UNIVERSITAT DE BARCELONA

Uncovering the functional organization of molecular interaction networks using network embeddings based on graphlet topology

Daniel Tello Velasco

UNIVERSITAT DE BARCELONA

FACULTAD DE BIOLOGÍA

DOCTORAL PROGRAMME IN BIOMEDICINE

UNCOVERING THE FUNCTIONAL ORGANIZATION OF MOLECULAR INTERACTION NETWORKS USING NETWORK EMBEDDINGS BASED ON GRAPHLET TOPOLOGY

Dissertation of:
Daniel Tello Velasco

Supervisor:
Prof. Nataša Pržulj

Tutor:
Prof. Josep Lluis Gelpi

Uncovering the functional organization of molecular interaction networks using network embeddings based on graphlet topology

Daniel Tello

Natasa

Nataša Pržulj
Thesis supervisor

Josep Lluis Gelpi
Thesis tutor

Universitat de Barcelona, Faculty of Biology

PhD Program: Biomedicine
Developed at: Barcelona Supercomputing Center

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

Abstract

Understanding the functional organization of molecular networks is an ongoing challenge. For this purpose, Spatial Analysis of Functional Enrichment (SAFE) framework was proposed to uncover functional regions in a network by embedding it in 2-dimensions (2D) using the Spring embedding algorithm. However, biological networks often have a heterogeneous degree distribution, i.e., nodes in the network have varying numbers of neighbours. In this case, the Spring embedding sometimes provides uninformative, densely packed embeddings best described as a 'hairball'. On the other hand, hyperbolic embeddings, such as the Coalescent embedding, maps a network onto a disk, so that nodes of high topological importance (i.e., of high node degree) are placed closer to the center of such disk. Additionally, these embedding methods only capture node connectivity information (i.e., which nodes are connected) but does not consider network structure (i.e., wiring or topology), which captures complementary information. The state-of-the-art methods to capture network structure are based on graphlets, which are small, connected, non-isomorphic, induced sub-graphs (e.g., triangles, paths). To better capture the functional organization of networks with heterogeneous degree distributions, taking into account different types of graphlet-based wiring patterns, in this work we introduce the graphlet-based Spring (GraSpring) and the graphlet-based Coalescent (GraCoal) embeddings. Furthermore, we extend the popular SAFE framework to take as input these two newly proposed embedding methods and we use SAFE to evaluate their performance on three types of molecular interaction networks (genetic interaction, protein-protein interaction and co-expression) of various model organisms. We show that the performance in terms of functional information uncovered by each of the embedding algorithms varies depending on the type of network considered and also the model organism considered. For instance, we show that GraCoals better capture the functional and spatial organization of the genetic interaction networks of four species (fruit fly, budding yeast, fission yeast and *E. coli*). Moreover, we discover that GraCoals capture different topology-function relationships depending on the species. We show that triangle-based GraCoals capture functional redundancy in GI networks of species whose genome is characterised by high counts of duplicated genes.

Acknowledgements

It is with deep gratitude and appreciation that I want to acknowledge the people who have supported me in my academic journey. Without their guidance and encouragement, the development of this thesis would not have been possible.

First, I would like to thank everyone in my research group, ICONBI. Specially my Thesis supervisor Nataša Pržulj for giving me the opportunity to advance my scientific career, for her guidance, patience and expertise throughout this project. Second, to my friends and family, specially my parents for supporting and being there for me throughout this process. Finally, I want to thank my beautiful wife for always being by my side during this roller coaster of good and bad moments.

Contents

Chapter 1

Introduction

1.1 Motivation

Systems biology is a discipline in biomedical sciences that studies complex biological interactions on different levels, allowing the identification of patterns that decode the complexity of the biological structure and the processes in the cell, tissues and organ systems (Ideker et al., 2001; Kirschner, 2005). It is the opposite to "reductionism" in biological research, which studies living phenomena at the lowest levels of complexity (Ayala, 1987; Barabasi & Oltvai, 2004) (e.g., studying a single molecule). However, it is evident that a particular biological process or function cannot be attributed to an individual molecule. Instead, biological entities interact with each other in complementary ways to produce a biological product (i.e., a particular phenotype). An important field in systems biology focuses on treating such a complex system of interactions as a network, where the nodes in the network correspond to a particular type of molecule, such as proteins, and the edges connecting them represent a type of interaction, such as the physical binding between the proteins in the cell. In this regard, network biology has been a relevant research area for studying the structure and dynamics of these complex interactions, allowing a better understanding of biological systems, such as the functional or structural properties of the cell (Barabasi & Oltvai, 2004; Baryshnikova, 2016; Emmert-Streib & Glazko, 2011; Ideker & Krogan, 2012; P. Wang, 2022).

In recent decades, advances in high-throughput technologies have increased the availability of genomic, metabolomic, proteomic and transcriptomic data, providing a valuable resource for the study of such complex biological systems in biology and medicine (Barabasi & Oltvai, 2004; Cahan et al., 2014; Emmert-Streib & Glazko, 2011; Silverman et al., 2020). This massive increase in omics data can be attributed primarily to biotechnological breakthroughs achieved in recent decades, including, but not limited to mass spectrometry (Y. Ho et al., 2002), chromatin immunoprecipitation (Iyer et al., 2001) and yeast two-hybrid (Rual et al., 2005; Stelzl et al., 2005). As molecular interaction data become more abundant, so does the complexity of the system it represents (i.e, the dimensionality, where each additional measured molecule adds a dimension to the data). Additionally, despite the biotechnological breakthroughs, many interaction data is characterized by being incomplete and prone to noise (Ning & Lo, 2010; Rajesh et al., 2021). Thus, a constant challenge in network biology, is the need for new computational methods that are more efficient and reliable for complex biological network processing and analysis. A common technique for processing and analyzing the data, is through network embeddings, which are methods for extracting a low-dimensional representation of the data (Nelson et al., 2019), while conserving the original similarity features in the data (Arsov & Mirceva, 2019). These lower representations of the data can later be used for downstream analysis such as for annotation of genes (García-Díaz et al., 2020), or for protein structure prediction (Dhingra et al., 2020).

1.2 Objectives

In this work, we extend popular network embedding methods such as the Spring embedding and the Coalescent embedding by combining these methods with graphlet topology. Additionaly, we improve and extend the Spatial Analysis of Functional Enrichment (SAFE) framework (Baryshnikova, 2016) to include these new methods and use SAFE to evaluate the performances of these graphlet based embeddings. Lastly, we use these graphlet based embeddings with SAFE to explore the functional organization of molecular interaction networks of model organisms to uncover new biological insights.

1.3 Contributions

In this thesis we introduce new methods for embedding molecular networks based on graphlets. We generalise the popular Spring embedding to graphlet-based Spring embedding. In brief, we use the graphlet adjacency matrix of a network instead of the standard adjacency matrix to embed the nodes in 2D using the Spring embedding. We also generalise the popular Coalescent embedding to the graphlet-based Coalescent (GraCoal) embedding. In brief, this method performs dimensionality reduction on the matrix representation of a network such as the Laplacian matrix to obtain an angular coordinate for each node in the network and computes a radial coordinate based on the degree of each node in the network. By extending this method to graphlets, we perform dimensionality reduction on the graphlet Laplacian matrix to obtain angular coordinates and compute the radial coordinate based on the graphlet degree of the nodes. We extend the SAFE framework to also consider graphlet-based Spring embedding and GraCoal embedding as additional embedding methods. Lastly, because our GraCoal embedding is based on the eigendecomposition of the Laplacian matrix of a particular graphlet, we also extend SAFE to consider graphlet based Spectral embedding. We compare the three graphlet based embeddings by annotating different types of molecular networks of various model organisms with SAFE. We show that when using graphlet based embeddings with SAFE, additional functional information can be captured as opposed to using SAFE with standard adjacency matrix (i.e., not based on graphlet information). Because graphlets capture different wiring patterns, we show for instance, that each GraCoal embedding used in SAFE captures unique biological functions. This demonstrates that GraCoal embeddings used in SAFE can be used in complementary ways to uncover the functional information encoded by molecular networks.

Resulting Papers

The following manuscript has been submitted for review to Bioinformatics:
Tello D., Windels S.F., Rotkevich M., Malod-Dognin N. and Pržulj N. Graphlet Coalescent embeddings capture the functional organization of the cell. *Bioinformatics*

Posters and Talks

The following poster was presented at ISMB-ECCB 2021:
Tello D., Windels S.F., René Böttcher, Malod-Dognin N. and Pržulj N. *Graphlet-based Coalescent embedding uncovers complementary biological information in yeast molecular networks.*

1.4 Thesis outline

The thesis is outlined as follows:

In **Chapter 2** we present all the relevant concepts and definitions related to molecular biology and network biology that are needed for the development of this work.

In **Chapter 3** we present the new methods developed during this work. In particular, we introduce the extend version of the Spring embedding used in SAFE (i.e., graphlet-based Spring embedding) and also the newly proposed GraCoal embedding. Furthermore, we describe the modifications done to SAFE to further extend and improve the framework, in particular to embed molecular networks using the newly proposed graphlet based embeddings.

In **Chapters 4-6**, we evaluate the graphlet-based embeddings with SAFE on three different molecular interaction network types for various model organisms. We first present, in **Chapter 4**, the results for the genetic interaction (GI) networks of *Drosophila melanogaster*, *Escherichia coli*, *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* and also for the genetic interaction similarity (GIS) network of *Saccharomyces cerevisiae.* In **Chapter 5** we present the results for the PPI networks of *Caenorhabditis elegans*, *Drosophila melanogaster*, *Escherichia coli*, *Homo sapiens*, *Mus musculus*, *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe.* Lastly, we present, in **Chapter 6**, the results for the co-expression (COEX) networks of *Caenorhabditis elegans*, *Drosophila melanogaster*, *Homo sapiens*, and *Saccharomyces cerevisiae.*

Chapter 2

Background

In this chapter we introduce key concepts in molecular biology and network biology. In brief, we present an overview of different types of molecular interaction networks used in network biology as well as the main types of experimental methods that allow to identify such interactions. Next, we define key concepts related to network structural properties and provide an overview of the model networks most commonly used in network biology that are particularly useful for understanding the global structure of a network. We also go over higher order network representations such as the graphlet adjacency matrix and the graphlet Laplacian matrix. Lastly, we define the three network embedding algorithms used in this work (Spring embedding, Coalescent embedding and Spectral embedding) and describe in detail the Spatial Analysis of Functional Enrichment (SAFE) framework.

A well-established approach in systems biology is to analyze large scale omics data by modeling them as networks, where molecules are represented as nodes that are connected by an edge if they express any type of interaction. Nodes that are connected by an edge are commonly referred to as “neighbors” in the network. For instance, one of the most widely studied types of interactions in the cell are the physical interactions that occur between proteins, which can be modeled as a protein-protein interaction (PPI) network. To this end, network biology has facilitated the understanding of large and complex interactions that describe biological systems, solving biological enigmas such as how and where these interactions occur (Baryshnikova, 2016; Luck et al., 2019; Niu et al., 2012; Rizzolo et al., 2017; Vissiennon et al., 2017; Yan et al., 2018; Youn et al., 2018). The massive increase in available omics data can be attributed primarily to biotechnological breakthroughs achieved in recent decades, including, but not limited to mass spectrometry (Y. Ho et al., 2002), chromatin immunoprecipitation (Iyer et al., 2001) and yeast two-hybrid (Rual et al., 2005; Stelzl et al., 2005). These methods have allowed for an increasing availability of molecular interaction data in particular across model organisms, including human (Huang et al., 2016; X. Li et al., 2010; Luck et al., 2020; Uetz et al., 2000). A common task in network biology is to study the structural properties of a network, providing valuable insights into the topology and geometry of the network (Bianconi & Rahmede, 2017; Knabe, 2013; Vella et al., 2018; Wu et al., 2015). This untangled information can later be used for downstream analysis, such as uncovering associations between disease and RNA molecules (G. Li et al., 2017), or predicting perturbation patterns in biological networks (Santolini & Barabási, 2018).

2.1 Molecular interaction networks

In recent decades, advances in high-throughput technologies have increased the availability of genomic, metabolomic, proteomic and transcriptomic data, providing a valuable resource for the study of such complex biological systems in biology and medicine (Barabasi & Oltvai, 2004; Cahan et al., 2014; Emmert-Streib & Glazko, 2011; Silverman et al., 2020).

In this section, we define the main types of molecular interaction networks used in network biology. In particular, we define protein-protein interaction networks, genetic interaction networks and co-expression networks. Furthermore, we discuss different types of annotation data, which are commonly used in parallel with molecular interaction networks.

2.1.1 Protein-protein interaction networks

Protein-protein interaction (PPI) networks represent the physical interactions between gene products (i.e., proteins). Most studies of PPI networks focus on the direct physical bindings between proteins, such as an enzyme physically interacting with another molecule to catalyze a specific reaction that occurs in a living cell or organism. However, indirect interactions such as those involving proteins in the same protein complex or level-2 interactions (i.e., proteins that share the same interaction neighbors) can also be used, and can be particularly useful for studying and predicting protein complexes (Chua et al., 2008). In a PPI network, the proteins are represented as nodes, and the links connecting pairs of nodes (i.e., edges) represent the physical binding (or indirect interaction) of the proteins (Gligorijević & Pržulj, 2015). When constructing a PPI network, the gene names/labels that encode each particular protein product are often used over protein names/labels, which may facilitate downstream analysis, for instance for comparison with other gene-based molecular networks (e.g., genetic interaction or co-expression). Analysis of PPI networks have proved to be useful for various tasks such as protein function prediction, protein complex prediction, drug discovery, uncovering disease mechanisms and uncovering the relationships between the proteins within the cell (Athanasios et al., 2017; Chua et al., 2008; Davis et al., 2015; Dobson et al., 2014; Piovesan et al., 2015; Safari-Alighiarloo et al., 2014; Vazquez et al., 2003), and to study different biological phenomena in the cell, such as gene regulation (Jiang et al., 2020; Mercatelli et al., 2020; J. Wang et al., 2006), disease mechanisms (Chakraborty et al., 2014; Kuzmanov & Emili, 2013; Safari-Alighiarloo et al., 2014) or signaling pathways (Giot et al., 2003; S. Li et al., 2004). Some of the most widely used experimental methods for detecting and identifying protein-protein interactions are the yeast two-hybrid (Y2H) system (Ito et al., 2001; Ito et al., 2000; Rual et al., 2005; Simonis et al., 2009; Van Criekinge & Beyaert, 1999) and techniques based on mass spectrometry (Collins et al., 2007; Gavin et al., 2002; Krogan et al., 2006; Rigaut et al., 1999), which we define in following sections.

2.1.2 Genetic interaction networks

Genetic interaction (GI) networks are a type of molecular interaction network that model the interactions between genes in a cell (Costanzo et al., 2010; Costanzo et al., 2016). In brief, two genes are said to genetically interact if a simultaneous mutation in both genes produces a phenotype that differs from the phenotype of each individual mutated gene (Mani et al., 2008). Thus, in a GI network, the genes are represented as nodes, and the links connecting pairs of nodes (i.e., edges) indicate there is a genetic interaction between them. Typically, these types of interactions are detected with experimental methods, such as genetic screens (Costanzo et al., 2010; Costanzo et al., 2016; Lehner et al., 2006; Yan Tong & Boone, 2006) that evaluate, if a particular phenotype in the cell varies significantly when two genes are simultaneously mutated (i.e., a double mutant) with respect to the phenotype of each mutated gene (Boucher & Jenna, 2013; Costanzo et al., 2010; Costanzo et al., 2016). Genetic interactions are classified in two major categories: positive genetic interactions and negative genetic interactions. Positive interactions occur when the combined effect of two mutated genes (double mutant) result in a fitness phenotype that is greater than the fitness phenotype of each individual mutation (Baryshnikova et al., 2010; Boucher & Jenna, 2013; Kuzmin et al., 2018). On the other hand, negative interactions occur when the combined effect of two or more mutations result in a phenotype that is more severe than the phenotype of each individual mutation (Baryshnikova et al., 2010; Boucher & Jenna, 2013; Kuzmin et al., 2018). For instance, synthetic lethality is an extreme example of a negative genetic interaction and occurs when two mutations, neither of which is lethal on its own, combine and lead to an inviable double mutant phenotype (Bender & Pringle, 1991; Kuzmin et al., 2018; Novick & Botstein, 1985). Lastly, the genetic interaction profiles of all the genes in a GI network are useful for constructing a similar type of molecular interaction network, a genetic interaction similarity (GIS) network. To construct a GIS network, the genetic interaction profiles of all the genes in the GI network are compared to one another to evaluate how similar their interaction patterns are. In a GIS network, two genes are connected by an edge if they share similar interaction profiles. To this end, interaction profiles of pairs of genes are compared by computing the Pearson correlation coefficient (PCC). Finally, gene pairs with profile similarity of PCC>0.2 are connected in the newly constructed GIS network. Larger values of the PCC threshold can be used to construct a more stringent GIS network (Costanzo et al., 2010; Costanzo et al., 2016).

2.1.3 Co-expression networks

Co-expression (COEX) networks are a type of molecular interaction network that model the interaction between genes based on their patterns of gene expression. In a COEX network, each gene is represented by a node and two pairs of nodes are connected by an edge if they are expressed simultaneously (Stuart et al., 2003). In this regard, a gene is said to be expressed if the information it encodes (i.e., DNA) is transcribed into an RNA molecule (i.e., a transcript), which later is translated and processed into a functional protein. Typically, gene expression can be detected using experimental methods such as RNA-seq (Z. Wang et al., 2009), which use deep sequencing technology to measure the level of transcript in a sample to determine if a particular gene is expressed. Thus, for constructing a COEX network, transcriptomics data is typically used, such as microarrays or RNA-seq, providing expresion values for all genes in a particular sample. Next, with the expression values from different experimental conditions, a pair-wise similarity score is computed for all the genes. For instance, a Pearson correlation coefficient (PCC) is usually used to construct a PCC matrix. For each gene, all other genes are ranked according to their correlation values. Finally, a threshold can be applied to the ranks to keep only the strongest correlation values. For instance, to build a highly reliable network, keeping the top 1% is typically used (Obayashi et al., 2019).

2.1.4 Annotation data

As molecular interaction data continue to increase, so does the knowledge about what characterizes each gene or protein in a cell, for instance which particular functions they carry out or in which part of the cell they might be localized. In this regard, molecular networks annotation data provides the necessary information that relates the nodes in a molecular network (e.g. proteins or genes) to particular properties or features such as biological processes, molecular functions or biological pathways and are used to understand how the nodes in the network relate to each other. To date, there exist multiple public databases that provide a valuable resource for molecular annotation data, such as KEGG (Kanehisa et al., 2023) or GeneOntology (Ashburner et al., 2000). The most common types of annotation data include functional information, such as Gene Ontology (GO) terms (i.e., GO biological processes, GO cellular components and GO molecular functions), metabolic pathways, and protein domains, as well as contextual information, such as subcellular localization and tissue expression patterns. Annotation data and molecular networks go hand by hand when analysing molecular interaction networks, as annotations are used to enrich molecular network models providing, for instance, biological insight into the functional organization of the network. In summary, molecular networks annotation data provides important information about the properties and functions of entities in a molecular network and is used to understand the relationships and processes within the network.

2.2 Experimental methods

Molecular interaction data has become a valuable resource for studying and understanding complex biological systems in the cell. The quantity and quality of these data continues to increase as biotechnological breakthroughs are achieved and experimental costs become cheaper. In this section we review some of the widely used experimental methods to detect molecular interactions.

2.2.1 Yeast 2-hybrid

The yeast two-hybrid (Y2H) system is a method for identifying PPIs in living cells of the budding yeast, Saccharomyces cerevisiae (Uetz et al., 2000). The Y2H system works by expressing two proteins, one as a bait and one as a prey. The bait protein is fused to a DNA-binding domain and the prey protein is fused to an activation domain. If the two proteins interact, they bring the DNA-binding and activation domains into close proximity, leading to the activation of a reporter gene. The reporter gene is usually a gene that confers an easy to appreciate phenotype, such as fluorescence or growth in a medium (Brückner et al., 2009; Uetz et al., 2000). The Y2H system has become one of the most widely used methods for identifying protein-protein interactions. One advantage of the Y2H system is that it can be used to detect interactions between proteins that are not easily detectable using other methods, such as interactions between intracellular proteins and membrane associated proteins (Brückner et al., 2009).

2.2.2 Mass spectrometry

Mass spectrometry (MS) is an experimental method used to measure the mass and abundance of proteins in a given sample (Glish & Vachet, 2003). To do this, it ionizes the sample molecules (i.e, charging it positively or negatively) and then analyzes the resulting ions by measuring the mass to charge ratio m/z. The m/z ratio of each protein can then be used to computationally identify the proteins in the sample by searching in large databases specific to the organism of interest (Richards et al., 2021). A widely used MS based technique for identifying protein-protein interactions is Crosslinking MS, which involves a chemical reagent (i.e., a cross-linker) between two functional groups in a protein or a protein complex. The cross-linker has a defined length, which allows for subsequent breaking of the cross-links and analyzing the resulting peptides by mass spectrometry (O'Reilly & Rappsilber, 2018; Piersimoni et al., 2021).

2.2.3 Affinity purification

Affinity purification methods are used to isolate a protein of interest or group of proteins from a given sample. To isolate the protein of interest (i.e., a target protein), a ligand, covalently attached to a solid support (i.e., a resin or bead), binds with the protein with high specificity (Kadonaga & Tjian, 1986). In brief, the sample with the target protein is passed through the affinity resin or bead, such that only the target protein is bound to it (i.e., to the ligand in the resin or bead). This allows non-interesting proteins and other molecules to continue passing through the resin. The resin may be washed to remove any particles or proteins that are not desired. Finally, to separate and isolate the target protein from the resin, an elusion solution might be used to break the binding of target and ligand in the resin (Kadonaga & Tjian, 1986). To detect protein-protein interactions, the proteins isolated with affinity purification are assessed with mass spectrometry for proper identification.

2.2.4 Tandem affinity purification (TAP)

Tandem Affinity Purification (TAP) is an extension of the affinity purification which isolates a protein of interest (i.e., a target protein) or group of proteins from a given sample (Puig et al., 2001). In a TAP, the target protein is sequentially bound to two different ligands which are covalently attached to two solid supports (i.e., resin or bead). The first step in a TAP is identical to the traditional affinity purification: target protein is bound to the resin via the ligand and subsequently purified with a washing step to remove any non-desired proteins or molecules. The second step consists in purifying the target protein further by repeating the process through a second resin with a different ligand which is also highly specific to the target protein. The washing step is repeated to remove non-desired proteins and molecules from the second resin and the target protein is eluted. These two rounds of binding to different ligands make TAP a powerful and highly specific method for protein purification. Lastly, mass spectrometry is typically used on the purified protein for proper identification.

2.2.5 Protein fragment complementation assay

Protein Fragment Complementation Assay (PCA) is a method used to study protein-protein interactions in vitro (Galarneau et al., 2002; Remy et al., 2007; Remy & Michnick, 1999). The intuition behind this method assumes that two fragments of a target protein will interact and reform the full-length functional protein when brought together in close proximity. The target protein is cleaved into two fragments, each of which is expressed as a fusion protein with a different, easily identifiable and measurable reporter protein, such as green fluorescent protein (GFP) or luciferase (Chalfie, 1995). If there is no interaction, the two fragments remain separate and the full-length functional protein is not reformed. On the other hand, when the interaction of the two fragments occur, the full-length functional protein is reformed, which exhibits the activity of the reporter protein (i.e., the fluoresence or luminiscence).

2.2.6 Synthetic genetic array (SGA)

Synthetic genetic array (SGA) is a high-throughput screening method for identifying genetic interactions in yeast. It involves creating libraries of yeast strains that each contain a deletion of a single gene, and then using these strains to systematically test all possible pairwise combinations of deleted genes to identify those that exhibit synthetic growth defects when deleted together. Thus, it is based on the concept of synthetic lethality, which is when two mutations that are individually viable (i.e., non lethal) become lethal when combined. By systematically testing pairs of mutants, SGA can be used to identify genetic interactions between genes, and to uncover new functional connections within cellular pathways and networks (Costanzo et al., 2010; Costanzo et al., 2016).

2.2.7 Microarrays

Microarrays are biotechnological tools used in molecular biology and genomics to study gene expression and detect genetic variations (Cheung et al., 1999). Two major types of microarrays are typically used, which are based either on DNA chips (Stoughton, 2005) or gene chips (Johnston, 1998). They consist of a solid support, usually a glass slide, coated with an array of microscopic spots of nucleic acid probes. During a microarray experiment, a sample of labeled nucleic acids, typically cDNA or RNA, is hybridized to the probes on the microarray. The hybridization signals are then measured and used to quantify the expression level of the genes represented by the probes. Microarrays can be used to study the expression of thousands of genes simultaneously, making them a powerful tool for gene expression profiling and functional genomics.

2.3 Network Analysis

Network analysis studies the patterns in connections (i.e., edges) and relationships between the elements (i.e., nodes) in a network, providing insights into the global and local structure of the network. In network biology, studying the structure of molecular interaction networks provides researchers valuable information to understand the underlying processes that occur in complex biological systems. In this section we review some of the most basic network descriptors, such as the size, diameter and clustering coefficient. Next, we define four common centrality measures used in network analysis to evaluate the importance of each node in the network. Lastly, we define graphlets and graphlet based methods for network analysis.

2.3.1 Global network descriptors

2.3.1.1 Size

The size of a network refers to the total number of nodes and edges in the network.

2.3.1.2 Density

The density of a network is the number of edges as a percentage of the max possible number of edges in the network. Real networks have low density

2.3.1.3 Shortest path lengths

The shortest path length is defined as the minimum number of edges along a path that needs to be traversed between a pair of any given nodes in a network. This measure provides a way of quantifying the distance or separation between nodes in a network. The shortest path length between two nodes is commonly used to study the structure and properties of networks. For instance, it can be used to analyze the robustness of a network by measuring how quickly information can be transmitted from one node to another, or to measure the average distance between all pairs of nodes in the network, known as the characteristic path length. Shortest path lengths can be calculated using algorithms such as Dijkstra's algorithm (Dijkstra et al., 2021) or Bellman-Ford algorithm (Bellman, 1958). The choice of algorithm depends on the type of network being analyzed and the specific problem being solved. In weighted networks, where each edge has an associated weight or cost, the shortest path length is calculated as the sum of the weights along the shortest path. In unweighted networks, where all edges have the same weight, the shortest path length is simply the number of edges in the shortest path.

2.3.1.4 Diameter

The diameter of a network is a measure of its breadth, defined as the longest shortest path between all pairs of nodes. It is measured as the number of edges along the shortest path between any given two nodes. Real networks are small world, meaning they have short diameter

2.3.1.5 Clustering coefficient

The clustering coefficient of a node in a network is a measure of the degree to which the neighbors of the node are interconnected. It is defined as the ratio of the number of actual connections between the neighbors of a node to the maximum number of connections that could exist between the neighbors. Finally, the clustering coefficient of a network is the average of the clustering coefficient over all nodes in the network. Real networks have large clustering coefficient.

2.3.1.6 Degree distribution

Degree distribution is a statistical property of a network that describes the distribution of the number of connections (i.e., edges) that each node in the network has (i.e., the degree of a node). It is typically represented as a histogram or a probability distribution function that shows the frequency or probability of nodes with a given degree (number of connections). It provides insight into the overall structure of the network and the way information or signals might propagate through it. For instance, a network with a power-law degree distribution, where the majority of the nodes in the network have one or very few connections and a small proportion of nodes have a lot of connections (Broido & Clauset, 2019; Moreira et al., 2009). These types of networks are said to have a scale-free structure and are characterized by a high degree of heterogeneity and robustness (Moreira et al., 2009; B. Wang et al., 2006). The power-law equation can be represented as:

$$P(k) \propto k^{-\gamma}$$

Where $P(k)$ is the probability that a node has degree k, and γ is the scaling exponent. The value of γ determines the shape of the degree distribution, with larger values of γ corresponding to more homogeneous distributions and smaller values of γ corresponding to more heterogeneous distributions. In scale-free networks, the exponent of the power-law distribution usually falls in the range of $2 < \gamma < 3$ (Ravasz & Barabási, 2003).

