where $Pa(x_i)$ are the ancestor of x_i in the Bayesian network, and $\theta = \{\theta_1, \theta_2, \ldots, \theta_n\}$ are the model's parameters that define the JPD. Not only does the Bayesian network capture the structure of the data, but also its sparsity also eases the computation of the JPD over the whole set of random variables. I.e., the number of parameters that are needed to characterize the JPD is reduced in the Bayesian network representation (Needham *et al.*, 2007; Ben-Gal, 2008).

Constructing a Bayesian networks requires (1) learning the network's wiring patterns, which is called *structure learning*, and (2) learning the parameters of its JPD, which is called *parameter learning* (Needham *et al.*, 2007; Ben-Gal, 2008). Structure learning consist of identifying the statistical dependencies (represented by edges in the DAG) between the variables. Because the number of possible wirings in a network is super-exponential in its number of nodes, learning the Bayesian network that best represents a dataset is an NP-hard problem for which heuristic algorithms have been proposed (Needham *et al.*, 2007). Once the structure and the parameters of a Bayesian network have been learned, it can be used for inference (prediction) about its variables. For ex-

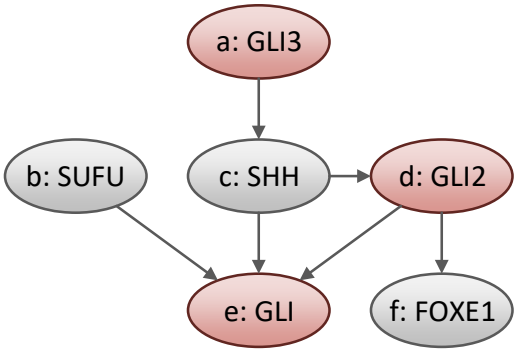


Figure 2.7: Bayesian Network: example of a gene regulatory network. The presented Bayesian network encodes part of the regulatory relations of GLI proteins (in pink), which are transcription factors whose mutations are involved in many congenital malformations. In this sparse representation, the expression of a gene depends only on its parents: e.g. the expression of SHH (c) only depends on GLI3 (a), so the conditional probability distribution of c is $p(c|a)$. The joint probability distribution of the Bayesian network is: $p(a,b,c,d,e,f) = p(a)p(b)p(c|a)p(d|c)p(e|b,c,d)p(f|d)$.

ample, with a Baysian network representing the regulatory interactions between genes (as illustrated in Figure 2.7), one can ask for the likelihood of a gene to be expressed given the expression status of the other genes. Note that exact inference, which requires the summation of the JPD over all possible values of unknown variables, is also an NP-hard problem (Cooper, 1990) for which approximate solutions have been proposed (Ben-Gal, 2008).

While Bayesian networks have been successfully used to integrate biological data (e.g. for gene regulatory network inference (Zhu *et al.*, 2008), or for cancer prognosis prediction (Gevaert *et al.*, 2006)), they suffer from the following limitations. First, in the structure learning, their sparse representation only captures the most important associations between the variables, discarding all other weaker associations. Second, the directed acyclic graph representation does not allow for loops, which are important components of biological networks (e.g. for representing control and feed-back loops). Finally, as already mentioned, their computational complexity limits the usage of Bayesian networks to small datasets.

2.3.3 Kernel-based methods

Kernel-based methods are machine learning approaches for pattern analysis. A kernel-based approach works by embedding the original data from its original input space, $\mathscr{X}$, into a higher dimensional space, called the feature space, $\mathscr{F}$, in which the analysis is performed (see illustration in Figure 2.8).

$\mathscr{F}$ is a vector space in which data points are represented by vectors, called *fea-*

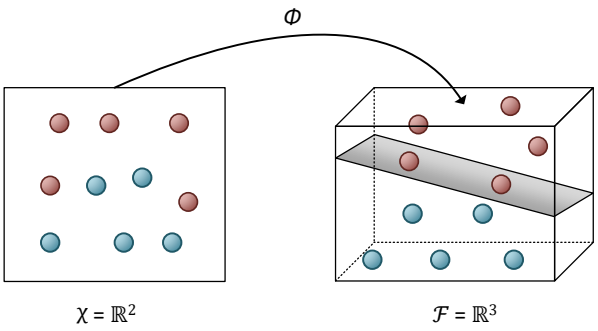


Figure 2.8: Illustration of kernel-based methods. To simplify a classification problem of datapoints in input space $\mathscr{X} = \mathbb{R}^2$, a kernel function ϕ is used to map $\mathscr{X}$ into a higher dimensional feature space $\mathscr{F} = \mathbb{R}^3$, in which the two clusters of datapoints are easily separated by a plane.

ture vectors. The embedding of $\mathscr{X}$ in $\mathscr{F}$ is represented by a *kernel matrix* (Schölkopf *et al.*, 2004), $\mathscr{K}$, which is a symmetric, positive semi-definite matrix whose entries $\mathscr{K}_{i,j} = k(x_i, x_j)$ represent the similarities between any two data points x_i and x_j, which are computed as the inner product between their representations $\phi(x_i)$ and $\phi(x_j)$ in the feature space:

$$k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle, \tag{2.15}$$

where ϕ maps data points from $\mathscr{X}$ to $\mathscr{F}$ (Schölkopf *et al.*, 2004; Borgwardt, 2011). In practice, only the definition of the *kernel function* $k(x_i, x_j)$ is required.

In kernel-based approaches, a network is frequently represented by using a diffusion kernel (see Section 2.2.2), where entries of the diffusion kernel quantify the closeness between any two nodes in the network. Alternatively, in relation to the graphlet based descriptors presented in Section 2.1.4, a network can be represented by a *graphlet kernel* (Vacic *et al.*, 2010), in which a node x_i of the network is represented by its graphlet degree vector, $GDV(x_i)$ in the feature space $\mathscr{F}$. The kernel matrix is then computed from the following graphlet-based kernel function:

$$k(x_i, x_j) = \langle GDV(x_i), GDV(x_j) \rangle, \tag{2.16}$$

To integrate multiple network datasets, *kernel data-fusion* (Yu *et al.*, 2013) consists of representing all network data in the same feature space and in linearly combining the corresponding kernel matrices before analysis. Kernel matrices are then mined using traditional statistical and machine learning methods, such as support vector machines (SVMs) (Vapnik and Vapnik, 1998), principal component analysis (PCA) (Jolliffe, 1986)

and canonical correlation analysis (CCA) (Hardoon *et al.*, 2004) to produce clusterings, rankings, principal components and correlations.

While kernel-based approaches have been used to integrate biological data (e.g. for cancer prognosis prediction (Daemen *et al.*, 2007), or for drug repurposing (Napolitano *et al.*, 2013)), they suffer from the following limitations. First, there is no guideline for choosing the right kernel function to best represent a given dataset. This short-coming is partially overcome by *multiple kernel learning* approaches, in which one linearly combines several kernel representations of the same dataset, each capturing different notions of similarity between the nodes of the network (Wang *et al.*, 2014b). Second, kernel data-fusion implies to transform, or project all network datasets into the same feature space, which may results in information loss.

2.3.4 Data integration with non-negative matrix factorization

Non-negative matrix factorization (NMF) is a machine learning method for clustering and dimensionality reduction. In NMF, a network is represented by a non-negative matrix, $A \in \mathbb{R}^{n_1 \times n_2}$, the adjacency matrix of the network. As illustrated in Figure 2.9, this matrix is approximated by the product of two lower-dimensional, non-negative matrix factors, $U \in \mathbb{R}^{n_1 \times k}$ and $V \in \mathbb{R}^{k \times n_2}$, where $k \ll min(n_1, n_2)$ (Lee and Seung, 1999):

$$A \simeq UV. \tag{2.17}$$

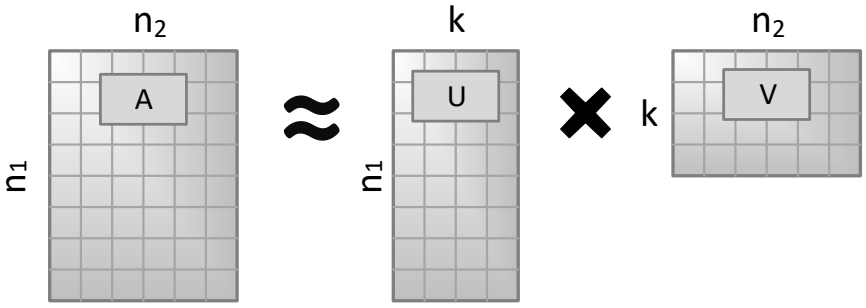


Figure 2.9: Illustration of NMF. In NMF, $n_1 \times n_2$-dimensional matrix A (e.g. the adjacency matrix of a dataset) is decomposed into the product of two lower dimensional matrix factors, U and V, where $k \ll min(n_1, n_2)$.

In NMF, setting the *rank parameter* $k \ll \min(n_1, n_2)$ provides dimensionality reduction (Cichocki *et al.*, 2009). NMF gained particular interest because of its relationship with k-means clustering (Ding *et al.*, 2005). From a clustering point of view, the non-negative matrix A represents n_1 datapoints (e.g. n_1 patients) by their n_2 dimensional feature vectors (e.g, for each patient, a vector representing the expression levels of n_2 genes),

this matrix is factorized into two matrices U and V, where U is the cluster indicator matrix, which assigns the n_1 datapoints into k clusters, and where V represents the cluster centroids (Ding *et al.*, 2005). (From a dimensionality reduction point of view, V is also called a base matrix, as it represents the k-dimensional space defined by the cluster centroids.) Note that the role of U and V can be exchanged, making V the cluster indicator matrix assigning the n_2 datapoints (e.g. genes) into k clusters, and making U represent the cluster centroids. Extracting clusters from V can be done with a procedure such as "hard clustering" (Zass and Shashua, 2005), in which each data point i is assigned to cluster j, $1 \leq j \leq k$, such that $U_{i,j}$ is the maximum value in row i (e.g. $j = argmax_{j=1}^{k} U_{i,j}$). Note that to improve clustering interpretation, one can add an orthogonality constraint on a cluster indicator matrix (e.g. for the n_1 datapoints, $U^T U = I$, where I is the identity matrix), resulting in so-called *Orthogonal NMF* (Ding *et al.*, 2006). It can happen that some datapoints (features in this context) do not contribute to the clustering. For instance, genes being expressed across all patients are uninformative when clustering patients based on gene expression. To identify the relevant features, a useful strategy is to compute the mutual information (between each input feature and the matrix factor (Kim and Park, 2007).

Apart from clustering, another important property of NMF is the *completion property*. Namely, after solving NMF, the reconstructed matrix $\hat{A} = UV$ features new entries, not observed in A, but emerging from the latent factors. (As opposed to observable factors, latent factors are factors that are not directly observed but are rather inferred.) Here, U and V are latent representations for dimension n_1 and n_2 respectively (e.g. patients and genes).

NMF was extended into Non-negative Matrix Tri-Factorization (NMTF) (Ding *et al.*, 2006) (illustrated in Figure 2.10). In NMTF, an $n_1 \times n_2$ matrix, A, which describes the relationships between two types of objects (e.g. the relationships between n_1 patients and their n_2 genes), is decomposed as the products of three non-negative, lower dimensional matrix factors:

$$A \simeq USV^T, \tag{2.18}$$

where $U \in \mathbb{R}^{n_1 \times k_1}$ is the cluster indicator matrix of the objects of type 1 (grouping the n_1 objects of type 1 into k_1 clusters), $V \in \mathbb{R}^{n_2 \times k_2}$ is the cluster indicator matrix of the objects of type two (grouping the n_2 objects of type 2 into k_2 clusters), and where $S \in \mathbb{R}^{k_1 \times k_2}$ is the compressed representation of A that relates the clusters in U to the clusters in V.

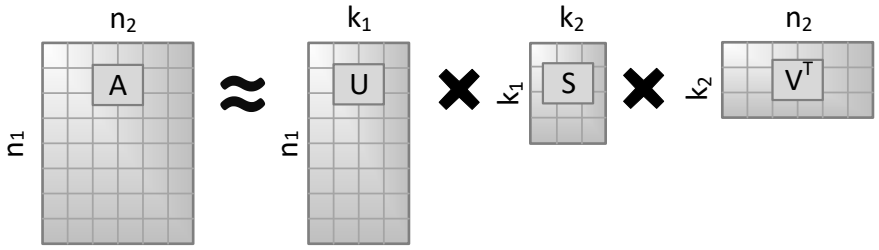


Figure 2.10: Illustration of NMTF. In NMTF, Matrix A (e.g. the adjacency matrix of a dataset) is decomposed as the product of three low-dimensional matrix factors, U, S and V, where $k_1, k_2 \ll min(n_1, n_2)$.

NMTF is a *co-clustering* approach, which means that one can extract from U clusters that group together objects of type 1 according to their relationships with objects of type 2, while extracting from V clusters that group together objects of type 2 according to their relationships with the objects of type 1. Such clusters can be extracted from U and V using the same hard-clustering procedure as described for NMF. Similar to NMF, NMTF can be used for matrix completion: after decomposition, the reconstructed matrix $\hat{A} = USV^T$ features new entries, not observed in data matrix A, which can be used for prediction.

2.3.5 Heterogeneous data integration with NMTF

Similar to NMF, NMTF can use both simultaneous decompositions and graph regularization penalties to integrate homogeneous and heterogeneous datasets. Given the NMTF of a single matrix $A_{1,2}$ representing the relationships between objects of types 1 and 2, which is decomposed as the product of three matrix factors: $A_{1,2} \simeq G_1 S_{1,2} G_2^T$, any new dataset can be simultaneously decomposed with $A_{1,2}$ as long as it relates to at least one of the object types that is already in the decomposition. E.g., $A_{2,3}$, which represents the relationships between objects of types 2 and 3, can be simultaneously decomposed with $A_{1,2}$ as $A_{2,3} \simeq G_2 S_{2,3} G_3^T$, while sharing the cluster indicator matrix of the objects of types 2, G_2, across the two decompositions. In this way, NMTF can be used to simultaneously decompose any combination of datasets. Furthermore, graph regularization penalties can be added so that each cluster indicator matrix, G_i, can benefit from the prior knowledge encoded in a network between nodes of type i (represented by Laplacian matrix $\mathscr{L}_i$). A generic formulation of this problem is:

$$f = \min_{G_i \geq 0 \forall i, S_{i,j} \geq 0 \forall i,j} \sum_{i \neq j} ||A_{i,j} - G_i S_{i,j} G_j^T||_F^2 + \sum_i \alpha_i tr(G_i^T \mathscr{L}_i G_i) \tag{2.19}$$

Thus, NMTF provide a principled framework for data integration. For instance, as illustrated in Figure 2.11, NMTF has been used in the patient specific data-fusion (PSDF)

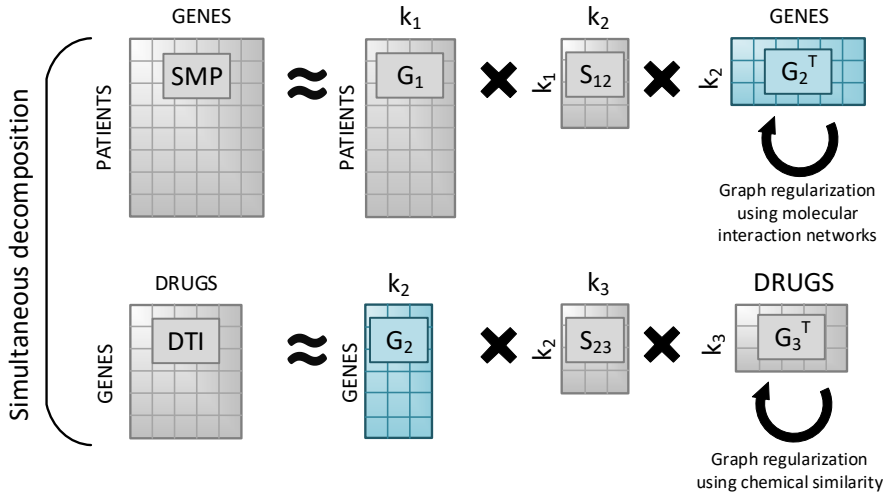


Figure 2.11: PDF as an example of heterogeneous integration with NMTF. In the patient specific data-fusion (PDF) framework, somatic mutation profiles (XMP) and drug target interactions (TI) are simultaneously decomposed as the product of three matrix factors. Both decompositions share the same cluster indicator matrix of genes, G_2 (in blue), which allows learning from all datasets. The cluster indicator matrix, G_2, is constrained by a graph regularization penalty to favour grouping together genes that interact in a molecular interaction network. Similarly, the cluster indicator matrix G_3 is constrained to favour grouping together drugs that are chemically similar.

framework (Gligorijević *et al.*, 2016b) in the context of cancer precision medicine. In PSDF, somatic mutation profiles (that represent relationships between patients and their mutated genes) and drug target interactions (that represent relationships between genes and the drugs that target their protein products) are simultaneously decomposed while sharing the cluster indicator matrix of genes (G_2 in Figure 2.11). PSDF uses graph regularization constraints so that the cluster indicator matrix of genes favors grouping together genes that interact in a molecular interaction network and that the cluster indicator matrix of the drugs favors grouping together drugs that are chemically similar. On serous ovarian cancer patients' data, the patient specific data-fusion framework allows for simultaneously uncovering patient subtypes having statistically significantly different disease outcomes, predicting novel cancer-related genes and predicting drugs for potential drug-repurposings.

2.3.6 Homogeneous data integration with NMTF

Like NMF, NMTF can be used for data integration of different types of networks that share the same set of nodes. This approach is taken in the iCell project, a side project on which I collaborated during my PhD (Malod-Dognin *et al.*, 2019b). In this project, we integrate molecular network data for a healthy and cancerous cell to learn a latent network representation for both cell states. We call such a latent network an integrated

cell or 'iCell'. Intuitively, by fusing the different types of molecular data, we aim to learn a latent network representation that better captures the cell's functional organisation than the constituent molecular networks individually. We identify cancer driver genes as those genes that have the biggest change in wiring between both iCells.

To create a pair of iCells, we first collect gene expression data for a given tissue type in a healthy and a diseased state and generic molecular network data: protein-protein interaction (PPI) networks, co-expression (COEX) networks and genetic interaction (GI) networks. Then, to create tissue and state-specific molecular networks, we take the sub-networks induced by the gene expression data on three types of networks. Next, for both tissue states, NMTF is applied to integrate these network data, as illustrated in Figure 2.12. Finally, to extract a latent network representation, i.e., an iCell, we interpret the cluster indicator matrix G_1 as an embedding of the genes in a latent space that captures the molecular organisation of the tissue (an assumption we confirm through cluster and enrichment analysis). For both tissue states, we extract a k-nearest neighbour network from the latent space, measuring the distance between all genes in the latent space as $D = G_1 G_1^T$, and connecting each gene with its 100 nearest neighbours in the latent space.

To identify cancer-related genes, we first observe that they tend to be always expressed, i.e., both when the cell is in a healthy and a diseased state. Then, for the always expressed genes, we quantify their wiring in both iCells using the graphlet degree vector (GDV) (see Section 2.1.4). Finally, we identify cancer-related genes as the always expressed genes with the biggest euclidean distance between their GDVs in a healthy and cancerous iCell. For details, see (Malod-Dognin *et al.*, 2019b).

I build on these results in chapter 5. There, I integrate pathway data and molecular network data to learn a latent representation of a healthy and diseased cell. Analogous to the iCell project, I first show that by integrating both types of data I can learn a latent representation that better captures the functional organisation of the healthy and diseased cell. Further, I identify cancer related pathways and genes by comparing the latent representation of the healthy and diseased cell. In contrast to the iCell project, however, I do this based on geometric properties of the learned latent space, rather than explicitly creating latent network representations for each cell state and computing changes in wiring patterns.

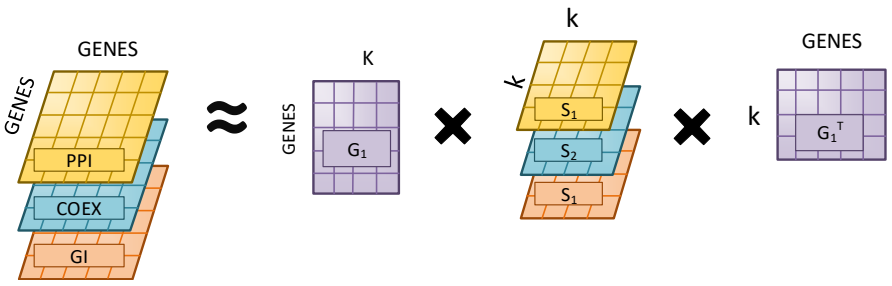


Figure 2.12: iCell as an example of Homogeneous integration with NMTF. NMTF is applied to integrate PPI, COEX and GI data. The cluster indicator matrix G_1 is interpreted as an embedding of the genes in the combined latent space spanned by $\{S_1, S_2, S_3\}$, the latent representations of the individual PPI, COEX and GI networks. An integrated latent network representation for the underlying tissue is created by computing the pairwise cosine distances between all genes in the embedding space spanned by G_1 and connecting each gene to its k nearest neighbours.

2.4 Pathway focused approaches to study cancer disease mechanism

Cancer is a genetic disease in which the accumulation of genetic mutations leads to the uncontrolled proliferation of tumour cells (Vogelstein and Kinzler, 2004; Vogelstein *et al.*, 2013). Specifically, mutations to cancer driver genes lead to the reprogramming of cellular *pathways*: functional subnetworks within the cell that once activated lead to a certain product, or a change within the cell through a series of consecutive interactions (Vogelstein and Kinzler, 2004; DeBerardinis and Chandel, 2016). This causes the cell to gain and lose functions that enable tumour growth and metastatic dissemination, such as gaining the ability to sustain proliferative signalling and resisting cell death, whilst losing the ability to respond to growth suppressors (Hanahan and Weinberg, 2011). To gain insight into the mechanisms underlying cancer, often pathway-based methods are considered, as they provide functional context to the observed gene mutations. This, in turn, helps to generate testable hypotheses, to identify drug targets and to determine tumour subtypes (Creixell *et al.*, 2015). Furthermore, pathway-based approaches offer a higher level point of view to uncover functional changes in cancer than the gene level. For instance, clinically similar cancer patients could have different sets of mutated genes, but have similar perturbed pathways (Vogelstein *et al.*, 2013).

All network-based pathway-focused approaches identify cancer implicated as those most (internally) perturbed by cancer driven gene expression changes. One can however distinguish two major classes of network-based pathway-focused approaches for studying cancer. *Pathway-topology-based* (PTB) measure the impact of gene expression changes

on a given pathway taking the topological importance of the genes to the pathway into account. Intuitively, if a gene has many interactions in the pathway (i.e. is topologically important), it is assumed to be important to the pathways normal functioning (i.e. is functionally important). So, changes in gene's expression should have a larger or lesser impact on a pathway's perturbation score dependent on its topological importance. As PTB methods consider pathways in isolation, i.e. gene perturbations outside the pathway do not affect its score, and current pathway annotation data is very incomplete, they are prone to producing large amounts of false negatives (Ogris *et al.*, 2017).

Crosstalk-enrichment (CE) methods acknowledge that pathways are part of a larger network. Given a large-scale network, CE methods prioritise pathways based on their association, i.e. *crosstalk*, with a set of differentially expressed genes. For instance, ANUBIX scores a given pathway by computing its edge-overlap with the subnetwork induced by a set of differentially expressed genes on the large-scale network and comparing this against the overlap that is expected by chance (Castresana-Aguirre and Sonnhammer, 2020).

2.5 Conclusion

In this chapter, I presented basic concepts from network analysis, spectral theory, data-integration applied machine learning methods and pathway-focused cancer analysis approaches. I highlighted that the Laplacian matrix, which underlies all spectral methods, captures the general connectivity structure of a network, i.e., what nodes are adjacent. Conversely, graphlets capture the wiring of a network around a node, but do not capture adjacency information. So, in Chapter 3, I combine both methods to define graphlet adjacency, which simultaneously captures connectivity information and topological information. In this background chapter, I also explained how pathway-focused approaches are often used to provide additional biological insight, as they are more descriptive. Hence, in Chapter 4, I take pathway focused approach to provide more biological insight into the results from chapter 3. Lastly, I also explained that methods aiming to identify cancer pathways do so by prioritising those pathways most internally perturbed in cancer. However, pathways do not operate in isolation in the cell. Therefore, in chapter 5, I integrate PPI data and pathway data using NMTF, to identify cancer pathways and genes, not based on their individual perturbation but on the extent their interactions with other pathways and genes change in cancer. There, graphlet adjacency is used to increase prediction accuracy.

Chapter 3

Graphlet Laplacian

The Laplacian matrix is the basis of many network integration and analysis frameworks, as it captures the global connectivity structure of the network (graph) (see Sections 2.2 and 2.3). All Laplacian based applications are based on the same underlying principle of *guilt by association*, inferring information on a given node based on its neighbours. However, graphlet-based approaches infer information on a given node based on the shape of its interaction pattern, typically independent of the identity of its neighbours (see Section 2.1.4). Therefore, to combine graphlet-based topological information and network neighbourhood information, I generalize the Laplacian to the Graphlet Laplacian by considering a pair of nodes to be 'adjacent' if they simultaneously touch a given graphlet. Below, through cluster enrichment analysis and diffusion of pan-cancer gene mutation scores, I demonstrate that Graphlet Laplacians capture topology-function and topology-disease relationships in biological networks. This analysis is summarised in Figure 3.1.

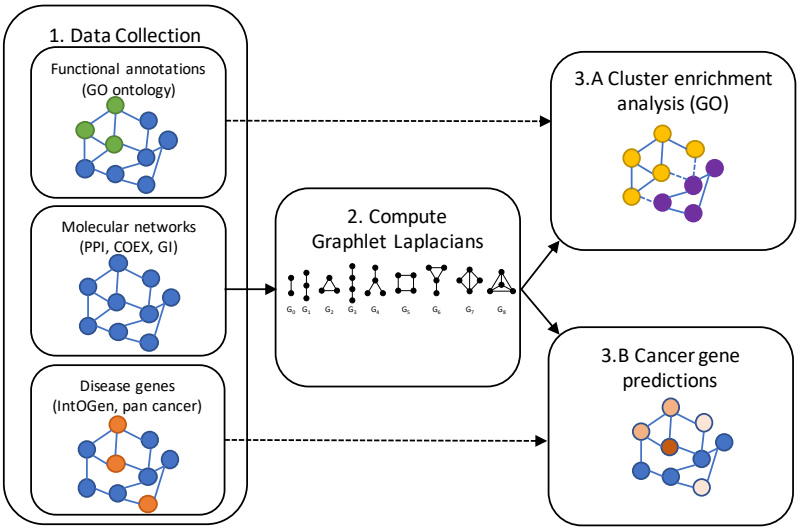


Figure 3.1: Chapter workflow summary. Step 1: collection of networks, GO annotations, cancer drivers and pan-cancer scores. Step 2: computation of graphlet laplacians for the different networks. Step 3.A: clustering of the graphlet laplacians. GO enrichment analysis to evaluate the function captured. Step 3.B: diffusion of pan-cancer scores.

Chapter impact

This chapter has lead to the following contributions:

Publications:

Windels, S. F. L., Malod-Dognin, N., and Pržulj, N. (2019). Graphlet Laplacians for topology-function and topology-disease relationships. *Bioinformatics*, **35**(24), 5226–5234.

Methodological contributions:

To capture the higher-order organisation of nodes in a network, I introduce *graphlet adjacency*, which considers two nodes connected based on their frequency of co-occurrence in a given graphlet (induced connected subgraphs of different 'shapes', such as paths and triangles). I use graphlet adjacency to generalise spectral methods spectral embedding and spectral clustering.

Biological contributions:

Graphlet adjacency provides complimentary views of the functional organisation of molecular networks. I show this for multiple types of molecular networks, species and functional annotations.

Software and data analysis:

The graphlet adjacency counter, as well as the data and scripts to perform the analysis presented in this chapter, have been made publicly available at: www.cs.ucl.ac.uk/staff/natasa/graphlet-laplacian .

3.1 Methods and data

3.1.1 Graphlet Laplacian definition

Here, I generalize the concept of the Laplacian (see Section 2.2) to that of a Graphlet Laplacian by generalizing the definitions of adjacency and degree to ones based on graphlets (see Section 2.1.4). First, I define two nodes u and v of G to be *graphlet-adjacent* with respect to a given graphlet, G_k, if they simultaneously touch G_k. For the example network presented in Figure 3.2-B, I find that nodes a and b are graphlet-adjacent w.r.t. graphlet G_1 twice, as G_1 can be induced on the dummy network twice: via paths a-b-c and a-b-e, each time including both nodes a and b. Similarly, nodes a and c and nodes a and e are graphlet adjacent only once, w.r.t. graphlet G_1. Given this extended definition of adjacency, I define the graphlet based adjacency matrix as:

$$A_k(u,v) = \begin{cases} a^k_{uv} & \text{if } u \neq v \\ 0 & \text{otherwise,} \end{cases} \tag{3.1}$$

where a^k_{uv} is equal to the number of times nodes u and v are graphlet-adjacent w.r.t graphlet G_k. Analogously, the *graphlet degree* generalizes the node degree as the number of times node u touches graphlet G_k. I extend the degree matrix to the *Graphlet Degree matrix* for graphlet G_k,

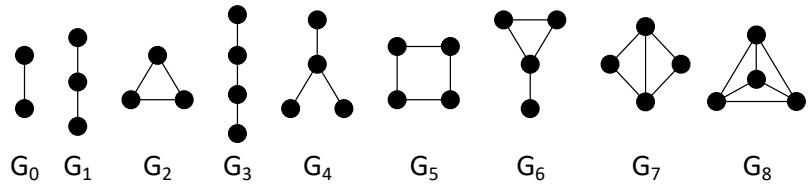
$$D_k(u,v) = \begin{cases} d^k_u & \text{if } u = v \\ 0 & \text{otherwise,} \end{cases} \tag{3.2}$$

where d^k_u is the number of times node u touches graphlet G_k. For an underlying graphlet G_k, I define the *Graphlet Laplacian* $\mathscr{L}^G_k$, as:

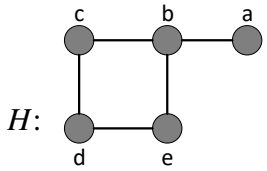
$$\mathscr{L}^G_k = D_k - (A_k/\theta). \tag{3.3}$$

where $\theta = size(G_k) - 1$. As opposed to the Laplacian simply capturing for each node its neighbours, the Graphlet Laplacian $\mathscr{L}^G_k$ captures for each node how strongly (i.e. frequently) each node is connected in the shape of G_k with each of the other nodes. $\mathscr{L}^G_0$ and $\mathscr{L}^G_1$ are illustrated in Figure 3.2-C. Finally, note that the Graphlet Laplacian for graphlet G_0, $\mathscr{L}^G_0$, is equivalent to the standard Laplacian, $\mathscr{L}$.

(A)



(B)



(C)

$$A_{G_0} = \begin{array}{c} \begin{matrix} a & b & c & d & e \end{matrix} \\ \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \end{bmatrix} \begin{matrix} a \\ b \\ c \\ d \\ e \end{matrix} \end{array} \qquad A_{G_1} = \begin{array}{c} \begin{matrix} a & b & c & d & e \end{matrix} \\ \begin{bmatrix} 0 & 2 & 1 & 0 & 1 \\ 2 & 0 & 3 & 2 & 3 \\ 1 & 3 & 0 & 2 & 2 \\ 0 & 2 & 2 & 0 & 2 \\ 1 & 3 & 2 & 2 & 0 \end{bmatrix} \begin{matrix} a \\ b \\ c \\ d \\ e \end{matrix} \end{array}$$

Figure 3.2: An illustration of graphlets and graphlet adjacencies. A: All graphlets with up to 4 nodes, labelled G_0 to G_8. **B:** Example network H. **C:** The graphlet adjacency matrices A_{G_0} and A_{G_1} for graphlets G_0 and G_1 of the example network H, shown in panel B. The off-diagonal elements correspond to the number of times two nodes touch a given graphlet together. $A_{G_0}(a,b) = 1$, as a and b form G_0 once. $A_{G_1}(a,b) = 2$, as a and b form G_1 twice, via paths *a*-*b*-*c* and *a*-*b*-*e*.

3.1.2 Graphlet Laplacian properties

To allow for an easy interpretation of the Graphlet Laplacian for each graphlet, G_k, I introduce the two-step transformation function, T, which maps graph G to its Graphlet Laplacian representation: $T(G,G_k) = \mathscr{L}_k^G$. First, T converts $G = \{V,E\}$ to a weighted network $G' = \{V,E'\}$, where the weight of each edge (u,v) in G' corresponds to $a_{uv}^k/(size(G_k)-1)$ measured in G. Next, T converts G' to its standard Laplacian representation. This shows that the Graphlet Laplacian can be interpreted as the Laplacian of an undirected weighted network. Therefore, the Graphlet Laplacian retains the following key properties of the Laplacian:

- The Graphlet Laplacian, $\mathscr{L}_k^G$, is symmetric and positive semi-definite.
- The smallest eigenvalue is 0 and the corresponding eigenvector is the constant vector $\mathbf{1}$. corresponding to the eigenvector
- The Graphlet Laplacian has n non-negative, real-valued eigenvalues: $0 = \lambda_1^k \leq \lambda_2^k, \ldots, \lambda_n^k$.
- The multiplicity of the eigenvalue 0 equals the number of connected components in

G', which I refer to as *graphlet based components*.