Scale-free networks are commonly observed in many complex systems, including the internet, social networks, biological networks, and technological networks, among others. They are considered to be robust and resilient to the removal or failure of nodes, as the hub nodes provide alternative paths for information or signals to flow (Barabási & Albert, 1999; Broido & Clauset, 2019; Moreira et al., 2009).

However, they are also vulnerable to targeted attacks on the hub nodes, as their removal can have a significant impact on the structure and function of the network. For instance, perturbing highly connected nodes in PPI networks is more likely to impact cell viability (Jeong et al., 2001b). The scale-free structure of networks can have important implications for the way information spreads, the way resources are distributed, and the way the network functions as a whole. Thus, understanding the scale-free structure of networks and how to manipulate it to achieve a desired outcome is an important area of research in many fields, including physics, computer science, biology, and sociology, among others.

2.3.2 Node centralities

Node centrality is a measure of the importance or significance of a node within a network. Centrality measures provide a way to quantify the relative influence or importance of nodes in a network, and can be used to identify essential nodes, such as genes or proteins in complex biological systems.

There are several different centrality measures, each with a slightly different interpretation and focus. In the following section we define two types of node centralities: based on connectivity, such as degree centrality and eigenvector centrality, and based on occurrence of paths in the network, such as betweenness centrality and closeness centrality.

2.3.2.1 Degree centrality

The degree centrality of a node is simply the number of connections (edges) it has to other nodes in the network. It thus considers highly connected nodes to be the most important nodes in the network.

It can be represented mathematically as:

$$DC(u) = \sum_{v=1}^{n} A_{u,v}$$

where $DC(u)$ is the degree centrality of node u, and $A_{u,v}$ is the element of the adjacency matrix that represents the connection between nodes u and v. The sum is taken over all nodes v in the network, and the resulting value represents the number of connections node u has to other nodes in the network.

2.3.2.2 Eigenvector centrality

The eigenvector centrality of a node is a measure of the influence of a node (i.e., how many connections the node has) but takes into account the influence of the nodes it is connected to. Thus, a node that has a large degree centrality (i.e., is well connected) will only have a large eigenvector centrality if the neighbors of the node also have a large degree centrality. In this way, nodes with large eigenvector centrality are considered to be influential not only because of their direct connections, but also because of their connections to other highly connected nodes. The eigenvector centrality of node u is defined as the average of the centralities of the n neighbors:

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

where EC is the eigenvector centrality of a node in the network, A is the adjacency matrix of the network, and λ is the eigenvalue associated with A.

2.3.2.3 Betweenness centrality

The betweenness centrality of a node measures the extent to which the node lies on the shortest paths between other nodes in the network. Nodes with high betweenness centrality are often considered to be important bottlenecks or intermediaries in the network. Formally, the betweenness centrality of a node is represented as:

$$BC(u) = \sum_{s,t \in V} \frac{\sigma_{st}(u)}{\sigma_{st}}$$

where $BC(u)$ is the betweenness centrality of node u, V is the set of all nodes in the network, σ_{st} is the number of shortest paths from node s to node t, and $\sigma_{st}(u)$ is the number of these shortest paths that pass through node u. The betweenness centrality of a node is proportional to the fraction of all shortest paths in the network that pass through the node.

2.3.2.4 Closeness centrality

The closeness centrality of a node measures the inverse of the sum of the shortest distances from the node to all other nodes in the network. Nodes with high closeness centrality are considered to be well-connected and influential, as they are able to reach many other nodes quickly. Thus, the closeness centrality is a measure of the accessibility of a particular node to all other nodes in the network. Formally the closeness centrality is represented as:

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

where $CC(u)$ is the closeness centrality of node i, n is the total number of nodes in the network, $d(u,v)$ is the distance between nodes u and v, and the sum is taken over all nodes v in the network.

2.3.3 Community structure

Community structure refers to the clustering of nodes in a network into groups or modules based on their connectivity patterns. In a network with a strong community structure, nodes within the same community are highly interconnected, while nodes in different communities have relatively few connections (Favila & Halffter, 1997; Girvan & Newman, 2002). Community structure is a common feature of many complex networks, such as social networks, biological networks, and technological networks. It is thought to reflect the underlying organization of the network and to play a critical role in the network's function and dynamics (Danon et al., 2005; M. E. Newman, 2006). The detection of community structure in a network is a central problem in network analysis, and there are many algorithms and techniques for detecting communities based on different criteria, such as modularity, cliques, or core-periphery structure (Ma et al., 2010; M. E. Newman, 2004; Sun et al., 2009).

Once detected, the community structure of a network can be used to study a variety of questions related to the network's organization and function, such as how information spreads, how resources are distributed, and how the network evolves over time. Thus, understanding the community structure of networks is an important area of research in many fields, including physics, computer science, biology, and sociology, among others.

2.3.4 Graphlets

A popular task in network science requires to quantify the network neighborhood of a node (i.e., the local topology of the node). To this end, one of the widely used measures for this purpose is the number of neighbors that each node has (i.e., the degree of a node). However, because it only considers the direct connections of a node, the information recovered from this measurement is very limited. Thus, graphlet-based methods have been proposed as state-of-the-art to quantify the local topology around each node in a network. Graphlets, illustrated in Fig 2.1, are defined as small, connected, non-isomorphic induced subgraphs in a graph, (Przulj et al., 2004) and have been used, for instance, to predict protein function (Davis et al., 2015) and to identify new cancer genes directly from their interaction patterns (Milenković et al., 2010) in PPI networks.

These powerful tools have been used not only for uncovering local structural (topological) patterns and their relation to biological function, but also for characterization and comparison of complex networks (Aparicio et al., 2017; Aparício et al., 2015; Cannoodt et al., 2018; Martin et al., 2017; Przulj, 2007; Sarajlić et al., 2016). Through functional analysis, it has been demonstrated that graphlets can capture the functional organization of biological networks (Dale, 2017; Hulovatyy et al., 2015; Winterbach et al., 2013; Yaveroğlu et al., 2014) and have also been generalized to other applications, such as the graphlet Laplacian matrix, to demonstrate that different graphlet topology can uncover different biological functions (Windels et al., 2019).

Formally, we define graphlets as follows. Let $G = (V, E)$ be a graph, where V is the set of vertices and E is the set of edges. A graphlet g is a subgraph of G, defined as $g = (V', E')$, where $V' \subseteq V$ and $E' \subseteq E$ such that the subgraph $G[V']$ is connected. Additionally, graphlets are induced subgraphs, meaning they contain all the vertices of the original graph that belong to the selected subset, as well as all the edges that connect those vertices. Moreover, graphlets are characterized by having orbits (See Figure 2.1), also called automorphism orbits, which are defined as symmetry groups of nodes within a graphlet, and are used to characterize different topological positions of a node in a graphlet (Yaveroğlu et al., 2014). One particular characteristic of the automorphism orbits of a graphlets, is such that swapping nodes within the orbit preserves the structure of the graphlet (Przulj, 2007). Some widely used graphlet-based measures include the graphlet degree vector (GDV), which can provide valuable local topological information at the node level; and others, such as the graphlet correlation matrix (GCM) and the graphlet correlation distance (GCD), which provide valuable topological information at the entire network level. We discuss these three widely used graphlet-based measures in the next sections of the the thesis.

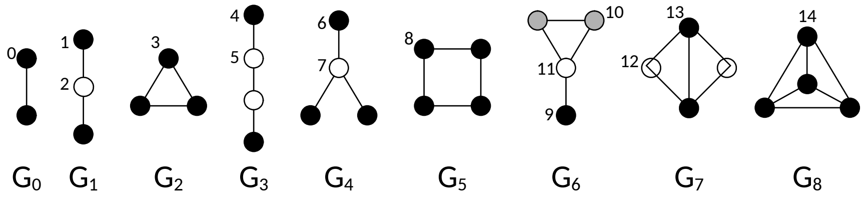


Figure 2.1. An illustration of graphlets. All nine graphlets with up to four nodes (G_0-G_8). Nodes of different shades correspond to the different orbits within each graphlet. Figure adapted from (Przulj et al., 2004).

2.3.5 Graphlet degree vector

The graphlet degree vector (GDV) is a measure that quantifies the local topology of a node in a network. The GDV of a particular node is the vector of the number of occurrences of each possible graphlet that is centered on that node (Milenkoviæ & Pržulj, 2008) (i.e. the number of times the node touches each particular graphlet). Thus, it provides a compact summary of the local network structure around each node, and can be used to compare the local topology of different nodes in a network. For instance, by comparing the GDVs of different nodes, it is possible to identify nodes that have similar local topology and to group nodes into functional modules or communities. The GDV approach has been used in several studies to analyze the topology of biological networks, such as protein-protein interaction networks (Sarajlić et al., 2013; Sarajlić et al., 2016) and metabolic networks, and also generalized to edge-based GDV instead of node GDV (Solava et al., 2012). It has been shown to be effective in detecting meaningful structural modules and in identifying key nodes in the network.

Let $G = (V, E)$ be a graph, where V is the set of vertices and E is the set of edges. Let $g_v = (g_{v,1}, g_{v,2}, ..., g_{v,k})$ be the Graphlet Degree Vector (GDV) of node v, where $g_{v,i}$ is the number of occurrences of the i-th type of graphlet in the neighborhood of node v. Thus, the GDV of a node provides a representation of the local neighborhood structure of around the node.

The calculation of GDVs typically requires the use of graphlet counting algorithms, which can be computationally expensive for large networks. However, there are several algorithms and methods available to reduce the computational complexity and make GDV analysis feasible for large networks. For calculating the GDVs of the nodes in a network, graphlet orbits, illustrated in Fig 2.1, are used to reduce the number of graphlets that need to be considered for the computation (Hočevar & Demšar, 2014).

2.3.6 Graphlet correlation matrix

The graphlet correlation matrix (GCM) is an 11 × 11 matrix that summarizes the network topology with the Spearman's correlations between eleven non-redundant graphlet orbit counts over all nodes in the network (Yaveroğlu et al., 2014). The GCM_{uv}, i.e., the (u, v)th element of the GCM matrix can be mathematically formulated as follows:

Let $GDV_u = (g_{u,0}, g_{u,1}, ..., g_{u,k})$ be the Graphlet Degree Vector (GDV) of node u, where $g_{u,i}$ is the number of occurrences of the i-th type of graphlet orbit in the neighborhood of vertex u. Then the $GCM_{u,v}$ is defined as:

$$GCM_{u,v} = \frac{\sum_{k=1}^{k} g_{u,k} g_{v,k}}{\sqrt{\sum_{k=1}^{k} (g_{u,k})^2} \sqrt{\sum_{k=1}^{k} (g_{v,k})^2}}$$

where $GCM_{u,v}$ is the (u, v)th element of the GCM, representing the similarity between vertices u and v.

2.3.7 Graphlet correlation distance

The graphlet correlation distance (GCD) is a measure of the similarity between two networks, based on the GCM (i.e., 11 x 11 matrix containing pairwise correlations between the 11 non-redundant orbits over all nodes in a network). Thus, the GCD between two networks G_1 and G_2, is defined as the Euclidean distance between both GCMs. This measure is used to quantify the global structural differences between two networks. The smaller the GCD between any two networks, the more similar the two networks are in terms of their graphlet profiles. The GCD is formally defined as:

$$GCD(G_1, G_2) = \|GCM(G_1) - GCM(G_2)\|_2 \tag{2.1}$$

where G_1 and G_2 are two networks, GDM(G) is the graphlet degree vector of graph G_1 and G_2, and $||.||_2$ is the Euclidean norm. Note that this is done on the upper triangle of the GCM matrices, as the diagonal is always 1, which indicates a perfect correlation between the entries.

2.4 Model networks

In this thesis we describe the structure of real world molecular interaction networks by performing model network fitting experiments to compare the real networks to different types of random model networks commonly used in network biology. In this section we define the eight random model networks used for this experiment.

2.4.1 ER

The Erdos–Renyi (ER) random graph model consists of a fixed set of nodes and a fixed set of links (i.e., edges) that are equally likely to exist (i.e., all interactions have the same probability) (Erdös & Rényi, 1959). To generate the ER networks, we set the number of nodes and edge density to match those of the real networks, and by randomly adding edges between uniformly chosen pairs of nodes (out of the $n(n-1)/2$ possible pairs of nodes) until a given density is reached.

2.4.2 ERDD

The ER-DD is the Generalized random graph model and an extension of the ER model. In the ER-DD, the node degree distribution matches that of an input data (i.e., a real network) (M. E. J. Newman, 2010). To generate ER-DD networks, we assign connection capacities (stubs, corresponding to the degree of a node) to the nodes of the network, and then add 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 number of nodes and the degree distributions in these model networks match those of the data networks

2.4.3 GEO

The geometric random graph model (GEO) consists of randomly placing points (i.e., nodes) in a k-dimensional space and connecting them by a link if the distance between them is below a certain threshold (Penrose, 2003). We generate GEO networks by distributing the set of nodes in three-dimensional space and connecting them by edges if the Euclidean distances between them are lower than or equal to threshold r. This value is set so that we obtain a given edge density. The number of nodes and edge density are set to match those of the real networks.

2.4.4 GEOGD

The GEOGD model is the GEO model with gene duplication, where the dispersion of nodes is no longer uniformly random, but according to duplication and divergence rules which mimics the gene duplication and mutation process in biology (Przulj et al., 2009). To generate a GEO-GD model network, we start from a seed network (i.e., two nodes connected by an edge) to which the duplication and mutation process is applied. First, a parent node is chosen at random and duplicated, and then the child node is randomly placed at a distance smaller than or equal to 2r (r is the same as in the GEO model). This process repeats itself until the required number of nodes matches that of the input data. The last step creates the edges with the same rules as in the GEO model until the edge density matches the input data.

2.4.5 NPSO

The Nonuniform Popularity-Similarity Optimization (nPSO) model simulates how random geometric graphs grow in the hyperbolic space with modular organization (also termed communities) (Muscoloni & Cannistraci, 2018). It is an extension of the PSO model, where the similarity between nodes is represented by the hyperbolic distance between them (i.e., the closer two nodes are in space according to the angular coordinates, the more likely they are to be connected by an edge). Similarly, the popularity of the nodes is represented by the radial coordinate in the hyperbolic plane, where nodes with larger degree are positioned closer to the center of the circle. To generate nPSO model networks, we set the number of nodes and number of communities to match those of the input data.

2.4.6 SF

The Barabàsi–Albert scale-free model (SF) is based on the preferential attachment principle and it is characterized by having a scale-free degree distribution (Barabási & Albert, 1999). To generate a SF network, we start from a seed network (i.e., two nodes connected by an edge), and nodes are subsequently added and attached to existing nodes of the network with a probability proportional to their node degrees. This is repeated until the desired number of nodes is reached.

2.4.7 SFGD

The SFGD is the scale-free model with gene duplication and divergence. Similar to the GEO-GD model, the SF-GD mimics the gene duplication and divergence processes in biology (Vazquez et al., 2001). The initial process is the same as in the SF model, starting with a single edge, which is grown through iterative duplication and divergence events. In brief, for each iteration, a parent node is randomly selected and duplicated into a child node. The newly produced node is connected to all the neighbors of the parent node as well as the parent node with probability p. For the divergence process, a single connection is removed with probability q between all the shared neighbors of the parent node and the newly duplicated node. Parameter q is set to match the edge density of the input data.

2.4.8 STICKY

The stickiness-index based (STICKY) model assigns a higher probability of interaction between two nodes the higher their degrees are (Pržulj & Higham, 2006). To generate a STICKY network, we start with n disconnected nodes and we randomly assign stickiness index values which are proportional to the node degrees of the input data. The probability of connecting two nodes is equal to the product of their stickiness indexes.

2.5 Higher order network representations

2.5.1 Graphlet adjacency matrix

To formally introduce graphlet adjacency, we first define the the adjacency matrix of a graph. Let G be a graph with the set of vertices V and the set of edges E, $G = (V, E)$. Two vertices, u and v, are adjacent (i.e., neighbors) if they are connected by and edge $(u, v) \in E$ in the graph. The adjacency matrix of G is a symmetric n x n matrix, A (where n is the total number of nodes in G) that indicates whether pairs of nodes are adjacent or not in the graph: $A(u, v) = 1$ if $(u, v) \in E$; 0 otherwise. The node degree represents the number of connections of a node, which is also the size of the neighborhood of said node. The degree matrix of G is the diagonal matrix, D_{nxn}, where $D(u, u)$ corresponds to the degree of node u; 0 otherwise (off diagonal elements are equal to 0). Finally, the graphlet adjacency matrix is an extension of the adjacency matrix that captures node connectivity patterns beyond simple direct node connectivity (Windels et al., 2019). It is defined as:

$$A_k(u, v) = \begin{cases} c^k_{uv}/\theta_k & \text{if } u \neq v \\ 0 & \text{otherwise,} \end{cases} \tag{2.2}$$

where c^k_{uv} is equal to the number of times the nodes u and v simultaneously touch graphlet k and θ_k is a scaling constant equal to the number of nodes in graphlet k minus 1. Similar to the adjacency matrix of a graph, the graphlet adjacency matrix represents the relationship information of the set of nodes with respect to graphlet k. The graphlet degree matrix of G for graphlet k is the diagonal matrix D_k, where $D_k(u, u)$ is the graphlet degree k of node u. The graphlet degree matrix contains on the diagonal, for each node u the number of times u touches graphlet k, with all non-diagonal elements being zero.

2.5.2 Graphlet Laplacian matrix

The graphlet Laplacian is a matrix representation of a graph that encodes information about the connectivity and node importance of a graph (i.e., connectivity). First we describe how the traditional graph Laplacian is defined: Let G be a graph with the set of vertices V and the set of edges E, $G = (V, E)$. Then, the Laplacian matrix of G, L is defined as $L = D - A$. Where A represents the adjacency matrix of the graph and D the degree matrix of the graph (formerly defined in section 2.5.1). It represents the global structure of a graph because it captures the adjacency relationship of the nodes as well as their importance in the network. Finally, the graphlet Laplacian matrix of a graph is an extension of the Laplacian matrix, generalized to graphlets. Hence, graphlet Laplacians, also capture the relationship information between the nodes, as well as how many connections they have with respect to a given graphlet (Windels et al., 2019). The graphlet Laplacian of a given graphlet k is defined as $L_k = D_k - A_k$. Where A_k and D_k are the graphlet adjacency and graphlet degree matrices with respect to graphlet k, respectively. The graphlet Laplacian has several important properties, such as being positive semidefinite and having real, non-negative eigenvalues. These properties make the graphlet Laplacian a useful tool for graph analysis and for characterizing graph structures such as communities, centrality, and connectivity.

In practical applications, these graphlet-based matrix representations (i.e., graphlet adjacency and graphlet Laplacian) are usually normalised (e.g., to achieve a more balanced graphlet-based spectral clustering (Windels et al., 2019)). The symmetrically normalised graphlet adjacency matrix for a given graphlet k, $\widetilde{A_k}$, is defined as: $\widetilde{A_k} = D_k^{1/2} A_k D_k^{1/2}$. Analogously, the symmetrically normalised graphlet Laplacian, $\widetilde{\mathcal{L}_k}$ is defined as: $\widetilde{\mathcal{L}_k} = D_k^{1/2} \mathcal{L}_k D_k^{1/2}$.

2.5.3 K-path Laplacians

The k-path Laplacian is a generalization of the graph Laplacian. It is a matrix of a graph that captures the connectivity of the nodes (i.e., the node degrees) but also takes into account paths or hops of up to length k between each pair of nodes (Estrada, 2012; Estrada et al., 2017). The entries of the k-path Laplacian matrix can be defined as:

$$L^{(k)}uv = \begin{cases} -1 & \text{if } d(u,v) = k \\ deg_k(u) & \text{if } u = v \\ 0 & \text{otherwise} \end{cases} \tag{2.3}$$

2.5.4 Vicus

Vicus matrix, V, is an alternative to the graph Laplacian and k-path Laplacian matrices of a graph that captures the local neighborhood structure of the graph based on network label diffusion (B. Wang et al., 2017). This label diffusion can be defined as $P = BQ$, where Q is a nxd matrix that assigns the n nodes of a network G to one of the d possible labels (in the case of labeled nodes), B is an nxn diffusion matrix. Lastly, P is the reconstructed matrix nxd that is 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 we define the Vicus matrix as $\mathcal{L}^V = (I - BT)(I - B)$. Next it was shown that Q can be learned as the eigenvectors of 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 $\mathcal{L}^V$, Vicus is interpreted as a Laplacian matrix.

Lastly, it shares many of the basic properties of the standard Laplacian matrix:

1. It is symmetric and positive semi-definite
2. The smallest eigenvalue is 0 and the corresponding eigenvector is the constant 1
3. It has n non-negative real valued eigenvalues $0 = \lambda_1 \leq \lambda_2 \leq \lambda_3 \leq \lambda_n$.
4. The multiplicity of the smallest eigenvalue (i.e., 0) of the Vicus matrix V equals the number of connected components in the graph.

2.6 Network embedding

Due to the complexity of the data, a growing trend in modern network analysis is to transform the network into a vector-based representation rather than to analyse the network directly, a process referred to as *network embedding* (Cai et al., 2018; Nelson et al., 2019). In brief, these methods extract a low-dimensional representation of the data while conserving the original similarity features in the data (Arsov & Mirceva, 2019). In this way, nodes that are in the same network neighbourhood have a similar vectorial representation. Often, this is interpreted as learning a low-dimensional embedding space, in which nodes in similar network neighbourhoods are embedded nearby in the space (i.e., have a similar position). What distinguishes different embedding algorithms, then, is the notion of what it means for two nodes to be in each others' network neighbourhood. For instance, *spectral embedding* learns an embedding space so that nodes that cluster in the network, i.e., that tend to share neighbours, are embedded nearby in space (Belkin & Niyogi, 2003). *Spring embedding*, on the other hand, imagines that all edges in the network are springs and places the nodes in a (euclidean) space so that the forces exerted by the springs on the nodes are in equilibrium (Kamada, Kawai, et al., 1989). Formally, a network embedding (also called representation learning) learns a vectorial representation of each element in the network (i.e., nodes) that captures the structure and semantics of the network (M. M. Li et al., 2022). Given a network $G(V, E)$, with V nodes and E edges, and its corresponding adjacency matrix, the goal is to learn a function $V \rightarrow R^d$ that maps each node to a $d-$dimensional ($d < |V|$) vector that captures its structural properties (M. M. Li et al., 2022).

To date, several embedding algorithms have been proposed, facilitating tasks such as classification, clustering, prediction and visualization across various fields including, biology, economy and social sciences (Cai et al., 2018; Chen et al., 2018; Grover & Leskovec, 2016; Gutiérrez-Gómez & Delvenne, 2019; Kulmanov et al., 2018; G. Li et al., 2017; Perozzi et al., 2014; Zong et al., 2017). Classical embedding approaches include Principal Component Analysis (PCA) and Multi-Dimensional Scaling (MDS). PCA is used to analyze the structure of a data matrix via dimensionality reduction that preserves most of the variance in the original data by performing eigendecomposition of the covariance matrix. The resulting eigenvectors, or "principal components" can be more easily visualized and analyzed, for instance in a 2-dimensional (2D) plot. Similarly, MDS is used to analyze the similarity of a set of objects by performing dimensionality reduction on a distance matrix of the objects from the set and attempting to map the objects in a geometric space by conserving the original distances between them (Borg & Groenen, 1997).

Force-directed algorithm, or Spring embedding, is a widely used method for visualization of complex networks. It works as a physical system where nodes act as charged particles that repel each other and edges as springs that keep everything together (Kobourov, 2012). In this way, between each pair of disconnected nodes there are repulsive forces inversely proportional to the distance between them, while for connected nodes, there are attractive forces. Other state-of-the-art embedding algorithms include: DeepWalk, which uses random walks to uncover the latent representation of a network by treating the walks as if they were sentences, and was originally used for classification of social networks (Perozzi et al., 2014); and Node2vec, which also uses random walks to explore the neighborhood of each node in a network and maximizes the likelihood to preserve these neighborhoods (Grover & Leskovec, 2016).

Other embeddings such as the Coalescent embedding map a network in a hyperbolic space by inferring the angular coordinates of the hyperbolic model and assigning a radius to each node (Muscoloni et al., 2017). In this method, the angular distances between nodes represent the similarity between them, while the radius represents how densely connected they are in the network (nodes of smaller radius, i.e., those that are more central in the embedding, have more connections in the network than the peripheral ones). One of the main findings in this study is that Coalescent embedding can significantly improve the community detection in complex networks (Muscoloni et al., 2017). A community is defined as a group of nodes or a region in a network that has densely connected nodes that are sparsely connected with the rest of the network. Community detection can provide additional biological insight, for instance to uncover functional molecular modules in biological networks (Yang et al., 2016).

In the following sections we explain in more detail the three main embedding methods used for the development of this work: Spring embedding, Coalescent embedding and Spectral embedding.

2.6.1 Spring embedding

Spring embedding is a type of force-directed layout algorithms, which model an input graph as a system of attractive and repulsive forces. Spring embedding imagines that all edges in the network are springs and places the nodes in a space so that the forces exerted by the springs on the nodes are in equilibrium (Kamada, Kawai, et al., 1989). The goal is to minimize a cost function that represents the total energy of the system, which depends on the length of edges and the distance between nodes. The resulting layout aims to provide a clear, readable representation of the graph structure, by spreading nodes out evenly, minimizing edge crossings and highlighting dense clusters or communities in the graph. Some limitations of Spring embedding is that it relies heavily on node connectivity and it is non-deterministic. When applied to biological networks, spring embedding is likely to produce uninformative, close-knit network embeddings resembling "hairball" (Bläsius et al., 2021). This is because many biological networks, including PPI (Jeong et al., 2001a) and GI networks (Tong et al., 2004), are *scale-free* (defined in section 2.3.3.1).In scale-free networks, the few high-degree nodes (i.e., nodes with many neighbours, known as 'hubs') connect to many low-degree nodes (Ravasz & Barabási, 2003). Spring embedding does not manage to spread the hub-nodes in the (Euclidean) embedding space, as they are pulled together by the many 'springs' connecting them to their shared low-degree neighbors (Bläsius et al., 2021).

2.6.2 Coalescent embedding

Due to the complexity of many real world data, a proper embedding representation is crucial for uncovering the latent geometry of complex networks, such as the Euclidean space, which has dominated many areas in science for data representation and visualization (Clauset et al., 2009; Cross et al., 2006; M. E. Newman, 2004; Zachary, 1977). More recently, the hyperbolic space has become a highly relevant space for network embedding representation (Cannistraci & Muscoloni, 2018; Muscoloni et al., 2017; Watts & Strogatz, 1998). In the hyperbolic space the data or network is usually represented in a hyperbolic disk where each data point or node in the network is assigned a radial coordinate and an angular coordinate. The radial coordinate of the nodes characterizes their hierarchy in the network, whereas the angular distance in the disk between the nodes represent their similarity in the network. In this way, the nodes of high topological importance (for instance highly connected nodes) are usually placed towards the center of the disk, and less important nodes (with fewer connections) are placed towards the periphery of such disk. In addition, the hyperbolic space has been shown to have more capacity than the Euclidean space, as its volume grows exponentially with the radius. Finally, hyperbolic geometry is better suited to embed data with tree-likeness or underlying hierarchical/heterogeneous structure (Adamic & Glance, 2005).

To date, several studies that make use of hyperbolic space have been proposed for representing the latent geometry behind complex network topologies (Alanis-Lobato et al., 2016a, 2016b; Bianconi & Rahmede, 2017; García-Pérez et al., 2019; Krioukov et al., 2010; Song & Wang, 2019). In this work we chose on focusing on Coalescent embedding (Muscoloni et al., 2017), which we generalise to graphlet based Coalescent (GraCoal) embedding (discussed in Chapter 3). In brief, Coalescent embedding maps a network onto a disk, assigning an angle and a radius to each node rather than a Euclidean coordinate. One of our motivations for focusing on hyperbolic embeddings, such as Coalescent embedding, is that the scale-freeness of many biological networks stems from an underlying latent hyperbolic geometry, which hyperbolic embeddings can uncover (Boguna et al., 2009). (Boguna et al., 2009). Additionally, Coalescent embedding (CE) in particular, was shown to detect successfully communities in many real world networks (Muscoloni et al., 2017).

Given a graph as input, the CE algorithm can be summarized as follows:

1. Perform dimensionality reduction on the graph, which can be done by using one of the following approaches: Minimum curvilinearity (MCE) (Cannistraci et al., 2013; Cannistraci et al., 2010); Isomap (ISO) (Tenenbaum et al., 2000); Noncentered minimum curvilinearity (ncMCE) (Cannistraci et al., 2013); Noncentered Isomap (ncISO) (Cannistraci et al., 2013); Laplacian eigenmaps (LE) (Belkin & Niyogi, 2001, 2003). For MCE, the first dimension is used; for ISO, the 1st and 2nd dimensions are used; for ncMCE, the 2nd dimension is used; finally, for ncISO and LE, the 2nd and 3rd dimensions are used.

2. Determine angular coordinates: The vectors obtained from the dimensionality

reduction are used as coordinates that correspond to every node in the network, which are transformed into angular coordinates. The vectors corresponding to the 1st and 2nd dimensions for ISO, and 2nd and 3rd dimensions for ncISO and LE algorithms (and which are commonly used as Cartesian coordinates), are converted to polar coordinates. For MCE and mcMCE, Cartesian to polar conversion is not applied because the vectors obtained by these dimensionality reduction techniques are already given as angles. Next, the angular vector obtained previously by either method is used for circular adjustment (CA) and equidistant adjustment (EA). First, CA sets the angular coordinates vector in the range (0, 2π), followed by EA, which reorganizes the coordinates equidistantly along the circle according to their original order learned by the dimensionality reduction.

3. Assign radial coordinates: For computing the radius, Coalescent embedding explicitly assumes that the degree distribution follows a power a power law: $P(d) \sim d^{\lambda}$. So first, coalescent embedding fits a power-law to the degree distribution (i.e., estimates λ). Then, the nodes are sorted in descending order according to their degree. Finally, the radial coordinate of the i$^{\text{th}}$ node, r_i, is calculated as:

$$r_i = \beta ln(i) + (1-\beta)ln(N), \tag{2.4}$$

where i is the rank of the node, N the number of nodes in the network and $\beta = 1/(\lambda - 1)$.

2.6.3 Graphlet Spectral embedding

Spectral embedding learns an embedding space such that nodes that share many neighbors in the network, are embedded close in space (Ng et al., 2001). It makes use of the Laplacian matrix of a graph to perform dimensionality reduction and subsequently use the vector coordinates corresponding to the second and third smallest eigenvalues to embed the graph in 2 dimensions (2D). More recently, the Laplacian matrix was generalized to the graphlet Laplacian and Spectral embedding was applied to capture functional information from the underlying networks, showing that different graphlets can uncover different biological functions (Windels et al., 2019).

Thus, here we recall our formal definition of graphlet Spectral embedding, which embeds nodes nearby in space if they frequently simultaneously touch a given graphlet (Windels et al., 2019). Formally, given an unweighted network H with n nodes, we find a low dimensional embedding, $Y = [y1, ..., yn] \ \epsilon \ R^{dxn}$ such that if nodes u and v are frequently graphlet-adjacent with respect to graphlet G$_k$, then y(u) and y(v) are close in the d-dimensional space by solving:

$$\begin{aligned} &\underset{Y}{\text{minimize}} && \sum_{u=1}^{n}\sum_{v=1}^{n} A_{G_k}(u,v)\mathbf{y}_u - \mathbf{y}_v{}^2 \\ &\text{subject to :} && YD_k\mathbf{1} = \mathbf{0} \text{ and } YD_kY^T = I, \end{aligned} \tag{2.5}$$

where A_{Gk} is the graphlet-based adjacency matrix of G for graphlet G_k, D_k is the graphlet-based degree matrix of G for graphlet G_k. The columns of Y are found as the generalized eigenvectors associated with the 2nd to $(d+1)^{th}$ smallest generalized eigenvalues solving $YL_{Gk} = YD_k$, where is the diagonal matrix with the generalized eigenvalues along its diagonal.

2.7 SAFE

Despite the abundance of biological network data, our knowledge of the functional organization of these networks remains incomplete. For instance, for the model organisms *C. elegans* and *S. pombe*, we find that the experimentally validated biological-process annotations in the Gene Ontology (GO) database cover only 14% and 15% of their genes (Ashburner et al., 2000). This illustrates the need for automated functional annotation algorithms for biological networks. A state-of-the-art network-based functional annotation algorithm is SAFE: Spatial Analysis of Functional Enrichment (Baryshnikova, 2016). Given a biological network and a set of node annotations, SAFE uncovers local network neighbourhoods where node annotations are over-represented or *enriched*. The annotations enriched in the same network neighbourhood are automatically aggregated into computationally generated *domains*, describing the general function of different local network neighbourhoods. SAFE provides an intuitive visualisation of the domains by placing the network in a 2D plane using Spring embedding and overlaying the network embedding with the different uncovered functional domains. This way, SAFE effectively creates a functional map of the cell that is powerful yet intuitive to biologists, enabling the study of the functional organization of the network at hand. As such, SAFE is widely used to study biological networks. Originally, SAFE was introduced to study the functional organization of the yeast GI network, and uncovered that cellular function is organised in hierarchical functional modules (Costanzo et al., 2016). Ho et al. applied SAFE on PPI data to show how protein abundance levels in the cell are dependent on the cellular functions they are involved in (B. Ho et al., 2018). For instance, high-abundance proteins were specifically over- represented biological processes related to morphogenesis and ribosome biogenesis, while low-abundance proteins were associated with DNA replication and repair, mitosis, and RNA processing. To validate their human proximity-dependent biotinylation PPI data, which detects PPIs in intact cells, Youn et al. applied SAFE and manage to recover the spatial organization of the human cell into cellular compartments (Youn et al., 2018).

The SAFE framework consists of 4 algorithmic steps.

In step 1, SAFE takes as an input an unweighted and undirected network H, and a set of node annotations of interest M to produce a 2D embedding E of this network.

In step 2, the local neighbourhood of each node is determined. SAFE does so taking both information from the embedding space into account, as well as information directly from the network. First, SAFE computes the pairwise shortest path distance between all nodes in the network. To take into account information from the embedding space, each edge between a pair of nodes in the network is weighted by their Euclidean distance in the embedding space. Then, SAFE considers the *local neighbourhood* of each node to be all nodes that are at a weighted shortest path distance (WSPD) less than a given threshold α.

In step 3, SAFE computes for the local neighbourhood of each individual node, the node annotations that occur more than expected by chance using a hypergeometric test, applying the Benjamini and Hochberg correction for multiple hypothesis testing (Benjamini & Hochberg, 1995).

In step 4, the annotations that are enriched in overlapping local neighbourhoods are aggregated into more descriptive groups. To do so, first, the attributes that are enriched in fewer than β local neighbourhoods are discarded (default: β=10). Then, agglomerative clustering with average linkage is applied on the remaining attributes, based on their Jaccard similarity in terms of the local neighbourhoods in which they are enriched. From this hierarchical clustering, clusters of annotations are extracted, cutting the tree at γ% of its height (default: γ=75%). The resulting clusters of annotations are referred to as *functional domains*. For each functional domain, the five most repeated words occurring in the annotations names are reported an aggregated description for the domain.

Chapter 3

New methods: embedding omics networks for new biological insights

In this chapter we first explain how we extend the Spring embedding algorithm and generalise it to graphlet-based Spring embedding. Next we explain how we extend the Coalescent embedding algorithm and generalise it to graphlet-based Coalescent (GraCoal) embedding. Finally, we briefly go over the modifications done to extend the SAFE pipeline to consider these newly graphlet-based embeddings, as well as the already established graphled-based Spectral embedding.

3.1 Graphlet based Spring embedding

We recall from section 2.6.1 how Spring embedding acts as a system of attractive and repulsive forces where it imagines that all edges in a network work as springs that attract connecting nodes to each other until the system reaches equilibrium (Kamada, Kawai, et al., 1989).

In particular, we focus on the Fruchterman-Reingold force-directed algorithm (Fruchterman & Reingold, 1991), which is the one that the SAFE framework uses. In this particular force-directed algorithm, all nodes in the network are assumed to repel each other by a repulsive force, while an attractive force pulls together pairs of nodes that are connected in the network. These repulsive and attractive forces between nodes and edges can be formally defined as follows:

Repulsive force Fr between two nodes u and v:

$$\mathbf{Fr}_{uv} = -\frac{k^2}{|\mathbf{x}_u - \mathbf{x}_v|}$$

Attractive force Fa between two nodes u and v connected by an edge:

$$\mathbf{Fa}_{uv} = \frac{|\mathbf{x}_u - \mathbf{x}_v|^2}{k}$$

Where k is a global hyperparameter that determines the strength of the repulsive and attractive forces and $|\mathbf{x}_u - \mathbf{x}_v|$ is the distance in embedding space between nodes u and v (Fruchterman & Reingold, 1991).

Having defined the attractive forces determined by the edges, and repulsive forces determined by the nodes, the following steps are repeated in the spring embedding algorithm until reaching equilibrium:

1. Calculate the repulsive forces between all pairs of nodes.
2. Calculate the attractive forces between all pairs of connected nodes.
3. Update the positions of all nodes based on the combined effect of the repulsive and attractive forces.
4. Limit the total displacement of the nodes by the temperature T, which is a global parameter that regulates the step size of node movement.

To reach equilibrium, the goal is to minimize a cost function that represents the total energy of the system, which depends on the length of edges and the distance between nodes. The resulting layout aims to provide a clear, readable representation of the graph structure, by spreading nodes out evenly, minimizing edge crossings and highlighting dense clusters or communities in the graph. The Spring embedding used in SAFE, typically receives as input the adjacency matrix of a given network G, which if is unweighted, sets the length of all edges in the network to 1 by default. We apply the Spring embedding to the graphlet adjacency matrix, formerly defined in section 2.5.4. Specifically, we use the symmetrically normalised graphlet adjacency matrix for a given graphlet k, $\widetilde{A_k}$, which is defined as: $\widetilde{A_k} = D_k^{1/2} A_k D_k^{1/2}$. In this way, because the edge weights between nodes in a particular graphlet adjacency matrix represent how well connected they are with respect to a given graphlet, we can obtain a different layout for each graphlet based adjacency matrix. We do this for all up to 4-node graphlets ($\widetilde{A_0}$ to $\widetilde{A_8}$) when using the Spring embedding as opposed to using the Spring embedding with only the traditional adjacency matrix of the same network. For instance, a pair of nodes that simultaneously touch many times the four node clique (i.e., they have a large in magnitud edge weight) might end up embedded close in space when applying graphlet based Spring embedding to $\widetilde{A_8}$, but not when applying graphletd based Spring embedding to $\widetilde{A_3}$, which is based on the four node path graphlet.

3.2 Graphlet based Coalescent embedding

Coalescent embedding maps a given network onto a hyperbolic circle, by assigning similar angles to nodes that are in the same network neighborhood (i.e., nodes that tend to form clusters in the network). Nodes with higher topological importance (e.g., have a higher degree), are embedded near the circle's centre (Muscoloni et al., 2017). After having formally defined the Coalescent embedding algorithm in section 2.6.2, in this section we present an overview of the newly proposed graphlet-based Coalescent embedding (GraCoal) approach.

1. For a given network and graphlet, we embed the given network into 2D space using graphlet spectral embedding (see Section: Graphlet spectral embedding).
2. We map the node Cartesian coordinates to an angular coordinate. This step is identical to step 2 in coalescent embedding.

3. We determine a radial coordinate for each node applying the following formula:

$$r_i = ln(i), \tag{3.1}$$

where i is the rank of the node based on its graphlet degree.

Note that our formula to determine the radius of a node (equation 3.1) is a simplified version of the equation applied in standard Coalescent embedding (equation 2.4). We do this because the graphlet degree distributions for our real networks do not all follow a power-law. In Figures 3.1-3.4 we show the graphlet degree distributions for the budding yeast GIS, GI, PPI and COEX networks, respectively. Fitting a power-law to graphlet node degree distributions lead to larger than usual values of the power-law exponent λ, which usually ranges between 2 and 3 (Ravasz & Barabási, 2003). This leads to large radial coordinates, pushing the nodes to the periphery of the hyperbolic space, as shown in Figure 3.5.

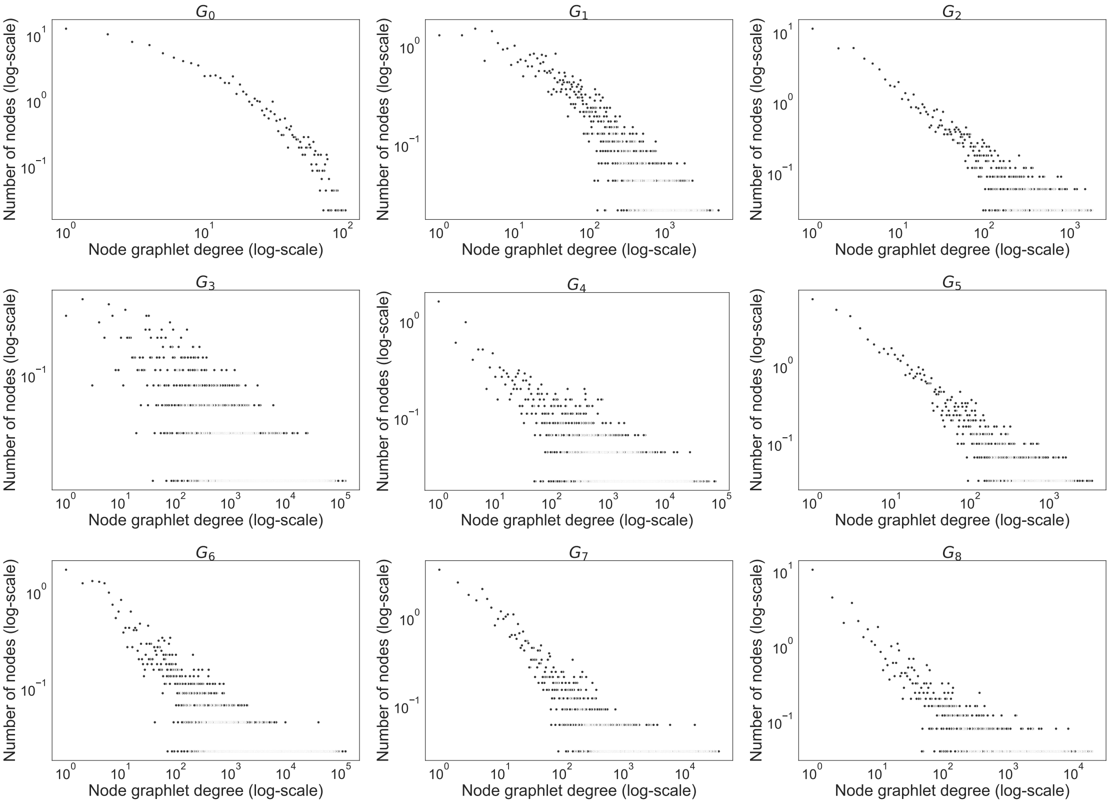


Figure 3.1. Node graphlet degree distributions for all up to 4-node graphlets (G_0-G_8) for the Budding yeast genetic interaction similarity (GIS) network. Graphlet G_0 is the only graphlet that appears to have a scale-free node graphlet degree distribution.

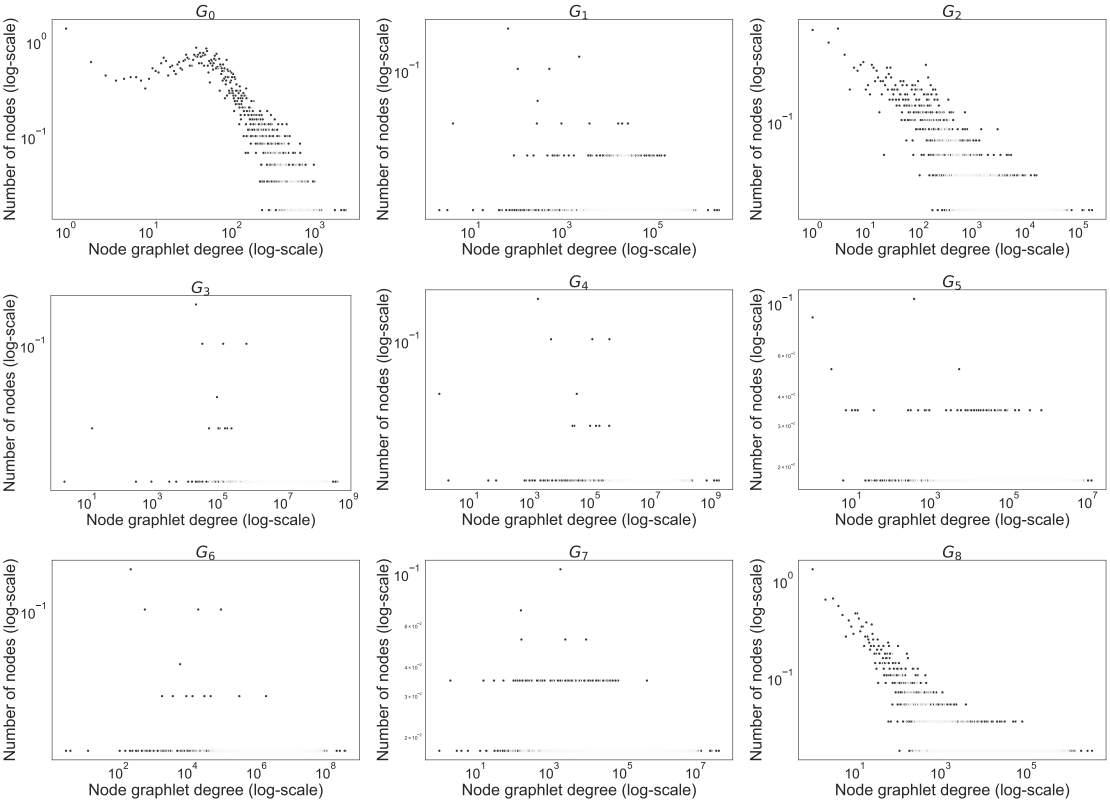


Figure 3.2. Node graphlet degree distributions for all up to 4-node graphlets (G_0-G_8) for the Budding yeast genetic interaction similarity (GI) network. None of the graphlets appear to have a scale-free node graphlet degree distribution.

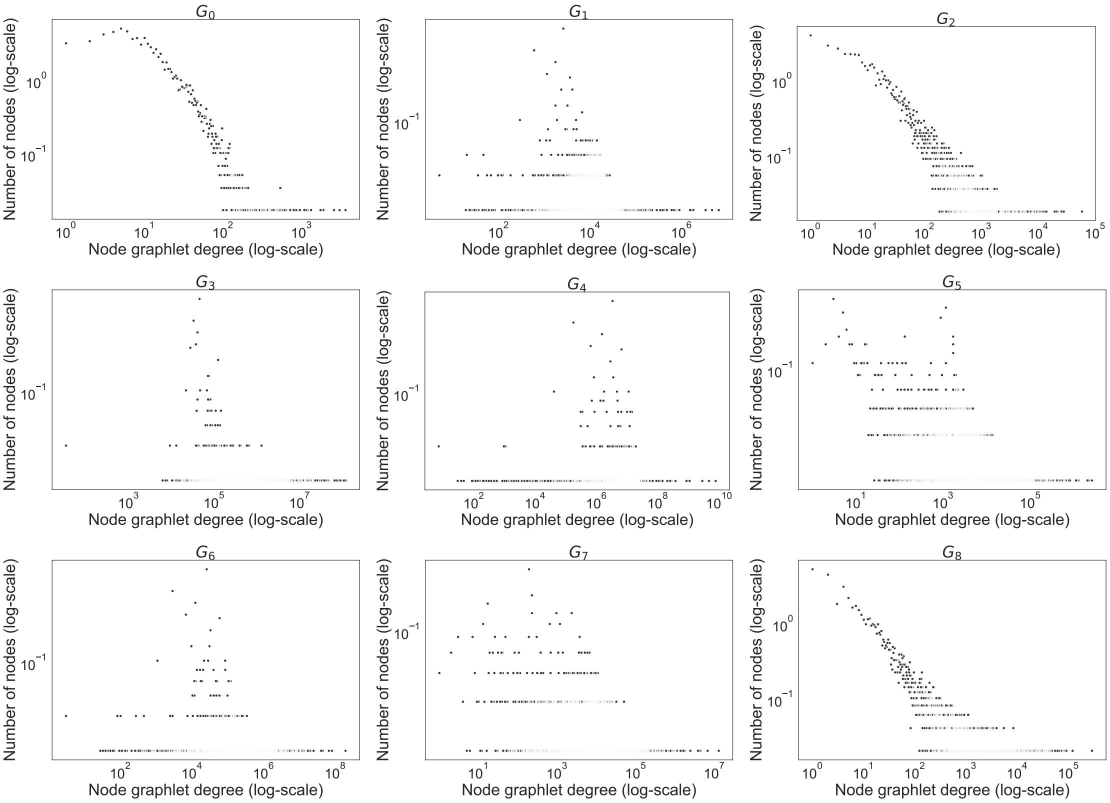


Figure 3.3. Node graphlet degree distributions for all up to 4-node graphlets (G_0-G_8) for the Budding yeast protein-protein interaction (PPI) network. None of the graphlets appear to have a scale-free node graphlet degree distribution.

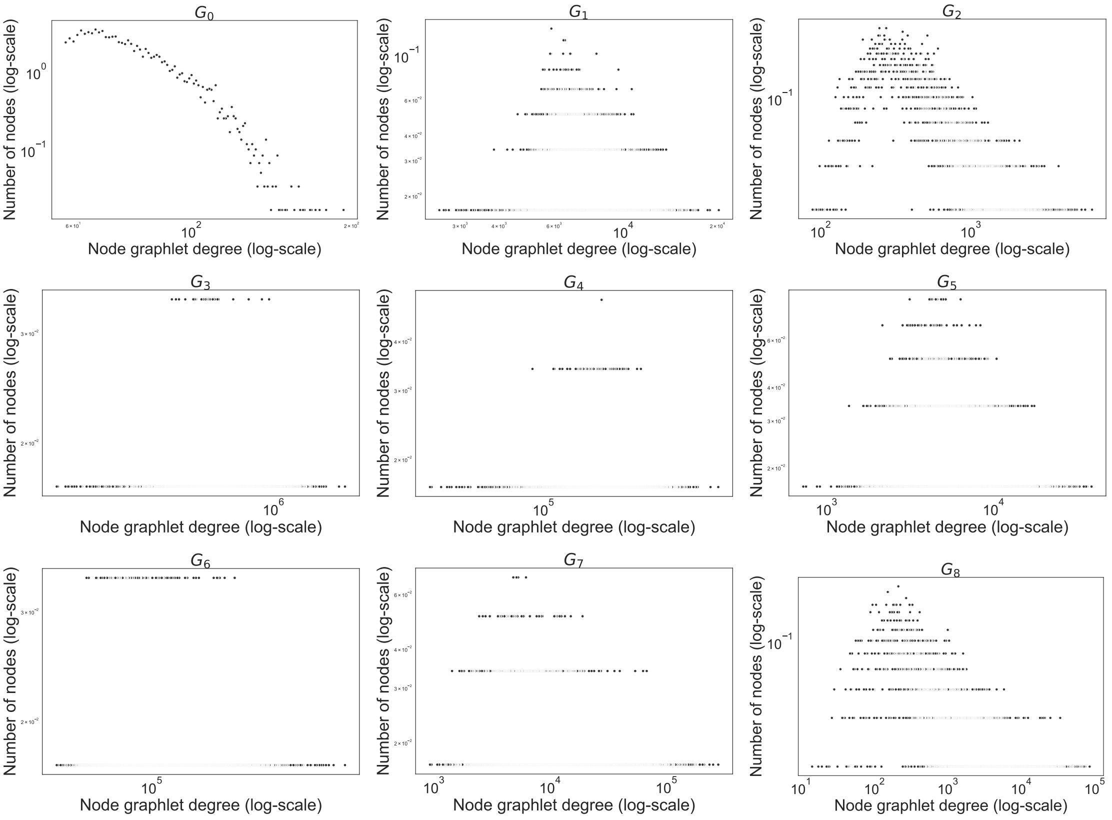


Figure 3.4. Node graphlet degree distributions for all up to 4-node graphlets (G_0-G_8) for the yeast co-expression (COEX) network. Graphlet G_0 is the only graphlet that appears to have a scale-free degree distribution.

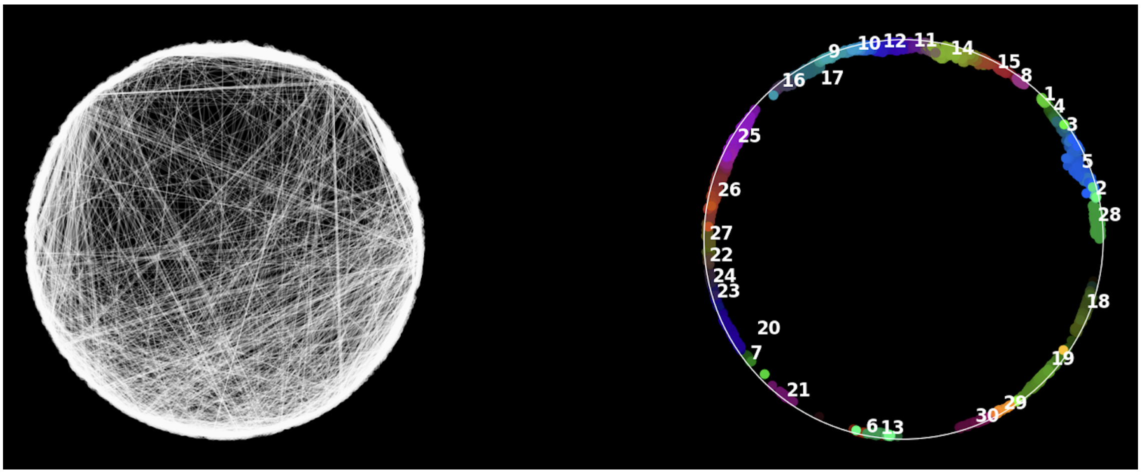


Figure 3.5. When using the Coalescent embedding, large values of λ lead to large radial coordinates, placing the nodes towards the periphery of the embedding space. We show the Coalescent embedding (left) and the corresponding enrichment landscape (right), visualized with SAFE when using Equation (3) (main document) with a large value of λ. The node importance in terms of graphlet degree (i.e., how well connected with respect to a given graphlet) is lost when applying this algorithm directly on graphlets.

3.3 Extension of SAFE

The SAFE framework takes as input a network and a set of annotations of the given network (Baryshnikova, 2016). Typically, a network is provided in edgelist format, which is, a two column file for unweighted networks where each row represents an interaction between a particular node (first column) and another node (second column). For weighted networks, a third column may be present, which contains the interaction magnitude of the two nodes. For instance, a GI network may be represented by a three column edgelist, where for each row, the first column contains the label of a particular gene, the second column the label of a gene that genetically interacts with the gene in the first column, and the third row the genetic interaction score of this interaction. We extended this step in the SAFE framework to optionally include a third input file: a graphlet adjacency matrix, in edgelist format. Similar to the edgelist previously described, the graphlet adjacency edgelist contains, for each row a gene in the first column, a different gene in the second column and in the third column, instead of an interaction score such as in a GI network, the number of times the two nodes touch a particular graphlet, previously symmetrically normalised.