3.1.3 Data

Real biological network data collection

I create five unweighted and undirected networks based on three types of generic molecular interactions in human and baker's yeast (*S. cerevisiae*). I collect validated protein-protein interactions (PPIs, validated using Two-hybrid or Affinity Capture based methods) from BioGRID version 3.5.178 to form PPI networks, where nodes represent genes and edges represent physical interactions between their protein products (Stark *et al.*, 2006). I collect gene co-expression (COEX) scores from COXPRESdb version 7.3 (Okamura *et al.*, 2015) to build COEX networks, where nodes represent genes and edges represent pairs of genes being co-expressed. I consider each gene to be co-expressed with its top 1% highest scoring co-expressed genes. For yeast, I collect experimentally validated genetic interactions (GIs) from BioGRID version 3.5.178 (Stark *et al.*, 2006). I exclude the human GI network from the analysis, as only a limited number of GIs is available. Basic network statistics of these networks are provided in table 3.1.

	Nodes	**Density**	**Diam.**
PPI yeast	5,881	0.0055	6
PPI human	17,380	0.0019	9
COEX yeast	5,363	0.0129	4
COEX human	15,373	0.0131	4
GI yeast	5,634	0.0273	6

Table 3.1: Network statistics. The columns 'nodes', 'Density' and 'Diameter' respectively report the number of nodes, density and diameter of each of the molecular networks (first column).

Random model network generation

I generate ten networks containing 2,000 nodes at edge density of 1.5% (to mimic the biological networks), for each of the following widely used seven random network models: Erdős-Rènyi random graphs (ER) (Erdős Paul and Rényi Alfréd, 1959), generalized random graphs with the degree distribution matching to the input graph (ER-DD) (Newman, 2010), Barabási-Albert scale-free networks (SF-BA) (Barabási and Albert, 1999), scale-free networks that model gene duplication and mutation events (SF-GD) (Vazquez *et al.*, 2001), geometric random graphs (GEO) (Penrose, 2003), geometric graphs that model gene duplications and mutations (GEO-GD) (Pržulj *et al.*, 2010), and stickiness-index based networks (Sticky) (Pržulj and Higham, 2006). As the real biological networks have

power-law degree distributions(Jeong *et al.*, 2001; Tong *et al.*, 2004), the set of model networks contains four types of networks with power-law degree distribution: ER-DD, SF-BA, SF-GD and Sticky. The GEO and GEO-GD random network models are generated using 3-dimensional space. A summary on the basic properties of these networks and how to generate them can be found in Section 2.1.5.

Functional annotations

I collect experimentally validated functional annotations from the Gene Ontology (i.e., evidence codes 'EXP', 'IDA', 'IPI', 'IMP', 'IGI', 'IEP'), that assign genes to biological process annotations (GO-BP), cellular component annotations (GO-CC) and molecular function annotations (GO-MF) (Ashburner *et al.*, 2000; Gene Ontology Consortium, 2017).

Cancer gene annotations

I collect the pan-cancer gene mutation frequency scores computed by Leiserson *et al.* (2015) for the purpose of detecting of pan-cancer disease modules. Leiserson *et al.* (2015) collected raw pan-cancer mutation data, such as SNV's, indels and CNA's, from the TCGA database (Kandoth *et al.*, 2013). These data were filtered to exclude statistical outliers and include only the samples (corresponding to a patient) for which SNV and CNA data were available. The resulting data set contains mutations on 11,565 genes across 3,110 patients in cancers across 20 different tissues. Additionally, I collect the sets of known cancer driver genes in all available tissues from IntOGen (Gonzalez-Perez *et al.*, 2013) and Cosmic (Futreal *et al.*, 2004).

3.2 Results

I investigate the potential usage of Graphlet Laplacians to analyse network data via embedding, clustering and network diffusion experiments. I consider Graphlet Laplacians for graphlets with up to four nodes. I compare the results to the state-of-the-art Laplacian matrices: the standard Laplacian, the k-path Laplacian and Vicus. I consider path lengths up to three for the k-path Laplacian, corresponding to the maximum size of the considered graphlets underlying the Graphlet Laplacian. I set Vicus' diffusion parameter to 0.9, as this value is recommended in the original paper (Wang *et al.*, 2017) and leads to the largest number of enriched functions (see Section 3.2.2). For each network, the numbers of clusters, d, is determined using the rule of thumb: $d = \sqrt{n/2}$ (Kodinariya and Makwana, 2013). In the Supplementary Section A.1, I present the justification for

this approach, based on inspection of the spectra of different Laplacian matrices of each network.

3.2.1 Graphlet Laplacians capture different local topology

While the standard Laplacian simply captures the direct neighbourhoods of nodes and can be used to cluster densely connected nodes together, the graphlet-based neighbourhood captured by the Graphlet Laplacian allows for clustering of nodes that strongly participate in a given graphlet of interest. Because different graphlets capture different local topologies around nodes in a network (e.g., G_3 involve paths while G_8 involves cliques), clusters obtained by using different Graphlet Laplacian are expected to possess different topological features, which I assess as follows.

To assess if two graphlet Laplacians, $\mathscr{L}_i^G$ and $\mathscr{L}_j^G$, capture different topologies, I apply each Laplacian to cluster nodes in a network using Graphlet Laplacian based spectral clustering. The resulting clusters are used to partition the network into two sets of sub-networks, by inducing the sub-networks from each clustering. $\mathscr{L}_i^G$ and $\mathscr{L}_j^G$ capture different topologies if the corresponding sets of sub-networks have significantly different topology, which I measure by the overlap of two distributions: the distribution of GCD-11 distances between the sub-networks produced from $\mathscr{L}_i^G$ with the sub-networks produced from $\mathscr{L}_j^G$ and distribution of GCD-11 distances between the sub-networks produced from $\mathscr{L}_i^G$ (see Section 2.1.4). The two Graphlet Laplacians capture statistically significantly different topologies if the Wilcoxon-Mann-Whitney U-test (MWU) between the two distributions of distances is lower than or equal to 5% (see Figure 3.3 for the case of $\mathscr{L}_0^G$ and $\mathscr{L}_4^G$). For each type of model network and pairwise combination of graphlet Laplacian, I perform this test ten times and report the least significant p-value for each pairwise comparison of Graphlet Laplacian based sub-networks in Figure 3.4. In general, clusters obtained from different Graphlet Laplacians are typically statistically significantly topologically different at the 5% significance level. This is true across all of the biological networks and most of the model networks, with some exceptions in geometric models which are known to have homogeneous structure.

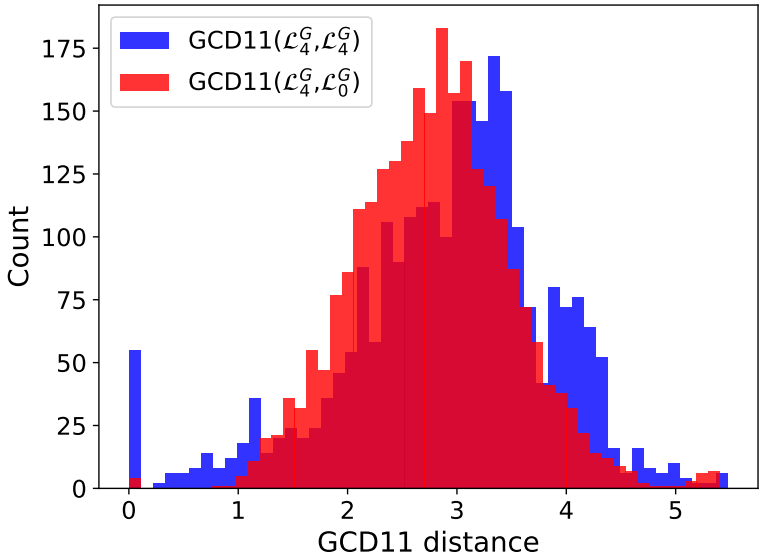


Figure 3.3: Comparison of topological distance distributions between sub-networks captured by two different Graphlet Laplacians in the human PPI network. The distribution of GCD-11 distances between the sub-networks from $\mathscr{L}_0^G$ (in blue) is statistically significantly different from the distribution of GCD-11 distances between the sub-networks from $\mathscr{L}_0^G$ and the sub-networks from $\mathscr{L}_4^G$ (in red) with MWU p-values $<5\%$. This means that $\mathscr{L}_0^G$ and $\mathscr{L}_4^G$ capture different topologies in the human PPI network.

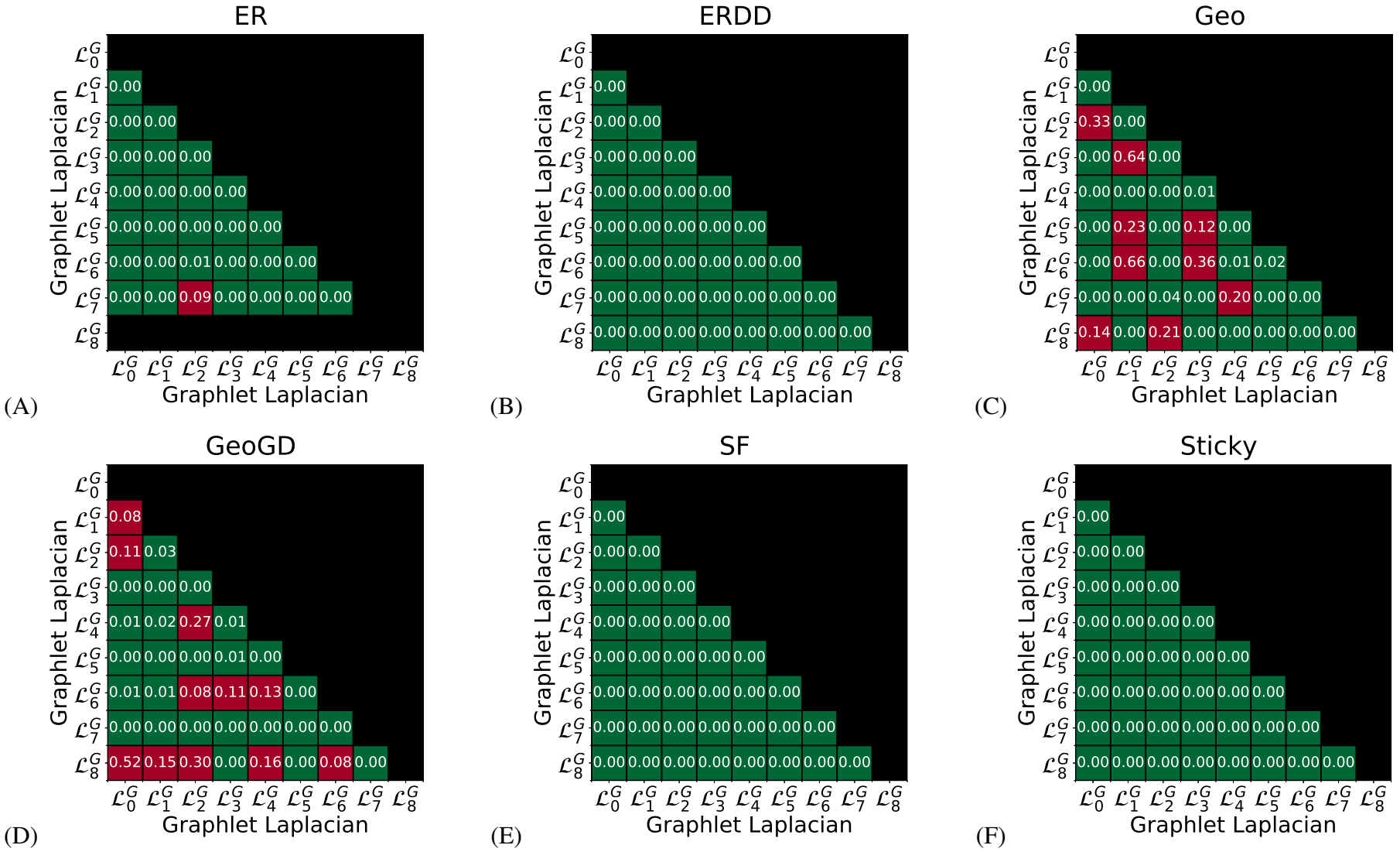


Figure 3.4: Graphlet Laplacians capture different topologies in model networks, as measured by GCD-11. Panels A-H show the p-values of the Mann-Whitney U tests, testing for topological difference between sub-networks captured by different Graphlet Laplacians in model networks (ER, ER-DD, GEO, GEO-GD, SF and Sticky).

3.2.2 Different Graphlet Laplacians capture complementary sets of biological functions

In addition to showing that Graphlet Laplacians capture different local topology, I assess their capacity to capture biological functions. To informally visualize this, I perform spectral embedding. I focus on the embedding of the yeast GI network, for which I use 14 core biological process annotations defined by Costanzo *et al.* (2016). I illustrate the spectral embedding of the symmetrically normalized $\mathscr{L}_3^G$ Graphlet Laplacian in Figure 3.5.

It is clear that the spectral embedding of $\mathscr{L}_3^G$ correctly groups and separates the biological processes of 'nuclear cytoplasmic transport', 'metabolism / mitochondria', 'Golgi / endosome / vacuole sorting' and 'Chrom. seg. / kinetoch. / spindle / micro tub.'.

I compare the spectral embeddings of the yeast GI network based on all graphlet Laplacians, Vicus and the k-path laplacian, in Figure 3.6. It is clear spectral embedding based on Vicus and the standard Laplacian fails to find any grouping at all, placing all of the nodes in the same dense cluster. Embeddings based on $\mathscr{L}_2^P$ and $\mathscr{L}_3^P$ succeed in separating different genes into different clusters, but without grouping them in a biologically meaningful way.

Next, I quantify how well the different graphlet Laplacians capture the functional organisation of the different molecular network. I apply Graphlet Laplacian based spectral clustering for each graphlet on the set of human molecular networks and assess the functional enrichments in terms of the percentage of clusters enriched and the total number of annotations enriched (Figures 3.7 and 3.8, respectively). Additionally, I create a baseline to validate the statistical significance of the enrichment results. I perform the same experiment 100 times with randomized GO-annotations. I do this by swapping the sets of gene annotations in the molecular networks such that no gene has its original set of annotations.

I observe that both in human and yeast and irrespective of the GO annotation type (BP, MF or CC), clusterings based on all Graphlet Laplacians but $\mathscr{L}_4^G$ tend to be of similar quality as those based on the standard Laplacian or Vicus, both in terms of the percentage of clusters enriched as well as total number of annotations enriched. k-path Laplacians $\mathscr{L}_2^P$ and $\mathscr{L}_3^P$ capture almost little function in PPI networks, as indicated by the low percentages of clusters enriched and few enriched GO-BP annotations found. I additionally observe that for each network and annotation type, there is always at least one Graphlet

$\mathcal{L}_3^G$

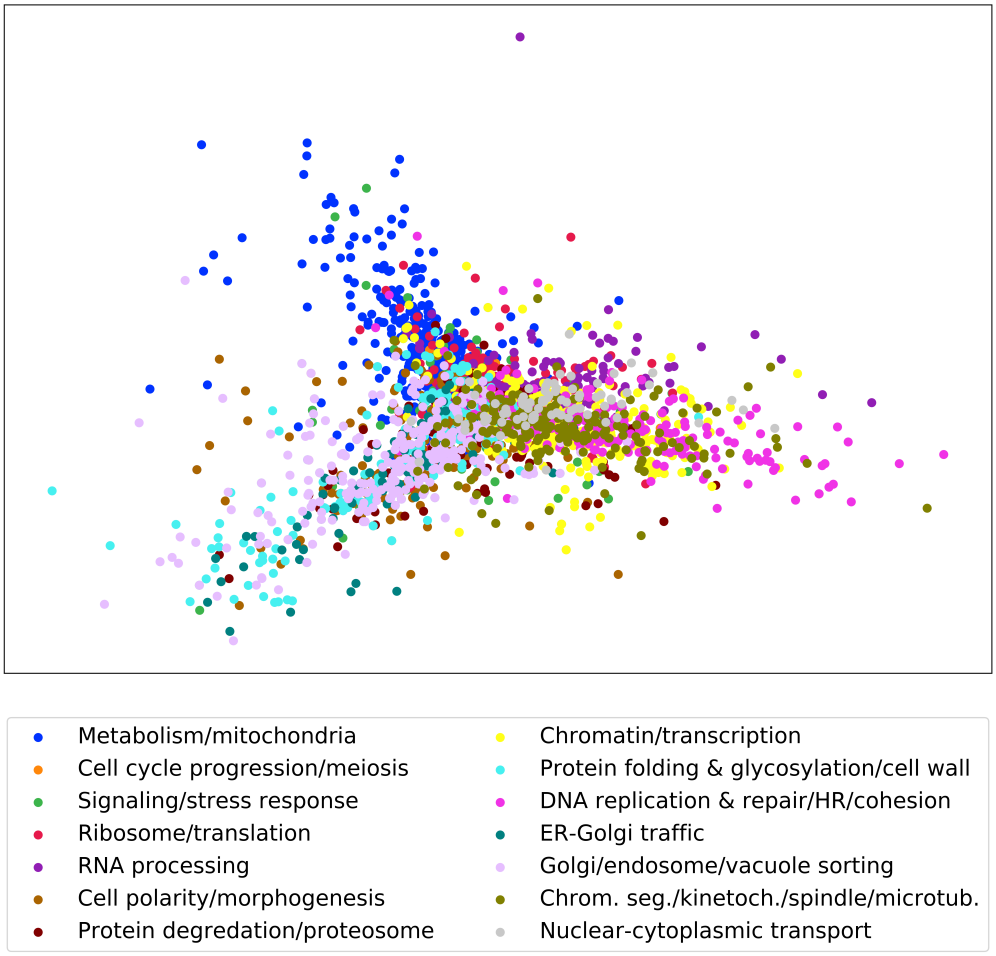


Figure 3.5: Capturing biological functions with Graphlet Laplacian $\mathcal{L}_3^G$. 2D spectral embedding of the yeast GI network using the Graphlet Laplacian for G_3. Points represent genes and are color-coded with 14 core biological process annotations defined by Costanzo *et al.* (2016).

Laplacian that shows a larger number of the total number of enriched annotations than Vicus. Hence, I conclude that Graphlet Laplacian based spectral clustering creates clusters that are at least as relevant as those achieved applying any other Laplacian matrix. This is true for all of the molecular networks for Gene Ontology BP, MF and CC annotations.

Having established that Graphlet Laplacian based clusters capture biological functions, I quantify the overlap in their enriched functions. To measure the overlap between two sets, I use the Jaccard index, which is defined as the norm of the intersection of the two sets over the norm of their union. For each type of annotation, I calculate the Jaccard Index between the sets of enriched functions corresponding to each Graphlet Laplacian. Results are presented for GO-BP, GO-MF and GO-CC in Figures 3.9, 3.10 and 3.11,

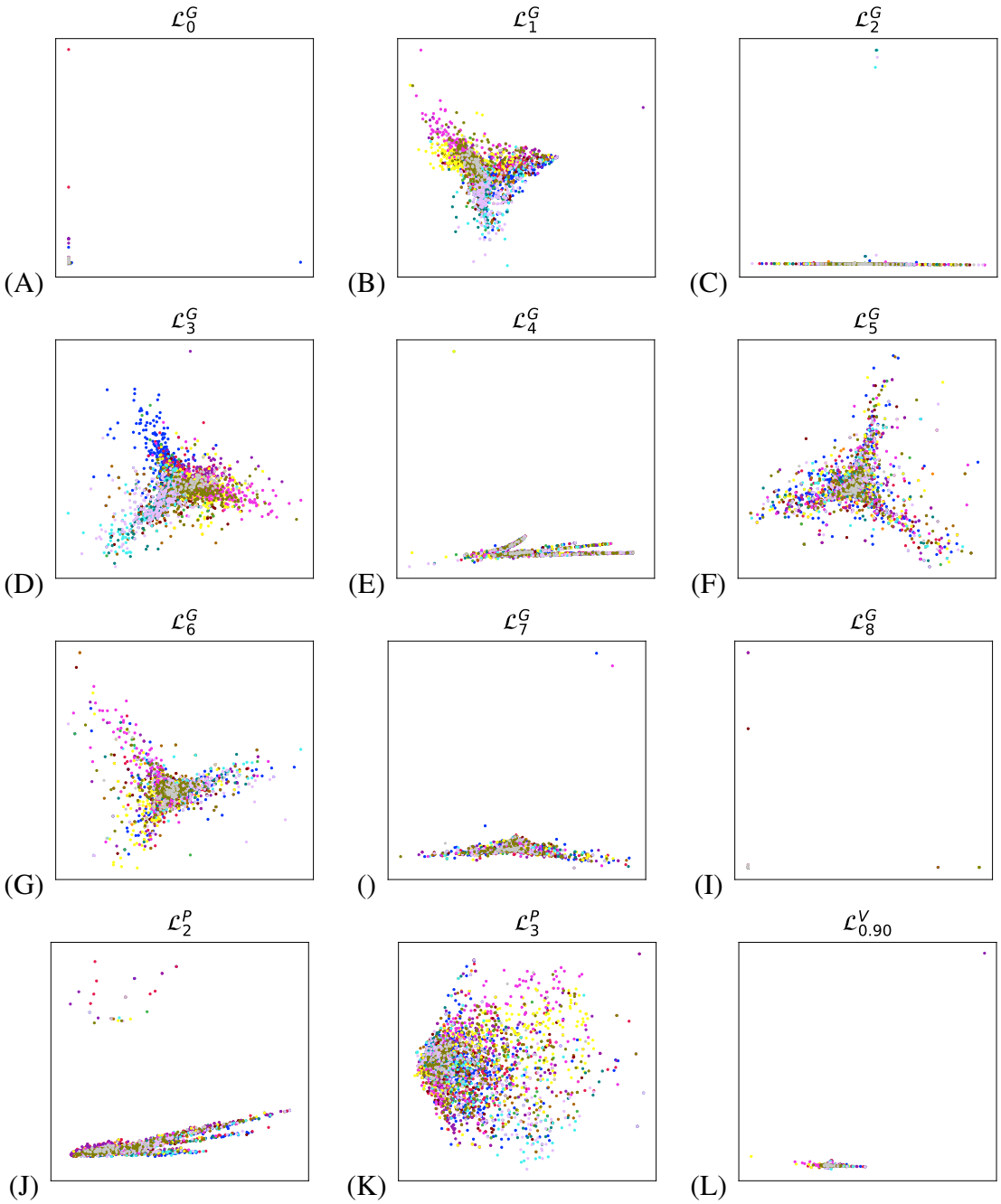
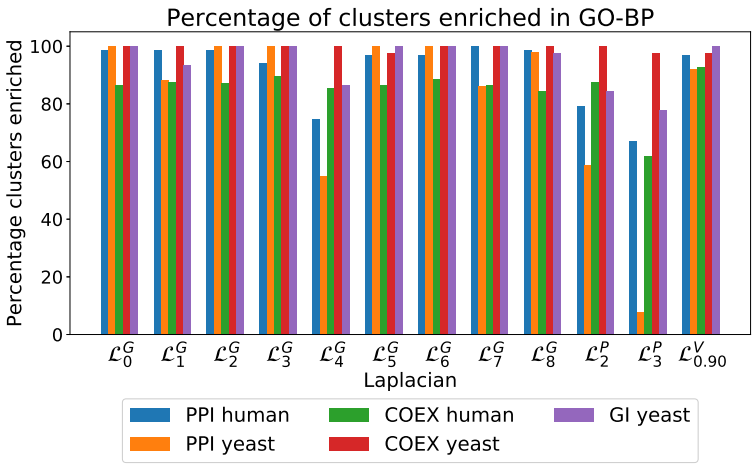
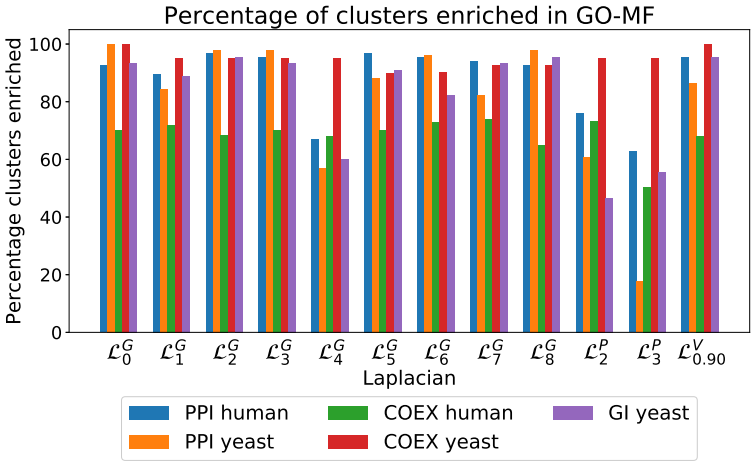


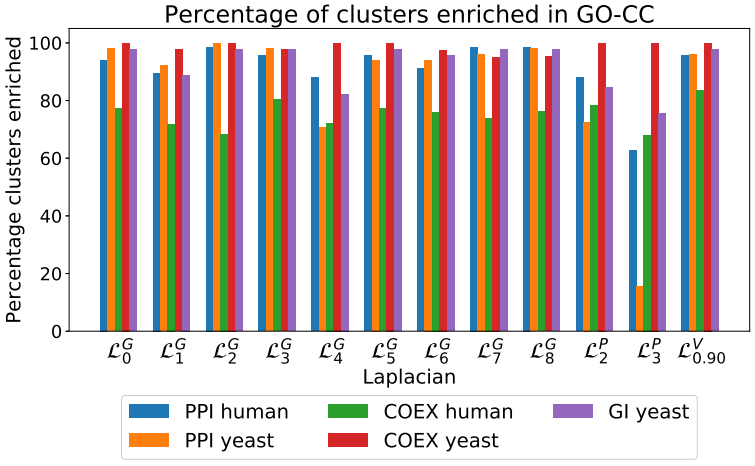
Figure 3.6: Graphlet Laplacians capture Gene Ontology biological process annotations in the yeast GI network. Panels A-K show the Spectral Embedding of the yeast GI network using different Laplacian matrices. Colour coding of the genes (represented by dots in the panels) is given in Figure 3.5.



(A)



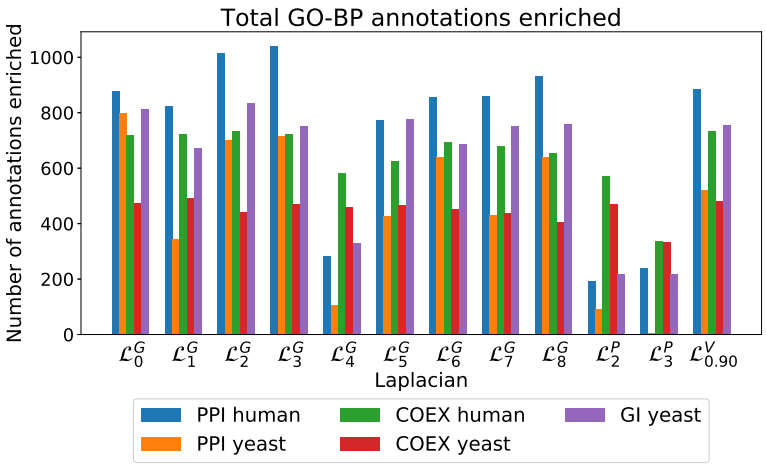
(B)



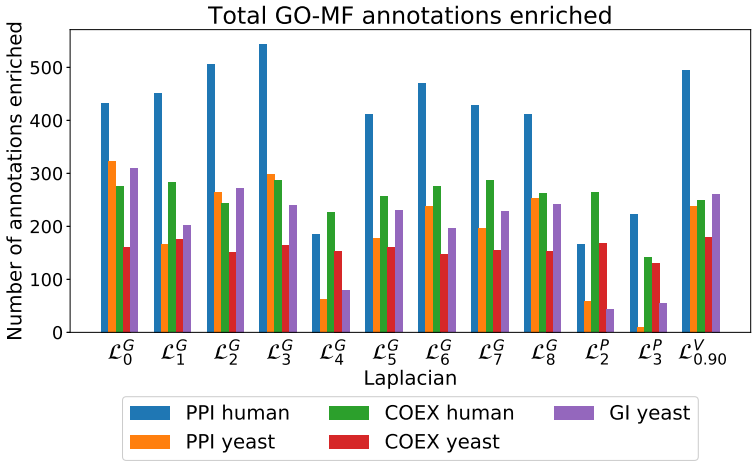
(C)

Figure 3.7: Percentage of enriched clusters. Panels A, B and C show for the set of human and yeast molecular networks (color-coded), the percentage of clusters enriched in GO BP, MF and CC annotations, respectively, with clusters obtained based on spectral clustering using different Laplacian matrices (x-axis).

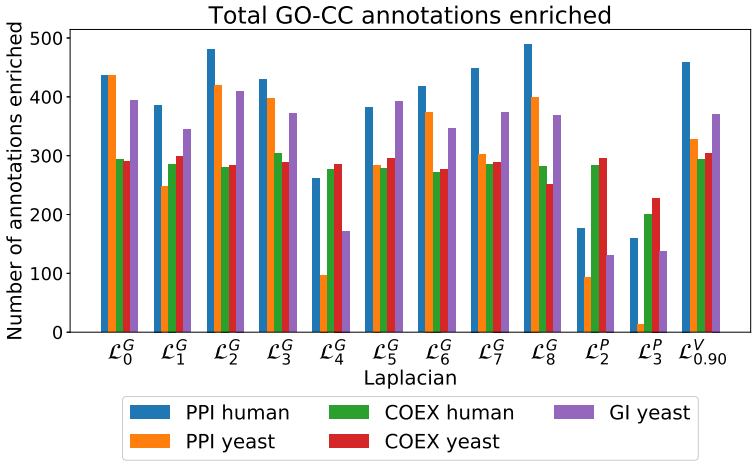
respectively. For all networks and over the different annotation types, I observe that clusterings based on different Graphlet Laplacians lead to different sets of enriched clusters, with the average Jaccard Index over all Graphlet Laplacians ranging from a minimum of



(A)



(B)



(C)

Figure 3.8: Number of functions enriched per network and per Laplacian. Panels A, B and C show for the set of human and yeast molecular networks (color-coded), the total number of enriched GO BP, MF and CC annotations, respectively, with clusters obtained based on spectral clustering using different Laplacian matrices (x-axis).

0.27 for GO-BP annotations in the human PPI network, to a maximum of 0.43 for GO-CC annotations in the yeast COEX network. For all three different types of annotations, I observe that clusterings of PPI and COEX networks based on $\mathscr{L}_4^G$, $\mathscr{L}_2^P$ and $\mathscr{L}_3^P$ are very

distinctly enriched from the clusterings based on other Laplacian matrices. In the case of $\mathscr{L}_3^P$, this is also true when clustering the yeast GI network. I conclude that different Graphlet Laplacians capture complementary sets of biological functions.

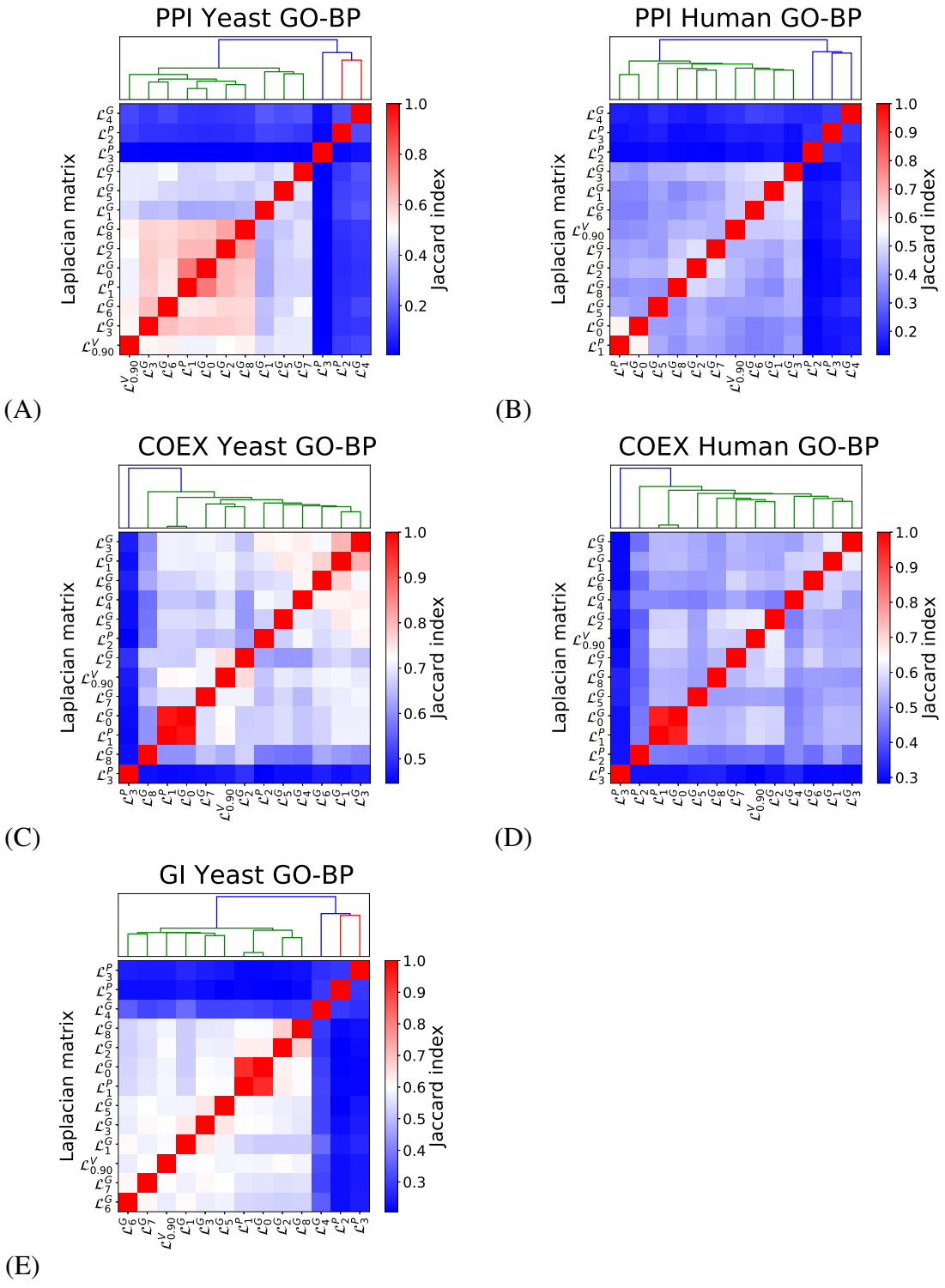


Figure 3.9: GO-BP enrichment overlap. Panels A-E show the overlap in enriched GO-BP annotations applying spectral clustering based on different Laplacian matrices (x-axis and y-axis) on the set of molecular networks (panel heading). Overlap is measured using the Jaccard Index.