We recall from section 2.7 that SAFE framework consists of 4 main algorithmic steps. Below we review the steps that we made modifications to in order to further extend and provide more functionality to the framework.

In step 1, a given network is embedded in a 2D space by applying the Spring embedding algorithm. For this step, we extended its functionality so the user can choose whether to use the default Spring on a the input network, or if a particular graphlet adjacency matrix (in edgelist format) is available, specify between 'GraSpring', 'GraCoal' and 'Spectral'.

In step 2, the local neighborhood of each node is determined. SAFE does so taking both information from the embedding space into account, as well as information directly from the network. First, SAFE computes the pairwise shortest path distance between all nodes in the network. To take into account information from the embedding space, each edge between a pair of nodes in the network is weighted by their Euclidean distance in the embedding space. Then, SAFE considers the *local neighbourhood* of each node to be all nodes that are at a weighted shortest path distance (WSPD) less than a given threshold α. To facilitate the downstream analysis of graphlet-based embeddings, we modified the way the local neighborhood of a node is computed. In brief, we fix the average neighborhood size to a user specified parameter, NS (neighborhood size) to avoid large discrepancies in average neighborhood sizes when using different graphlet-based embeddings. In this regard, before evaluating the performance of the different graphlet based embeddings, we choose an optimal NS based on the enrichment results obtained with SAFE and fix this value to allow for a comparison across our methods. We run SAFE with different values of this new user specified hyperparameter with the three embedding algorithms and compare the percentages of genes enriched in at least one annotation and percentages of annotations enriched with respect to different neighborhood sizes. In brief, we discover that setting the NS to values above 50 provides no additional enrichment results in terms of annotations when using SAFE with the different graphlet-based embeddings (Fig 3.6).

Lastly, in addition to the three output files produced by the SAFE framework, we added further relevant data as output files. This is particularly useful for when having to run the framework over multiple networks and/or graphlet adjacencies. SAFE now stores all node embedding coordinates for each run as a two column file where each row represents a node and the two columns represent the X and Y Cartesian coordinates. An output file named *attribute2enrichedgenes* containing, for each enriched annotation, a list of gene indices for which the annotation is enriched in the neighborhood. All the WSPD are also saved to an output file. Finally, we also save the plot corresponding to the embedding, as opposed to only the plot with the functional domains.

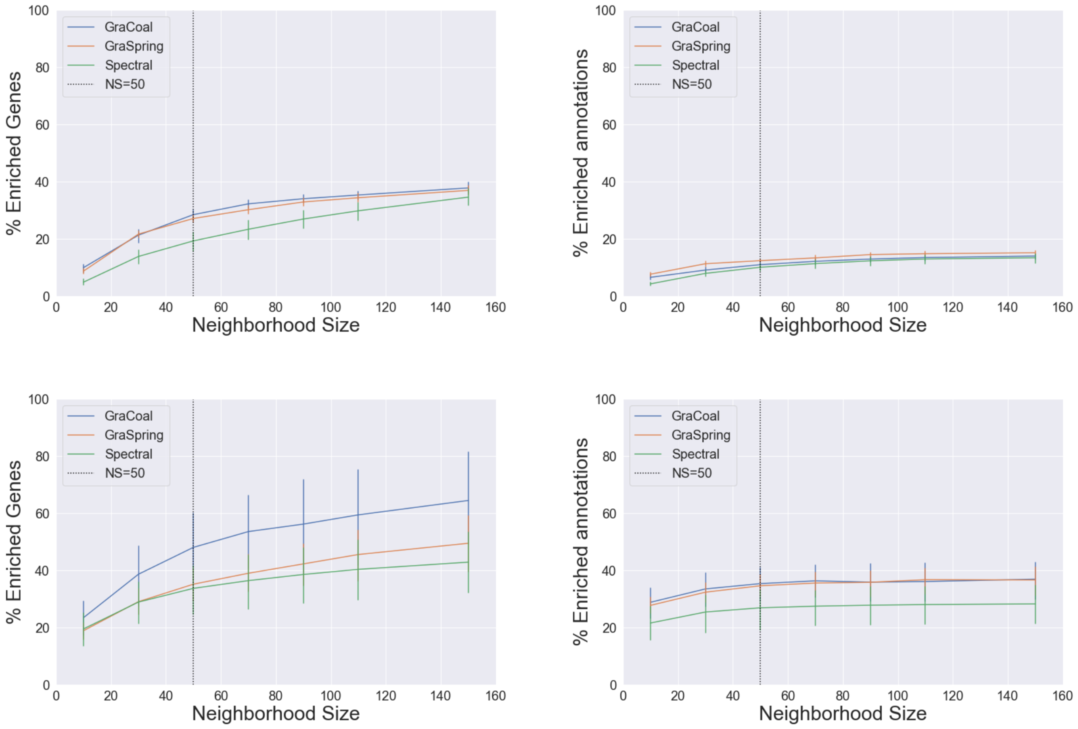


Figure 3.6. SAFE enrichment statistics with respect to neighborhood size, Part 1. We show the percentages of genes enriched in at least one GO-BP (left) and percentages of enriched GO-BP (right) for different neighborhood sizes used in SAFE (x-axis) for the GI network of *E. coli* (top) and Fruit fly (bottom).

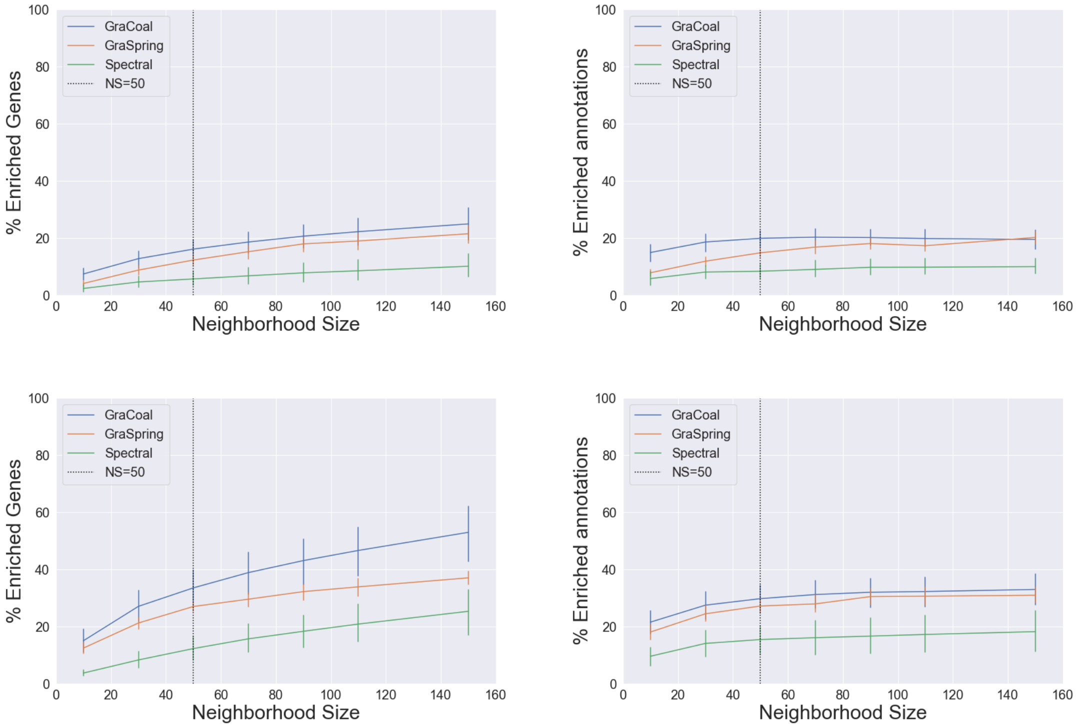


Figure 3.6. SAFE enrichment statistics with respect to neighborhood size, Part 2. We show the percentages of genes enriched in at least one GO-BP (left) and percentages of enriched GO-BP (right) for different neighborhood sizes used in SAFE (x-axis) for the GI network of Fission yeast (top) and Budding yeast (bottom).

Chapter 4

Application 1: Analysis of GI networks

In this chapter we evaluate the performance of the graphlet-based embeddings (i.e., GraCoal, GraSpring and graphlet based Spectral) with the Spatial Analysis of Functional Enrichment (SAFE) framework on the GI networks of the following species: *Drosophila melanogaster*, *Escherichia coli*, *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*, which throughout the text we will refer to as fruit fly, *E. coli*, budding yeast and fission yeast, respectively. We present the GI network statistics in Table 4.1. For more information on how we built these molecular networks please refer to section A.1 in Appendix A. Moreover, we focus mainly on analysing results based on Gene Ontology Biological Processes (GO-BP), as is one of the most complete set of annotations. For detailed results corresponding to our other annotations (e.g., GO molecular functions and GO cellular components), please refer to section A.2 in Appendix.

		GI	
	Nodes	**Edges**	**Density**
Budding yeast	5,842	447,747	0.03
E. coli	3,973	169,594	0.02
Fission yeast	3,577	52,402	0.008
Fruit fly	3,159	10,687	0.002

Table 4.1. GI molecular network data statistics. For each species (row), we report the number of nodes, the number of edges and the density of the corresponding GI network (columns 1-3).

In general, we find that GraCoal embeddings outperform both GraSpring embeddings and graphlet based Spectral embeddings on every GI network. Additionally, some GraCoals lead to better enrichments than others, and thus, we perform a detailed investigation of the topology-function relationship captured by the different GraCoal embeddings. When providing specific examples, we focus mainly on the budding yeast GI network, as it is the most complete and best annotated GI network. Moreover, we choose to study GI networks in this section over GIS networks, as the latter is only available for budding yeast, which would excessively limit the scope of our study in terms of species coverage.

4.1 Results for GraCoal with SAFE

GraCoal best uncovers the functional organization of the cell

In this section, we evaluate: (1) how well GraCoal embeddings capture the functional organization of genetic interaction networks and (2) which higher-order topologies (i.e., graphlets) capture the most function. In both experiments, we apply SAFE-based enrichment analysis to quantify how well a given embedding captures the functional organization of a given network. An annotation (e.g., GO-BP) is enriched if it is over-represented in the local neighborhood of at least one gene. Similarly, a gene is enriched if it has at least one annotation enriched in its local neighborhood. As our conclusions are the same for both experiments, whether we focus on gene enrichment or annotation enrichment, we focus on gene enrichment.

To evaluate how well GraCoals capture the functional organization of our GI networks, we compare against GraSpring embedding (as GraSpring for graphlet G_0 corresponds to standard Spring embedding, used in the original SAFE) and Graphlet Spectral embedding (as it underlies our GraCoal embeddings). At this stage, we want to evaluate which of these embedding methods is best in general, regardless of the chosen graphlet-based topology. Therefore, we consider for each embedding method the union of the enriched genes across the different types of underlying graphlet adjacencies (i.e, $\widetilde{A_{G_i}}$ to $\widetilde{A_{G_8}}$). We show the results in Figure 4.1. We observe that for all four of our species, GraCoal outperforms both GraSpring and graphlet Spectral embedding. In particular, we find GraCoal captures the functional organization of the fruit fly and budding yeast (% enriched genes 90.3 and 71.4) exceptionally well, greatly outperforming GraSpring embedding (% enriched genes 61.3 and 42.6) and Graphlet Spectral embedding (% enriched genes 76.72 and 61.13). GraCoal also outperforms GraSpring and graphlet Spectral embedding in fruit fly, budding yeast and fission yeast GI networks when we consider the two alternative annotation types that describe the spatial organization of the cell (GO cellular components) and the function of the cell (GO molecular functions) (see Appendix A.2). Lastly, GraCoal embedding also best capture the functional and spatial organization of our GIS network, achieving the best enrichment scores for all three annotation types, both in terms of gene enrichment as well as annotation enrichment (Appendix Figures A.6, A.9 and A.16).

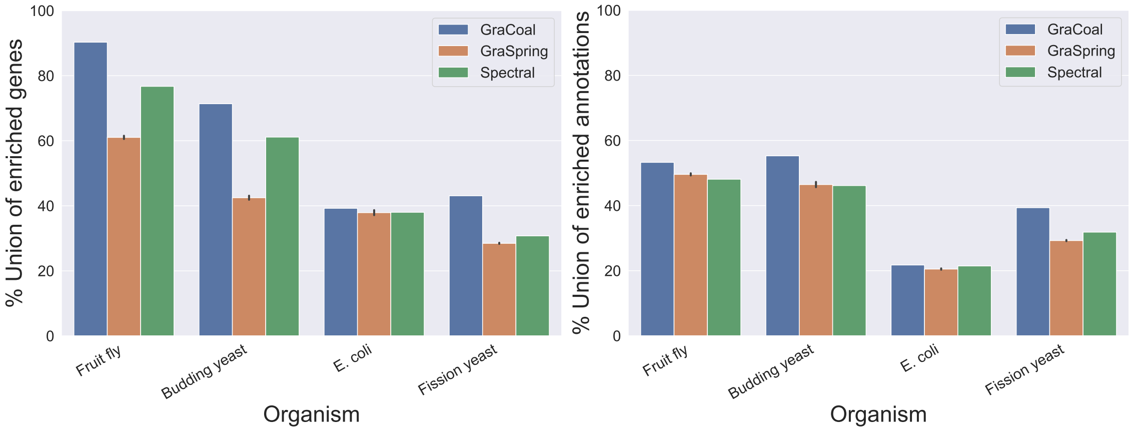


Figure 4.1. SAFE GO-BP enrichment analysis for GI networks. For the GI networks of our four species (x-axis), we show the percentage of enriched genes (y-axis) and percentage of enriched annotations for each of the embedding algorithms considered (legend). In the case of GraSpring, we show the average across ten randomised runs and the standard deviation (error-bars).

GraCoal embeddings spread the nodes more evenly in the embedding space

To explain why GraCoals best capture the functional organization of GI networks, we use SAFE's functionality to visualise the GO-BP enrichment landscapes for our three different types of graphlet based embeddings. In general, we observe that when using GraCoal embeddings, the nodes are spread much more evenly than when using graphlet-based Spring or graphlet-based Spectral embeddings, regardless of the graphlet adjacency used. For Spring embedding, this is expected, as the budding yeast GI network is scale-free, which is known to lead to dense, entangled Spring embeddings (Bläsius et al., 2021). As a consequence, the functional domains based on GraCoal embeddings are also much more spread out and discernible than those based on GraSpring embedding and Spectral embedding.

Below we compare the 2D network embedding layouts (left) and functional landscapes (right) produced by SAFE of the three graphlet based embeddings for the Budding yeast GI network. To have a baseline comparison, we show the embeddings corresponding to the normal graphlet adjacency in Figure 4.2 (i.e., $\widetilde{A}_{G_0}$). Next, to compare the embeddings corresponding to both a densely connected graphlet and a long path, we show, in Figures 4.3 and 4.4, the embeddings based on graphlet adjacency $\widetilde{A}_{G_2}$ (i.e., the three-node clique), and graphlet adjacency $\widetilde{A}_{G_3}$ (i.e., the four-node path), respectively.

Moreover, as a measure of how well the nodes are spread in the space, we compute the average distance between each pair of nodes in the embedding space for all graphlet based embeddings. In Table 4.2, we report the average Euclidean distance between all nodes when using the three graphlet based embeddings in SAFE over all GI networks. We observe that the average distance between nodes when using GraCoal embedding, is 2.11 to 3.56 times larger than when using GraSpring, and 475 to 2,850 times larger than when using graphlet based Spectral.

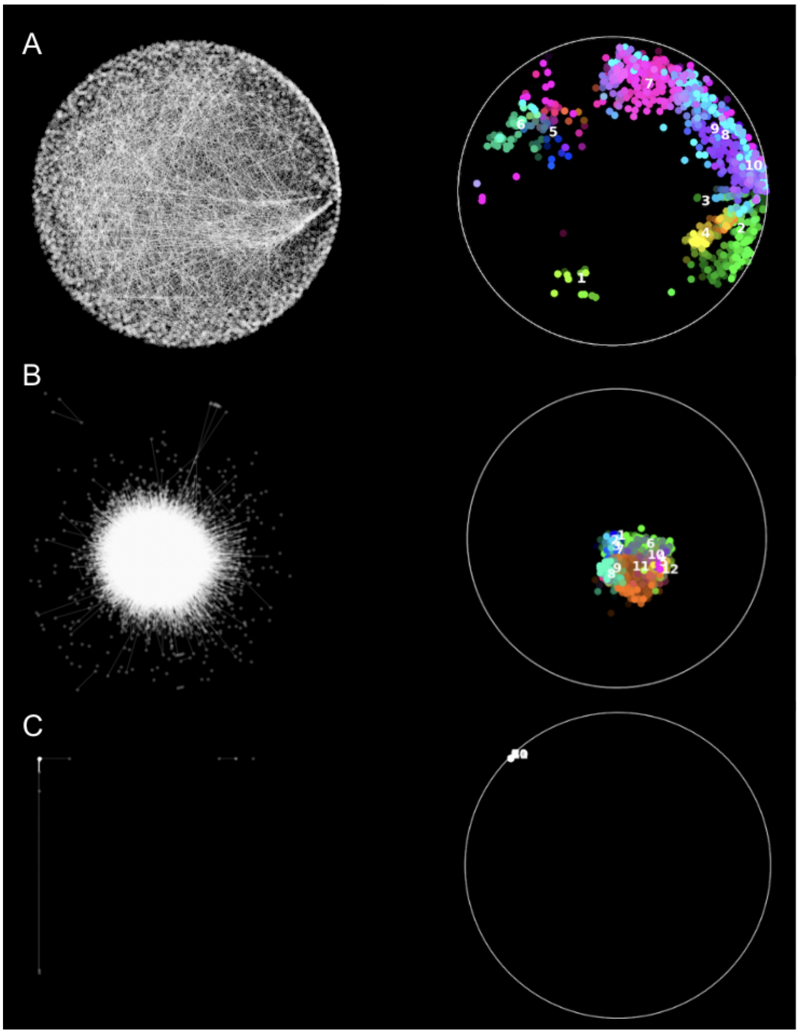


Figure 4.2. Functional landscape of the Budding yeast GI network for different types of network embedding based on graphlet adjacency $\widetilde{A}_{G_0}$. We use SAFE to annotate the Budding yeast GI network with GO-BP for (A) GraCoal embedding, (B) Spring embedding and (C) Spectral embedding. For each type of network embedding, we show the network embedding on the left and the SAFE enrichment domains highlighted in colour on the right.

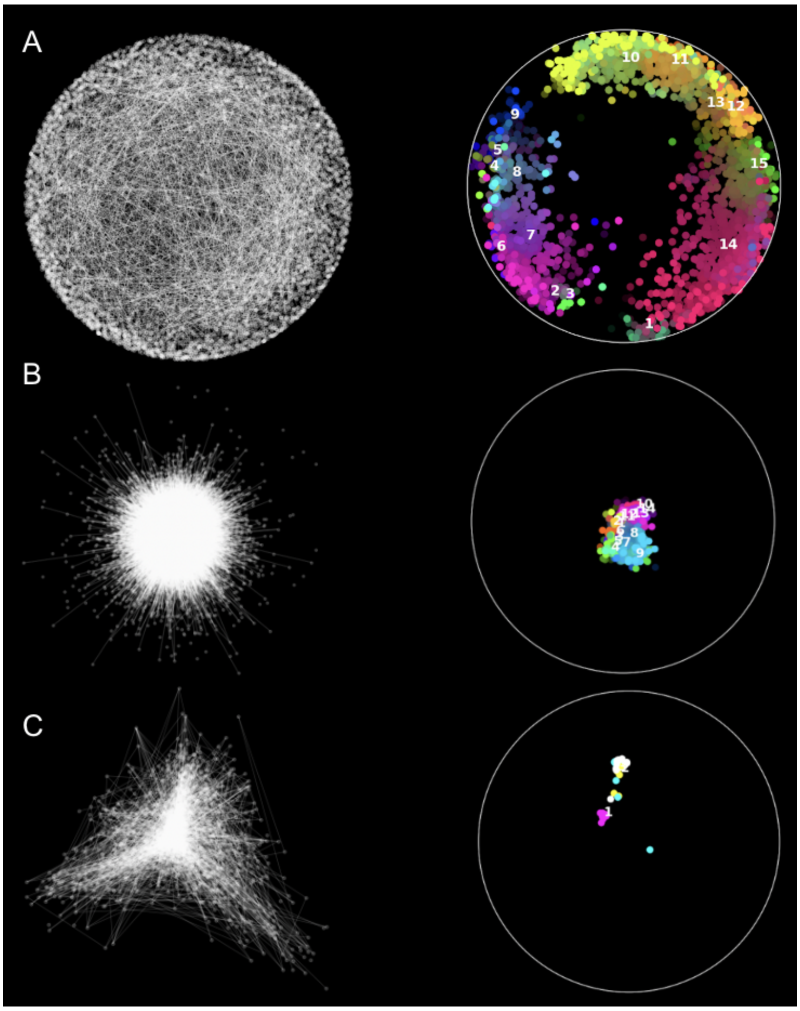


Figure 4.3. Functional landscape of the Budding yeast GI network for different types of network embedding based on graphlet adjacency $\widetilde{A}_{G_2}$. We use SAFE to annotate the Budding yeast GI network with GO-BP for (A) GraCoal embedding, (B) Spring embedding and (C) Spectral embedding. For each type of network embedding, we show the network embedding on the left and the SAFE enrichment domains highlighted in colour on the right.

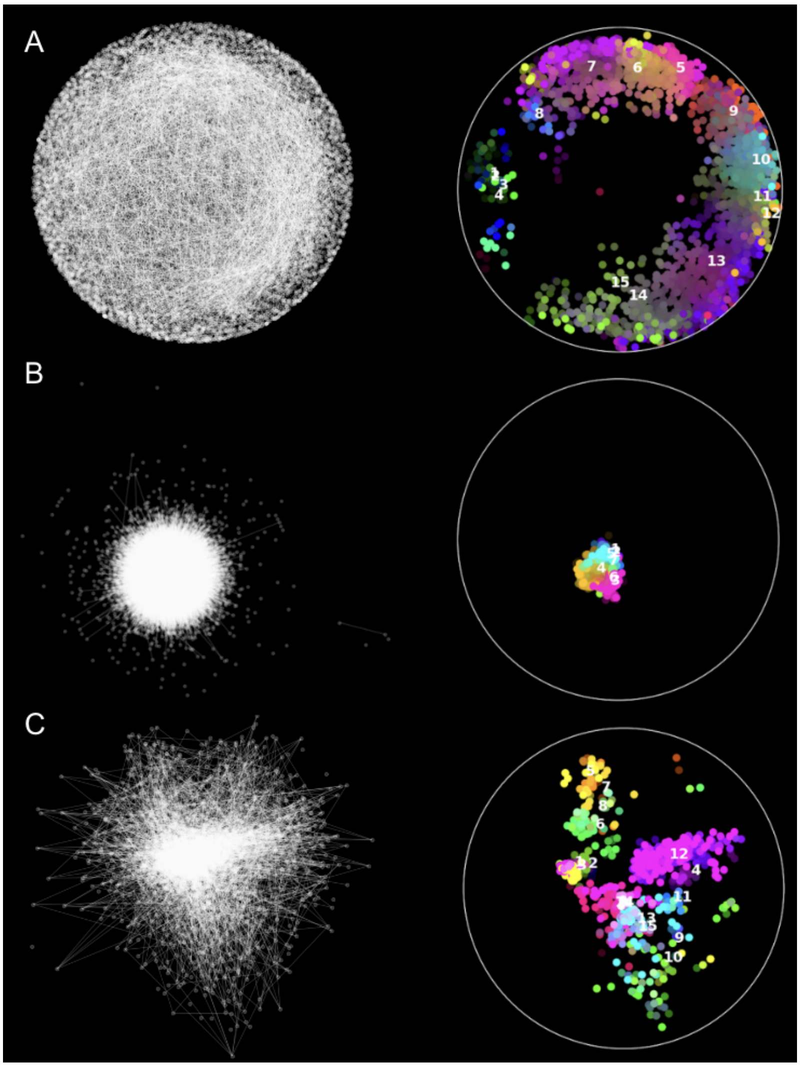


Figure 4.4. Functional landscape of the Budding yeast GI network for different types of network embedding based on graphlet adjacency $\widetilde{A}_{G_3}$. We use SAFE to annotate the Budding yeast GI network with GO-BP for (A) GraCoal embedding, (B) Spring embedding and (C) Spectral embedding. For each type of network embedding, we show the network embedding on the left and the SAFE enrichment domains highlighted in colour on the right.

	GraCoal	**GraSpring**	**Spectral**
Budding yeast	0.57 (std=0.00)	0.16 (std=0.03)	0.08 (std=0.05)
E. coli	0.57 (std=0.00)	0.27 (std=0.04)	0.12 (std=0.03)
Fission yeast	0.57 (std=0.00)	0.21 (std=0.05)	0.02 (std=0.02)
Fruit fly	0.56 (std=0.00)	0.22 (std=0.03)	0.02 (std=0.01)

Table 4.2. Average Euclidean distance between nodes in graphlet based embeddings. For each of our GI molecular networks (rows), we report the average Euclidean distance between all pairs of nodes across all graphlet adjacencies (i.e., $\widetilde{A}_{G_0}$ - $\widetilde{A}_{G_8}$) for GraCoal, GraSpring and graphlet based Spectral embeddings (columns 1-3).

GraCoal embeddings uncover complementary biological information

After having shown that GraCoal embeddings work best for the GI molecular networks when using SAFE, we then investigate which topologies (i.e., graphlets) capture the most function in GI networks by comparing our results between the different GraCoals. We present our results in Figure 4.5 and observe that for the two species where GraCoals capture the most function, i.e., fruit fly and budding yeast, there are clear top performing GraCoals. For budding yeast for instance, the top performing GraCoals, $GraCoal_{2,3,6,7}$ achieve between 42.0% and 45.2% enriched genes, which is distinctly better then to the low performing GraCoals, $GraCoal_{0,1,4,5,8}$, which achieve between 17.4% and 34.9% enriched genes. Additionally, we observe that the top performing GraCoals are not the same across species, as those for fruit fly (GraCoal$_{0,1,3,4,6}$) are clearly distinct of those for budding yeast ($\widetilde{A}_{G_{2,3,6,7}}$). Notably, GraCoals based uniquely on triangles, GraCoal$_{2,7}$, perform particularly well in budding yeast but not in fruit fly. Conversely, GraCoals void of triangles, GraCoal$_{0,1,3,4}$ perform particularly well in fruit fly but not in yeast. For Fission yeast, the best performing GraCoals (GraCoals$_{2,3,6}$) largely follow those for Budding yeast, although the differences in performance between the different GraCoals is less pronounced. For *E. coli*, there are no clear best GraCoals. In all, these results imply that the same GraCoals capture different topology-function relationships in GI networks depending on the species.

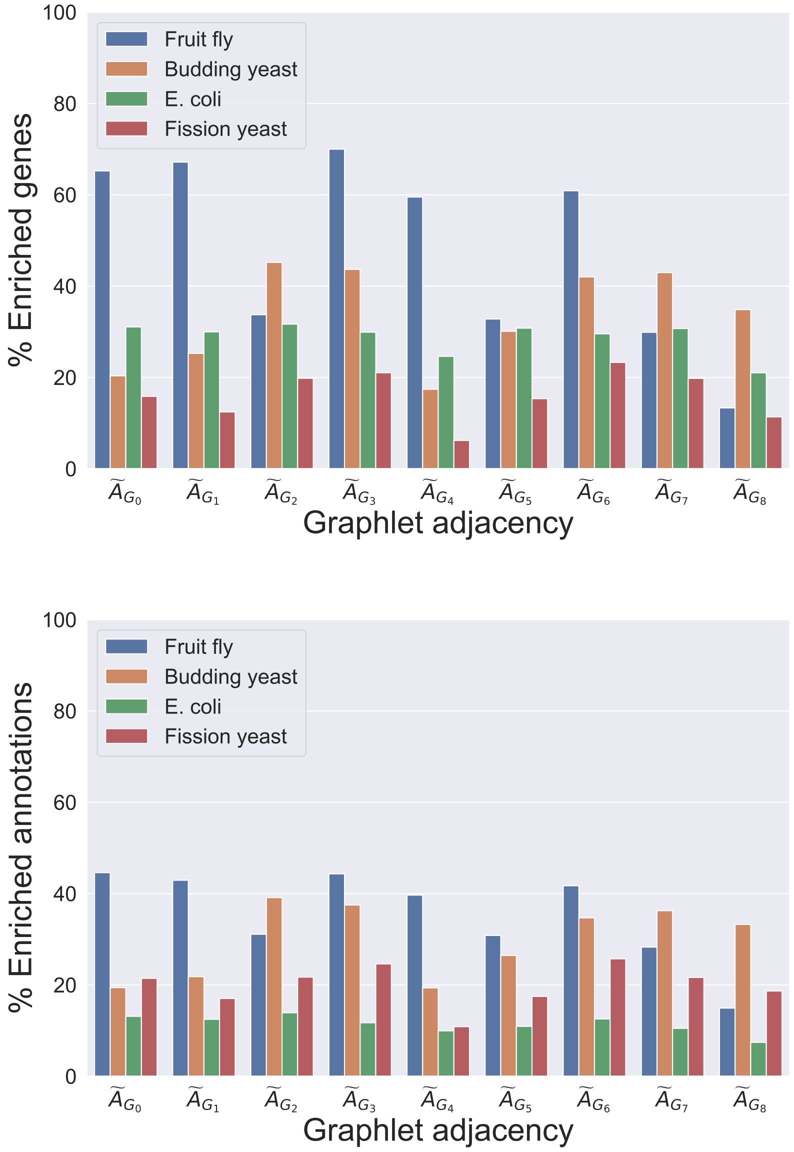


Figure 4.5. SAFE GO-BP enrichment analysis comparing GraCoals in GI networks. For the GI networks of our four species (legend), we show, on the y-axis, the percentage of enriched genes (top) and the percentage of enriched annotations (bottom) for each of the different GraCoal embeddings (x-axis).

Next, for each species, we focus on identifying what characterizes each particular GraCoal (i.e., $\widetilde{A}_{G_0}$-$\widetilde{A}_{G_8}$) from a biological perspective. To this end, we explore in more detail the functional information uncovered by the GraCoal embeddings when used in SAFE. We do this at the annotation level (i.e., identifying particular GO-BPs that are characteristic of each GraCoal) and at the functional domain level (i.e., identifying particular functional domains that are characteristic of each GraCoal).

First, we identify the uniquely enriched annotations for each GraCoal (i.e., annotations enriched in a particular GraCoal that are not enriched in any of the other GraCoals). For the budding yeast, we find that on average, 22 GO-BPs are uniquely enriched for each particular GraCoal. This is in line with the literature, as different graphlet adjacencies are known to capture complementary topology-function relationships in molecular networks (Windels et al., 2019).

We observe that each GraCoal embedding not only uncovers biological information that is not uncovered by the other GraCoal embeddings, but that there is some degree of functional similarity in the information uncovered. For instance, five of the ten largest GO-BP enriched for GraCoal_2 are related to nuclease activity. This implies some larger function being uniquely captured by each GraCoal. To further asses that this is the case, we evaluate the biological relevance of the uniquely enriched annotations (i.e., GO-BPs) for a particular GraCoal, by computing the semantic similarity of each pair of annotations in the set of annotations. We define the semantic similarity (SS) as the inverse of the semantic distance between a given pair of GO terms, where the semantic distance is the minimum number of connecting branches between the pair of GO terms in the ontology directed acyclic graph (DAG) (Dessimoz & Škunca, 2017; Rada et al., 1989). Finally, we rank the uniquely enriched annotations according to their size, defined as the total number of neighborhoods they are enriched in, as a measure of how well they are captured by each particular GraCoal. In table (Table 4.3 we report for the budding yeast, the number of uniquely enriched annotations (column 1), the mean SS for the uniquely enriched annotations (column 2) as well as the mean SS for the top 10 largest uniquely enriched annotations (column 3) for each GraCoal used in SAFE. Finally, in column 5, we report the names of the top 10 uniquely enriched annotations and their corresponding size in terms of enriched neighborhoods (column 4). The lowest and maximum average SS for the sets of uniquely enriched annotations for the budding yeast GI network are 0.15 (Std=0.06) and 0.29 (Std=0.05) for $\widetilde{A}_{G_1}$ and $\widetilde{A}_{G_8}$, respectively. In general, when evaluating the top 10 enriched annotations, we observe a larger degree of functional similarity between the annotations, which ranges from 0.21 (Std=0.04) to 0.33 (Std=0.05) for the budding yeast ($\widetilde{A}_{G_1}$, and $\widetilde{A}_{G_2}$, respectively). For the summary of uniquely enriched annotations for our other GI networks please refer to Tables A.3 to A.5 in Appendix A.

$\widetilde{A}_{G_i}$	**Total annotations**	**Mean SS**	**Mean SS Top 10**	**EN**	**Annotation**
$\widetilde{A}_{G_0}$	17	0.21 (Std = 0.05)	0.28 (Std = 0.05)	193	protein localization to mitochondrion
				193	establishment of protein localization to mitochondrion
				192	mitochondrial transport
				118	ribosome disassembly
				110	protein insertion into membrane
				95	RNA methylation
				88	protein insertion into mitochondrial membrane
				85	establishment of protein localization to mitochondrial membrane
				82	regulation of DNA double-strand break processing
				72	tRNA methylation
$\widetilde{A}_{G_1}$	25	0.15 (Std = 0.06)	0.21 (Std = 0.04)	176	nuclear pore localization
				174	tRNA gene clustering
				161	positive regulation of attachment of spindle microtubules to kinetochore
				112	attachment of spindle microtubules to kinetochore involved in meiotic chromosome segregation
				112	monopolar spindle attachment to meiosis I kinetochore
				111	spliceosomal complex assembly
				103	DNA unwinding involved in DNA replication
				91	positive regulation of chromosome segregation
				72	positive regulation of DNA-templated transcription, initiation
				70	U1 snRNA 3'-end processing

Table 4.3. Summary of uniquely enriched GO-BPs for Gracoal embeddings, Part 1. We report, for the budding yeast GI network, the number of uniquely enriched GO-BPs and the mean semantic similarity (SS) between the uniquely enriched annotations (GO-BPs) (columns 2 and 3) for GraCoals based on all graphlet adjacencies for up to four node graphlets, i.e. $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$ (column 1). In column 4, we report the mean SS for the top ten largest enriched annotations (column 6), i.e., ranking them in descending order according to the number of neighborhoods that the annotations are enriched in (column 5).

	Total annotations	**Mean SS**	**Mean SS Top 10**	**EN**	**Annotation**
$\widetilde{A}_{G_2}$	46	0.17 (Std = 0.04)	0.33 (Std = 0.05)	270	double-strand break repair via nonhomologous end joining
				232	positive regulation of DNA metabolic process
				231	regulation of reproductive process
				195	endonucleolytic cleavage in ITS1 upstream of 5.8S rRNA from tricistronic rRNA transcript
				194	regulation of nuclease activity
				190	regulation of deoxyribonuclease activity
				173	regulation of endodeoxyribonuclease activity
				166	regulation of transcription by RNA polymerase I
				161	positive regulation of deoxyribonuclease activity
				161	positive regulation of nuclease activity
$\widetilde{A}_{G_3}$	21	0.19 (Std = 0.04)	0.25 (Std = 0.06)	182	retrograde vesicle-mediated transport Golgi to endoplasmic reticulum
				165	protein methylation
				165	protein alkylation
				157	peptidyl-lysine methylation
				154	replication fork arrest
				146	Golgi organization
				140	mitotic DNA damage checkpoint signaling
				128	mitotic intra-S DNA damage checkpoint signaling
				124	histone H3-K79 methylation
				120	organophosphate biosynthetic process
$\widetilde{A}_{G_4}$	7	0.29 (Std = 0.05)	0.29 (Std = 0.06)	145	resolution of meiotic recombination intermediates
				93	positive regulation of cell cycle process
				90	positive regulation of cell cycle
				33	cellular component disassembly
				28	transition metal ion transport
				13	phosphatidylcholine biosynthetic process
				13	phosphatidylcholine metabolic process
$\widetilde{A}_{G_5}$	35	0.20 (Std = 0.05)	0.30 (Std = 0.05)	200	response to cell cycle checkpoint signaling
				200	cellular response to biotic stimulus
				196	response to biotic stimulus
				192	cellular response to endogenous stimulus
				192	response to endogenous stimulus
				74	vacuole organization
				74	regulation of signal transduction
				74	regulation of signaling
				70	regulation of intracellular signal transduction
				70	vacuole fusion, non-autophagic

Table 4.3. Summary of uniquely enriched GO-BPs for Gracoal embeddings, Part 2. We report, for the Budding yeast GI network, the number of uniquely enriched GO-BPs and the mean semantic similarity (SS) between the uniquely enriched annotations (GO-BPs) (columns 2 and 3) for GraCoals based on all graphlet adjacencies for up to four node graphlets, i.e. $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$ (column 1). In column 4, we report the mean SS for the top ten largest enriched annotations (column 6), i.e., ranking them in descending order according to the number of neighborhoods that the annotations are enriched in (column 5).

	Total annotations	Mean SS	Mean SS Top 10	EN	Annotation
$\widetilde{A}_{G_6}$	1	Na	Na	101	ribosomal small subunit export from nucleus
$\widetilde{A}_{G_7}$	13	0.25 (Std = 0.06)	0.27 (Std = 0.05)	184	leading strand elongation
				152	regulation of cellular response to stress
				131	regulation of response to stress
				100	resolution of recombination intermediates
				95	reproduction
				87	regulation of response to endoplasmic reticulum stress
				85	protein targeting to membrane
				83	regulation of endoplasmic reticulum unfolded protein response
				65	response to unfolded protein
				34	organic hydroxy compound metabolic process
$\widetilde{A}_{G_8}$	34	0.17 (Std = 0.04)	0.22 (Std = 0.03)	239	transcription by RNA polymerase II
				159	cellular chemical homeostasis
				151	regulation of cellular component organization
				139	regulation of microtubule-based process
				126	ubiquitin-dependent ERAD pathway
				116	autophagy of peroxisome
				114	chromosome organization involved in meiotic cell cycle
				111	regulation of microtubule cytoskeleton organization
				110	sphingolipid metabolic process
				110	chemical homeostasis

Table 4.3. Summary of uniquely enriched GO-BPs for Gracoal embeddings, Part 3. We report, for the Budding yeast GI network, the number of uniquely enriched GO-BPs and the mean semantic similarity (SS) between the uniquely enriched annotations (GO-BPs) (columns 2 and 3) for GraCoals based on all graphlet adjacencies for up to four node graphlets, i.e. $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$ (column 1). In column 4, we report the mean SS for the top ten largest enriched annotations (column 6), i.e., ranking them in descending order according to the number of neighborhoods that the annotations are enriched in (column 5).

Thus far, we have shown that GraCoal embeddings outperform GraSpring embeddings and Graphlet Spectral embedding in capturing the functional organization of genetic interaction networks. Additionally, we have shown that different GraCoals capture different topology-function relationships when applied to different GI networks (i.e., for different species). Lastly, we observed that triangle based graphlets or graphlets void of triangles tend to perform very well depending on the species of the GI network being looked at. In light of these observations, in the next section we perform topological analysis to explain why some GraCoals work better depending on the species and we explore in more detail the functional information uncovered at the functional domain level.

The topology-function relationships captured by GraCoals

We observed that triangle-based GraCoals (GraCoals$_{2,7}$) or GraCoals void of triangles (GraCoals$_{0,1,3,4}$) tend to best capture the functional organization of GI networks depending on the species. Here, we investigate when triangle based GraCoals work best. For ease of readability we focus on GraCoal$_2$, although the same conclusion can be reached based on GraCoals$_7$. In our analysis, we first characterise the organizational principles of our GI networks by comparing their topology (wiring) to that of model networks (see Section 2.4: Model networks and section A.5 in Appendix A) and then relate the organizational principles of the GI network to our enrichment results.

Our model-fitting results for our GI networks are presented in Figure A.44 in Appendix section A.5. We observe that all four GI networks have non-random topology, as they can be distinguished from ER networks (least significant p-value 1.0E-6, see Table A.21 in Appendix A), implying they are functionally organised. Additionally, we observe that the topologies of the GI networks for budding yeast, fission yeast and *E. coli* are almost indistinguishable from Scale-Free Gene Duplication (SF-GD) networks (Figure A.45 in Appendix A). Firstly, the scale-freeness of these networks is in line with the literature, as GI networks are known to be scale-free (Tong et al., 2004). Secondly, this result implies that numerous gene-duplication events along the DNA have influenced these GI networks' topologies. This is consistent with the literature in case of budding yeast, as its genome has undergone a whole genome duplication event (Kellis et al., 2004). Similarly, for *E. coli*, 60% of its genes have been reported to have at least one paralogous gene (i.e., a homologous gene that has diverged within one species due to gene duplication events) (Blattner et al., 1997; Bratlie et al., 2010; Patterson, 1988). Thus, to enable further investigation, we determine for each species a set of *gene-paralogs*.

Gene-paralog assessment

After having shown that the SFGD network model is the best fit for the GI molecular networks, we first identify the sets of gene-paralogs. In brief, for each species, we collect the corresponding proteome from Ensembl (Cunningham et al., 2022) and compute the pairwise sequence alignments between all proteins using BlastP (Altschul et al., 1990) using the procedure and thresholds outlined in Pearson, 2013. We consider pairs of genes with a percentage of sequence identity of at least 85%, an E-value equal to or less than 0.001 and a bit score of at least 50 as paralogous genes. For details on the gene-paralogs per species and network, see Table A.24 in Appendix A.

Genes enriched in *E. coli*, Fission yeast and Budding yeast GI networks cover more paralogous genes than the Fruit fly GI network

Below, we report the total number of genes that have at least one GO-BP enriched in the neighborhood (i.e., "Enriched genes" in Table 4.4) and the number of genes that are enriched and are paralogous (i.e., "Paralogs" in Table 4.4). For the fruit fly GI network, we observe that even when obtaining the largest percentages of genes enriched, the number of paralogous genes covered are relatively low. In particular, for GraCoal$_2$ we observe a large difference between the number of paralogs enriched in budding yeast with respect to the number of paralogs enriched in fruit fly.

We find that the budding yeast, *E. coli* and fission yeast have close to three times more paralogous genes in their GI network than fruit fly (see Table A.24 in Appendix A), whose GI network topology can be distinguished from SF-GD (p-value 9.63E-7); although even for fruit fly it is still the best-fitted model network.

When relating GI topology to our enrichment results, we immediately observe that our topological findings correlate with our GO-BP enrichment results, as the triangle based GraCoals, GraCoal$_2$ achieve among the best GO-BP enrichment scores in budding yeast, fission yeast and *E. coli* (GI networks indistinguishable from SF-GD) but poor scores in fruit fly (GI network distinguishable from SF-GD). This observation could imply that, in GI networks, GraCoals$_2$ capture GO-BPs that involve functional paralogous genes in species with many duplicated genes in their genome. We further validate this hypothesis by showing that the genes enriched based on GraCoals$_2$ for budding yeast, fission yeast and *E. coli* cover relatively more paralogous gene pairs than for fly (Table 4.4). Moreover, for budding yeast, fission yeast and *E. coli*, the genes enriched based on GraCoals$_2$ cover more paralogous genes than any other GraCoal embedding (except for GraCoal$_6$ in *E. coli*).

	Budding yeast		***E. coli***		**Fission yeast**		**Fruit fly**	
	Enriched Genes	Paralogs	Enriched Genes	Paralogs	Enriched Genes	Paralogs	Enriched Genes	Paralogs
$\widetilde{A}_{G_0}$	1,189	10.09	1,234	23.26	567	17.64	2,061	7.13
$\widetilde{A}_{G_1}$	1,476	11.04	1,191	26.87	445	16.85	2,121	6.51
$\widetilde{A}_{G_2}$	2,640	21.74	1,258	25.51	709	26.23	1,066	8.92
$\widetilde{A}_{G_3}$	2,551	18.35	1,188	25.67	752	19.28	2,211	7.92
$\widetilde{A}_{G_4}$	1,017	6.19	978	26.48	222	4.50	1,879	10.27
$\widetilde{A}_{G_5}$	1,759	16.09	1,223	26.49	549	20.22	1,036	7.24
$\widetilde{A}_{G_6}$	2,454	17.36	1,174	30.66	833	19.93	1,923	7.28
$\widetilde{A}_{G_7}$	2,509	20.65	1,220	25.49	708	22.46	944	7.31
$\widetilde{A}_{G_8}$	2,036	14.00	835	14.73	406	16.26	421	10.22

Table 4.4. Statistics for paralogous genes enriched using SAFE with GraCoal embeddings. For each of the four GI networks (Budding yeast, *E. coli*, Fission yeast and Fruit fly), we show the number of enriched genes when using SAFE with GraCoal embeddings (i.e., genes that have at least one annotation enriched in their neighborhood) and the percentages of genes enriched that are paralogs.

Next, we explain why GraCoal$_2$ best captures GO-BP involving paralogous genes. First, we show that paralogous genes are statistically significantly more likely to genetically interact than non-paralogous genes (using a hypergeometric test, least significant p-value 2.02E-2 over all four species). This observation is consistent with the literature, as one of the key drivers for the retention of duplicated genes in the genome is functional redundancy (Kuzmin et al., 2020), in which case duplicated genes are also likely to genetically interact and interact with the same genes. Consequently, paralogous genes should tend to form triangles in the GI network, in which two nodes are the two paralogs and a third is a shared neighbour. We confirm this by showing in Figure 4.6 that duplicated genes occur on statistically significantly more triangles, i.e., graphlet G_2, than randomly selected pairs from the network (i.e., the background) (Table 4.5). Moreover, we assess if gene-paralog pairs share similar wiring patterns by computing the graphlet degree vector similarity (GDV-sim) between all 'gene to paralog' pairs in the network and comparing it to randomly chosen pairs. We perform a one-sided Mann-Whitney-U test to see if gene-paralog pairs have a larger GDV similarity with respect to the background. Below we show that paralogous pairs of genes have statistically larger GDV similarities with respect to randomly chosen pairs of genes from any of the GI networks (Figure 4.7, left panels and Table 4.5). As a consequence of having more similar wiring patterns than random, we also show how paralogous pairs are closer in embedding space than random (Figure 4.7, right panels). Thus, we also perform a one-sided MWU test to evaluate if paralogous pairs have statistically shorter distances (i.e., shortest weighted path lengths) than random pairs. We do this for all four GI networks and in every case the distances for paralogous pairs is statistically shorter with respect to randomly chosen pairs (Table 4.5). In all, these findings explain why GO-BP involving paralogs tend to be enriched using SAFE.

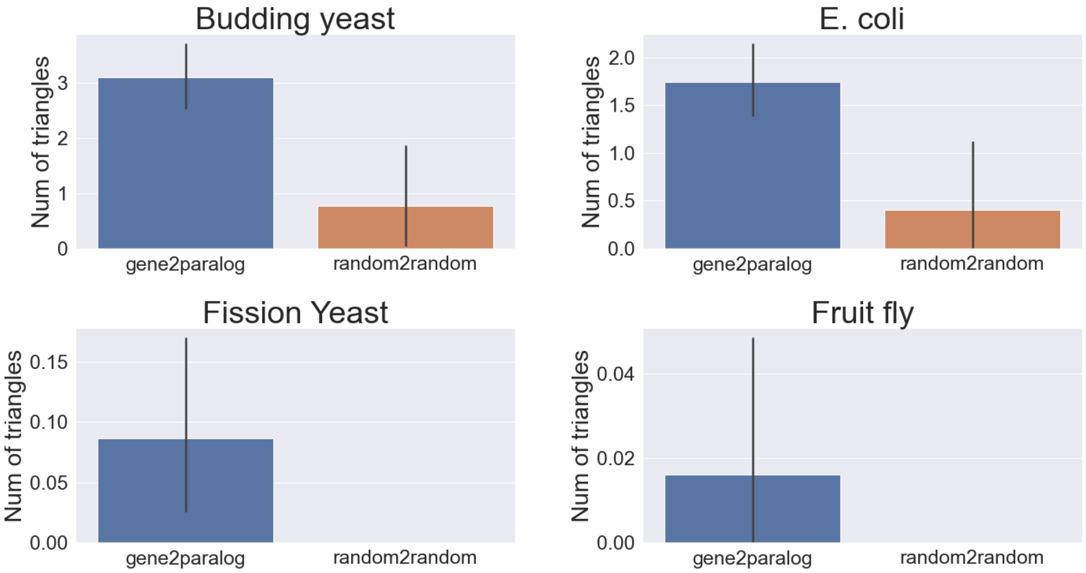


Figure 4.6. Triangle count distribution for pairs of paralogous genes. For each of our GI molecular networks, we show the triangle count distribution for all pairs of paralogous genes (i.e., blue 'gene2paralog') and the triangle count distribution for random pairs of genes in the network (i.e., orange 'random2random').

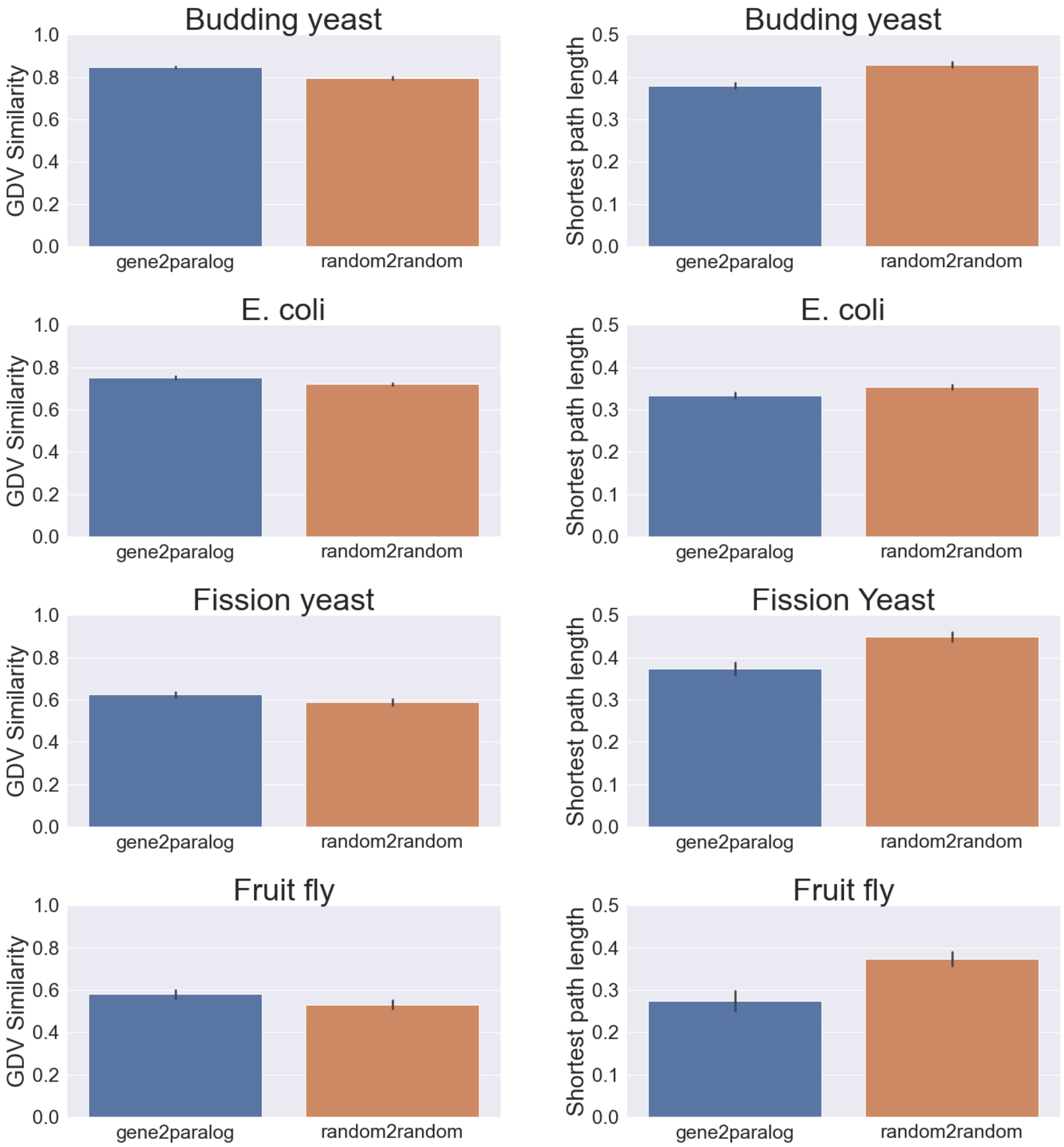


Figure 4.7. GDV similarity and shortest weighted path lengths. On the left panels we show the GDV similarity distribution between all pairs of paralogous genes (i.e., blue 'gene2paralog') and the GDV similarity distribution for random pairs of genes in the network (i.e., orange 'random2random'). P-values corresponding to the one-sided MWU test between these two distributions is statistically significant ($p<0.05$) in every case. On the right panels we show the distribution of the shortest weighted path lengths between all pairs of paralogous genes (blue) and the distribution of shortest weighted path lengths between random pairs (orange). P-values are also statistically significant in every case ($p<0.05$). From top to bottom: results for the Budding yeast, *E. coli*, Fission yeast and Fruit fly GI networks.

	Triangle counts	GDV-sim	SWPL
Budding yeast	1.78E-23	4.90E-32	2.79E-21
E. coli	3.77E-07	1.27E-10	7.29E-07
Fission yeast	0.001	7.38E-05	6.72E-15
Fruit fly	0.16	0.002	6.53E-10

Table 4.5. MWU p-values. We show the resulting p-values after comparing different distributions by performing a one-sided MWU test. Column "Triangle counts" shows the p-values for a one-sided MWU test to compare triangle counts in the network for paralogous pairs vs random pairs of nodes to see if the paralogous pairs participate in statistically significant more triangles than random. Column 'GDV-sim' shows p-values for one-sided MWU test to see if the GDV similarity is statistically larger in between paralogous pairs than random pairs. Finally, 'SWPL' shows p-values for one-sided MWU test to see if paralogous pairs of nodes are statistically closer in space than random pairs.

In summary, we find that when the genome of a species contains many duplicated genes, this is reflected in the topology of their GI networks, as paralogs tend to interact and share many of their neighbours in the network, leading to dense patches of triangles in the GI networks. This topology is well captured by GraCoal$_2$ (i.e., GraCoal based on the triangle shaped graphlet), leading to high percentages of GO-BP enrichments, driven by the high enrichment of GO-BPs that including paralogous genes.

Biological insights of GraCoals at a functional domain level

Lastly, we aim to give insight into the biological function captured by our GraCoals across species, and in particular when using GraCoal$_2$. To this end, we identify the most characteristic functional domains in each species, i.e., the domains that could not be captured by any of the other GraCoals. To quantify this, we measure the uniqueness of all functional domains obtained with SAFE. To do this, we first compute the Jaccard similarity index (JI) (Jaccard, 1912; Tanimoto, 1958), which is defined as the size of the intersection between two sets of elements divided by the size of the union of the two sets of elements. We compute this between the sets of enriched annotations of each functional domain in a particular GraCoal and the sets of enriched annotations of every other functional domain in the other GraCoal embeddings. Next, for each functional domain, we keep the maximum JI, as this value represents the maximum overlap to any other functional domain in the other GraCoal embeddings and thus, reflects how unique to a particular GraCoal a functional domain is. Additionally, we compute the paralog ratio for each functional domain, which we define as follows: for a given functional domain, the paralog ratio is the number of paralogs that are annotated by the enriched annotations in the functional domain over the total number of genes annotated by the enriched annotations in the functional domain. In this way, we can evaluate how well a given functional domain is capturing biological information that involves paralogs.