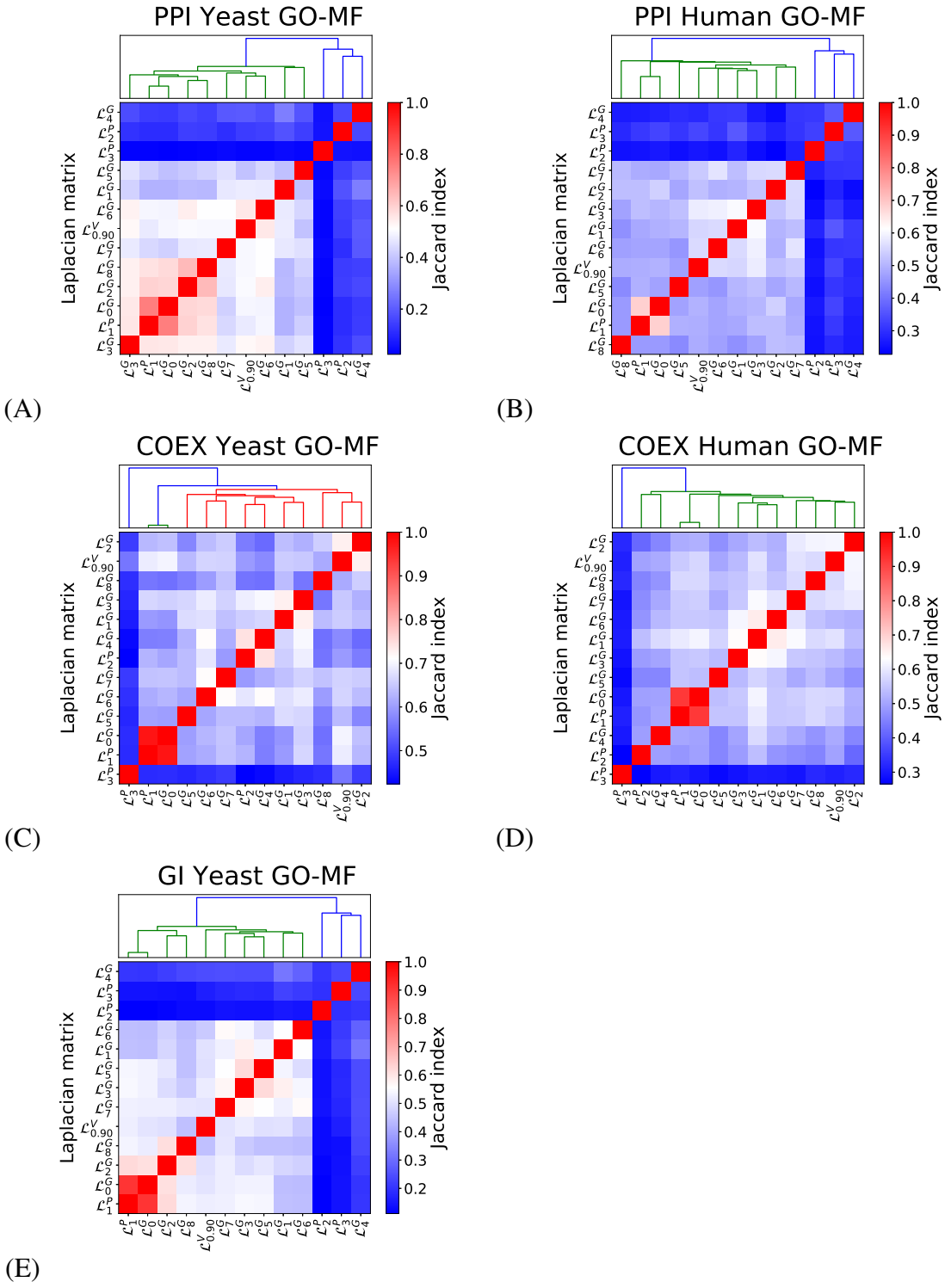


Figure 3.10: GO-MF enrichment overlap. Panels A-E show the overlap in enriched GO-MF annotations applying spectral clustering based on different Laplacian matrices (x-axis and y-axis) on the set of molecular networks (panel heading). Overlap is measured using the Jaccard Index.

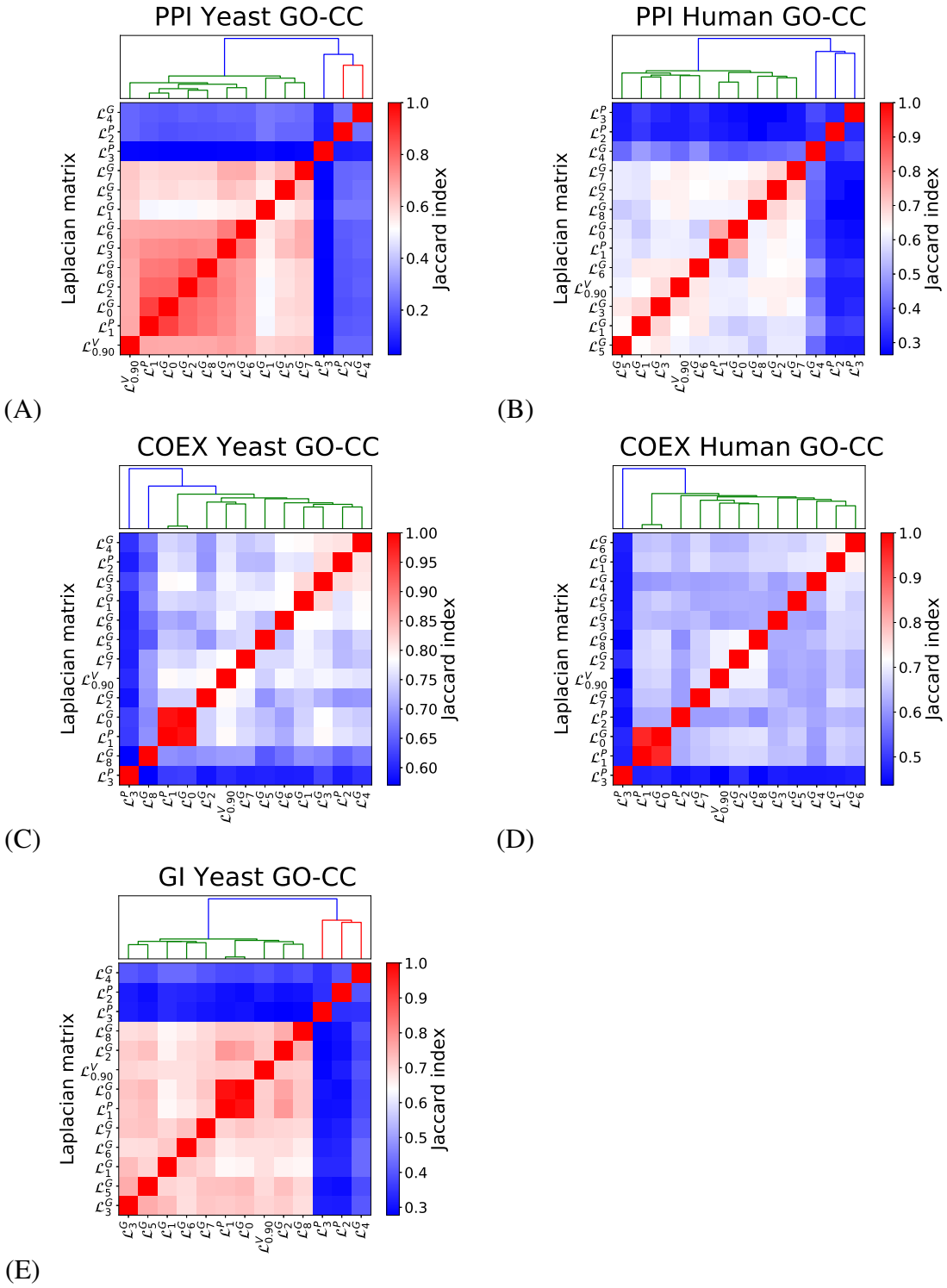


Figure 3.11: GO-CC enrichment overlap. Panels A-E show the overlap in enriched GO-CC annotations applying spectral clustering based on different Laplacian matrices (x-axis and y-axis) on the set of molecular networks (panel heading). Overlap is measured using the Jaccard Index.

3.2.3 Different Graphlet Laplacians capture complementary sets of pan-cancer driver genes

Laplacian based approaches towards predicting cancer related genes are based on guilt by association: genes that tend to be connected to frequently mutated genes are used

as cancer gene predictions. Here I show that by considering the different shapes (i.e. graphlets) by which genes can be connected to frequently somatically mutated genes, complementary cancer mechanisms can be captured.

I do this by diffusing (see section 2.2.2) the gene mutation frequency scores (see section 3.1.3) on the human PPI and COEX networks based on different Graphlet Laplacian matrices. Network diffusion is a method underlying many of the different approaches of cancer gene prioritization (Cowen *et al.*, 2017). I prioritize genes as potential cancer related genes according the highest diffused score first. I measure the quality of these scores using the area under the Precision-Recall (PR) curve and the area under the Receiver Operator Characteristic (ROC) curve. I assume a gene is correctly classified as a cancer related gene if it is known to be a cancer driver gene in at least one type of cancer in the intOGen cancer driver database (see section 3.1.3). When applying network diffusion, diffusion parameter α is set to 0.5 and 0.6 on the PPI and COEX network, respectively, as these values provided the highest area under the PR and ROC curve for each type of Laplacian matrix considered. Results are presented in Figure 3.12.

For the human PPI network, I find that the area under the ROC curve for all Laplacian matrices is higher than the expected score of 0.5 in the case of random predictions, as for each one of them the area under the ROC curve is at least 0.73. I observe that both in therms of area under the PR and ROC, Vicus and k-path Laplacian provide slightly worse scores than any Graphlet Laplacian.

For the human COEX network, I observe that all Laplacians again provide an area under the ROC curve better than random, the lowest area under the ROC curve being associated with the k-path Laplacian for paths of length three. However, differences between different Laplacian matrices are negligible.

Finally, when comparing the same Laplacian matrix in both networks, I observe that the accuracy is typically a couple of percentage points higher in the PPI network. I conclude that different Graphlet Laplacian matrices provide cancer driver gene scores of the same or better quality than the standard Laplacian and other alternative matrices.

In Figure 3.13, for the PPI and COEX network, respectively, I evaluate the overlap between the top hundred highest ranking cancer related genes per Laplacian, measured using the Jaccard Index.

For the PPI network, I observe five clusters of different Laplacian matrices producing similar sets of cancer driver genes scores. Importantly, diffusions based on three sets of Graphlet Laplacians ($\mathscr{L}^G_{\{2,5,7\}}$, $\mathscr{L}^G_{\{1,3,4,6\}}$ and $\mathscr{L}^G_{\{8\}}$) provide gene scores dissimilar to

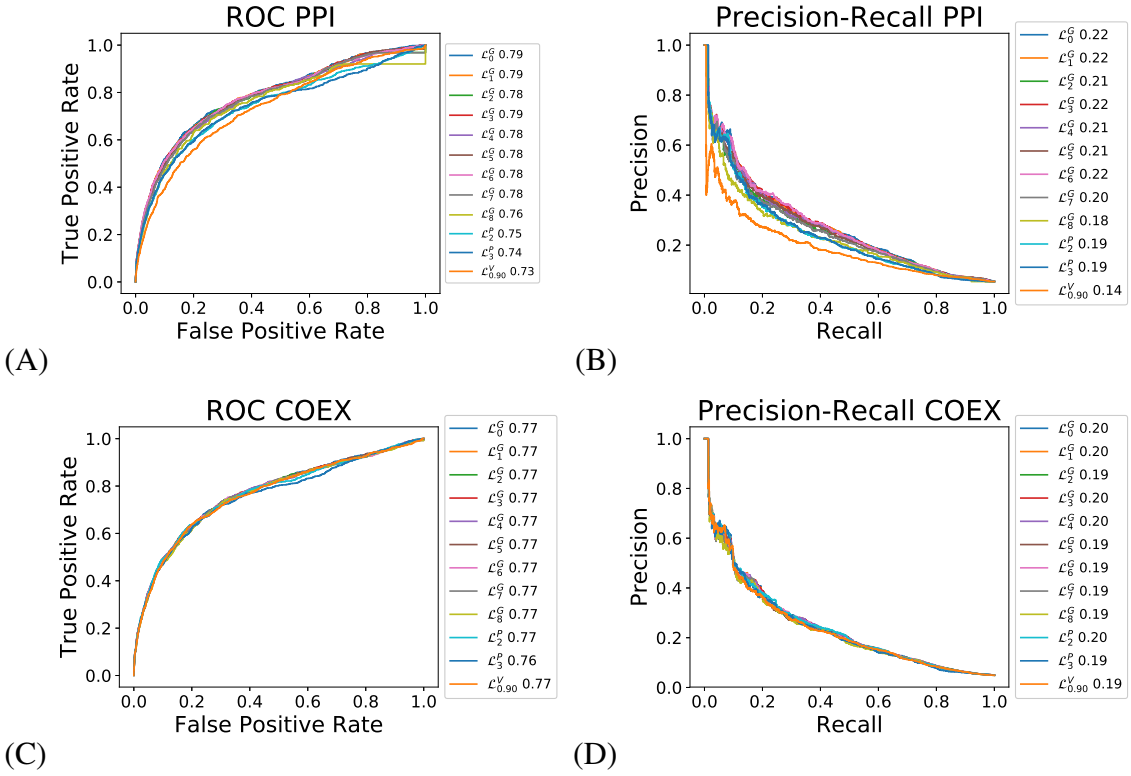
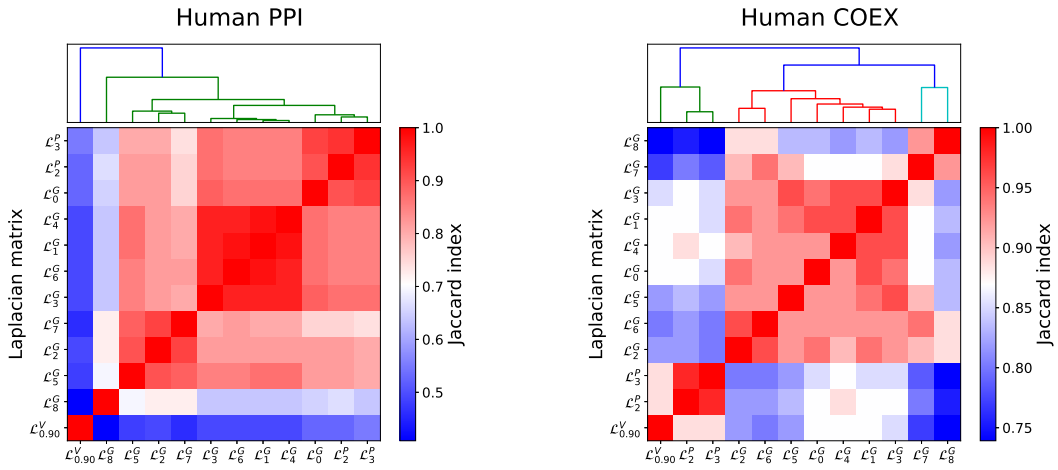


Figure 3.12: Cancer gene prediction accuracy comparison. For the human PPI network (top) and COEX network (bottom), I show the accuracy for predicting cancer related genes applying Graphlet Laplacian based diffusion, measured using the area under the ROC curve (left) and the area under the PR curve (right), respectively. For each Laplacian and measure (ROC and PR), the area under the curve is given in the legend.

those achieved using the standard Laplacian (the average Jaccard Index of each gene cluster with the clusters obtained by the standard Laplacian based scores being 0.79, 0.87 , 0.65, respectively). In contrast, the highest scoring genes based on k-path Laplacians $\mathscr{L}^{P}_{\{2,3\}}$ overlap greatly with those based on the standard Laplacian (the average Jaccard Index being 0.93). Vicus based diffusion provides cancer driver gene scores dissimilar from all other Laplacian matrices.

For the COEX network, I observe six clusters of different Laplacian matrices associated with similar sets of cancer driver genes scores. Importantly, diffusions based on three sets of Graphlet Laplacians ($\mathscr{L}^{G}_{\{2,6,7\}}$, $\mathscr{L}^{G}_{\{5\}}$ and $\mathscr{L}^{G}_{\{8\}}$) provide gene scores dissimilar to those achieved using the standard Laplacian (the average Jaccard Index of each gene cluster with the clusters obtained by the standard Laplacian based scores being 0.85, 0.85 and 0.76, respectively). Conversely, the highest scoring genes based on $\mathscr{L}^{G}_{\{1,3,4\}}$ overlap greatly with those based on the standard Laplacian (the average Jaccard Index being 0.91). Vicus and k-path Laplacian based diffusion provides cancer driver gene scores dissimilar



(A) (B)

Figure 3.13: Complementarity of cancer driver gene scores. Panels A and B show the overlap between the top 100 highest ranking cancer driver genes computed using network diffusion based on different Laplacian matrices, in the human PPI network and COEX network, respectively.

from all other Laplacian matrices.

I conclude that Graphlet Laplacian based diffusion can be used to find complementary sets of cancer driver genes.

3.3 Conclusion

In this chapter, I introduce graphlet adjacency for unweighted and undirected networks to simultaneously capture graphlet-based topological information and neighbourhood membership information. I demonstrate that they can straightforwardly be plugged into current Laplacian based network analysis methods widely used in systems biology, using spectral clustering, spectral embedding and network diffusion as example applications.

Through the generalized spectral embedding and spectral clustering methods on real and model networks, I show that different Graphlet Laplacians capture sub-networks having distinct local topologies and that are enriched in different, but complementary sets of biological annotations. Finally, I show that the generalized network diffusion of pan-cancer gene mutation scores resulted in complementary sets of cancer related genes for gene prioritization dependent on the underlying graphlet. In all the tested applications, the Graphlet Laplacians perform as good as and often better than k-path and Vicus Laplacians, while being directly interpretable.

Chapter 4

Graphlet eigencentralities capture novel central roles of genes in pathways

The previous chapter introduced the graphlet adjacency, a graphlet-based generalization of the regular adjacency to capture the higher-order wiring patterns in local network neighbourhoods. I showed that graphlet adjacency captures topology-function and topology-disease relationships in biological networks through cluster enrichment analysis and pan-cancer gene mutation scores diffusion. A question that arises is what exactly is captured by graphlet-adjacencies in these contexts. Therefore, I perform a more descriptive pathway-focused analysis to investigate further the relationships between the topological features of genes participating in molecular networks, as captured by graphlet adjacencies, and the biological functions and disease mechanisms they capture. I introduce a new graphlet-based definition of eigencentrality, *graphlet eigencentrality*, to identify pathways and cancer mechanisms described by a given graphlet adjacency. I show that pathways are best described by the graphlet adjacencies that capture the importance of their functionally critical genes. I also show that cancer driver genes characteristically perform hub roles between pathways. This analysis is summarised in Figure 4.1.

Chapter impact

This chapter has lead to the following contributions:

Publications:

Windels, S. F. L., Malod-Dognin, N., and Pržulj, N. (2022). Graphlet eigencentralities capture novel central roles of genes in pathways. *PLoS One*, **35**(24), 5226–5234.

Methodological contributions:

To capture the topological importance of nodes in a network based on their higher-order topology, as captured by graphlet adjacency, I introduce graphlet eigencentrality. For a given graphlet, a node has a high graphlet eigencentrality if it and its neighbours frequently occur on that graphlet.

Biological contributions:

Pathway-based graphlet eigencentralities allow to accurately predict which genes participate in a given pathway. Pathway-based graphlet eigencentralities based on graphlets that capture hub-roles allow for an accurate prediction of cancer related genes.

Software and data analysis:

The data and scripts to perform the analysis presented in this chapter have been made publicly available at: https://gitlab.bsc.es/swindels/graphlet_eigencentrality .

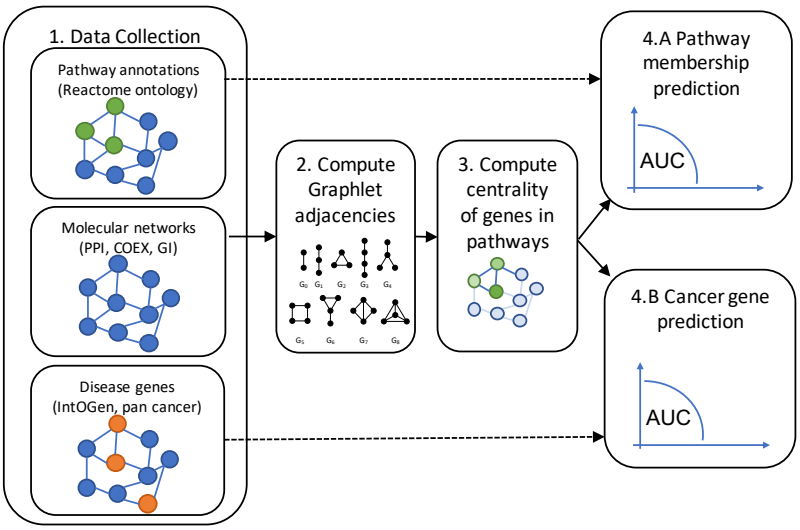


Figure 4.1: Chapter workflow summary. Step 1: collection of networks, pathway annotations and cancer drivers. Step 2: computation of graphlet adjacencies. Step 3: computation of gene-pathway centralities based on different graphlet eigencentralities. Step 4: use the centrality scores to predict: A) pathway membership, B) cancer related genes.

4.1 Methods

Network centrality measures quantify the importance of a node in a network (see Section 2.1.2). I first extend eigencentrality, which considers a node as important if it clusters with other highly clustered nodes (see Section 2.2.4), to graphlet eigencentrality (Section 4.1.1). Next, I explain how I use graphlet eigencentralities to measure the centrality of a gene in a pathway, or its *pathway centrality*. I can measure pathway centrality from the pathway perspective (the centrality of the genes is computed on the genes known to participate in the pathway, see Section 4.1.2), or from the global network perspective (the centrality of the genes is computed on the full network before inducing the set of nodes corresponding to genes participating in the pathway, see Section 4.1.2). Finally, I explain how I use pathway centrality to predict which genes participate in a given pathway (see Section 4.1.3).

4.1.1 Graphlet eigencentrality

Graphlets are small, connected, non-isomorphic, induced subgraphs of a large network (Pržulj *et al.*, 2004). Two nodes u and v of H are considered *graphlet adjacent* with respect to a given graphlet, G_i, if they simultaneously touch G_i (Windels *et al.*, 2019). Given this extended definition of adjacency, the graphlet based adjacency matrix is defined as:

$$A_{G_i}(u,v) = \begin{cases} c_{uv}^{G_i}/\theta_{G_i} & \text{if } u \neq v \\ 0 & \text{otherwise,} \end{cases} \tag{4.1}$$

where $c_{uv}^{G_i}$ is equal to the number of times the nodes u and v are graphlet adjacent w.r.t. graphlet G_i and θ_{G_i} is equal to the number of nodes in graphlet G_i minus 1. I generalize normalised eigencentrality to graphlet eigencentrality by replacing A with the normalised graphlet adjacency matrix, $\widetilde{A_{G_i}}$, in equation 2.10:

$$\widetilde{A_{G_i}}\mathbf{c}_{G_i} = \lambda_{G_i}\mathbf{c}_{G_i}, \tag{4.2}$$

4.1.2 Pathway centrality

I aim to measure the centrality of the set of genes that participate in a given pathway. I can do this from the pathway perspective, which I will refer to as 'Local pathway centrality', or from the perspective of the entire network, which I will refer to as 'Global pathway centrality'.

Local pathway centrality

I take the submatrix of the adjacency matrix of the full network corresponding to the m genes participating in the pathway, to create the $m \times m$ dimensional local adjacency matrix P. Then, for a given underlying graphlet G_i, I compute the corresponding graphlet adjacency matrix, P_{G_i}, and compute the normalised graphlet eigencentrality applying equation 4.2.

Global pathway centrality

For a given underlying graphlet G_i, I compute the global graphlet eigencentrality vector, $\mathbf{c}_{G_i}$, on the normalised graphlet adjacency matrix, $\widetilde{A_{G_i}}$, applying equation 4.2. Then, I take the subvector of the global eigenvector corresponding the m to genes participating in the pathway , $\mathbf{c}_{G_i}$, to determine their pathway centrality.

4.1.3 Predicting pathway participation

Pathways are functional subgraphs in which a group of genes work together to perform a given biological function. I assume that for a pathway to perform its function, each gene is important. So, I consider a pathway to be described by a given graphlet adjacency if the topology captured by it correctly recognises that all genes in the pathway as important. To evaluate which pathways are described by a given graphlet adjacency, I use local and global pathway centrality eigencentralities to predict which genes belong to a given pathway, as described below. I consider the pathways for which I achieve the highest prediction accuracies as being described by that graphlet adjacency; as for those pathways I can best distinguish between the genes are relevant w.r.t. the pathway and those are not. To show this approach captures biological signal, I compare the prediction accuracy to that of the label propagation algorithm GeneMANIA, (Franz *et al.*, 2018).

Given a molecular network and graphlet adjacency, I apply for each pathway ten iterations of 5-fold cross-validation, where I predict which genes participate in it based on their pathway-based graphlet eigencentrality. I evaluate prediction performance per pathway. That is, for each pathway and fold, I randomly hold out 20% of the genes known to participate in the pathway to form the positive examples in the test-set. The negative examples in the test-set are all genes in the full network that directly interact with one of the m (i.e., 80%) of the remaining genes known to participate in the pathway.

Prediction based on local pathway centrality

For each gene in the test set, I compute how central it would be in the pathway if it were to participate in it. That is, for each gene in the test set, I induce the nodes corresponding to the gene and the m remaining genes known to participate in the pathway on the full network to define a local $(m+1) \times (m+1)$ dimensional adjacency matrix P, based on which I compute the local pathway centrality of the gene (see Section 4.1.2). In this way, the centrality of each gene in the test set is based on local pathway topology, avoiding taking into account the 'noise' coming from interactions with nodes outside the pathway.

Prediction based on global pathway centrality

For a given pathway, the underlying graphlet and a given fold, I compute the global pathway-based graphlet eigencentralities for all the genes in the test set (see Section 4.1.2). I consider genes with a higher global pathway-based graphlet eigencentrality to be more likely to be participating in the pathway.

GeneMANIA

GeneMANIA is a supervised approach that uses a label propagation algorithm to predict gene annotations (see Section 2.2.2). I choose to compare against GeneMANIA as it: (1) is one of the few gene annotation predictors that, like pathway eigencentrality, can be trained using only positive examples and (2) allows for sampling annotations at the pathway level rather than at the gene level (i.e. 20% of the genes of each pathway can be held out, instead of holding out 20% of the pathway annotated genes), such that when applying 5-fold cross-validation, all pathways have exactly 20% of the nodes withheld.

4.1.4 Predicting cancer-related genes

I hypothesise that cancer-related genes play central roles in pathways and hence can be predicted based on their pathway based graphlet eigencentralities. For each pathway and graphlet adjacency, I directly use global or local graphlet eigencentrality to rank the genes participating in a given pathway, assuming that genes with a higher eigencentrality are more likely to be cancer-related. For each pathway, I consider the set of known cancer driver genes participating in the pathway as the set of true positives. As here the approach is unsupervised (i.e. I do not use the information of which genes are known cancer drivers when computing pathway centralities), no cross-fold validation is needed.

4.1.5 Evaluating prediction performance

I evaluate prediction performance on a per pathway and per graphlet adjacency basis using the area under the precision-recall curve (AUC-PR) and the area under the receiver operating characteristic curve (AUC-ROC), which are defined as follows.

For a given prediction, the *true positive rate* (TPR) is the number of correctly predicted true positives (i.e., the genes correctly predicted as part of the pathway or to be cancer driver genes) over all known true positives (i.e., all genes known to be part of the pathway or all cancer drivers in the pathway). The *false positive rate* (FPR) is defined as the number of genes falsely predicted as positive (i.e., the genes falsely predicted to be participating in a pathway or to be cancer driver genes). The ROC curve sets out the relationship between the TPR and FPR for predicting pathway participation at various cut-offs. The AUC-ROC is used as a single number summary of the ROC curve, as a measure of prediction accuracy.

Similarly, for a given prediction, the *precision* is defined as the number of correctly predicted true positives (i.e., the number of genes correctly predicted to participate in the pathway, or the number of genes correctly predicted as cancer drivers) over the total number of genes in the prediction set (e.g., the known genes participating in the pathways and the genes they are directly connected with outside the pathway, or the all genes known to participate in the pathway). *Recall* is synonymous to the TPR, defined above. The precision-recall curve sets out the relationship between the precision and recall at various cut-offs. The area under the PR curve is then used as a single number summary of the precision-recall curve, as a measure of prediction accuracy.

To be able to identify the pathways or cancer mechanisms that are exceptionally well captured by a given graphlet adjacency, I define the normalized AUC-PR. For each graphlet adjacency and a given prediction task, I normalize the distribution of AUC-PR scores over all pathways by subtracting the median and dividing by the mean absolute deviation.

Assigning ancestor annotations to pathways

The Reactome Ontology is a collection of 23 direct acyclic graphs (dags), where nodes represent pathway annotations and directed edges represent ‘is a’- relationships. I annotate each pathway with its ancestor terms found 1 step away from the root node of the corresponding dag. That is, to annotate a given pathway with its ancestor(s), I first find that pathway in the Reactome dag, from there trace the Reactome Ontology dag upwards

(against the direction of the 'is_a' relationships) until I reach the pathway annotation(s) that is(are) one step away from the root node(s), and use the annotations corresponding to these nodes as ancestor annotations.

Pathway set enrichment

To assess if a set of pathways is statistically significantly enriched by pathways sharing ancestor annotations, I apply the hyper-geometric test. That is, I consider a set of pathways as a 'sampling without replacement' experiment, in which each time I find a given ancestor or GO-term annotation, I count that as a 'success'.

The probability of observing the same or higher enrichment (i.e. successes) of the given annotation purely by chance is equal to:

$$p = 1 - \sum_{i=0}^{X-1} \binom{K}{i} \binom{M-K}{N-i} / \binom{M}{N}. \tag{4.3}$$

where N is the number of ancestor annotated pathways in the pathway-set, X is the number of pathways annotated with the given ancestor annotation in the pathway, M is the number of ancestor annotated pathways pathways and K is the number of pathways annotated with the given ancestor over all pathways in the pathway-set. An ancestor annotation is considered to be statistically significantly enriched if its enrichment p-value is lower than or equal to 5% after application of the Benjamini and Hochberg correction for multiple hypothesis testing.

4.1.6 Data

Biological networks

I reuse the biological networks collected in the previous chapter, see Section 3.1.3 for details.

Annotation data

I collect pathway annotation data assigning genes to pathways, from the Reactome pathway ontology (Jassal *et al.*, 2019). For each of the five molecular networks, I create a set of pathway networks by inducing the gene set of each pathway on the network. For each molecular network, I consider those pathways that form a connected subgraph of a size of at least 10 and up to 100 nodes when induced on the full network. The number of pathways considered per molecular network is summarized in Table 4.1. The distribution of pathway sizes for each of the molecular networks is shown in Figure 4.2.

	No. of pathways
PPI yeast All	187
PPI human All	969
PPI human Disease	92
PPI human Non-disease	877
COEX yeast All	141
COEX human All	712
COEX human Disease	68
COEX human Non-disease	644
GI yeast All	241

Table 4.1: Number of pathways considered for each molecular network.