In Table 4.6 for budding yeast, and Tables A.12 to A.14 in Appendix A for *E. coli*, fission yeast and fruit fly, respectively, we summarize, the number of functional domains (column 1) and the mean paralog ratio (column 2), over each GraCoal embedding (i.e., $\widetilde{A}_{G_0}$-$\widetilde{A}_{G_8}$). Furthermore, we report the top three most characteristic functional domains for the GI molecular networks (column 5) according to the lowest maximum JI (column 4) and the corresponding paralog ratio.

Firstly, we observe that for fruit fly, budding yeast and fission yeast, many domains are highly characteristic to a particular GraCoal, with many domains (7, 14 and 5, respectively) being completely unique, scoring a maximum JI of 0. This is in line with our earlier results, as we observed big discrepancies in gene enrichment performance between different GraCoals for these species, indicating we capture strong topology-function relationships. Secondly, we observe that for budding yeast, fission yeast and *E. coli*, the top three most characteristic domains of GraCoal$_2$ show on average the highest paralog ratios of all GraCoals (except in *E. coli*, where GraCoal$_2$ is just behind GraCoal$_8$ in this regard). This is in line with our previous observation that GraCoal$_2$ tends to capture GO-BP involving paralogs. Lastly, we can also find literature support that the paralogs combined in our domains are functional. For instance, if we focus on the 'secretion, cell, exocytosis, export' domain uncovered by GraCoal$_2$ in budding yeast. This domain has the largest paralog ratio measured for all domains in budding yeast (0.43) and is strongly characteristic of GraCoal$_2$ (JI=0.12). This domain is composed of GO-BPs such as 'export from cell', 'secretion by cell', 'secretion' and 'exocytosis', which are all vesicle traffic related functions. Previous studies suggest that gene duplication events enabled the expansion and diversification of the vesicle traffic pathway (Purkanti & Thattai, 2022). The authors show that gene duplications allowed for the formation of paralogous modules. As paralogs can be differentially expressed or regulated, or can have different interaction partners, paralogous modules contribute to the robustness and versatility of the vesicle traffic pathway. We can make similar observations for the 'cell wall chitin biosynthetic process' domain, the most characteristic domain of GraCoal$_8$ (JI=0.0, paralog ratio 0.43). A key element of the biosynthetic process is the 'exomer' protein complex, a heterotetrameric complex assembled at the trans-Golgi network, that is required for the delivery of a distinct set of proteins to the plasma membrane. Its cargo adaptors consist of two Chs5 proteins and two out of four paralogous proteins: Bud7, Bch1, Bch2 and Chs6. The paralogs part of the exomer complex determine which proteins it can transport (Anton et al., 2018). For instance, transport of Chs3 is completely dependent on the presence of Chs6 in the exomer. So, in the chitin biosynthetic process, gene duplication enabled different specialisations of the exomer to transport different proteins, which is captured by GraCoal$_8$.

In conclusion, we have shown that triangle based GraCoals capture functional redundancy and functional specialisation in GI networks of species whose GI network is characterised by many paralogs.

$\widetilde{A}_{G_i}$	**Num functional domains**	**Mean paralog ratio**	**Domain paralog ratio**	**Domain max JI**	**Domain description**
$\widetilde{A}_{G_0}$	10	0.16 (std=0.08)	0.32	0.00	cytokinesis, cytoskeleton, septin, organization, histone
$\widetilde{A}_{G_0}$	10	0.16 (std=0.08)	0.15	0.07	regulation, attachment, spindle, microtubules, kinetochore
$\widetilde{A}_{G_0}$	10	0.16 (std=0.08)	0.07	0.13	metabolic, process, glycolipid, liposaccharide
$\widetilde{A}_{G_1}$	15	0.15 (std=0.06)	0.25	0.00	process, amino, acid, biosynthetic, glutamine
$\widetilde{A}_{G_1}$	15	0.15 (std=0.06)	0.11	0.05	histone, methylation, lysine, H3, K4
$\widetilde{A}_{G_1}$	15	0.15 (std=0.06)	0.06	0.13	acetylation, peptidyl, lysine, modification, internal
$\widetilde{A}_{G_2}$	15	0.20 (std=0.09)	0.26	0.00	growth, in, filamentous, conjugation, with
$\widetilde{A}_{G_2}$	15	0.20 (std=0.09)	0.17	0.00	biosynthetic, process, purine, ribonucleotide, nucleotide
$\widetilde{A}_{G_2}$	15	0.20 (std=0.09)	0.43	0.12	secretion, cell, exocytosis, export, by
$\widetilde{A}_{G_3}$	15	0.16 (std=0.07)	0.25	0.00	purine, containing, compound, metabolic, process
$\widetilde{A}_{G_3}$	15	0.16 (std=0.07)	0.16	0.00	electron, transport, chain, aerobic, respiratory
$\widetilde{A}_{G_3}$	15	0.16 (std=0.07)	0.18	0.02	receptor, recycling, protein, import, peroxisome
$\widetilde{A}_{G_4}$	11	0.19 (std=0.12)	0.24	0.00	transition, metal, ion, transport
$\widetilde{A}_{G_4}$	11	0.19 (std=0.12)	0.22	0.00	phosphatidylcholine, process, metabolic, biosynthetic
$\widetilde{A}_{G_4}$	11	0.19 (std=0.12)	0.26	0.08	transport, retrograde, endosome, Golgi, endosomal
$\widetilde{A}_{G_5}$	11	0.18 (std=0.19)	0.15	0.00	transmembrane, transport, hexose, monosaccharide, small
$\widetilde{A}_{G_5}$	11	0.18 (std=0.19)	0.13	0.00	regulation, heterochromatin, assembly, negative, organization
$\widetilde{A}_{G_5}$	11	0.18 (std=0.19)	0.17	0.04	positive, regulation, process, cellular, biological
$\widetilde{A}_{G_6}$	10	0.18 (std=0.19)	0.18	0.00	rRNA, RNA, splicing, transesterification, LSU
$\widetilde{A}_{G_6}$	10	0.18 (std=0.19)	0.12	0.57	rRNA, processing, SSU, RNA, endonucleolytic
$\widetilde{A}_{G_6}$	10	0.18 (std=0.19)	0.14	0.62	catabolic, process, dependent, macromolecule, protein
$\widetilde{A}_{G_7}$	16	0.18 (std=0.19)	0.15	0.00	cellular, response, stimulus, abiotic, osmotic
$\widetilde{A}_{G_7}$	16	0.18 (std=0.19)	0.08	0.08	mRNA, cleavage, polyadenylation, processing, response
$\widetilde{A}_{G_7}$	16	0.18 (std=0.19)	0.26	0.44	rRNA, SSU, processing, endonucleolytic, cleavage
$\widetilde{A}_{G_8}$	10	0.22 (std=0.06)	0.43	0.00	actin, cytoskeleton, organization, filament, based
$\widetilde{A}_{G_8}$	10	0.22 (std=0.06)	0.21	0.00	membrane, cell, wall, chitin, process
$\widetilde{A}_{G_8}$	10	0.22 (std=0.06)	0.21	0.12	response, compound, organonitrogen, ERAD, pathway

Table 4.6. Summary of most unique functional domains for Gracoal embeddings. We report for the budding yeast GI network, for each GraCoal embedding used with SAFE (column 1), i.e., based on graphlet adjacencies for up to four node graphlets ($\widetilde{A}_{G_0}$-$\widetilde{A}_{G_8}$), the number of functional domains (column 2) the mean paralog ratio (column 3) and the top three most characteristic functional domains (column 6). Lastly, for each functional domain we report the paralog ratio (column 4) and the maximum Jaccard similarity index (JI) (column 5).

4.2 Results for GraSpring with SAFE

In this section we present our results for the different GraSpring embeddings over the GI networks. Because of Spring embedding's non-deterministic nature, all of our results for GraSpring embeddings are based on the average results over 10 independent runs.

GraSpring uncovers the functional organization of the cell

In this section, we evaluate: (1) how well GraSpring embeddings capture the functional organization of genetic interaction networks and (2) which higher-order topologies (i.e., graphlets) capture the most function. In both experiments, we apply SAFE-based enrichment analysis to quantify how well a given embedding captures the functional organization of a given network.

To evaluate how well GraSpring capture the functional organization of our GI networks, we compare against Spring embedding (as GraSpring for graphlet G_0, which corresponds to standard Spring embedding, used in the original SAFE), GraCoal embedding and graphlet Spectral embedding (as it underlies our GraCoal embeddings). In Figure 4.1 we show the union of the enriched genes and the union of the enriched GO-BPs across the different types of underlying graphlet adjacencies (i.e., $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$). When comparing the three graphlet based embeddings, we have already established that GraCoal embeddings outperform both GraSpring and graphlet Spectral embeddings when uncovering the functional organization of GI networks using SAFE. Additionally, GraSpring embeddings are outperformed most of the time by graphlet Spectral embeddings as well. For instance, for the fruit fly GI network, the union of genes enriched in GO-BPs in terms of percentage (Figure 4.1, left), from highest to lowest (i.e., best to worst) are 90.3%, 76.72% and 61.3% for GraCoal, graphlet Spectral and GraSpring, respectively. Similarly, for budding yeast, these values are 71.4%, 61.13% and 42.6% for GraCoal, graphlet Spectral and GraSpring, respectively. This is also the case when considering a different annotation type such as GO-CC, except for the union of enriched GO-CCs for *E. coli*, where GraSpring outperforms both GraCoal and graphlet Spectral (Figure A.10). Finally, for our third annotation type, GO-MF, GraSpring is outperformed again by the other two graphlet based embeddings, except for the *E. coli* GI network, where GraSpring works best both in terms of the union of genes enriched in GO-MF and the union of enriched GO-MFs.

GraSpring embeddings uncover complementary biological information

Similar to GraCoal embeddings, we can use GraSpring embeddings to uncover biological information in complementary ways. Even when obtaining the lowest percentages of genes enriched and the lowest percentages of annotations enriched with respect to GraCoals and graphlet Spectrals across all species and annotations types, we still recover information across all GraSprings in each GI network. In Figure 4.8 we show the percentages of genes enriched (top) and percentages of GO-BPs enriched (bottom) for all GraSpring embeddings (i.e., $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$) across our four GI networks. We observe the largest percentages of both genes and annotations enriched in the fruit fly GI network, which is consistent with the previous results with GraCoal embeddings. On the opposite side, we obtain the lowest percentages of genes enriched and lowest percentages of annotations enriched for the fission yeast and *E. coli* GI networks, respectively. Interestingly, the second best enrichments in terms of genes are achieved by the budding yeast and *E. coli*, even though the latter is clearly the worst in terms of enriched GO-BPs. For GO-CC and GO-MF we present similar results in Figures A.12 and A.19 in Appendix A.

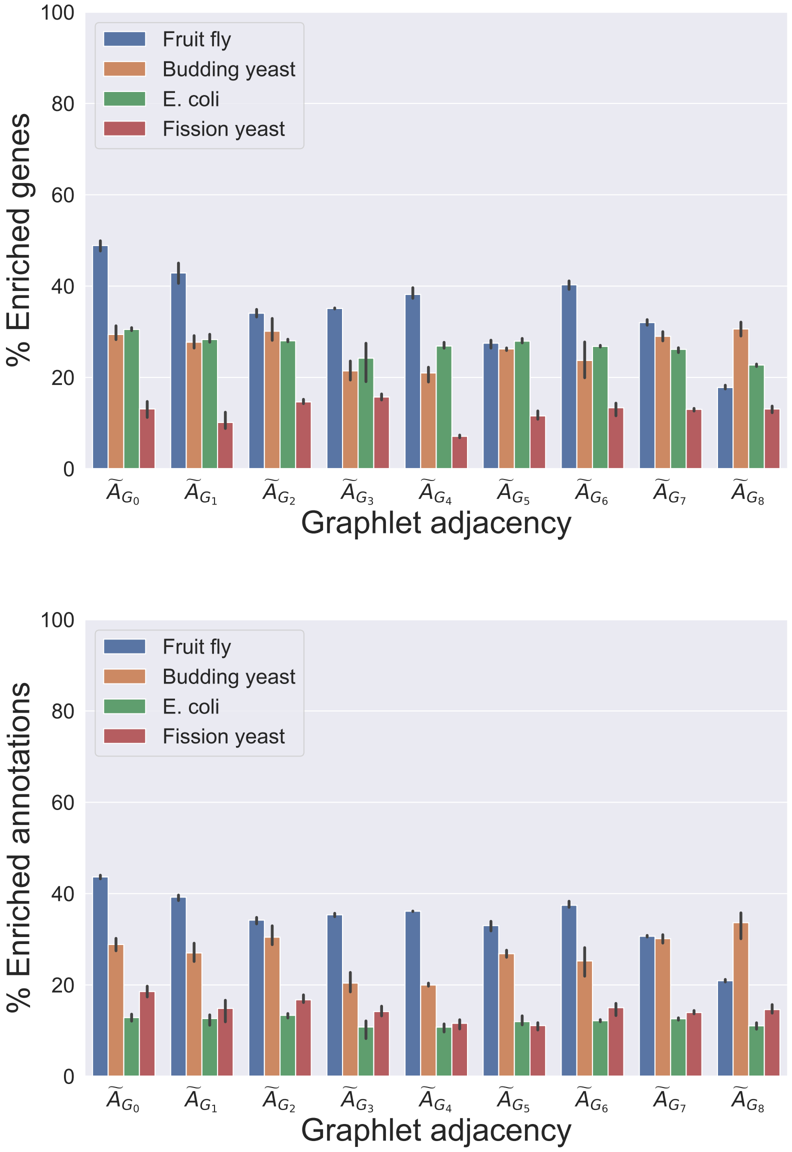


Figure 4.8. SAFE GO-BP enrichment analysis comparing Graspring in GI networks. For the GI networks of our four species (legend), we show, on the y-axis, the percentage of enriched genes (top) and percentage of enriched annotations (bottom) for each of the different Graspring embeddings (x-axis).

Next, we investigate which topologies capture the most function in GI networks between different GraSpring embeddings. We observe that in terms of percentages of genes enriched, the top performing GraSprings for fruit fly, $GraSpring_{0,1,4,5}$, achieve between 38.3% and 47.2%, while the low performing, $GraSpring_{2,3,6,7,8}$ achieve between 18.4% and 36.1%. This is also consistent for the percentages of enriched GO-BPs (except for $GraSpring_5$). Similarly, for budding yeast, even though the differences between the percentages of genes enriched are not as noticeable than for the fruit fly, we can also distinguish between top and low performing GraSprings. The top performing GraSprings, $GraSpring_{0,1,2,5,7,8}$, achieve between 25.1% and 32% and the low performing, $GraSpring_{3,4,6}$ achieve between 21.5% and 24.8%. This is consistent for percentages of GO-BPs enriched (Figure 4.8, bottom). Next, for *E. coli*, the top performing GraSprings, $GraSpring_{0,1,2,5}$, achieve between 26.2% and 29.7%, while the low performing, $GraSpring_{3,4,6,7,8}$, achieve between 22.3% and 25.8%. For percentages of enriched GO-BPs the differences are almost negligible, achieving between 13.8% and 16.1% across all GraSprings. For fission yeast, we observe two clear top performing GraSprings, $GraSpring_{2,3}$ in terms of genes enriched, achieving 17.7% and 18.3%, respectively, while the lowest, $GraSpring_{0,1,4,5,6,7,8}$ achieve between 7.3% and 16.2%. In terms of enriched GO-BPs, these results are somewhat consistent, as $GraSpring_{2,3}$ are amongst the top performing, even though $GraSpring_0$ is clearly the best. In all, when comparing between species, the results for top and low performing GraSprings are not always the same, though some overlap exist, which is consistent with our previous results for GraCoal embeddings across the GI networks. This implies that the same GraSprings capture different topology-function relationships in the GI networks, depending on the species.

Additionally, for each species, we focus on identifying what characterizes each particular GraSpring (i.e., $\widetilde{A}_{G_0}$-$\widetilde{A}_{G_8}$) from a biological perspective, just as we did previously for our GraCoal embeddings. To this end, in table 4.7 we present the same format as previously presented for GraCoals: we report for the budding yeast, the number of uniquely enriched annotations (column 1), the mean SS for the uniquely enriched annotations (column 2) as well as the mean SS for the top 10 largest uniquely enriched annotations (column 3) for each GraCoal used in SAFE. Finally, in column 5, we report the names of the top 10 uniquely enriched annotations and their corresponding size in terms of enriched neighborhoods (column 4).

In brief, we first identify the uniquely enriched annotations for each GraSpring (i.e., annotations enriched in a particular GraSpring that are not enriched in any of the other GraSprings). For the budding yeast, we find that on average, 29 GO-BPs are uniquely enriched for each particular GraSpring (Table 4.7), which is more than we can uncover with GraCoals (Table 4.3). Similar to our GraCoal embedding findings, we observe that each GraSpring embedding (except for $GraSpring_4$, uncovers biological information that is not uncovered by the other GraSpring embeddings, which is again in line with the literature, as different graphlet adjacencies are known to capture complementary topology-function relationships in molecular networks (Windels et al., 2019). Aditionally, there is some degree of functional similarity in the information uncovered, as shown by the mean semantic similarity of the set of enriched annotations that each GraSpring uncovers (Table 4.7, column 2). Finally, we rank the uniquely enriched annotations according to their size, defined as the total number of neighborhoods they are enriched in, as a measure of how well they are captured by each particular GraSpring. The lowest and maximum average SS for the sets of uniquely enriched annotations for the budding yeast GI network are 0.11 (Std=0.02) and 0.62 (Std=0.01) for $\widetilde{A}_{G_8}$ and $\widetilde{A}_{G_5}$, respectively. Finally, when evaluating the top 10 enriched annotations when possible, we observe a larger degree of functional similarity between the annotations, which ranges from 0.23 (Std=0.05) to 0.62 (Std=0.01) for the budding yeast ($\widetilde{A}_{G_2}$, and $\widetilde{A}_{G_5}$, respectively). Moreover, the average size of the uniquely enriched GO-BPs uncovered by GraSprings is 149.8 (std=84.1), while the average size of the uniquely enriched GO-BPs for GraCoals is 129.7 (std=54.9). This implies that the biological information that is uniquely captured by the different GraSpring embeddings tends to be more generic than the biological information that is uniquely captured by the different GraCoals in budding yeast. For the summary of uniquely enriched annotations for our other GI networks please refer to Tables A.6 to A.8 in Appendix A.

$\widetilde{A}_{G_i}$	**Total annotations**	**Mean SS**	**Mean SS Top 10**	**EN**	**Annotation**
$\widetilde{A}_{G_0}$	1	1.00 (std=nan)	1.00 (std=nan)	172.0	protein insertion into ER membrane
$\widetilde{A}_{G_1}$	8	0.26 (std=0.06)	0.26 (std=0.06)	152.0	gene conversion
$\widetilde{A}_{G_1}$	8	0.26 (std=0.06)	0.26 (std=0.06)	91.0	maturation of 5.8S rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)
$\widetilde{A}_{G_1}$	8	0.26 (std=0.06)	0.26 (std=0.06)	86.0	maturation of 5.8S rRNA
$\widetilde{A}_{G_1}$	8	0.26 (std=0.06)	0.26 (std=0.06)	67.0	lipid droplet organization
$\widetilde{A}_{G_1}$	8	0.26 (std=0.06)	0.26 (std=0.06)	44.0	thioester biosynthetic process
$\widetilde{A}_{G_1}$	8	0.26 (std=0.06)	0.26 (std=0.06)	44.0	acyl-CoA biosynthetic process
$\widetilde{A}_{G_1}$	8	0.26 (std=0.06)	0.26 (std=0.06)	44.0	acetyl-CoA biosynthetic process
$\widetilde{A}_{G_1}$	8	0.26 (std=0.06)	0.26 (std=0.06)	33.0	nuclear mRNA surveillance
$\widetilde{A}_{G_2}$	92	0.14 (std=0.02)	0.23 (std=0.05)	258.0	regulation of microtubule cytoskeleton organization
$\widetilde{A}_{G_2}$	92	0.14 (std=0.02)	0.23 (std=0.05)	254.0	regulation of microtubule-based process
$\widetilde{A}_{G_2}$	92	0.14 (std=0.02)	0.23 (std=0.05)	210.0	sno(s)RNA metabolic process
$\widetilde{A}_{G_2}$	92	0.14 (std=0.02)	0.23 (std=0.05)	203.0	mitotic spindle checkpoint signaling
$\widetilde{A}_{G_2}$	92	0.14 (std=0.02)	0.23 (std=0.05)	203.0	spindle checkpoint signaling
$\widetilde{A}_{G_2}$	92	0.14 (std=0.02)	0.23 (std=0.05)	199.0	sno(s)RNA processing
$\widetilde{A}_{G_2}$	92	0.14 (std=0.02)	0.23 (std=0.05)	194.0	DNA conformation change
$\widetilde{A}_{G_2}$	92	0.14 (std=0.02)	0.23 (std=0.05)	185.0	regulation of exit from mitosis
$\widetilde{A}_{G_2}$	92	0.14 (std=0.02)	0.23 (std=0.05)	178.0	chromosome segregation
$\widetilde{A}_{G_2}$	92	0.14 (std=0.02)	0.23 (std=0.05)	169.0	establishment of mitotic spindle localization
$\widetilde{A}_{G_3}$	17	0.23 (std=0.06)	0.25 (std=0.05)	333.0	cellular process
$\widetilde{A}_{G_3}$	17	0.23 (std=0.06)	0.25 (std=0.05)	299.0	regulation of telomere maintenance
$\widetilde{A}_{G_3}$	17	0.23 (std=0.06)	0.25 (std=0.05)	220.0	ribophagy
$\widetilde{A}_{G_3}$	17	0.23 (std=0.06)	0.25 (std=0.05)	189.0	positive regulation of glucose metabolic process
$\widetilde{A}_{G_3}$	17	0.23 (std=0.06)	0.25 (std=0.05)	189.0	positive regulation of cellular carbohydrate metabolic process
$\widetilde{A}_{G_3}$	17	0.23 (std=0.06)	0.25 (std=0.05)	169.0	regulation of cell cycle G2/M phase transition
$\widetilde{A}_{G_3}$	17	0.23 (std=0.06)	0.25 (std=0.05)	169.0	regulation of G2/M transition of mitotic cell cycle
$\widetilde{A}_{G_3}$	17	0.23 (std=0.06)	0.25 (std=0.05)	139.0	lipid translocation
$\widetilde{A}_{G_3}$	17	0.23 (std=0.06)	0.25 (std=0.05)	122.0	regulation of membrane lipid distribution
$\widetilde{A}_{G_3}$	17	0.23 (std=0.06)	0.25 (std=0.05)	111.0	DNA double-strand break processing
$\widetilde{A}_{G_5}$	3	0.62 (std=0.01)	0.62 (std=0.01)	163.0	nucleotide-excision repair
$\widetilde{A}_{G_5}$	3	0.62 (std=0.01)	0.62 (std=0.01)	72.0	protein maturation
$\widetilde{A}_{G_5}$	3	0.62 (std=0.01)	0.62 (std=0.01)	44.0	protein processing

Table 4.7. Summary of uniquely enriched GO-BPs for GraSpring embeddings, Part 1. We report, for the Budding yeast GI network, the number of uniquely enriched GO-BPs and the mean semantic similarity (SS) between the uniquely enriched annotations (GO-BPs) (columns 2 and 3) for GraCoals based on all graphlet adjacencies for up to four node graphlets, i.e. $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$ (column 1). In column 4, we report the mean SS for the top ten largest enriched annotations (column 6), i.e., ranking them in descending order according to the number of neighborhoods that the annotations are enriched in (column 5).

$\widetilde{A}_{G_i}$	**Total annotations**	**Mean SS**	**Mean SS Top 10**	**EN**	**Annotation**
$\widetilde{A}_{G_6}$	13	0.52 (std=0.04)	0.54 (std=0.04)	43.0	phytosteroid metabolic process
$\widetilde{A}_{G_6}$	13	0.52 (std=0.04)	0.54 (std=0.04)	43.0	ergosterol metabolic process
$\widetilde{A}_{G_6}$	13	0.52 (std=0.04)	0.54 (std=0.04)	43.0	cellular alcohol metabolic process
$\widetilde{A}_{G_6}$	13	0.52 (std=0.04)	0.54 (std=0.04)	39.0	secondary alcohol metabolic process
$\widetilde{A}_{G_6}$	13	0.52 (std=0.04)	0.54 (std=0.04)	28.0	cellular alcohol biosynthetic process
$\widetilde{A}_{G_6}$	13	0.52 (std=0.04)	0.54 (std=0.04)	28.0	cellular lipid biosynthetic process
$\widetilde{A}_{G_6}$	13	0.52 (std=0.04)	0.54 (std=0.04)	28.0	phytosteroid biosynthetic process
$\widetilde{A}_{G_6}$	13	0.52 (std=0.04)	0.54 (std=0.04)	28.0	ergosterol biosynthetic process
$\widetilde{A}_{G_6}$	13	0.52 (std=0.04)	0.54 (std=0.04)	24.0	secondary alcohol biosynthetic process
$\widetilde{A}_{G_6}$	13	0.52 (std=0.04)	0.54 (std=0.04)	18.0	sterol biosynthetic process
$\widetilde{A}_{G_7}$	15	0.26 (std=0.05)	0.27 (std=0.05)	228.0	mRNA polyadenylation
$\widetilde{A}_{G_7}$	15	0.26 (std=0.05)	0.27 (std=0.05)	210.0	pre-mRNA cleavage required for polyadenylation
$\widetilde{A}_{G_7}$	15	0.26 (std=0.05)	0.27 (std=0.05)	207.0	RNA polyadenylation
$\widetilde{A}_{G_7}$	15	0.26 (std=0.05)	0.27 (std=0.05)	206.0	mRNA cleavage involved in mRNA processing
$\widetilde{A}_{G_7}$	15	0.26 (std=0.05)	0.27 (std=0.05)	203.0	mRNA cleavage
$\widetilde{A}_{G_7}$	15	0.26 (std=0.05)	0.27 (std=0.05)	181.0	protein lipidation
$\widetilde{A}_{G_7}$	15	0.26 (std=0.05)	0.27 (std=0.05)	155.0	carbohydrate derivative biosynthetic process
$\widetilde{A}_{G_7}$	15	0.26 (std=0.05)	0.27 (std=0.05)	108.0	organonitrogen compound catabolic process
$\widetilde{A}_{G_7}$	15	0.26 (std=0.05)	0.27 (std=0.05)	86.0	protein-containing complex localization
$\widetilde{A}_{G_7}$	15	0.26 (std=0.05)	0.27 (std=0.05)	47.0	endonucleolytic cleavage in ITS1 to separate SSU-rRNA from 5.8S rRNA and LSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)
$\widetilde{A}_{G_8}$	120	0.11 (std=0.02)	0.28 (std=0.04)	263.0	protein modification by small protein conjugation
$\widetilde{A}_{G_8}$	120	0.11 (std=0.02)	0.28 (std=0.04)	254.0	protein acetylation
$\widetilde{A}_{G_8}$	120	0.11 (std=0.02)	0.28 (std=0.04)	234.0	actin cytoskeleton organization
$\widetilde{A}_{G_8}$	120	0.11 (std=0.02)	0.28 (std=0.04)	231.0	protein acylation
$\widetilde{A}_{G_8}$	120	0.11 (std=0.02)	0.28 (std=0.04)	229.0	tubulin complex assembly
$\widetilde{A}_{G_8}$	120	0.11 (std=0.02)	0.28 (std=0.04)	228.0	protein ubiquitination
$\widetilde{A}_{G_8}$	120	0.11 (std=0.02)	0.28 (std=0.04)	225.0	protein methylation
$\widetilde{A}_{G_8}$	120	0.11 (std=0.02)	0.28 (std=0.04)	225.0	protein alkylation
$\widetilde{A}_{G_8}$	120	0.11 (std=0.02)	0.28 (std=0.04)	222.0	negative regulation of cellular protein metabolic process
$\widetilde{A}_{G_8}$	120	0.11 (std=0.02)	0.28 (std=0.04)	221.0	mitotic DNA damage checkpoint signaling

Table 4.7. Summary of uniquely enriched GO-BPs for GraSpring embeddings, Part 2. We report, for the Budding yeast GI network, the number of uniquely enriched GO-BPs and the mean semantic similarity (SS) between the uniquely enriched annotations (GO-BPs) (columns 2 and 3) for GraCoals based on all graphlet adjacencies for up to four node graphlets, i.e. $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$ (column 1). In column 4, we report the mean SS for the top ten largest enriched annotations (column 6), i.e., ranking them in descending order according to the number of neighborhoods that the annotations are enriched in (column 5).

Genes enriched in *E. coli*, Fission yeast and Budding yeast GI networks cover more paralogous genes than the Fruit fly GI network

In our analysis of GraCoal embeddings for the GI networks in the previous section, we discussed how GraCoals, in particular $GraCoal_{2,7,8}$ (i.e., based on triangle topology), lead to higher enrichments because of the presence of paralogs in the networks. Below we report gene enrichment and paralog enrichment statistics for the GI networks in the same format as before: total number of genes that have at least one GO-BP enriched in the neighborhood (i.e., "Enriched genes" in Table 4.8) and the number of genes that are enriched and are paralogs (i.e., "Paralogs" in Table 4.8). Our observations for the fruit fly are consistent with previous results as it achieves the lowest percentages of paralogs enriched across all GraSprings, which is easily explained by the fact that it has the least amount of paralogs in the GI network. On the other hand, the percentages of paralogs enriched for the budding yeast, *E. coli* and fission yeast tend to be higher in GraSprings corresponding to the triangle topology. For instance, for budding yeast and fission yeast, $GraSpring_2$ and $GraSpring_7$ achieve the highest and second highest, respectively, percentages of enriched paralogs. For *E. coli*, even though the percentages of enriched paralogs are very close in all GraSprings, $GraSpring_2$ achieves the highgest value. This further validates our previous observations that when there are lots of duplicated genes (i.e., paralogs) in a given GI network, the enrichments tend to be the best when based on triangle topology (e.g., $\widetilde{A}_2$, $\widetilde{A}_7$ or $\widetilde{A}_8$). Finally, we observe that these results for GraSprings based on triangle topology, $GraSpring_{2,7,8}$ are of lower performance than the ones for $GraCoal_{2,7,8}$, except for $GraSpring_2$ for *E. coli*, which further validates that GraCoals are a better approach for uncovering biological information from GI networks that contain many paralogs.

	Budding yeast		***E. coli***		**Fission yeast**		**Fruit fly**	
	Enriched Genes	Paralogs	Enriched Genes	Paralogs	Enriched Genes	Paralogs	Enriched Genes	Paralogs
$\widetilde{A}_{G_0}$	1,652	11.38	1,206	22.69	402	6.47	1,575	7.68
$\widetilde{A}_{G_1}$	1,610	12.52	1,170	22.60	342	9.36	1,282	7.18
$\widetilde{A}_{G_2}$	1,643	18.32	1,108	27.41	517	10.67	1,101	7.81
$\widetilde{A}_{G_3}$	1,378	14.13	1,092	26.74	586	9.41	1,103	7.71
$\widetilde{A}_{G_4}$	1,299	11.55	1,049	25.93	255	5.88	1,185	6.67
$\widetilde{A}_{G_5}$	1,513	13.66	1,095	26.12	388	3.09	834	7.31
$\widetilde{A}_{G_6}$	1,622	16.46	1,059	27.20	505	9.31	1,299	7.39
$\widetilde{A}_{G_7}$	1,693	18.19	1,011	21.27	471	9.55	1,031	7.76
$\widetilde{A}_{G_8}$	1,791	13.70	911	15.30	491	8.76	576	8.33

Table 4.8. Statistics for paralogous genes enriched using SAFE with GraSpring embeddings. For each of the four GI networks (Budding yeast, *E. coli*, Fission yeast and Fruit fly), we show the number of enriched genes when using SAFE with GraSpring embeddings (i.e., genes that have at least one annotation enriched in their neighborhood) and the percentages of genes enriched that are paralogs.

Biological insights of GraSprings at a functional domain level

Lastly, we aim to give insight into the biological function captured by our GraSpring embeddings across species at a functional domain level just as we did for our GraCoal embeddings in previous sections. To this end, we identify the most characteristic functional domains in each species, i.e., the domains that could not be captured by any of the other GraSprings. In brief, we measure the uniqueness of all functional domains obtained with SAFE by computing the Jaccard similarity index (JI) between the sets of enriched annotations of each functional domain in a particular GraSpring and the sets of enriched annotations of every other functional domain in the other GraSpring embeddings. For each domain, we report the maximum JI, which represents the maximum overlap to any other functional domain in the other GraSpring embeddings, and thus reflects how unique the functional domain is to its corresponding GraSpring. Finally, we also report the paralog ratio, which we already defined, as a way to evaluate how well a given functional domain is capturing biological information that involves paralogs.

In Table 4.9 for budding yeast, and Tables A.15 to A.17 in Appendix A for *E. coli*, fission yeast and fruit fly, respectively, we summarize, the number of functional domains (column 1) and the mean paralog ratio (column 2), over each GraSpring embedding (i.e., $\widetilde{A}_{G_0}$-$\widetilde{A}_{G_8}$). Furthermore, we report the top three most characteristic functional domains for the GI molecular networks (column 5) according to the lowest maximum JI (column 4) and the corresponding paralog ratio.

Our first observation is that across all of our GI networks, there are less domains that are highly characteristic to a particular GraSpring. For instance, for budding yeast (Table 4.9), only 5 functional domains are completely unique with a Max JI of 0, while for GraCoals for budding yeast, a total of 14 unique functional domains could be achieved (Table 4.6). Secondly, on average, fewer functional domains can be achieved with GraSprings with respect to GraCoals. This means that GraCoals, with budding yeast at least, uncover biological information at a functional domain level that is more functionally organised (i.e., more functional domains than with GraSprings) and that is less redundant (i.e., there is less overlap between the functional domains of each GraSpring), in addition to uncovering more functional information overall (as seen by the overall enrichments in Figure 4.1).

In general, we observe lower paralog ratios in GraSprings when compared to GraCoals. For instance, the lowest and highest average paralog ratio previously reported for our GraCoals for the budding yeast is 0.15 (std=0.06) and 0.22 (std=0.06), respectively. On the other hand, with GraSprings we obtain paralog ratios than range from 0.12 (std=0.07) and 0.20 (std=0.06). Additionally, the largest paralog ratio obtained from the top three most characteristic functional domains across all GraSprings, the largest value we obtain is 0.32 with a Max JI of 0.0. Interestingly, this functional domain is uncovered by $GraSpring_2$, which is consistent with our previous results that indicate that triangle topology is the best for capturing biological function that involves paralogs. Similarly, the most characteristic functional domain of $GraSpring_8$, with a Max JI of 0.0, also has one of the highest paralog ratios (0.22).

In conclusion, we have shown that information captured by GraSprings is less functionally organised and thus leads to fewer functional domains uncovered with SAFE. As a consequence, there is more overlap between the different GraSprings, as shown by the maximum JIs in table 4.9. Finally, even when performing worse than GraCoals, GraSprings corresponding to triangle topology, for instance $GraSpring_{2,8}$, still uncover biological functions that involve many paralogs, as shown by their paralog ratios.

$\widetilde{A}_{G_i}$	Num functional domains	Mean paralog ratio	Domain paralog ratio	Domain max JI	Domain description
$\widetilde{A}_{G_0}$	12	0.19 (std=0.05)	0.09	0.09	process, metabolic, biosynthetic, membrane, lipid
$\widetilde{A}_{G_0}$	12	0.19 (std=0.05)	0.20	0.12	lipid, process, metabolic, cellular, biosynthetic
$\widetilde{A}_{G_0}$	12	0.19 (std=0.05)	0.23	0.18	aerobic, respiration, generation, precursor, metabolites
$\widetilde{A}_{G_1}$	16	0.16 (std=0.06)	0.23	0.17	biosynthetic, process, CoA, thioester, acetyl
$\widetilde{A}_{G_1}$	16	0.16 (std=0.06)	0.09	0.24	DNA, regulation, replication, heterochromatin, assembly
$\widetilde{A}_{G_1}$	16	0.16 (std=0.06)	0.21	0.27	capping, RNA, 7, methylguanosine, actin
$\widetilde{A}_{G_2}$	14	0.15 (std=0.09)	0.32	0.00	regulation, kinase, activity, protein, G1
$\widetilde{A}_{G_2}$	14	0.15 (std=0.09)	0.03	0.01	regulation, mitotic, negative, sister, chromatid
$\widetilde{A}_{G_2}$	14	0.15 (std=0.09)	0.19	0.01	regulation, cell, communication, signal, transduction
$\widetilde{A}_{G_3}$	7	0.17 (std=0.05)	0.22	0.00	pH, regulation, monovalent, inorganic, cation
$\widetilde{A}_{G_3}$	7	0.17 (std=0.05)	0.13	0.41	process, catabolic, protein, G2, M
$\widetilde{A}_{G_3}$	7	0.17 (std=0.05)	0.14	0.53	regulation, process, positive, metabolic, cellular
$\widetilde{A}_{G_4}$	7	0.20 (std=0.06)	0.08	0.19	DNA, maintenance, repair, checkpoint, signaling
$\widetilde{A}_{G_4}$	7	0.20 (std=0.06)	0.24	0.39	regulation, assembly, positive, complex, organization
$\widetilde{A}_{G_4}$	7	0.20 (std=0.06)	0.24	0.41	catabolic, process, protein, dependent, macromolecule
$\widetilde{A}_{G_5}$	11	0.14 (std=0.07)	0.16	0.32	protein, process, transport, localization, Golgi
$\widetilde{A}_{G_5}$	11	0.14 (std=0.07)	0.15	0.33	process, regulation, DNA, protein, metabolic
$\widetilde{A}_{G_5}$	11	0.14 (std=0.07)	0.08	0.40	membrane, protein, tethering, processing, organelle
$\widetilde{A}_{G_6}$	11	0.14 (std=0.06)	0.19	0.00	process, biosynthetic, metabolic, alcohol, cellular
$\widetilde{A}_{G_6}$	11	0.14 (std=0.06)	0.08	0.33	fusion, vesicle, Golgi, membrane, mediated
$\widetilde{A}_{G_6}$	11	0.14 (std=0.06)	0.20	0.53	regulation, process, protein, assembly, transcription
$\widetilde{A}_{G_7}$	10	0.15 (std=0.06)	0.08	0.08	rRNA, endonucleolytic, cleavage, SSU, 5'
$\widetilde{A}_{G_7}$	10	0.15 (std=0.06)	0.14	0.12	regulation, signaling, positive, TORC1, TOR
$\widetilde{A}_{G_7}$	10	0.15 (std=0.06)	0.04	0.50	mitochondrion, organization
$\widetilde{A}_{G_8}$	13	0.12 (std=0.07)	0.22	0.00	process, purine, ribonucleotide, metabolic, biosynthetic
$\widetilde{A}_{G_8}$	13	0.12 (std=0.07)	0.01	0.00	regulation, actin, filament, negative, depolymerization
$\widetilde{A}_{G_8}$	13	0.12 (std=0.07)	0.14	0.09	response, process, stimulus, cellular, biosynthetic

Table 4.9. Summary of most unique functional domains for GraSpring embeddings. We report for the budding yeast GI network, for each GraSpring embedding used with SAFE (column 1), i.e., based on graphlet adjacencies for up to four node graphlets ($\widetilde{A}_{G0}$-$\widetilde{A}_{G8}$), the number of functional domains (column 2) the mean paralog ratio (column 3) and the top three most characteristic functional domains (column 6). Lastly, for each functional domain we report the paralog ratio (column 4) and the maximum Jaccard similarity index (JI) (column 5).

4.3 Results for graphlet Spectral with SAFE

In this section we present our results for the different graphlet Spectral embeddings over the GI networks.

Graphlet Spectral uncovers the functional organization of the cell

In this section, we evaluate: (1) how well graphlet Spectral embeddings capture the functional organization of genetic interaction networks and (2) which higher-order topologies (i.e., graphlets) capture the most function. In both experiments, we apply SAFE-based enrichment analysis to quantify how well a given embedding captures the functional organization of a given network.

To evaluate how well graphlet Spectral capture the functional organization of our GI networks, we compare against Spectral embedding (as graphlet Spectral for graphlet G_0, which corresponds to standard Spectral embedding), GraCoal embedding and Graspring embedding. In Figure 4.1 we show the union of the enriched genes and the union of the enriched GO-BPs across the different types of underlying graphlet adjacencies (i.e., $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$). When comparing the three graphlet based embeddings, we have already established that GraCoal embeddings outperform both GraSpring and graphlet Spectral embeddings when uncovering the functional organization of GI networks using SAFE. However, graphlet Spectral embeddings outperforms GraSprings most of the time, both in terms of genes enriched and GO-BPs enriched. For instance, for the fruit fly GI network, the union of genes enriched in GO-BPs in terms of percentage (Figure 4.1, left), from highest to lowest (i.e., best to worst) are 90.3%, 76.72% and 61.3% for GraCoal, graphlet Spectral and GraSpring, respectively. Similarly, for budding yeast, these values are 71.4%, 61.13% and 42.6% for GraCoal, graphlet Spectral and GraSpring, respectively. This is also the case when considering a different annotation type such as GO-CC, except for the union of enriched GO-CCs for *E. coli*, where GraSpring outperforms both GraCoal and graphlet Spectral (Figure A.10). Finally, for our third annotation type, GO-MF, graphlet Spectral embeddings outperform again GraSpring embeddings, except for the *E. coli* GI network, where GraSpring works best both in terms of the union of genes enriched in GO-MF and the union of enriched GO-MFs (Figure A.17 in Appendix A).

Graphlet Spectral embeddings uncover complementary biological information

Similar to GraCoal embeddings and GraSpring embeddings, we can use graphlet Spectral embeddings to uncover biological information in complementary ways. In Figure 4.9 we show the percentages of genes enriched (top) and percentages of GO-BPs enriched (bottom) for all graphlet Spectral embeddings (i.e., $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$) across our four GI networks. We observe the same pattern as with GraCoals or GraSprings, that is, the largest percentages of both genes enriched and annotations enriched are obtained for the fruit fly GI network. On the other hand, the lowest percentages of genes enriched and lowest percentages of annotations enriched when using graphlet Spectral embeddings are obtained for the fission yeast, although *E. coli* is not far behind in terms of enriched annotations. Interestingly, the second best enrichments in terms of genes are achieved by *E. coli*, and by the budding yeast in terms of annotations enriched. For GO-CC and GO-MF we present similar results in Figures A.13 and A.20 in Appendix A.

Next, we investigate which topologies capture the most function in GI networks between different graphlet Spectral embeddings. We observe that in terms of percentages of genes enriched, the top performing graphlet Spectrals for fruit fly, $Spectral_{0,1,3,4,6}$, achieve between 40.3% and 49.6%, while the low performing, $Spectral_{2,5,7,8}$ achieve between 6.2% and 24.1%. This is also consistent for the percentages of enriched GO-BPs. For the second best in terms of genes enriched, *E.*

coli, the top performing graphlet Spectrals, $Spectral_{0,3,4}$ all achieve around 24%, while the low performing graphlet Spectrals all achieve less than 20 with $Spectral_8$ performing the worst at 6.8%. Similarly, for budding yeast, only $Spectral_3$ achieves more than 20% (24.3%), while the lowest, $Spectral_4$, achieves only 2%. This is consistent for enriched annotations, with $Spectral_3$ and $Spectral_4$ performing the best and worst, respectively. Finally, for fission yeast, the enrichments over all species are in general very low. However, we observe that $Spectral_4$ is clearly the top performing both in terms of genes enriched (16.8%) and annotations enriched (18.3%). On the other hand, $Spectral_{2,3,7,8}$ are the worst performers, achieving between 1.7% and 2.5% enriched genes and between 3.7% and 6.8% enriched annotations. In all, when comparing between species, the results for top and low performing graphlet Spectrals are not always the same, though some overlap exist, which is consistent with our previous results for GraCoal embeddings and GraSpring embeddings across the GI networks. This implies that the same graphlet Spectrals capture different topology-function relationships in the GI networks, depending on the species.

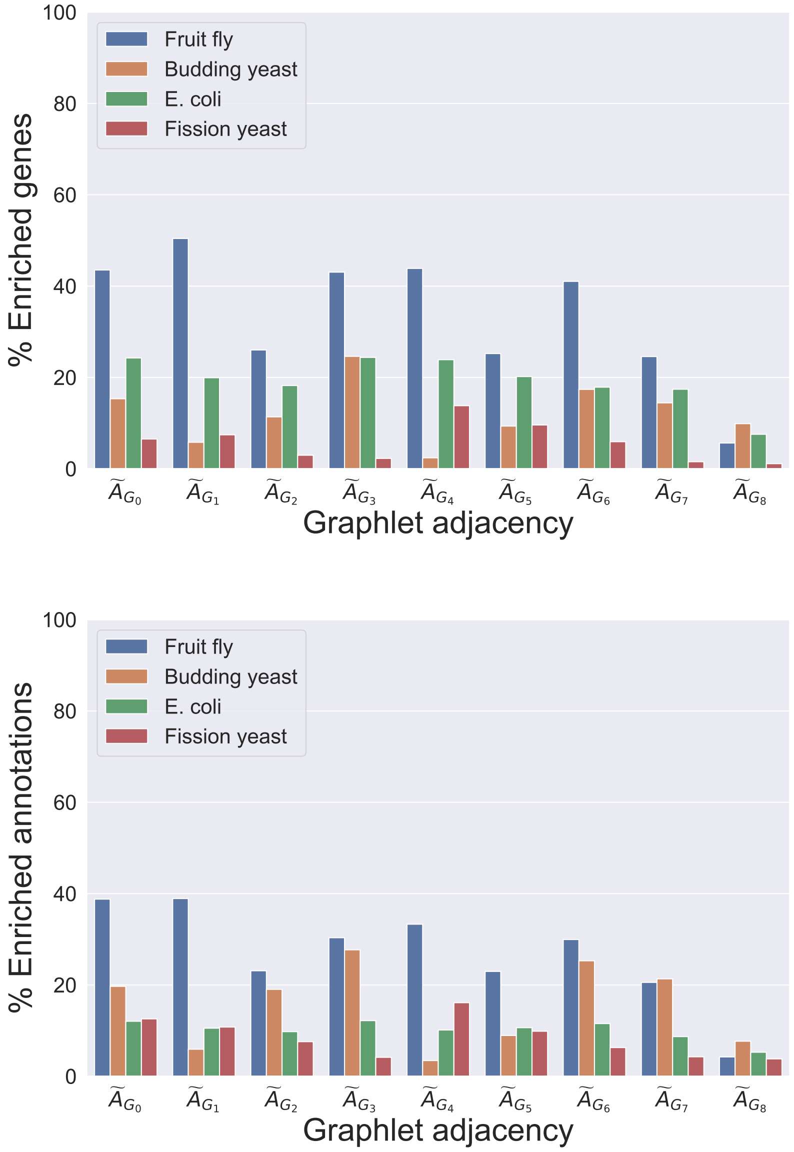


Figure 4.9. SAFE GO-BP enrichment analysis comparing graphlet based Spectral in GI networks. For the GI networks of our four species (legend), we show, on the y-axis, the percentage of enriched genes (top) and the percentage of enriched annotations (top) for each of the different Spectral embeddings (x-axis).

Additionally, for each species, we focus on identifying what characterizes each particular graphlet Spectral (i.e., $\widetilde{A}_{G_0}$-$\widetilde{A}_{G_8}$) from a biological perspective, just as we did previously for our GraCoal embeddings and GraSpring embeddings. To this end, in table 4.10 we present the same format as previously presented for GraCoals: we report for the budding yeast, the number of uniquely enriched annotations (column 1), the mean SS for the uniquely enriched annotations (column 2) as well as the mean SS for the top 10 largest uniquely enriched annotations (column 3) for each graphlet Spectral used in SAFE. Finally, in column 5, we report the names of the top 10 uniquely enriched annotations and their corresponding size in terms of enriched neighborhoods (column 4).

In brief, we first identify the uniquely enriched annotations for each graphlet

Spectral (i.e., annotations enriched in a particular graphlet Spectral that are not enriched in any of the other graphlet Spectrals). For the budding yeast, we find that on average, 18 GO-BPs are uniquely enriched for each particular graphlet Spectral (Table 4.10), which is less than we can uncover with GraCoals (average 22 GO-BPs, Table 4.3) or GraSprings (average 29 GO-BPs, Table 4.7. We observe uniquely enriched GO-BPs in every graphlet Spectral except for $Spectral_4$, which is consistent with previous results for the other graphlet based embeddings, i.e., different graphlet adjacencies are known to capture complementary topology-function relationships in molecular networks (Windels et al., 2019). Interestingly, $GraSpring_4$ is the only GraSpring that could not uncover any uniquely enriched GO-BPs in the budding yeast GI. For our other species, $GraSpring_4$ also fails to uncover unique information for the fission yeast. This could indicate that these graphlet based embeddings are not suitable for uncovering unique information from the star topology.

Additionally, the mean semantic similarities reported also indicate some degree of functional similarity in the information uncovered by each graphlet Spectral (Table 4.10, column 2). For instance, the lowest and maximum average SS for the sets of uniquely enriched annotations of graphlet Spectrals for the budding yeast GI network are 0.14 (Std=0.03) and 0.63 (Std=0.01) for $\widetilde{A}_{G_3}$ and $\widetilde{A}_{G_8}$, respectively. Finally, when evaluating the top 10 enriched annotations when possible, we observe a larger degree of functional similarity between the annotations, which ranges from 0.24 (Std=0.05) to 0.63 (Std=0.01) for the budding yeast ($Spectral_{0,5}$, and $Spectral_8$, respectively). Next, we rank the uniquely enriched annotations according to their size, defined as the total number of neighborhoods they are enriched in, as a measure of how well they are captured by each particular graphlet Spectral. The average size of the uniquely enriched GO-BPs uncovered by graphlet Spectrals is 38.6 (std=29.6), while the average size of the uniquely enriched GO-BPs for GraCoals and GraSprings are 129.7 (std=54.9 and 149 (std=84.1), respectively. This implies that the biological information that is uniquely captured by the different graphlet Spectral embeddings could correspond to more specific biological functions as opposed to what is uncovered by both GraCoals and GraSprings. to be more generic than the biological information that is uniquely captured by the different GraCoals in budding yeast. For the summary of uniquely enriched annotations for our other GI networks please refer to Tables A.9 to A.11 in Appendix A.

$\widetilde{A}_{G_i}$	**Total annotations**	**Mean SS**	**Mean SS Top 10**	**EN**	**Annotation**
$\widetilde{A}_{G_0}$	44	0.17 (std=0.05)	0.24 (std=0.06)	86.0	negative regulation of biological process
$\widetilde{A}_{G_0}$	44	0.17 (std=0.05)	0.24 (std=0.06)	82.0	negative regulation of macromolecule metabolic process
$\widetilde{A}_{G_0}$	44	0.17 (std=0.05)	0.24 (std=0.06)	74.0	cytoplasm to vacuole transport by the Cvt pathway
$\widetilde{A}_{G_0}$	44	0.17 (std=0.05)	0.24 (std=0.06)	70.0	negative regulation of metabolic process
$\widetilde{A}_{G_0}$	44	0.17 (std=0.05)	0.24 (std=0.06)	50.0	maintenance of DNA trinucleotide repeats
$\widetilde{A}_{G_0}$	44	0.17 (std=0.05)	0.24 (std=0.06)	44.0	organelle organization
$\widetilde{A}_{G_0}$	44	0.17 (std=0.05)	0.24 (std=0.06)	39.0	septin ring organization
$\widetilde{A}_{G_0}$	44	0.17 (std=0.05)	0.24 (std=0.06)	37.0	deoxyribonucleoside triphosphate biosynthetic process
$\widetilde{A}_{G_0}$	44	0.17 (std=0.05)	0.24 (std=0.06)	36.0	positive regulation of RNA polymerase II transcription preinitiation complex assembly
$\widetilde{A}_{G_0}$	44	0.17 (std=0.05)	0.24 (std=0.06)	36.0	positive regulation of transcription initiation from RNA polymerase II promoter
$\widetilde{A}_{G_1}$	8	0.38 (std=0.06)	0.38 (std=0.06)	14.0	homeostatic process
$\widetilde{A}_{G_1}$	8	0.38 (std=0.06)	0.38 (std=0.06)	13.0	cellular homeostasis
$\widetilde{A}_{G_1}$	8	0.38 (std=0.06)	0.38 (std=0.06)	13.0	chemical homeostasis
$\widetilde{A}_{G_1}$	8	0.38 (std=0.06)	0.38 (std=0.06)	11.0	negative regulation of chromosome organization
$\widetilde{A}_{G_1}$	8	0.38 (std=0.06)	0.38 (std=0.06)	10.0	regulation of chromosome separation
$\widetilde{A}_{G_1}$	8	0.38 (std=0.06)	0.38 (std=0.06)	10.0	regulation of mitotic sister chromatid separation
$\widetilde{A}_{G_1}$	8	0.38 (std=0.06)	0.38 (std=0.06)	10.0	regulation of sister chromatid segregation
$\widetilde{A}_{G_1}$	8	0.38 (std=0.06)	0.38 (std=0.06)	10.0	regulation of chromosome segregation
$\widetilde{A}_{G_2}$	9	0.34 (std=0.05)	0.34 (std=0.05)	15.0	autophagy of nucleus
$\widetilde{A}_{G_2}$	9	0.34 (std=0.05)	0.34 (std=0.05)	14.0	autophagy of mitochondrion
$\widetilde{A}_{G_2}$	9	0.34 (std=0.05)	0.34 (std=0.05)	14.0	mitochondrion disassembly
$\widetilde{A}_{G_2}$	9	0.34 (std=0.05)	0.34 (std=0.05)	14.0	piecemeal microautophagy of the nucleus
$\widetilde{A}_{G_2}$	9	0.34 (std=0.05)	0.34 (std=0.05)	12.0	organelle disassembly
$\widetilde{A}_{G_2}$	9	0.34 (std=0.05)	0.34 (std=0.05)	11.0	RNA splicing, via transesterification reactions
$\widetilde{A}_{G_2}$	9	0.34 (std=0.05)	0.34 (std=0.05)	11.0	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile
$\widetilde{A}_{G_2}$	9	0.34 (std=0.05)	0.34 (std=0.05)	11.0	mRNA splicing, via spliceosome
$\widetilde{A}_{G_2}$	9	0.34 (std=0.05)	0.34 (std=0.05)	10.0	autophagosome organization
$\widetilde{A}_{G_3}$	76	0.14 (std=0.03)	0.33 (std=0.05)	61.0	chromatin assembly
$\widetilde{A}_{G_3}$	76	0.14 (std=0.03)	0.33 (std=0.05)	56.0	positive regulation of macromolecule metabolic process
$\widetilde{A}_{G_3}$	76	0.14 (std=0.03)	0.33 (std=0.05)	56.0	positive regulation of biosynthetic process
$\widetilde{A}_{G_3}$	76	0.14 (std=0.03)	0.33 (std=0.05)	56.0	positive regulation of cellular biosynthetic process
$\widetilde{A}_{G_3}$	76	0.14 (std=0.03)	0.33 (std=0.05)	54.0	regulation of gene expression, epigenetic

Table 4.10. Summary of uniquely enriched GO-BPs for graphlet Spectral embeddings, Part 1. We report, for the Budding yeast GI network, the number of uniquely enriched GO-BPs and the mean semantic similarity (SS) between the uniquely enriched annotations (GO-BPs) (columns 2 and 3) for graphlet Spectral embeddings based on all graphlet adjacencies for up to four node graphlets, i.e. $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$ (column 1). In column 4, we report the mean SS for the top ten largest enriched annotations (column 6), i.e., ranking them in descending order according to the number of neighborhoods that the annotations are enriched in (column 5).