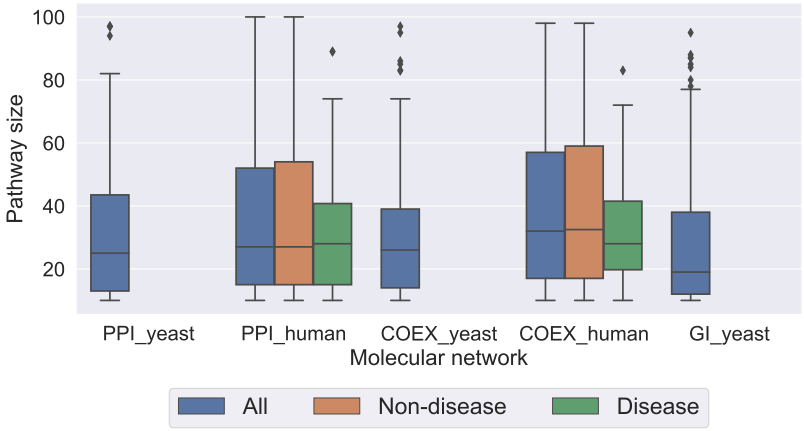


Figure 4.2: Distribution of pathway sizes per molecular network. Each box plot represents the distribution of the pathway sizes for each of the molecular networks (x-axis) considering all pathways, non-disease pathways and disease pathways (colour) in the Reactome ontology.

I reuse the 586 cancer driver annotations collected in the previous chapter from the intOGen database (Gonzalez-Perez *et al.*, 2013). I consider a gene to be a cancer driver if it is a known cancer driver in at least one cancer type.

4.2 Results and discussion

4.2.1 Graphlet adjacencies describe topologically and biologically distinct pathways

First, I validate that graphlet adjacencies can capture topological relationships between the nodes in a pathway by evaluating pathway participation prediction accuracy (Section 4.2.1). Then, I show that the pathways that are described by the same graphlet adjacency, share biological functional similarities that are different dependent on the graphlet adja-

cency considered (Section 4.2.1). I conclude this section with a case study, where I focus on the 'Receptor mediated mitophagy' pathway and explain why some graphlet adjacencies best capture the topological-functional relationships between nodes in the pathway (Section 4.2.1).

Graphlet adjacency captures pathway specific topology

I assess if graphlet adjacencies capture pathway topological signal by evaluating the performance of graphlet eigencentrality for the purpose pathway participation prediction. In Figure 4.3, I observe that regardless of the underlying graphlet adjacency and molecular network type, the local approach and GeneMANIA consistently perform better than random (AUC-ROC=0.5), achieving median AUC-ROC scores higher than 0.6. This, except in the yeast GI network, where GeneMANIA performs close to random and in the yeast PPI network, where the local approach performs close to random when applied on graphlet adjacency $\widetilde{A_{G_4}}$. The global approach performs as by random when applied on graphlet adjacencies for $\widetilde{A_{G_1}}$, $\widetilde{A_{G_2}}$ and $\widetilde{A_{G_8}}$ in PPI and GI networks, with median AUC-ROC scores around 0.5. Given that the ratio of positive examples in each test-set is only 0.15 on average, AUC-PR is a better measure for comparison. In terms of AUC-PR, I observe that the local approach consistently outperforms the global approach, as well as GeneMANIA. To explain this result, I found that each pathway annotated gene participates in 6 pathways on average. Furthermore, on average, these 6 pathways are descendants of 2 (of the 23) different root nodes of the pathway ontology. This implies that from the perspective of the global network, pathways are intertwined, even functionally very distinct ones, making it harder to predict if a gene participates in a pathway or not. The local approach, however, considers each pathway as an individual entity, disentangled from the rest of the network. This validates the intuition that, from the perspective of the pathway, all genes participating in it are important.

Next, to validate that different graphlet adjacencies best capture different sets of pathways, I compare the set of top-scoring pathways of each graphlet adjacency. I will be referring to the pathways for which I achieve the highest prediction accuracy considering a given graphlet adjacency as *described* by that graphlet adjacency.

Formally, for each graphlet adjacency, I consider those pathways with a normalised AUC-PR score larger than 3.0 (in analogy to the 99.7% confidence interval for variables following a standard normal distribution) to be described by that graphlet adjacency. On average, 55 pathways are found to be described by a graphlet adjacency. By measuring

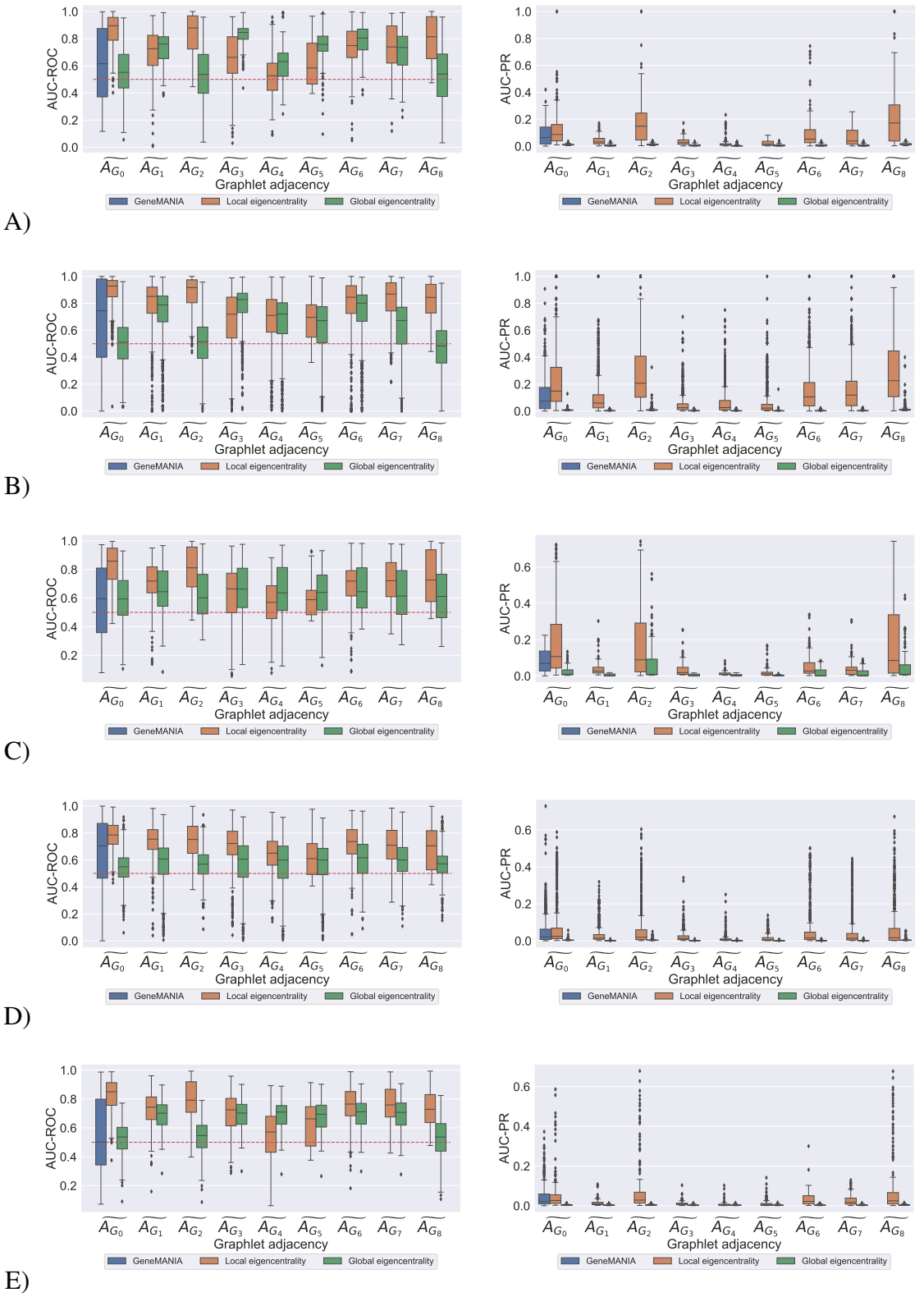


Figure 4.3: Pathway participation prediction accuracy in different molecular networks. I show the pathway participation prediction accuracy measured using AUC-ROC (left) and AUC-PR (right), for three methods (see legend), applied on different graphlet adjacencies (x-axis), in the yeast PPI network (A), hyman PPI network (B), yeast COEX network (C), human COEX network (D), yeast GI network (E). Each box plot represents the distribution of prediction accuracies over all pathways using the indicated method and graphlet adjacency.

the pairwise overlap between the set of pathways described by the different graphlet adjacencies, I find that the average of the Jaccard indices is 0.17. I conclude that graphlet adjacencies capture pathway topologies that are different and described by the underlying graphlet.

Graphlet adjacencies for different graphlets captures complementary groups of functionally related pathways

Having shown that graphlet adjacencies capture pathway topologies, I assess if any graphlet adjacency describes functionally similar pathways and compare the biological functions captured by different graphlet adjacencies.

To assess if a given graphlet adjacency captures similar pathways, I annotate each pathway with its second level *ancestors*, i.e. annotations in the second most general level of the pathway ontology, one step away from the root nodes (see Section 4.1.5) and perform pathway set enrichment analysis (see Section 4.1.5).

In the bar charts at the top of Figure **??** , I observe that in all five of the molecular networks, each graphlet adjacency describes pathways that are enriched in at least one ancestor annotation. This means that in all five of the molecular networks, different graphlet adjacencies describe pathways that are functionally similar in terms of the types of ancestor annotations. For instance, the set of pathways described by graphlet adjacency $\widetilde{A_{G_3}}$ in the yeast PPI network, is enriched in pathways related to 'Signaling by GPCR' (9 out of 59 pathways are descendants of this ancestor, adjusted p-value 2.23E−5), 'Transmission across Chemical Synapses' (9 out of 59 pathways are descendants of this ancestor, adjusted p-value 2.23E−5) and 'Platelet activation, signalling and aggregation' (6 out of 59 pathways are descendants of this ancestor, adjusted p-value 7.61E−23). There is one exception to this conclusion in the yeast COEX network, where the set of pathways described by graphlet adjacency A_{G_1} is not enriched in any ancestor annotations, meaning these pathways are not statistically significantly similar in terms of the type of pathways they represent.

For the set of yeast molecular networks, in the heat maps presented in panels A, C and E, I generally find very low overlap between the functional annotations enriched in the pathway sets described by different graphlet adjacencies. The lowest overlap in terms of enriched functional annotations is achieved in the COEX network, where the average Jaccard index between the ancestors enriched in the pathways described by two different graphlet adjacencies 0.11. The highest overlap in terms of enriched functional

annotations is achieved in the PPI network, where the average Jaccard index between ancestors enriched in the pathways described by two different graphlet adjacencies is 0.20. For the set of human molecular networks, in the heat maps presented in panels B and D, I generally find low overlap between the ancestor annotations enriched in the pathway sets described by different graphlet adjacencies. The lowest overlap is achieved in the PPI network, where the average Jaccard index between the ancestors enriched in the pathways described by two different graphlet adjacencies 0.17. The highest average overlap is achieved in the human COEX network, where the average Jaccard index between the GO-BP terms enriched in the pathways described by two different graphlet adjacencies 0.40. I conclude that, pathways described by different graphlet adjacencies are functionally different in terms of the ancestor annotations in which they are enriched.

Case study: Receptor mediated mitophagy

'Receptor Mediated Mitophagy' (RMM) is a degradation process in the cell focused on the degradation of damaged mitochondria. I found the pathway to be highly described by graphlet adjacency $\widetilde{A_{G_6}}$ (normalised AUC-PR 5.98) and not described by $\widetilde{A_{G_0}}$ (normalised AUC-PR 0.04) in the yeast PPI network, and will be focusing on this pathway to explain why some graphlet adjacencies better capture some pathways than others.

In Figure 4.5 I show the spring embedding of RMM based on normalised graphlet adjacencies $\widetilde{A_{G_0}}$ and $\widetilde{A_{G_6}}$. For graphlet adjacency $\widetilde{A_{G_0}}$, the RMM pathway is composed of two densely connected modules, the control mechanism (genes CSNK2A, CSNK21, CSNK2B, SRC) and the phagophore formation process (genes ATG5, ATG12, MAP1LCA, MAP1LCB, ULK1), which interact through a single hub gene, FUNDC1. This is unfavourable for prediction, as a gene would be predicted to be part of the pathway if it is densely connected with just one of the two clusters. Graphlet adjacency $\widetilde{A_{G_6}}$, however, does capture the fact that, through hub node FUNDC1, all the genes in the control mechanism and the phagophore formation process are functionally related (i.e. executing the RMM process), as both groups of genes are now highly connected. This also better captures the pathway from a topological perspective, as genes predicted to be part of the pathway would have to interact (in the form of graphlet G_6) with all pathway members. I conclude that graphlet adjacency allows to describe the functional organization of the pathway.

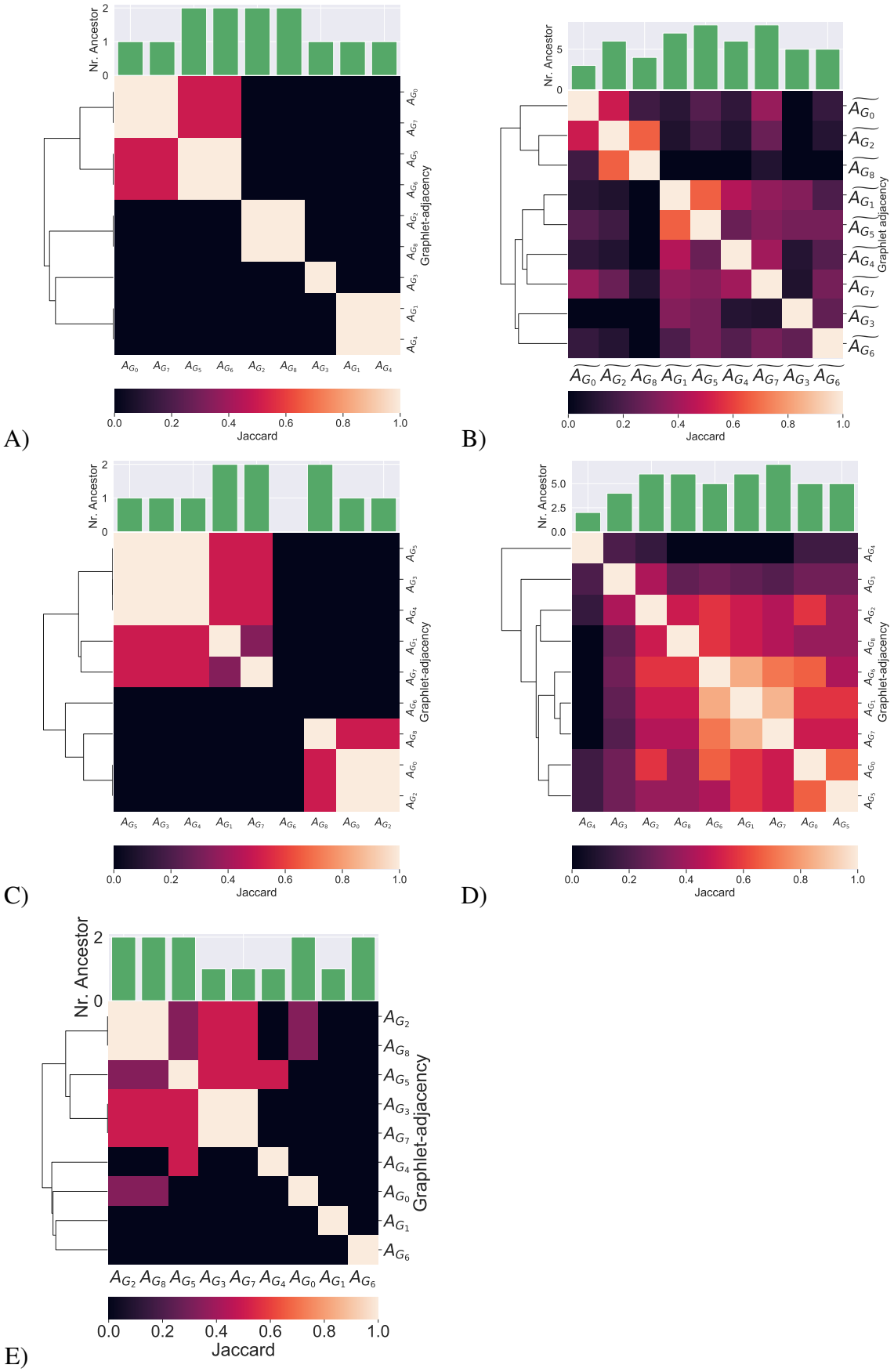


Figure 4.4: Functional similarity between pathways described by different graphlet adjacencies. For the yeast PPI network (A), human PPI network (B), yeast COEX network (C), human COEX network (D) and yeast GI network (E), I show a clustered heat map of the Jaccard similarity indices between the sets of ancestor annotations enriched in the sets of pathways described by different types of graphlet adjacencies. Above each heat map, a bar-chart indicates the number of ancestor annotations enriched in the pathways described by each corresponding graphlet adjacency.

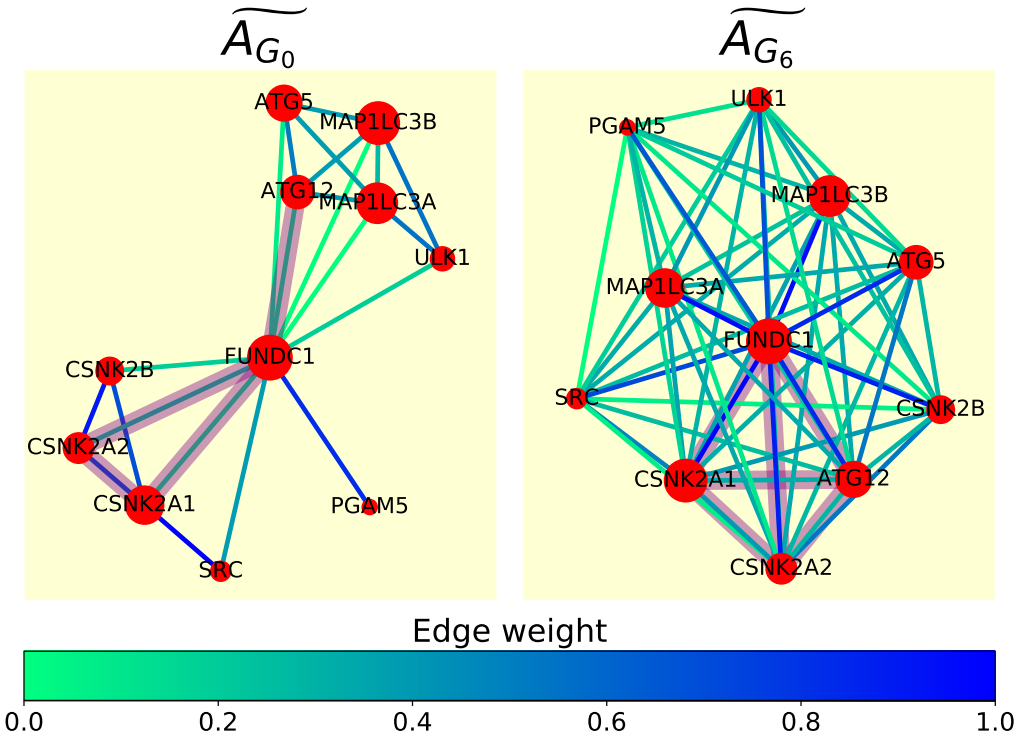


Figure 4.5: Graphlet adjacency $\widetilde{A_{G_6}}$ captures RMM functional organisation Spring embedding of RMM based on normalised graphlet adjacency $\widetilde{A_{G_0}}$ (left) and $\widetilde{A_{G_6}}$ (right), where nodes represent genes (red) and edges represent weighted normalised graphlet adjacency (see legend). Graphlet G_6 is indicated in purple in the spring embedding based on G_0, connecting genes CSNK2A1, CSNK2A2, FUNDC1 and ATG12. The subnetwork obtained by inducing these same nodes is also indicated in purple in the spring embedding based on graphlet adjacency $\widetilde{A_{G_6}}$. Although only connected via FUNDC1 when considering regular adjacency, ATG12 is directly connected to CSNK2A1 and CSNK2A2 in the spring embedding based on graphlet adjacency $\widetilde{A_{G_6}}$, illustrating how graphlet adjacencies capture functionally relevant indirect relationships between nodes.

4.2.2 Graphlet adjacency based pathway centrality captures complementary cancer mechanisms

Here, I illustrate how graphlet eigencentralities enables to relate specific local wiring patterns of genes in a pathway with their individual biological function. I focus on predicting cancer driver genes. I first verify that cancer-related genes play central roles in pathways (Section 4.2.2). Then, I show that the set of cancer driver genes recognised for playing central roles in pathways are different based on the graphlet adjacency considered (Section 4.2.2). To explain this, I illustrate it with a case study, where I show why graphlet adjacency $\widetilde{A_{G_6}}$ best captures the central roles of cancer driver genes, TP53 and RB1, in the 'Formation of Senescence-Associated Heterochromatin Foci' pathway (Section 4.2.2). In this part of the study, I consider all non-disease specific pathways in Reactome.

Cancer related genes play central roles in pathways, as captured by their graphlet eigencentrality

I assess if cancer driver genes tend to have central positions in pathways by performing the following analysis. For each pathway, I predict its genes to be cancer-related according to their pathway centrality. I consider the set of cancer driver genes provided by intOGen as the set of true positives (Section 4.1.6).

For the PPI network, I observe that local and global graphlet eigencentralities approaches perform better than the expected AUC-ROC of 0.5 in case of random prediction accuracy, with median AUC-ROC scores over all pathways typically over 0.60, for each of the different underlying graphlets. I observe that global graphlet eigencentrality consistently outperforms local graphlet eigencentrality. To explain this, I perform an MWU test comparing the distribution of the number of pathways each cancer driver genes occurs in, with the distribution of the number of pathways each non-cancer driver gene occurs in. I find that, in the human PPI network, cancer driver genes occur on average in almost twice as many pathways as non-cancer driver genes (10.56 compared to 6.07), which is statistically significantly different with a p-value of $5.19\text{E}{-20}$. Therefore, while cancer genes tend to have a central role in pathways (as indicated by the local graphlet eigencentralities), the results also suggest that they play a more critical role in the crosstalk between the pathways regulating the cell (as indicated by global graphlet eigencentralities). These results are in line with the existing literature, as cancer driver genes have been shown to have a statistically significantly higher betweenness centrality than other genes in the PPI network (Iakoucheva *et al.*, 2002). Looking for specific examples of cancer driver genes playing a role in cancer through crosstalk, I find, for instance, that the crosstalk between cancer driver STAT3 and the p53/RAS signaling pathway controls cancer cell metastasis (Liang *et al.*, 2019). Similarly, crosstalk between p53 and the IGF-1R/AKT/mTORC1 pathway can lead to chemo resistance (Davaadelger *et al.*, 2016).

I find similar results in the COEX network. In the COEX network, I observe that only global graphlet eigencentrality based on $\widetilde{A_{G_1}}$, $\widetilde{A_{G_3}}$, $\widetilde{A_{G_4}}$ and $\widetilde{A_{G_6}}$ performs better than the expected AUC-ROC of 0.5 in case of random prediction accuracy, achieving median AUC-ROC scores over all pathways of at least 0.60. For these graphlet adjacencies, I achieve a median AUC-PR 0.23 in all four cases. Again I observe that the global approach greatly outperforms the local approach in terms of median AUC-PR as well as median AUC-ROC. I validate that cancer driver genes occur in statistically significantly more

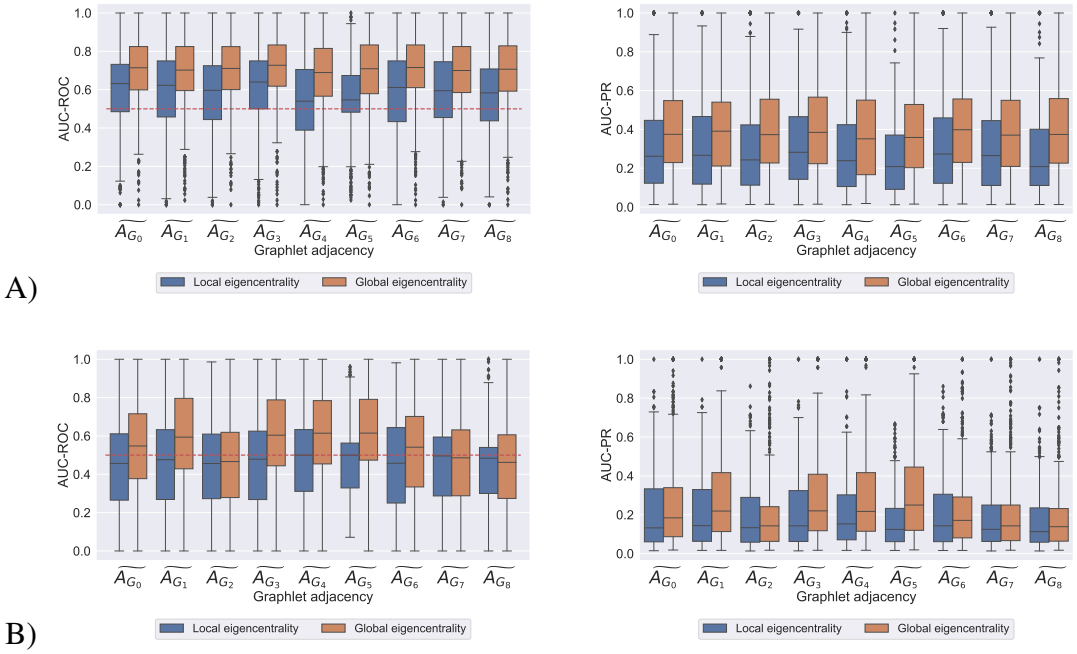


Figure 4.6: Cancer-related gene prediction accuracy. For the human PPI network (A) and the human COEX network (B), I show the distribution of cancer-related gene prediction accuracies over all pathways as box plots, measured using AUC-ROC (left, y-axis) and AUC-PR respectively (right,y-axis), applying local and global graphlet eigencentrality (colour, see legend), applied on different types of graphlet adjacencies (x-axis), in the human PPI network. A dashed red line at 0.5 indicates the expected AUC-ROC in case of random performance.

pathways than non-driver genes. As before, I apply a one sided Mann–Whitney U test in which I compare the distribution the number of pathways driver genes occur in, with the distribution of the number of pathways non-driver genes occur in. Doing so, I achieve a significant p-value 3.44E−13. On average, cancer driver genes occur in 6.31 different pathways, whereas non cancer driver genes occur in only 4.64 different pathways.

I conclude global eigencentrality is the best approach for finding pathways in which cancer-related genes play a central role.

I consider the study of how cancer related genes interact between pathways as future work and will be focussing on illustrating how graphlet eigencentralities capture pathway mechanisms within pathways.

Different graphlet adjacencies allow for the discovery of different cancer-related genes.

First, I focus on those pathways *described* by central cancer driver genes, i.e. those pathways for which I achieve a normalized AUC-PR score larger than 3 applying local graphlet eigencentralities. Additionally, I determine for each pathway a set of *correctly*

predicted cancer-related genes. For each pathway, I determine the threshold such that the F1 score for predicting cancer drivers in that pathway is maximal and consider all the known cancer driver genes with a centrality score higher than this threshold as correctly predicted cancer-related genes. In Figure 4.7, I show the pairwise Jaccard indices between the sets of correctly predicted genes uncovered based on different graphlet adjacencies. With an average Jaccard index of 0.30 in the human PPI network and 0.45 in the human COEX network, I conclude that different graphlet adjacencies describe the role in cancer of different sets of cancer related genes.

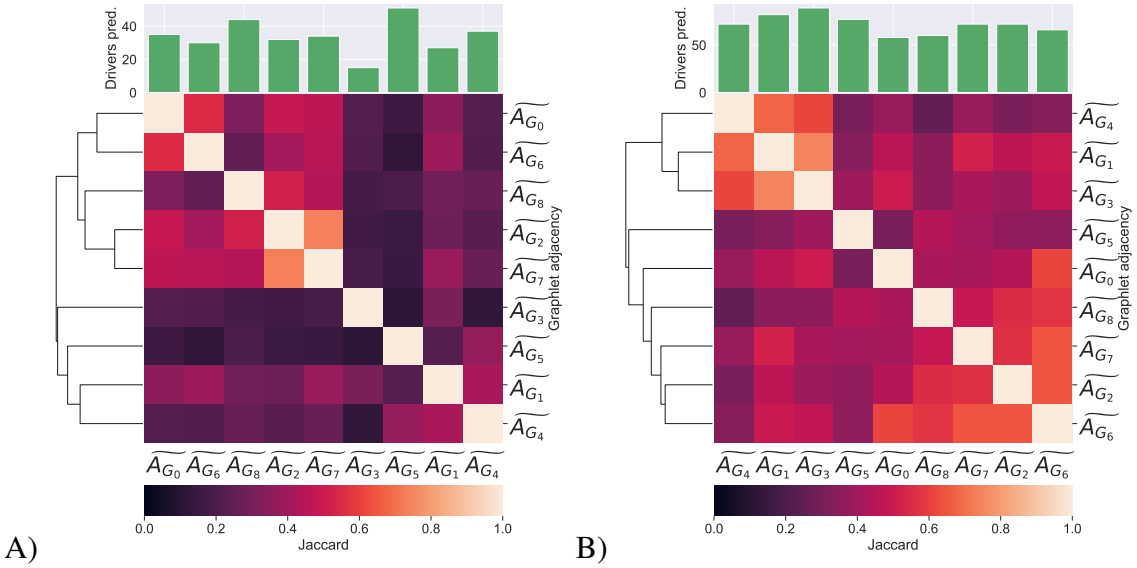


Figure 4.7: The overlap between correctly predicted cancer genes in pathways based on different graphlet eigencentralities. For the human PPI (A) and the human COEX network (B), a clustered heat map of the Jaccard similarity indices between the sets of correctly predicted cancer genes found in pathways described by central driver genes based on different types of graphlet adjacencies. At the top of each heat map, the bar-chart indicates the number of correctly predicted genes corresponding to each graphlet adjacency.

Case study: Formation of Senescence-Associated Heterochromatin Foci (FSAHF)

The formation of senescence-associated heterochromatin foci (FSAHF), contributes to senescence (permanent interruption of cell division) by repressing the expression of proliferation-promoting genes through reorganisation of chromatin (Narita *et al.*, 2003). Cellular senescence plays a vital role in permanently restricting the propagation of damaged and defective cells and forms a natural tumour-suppressor mechanism. I found the cancer mechanism in the FSAHF pathway to be described by graphlet adjacency $\widetilde{A_{G_6}}$ (normalised AUC-PR 3.2) and poorly described by $\widetilde{A_{G_0}}$ (normalised AUC-PR -0.56). I

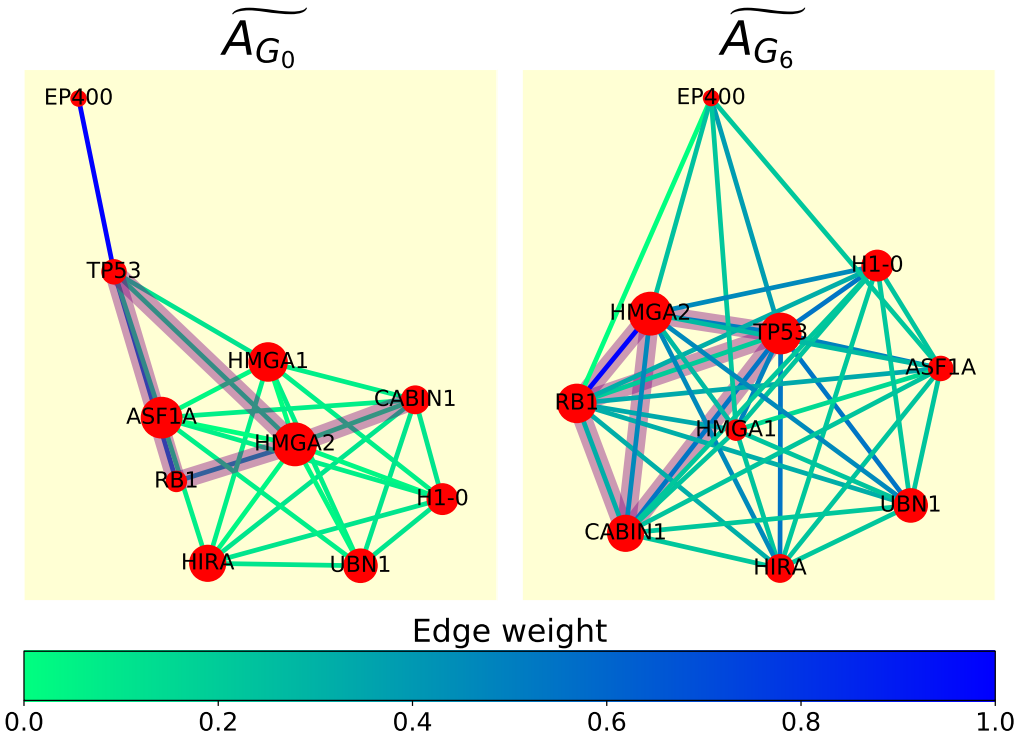


Figure 4.8: Graphlet adjacency $\widetilde{A_{G_6}}$ captures centrality of cancer driver genes in the FSAHF pathway. Spring embedding of FSAHF based on normalised graphlet adjacency $\widetilde{A_{G_0}}$ (left) and $\widetilde{A_{G_6}}$ (right), where nodes represent non-cancer-related genes (red) and known cancer driver genes RB1 and TP53 (yellow), and edges represent weighted normalised graphlet adjacency (see legend). Graphlet G_6 is indicated in translucent purple in the spring embedding based on G_0, connecting genes RB1, TP53, HMGA2 and CABIN1. The subnetwork obtained by inducing these same nodes is also indicated in translucent purple in the spring embedding based on graphlet adjacency $\widetilde{A_{G_6}}$. Although only connected via HMGA2 when considering regular adjacency, TP53 and RB1 are directly connected to CABIN1 in the spring embedding based on graphlet adjacency $\widetilde{A_{G_6}}$, illustrating how graphlet adjacencies can capture functionally relevant indirect relationships between nodes.

will be focusing on this pathway to explain how graphlet adjacencies can capture different cancer mechanisms in pathways.

In Figure 4.8, I show the spring embedding of the SAHF formation pathway-based on normalised graphlet adjacency $\widetilde{A_{G_0}}$ and $\widetilde{A_{G_6}}$. From the perspective of graphlet adjacency $\widetilde{A_{G_0}}$, cancer drivers RB1 and TP53 do not play a central role in this pathway, as they appear peripheral to the other nodes in the pathway. The mediating role of TP53 and RB1 trough hub node HMGA2 is well captured by graphlet adjacency $\widetilde{A_{G_6}}$, connecting them with all nodes in the pathway. Additionally, through literature curation, I find that HMGA2, the most central node in the pathway according to graphlet adjacency $\widetilde{A_{G_6}}$ and predicted as cancer-related in Section 4.2.2, is also a driver of tumour metastasis (Morishita *et al.*, 2013). I conclude that graphlet eigencentrality enables considering different notions of the centrality of genes in pathways, allowing the capturing of different func-

tional roles of genes in pathways. As centrality measures are widely used to uncover disease-related genes and graphlet eigencentrality captures notions of centrality different from those based on regular adjacency, graphlet eigencentrality opens up the opportunity of uncovering novel disease related genes.

4.3 Conclusion

In this chapter, I introduce graphlet eigencentrality, which captures different notions of the centrality of nodes in a network. I apply it on measuring the centrality of genes in pathways, enabling a detailed investigation of how different graphlet adjacencies capture different biological functions. I apply this method at two levels: from the local pathway perspective or the global network perspective.

I apply graphlet eigencentralities to identify pathways described by different graphlet adjacencies, i.e. all genes participating in a pathway are also important from the topological perspective. To do so, I use graphlet eigencentralities to predict which genes belong to a given pathway, considering the pathways for which I achieve the highest prediction accuracies as being described by that graphlet adjacency. I find that local pathway-based graphlet eigencentralities well predict which genes participate in a given pathway, outperforming state-of-the-art predictor GeneMANIA. To explain this result, I show that pathways, even when functionally unrelated, show large amounts of overlap (see Section 4.2.1). As the local approach considers each pathway as an individual entity disentangled from the full network, it is able to best capture the topological essence of a pathway. I go on to show that pathways that are described by a given graphlet adjacency are biologically functionally similar in terms of the ancestral, GO-BP, GO-CC and GO-MF terms in which they are enriched, and that these functional similarities depend on the graphlet adjacency. I illustrate these results by a case study of the 'Receptor mediated mitophagy' pathway, where I show how graphlet adjacency $\widetilde{A_{G_6}}$ captures the hub-role of FUNDC1, allows to capture the functional organisation of the pathway.

Secondly, I apply graphlet eigencentrality at predicting cancer-related genes in pathways. I observe that global graphlet eigencentrality consistently outperforms local graphlet eigencentrality. To explain this result, I show that cancer driver genes participate in statistically significantly more pathways than non-cancer-related genes. Therefore, while cancer genes tend to have central roles in pathways (as indicated by local graphlet eigencentralities), the results also suggest that they play a more essential role in the crosstalk that occurs between pathways to regulate the cell (as indicated by the global

graphlet eigencentralities). This is a key insight, as it indicates that pathway-focused approaches for studying cancer should focus on the interactions between pathways, although most current state-of-the-art approaches focus on their individual differential expression or rewiring. In chapter 5, I build on this observation to predict cancer implicated pathways based on their changing pathway-pathway interactions.

Additionally, I show that cancer genes that can be uncovered by their pathway centrality are different depending on the graphlet eigencentrality. I illustrate these results by a case study of the FSAHF pathway, where I show how graphlet adjacency captures the central roles of cancer driver genes, RB1 and TP53. I conclude that graphlet eigencentralities capture different functional roles of genes in and between pathways.

Chapter 5

Pathway-driven NMTF captures the reorganisation of pathways in cancer

In chapter 3, I introduced graphlet adjacency, a graphlet based generalization of regular adjacency designed to capture the higher order wiring of nodes in local network neighbourhoods. I showed that graphlet laplacians capture topology-function and topology-disease relationships in biological networks. Subsequently in chapter 4, I took a more descriptive pathway focused approach to investigate how different graphlet adjacencies capture different biological functions and disease mechanisms. I showed that different graphlet adjacencies better capture the different functional roles of genes in pathways. During the pathway focused analysis, I observed that driver genes characteristically perform hub roles between pathways. Therefore, I hypothesise that cancer pathways should be identified by changes in their pathway-pathway relationships, which is the focus of this chapter.

In this chapter, to learn an embedding space that captures the relationships between pathways in a healthy cell, I propose pathway-driven non-negative matrix tri-factorisation (PNMTF). I apply PNMTF to predict 15 genes and pathways involved in four major cancers, predicting 60 gene-cancer associations in total, covering 28 unique genes. To further exploit driver genes' tendency to perform hub roles, I model the network data using graphlet-adjacency, which considers nodes adjacent if their interaction patterns form specific shapes (e.g., paths or triangles). I find that the predicted genes rewire pathway-pathway interactions in the immune system and provide literary evidence that many are druggable (15/28) and implicated in the associated cancers (47/60). I predict six druggable cancer-specific drug targets.

This analysis is summarised in Figure 5.1.

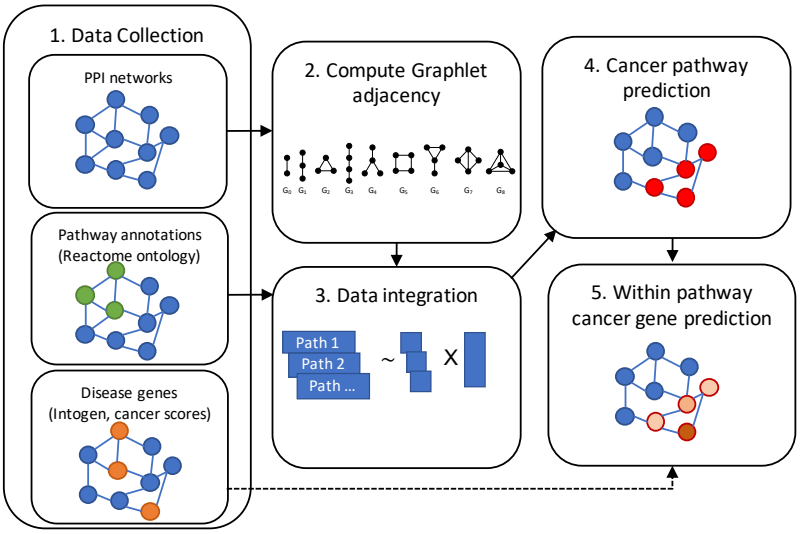


Figure 5.1: Chapter workflow summary. Step 1: collection of the PPI network, Reactome pathways and cancer drivers. Step 2: computation of graphlet adjacencies. Step 3: integrate a graphlet adjacency and the entire pathway ontology. Step 4: predict cancer pathways. Step 5: predict cancer genes in cancer pathways.

Chapter impact

This chapter has lead to the following contributions:

Publications:

Windels, S. F. L., Malod-Dognin, N., and Pržulj, N. (2022). Identifying cellular cancer mechanisms through pathway-driven data integration. *Bioinformatics*, btac493.

Methodological contributions:

To capture the organisation of a network as a composition of subnetworks, I propose pathways-driven non-negative matrix tri-factorisation (PNMTF), which fuses network data and prior domain-specific knowledge assigning nodes to subnetworks in the network. PNMTF allows to measure the topological importance of nodes and subnetworks in a network and the rewiring of their interactions between two different network states.

Biological contributions:

PNMTF allows to predict genes and the pathways involved in four major cancers based on their functional importance in the healthy cell and their changing functional relationships in cancer. By combining PNMTF with graphlet adjacency, I can exploit the tendency of driver genes to perform hub roles and increase prediction accuracy. I provide strong literature support for the top predicted genes, of

which I prioritise six as potential cancer-specific drug targets. These results are in submission.

Software and data analysis:

PNMTF, as well as the data and scripts to perform the analysis presented in this chapter, have been made publicly available at: gitlab.bsc.es/swindels/pnmtf .

5.1 Methods

To capture the higher-order organisation encoded by graphlet adjacencies in a lower-dimensional space, I define the baseline model Global-NMTF (see Section 5.1.1). Then, I extend the NMTF model to Pathway-driven NMTF to benefit from the known functional organisation of pathways in Reactome (see Section 5.1.1). Finally, to identify pathways and genes implicated in cancer, I define the embedding based centrality and moving distance measures, which respectively measure the (topological) importance of a pathway or gene and how much the their functional relationships change between a healthy to a diseased state (see Sections 5.1.2 and 5.1.2).

5.1.1 NMTF models

To capture the functional organisation of the cell as an embedding space, I define the baseline NMTF model called Global NMTF (GNMTF) (see Section 5.1.1). Then, I extend the NMTF model to benefit from the known functional organisation of pathways (see Section 5.1.1). The solvers for both models are presented in Supplementary Section **??**.

Global NMTF (GNMTF) model: a basic approach to learning the organisation of the healthy cell

To learn an embedding space that captures the higher-order functional organisation of a healthy (control) cell as captured by a given graphlet adjacency matrix, $\widetilde{A}_{G_i}$, which decomposes $\widetilde{A}_{G_i}$ as the product of three non-negative matrix factors, $U^{n\times d}, S^{d\times d}$ and $V^{n\times d}$: $\widetilde{A}_{G_i} \approx USV^T$. This corresponds to solving the following optimisation problem:

$$\min_{U,S,V\geq 0} \sum_{i=0}^{8} \left\| \widetilde{A}_{G_i} - USV^T \right\|_F^2, \text{ s.t.: } V^TV = I, \tag{5.1}$$

where F denotes the Frobenius norm. I interpret V as an orthogonal basis that captures the functional organisation of the cell as captured by $\widetilde{A}_{G_i}$, and $E = US$ as embedding all genes in common space V. Each row of E corresponds to the embedding of a gene, which I denote $\vec{g}$, in the space spanned by V. Then, analogous to NLP, where sentences can be represented by the average embedding of their constituent words (Le and Mikolov, 2014), I define the embedding of a pathway, which I denote $\vec{p}$, as the average embedding of its genes: $\vec{p} = \frac{1}{|m_p|}\sum_{\vec{g}\in m_p}\vec{g}$, where m_p is the set of gene embeddings for genes in a given pathway p.

Pathway-driven NMTF (PNMTF) model: improved learning of the organisation of the healthy cell

I extend GNMTF model to benefit from the known functional organisation of pathways in Reactome. PNMTF learns a latent representation for each pathway and an embedding space that organises these discrete latent representations. Specifically, how each pathway p interacts within the healthy cell is encoded by taking the rectangular submatrix, $H_p^{|m_p|\times n}$, induced by the $|m_p|$ genes in the pathway and n genes in the cell on $\widetilde{A}_{G_i}$. Then, PNMTF simultaneously decomposes the H_p-matrices for all of the r pathways in Reactome into r pairs of non-negative latent matrices $U_p^{|m_p|\times 1}$ and $S_p^{1\times d}$ and one orthogonal non-negative latent matrix $V^{d\times n}$: $H_p \approx U_p S_p V^T$ for all $p \in [1,r]$. This corresponds to solving the following optimisation problem:

$$\min_{U_p,S_p,V\geq 0} \sum_{p=1}^{r} \left\| H_p - U_p S_p V^T \right\|_F^2, \text{ s.t.: } V^T V = I, \tag{5.2}$$

$E_p = U_p S_p$ is interpreted as embedding the genes of pathway p in the orthogonal space spanned by V. Each row of E_p corresponds to the embedding of a gene in the (functional) context of a given pathway p, which I denote by $\vec{g}_p$. Analogous to the GNTMF model, I define the embedding of a pathway p as the average embedding of its genes: $\vec{p} = \frac{1}{|m_p|}\sum_{\vec{g}_p \in m_p} \vec{g}_p$, where m_p is the set of gene embeddings for genes in pathway p.

Extending PNMTF: learning representations for cancer-affected pathways

To enable the identification of pathways whose functional relationships change the most in cancer (see Section 5.1.2), the aims is to learn how cancer-affected pathways change their interactions with other pathways. To do so, I fix the basis of the common space V learned in Eqn 5.2 based on the control PPI network, and solve PNMTF based on the case (cancer) PPI network to learn a second latent representation for each pathway, this time in a diseased state.

5.1.2 NMTF scores for cancer predictions

To identify pathways and genes implicated in cancer, I define three heuristics based on PNMTF pathway and gene embeddings.

NMTF centrality

Here I define how to measure the topological importance of a pathway or gene based on its embedding. To do so, I take inspiration from the eigencentrality, which considers a node important if it is highly connected to other highly connected nodes, i.e, if it is part of a cluster of nodes in the network. For a formal definition, see Section 2.1.2.

In NMTF, the left and right latent matrices' rows can also be interpreted as cluster-indicator vectors, where the entity corresponding to the row is assigned to the cluster corresponding to the column containing the highest valued entry. Therefore, following the proposition that an entity is considered central if it is part of one or more clusters, I measure the centrality of a pathway or gene by the Euclidean norm of its embedding:

$$\text{NMTF centrality}(\vec{E}) = \left\|\vec{E}\right\|_2, \tag{5.3}$$

where $\vec{E}$ is the embedding of a healthy pathway (i.e. $\vec{P}$) or gene (i.e. $\vec{G}$) (see Section 5.1.1).

Moving distance

Here I define the *moving distance*, which measures how a pathway's or gene's functional relationships change when moving from a healthy to a diseased state. To do so, I take the Manhattan distance between a pathway's or gene's embedding in a healthy and diseased state (see Sections 5.1.1 and 5.1.1):

$$\text{moving distance}(\vec{E}_1, \vec{E}_2) = \left\|\vec{E}_1 - \vec{E}_2\right\|_1, \tag{5.4}$$

where $\vec{E}_1$ and $\vec{E}_2$ are the embeddings of a pathway or gene in a healthy and cancerous state, respectively (see Sections 5.1.1 and 5.1.1).

Hybrid score

I use the geometric mean to combine the centrality and moving distance:

$$\text{hybrid}(\vec{E}_1, \vec{E}_2) = \sqrt{\text{NMTF centr.}(\vec{E}_1) * \text{mov.dist.}(\vec{E}_1, \vec{E}_2)}, \tag{5.5}$$

where $\vec{E}_1$ and $\vec{E}_2$ are the embeddings of a pathway or gene in a healthy and cancerous state, respectively (see Sections 5.1.1 and 5.1.1).

5.1.3 Measuring prediction accuracy

I apply three different NMTF-scores to predict cancer implicated pathways and genes in Sections 5.2.2 and 5.2.3, respectively. As wet-lab validation is expensive, we are predominantly interested in the top-scoring entities that are highly likely to be cancer implicated for both types of predictions. So, for both types of predictions, I consider the set of top-scoring entities as a prediction and use the Matthew Correlation Coefficient (MCC) to measure the prediction accuracy:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}, \quad (5.6)$$

where TP is he number of true positives, TN is the number of true negatives, FP is the number of false positives and FN is the number of false negatives. For pathway focused predictions, I use the set of known 'cancer pathways' in Reactome as the set of true positives (see Section 5.1.4). For gene focused predictions, I use the set of known cancer drivers in COSMIC as the set of true positives (see Section 5.1.4). The MMC ranges from -1 to 1, where 1 indicates a perfect prediction, 0 indicates random performance and -1 indicates an inverse prediction.

5.1.4 Data

Case and control protein-protein interaction (PPI) networks

I create four pairs of case and control PPI networks (i.e. cancerous and healthy), one pair for each of the four cancers considered. First, I create a generic PPI network by combining the experimentally validated PPI, only those captured using Two-hybrid or Affinity Capture based methods, from BioGRID version 4.4.197 (Stark *et al.*, 2006) and the PPI from Reactome (Jassal *et al.*, 2019). Then, I overlay the RNA-SEQ gene expression data for four cancer cell lines and their corresponding control tissue from the Human Protein Atlas, on the generic PPI network (Uhlén *et al.*, 2015). Prostate cancer (cell line PC-3), lung cancer (cell line A549), colon cancer (cell line CACO-2) and ovarian cancer (cell line EFO-21) are considered. Basic network statistics are presented in Tables 5.1 and 5.2.

	Nodes	Edges	Density
Lung Case	11,635	274,202	0.41%
Lung Control	13,590	311,754	0.34%
Colon Case	11,298	267,833	0.42%
Colon Control	13,480	316,247	0.35%
Prostate Case	11,651	275470	0.41%
Prostate Control	13,654	312,201	0.33%
Ovary Case	12,027	286,596	0.40%
Ovary Control	12,626	288,960	0.36%

Table 5.1: Details case and control PPI networks. The number of nodes, the number of edges and the density (columns 1-3, respectively), for each of the case and control PPI networks (rows).

	Node overlap	Edge overlap
Lung	0.78	0.80
Colon	0.78	0.80
Prostate	0.79	0.79
Ovary	0.81	0.80

Table 5.2: Overlap between case and control networks. For each case and control network (rows), the overlap in terms of nodes and edges measured using the Jaccard Index (columns).

The Reactome Pathway Ontology

The Reactome Ontology is a collection of 23 directed acyclic graphs (DAG's), encoding the relationships between 2,516 pathway annotations from the most generic to the most specific (Jassal *et al.*, 2019). For each of the four pairs of case and control networks, I determine the set of pathways that induce a subnetwork of at least 10 and up to 100 nodes on either of the networks. The number of pathways per pair of networks (case and control) is presented in Table 5.4.

	No. of pathways	**Coverage of nodes in union of case and control PPI**
Lung	1,025	5,266 / 14,118 (37.30%)
Colon	1,027	5,269 / 13,908 (37.88%)
Prostate	1,024	5,227 / 14,096 (37.09%)
Ovary	1,002	5,100 / 13,627 (37.43%)

Table 5.3: Pathway statistics. For each of the four tissues considered (rows), I indicate the number of pathways considered (column 1) and how many genes they annotate in the union of the case and control network for the corresponding tissue.

Cancer annotation data

For the pathways and genes considered for each cancer type (see Sections 5.1.4 and 5.1.4), I collect cancer annotation data. At the pathway level, I collect 'cancer pathway'-annotations from Reactome, which indicate if a given pathway is considered to be a cancer pathway. At the gene level, I collect driver genes from the COSMIC database (Tate *et al.*, 2019). I consider a gene to be a cancer driver if it is a known cancer driver in at least one cancer type, with strong evidence (i.e., 'Tier 1') in the literature. Also, I collect a set of tissue-specific prognostic genes from the Pathology Atlas (Uhlen *et al.*, 2017). The number of cancer pathways, driver genes and prognostic genes per cancer are presented in Table 5.3.

	Cancer pathways	**Driver genes**	**Prognostic genes**
Lung	61	652	614
Colon	61	647	575
Prostate	61	647	152
Ovary	61	632	475

Table 5.4: Cancer annotation data statistics. For each tissue, the number of cancer pathways, driver genes and prognostic genes considered.

5.2 Results and discussion

I apply PNMTF to uncover novel pathways and genes involved in lung, colorectal, prostate and ovarian cancer. Specifically, for a given cancer type, I construct a case and a control network, representing a cancerous and a healthy cell (see Section 5.1.4). For the case and control networks, I compute the different graphlet adjacency matrices for graphlets up to four nodes (see Section 3.1.1). Then, for a given graphlet adjacency, I learn the higher order functional organisation of the healthy cell as an embedding space using PNMTF, into which I embed pathways and genes (see Section 5.1.1). Next, in this same space, I also compute embeddings for pathways and genes of a cancer affected cell, by fixing the basis trained for the control cell and solving PNMTF for the case PPI network (see Section 5.1.1). Finally, having computed a pair of embeddings for each pathway and gene based on the cell's healthy and cancerous state, I apply three NMTF-scores: NMTF centrality, moving distance and hybrid score (see Section 5.1.2) to predict their cancer relatedness.

In my analysis, I first validate that PNMTF captures the functional organisation of pathways in the cell (Section 5.2.1). Then I show that using the NMTF-scores I can prioritise pathways and genes implicated in cancer (Sections 5.2.2 and 5.2.3). Finally, for each of the four cancers, I validate the top 15 predicted cancer genes and pathways involved in the literature (i.e., predicting 60 cancer-specific gene-pathway pairs in total, see Section 5.2.4).

5.2.1 PNMTF captures the functional organisation of the cell described by the Reactome pathway ontology

First, I validate that PNMTF best captures the functional organisation of pathways in the healthy (control) cell, compared to GNMTF (essentially a standard NMTF model). To do so, for a given control network and graphlet adjacency, I train PNMTF and GNMTF (see Sections 5.1.1 and 5.1.1), embed all pathways in the shared space V and apply agglomerative hierarchical clustering on their pairwise Euclidean distances. Then, I confirm that pathway embeddings based on PNMTF form better separable and more functionally coherent clusters than those based on GNMTF.

Intrinsic quality of pathway-clusters in space: cophenetic correlation

For a given network and graphlet adjacency, to measure if the pathway embeddings form dense and well-separable clusters in the embedding space (i.e. have high intrinsic cluster-

ing quality), I first embed all pathways in shared space, apply agglomerative hierarchical clustering and measure the intrinsic quality of the hierarchical clustering using the cophenetic correlation (i.e. the Pearson correlation between the cosine distance between two pathways and the height in the linkage tree where their corresponding branches meet). I present the results for lung cancer based on graphlet adjacency $\widetilde{A}_{G1}$ in Figure 5.2. I observe that the agglomerative clustering uncovers a better separable clustering when applying PNMTF than GNMTF (cophenetic correlation 86.5% vs 66%).

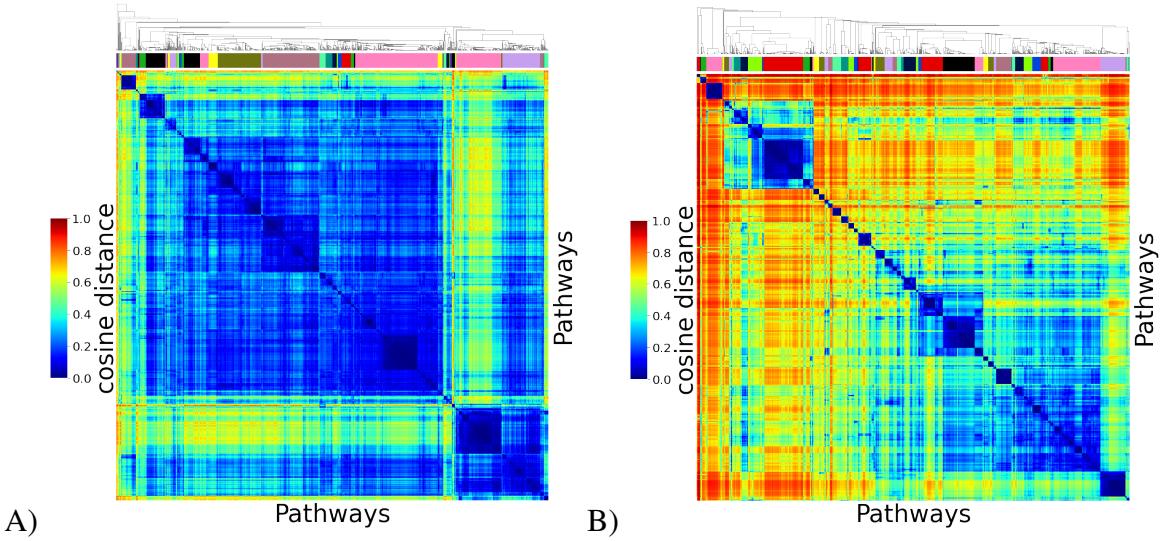


Figure 5.2: PNMTF best captures the functional organisation of pathways in the healthy lung cell. (A) and (B) show a clustered heat map of the pairwise cosine distances between all pathway embeddings in the shared space V learned based on graphlet adjacency $\widetilde{A}_{G_1}$ by GNMTF and PNMTF, respectively. For each heat map, the colour bar under the hierarchical tree on the top indicates the 65 pathway clusters.

Results averaged over the four control networks across all graphlet adjacencies are presented in Figure 5.3. I observe that, with the exception of when based on graphlet adjacency $\widetilde{A}_{G_5}$, PNMTF outperforms GNMTF (average cophenetic correlation 89% compared to 83%). The best intrinsic clustering quality, averaged over the 4 control networks, is measured based on graphlet adjacency $\widetilde{A}_{G_8}$ using PNMTF at 96%. I conclude that as PNMTF based pathway embeddings show the clearest clustering structure in the embedding space.

Extrinsic quality of pathway-clusters in space: enrichment analysis

For a given network and graphlet adjacency, to measure how well both methods group functionally related pathways, I extract 65 clusters from both hierarchical clusterings (I determine this is the optimal number of clusters applying an elbow method in Supplementary Section B.1), and check their enrichment in pathway *ancestors*, less specific

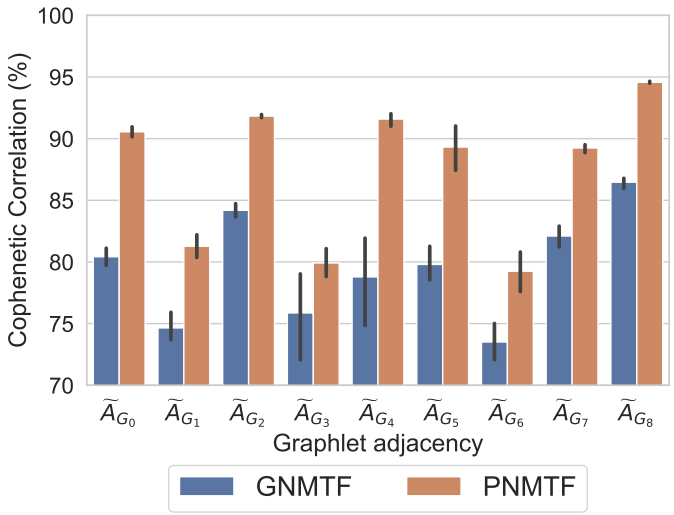


Figure 5.3: PNMTF best captures the functional organisation of pathways in the cell. For GNMTF and PNMTF, I present the cophenetic correlation averaged over the four tissues (y-axis) for the pathway embeddings. The distance between pathways is measured using 'cosine distance'. Hierarchical clustering is performed using 'average linkage'.

pathways higher up in the Reactome ontology (see ancestor-pathway enrichment in the previous chapter, Section 4.1.5). The results averaged over the four control networks across the different graphlet adjacencies are presented in Figure 5.4. Averaged over the four control networks, I observe that pathway clusters in the shared space trained using PNMTF are at least as much or more enriched in ancestor annotations as pathway clusters using GNMTF; this is true across all graphlet adjacencies in terms of percentage of clusters enriched (on average 94% compared to 89%), percentage of ancestor pathway annotations enriched (on average 94% compared to 84%) and percentage of pathways with at least one ancestor enriched (on average 76% compared to 72%). This means that PNMTF outperforms GNMTF in capturing the functional organisation of pathways as described by Reactome.

In conclusion, compared to GNMTF, PNMTF based pathway embeddings form clusters that are better separable (indicated by the high cophenetic correlation coefficient) and more functionally coherent (indicated by the high percentage of ancestor enriched clusters), hence I conclude that PNMTF better captures the functional organisation of pathways in the (healthy) cell than the standard GNMTF model.

5.2.2 PNMTF identifies pathways implicated in cancer

Having validated that PNMTF captures the functional organisation of pathways in the healthy cell, I assess if the three NMTF-scores, centrality, moving distance and hybrid score (defined Sections 5.1.2-5.1.2) can be used to prioritise pathways implicated in can-

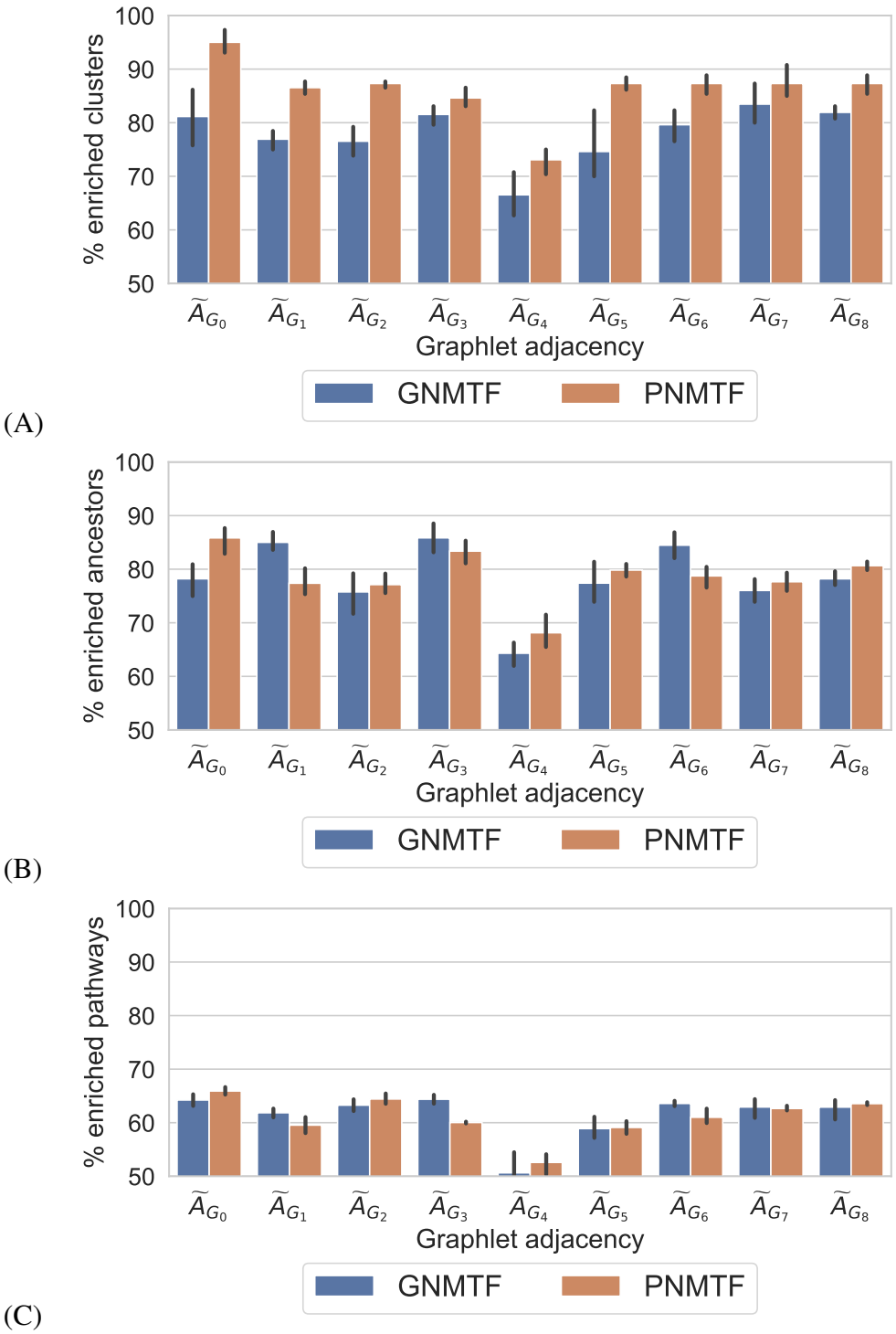


Figure 5.4: Pathway clustering ancestor enrichment analysis. For the GNMTF model and PNMTF model, I apply ancestor enrichment analysis on the pathway clusters and report: the percentage of clusters that contain at least one enriched ancestor annotation, the percentage of ancestor annotations that are enriched in at least one cluster and the percentage of pathways that have at least one ancestor annotation enriched.

cer. Specifically, for a given NMTF-score, cancer and graphlet-adjacency, I measure the Matthews Correlation Coefficient (MCC) using the set of known cancer pathways in Reactome as a gold standard and a set of top-scoring pathways for each method as predictions for pathways implicated in cancer (see Section 5.1.3). To determine the set of top-scoring pathways for each cancer type, graphlet-adjacency and NMTF-score, I apply an elbow method. The results are presented in Supplementary Figure B.2. As all three NMTF-scores plateau beyond 100 pathways, regardless of the cancer type and graphlet adjacency, I consider the top 100 highest scoring pathways as the prediction set for pathways implicated in cancer. Applying a hypergeometric test, I find that this set of pathways is enriched in Reactome cancer pathways (least significant p-value $\approx$ 4.67e-08). Additionally, I acknowledge that many pathways not labelled as cancer pathways in Reactome might overlap with cancer-mechanisms. For that reason, I also consider the ratio of driver genes in a pathway as an indication of a its engagement in cancer. Then, to evaluate a given pathway prediction method, I measure the Spearman rank correlation between this ratio and a pathway's score.

First, I compare the results for all different graphlet adjacencies, averaged over the four cell types, in Figure 5.5. I observe that in terms of MCC (Fig. 5.5.A), the best performance is achieved when using the moving distance and regular adjacency (0.244), just outperforming the hybrid score with graphlet adjacencies A_{G_1},A_{G_3}, and A_{G_6} (0.237, 0.220 and 0.233, respectively). Looking at the correlation results (Fig. 5.5.B), I find that the hybrid score with graphlet adjacencies A_{G_1}, A_{G_3}, and A_{G_6} greatly outperform the moving distance with regular adjacency (0.421, 0.443 and 0.461, compared to 0.312). Based on these results, I chose to focus on graphlet adjacencies AG_0, AG_1, AG_3 and AG_6 for the comparison against the state-of-the-art in the main paper.

Next, I compare the results based on these top-scoring graphlet adjacencies and those based on pathway prediction methods 'ORA' (which despite its simplicity is still widely used) and 'ANUBIX' (a state-of-the-art crosstalk enrichment method), see Figure 5.6. I can not compare against FCS and PTB methods as they require multiple case and control samples for a given cancer type. Firstly, I observe that in terms of MCC (see Figure 5.6.A), the best performance is achieved using the moving distance and regular adjacency (0.244), just outperforming hybrid score with graphlet adjacencies A_{G_1},A_{G_3}, and A_{G_6} (0.237, 0.220 and 0.233, respectively). All of NMTF-scores outperform ORA (-0.054) and ANUBIX (0.032). This renders the PNMTF scores more practical for further downstream analysis (i.e., gene focused predictions) than ORA and ANUBIX, as the top ranked

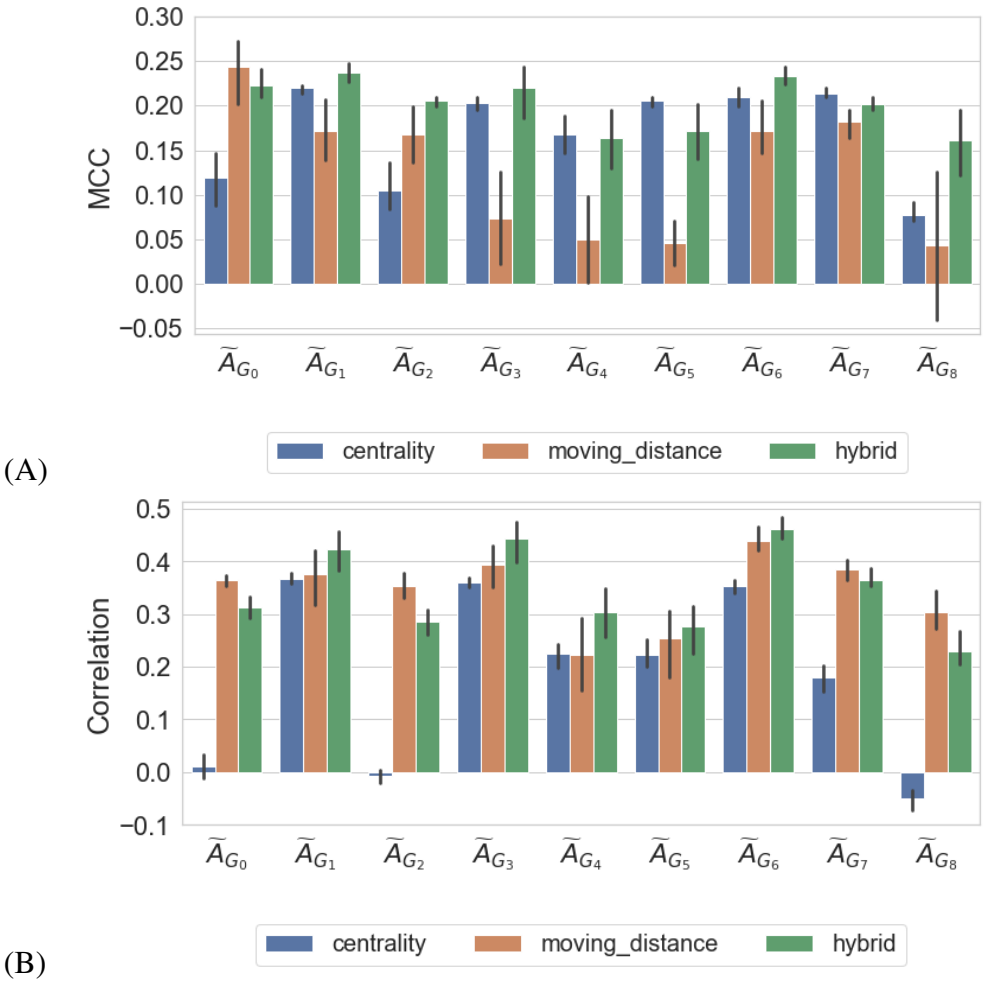


Figure 5.5: Comparing cancer pathway prediction accuracy for different NMTF-scores across graphlet adjacencies. Sub-plots (A) and (B) respectively show the MCC and rank-correlation scores for predicting Reactome cancer pathways. From left to right, I compare results for PNMTF based on different graphlet adjacencies (x-axis) and different NMTF-scores (legend).

pathways are more likely to be cancer related. Looking at the correlation results (see Figure 5.6.B), I find that the hybrid score based on graphlet adjacencies A_{G_1},A_{G_3}, and A_{G_6} greatly outperform the moving distance with regular adjacency (0.42, 0.443 and 0.461, compared to 0.363). I also observe ANUBIX scores drastically better in terms of correlation (0.302) than in terms of MMC, which indicates that ANUBIX is able to rank pathways according to their likely involvement in cancer in general, although the set of top 100 highest ranked pathways is not particularly enriched in cancer pathways. I consider the hybrid score-based on graphlet adjacency A_{G_1} as the best approach, as it is only marginally behind the moving distance with A_{G_0}, the best method in terms of MMC, but greatly outperforms this method in terms of rank correlation. Finally, I note that the highest scoring graphlet adjacencies, A_{G_1},A_{G_3}, and A_{G_6} happen to be based on graphlets capturing betweenness and hubness, suggesting that cancer-related pathways tend to have

hub-roles. This is in line with the previous results, where I observed that cancer driver genes occur in statistically significantly more pathways than non-driver genes (Windels *et al.*, 2022).

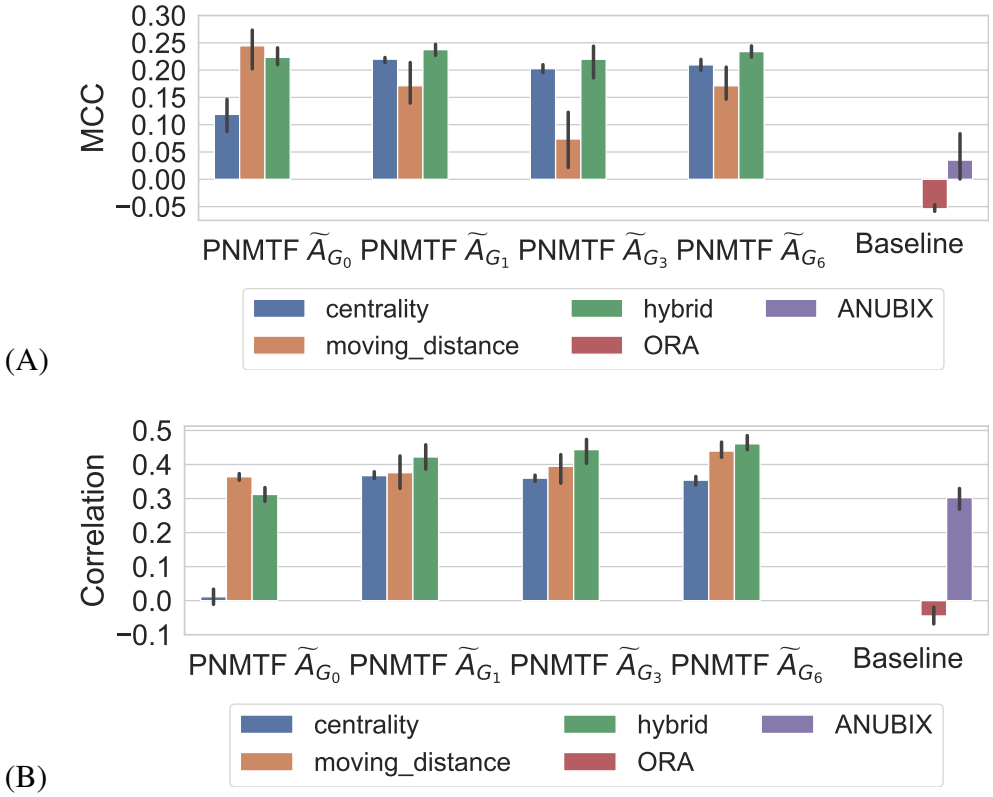


Figure 5.6: PNMTF identifies pathways implicated in cancer. Sub-plots (A) and (B) show the MCC and rank-correlation scores for predicting Reactome cancer pathways, respectively. From left to right, I present the results for PNMTF based on different graphlet adjacencies (x-axis) and different NMTF-scores (legend), against the state-of-the-art (far right).

To further validate the hybrid score captures cancer-implicated pathways, I present the top 10 highest scoring pathways in each of the four tissues, using the hybrid score based on graphlet adjacency A_{G_1}, in Tables 5.5 to 5.8. I observe, for each of the four tissues, that between 5 and 7 out of 10 pathways are cancer pathways. For all four tissues, I observe that all top 10 pathways are related to the RAS-MAPK pathway, which transduces extracellular signals to the cell nucleus, regulating cell growth, cell division and cell repair. The RAS-MAPK pathway is frequently associated with oncogenesis, tumour progression and drug resistance, and is a frequent subject of therapeutic studies (Braicu *et al.*, 2019). Given that a high ratio of prioritised pathways is a known cancer pathway and the fact that the prioritised pathways are involved in mechanisms known to be rewired in cancer, I am confident that the remaining prioritised pathways that are not known as cancer pathways could be indeed cancer related.

Rank	Pathway	Known cancer pathway
1	Constitutive Signaling by EGFRvIII	✓
2	Signaling by EGFRvIII in Cancer	×
3	Signaling by ERBB2 ECD mutants	✓
4	GAB1 signalosome	×
5	PI3K events in ERBB2 signaling	×
6	HSF1 activation	×
7	Constitutive Signaling by Ligand-Responsive EGFR Cancer Variants	✓
8	Signaling by Ligand-Responsive EGFR Variants in Cancer	✓
9	SHC1 events in EGFR signaling	×
10	Signaling by ERBB2 KD Mutants	✓

Table 5.5: Top 10 highest scoring pathways in lung cancer. The table ranks pathways according to their hybrid score using PNMTF based on graphlet adjacency A_{G_1}. The final column indicates if the pathway (2nd column) is a known cancer pathway in Reactome.

Rank	Pathway	Known cancer pathway
1	Constitutive Signaling by EGFRvIII	✓
2	Signaling by EGFRvIII in Cancer	✓
3	HSF1 activation	×
4	Signaling by ERBB2 ECD mutants	✓
5	Constitutive Signaling by Ligand-Responsive EGFR Cancer Variants	✓
6	Signaling by Ligand-Responsive EGFR Variants in Cancer	✓
7	Signaling by FGFR4 in disease	✓
8	Signaling by FGFR3 fusions in cancer	✓
9	Downstream signaling of activated FGFR4	×
10	Role of LAT2/NTAL/LAB on calcium mobilization	×

Table 5.6: Top 10 highest scoring pathways in colon cancer. The table ranks pathways according to their hybrid score using PNMTF based on graphlet adjacency A_{G_1}. The final column indicates if the pathway (2nd column) is a known cancer pathway in Reactome.

5.2.3 PNMTF identifies genes implicated in cancer

Having shown that NMTF-scores can identify pathways implicated in cancer, I move on to find cancer-related genes within the set of 100 top-scoring pathways for each cancer type. To identify a set of top scoring genes, I apply an elbow method to the three NMTF-scores: centrality, moving distance and hybrid score (defined Sections 5.1.2-5.1.2). The results are presented in Supplementary Figure B.3. I observe that the gene scores plateau beyond the top 100 scoring genes, hence I choose to focus on the top 100 highest scoring genes in the previously identified set of top 100 highest scoring pathways (see Supplementary Section B.4). Then, I measure the MCC score using the set of top-scoring genes as a

Rank	Pathway	Known cancer pathway
1	Signaling by EGFRvIII in Cancer	✓
2	Constitutive Signaling by EGFRvIII	✓
3	HSF1 activation	×
4	Signaling by ERBB2 ECD mutants	✓
5	Constitutive Signaling by Ligand-Responsive EGFR Cancer Variants	✓
5	Signaling by Ligand-Responsive EGFR Variants in Cancer	✓
7	GAB1 signalosome	×
8	Constitutive Signaling by Overexpressed ERBB2	✓
9	Downstream signaling of activated FGFR4	×
10	Signalling to RAS	×

Table 5.7: Top 10 highest scoring pathways in prostate cancer. The table ranks pathways according to their hybrid score using PNMTF based on graphlet adjacency A_{G_1}. The final column indicates if the pathway (2nd column) is a known cancer pathway in Reactome.

Rank	Pathway	Known cancer pathway
1	Signaling by EGFRvIII in Cancer	✓
2	Constitutive Signaling by EGFRvIII	✓
3	GAB1 signalosome	×
4	Signaling by ERBB2 ECD mutants	✓
5	Transcriptional regulation by the AP-2 (TFAP2) fam. of TF.	×
6	Signaling by EGFR in Cancer	✓
7	Downstream signaling of activated FGFR3	×
8	Retrograde neurotrophin signalling	×
9	Constitutive Signaling by Ligand-Responsive EGFR Cancer Variants	✓
10	Signaling by Ligand-Responsive EGFR Variants in Cancer	✓

Table 5.8: Top 10 highest scoring pathways in ovarian cancer. The table ranks pathways according to their hybrid score using PNMTF based on graphlet adjacency A_{G_1}. The final column indicates if the pathway (2nd column) is a known cancer pathway in Reactome.

prediction and the driver genes in COSMIC as the gold standard (see Section 5.1.4). I compare PNMTF for graphlet adjacency $\widetilde{A}_{G_1}$ against: PNMTF with regular adjacency, graphlet eigencentrality for $\widetilde{A}_{G_1}$ (which predicts cancer genes based on their topological importance, see Section 5.1.2) and network diffusion for $\widetilde{A}_{G_1}$ (which predicts genes as cancer related if they are in the network neighbourhood of differentially expressed genes, see Section 2.2.2. I tune diffusing parameter α to 1.9, as it leads to the highest MCC scores when ranging α from 0.1 to 2.0 in increments of 0.1). The results are presented in Figure 5.7.

I observe that by using PNMTF based on $\widetilde{A}_{G_1}$ and by using the hybrid score heuristic, I achieve the highest score (average MCC of 0.18). This implies that cancer-related genes are best predicted when they are simultaneously of high importance in the control (healthy) networks (i.e. have a high centrality) and have a large shift in functional relations between case and control (i.e. have a high moving distance). Additionally, I observe that by considering the higher-order topology of pathways, as captured by $\widetilde{A}_{G_1}$, to take advantage of cancer drivers performing hub-roles between pathways, I manage to increase the performance compared to regular adjacency by 40% (average MCC with hybrid heuristic of 0.12). Lastly, I observe that the hybrid score outperforms graphlet eigencentrality and diffusion (average MCC of 0.09 and 0.10). Given that hybrid scores greatly outperforms the baseline methods, I conclude that PNMTF allows to predict cancer-related genes with high accuracy, whilst indicating the pathways involved. In the next section, I investigate these results more in detail and perform literature validation.

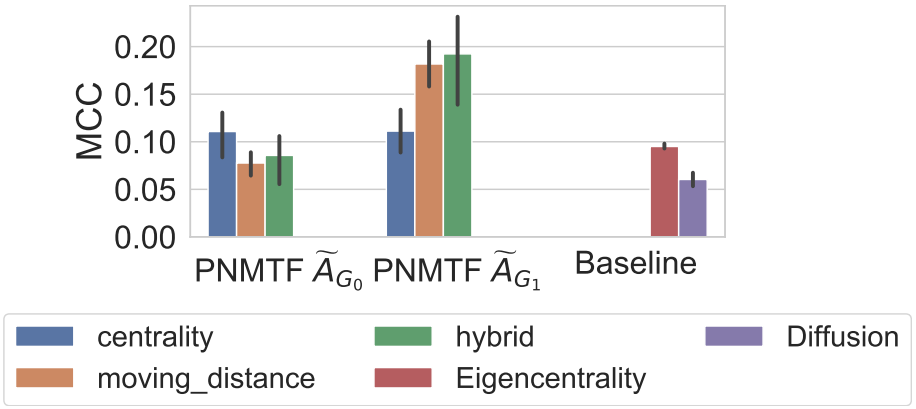


Figure 5.7: PNMTF identifies genes implicated in cancer. From left to right, I present the MCC scores for predicting cancer-related genes using PNMTF based on graphlet adjacencies $\widetilde{A}_{G_1}$ and $\widetilde{A}_{G_1}$ (x-axis) using different NMTF-scores (colour-coded, see legend), and compare against the state-of-the-art (far right)

5.2.4 Case study: identifying most rewired genes in changing pathway-pathway interactions in lung cancer

I showed that by applying PNMTF-scores consecutively at the pathway and gene level, I can predict cancer implicated pathways and cancer implicated genes within those pathways (see Sections 5.2.2 and 5.2.3). In other words, PNMTF allows to predict cancer implicated genes, whilst predicting for each gene the main pathway involved. Next, I validate in the literature the top 15 predicted genes in section 5.2.3 based on the hybrid PNMTF-score with graphlet adjacency A_{G_1}. To validate a predicted gene-cancer association in the literature, I first consider a gene to have a known role in cancer if it is listed in

the COSMIC cancer driver database or if any wet-lab experiment demonstrates it has one of the following properties in the cancer:

1. enhances/diminishes susceptibility to anti-cancer agents,
2. promotes cell proliferation and cell survivability,
3. promotes migration and invasion,
4. inhibits tumour genesis (i.e. has a suppressor role).

Note that differential expression is not accepted as evidence. For genes that I can not validate, I consider whether they are prognostic in the given cancer per the Pathology Atlas (Uhlen *et al.*, 2017) and if they are implicated in other cancer(s) based on wet-lab experiments or the COSMIC cancer driver database.

Here, I discuss the results for lung adenocarcinoma (see Table 5.9). For the other cancers, see Suppl. Tables B.1-B.3.

Rank	Symbol	Pathway	Validation in Lung Adenocarcinoma	Prognostic	Validation in non-lung cancer	Drugability	Immune System
1	GRB2	SHC-mediated cascade: FGFR2	PMID: 26693065 (2, 3, mouse model)			Approved	×
2	CSK	Phosphorylation of CD3 and TCR zeta chains			PMID: 11054667 (colon cancer, 4, in vitro)	Trial	✓
3	PTPN11	PD-1 signaling	PMID: 25730908 (2,3, mouse model)			×	✓
4	FYN	Dectin-2 family	PMID: 21371426 (3, in vitro)			Trial	✓
5	HSP90AA1	Attenuation phase			COSMIC (non-Hodgkins lymphoma)	Trial	×
6	PIK3R1	RHOF GTPase cycle	PMID: 24550137 (1, mouse model)			Trial	×
7	EGFR	Transcriptional reg. by the AP-2 family of TF	PMID: 20979469 (1, patient data)			Approved	×
8	SRC	FCGR activation	PMID: 17200208 (1, in vitro)			Approved	✓
9	GNB1	G beta:gamma signalling through PI3Kgamma				×	×
10	MYC	Transcriptional reg. by the AP-2 family of TF	PMID: 19551151 (2,3, mouse model)			Trial	×
11	CUL1	Prolactin receptor signaling	PMID: 33478195 (4, in vitro)			×	✓
12	XPO1	Extra-nuclear estrogen signaling	PMID: 27680702 (1, in vitro)			Trial	×
13	HNRNPH1	Signaling by FGFR2			PMID: 29362363 (Rhabdomyosarcoma, 2, mouse model)	Trial	×
14	GNG2	G-protein beta:gamma signalling			PMID: 24660107 (melanoma, 4, in vitro)	×	×
15	LYN	Dectin-2 family	PMID: 23866081 (1, in vitro)			Approved	✓

Table 5.9: Validation of top-scoring genes in lung cancer. The table ranks the top 15 genes and corresponding pathways according to their hybrid score in lung cancer using PNMTF based on graphlet adjacency A_{G_1}. Genes in black have literature support for their role in Lung Adenocarcinoma (PubMed IDs or references to COSMIC are given in the column 'Validation in Lung Adenocarcinoma'). Genes without literature validation, i.e. predictions, are highlighted in blue. For predicted genes, additional literary evidence highlights their potential role in cancer. Firstly, I provide the p-values for genes that statistically significantly impact patient survival in Lung Adenocarcinoma based on Kaplan Meier survival curves in the column 'Prognostic'. Secondly, when available, I provide PubMed IDs or references to COSMIC for 'predicted' genes that have a proven role in cancers other than Lung Adenocarcinoma in the column 'Validation non-lung cancer'. In the column 'Druggability', I indicate which genes are a known drug target for at least one approved drug or one drug in clinical trial according to DrugBank (Wishart *et al.*, 2018). In the column 'Immune System', I indicate which pathways are immune system pathways according to Reactome.

First, I validate that the hybrid score prioritises genes with hub-roles between pathways. I apply a Man-Whitney U (MWU) test to confirm that the prioritised genes, i.e. the top 15 genes based on the hybrid score (see Table 5.9, column 2), participate more frequently in the set of prioritised pathways (see Table 5.9, column 3) compared to the remaining, non-prioritised genes in those pathways. The MWU test yields a significant result (the prioritised genes participate on average in 2.0 of the prioritised pathways compared to 1.4 for the remaining, non-prioritised genes, p-value $\approx$ 2.22e-04). Therefore, as the prioritised genes occur more frequently in the prioritised pathways, they are the genes connecting those pathways, validating the approach. Moreover, this result is not observed when applying PNMTF on regular adjacency, highlighting that graphlet-adjacency G_1 enables the capturing of hub-roles of potential cancer drivers. Further, I validate in the literature that 11/15 (73%) of the prioritised genes are implicated in lung cancer (see Suppl. Table 5.9, column 4). For the four unvalidated genes: CSK, HSP90AA1, HNRNPH1 and GNG2, I find strong web-lab evidence in the literature that they are involved in other cancers (see Table 5.9, column 6). I find that HSP90AA1 is a known cancer driver in non-Hodgkins lymphoma (COSMIC), HNRNPH1 supports cancer-cell proliferation in rhabdomyosarcoma (Li *et al.*, 2018) and CKS and GNG2 have tumour suppressing roles in colon cancer and melanoma, respectively (Nakagawa *et al.*, 2000; Yajima *et al.*, 2014). Moreover, I find that 11/15 (73%) of the prioritised genes are known drug targets, including CSK, HSP90AA1 and HNRNPH1 (see Suppl. Table 5.9 , column seven).

Next, I focus on the associated prioritised pathways, which serve as a functional contexts to the gene predictions (see Table 5.9, column three). From the network perspective, I observe that the union of the prioritised pathways induces on both the case and control PPI network a connected sub-network that is denser than expected by chance (both p-values $\approx$ 1.00e-4, based on bootstrapping, obtained by sampling 10,000 sets of pathways that are of size within the range of those in the prioritised pathway list). This indicates that the prioritised pathways are likely functionally related (as they are overlapping in the PPI network) and that the method is capturing an underlying disease-related signal (as the pathways are more intertwined than expected by chance). Further, I validate that the hybrid score prioritises pathways based their altered pathway-pathway interactions rather than their internal perturbation. To do so, I assess if the prioritised pathways have significantly more edges rewired (i.e., added or deleted) that connect them to the other prioritised pathways compared to the number of edges they have rewired within them, by applying a hypergeometric test. I find that edges between pathways are twelve times more

rewired (p-value $\approx$ 3.40e-19), validating the method.

From a functional perspective, I find that the prioritised pathways are enriched in Reactome 'Immune System' pathways (pathways ranked 2, 3, 4, 8 and 15 in Suppl. Table 5.9, p-value $\approx$ 2.31e-2). Furthermore, the remaining pathways can easily be related to the immune system. For instance, pathways 9 and 14 are downstream of GCPR signalling, which regulates T-cell immunity (Wang, 2018). Pathway 12, 'Receptors for oestrogens signalling', regulates immune system pathways, as well as immune cell development (Kovats, 2015). These results are in line with the cancer literature, as immune system rewiring is necessary for cancer cells to evade immunological response and to enable them to abuse inflammatory responses as a source for bioactive molecules (e.g., growth factors) (Hanahan and Weinberg, 2011). Combined with the results at the gene level, I conclude that PNMTF uncovers a cancer induced rewiring of the proteins linking immune system pathways in lung cancer.

I obtain similar results across all four cancers, see Suppl. Section B.6. When considering the top 15 predicted genes for the four cancers collectively, I can validate 47/60 (78%) of these gene-cancer associations in the literature. I show that the genes involved in the 13 unvalidated gene-cancer associations are implicated in other cancers. As the top 15 predicted genes across the four cancers overlap, which is expected as cancers can share the same disease mechanisms, I predict 28 unique genes in total. Of these genes, 15/28 (54%) are known drug targets (see Suppl. Tables 5.9-B.3, column 7). As six of the 13 unvalidated gene-cancer associations involve druggable genes, I suggest them as cancer-specific drug targets: CSK, HSP90AA1 and HNRNPH1 for lung cancer, HSP90AA1 for colon cancer and prostate cancer, and HNRNPH1 for ovarian cancer. At the pathway level, I find that the hybrid score uncovers a cancer-induced rewiring of the proteins connecting pathways involved in the immune system in colon and prostate cancer.

5.3 Conclusion

In this chapter, I suggest PNMTF, which learns an embedding space that captures the functional organisation of pathways in a cell. In this embedding space, I define two heuristics, NMTF centrality and moving distance, which measure the importance and disruption of functional relationships of a pathway or gene in cancer, respectively. I apply these heuristics to predict cancer implicated pathways and genes in four cancers. Additionally, I exploit cancer genes tending to perform hub roles between pathway interactions by considering graphlet-based higher-order topologies that encode hub roles. I find that PNMTF

uncovers a cancer-induced rewiring of the genes linking pathways involved in the immune system for three out of the four cancers. This is in line with the literature, where the immune system's rewiring is considered a hallmark of cancer. Finally, I provide literary evidence indicating the top predicted genes are likely involved in cancer and find many are known drug targets, allowing me to predict six druggable cancer-specific drug targets.

Further, this analysis opens up multiple research questions:

1. To uncover emerging (disappearing) functional relationships in cancer and thus provide insight into cancer development, it could be interesting to see what pathways become (less) central and form new (no longer form) dense clusters in cancer.

2. To enable a de novo pathway analysis I propose to extend PNMTF as a subspace-clustering algorithm, as explained next.

3. As Reactome pathway annotations only cover 37% of the genes in the PPI network, PNMTF ignores 63% of the human genes when predicting cancer implicated genes. Therefore, to extend the coverage of the gene predictions, I suggest to extend the PNMTF model as a subspace clustering model, so that rather than taking the assignment of genes to pathways as a prior input, PNMTF model would learn the assignment of genes to subspaces, which I would interpret as *de novo* pathways. I discuss this suggestion in detail in the closing notes of the thesis, see Section 6.2.3.

4. To provide insight in how different drugs affect pathways in cancer, gene-drug data could be integrated in the PNMTF model.

5. Lastly, PNMTF can be applied outside of biology, when the input data is a network and domain knowledge categorising the nodes. For instance, PNMTF could be applied on trade networks, where nodes are countries and edges are the value of the trade between them, while trade agreements form a prior grouping of the nodes.

Chapter 6

Conclusions

In this chapter, I summarise the results and contributions presented in this thesis, both in terms of methodology and biological aspects. Further, having briefly touched on future research directions at the end of chapters 3 and 4, here I provide more in-depth detail on some of the key suggested methodological improvements.

6.1 Summary of thesis achievements

Systems biology is flooded with large scale "omics" data, typically modelled as networks. This abundance of network data started the fields of network biology, allowing the elucidation of molecular mechanisms of a broad range of diseases, such as rare Mendelian disorders (Smedley *et al.*, 2014), cancer (Leiserson *et al.*, 2015), and metabolic diseases (Baumgartner *et al.*, 2018). In this thesis, I focus on studying the higher-order organisation of molecular networks at the node-level (genes) and at the sub-graph level (pathways).

In chapter 3, to study the higher-order organisation of networks at the node-level, I introduce graphlet adjacency, which considers a pair of nodes to be adjacent if they simultaneously touch a given graphlet. I demonstrate that graphlet adjacency captures topology-function and topology-disease relationships in molecular networks. In particular through graphlet-generalized spectral clustering of model networks and biological networks, I show that different Graphlet Laplacians capture different local topologies. By applying graphlet-generalized spectral embedding, I demonstrate that Graphlet Laplacians capture biological functions as well. I quantify this through graphlet-generalized spectral clustering analysis. I show that Graphlet Laplacians are not only as biologically relevant as alternative Laplacian matrices, but also capture complementary biological functions. Finally, by graphlet-generalized diffusing of pan-cancer gene mutation scores on the human PPI network, I show that Graphlet Laplacians capture complementary disease mechanisms.

In chapter 4, to further investigate the relationships between the topological features of genes in molecular networks in human and yeast, as captured by graphlet adjacencies, and the biological functions of the genes, I build more descriptive pathway-based approaches. Specifically, I extend eigencentrality to graphlet eigencentrality, to study the importance (centrality) of genes in pathways, either from the local pathway perspective or the global perspective of the entire network.

First, I identify the pathways that are described by each graphlet adjacency, i.e. all genes participating in a pathway are also captured as topologically important by the graphlet adjacency, using graphlet eigencentralities to predict which genes belong to a given pathway. I consider the pathways for the highest prediction accuracies are achieved as being described by that graphlet adjacency. I show that the pathways that are described by a given graphlet adjacency, are functionally similar, implying that each graphlet adjacency uncovers different pathway topology and function relationships. I illustrate this relationship by means of a case study in the 'Receptor Mediated Mitophagy' pathway, where I show how, unlike regular adjacency, graphlet adjacencies capture the relevance of all genes in the pathway.

Second, considering different graphlet adjacencies, from the local and global perspective, uncovers complementary sets of cancer driver genes (known to be drivers in at least one type of cancer) that are described by playing central roles in pathways and the crosstalk between them. This suggests that by considering different graphlet eigencentralities, one can capture different functional roles of genes in and between pathways. I illustrate this relationship by means of a case study, this time in the 'Formation of Senescence-Associated Heterochromatin Foci'-pathway, where I show how, unlike regular adjacency, graphlet adjacencies capture the central roles of cancer driver genes TP53 and RB1. This result shows that, as centrality measures are widely used to uncover disease-related genes and graphlet eigencentrality captures notions of centrality different from those based on regular adjacency, graphlet eigencentrality opens up the opportunity of uncovering novel disease related genes.

In chapter 5, I study the higher-order organisation of functional sub-graphs in a network. In particular, I prioritise cancer-implicated pathways whilst simultaneously providing insight into the key genes involved, in four cancer types: lung cancer and colorectal cancer, respectively the deadliest cancer types, and prostate cancer and ovarian cancer, the most prevailing gender-specific cancer types (Sung *et al.*, 2021).

To do so, I introduce pathway-driven non-negative matrix factorization (PNMTF),

which learns the functional organisation of a healthy cell by decomposing 'healthy' curated pathways, encoded as induced subgraphs of a control PPI network. In this space, I define pathway and gene embeddings, based on the pathways in a healthy and diseased state (represented by the subgraphs induced by curated pathways on control and case PPI network). Based on these embeddings, I define 'NMTF centrality', which measures the functional importance of a pathway or gene as the norm of its healthy embedding and the 'moving distance', which measures the disruption of a pathway's or gene's functional relationships, as the distance between its healthy and cancerous embedding.

I validate that PNMTF captures the functional organisation of pathways in the cell: embedding all pathways in the common space I find that their embeddings form easily separable clusters that are functionally coherent. Then, I show that pathways or genes with high centralities and moving distances are likely to be cancer-related; effectively identifying cancer-related pathways and genes not based on their (internal) perturbation in cancer, but based on their functional relationships with other pathways and genes in the cell. Additionally, I show that higher-order topologies based on graphlets that allow for capturing different hubness properties, such as occurrence on shortest-paths, improve the prediction accuracy, indicating that they improve the capturing of cancer mechanisms. Finally, I consider the top 50 predicted cancer genes for each cancer type and validate 59/65 of the uniquely predicted genes through literature curation. I present a case study for lung cancer, where I confirm through literature investigation that the only non-validated gene, SLC47A1, and the corresponding pathway on which its score is based, 'SLC-mediated transmembrane transport' (not a known cancer pathway), are potentially cancer-related.

6.2 Methodological extensions

6.2.1 Alternative notions of higher-order topology

Graphlet-adjacency quantifies the association between two nodes based on how frequently they form a given graphlet. However, alternative notions that capture higher-order information could be considered. Currently, our group is generalising the clustering coefficient based on graphlets, defining the graphlet clustering coefficient. Building on this work, I define the graphlet clustering matrix below.

For a given node, the clustering coefficient measures the extent to which its neighbours cluster together, i.e. a tend to form triangles. Formally, the clustering coefficient of a node is defined as the ratio of the number of edges between its neighbours over the pos-

sible number of edges between them (see Section 2.1.3). Alternatively, from a graphlet perspective, the clustering coefficient can be defined as as the number of times a node touches orbit 3 (i.e. forms a triangle), over the number of times it touches orbit 2 (i.e. is at the centre of a wedge) (see Section 2.1.4 for an illustration on orbits).

To extend regular adjacency, graphlet adjacency defines two nodes u and v of a network G to be graphlet-adjacent based on how frequently they simultaneously touch a given graphlet G_k. Analogously, I define *clustering adjacency*, which considers a node v to be cluster-adjacent to a node u, based on how much v contributes to the clustering coefficient of u, i.e. how many times u and v form triangles together (i.e., count for orbit 3) relative to the total number of wedges u participates in (i.e., orbit count 2). Note that orbit count o_3 is equivalent to graphlet count G_2, as graphlet G_2 only has a single orbit.

Given this extended definition of adjacency, I define the *graphlet clustering matrix* as:

$$A_C(u,v) = \begin{cases} o^3_{uv}/o^2_u & \text{if } u \neq v \\ 0 & \text{otherwise,} \end{cases} \tag{6.1}$$

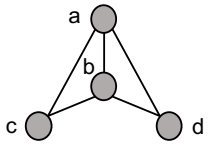
where o^3_{uv} is the number of times nodes u and v simultaneously touch orbit 3 and o^2_u the number of times node u touches orbit 2. I illustrate the clustering adjacency matrix for a toy network and compare it to graphlet adjacency for graphlet G_2 in Figure 6.1. Note that A_C, unlike A_{G_2}, is not a symmetric matrix. Additionally, the entries of A_C automatically scale between 0 and 1, absolving the need to normalise the adjacency matrix. It should be clear that by considering different pairs of orbits, alternative graphlet clustering matrices can be defined, allowing to capture different notions of clustering connectivity patterns.

6.2.2 Faster graphlet counting

The current graphlet adjacency counter is based on the relatively inefficient graphlet counter of GraphCrunch 2, which counts graphlets through brute force enumeration (Kuchaiev *et al.*, 2011). When considering four node graphlets, for each of the n nodes in the network I perform a breadth-first search three nodes deep, counting the formed graphlets along the way. So, if d is the maximal degree in the network, the time complexity of the counting algorithm is $\mathcal{O}(nd^3)$. In practical terms, the counter takes one day to compute for the human PPI network and five days for the human COEX network. So, the current counter is impractical for larger or denser networks than those considered in this thesis.

To speed up computations, I suggest taking inspiration from the Orbit Counting Al-

(A)



(B)

$$A_{G_2} = \begin{array}{c} \begin{matrix} a & b & c & d \end{matrix} \\ \begin{bmatrix} 0 & 2 & 1 & 1 \\ 2 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix} \end{array} \begin{matrix} a \\ b \\ c \\ d \end{matrix} \qquad A_C = \begin{array}{c} \begin{matrix} a & b & c & d \end{matrix} \\ \begin{bmatrix} 0 & 2/3 & 1/3 & 1/3 \\ 2/3 & 0 & 1/3 & 1/3 \\ 1/2 & 1/2 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \end{bmatrix} \end{array} \begin{matrix} a \\ b \\ c \\ d \end{matrix}$$

Figure 6.1: An illustration comparing graphlet adjacency A_{G_2} (triangle) to clustering adjacency. **A:** Example network H. **B:** The adjacency matrices A_{G_2} and A_C for the example network H, shown in panel (A). In the case of A_{G_2}, the off-diagonal elements correspond to the number of times two nodes touch graphlet G_2 (triangle) together. For A_C, the off-diagonal elements corresponds to the ratio of same counts, divided by the number time the node corresponding to the row touches orbit o_2 (i.e., is at the center of a wedge). For instance, $A_{G_2}(a,b) = 2$, as a and b form G_2 twice, via triangles : *a*-*b*-*c* and *a*-*b*-*d*. $A_C(a,b) = 2/3$, as a touches o_2 three times: at the centre of wedges *c*-*a*-*d*, *c*-*a*-*b* and *b*-*a*-*d*.

gorithm (Orca) (Hočevar and Demšar, 2014). To count four-node graphlets, the authors of Orca provide ten *redundancy equations*, linear relationships between the eleven different orbit counts (see Section 2.1.4 for an illustration on orbits). Therefore, the counts for only one type of orbit need to be computed to make this a determined system, such that all other orbit counts can be inferred. Then, the authors suggest a novel heuristic for counting four-node cliques, i.e. graphlet G_8, that takes advantage of these rarely occurring in biological networks, as they are typically sparse. Although the theoretical time complexity of Orca is still $\mathcal{O}(nd^3)$, in practice, it is a lot faster than GraphCrunch 2 (Hočevar and Demšar, 2014). In practice, the Orca counter takes less than ten minutes to compute the GDVs for the human PPI or COEX network. So, I suggest extending the Orca redundancy equations to be applicable for the graphlet adjacency counter to enable graphlet-adjacency based higher-order network analysis for larger networks

6.2.3 Uncovering de novo cancer pathways with NMTF based subspace clustering

In chapter 5, I propose PNMTF to identify cancer implicated pathways and the genes involved. Pathway based approaches, such as PNMTF, rely on a set of predefined curated pathways that have a well defined function within the cell. This has the benefit of leading to highly interpretable results which in turns supports to the creation of testable hypotheses. However, curated pathway annotations are very incomplete, meaning that pathway-based approaches cover only a subset of the human proteins. For instance, I

find that the curated pathway database Reactome only covers 37% of all human proteins with known PPI (Jassal *et al.*, 2019; Oughtred *et al.*, 2019). To counter this issue, *de novo* pathway based methods aim at identifying significantly perturbed subnetworks in disease, avoiding the need for a set of pre-defined pathways. Note that, unlike biological pathways, de novo pathways lack an a priori known function. For instance, given a large biological interaction network, KeyPathwayMiner extracts connected subnetworks enriched in differentially expressed genes in cancer and interprets them as functional modules or *de novo* pathways (Alcaraz *et al.*, 2016). Similarly, Hierarchical HotNet diffuses somatic mutation scores that quantify for each gene how likely it is to be somatically mutated in cancer on a protein-protein interaction (PPI) network. After diffusion, neighbourhoods enriched in frequently mutated genes are predicted as being cancer-related modules or 'pathways' (Reyna *et al.*, 2018). To enable a de novo pathway analysis I propose to extend PNMTF as a subspace-clustering algorithm, as explained next.

In classical clustering, each data points is assigned to a given cluster based on all of its features. Sub-space algorithms on the other hand learn for each data point multiple multivariate latent feature representations, which can be based on all or a subset of all features in the data. Data points are then segmented based on those learned features, meaning that for a given data point, irrelevant features can be ignored. For instance, classical k-means iteratively learns cluster centroids (as points in a global space) and assigns each individual data point to the nearest centroid. Note that in K-means, the centroid is computed as the average of all data points assigned to it, so all features are taken into account. It's subspace generalisation, K-subspaces, iteratively learns cluster subspaces (hyperplanes in a global space) and assigns points to the nearest subspace (Vidal, 2011). As the name entails, each individual subspace does not need to cover the entire global space, i.e. the learned hyperplanes can be based on only a subset of the features in the data. A synthetic example is given in Figure 6.2, where data points visibly lie on different hyperplanes. A classical clustering algorithm, such as k-means, would have issues correctly segmenting data points that lie near the intersection of one or more clusters. Subspace clustering algorithms would correctly segment those data points, as they learn the features that are relevant to segmenting them: the axes of the hyperplanes.

PNMTF can be extended as a sub-space clustering problem by decomposing an adjacency matrix $A^{n\times n}$ into m pairs of non-negative latent matrices $U_p^{n\times d_p}$ and $S_p^{d_p\times d}$ and one orthonormal non-negative latent matrix $V^{n\times d}$: $A \approx \sum_{p=0}^{m} U_p S_p V^T$. This corresponds to solving the following objective function

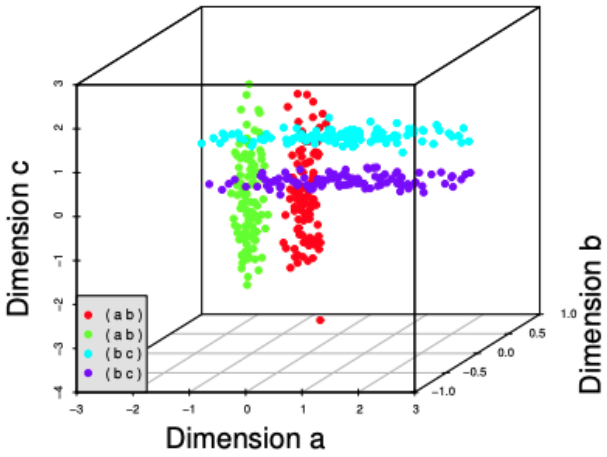


Figure 6.2: Illustration of the merits of subspace clustering. Show is a sample dataset with four clusters, each in two dimensions with the third dimension being noise. Points near the intersection of two clusters would confuse many traditional clustering algorithms. Additionally, only subspace clustering algorithms would (correctly) ignore the noise in the third dimension. This image is taken from (Parsons *et al.*, 2004)

$$\min_{U_p,S_p,V\geq 0,\forall p\in[0,m[} \left\| A - \sum_{p=0}^{m} U_p S_p V^T \right\|_F^2, \text{ s.t.: } V^T V = I. \tag{6.2}$$

Here, each matrix $U_p^{n\times d_p}$ is a latent subspace, that can be projected in the common space V by multiplying it with S_p. The time complexity of solving this optimisation problem is $\mathcal{O}(mn^2 d_p)$. Note that each matrix U_p covers all nodes in the network, so a post-processing step is needed to decided if a gene is assigned to one or more subspaces. Having assigned all the genes to one or more sub-spaces, can interpret each sub-space U_p as a latent de novo pathway and use the NMTF centrality and moving distance to predict cancer implicated de novo pathways and identify genes involved.

Bibliography

Alcaraz, N., List, M., Dissing-Hansen, M., Rehmsmeier, M., Tan, Q., Mollenhauer, J., Ditzel, H. J., and Baumbach, J. (2016). Robust de novo pathway enrichment with KeyPathwayMiner 5. *F1000Research*, **5**.

Asensio, N. C., Giner, E. M., De Groot, N. S., and Burgas, M. T. (2017). Centrality in the host–pathogen interactome is associated with pathogen fitness during infection. *Nature communications*, **8**(1), 1–6.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, **25**(1), 25–29.

Barabási, A. L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, **286**(5439), 509–512.

Baumgartner, C., Spath-Blass, V., Niederkofler, V., Bergmoser, K., Langthaler, S., Lassnig, A., Rienmüller, T., Baumgartner, D., Asnani, A., and Gerszten, R. E. (2018). A novel network-based approach for discovering dynamic metabolic biomarkers in cardiovascular disease. *PloS one*, **13**(12), e0208953.

Belkin, M. and Niyogi, P. (2003). Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Computation*, **15**(6), 1373–1396.

Ben-Gal, I. (2008). Bayesian networks. *Encyclopedia of statistics in quality and reliability*.

Borgwardt, K. M. (2011). Kernel methods in bioinformatics. In *Handbook of statistical bioinformatics*, pages 317–334. Springer.

Braicu, C., Buse, M., Busuioc, C., Drula, R., Gulei, D., Raduly, L., Rusu, A., Irimie, A., Atanasov, A. G., Slaby, O., and Others (2019). A comprehensive review on MAPK: a promising therapeutic target in cancer. *Cancers*, **11**(10), 1618.

Broido, A. D. and Clauset, A. (2019). Scale-free networks are rare. *Nature Communications*, **10**(1), 1–10.

Castresana-Aguirre, M. and Sonnhammer, E. L. L. (2020). Pathway-specific model estimation for improved pathway annotation by network crosstalk. *Scientific Reports*, **10**(1), 1–12.

Chen, Y., Hao, J., Jiang, W., He, T., Zhang, X., Jiang, T., and Jiang, R. (2013). Identifying potential cancer driver genes by genomic data integration. *Scientific Reports*, **3**, 3538.

Chung, F. R. K. and Graham, F. C. (1997). *Spectral graph theory*. Number 92. American Mathematical Soc.

Cichocki, A., Zdunek, R., Phan, A. H., and Amari, S.-i. (2009). *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons.

Cohen, R., Erez, K., Ben-Avraham, D., and Havlin, S. (2000). Resilience of the internet to random breakdowns. *Physical review letters*, **85**(21), 4626.

Cooper, G. F. (1990). The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial intelligence*, **42**(2-3), 393–405.

Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E. D., Sevier, C. S., Ding, H., Koh, J. L., Toufighi, K., Mostafavi, S., Prinz, J., St. Onge, R. P., VanderSluis, B., Makhnevych, T., Vizeacoumar, F. J., Alizadeh, S., Bahr, S., Brost, R. L., Chen, Y., Cokol, M., Deshpande, R., Li, Z., Lin, Z.-Y., Liang, W., Marback, M., Paw, J., San Luis, B.-J., Shuteriqi, E., Tong, A. H. Y., van Dyk, N., Wallace, I. M., Whitney, J. A., Weirauch, M. T., Zhong, G., Zhu, H., Houry, W. A., Brudno, M., Ragibizadeh, S., Papp, B., Pál, C., Roth, F. P., Giaever, G., Nislow, C., Troyanskaya, O. G., Bussey, H., Bader, G. D., Gingras, A.-C., Morris, Q. D., Kim, P. M., Kaiser, C. A., Myers, C. L., Andrews, B. J., and Boone, C. (2010). The Genetic Landscape of a Cell. *Science*, **327**(5964).

Costanzo, M., VanderSluis, B., Koch, E. N., Baryshnikova, A., Pons, C., Tan, G., Wang, W., Usaj, M., Hanchard, J., Lee, S. D., Pelechano, V., Styles, E. B., Billmann, M.,

Van Leeuwen, J., Van Dyk, N., Lin, Z. Y., Kuzmin, E., Nelson, J., Piotrowski, J. S., Srikumar, T., Bahr, S., Chen, Y., Deshpande, R., Kurat, C. F., Li, S. C., Li, Z., Usaj, M. M., Okada, H., Pascoe, N., Luis, B. J. S., Sharifpoor, S., Shuteriqi, E., Simpkins, S. W., Snider, J., Suresh, H. G., Tan, Y., Zhu, H., Malod-Dognin, N., Janjic, V., Przulj, N., Troyanskaya, O. G., Stagljar, I., Xia, T., Ohya, Y., Gingras, A. C., Raught, B., Boutros, M., Steinmetz, L. M., Moore, C. L., Rosebrock, A. P., Caudy, A. A., Myers, C. L., Andrews, B., and Boone, C. (2016). A global genetic interaction network maps a wiring diagram of cellular function. *Science*, **353**(6306), aaf1420.

Cowen, L., Ideker, T., Raphael, B. J., and Sharan, R. (2017). Network propagation: a universal amplifier of genetic associations. *Nature Reviews Genetics*, **18**(9), 551.

Creixell, P., Reimand, J., Haider, S., Wu, G., Shibata, T., Vazquez, M., Mustonen, V., Gonzalez-Perez, A., Pearson, J., Sander, C., and Others (2015). Pathway and network analysis of cancer genomes. *Nature Methods*, **12**(7), 615.

Daemen, A., Gevaert, O., and De Moor, B. (2007). Integration of clinical and microarray data with kernel methods. In *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE*, pages 5411–5415. IEEE.

Davaadelger, B., Duan, L., Perez, R. E., Gitelis, S., and Maki, C. G. (2016). Crosstalk between the IGF-1R/AKT/mTORC1 pathway and the tumor suppressors p53 and p27 determines cisplatin sensitivity and limits the effectiveness of an IGF-1R pathway inhibitor. *Oncotarget*, **7**(19), 27511.

Davis, C. and Kahan, W. M. (1970). The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, **7**(1), 1–46.

Davis, D. A. and Chawla, N. V. (2011). Exploring and exploiting disease interactions from multi-relational gene and phenotype networks. *PloS one*, **6**(7), e22670.

DeBerardinis, R. J. and Chandel, N. S. (2016). Fundamentals of cancer metabolism. *Science Advances*, **2**(5), e1600200.

Ding, C., He, X., and Simon, H. (2005). On the Equivalence of Nonnegative Matrix Factorization and Spectral Clustering. *Proceedings of the 2005 SIAM ICDM*, (4), 126–135.

Ding, C., Li, T., Peng, W., and Park, H. (2006). Orthogonal nonnegative matrix tri-factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 126–135. ACM.

Dutkowski, J., Kramer, M., Surma, M. A., Balakrishnan, R., Cherry, J. M., Krogan, N. J., and Ideker, T. (2013). A gene ontology inferred from molecular networks. *Nature Biotechnology*, **31**(1), 38–45.

Ehrlinger, L. and Wöß, W. (2016). Towards a definition of knowledge graphs. *SEMANTiCS (Posters, Demos, SuCCESS)*, **48**(1-4), 2.

Erdős Paul and Rényi Alfréd, S. (1959). On random graphs. *Publicationes Mathematicae*, **6**, 290–297.

Estrada, E. (2012). Path Laplacian matrices: Introduction and application to the analysis of consensus in networks. *Linear Algebra and Its Applications*, **436**(9), 3373–3391.

Franz, M., Rodriguez, H., Lopes, C., Zuberi, K., Montojo, J., Bader, G. D., and Morris, Q. (2018). GeneMANIA update 2018. *Nucleic Acid Research*, **46**(W1), W60—-W64.

Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M. R. (2004). A census of human cancer genes. *Nature Reviews Cancer*, **4**(3), 177.

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

Gene Ontology Consortium, T. (2017). Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Research*, **45**, D331–D338.

Gevaert, O., Smet, F. D., Timmerman, D., Moreau, Y., and Moor, B. D. (2006). Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. *Bioinformatics*, **22**(14), e184—-e190.

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

Gligorijević, V., Malod-Dognin, N., and Pržulj, N. (2016a). Integrative methods for analyzing big data in precision medicine. *Proteomics*, **16**(5), 741–758.

Gligorijević, V., Malod-Dognin, N., and Pržulj, N. (2016b). Patient-Specific Data Fusion for Cancer Stratification and Personalised Treatment. *Pacific Symposium on Biocomputing.*, **21**, 321–32.

Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Tamborero, D., Schroeder, M. P., Jene-Sanz, A., Santos, A., and Lopez-Bigas, N. (2013). IntOGen-mutations identifies cancer drivers across tumor types. *Nature Methods*, **10**(11), 1081.

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

Guo, X., Gao, L., Wei, C., Yang, X., Zhao, Y., and Dong, A. (2011). A computational method based on the integration of heterogeneous networks for predicting disease-gene associations. *PloS one*, **6**(9), e24171.

Guo, Z., Liu, X., Hou, H., Wei, F., Liu, J., and Chen, X. (2016). Abnormal degree centrality in Alzheimer's disease patients with depression: A resting-state functional magnetic resonance imaging study. *Experimental Gerontology*, **79**, 61–66.

Hagen, L. and Kahng, A. B. (1992). New spectral methods for ratio cut partitioning and clustering. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, **11**(9), 1074–1085.

Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell*, **144**(5), 646–74.

Hardoon, D. R., Szedmak, S., and Shawe-Taylor, J. (2004). Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, **16**(12), 2639–2664.

Hočevar, T. and Demšar, J. (2014). A combinatorial approach to graphlet counting. *Bioinformatics*, **30**(4), 559–565.

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

Iakoucheva, L. M., Brown, C. J., Lawson, J. D., Obradović, Z., and Dunker, A. K. (2002). Intrinsic disorder in cell-signaling and cancer-associated proteins. *Journal of Molecular Biology*, **323**(3), 573–584.

Isik, Z., Baldow, C., Cannistraci, C. V., and Schroeder, M. (2015). Drug target prioritization by perturbed gene expression and network information. *Scientific Reports*, **5**, 17417.

Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., Sidiropoulos, K., Cook, J., Gillespie, M., Haw, R., and Others (2019). The reactome pathway knowledgebase. *Nucleic Acids Research*, **48**(1), 498–503.

Jeong, H., Mason, S. P., Barabási, A.-L., and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, **411**(6833), 41–42.

Jolliffe, I. T. (1986). Principal Component Analysis and Factor Analysis. In *Principal component analysis*, pages 115–128. Springer.

Kandoth, C., McLellan, M. D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J. F., Wyczalkowski, M. A., and Others (2013). Mutational landscape and significance across 12 major cancer types. *Nature*, **502**(7471), 333.

Kim, H. and Park, H. (2007). Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*, **23**(12), 1495–1502.

Kodinariya, T. M. and Makwana, P. R. (2013). Review on determining number of Cluster in K-Means Clustering. *International Journal of Advance Research in Computer Science and Management Studies*, **1**(6), 2321–7782.

Kondor, R. I. and Lafferty, J. (2002). Diffusion kernels on graphs and other discrete input spaces. In *ICML*, volume 2, pages 315–322.

Kotlyar, M., Pastrello, C., Malik, Z., and Jurisica, I. (2019). IID 2018 update: context-specific physical protein–protein interactions in human, model organisms and domesticated species. *Nucleic Acids Research*, **47**(D1), D581—-D589.

Kovats, S. (2015). Estrogen receptors regulate innate immune cells and signaling pathways. *Cellular immunology*, **294**(2), 63–69.

Kuchaiev, O., Stevanović, A., Hayes, W., and Pržulj, N. (2011). GraphCrunch 2: Software tool for network modeling, alignment and clustering. *BMC Bioinformatics*, **12**(1), 1–13.

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *International Conference on Machine L earning*, pages 1188–1196.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, **401**(6755), 788–791.

Leiserson, M. D. M., Vandin, F., Wu, H.-T., Dobson, J. R., Eldridge, J. V., Thomas, J. L., Papoutsaki, A., Kim, Y., Niu, B., McLellan, M., and Others (2015). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature Genetics*, **47**(2), 106.

Li, Q. and Milenković, T. (2019). Supervised prediction of aging-related genes from a context-specific protein interaction subnetwork. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 130–137. IEEE.

Li, Y., Bakke, J., Finkelstein, D., Zeng, H., Wu, J., and Chen, T. (2018). HNRNPH1 is required for rhabdomyosarcoma cell growth and survival. *Oncogenesis*, **7**(1), 1–13.

Liang, F., Ren, C., Wang, J., Wang, S., Yang, L., Han, X., Chen, Y., Tong, G., and Yang, G. (2019). The crosstalk between STAT3 and p53/RAS signaling controls cancer cell metastasis and cisplatin resistance via the Slug/MAPK/PI3K/AKT-mediated regulation of EMT and autophagy. *Oncogenesis*, **8**(10), 1–15.

Malod-Dognin, N., Windels, S. F., and Pržulj, N. (2019a). *Machine Learning for Data Integration in Cancer Precision Medicine: Matrix Factorization Approaches*, pages 286–312. Cambridge University Press.

Malod-Dognin, N., Petschnigg, J., Windels, S. F. L., Povh, J., Hemmingway, H., Ketteler, R., and Pržulj, N. (2019b). Towards a data-integrated cell. *Nature Communications*, **10**(1), 805.

Milenković, T. and Pržulj, N. (2008). Uncovering biological network function via graphlet degree signatures. *Cancer Informatics*, **6**, 257–273.

Milenković, T. and Pržulj, N. (2008). Uncovering biological network function via graphlet degree signatures. *Cancer informatics*, **6**, CIN–S680.

Mohar, B. (1997). Some applications of Laplace eigenvalues of graphs. In *Graph symmetry*, pages 225–275. Springer.

Morishita, A., Zaidi, M. R., Mitoro, A., Sankarasharma, D., Szabolcs, M., Okada, Y., D'Armiento, J., and Chada, K. (2013). HMGA2 is a driver of tumor metastasis. *Cancer Research*, **73**(14), 4289–4299.

Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C., and Morris, Q. (2008). GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome biology*, **9**(1), S4.

Muscoloni, A. and Cannistraci, C. V. (2018). A nonuniform popularity-similarity optimization (nPSO) model to efficiently generate realistic complex networks with communities. *New Journal of Physics*, **20**(5), 52002.

Nakagawa, T., Tanaka, S., Suzuki, H., Takayanagi, H., Miyazaki, T., Nakamura, K., and Tsuruo, T. (2000). Overexpression of the csk gene suppresses tumor metastasis in vivo. *International Journal of Cancer*, **88**(3), 384–391.

Napolitano, F., Zhao, Y., Moreira, V. M., Tagliaferri, R., Kere, J., D'Amato, M., and Greco, D. (2013). Drug repositioning: a machine-learning approach through data integration. *Journal of cheminformatics*, **5**(1), 30.

Narita, M., Nuñez, S., Heard, E., Narita, M., Lin, A. W., Hearn, S. A., Spector, D. L., Hannon, G. J., and Lowe, S. W. (2003). Rb-mediated heterochromatin formation and silencing of E2F target genes during cellular senescence. *Cell*, **113**(6), 703–716.

Needham, C. J., Bradford, J. R., Bulpitt, A. J., and Westhead, D. R. (2007). A primer on learning in Bayesian networks for computational biology. *PLoS computational biology*, **3**(8), e129.

Newman, M. E. J. M. E. J. (2010). *Networks: An Introduction*. Oxford University Press.

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

Obayashi, T., Kagaya, Y., Aoki, Y., Tadaka, S., and Kinoshita, K. (2019). COXPRESdb v7: a gene coexpression database for 11 animal species supported by 23 coexpression platforms for technical evaluation and evolutionary inference. *Nucleic Acids Research*, **47**(D1), D55—-D62.

Ogris, C., Guala, D., Helleday, T., and Sonnhammer, E. L. L. (2017). A novel method for crosstalk analysis of biological networks: improving accuracy of pathway annotation. *Nucleic Acids Research*, **45**(2), e8.

Okamura, Y., Aoki, Y., Obayashi, T., Tadaka, S., Ito, S., Narise, T., and Kinoshita, K. (2015). COXPRESdb in 2015: Coexpression database for animal species by DNA-microarray and RNAseq-based expression data with multiple quality assessment systems. *Nucleic Acids Research*, **43**(D1), D82–D86.

Oldham, S., Fulcher, B., Parkes, L., Arnatkevic iūtė, A., Suo, C., and Fornito, A. (2019). Consistency and differences between centrality measures across distinct classes of networks. *PloS One*, **14**(7), e0220061.

Oughtred, R., Stark, C., Breitkreutz, B.-J., Rust, J., Boucher, L., Chang, C., Kolas, N., O'Donnell, L., Leung, G., McAdam, R., and Others (2019). The BioGRID interaction database: 2019 update. *Nucleic Acid Research*, **47**(D1), D529—-D541.

Papadopoulos, F., Kitsak, M., Serrano, M. Á., Boguná, M., and Krioukov, D. (2012). Popularity versus similarity in growing networks. *Nature*, **489**(7417), 537.

Parsons, L., Haque, E., and Liu, H. (2004). Subspace clustering for high dimensional data: a review. *Acm Sigkdd Explorations Newsletter*, **6**(1), 90–105.

Pearl, J. (2014). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann.

Penrose, M. D. (2003). *Random Geometric Graphs*. Oxford University Press.

Pržulj, N. (2007). Biological network comparison using graphlet degree distribution. In *Bioinformatics*, volume 26, pages 177–183.

Pržulj, N. and Higham, D. J. (2006). Modelling protein–protein interaction networks via a stickiness index. *Journal of the Royal Society Interface*, **3**(10), 711–716.

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

Pržulj, N., Kuchaiev, O., Stevanović, A., and Hayes, W. (2010). Geometric evolutionary dynamics of protein interaction networks. In *Biocomputing 2010*, pages 178–189. World Scientific.

Reyna, M. A., Leiserson, M. D. M., and Raphael, B. J. (2018). Hierarchical HotNet: identifying hierarchies of altered subnetworks. *Bioinformatics*, **34**(17), i972—-i980.

Roweis, S. T. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, **290**(5500), 2323–2326.

Schölkopf, B., Tsuda, K., and Vert, J.-P. (2004). *Kernel methods in computational biology*. MIT press.

Smedley, D., Köhler, S., Czeschik, J. C., Amberger, J., Bocchini, C., Hamosh, A., Veldboer, J., Zemojtel, T., and Robinson, P. N. (2014). Walking the interactome for candidate prioritization in exome sequencing studies of Mendelian diseases. *Bioinformatics*, **30**(22), 3215–3222.

Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). BioGRID: a general repository for interaction datasets. *Nucleic Acids Research*, **34**(90001), D535–D539.

Sun, K., Buchan, N., Larminie, C., and Pržulj, N. (2014). The integrated disease network. *Integrative Biology*, **6**(11), 1069–1079.

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K. P., and Others (2014). STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic acids research*, **43**(D1), D447—-D452.

Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., Boutselakis, H., Cole, C. G., Creatore, C., Dawson, E., and Others (2019). COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Research*, **47**(D1), D941–D947.

Tong, A. H. Y., Lesage, G., Bader, G. D., Ding, H., Xu, H., Xin, X., Young, J., Berriz, G. F., Brost, R. L., Chang, M., *et al.* (2004). Global mapping of the yeast genetic interaction network. *science*, **303**(5659), 808–813.

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., and Others (2015). Tissue-based map of the human proteome. *Science*, **347**(6220), 1260419.

Uhlen, M., Zhang, C., Lee, S., Sjöstedt, E., Fagerberg, L., Bidkhori, G., Benfeitas, R., Arif, M., Liu, Z., Edfors, F., and Others (2017). A pathology atlas of the human cancer transcriptome. *Science*, **357**(6352).

Vacic, V., Iakoucheva, L. M., Lonardi, S., and Radivojac, P. (2010). Graphlet kernels for prediction of functional residues in protein structures. *Journal of Computational Biology*, **17**(1), 55–72.

Vapnik, V. N. and Vapnik, V. (1998). *Statistical learning theory*, volume 1. Wiley New York.

Vazquez, A., Flammini, A., Maritan, A., and Vespignani, A. (2001). Modeling of protein interaction networks. *Complexus*, **1**(1), 38–44.

Vazquez, A., Dobrin, R., Sergi, D., Eckmann, J.-P., Oltvai, Z. N., and Barabási, A.-L. (2004). The topological relationship between the large-scale attributes and local interaction patterns of complex networks. *Proceedings of the National Academy of Sciences*, **101**(52), 17940–17945.

Vidal, R. (2011). Subspace clustering. *IEEE Signal Processing Magazine*, **28**(2), 52–68.

Vidal, R., Ma, Y., and Sastry, S. S. (2016). Nonlinear and Nonparametric Extensions. In *Generalized principal component analysis*, pages 25–62. Springer.

Vogelstein, B. and Kinzler, K. W. (2004). Cancer genes and the pathways they control. *Nature medicine*, **10**(8), 789.

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., and Kinzler, K. W. (2013). Cancer Genome Landscapes. *Science*, **339**(6127), 1546–1558.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, **17**(4), 395–416.

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., and Goldenberg, A. (2014a). Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods*, **11**(3), 333–337.

Wang, B., Huang, L., Zhu, Y., Kundaje, A., Batzoglou, S., and Goldenberg, A. (2017). Vicus: Exploiting local structures to improve network-based analysis of biological data. *PLoS Computational Biology*, **13**(10), e1005621.

Wang, D. (2018). The essential role of G protein-coupled receptor (GPCR) signaling in regulating T cell immunity. *Immunopharmacology and Immunotoxicology*, **40**(3), 187–192.

Wang, J., Chen, G., Li, M., and Pan, Y. (2011). Integration of breast cancer gene signatures based on graph centrality. *BMC Systems Biology*, **5**(3), S10.

Wang, X., Xing, E. P., and Schaid, D. J. (2014b). Kernel methods for large-scale genomic data analysis. *Briefings in bioinformatics*, **16**(2), 183–192.

West, D. B. (2001). *Introduction to graph theory*. Prentice Hall.

Windels, S. F. L., Malod-Dognin, N., and Pržulj, N. (2019). Graphlet Laplacians for topology-function and topology-disease relationships. *Bioinformatics*, **35**(24), 5226–5234.

Windels, S. F. L., Malod-Dognin, N., and Pržulj, N. (2022). Graphlet eigencentralities capture novel central roles of genes in pathways. *PLOS ONE*, **17**(1), e0261676.

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., and Others (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research*, **46**(D1), D1074–D1082.

Wu, M. and Scholkopf, B. (2006). A local learning approach for clustering. *Advances in neural information processing systems 18*, pages 1529–1536.

Yajima, I., Kumasaka, M. Y., Yamanoshita, O., Zou, C., Li, X., Ohgami, N., and Kato, M. (2014). GNG2 inhibits invasion of human malignant melanoma cells with decreased FAK activity. *American Journal of Cancer Research*, **4**(2), 182.

Yaveroğlu, Ö. N., Malod-Dognin, N., Davis, D., Levnajic, Z., Janjic, V., Karapandza, R., Stojmirovic, A., and Pržulj, N. (2014). Revealing the hidden language of complex networks. *Scientific Reports*, **4**, 4547.

Yaveroğlu, Ö. N., Milenković, T., and Pržulj, N. (2015). Proper evaluation of alignment-free network comparison methods. *Bioinformatics*, **31**(16), 2697–2704.

Yaveroğlu, Ö. N., Malod-Dognin, N., Milenković, T., and Pržulj, N. (2017). Rebuttal to the Letter to the Editor in response to the paper: proper evaluation of alignment-free network comparison methods. *Bioinformatics*, **33**(7), 1107–1109.

Yu, S., Tranchevent, L.-C., De Moor, B., and Moreau, Y. (2013). *Kernel-based data fusion for machine learning*. Springer.

Zass, R. and Shashua, A. (2005). A unifying approach to hard and probabilistic clustering. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 294–301. IEEE.

Zhu, J., Zhang, B., Smith, E. N., Drees, B., Brem, R. B., Kruglyak, L., Bumgarner, R. E., and Schadt, E. E. (2008). Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nature genetics*, **40**(7), 854–861.

Zitnik, M., Agrawal, M., and Leskovec, J. (2018). Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, **34**(13), i457—-i466.

Supplement A

Graphlet Laplacian

A.1 Determining number of clusters, d

When applying Graphlet Laplacian based spectral clustering, the number of clusters, d, needs to be determined for each of the molecular networks. I determine d by using the rule of thumb: $d = \sqrt{n/2}$ (Kodinariya and Makwana, 2013). The intuition behind this heuristic is that it provides a 'reasonable number' of clusters relative to the size of the network. I validate this approach by inspecting the spectra of the networks. In Section 2.2.1, I explained how the eigenvectors of a Laplacian solve an approximation to the ratio-cut problem, in which a graph is partitioned into similarly sized partitions whilst minimizing the numbers of edges being cut. Additionally, the eigenvalues corresponding to those eigenvectors related to the number of edges cut, as their value provides an upper-bound on the graph-cut suggested by the eigenvector (Mohar, 1997). In Supplementary Figure A.1 I present the first 100 eigenvalues of different Graphlet Laplacians for each network. For each network I indicated the number of clusters d by means of a vertical line. I observe that typically beyond the d^{th} eigenvalue, the curves of each spectrum are relatively flat, i.e. $\lambda_{i+1} - \lambda_i \approx 0$. Hence, clustering beyond d clusters is not warranted, as there is little motivation to choose one additional graph-cut over another.

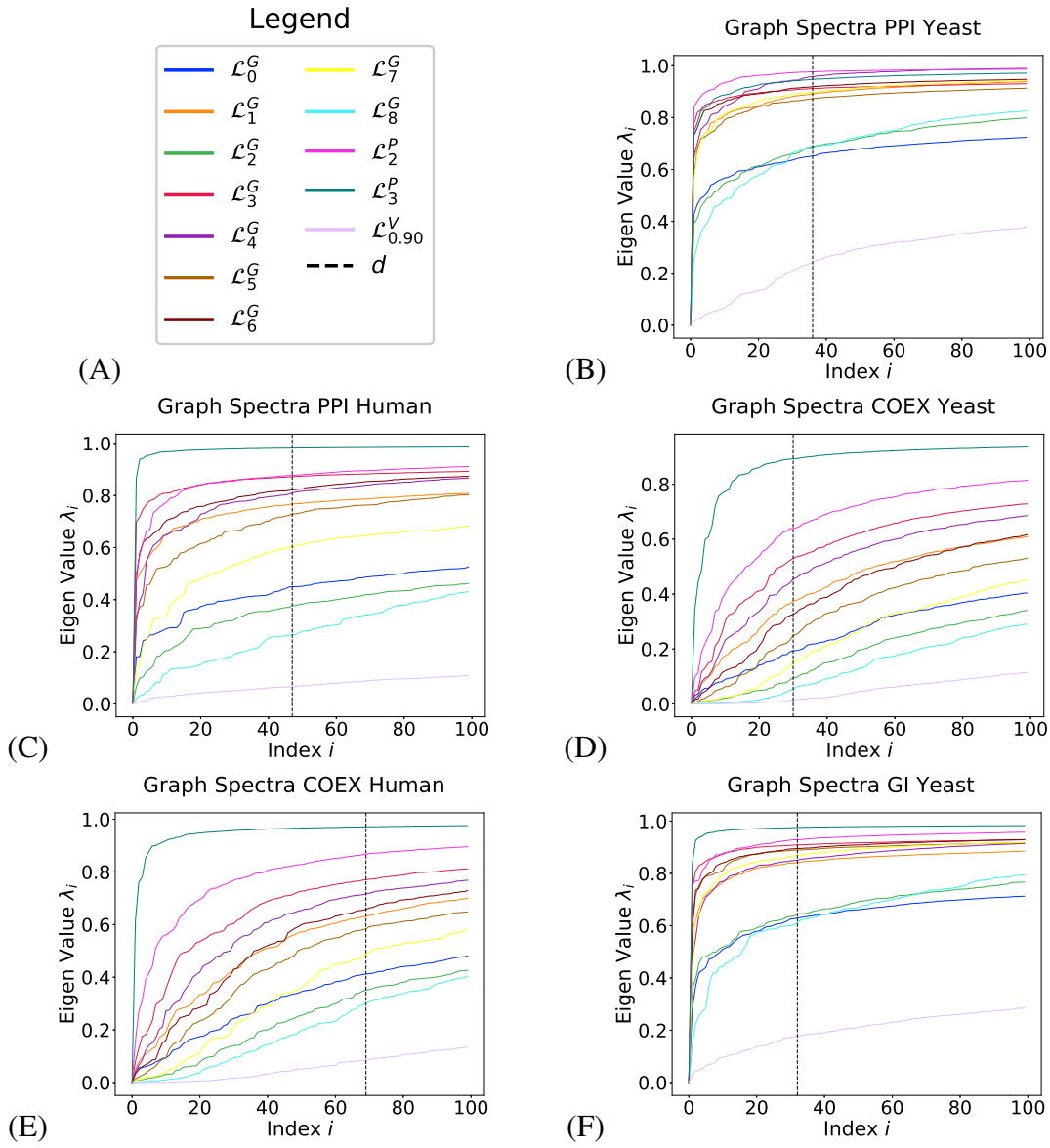


Figure A.1: Spectra of molecular networks. Panel A shows the legend for subsequent panels. Panels B-F show the graph spectra of the Graphlet Laplacians ($\mathscr{L}_i^G$), k-path Laplacian ($\mathscr{L}_2^P$ and $\mathscr{L}_3^P$) and Vicus ($\mathscr{L}_{0.90}^V$), for the set of molecular networks (PPI and COEX for yeast and human, GI for yeast). For each network, the suggested number of clusters, d, is indicated by a vertical line. Beyond the d^{th} eigenvalue, the spectrum is flat.

Supplement B

Pathway-driven NMTF captures the reorganisation of pathways in cancer

B.1 Solving GNMTF

Below, I detail how GNMTF is solved using multiplicative update rules, and how the solver is initialised using Singular Value Decomposition (SVD).

B.1.1 GNMTF Multiplicative update rules

The algorithm to solve GNMTF is presented below in Algorithm 2.

Algorithm 2 Multiplicative update rules GNMTF

Initialise U, S, V using SVD (see Supplementary Section B.1.2)
for all $t = 0, 1, \ldots, t_{limit} - 1$, or until $S^{t+1} = S^t$, $U^{t+1} = U^t$ and $V^{t+1} = V^t$ **do**

$$S_p^{t+1} = S_p^t \odot \sqrt{\frac{((U_p^t)^T A_{pG_i})^+ + (U_p^t S_p^t)^-}{((U_p^t)^T A_{pG_i})^- + (U_p^t S_p^t)^+}};$$

$$U_p^{t+1} = U_p^t \odot \sqrt{\frac{(A_{pG_i} V^t (S_p^{t+1})^T)^+ + U_p^t (S_p^{t+1} (S_p^{t+1})^T)^-}{(A_{pG_i} V^t (S_p^{t+1})^T)^- + U_p^t (S_p^{t+1} (S_p^{t+1})^T)^+}};$$

$$V^{t+1} = V^t \odot \sqrt{\frac{(A_{pG_i}^T U_p^{t+1} S_p^{t+1})^+ + (V^t (S_p^t + 1)^T (U_p^{t+1})^T U^{t+1} S^{t+1})^-}{(A_{pG_i}^T U^{t+1} S^{t+1})^- + (V^t (S^{t+1})^T (U^{t+1})^T U^{t+1} S^{t+1})^+}};$$

Return the last computed (U, S, V);

B.1.2 GNMTF initialisation

To initialise U, S, and V, a truncated SVD is applied on graphlet adjacency matrix $\widetilde{A}_{G_i}$, for the d largest singular values:

$$W \Sigma Z^T = SVD(\widetilde{A}_{G_i}), \tag{B.1}$$

where W and Z^T are $n \times d$ dimensional matrices of which the columns are respectively the d left and right singular vectors of $\widetilde{A}_{G_i}$, and Σ is a $d \times d$ diagonal matrix with the d largest singular values of $\widetilde{A}_{G_i}$ on the diagonal.

The columns of U are based on the columns of W. Specifically, the j^{th} column vector of U, $U[j]$, is initialised based on the j^{th} column vector of W, $W[j]$. To do so, $W[j]$ is split into two non-negative vectors: $W[j]^+$, which is a copy of $W[j]$ with all negative values set to 0, and $W[j]^-$, which is a copy of $W[j]$ where all positive values are set to 0 and all negative entries are set to their absolute value. Either $W[j]^+$ or $W[j]^-$ is assigned to $U[j]$, depending on which one has the highest euclidean norm. Analogously, V is initialised based on Z. S is simply initialized by matrix Σ.

B.2 Solving PNMTF

Below, I detail how PNMTF is solved using multiplicative update rules, and how the solver is initialised using Singular Value Decomposition (SVD).

B.2.1 PNMTF multiplicative update rules

The algorithm used to solve PNMTF is presented below in Algorithm 3.

Algorithm 3 Multiplicative update rules PNMTF

Initialise U, S, V using SVD (see Supplementary Section B.2.2)
for all $t = 0, 1, \ldots, t_{limit} - 1$, or until $S^{t+1} = S^t$, $U^{t+1} = U^t$ and $V^{t+1} = V^t$ **do**

$$S_p^{t+1} = S_p^t \odot \sqrt{\frac{((U_p^t)^T H_{pG_i})^+ + (U_p^t S_p^t)^-}{((U_p^t)^T H_{pG_i})^- + (U_p^t S_p^t)^+}};$$

$$U_p^{t+1} = U_p^t \odot \sqrt{\frac{(H_{pG_i} V^t (S_p^{t+1})^T)^+ + U_p^t (S_p^{t+1} (S_p^{t+1})^T)^-}{(H_{pG_i} V^t (S_p^{t+1})^T)^- + U_p^t (S_p^{t+1} (S_p^{t+1})^T)^+}};$$

$$V^{t+1} = V^t \odot \sqrt{\frac{(H_{pG_i}^T U_p^{t+1} S_p^{t+1})^+ + (V^t (S_p^t + 1)^T (U_p^{t+1})^T U^{t+1} S^{t+1})^-}{(H_{pG_i}^T U^{t+1} S^{t+1})^- + (V^t (S^{t+1})^T (U^{t+1})^T U^{t+1} S^{t+1})^+}};$$

Return the last computed (U, S, V);

B.2.2 PNMTF initialisation

The initialisation of PNMTF is analogous to that of GNMTF. To initialize U_p for a given pathway p, a truncated SVD is applied on graphlet adjacency matrix H_{pG_i}, for the d_p largest singular values:

$$W \Sigma Z = SVD(H_{pG_i}). \tag{B.2}$$

The columns of U_p are initialised based on the columns of W. Specifically, the j^{th} column vector of U_p, $U_p[j]$, is initialised based on the j^{th} column vector of W, $W[j]$. To do so, $W[j]$ is split into two non-negative vectors: $W[j]^+$, which is a copy of $W[j]$ with all negative values set to 0, and $W[j]^-$, which is a copy of $W[j]$ where all positive values are set to 0 and all negative entries are set to their absolute value. Either $W[j]^+$ or $W[j]^-$ is assigned to $U_p[j]$, depending on which one has the highest euclidean norm. S_p is simply initialized by matrix Σ.

To initialize V, a truncated SVD is applied on normalized graphlet adjacency matrix A_{G_i}, computed for the d components corresponding to the d largest singular values:

$$W\Sigma Z = SVD(\widetilde{A}_{G_i}). \tag{B.3}$$

V based is initialised based on Z using the same procure applied to initialize V in GNMTF.

B.3 Determining number of clusters, d

The aim is to extract clusters of pathways from the PNMTF and GNMTF based hierarchical clusterings of pathways so that subsequently I can validate that pathways embedded/clustered together are functionally related (see Supplementary Section 5.2.1). To that end, for both GNMTF and PNMTF, a threshold that defines the number of clusters to be extracted from the hierarchical tree of pathways. To determine the optimal threshold, I cut the tree at different heights, such that the number of exacted clusters, k, varies from 25 to 100. For each value of k, I compute the corresponding silhouette score, which measures how well separated the extracted clusters are as a measure of intrinsic clustering quality. Results for PNMTF and GNMTF based on the different graphlet adjacencies are shown in Supplementary Figure B.1.

For PNMTF, I find that the silhouette scores are non-decreasing and plateau from 65 clusters onwards for all graphlet adjacencies, except for $\widetilde{A}_{G_4}$ and $\widetilde{A}_{G_5}$. For GNMTF, I find that the silhouette scores do not climb and plateau as for PNMTF but rather remain stable regardless of the value of k (except for graphlet adjacencies $\widetilde{A}_{G_2}$ and $\widetilde{A}_{G_8}$). This is because the pathway embeddings do not form easily separable clusters in space with GNMTF. Therefore, for both methods, I cut the tree such that I achieve 65 clusters.

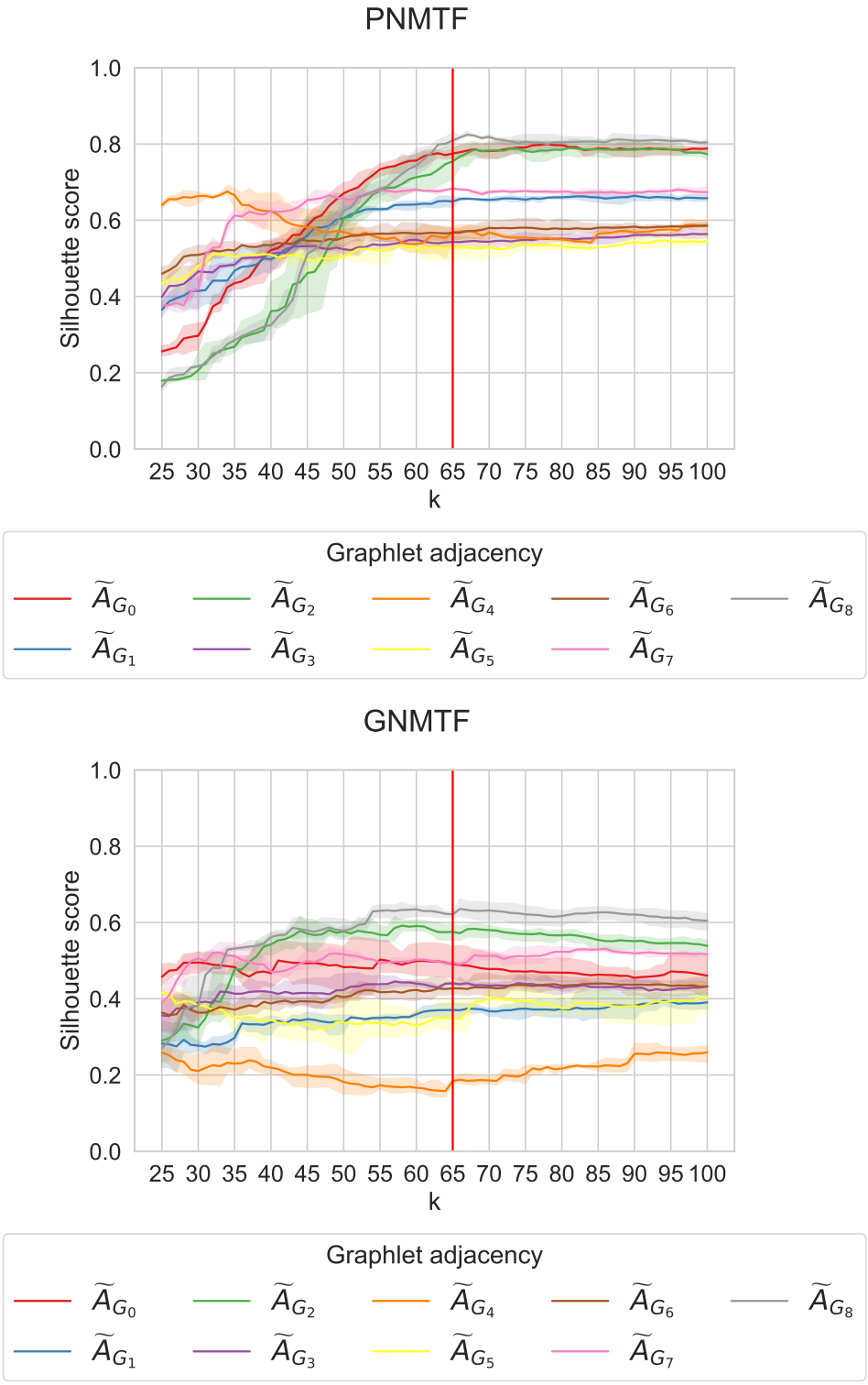


Figure B.1: Determining the optimal number of pathways to extract. For PNMTF (top) and GNMTF (bottom), the silhouette scores (y-axis) extracting k (x-axis) clusters of pathways from the embedding spaces, based on different graphlet adjacencies (colour, legend).

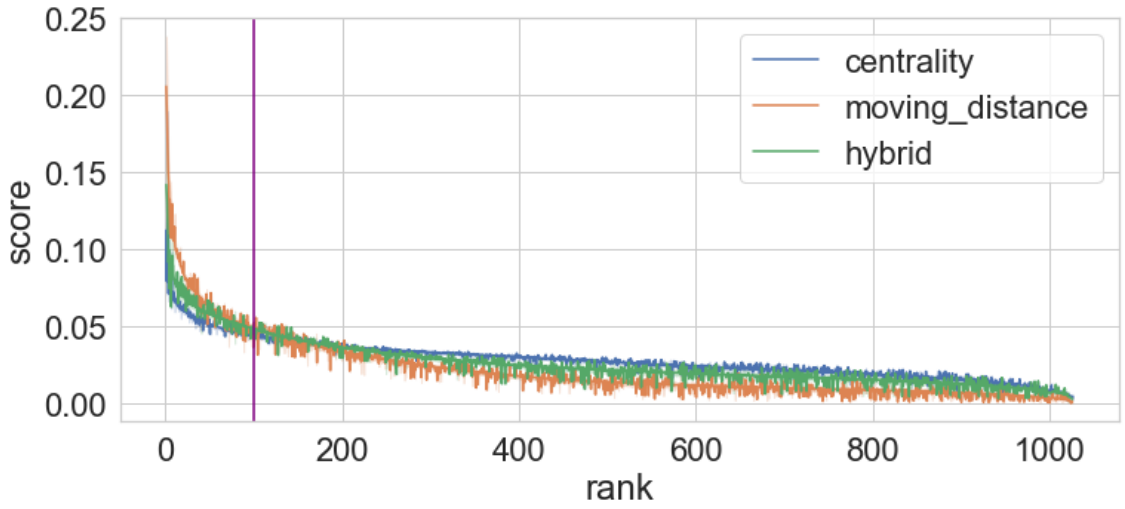


Figure B.2: Determining a threshold for identifying top-scoring pathways through an elbow method. For each of the three different measures (legend), the pathway scores (y-axis) are sorted in descending order (x-axis), averaged over four cell types and graphlet adjacencies (represented as error bands). A vertical purple line indicates the top 100 highest ranked pathways.

B.4 Identifying the set of top-scoring pathways: defining a threshold

Here, I want to define for each of the different NMTF scores applied at the pathway level, a threshold to select top-scoring pathways. To do so, for each of the three different NMTF-scores, graphlet adjacency and tissue, I compute the pathway scores and sort them in descending order. I present the results per NMTF-score, i.e., averaged over the different tissues and graphlet adjacencies, in Supplementary Figure B.2.

The narrow error bands, representing the 95% confidence interval across different tissues and graphlet adjacencies, indicate that the trends across different graphlet adjacencies are similar, allowing to pick a single threshold that holds for all graphlet adjacencies. I choose to consider the top 100 pathways as the set of pathways predicted to be implicated in cancer, as that is where the centrality score, which also underlies the hybrid score, flattens out (see purple line).

B.5 Identifying the set of top-scoring genes: defining a threshold

Here, I aim to define for each of the NMTF scores applied at the gene level, a threshold to select the top-scoring genes that participate in pathways that are in the set of 100 top-scoring pathways. I focus on graphlet adjacency A_{G_1}, as I show it best captures cancer

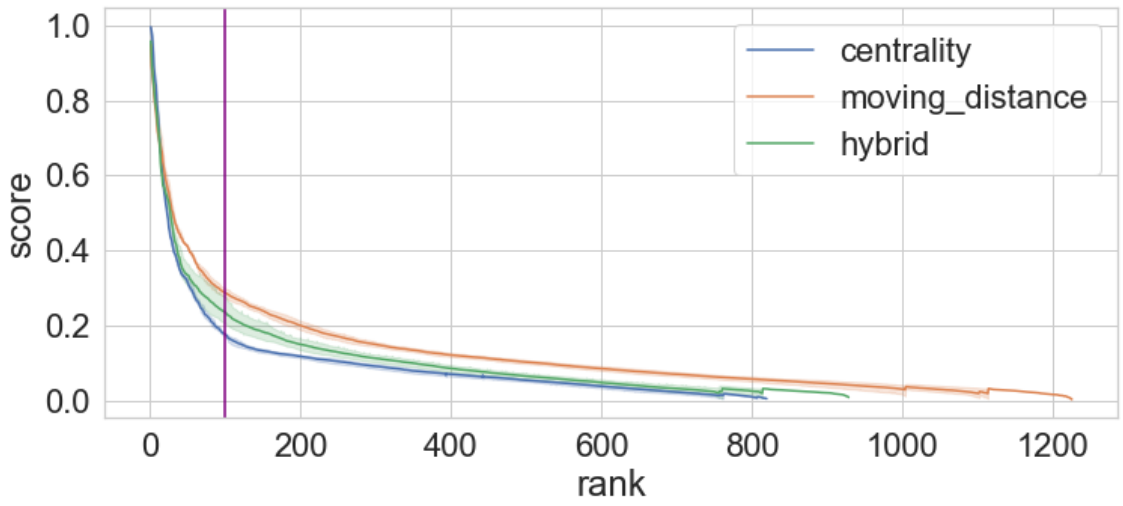


Figure B.3: Determining a threshold for identifying top-scoring genes using an elbow method. Limited to the top 100 highest scoring pathways, for each of the three different NMTF-scores (legend) based on graphlet adjacency A_{G_1}, the gene scores (y-axis), averaged over four cancer types (represented as error bands), in descending order (x-axis).

mechanisms at the pathway level in Section 5.2.2

Specifically, for each of the three NMTF-scores and four different tissues, I compute the gene scores for the genes participating in the top 100 highest scoring pathways and rank them in descending order. Results averaged over the four tissues are presented in Supplementary Figure B.3. I choose to consider the top 100 highest ranked genes (purple vertical line) as the set of predicted cancer-related genes, as from there all three NMTF-scores plateau.

B.6 Gene-level validation

Rank	Symbol	Pathway	Validated in colon cancer	Prognostic	Validation in non-colon cancer	Drugability	Immune System
1	GRB2	CD28 dependent Vav1 pathway	PMID: 12134161 (3, in vitro)			Approved	✓
2	PTPN11	PD-1 signaling	PMID: 32467571 (4, in vitro)			×	✓
3	PIK3R1	RND3 GTPase cycle	COSMIC			Trial	×
4	CSK	Phosphorylation of CD3 and TCR zeta chains	PMID: 20010872 (3, in vitro)			Trial	✓
5	TRAF2	TNF receptor superfamily mediating non-canonical NF-kB pathway		9.31e-4	PMID: 28667915 (1, in vitro)	×	✓
6	SRC	FCGR activation	COSMIC			Approved	✓
7	HSP90AA1	HSF1 activation			COSMIC (non-Hodgkins lymphoma)	Trial	×
8	LMNA	Meiotic synapsis			PMID: 22301279 (prostate, 2,3, in vitro)	×	×
9	EGFR	GRB2 events in EGFR signaling	PMID: 15863375 (1, in vivo)			Approved	×
10	XPO1	Extra-nuclear estrogen signaling	PMID: 26603256 (1, mouse model)			Trial	×
11	CUL1	Prolactin receptor signaling	PMID: 29475926 (2, patient data)			×	✓
12	PTPRJ	Phosphorylation of CD3 and TCR zeta chains	PMID: 12089527 (2, mouse model)			×	✓
13	FN1	p130Cas linkage to MAPK signaling for integrins	PMID: 29274284 (2, 3, in vitro)			Approved	×
14	BIRC3	TNF receptor superfamily mediating non-canonical NF-kB pathway		9.35e-4	COSMIC (leukemia, non-Hodgkins lymphoma, mantele cell lymphoma, multiple myeloma)	×	✓
15	GNB1	G beta:gamma signalling through PI3Kgamma			PMID: 25485910 (leukemia, 1, in vitro)	×	×

Table B.1: Validation of top-scoring genes in colon cancer. The table ranks the top 15 genes and corresponding pathways according to their hybrid score in colon cancer using PNMTF based on graphlet adjacency A_{G_1}. Genes in black have literature support for their role in colon cancer (PubMed IDs or references to COSMIC are given in the column 'Validation in colon cancer'). Genes without literature validation, i.e. predictions, are highlighted in blue. For predicted genes, additional literary evidence highlights their potential role in cancer. Firstly, the p-values for genes that statistically significantly impact patient survival in colon cancer based on Kaplan Meier survival curves are provided in the column 'Prognostic'. Secondly, when available, PubMed IDs or references to COSMIC for 'predicted' genes that have a proven role in cancers other than colon cancer are provided in the column 'Validation non-colon cancer'. The column 'Druggability' indicates which genes are a known drug target for at least one approved drug or one drug in clinical trial according to DrugBank (Wishart *et al.*, 2018). The column 'Immune System' indicates which pathways are immune system pathways according to Reactome.

Rank	Symbol	Pathway	Validated in prostate cancer	Prognostic	Validated in non-prostate cancer	Drugability	Immune System
1	HSP90AA1	Binding and Uptake of Ligands by Scavenger Receptors			COSMIC (non-Hodgkins lymphoma)	Trial	×
2	GRB2	SHC-mediated cascade:FGFR2	PMID: 17372910 (1, mouse model)			Approved	×
3	SHC1	Interleukin-2 signaling	PMID: 29462661 (3, in vitro)			×	✓
4	PTPN11	PD-1 signaling	PMID: 21442024 (2, mouse model)			×	✓
5	PIK3R1	Interleukin-7 signaling			PMID: 20530665 (liver cancer, 4, mouse model) + downregulated in prostate cancer	×	✓
6	TRAF2	TNF receptor superfamily mediating non-canonical NF-kB pathway	PMID: 28667915 (1, in vitro)			×	✓
7	SRC	Nuclear signaling by ERBB4	PMID: 14662770(2, mouse model)	1.63e-4		Approved	×
8	LMNA	Diseases of programmed cell death	PMID: 22301279 (2,3, in vitro)			×	×
9	EGFR	Transcriptional regulation by the AP-2 family of transcription factors	PMID: 32574928 (1, in vitro)			Approved	Approved
10	NTRK1	Signalling to RAS	PMID: 17143529 (1, in vitro)			Approved	×
11	HMGB1	Regulation of TLR by endogenous ligand	PMID: 31410208 (3, in vitro)			Trial	✓
12	UBE2I	SUMOylation of intracellular receptors	PMID: 30631151 (3, in vitro)			×	×
13	PRC1	RHO GTPases activate CIT	PMID: 31327655 (3, in vitro)			×	×
14	TLR4	Regulation of TLR by endogenous ligand	PMID: 18092352 (2, in vitro)			Approved	✓
15	BIRC3	TNF receptor superfamily mediating non-canonical NF-kB pathway	PMID: 31511829 (2, in vitro)			×	✓

Table B.2: Validation of top-scoring genes in prostate cancer. The table ranks the top 15 genes and corresponding pathways according to their hybrid score in colon cancer using PNMTF based on graphlet adjacency A_{G_1}. Genes in black have literature support for their role in prostate cancer (PubMed IDs or references to COSMIC are given in the column 'Validation in prostate cancer'). Genes without literature validation, i.e. predictions, are highlighted in blue. For predicted genes, additional literary evidence highlights their potential role in cancer. Firstly, the p-values for genes that statistically significantly impact patient survival in prostate cancer based on Kaplan Meier survival curves are provided in the column 'Prognostic'. Secondly, when available, PubMed IDs or references to COSMIC for 'predicted' genes that have a proven role in cancers other than prostate cancer are provided in the column 'Validation non-prostate cancer'. Te column 'Druggability' indicates genes that are a known drug target for at least one approved drug or one drug in clinical trial according to DrugBank (Wishart *et al.*, 2018). The column 'Immune System' indicates which pathways are immune system pathways according to Reactome.

Rank	Symbol	Pathway	Validated in ovarian cancer	Prognostic	Validated in non-ovarian cancer	Drugability	Immune System
1	GRB2	Signal attenuation	PMID: 32754300 (1, mouse model)			Approved	×
2	TRIM25	Ovarian tumor domain proteases	PMID: 32826889(2, mouse model)			×	×
3	PTPN11	Signaling by Leptin	PMID: 28814887 (3, in vitro + mouse model)			×	×
4	HSP90AA1	Attenuation phase	PMID: 23135731 (1, 2, in vitro)			Trial	×
5	SRC	RUNX2 regulates osteoblast differentiation	PMID: 27526105 (1, in vitro)			Approved	×
6	EGFR	GRB2 events in ERBB2 signaling	PMID: 22416774 (1, review paper)			Approved	×
7	PIK3R1	RHOF GTPase cycle	PMID: 30755611 (1, mouse model)			Trial	×
8	YWHAB	Frs2-mediated activation	PMID: 30535456 (3, mouse model)			Trial	×
9	LMNA	Diseases of programmed cell death	PMID: 30384980 (4, in vitro)			Trial	×
10	XPO1	Extra-nuclear estrogen signaling	PMID: 27649553 (1, mouse model)			×	×
11	UBE2I	SUMOylation of intracellular receptors			PMID: 30631151 (prostate, 3, in vitro)	×	×
12	HNRNPH1	Signaling by FGFR2			PMID: 34295818 (leukemia, 2, mouse model)	Trial	×
13	MYC	Transcriptional regulation by the AP-2 (TFAP2) fam. of TFs	PMID: 8314536 (1, in vitro)	7.52e-4		Trial	×
14	FN1	p130Cas linkage to MAPK signaling for integrins	PMID: 34093898 (3, in vitro)			Approved	×
15	SYK	Interleukin-2 signaling	PMID: 29643476 (1, 3, in vitro)			Trial	×

Table B.3: Validation of top-scoring genes in ovarian cancer. The table ranks the top 15 genes and corresponding pathways according to their hybrid score in colon cancer using PNMTF based on graphlet adjacency A_{G_1}. Genes in black have literature support for their role in ovarian cancer (PubMed IDs or references to COSMIC are given in the column 'Validation in ovarian cancer'). Genes without literature validation, i.e. predictions, are highlighted in blue. For predicted genes, additional literary evidence highlights their potential role in cancer. Firstly, the p-values for genes that statistically significantly impact patient survival in ovarian cancer based on Kaplan Meier survival curves are provided in the column 'Prognostic'. Secondly, when available, PubMed IDs or references to COSMIC for 'predicted' genes that have a proven role in cancers other than ovarian cancer are provided in the column 'Validation non-ovarian cancer'. The column 'Druggability' indicates which genes are a known drug target for at least one approved drug or one drug in clinical trial according to DrugBank (Wishart *et al.*, 2018). Thee column 'Immune System' indicates which pathways are immune system pathways according to Reactome.