$\widetilde{A}_{G_i}$	**Total annotations**	**Mean SS**	**Mean SS Top 10**	**EN**	**Annotation**
$\widetilde{A}_{G_3}$	76	0.14 (std=0.03)	0.33 (std=0.05)	54.0	heterochromatin assembly
$\widetilde{A}_{G_3}$	76	0.14 (std=0.03)	0.33 (std=0.05)	54.0	negative regulation of gene expression, epigenetic
$\widetilde{A}_{G_3}$	76	0.14 (std=0.03)	0.33 (std=0.05)	54.0	heterochromatin organization
$\widetilde{A}_{G_3}$	76	0.14 (std=0.03)	0.33 (std=0.05)	53.0	mitochondrion organization
$\widetilde{A}_{G_3}$	76	0.14 (std=0.03)	0.33 (std=0.05)	45.0	positive regulation of biological process
$\widetilde{A}_{G_5}$	10	0.24 (std=0.05)	0.24 (std=0.05)	56.0	RNA biosynthetic process
$\widetilde{A}_{G_5}$	10	0.24 (std=0.05)	0.24 (std=0.05)	31.0	translational initiation
$\widetilde{A}_{G_5}$	10	0.24 (std=0.05)	0.24 (std=0.05)	30.0	mannosyl-inositol phosphorylceramide metabolic process
$\widetilde{A}_{G_5}$	10	0.24 (std=0.05)	0.24 (std=0.05)	28.0	sister chromatid cohesion
$\widetilde{A}_{G_5}$	10	0.24 (std=0.05)	0.24 (std=0.05)	27.0	post-Golgi vesicle-mediated transport
$\widetilde{A}_{G_5}$	10	0.24 (std=0.05)	0.24 (std=0.05)	24.0	rDNA heterochromatin assembly
$\widetilde{A}_{G_5}$	10	0.24 (std=0.05)	0.24 (std=0.05)	24.0	facultative heterochromatin assembly
$\widetilde{A}_{G_5}$	10	0.24 (std=0.05)	0.24 (std=0.05)	20.0	ribonucleoprotein complex disassembly
$\widetilde{A}_{G_5}$	10	0.24 (std=0.05)	0.24 (std=0.05)	20.0	spliceosomal complex disassembly
$\widetilde{A}_{G_5}$	10	0.24 (std=0.05)	0.24 (std=0.05)	15.0	cellular component disassembly
$\widetilde{A}_{G_6}$	7	0.35 (std=0.05)	0.35 (std=0.05)	61.0	RNA phosphodiester bond hydrolysis
$\widetilde{A}_{G_6}$	7	0.35 (std=0.05)	0.35 (std=0.05)	44.0	nucleic acid phosphodiester bond hydrolysis
$\widetilde{A}_{G_6}$	7	0.35 (std=0.05)	0.35 (std=0.05)	38.0	positive regulation of nucleobase-containing compound metabolic process
$\widetilde{A}_{G_6}$	7	0.35 (std=0.05)	0.35 (std=0.05)	37.0	cleavage involved in rRNA processing
$\widetilde{A}_{G_6}$	7	0.35 (std=0.05)	0.35 (std=0.05)	34.0	monocarboxylic acid metabolic process
$\widetilde{A}_{G_6}$	7	0.35 (std=0.05)	0.35 (std=0.05)	30.0	nuclear transport
$\widetilde{A}_{G_6}$	7	0.35 (std=0.05)	0.35 (std=0.05)	30.0	nucleocytoplasmic transport
$\widetilde{A}_{G_7}$	4	0.62 (std=0.01)	0.62 (std=0.01)	123.0	NAD metabolic process
$\widetilde{A}_{G_7}$	4	0.62 (std=0.01)	0.62 (std=0.01)	121.0	NADH metabolic process
$\widetilde{A}_{G_7}$	4	0.62 (std=0.01)	0.62 (std=0.01)	117.0	NADH oxidation
$\widetilde{A}_{G_7}$	4	0.62 (std=0.01)	0.62 (std=0.01)	47.0	late nucleophagy
$\widetilde{A}_{G_8}$	5	0.63 (std=0.00)	0.63 (std=0.00)	124.0	ion transport
$\widetilde{A}_{G_8}$	5	0.63 (std=0.00)	0.63 (std=0.00)	13.0	pyridoxine metabolic process
$\widetilde{A}_{G_8}$	5	0.63 (std=0.00)	0.63 (std=0.00)	13.0	vitamin B6 metabolic process
$\widetilde{A}_{G_8}$	5	0.63 (std=0.00)	0.63 (std=0.00)	13.0	pyridoxine biosynthetic process
$\widetilde{A}_{G_8}$	5	0.63 (std=0.00)	0.63 (std=0.00)	13.0	vitamin B6 biosynthetic process

Table 4.10. Summary of uniquely enriched GO-BPs for graphlet Spectral embeddings, Part 2. We report, for the Budding yeast GI network, the number of uniquely enriched GO-BPs and the mean semantic similarity (SS) between the uniquely enriched annotations (GO-BPs) (columns 2 and 3) for graphlet Spectral embeddings based on all graphlet adjacencies for up to four node graphlets, i.e. $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$ (column 1). In column 4, we report the mean SS for the top ten largest enriched annotations (column 6), i.e., ranking them in descending order according to the number of neighborhoods that the annotations are enriched in (column 5).

Genes enriched in *E. coli*, Fission yeast and Budding yeast GI networks cover more paralogous genes than the Fruit fly GI network

In our analysis of GraCoal embeddings for the GI networks, we discussed how GraCoals, in particular $GraCoal_{2,7,8}$ (i.e., based on triangle topology), lead to higher enrichments because of the presence of paralogs in the networks. Here we assess if this is also the case for graphlet Spectral embeddings. Below we report gene enrichment and paralog enrichment statistics for the GI networks in the same format as before: total number of genes that have at least one GO-BP enriched in the neighborhood (i.e., “Enriched genes” in Table 4.11) and the number of genes that are enriched and are paralogs (i.e., “Paralogs” in Table 4.11). In general, our observations for the fruit fly are consistent with previous results as it achieves low percentages of paralogs enriched across all graphlet Spectrals. Graphlet Spectrals corresponding to triangle topology (i.e., $Spectral_{2,7,8}$) in general perform really well, achieving the best paralog enrichments for budding yeast and *E. coli*. For fission yeast, $Spectral_2$ achieves the highest percentage of enriched paralogs, while $Spectral_{7,8}$ achieve both 0%, which can be easily explained by an extremely low count of enriched genes in these two embeddings (55 and 40 enriched genes, respectively). This is consistent with previous results both for GraSpring embeddings and GraCoal embeddings on GI networks. We observe that when there are lots of duplicated genes (i.e., paralogs) in a given GI network, the enrichments tend to be the best when based on triangle topology (e.g., $\widetilde{A}_2$, $\widetilde{A}_7$ or $\widetilde{A}_8$). Finally, we observe that these results for graphlet Spectral embeddings based on triangle topology, $Spectral_{2,7,8}$ are of lower performance than the ones for $GraCoal_{2,7,8}$, which further validates that GraCoals are a better approach for uncovering biological information from GI networks that contain many paralogs.

	Budding yeast		***E. coli***		**Fission yeast**		**Fruit fly**	
	Enriched Genes	Paralogs	Enriched Genes	Paralogs	Enriched Genes	Paralogs	Enriched Genes	Paralogs
$\widetilde{A}_{G_0}$	896	14.51	964	22.32	233	6.30	1,373	6.18
$\widetilde{A}_{G_1}$	339	9.59	792	23.99	266	6.77	1,592	5.99
$\widetilde{A}_{G_2}$	664	16.96	724	28.97	106	11.89	821	6.58
$\widetilde{A}_{G_3}$	1,438	12.52	969	21.28	81	1.23	1,361	7.27
$\widetilde{A}_{G_4}$	140	0.71	949	22.13	494	10.12	1,388	4.10
$\widetilde{A}_{G_5}$	546	4.21	802	24.55	343	11.08	798	9.27
$\widetilde{A}_{G_6}$	1,016	11.12	710	22.96	212	4.72	1,296	6.64
$\widetilde{A}_{G_7}$	843	18.03	692	24.11	55	0.00	777	6.82
$\widetilde{A}_{G_8}$	577	20.10	300	19.00	40	0.00	179	2.79

Table 4.11. Statistics for paralogous genes enriched using SAFE with graphlet Spectral embeddings. For each of the four GI networks (Budding yeast, *E. coli*, Fission yeast and Fruit fly), we show the number of enriched genes when using SAFE with graphlet Spectral embeddings (i.e., genes that have at least one annotation enriched in their neighborhood) and the percentages of genes enriched that are paralogs.

Biological insights of graphlet Spectral embeddings at a functional domain level

Lastly, we aim to give insight into the biological function captured by our graphlet Spectral embeddings across species at a functional domain level just as we did for our GraCoal embeddings and GraSpring embeddings in previous sections. To this end, we identify the most characteristic functional domains in each species across all graphlet Spectrals, i.e., the domains that could not be captured by any of the other graphlet Spectrals. In brief, we measure the uniqueness of all functional domains obtained with SAFE by computing the Jaccard similarity index (JI) between the sets of enriched annotations of each functional domain in a particular graphlet Spectral embedding and the sets of enriched annotations of every other functional domain in the other graphlet Spectral embeddings. For each domain, we report the maximum JI, which represents the maximum overlap to any other functional domain in the other graphlet Spectral embeddings, and thus reflects how unique the functional domain. Finally, we also report the paralog ratio, which we already defined, as a way to evaluate how well a given functional domain is capturing biological information that involves paralogs.

In Table 4.12 for budding yeast, and Tables A.18 to A.20 in Appendix A for *E. coli*, fission yeast and fruit fly, respectively, we summarize, the number of functional domains (column 1) and the mean paralog ratio (column 2), over each graphlet Spectral embedding (i.e., $\widetilde{A}_{G_0}$-$\widetilde{A}_{G_8}$). Furthermore, we report the top three most characteristic functional domains for the GI molecular networks (column 5) according to the lowest maximum JI (column 4) and the corresponding paralog ratio. Our first observation is that we obtain more unique functional domains (i.e., Max JI = 0.0) for budding yeast when using graphlet Spectrals than using GraSprings or GraCoals. For instance, for budding yeast we obtain 17 completely unique functional domains while with GraSpring or GraCoal we obtain 5 and 14 completely unique functional domains, respectively. Moreover, on average, we obtain 6.22 functional domains for each graphlet Spectral on budding yeast, while this number is almost twice as many when using GraCoal (average = 12.6). By having less functional domains overall, this can easily explain the low overlap between the different graphlet Spectrals, as the Max JI = 0.0 on most of the most characteristic funtionanl domains (Table 4.12). In general, we observe lower paralog ratios in graphlet Spectrals when compared to GraCoals. For instance, the lowest and highest average paralog ratio previously reported for our GraCoals for the budding yeast is 0.15 (std=0.06) and 0.22 (std=0.06), respectively. On the other hand, with graphlet Spectrals we obtain paralog ratios than range from 0.13 (std=0.07) and 0.21 (std=0.28). Additionally, the largest paralog ratio obtained from the top three most characteristic functional domains across all graphlet Spectrals, is 0.39 with a Max JI of 0.0. Interestingly, this functional domain is uncovered by $Spectral_7$, which is consistent with our previous results that indicate that triangle topology is the best for capturing biological function that involves paralogs. Similarly, the most characteristic functional domain of $GraSpring_8$, with a Max JI of 0.0, also has one of the highest paralog ratios (0.26).

In conclusion, we have shown that information captured by graphlet Spectrals is less functionally organised, as evidenced by low percentages of enriched genes and low percentages of annotations enriched, which in turn leads to fewer functional domains uncovered with SAFE. Next, we showed that this biological information uncovered is different across all graphlet Spectrals (i.e., complementarity). Finally, even when performing worse than GraCoals, graphlet Spectrals based on triangle topology, for instance $GraSpring_{7,8}$, still uncover biological functions that involve many paralogs, as shown by the mean paralog ratios of the functional domains.

$\widetilde{A}_{G_i}$	Num functional domains	Mean paralog ratio	Domain paralog ratio	Domain max JI	Domain description
$\widetilde{A}_{G_0}$	11	0.19 (std=0.24)	0.33	0.00	transport, transmembrane, hexose, carbohydrate, monosaccharide
$\widetilde{A}_{G_0}$	11	0.19 (std=0.24)	0.23	0.00	transport, retrograde, endosome, Golgi, cytosolic
$\widetilde{A}_{G_0}$	11	0.19 (std=0.24)	0.09	0.00	process, metabolic, biosynthetic, glycolipid, glycerolipid
$\widetilde{A}_{G_1}$	3	0.15 (std=0.12)	0.30	0.00	homeostasis, cellular, homeostatic, process, chemical
$\widetilde{A}_{G_1}$	3	0.15 (std=0.12)	0.01	0.00	regulation, chromosome, separation, sister, chromatid
$\widetilde{A}_{G_1}$	3	0.15 (std=0.12)	0.12	0.09	protein, containing, complex, organization
$\widetilde{A}_{G_2}$	3	0.18 (std=0.06)	0.07	0.07	mitochondrion, disassembly, autophagy, nucleus, microautophagy
$\widetilde{A}_{G_2}$	3	0.18 (std=0.06)	0.22	0.40	rRNA, metabolic, process, processing, ncRNA
$\widetilde{A}_{G_2}$	3	0.18 (std=0.06)	0.16	0.44	metabolic, process, compound, RNA, mRNA
$\widetilde{A}_{G_3}$	15	0.16 (std=0.09)	0.32	0.00	localization, transport, establishment, cellular
$\widetilde{A}_{G_3}$	15	0.16 (std=0.09)	0.20	0.00	glycosylation, macromolecule, protein
$\widetilde{A}_{G_3}$	15	0.16 (std=0.09)	0.16	0.00	electron, transport, chain, aerobic, respiratory
$\widetilde{A}_{G_5}$	7	0.13 (std=0.07)	0.14	0.00	translational, initiation
$\widetilde{A}_{G_5}$	7	0.13 (std=0.07)	0.12	0.00	RNA, biosynthetic, process
$\widetilde{A}_{G_5}$	7	0.13 (std=0.07)	0.00	0.00	heterochromatin, assembly, rDNA, facultative
$\widetilde{A}_{G_6}$	10	0.19 (std=0.07)	0.34	0.00	monocarboxylic, acid, metabolic, process
$\widetilde{A}_{G_6}$	10	0.19 (std=0.07)	0.10	0.00	phosphodiester, bond, hydrolysis, RNA, nucleic
$\widetilde{A}_{G_6}$	10	0.19 (std=0.07)	0.15	0.05	chromatin, organization
$\widetilde{A}_{G_7}$	5	0.21 (std=0.28)	0.39	0.00	NADH, metabolic, process, oxidation, NAD
$\widetilde{A}_{G_7}$	5	0.21 (std=0.28)	0.00	0.00	late, nucleophagy
$\widetilde{A}_{G_7}$	5	0.21 (std=0.28)	0.20	0.40	metabolic, process, RNA, rRNA, ncRNA
$\widetilde{A}_{G_8}$	2	0.20 (std=0.08)	0.26	0.00	process, pyridoxine, metabolic, vitamin, B6
$\widetilde{A}_{G_8}$	2	0.20 (std=0.08)	0.15	0.00	ion, transport

Table 4.12. Summary of most unique functional domains for graphlet Spectral embeddings. We report for the budding yeast GI network, for each graphlet Spectral embedding used with SAFE (column 1), i.e., based on graphlet adjacencies for up to four node graphlets ($\widetilde{A}_{G_0}$-$\widetilde{A}_{G_8}$), the number of functional domains (column 2) the mean paralog ratio (column 3) and the top three most characteristic functional domains (column 6). Lastly, for each functional domain we report the paralog ratio (column 4) and the maximum Jaccard similarity index (JI) (column 5).

Chapter 5

Application 2: Analysis of PPI networks

In this chapter we evaluate the performance of the graphlet-based embeddings (i.e., GraCoal, GraSpring and graphlet based Spectral) with the Spatial Analysis of Functional Enrichment (SAFE) framework on the PPI networks of the following species: *Drosophila melanogaster*, *Escherichia coli*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Homo sapiens*, *Mus musculus* and *Caenorhabditis elegans* which throughout the text we will refer to as fruit fly, *E. coli*, budding yeast, fission yeast, human, mouse and roundworm, respectively. We present the PPI network statistics in Table 5.1. For more information on how we built these molecular networks please refer to section A.1 in Appendix A. Moreover, we focus mainly on analysing results based on Gene Ontology Biological Processes (GO-BP), as is one of the most complete set of annotations. For detailed results corresponding to our other annotations (e.g., GO molecular functions and GO cellular components), please refer to section A.3 in Appendix A. In the next sections, we summarize the results obtained by our GraCoal embedding, GraSpring embedding and finally graphlet based Spectral embedding. Addidionally, we perform model fitting experiments, as we did for our GI networks in the previous chapter. However, none of our PPI networks were well fitted by any of the model networks, as seen in section A.5 in Appendix A.

		PPI	
	Nodes	**Edges**	**Density**
Budding yeast	5,726	92,930	0.006
E. coli	2,022	12,788	0.006
Fission yeast	3,530	12,757	0.002
Fruit fly	8,864	54,722	0.001
Human	18,614	398,713	0.002
House mouse	10,164	55,640	0.001
Roundworm	7,628	32,502	0.001

Table 5.1. PPI molecular network data statistics. For each species (row), we report the number of nodes, the number of edges and the density of the corresponding PPI network (columns 1-3).

5.1 Results for GraCoal with SAFE

In general, we observe that GraCoal embeddings appear to be the best performing in terms of genes enriched with respect to GraSpring embeddings and graphlet Spectrals. However, this is not always the case in terms of GO-BP annotations enriched. For instance, when we consider the union of the enrichments over the different embeddings (i.e., over $\widetilde{A}_0$ to $\widetilde{A}_8$), we observe GraCoal embeddings outperform the union of the GraSpring embeddings, on average by 14.1% and 3.27%, in terms of genes enriched and GO-BP annotations enriched, respectively. The union over all GraCoal embeddings also outperform the union of graphlet Spectral embeddings in terms of genes enriched but not in terms of GO-BP annotations enriched (0.7% and -1.49%, respectively) (Figure 5.1).

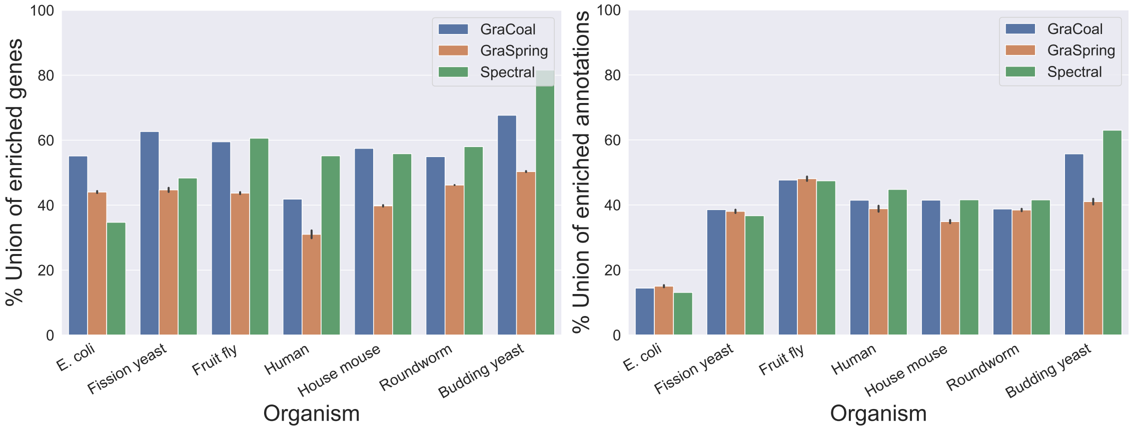


Figure 5.1. SAFE GO-BP enrichment analysis for PPI networks. For the PPI networks of our seven species (x-axis), we show the percentage of enriched genes (y-axis) and percentage of enriched annotations for each of the embedding algorithms considered (legend). In the case of GraSpring, we show the average across ten randomised runs and the standard deviation (error-bars).

On the other hand, when we take individual enrichments as opposed to the union over all graphlet based embeddings, we observe that the best scoring GraCoal embeddings (Figure 5.2) outperform the best GraSpring embeddings (Figure 5.3) and graphlet Spectral embeddings (Figure 5.4), on average by 4.50% and 4.29, respectively in terms of genes enriched. For the percentages of enriched GO-BP annotations enriched, the best embedding algorithm is not consistent, as it depends on the species. Finally, the best performing GraCoal embeddings outperform GraCoal embedding based on standard graphlet adjacency (i.e., $\widetilde{A}_{G_0}$) on average by 2.64% and 1.11% in terms of genes and GO-BP annotations enriched, respectively.

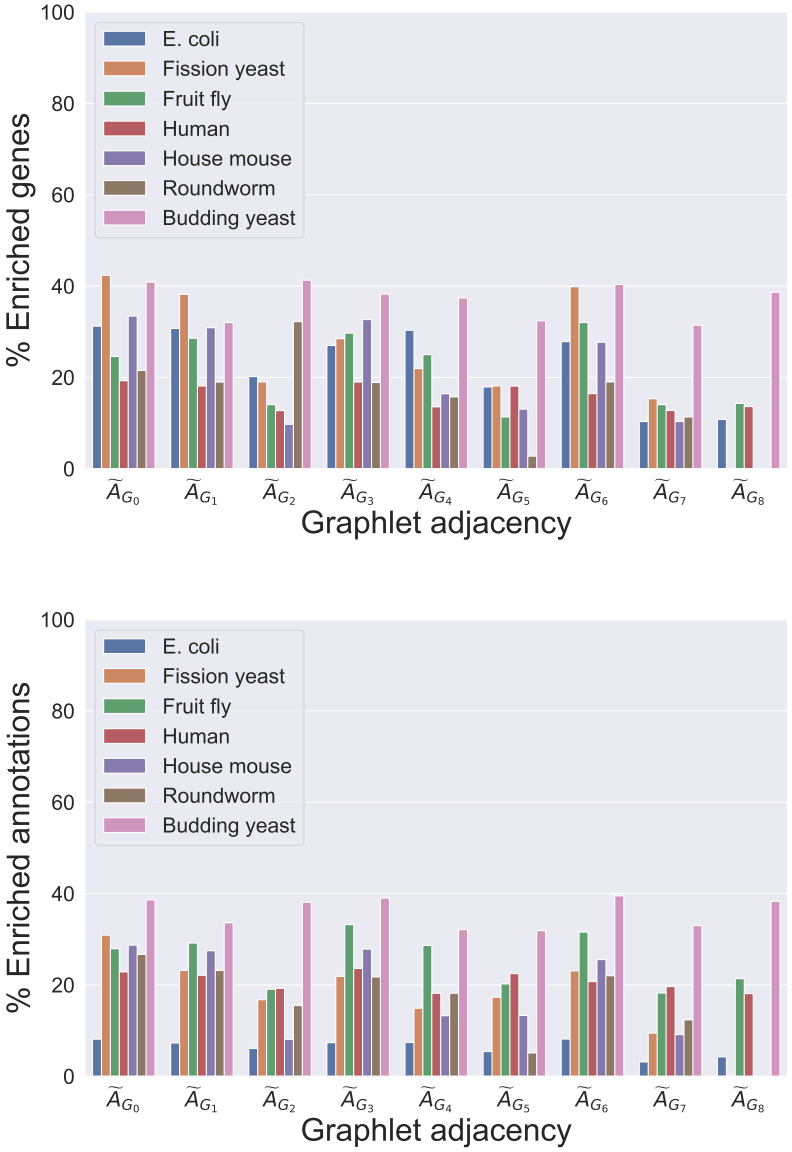


Figure 5.2. SAFE GO-BP enrichment analysis comparing GraCoals in PPI networks. For the PPI networks of our seven species (legend), we show, on the y-axis, the percentage of enriched genes (top) and the percentage of enriched annotations (bottom) for each of the different GraCoal embeddings (x-axis).

We observe similar results for GO-CC annotations over our PPI networks in Figure A.26 in Appendix A. The union of the enriched genes and annotations over the different GraCoal embeddings outperforms those based on the union of the Spring embedding (on average by 16.6% and 3.43% in terms of genes and GO-CC annotations enriched, respectively) or Spectral embedding based results (on average by 3.5% and 1.32% in terms of genes and GO-CC annotations enriched, respectively). The best scoring GraCoal embeddings (Figure A.27 in Appendix A) outperform the best GraSpring embeddings (Figure A.28 and graphlet Spectral embeddings (Figure A.29 in terms of genes enriched (on average by 3.63 % and 4.05%, respectively) and perform close to the best Spring embeddings and Spectral embeddings in terms of GO-CC annotations enriched (on average by -1.37% and -2.97%, respectively).

Finally, the best performing GraCoal embeddings outperform GraCoal based on standard adjacency (on average by 3.80% and 0.88% in terms of genes and GO-CC annotations enriched, respectively).

For GO-MF annotations, we observe are not well captured by any of the embedding algorithms for any of our networks, with all embeddings achieving less than 25% enriched annotations across all networks and graphlet adjacencies (i.e., $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$).

GraCoals uncover complementary biological information in PPI networks

When applying GraCoal embeddings with SAFE to our PPI molecular networks, we can uncover biological information in complementary ways, just as we previously discussed for our GI networks. In Figure 5.2 we show the percentages of genes enriched (top) and percentages of GO-BPs enriched (bottom) for all GraCoals (i.e., $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$) across our seven PPI networks. We observe the largest percentages of both genes enriched and annotations enriched for the budding yeast PPI network (pink label). In terms of genes enriched, we find that the best performing GraCoals for budding yeast correspond to $GraCoal_{0,2,6,8}$, achieving between 38.4.2% and 41% enriched genes, while the low performing, $GraCoal_{1,3,4,5,7}$, achieve between 31.2% and 37.4% enriched genes. This is consistent in terms of GO-BP annotations enriched, although we can observe an additional GraCoal in the top performing ones ($GraCoal_3$), which achieve between 37.3% and 39.5%. Next, in terms of genes enriched, GraCoals perform best on the fission yeast PPI network, achieving between 17.1% on the lower end with $GraCoal_7$ and 42.4% on the high end with $GraCoal_0$. However, this is not consistent with respect to enriched GO-BPs, as the second best percentages of enriched annotations are achieved on the fruit fly PPI network, between 18.2% ($GraCoal_7$) and 36.1% ($GraCoal_3$). Interestingly, GraCoals perform relatively well in terms of genes enriched for the *E. coli* PPI network, but not in terms of enriched GO-BPs, achieving the worst percentages of enriched annotations across all PPI networks. In general, besides the top performing species and the worse performing in the case of *E. coli* (in terms of GO-BPs), the results for the other PPI networks are not that different. Finally, we observe that GraCoals based on paths (i.e., $GraCoal_{0,1,3}$ tend to capture more biological information than GraCoals based on more densely connected graphlets such as the two-node and three-node cliques (i.e., $GraCoal2, 8$). In all, when comparing between our different PPI networks, there is no clear top performing GraCoal, indicating that GraCoals capture different topology-function relationships that depends on the species.

Additionally, for each specie, we focus on identifying what characterizes each particular GraCoal (i.e., $\widetilde{A}_{G_0}$-$\widetilde{A}_{G_8}$) from a biological perspective, just as we did previously for our GI networks with our graphlet based embeddings. In table 5.2 we present the same format as previously presented for GI networks: we report for the budding yeast, the number of uniquely enriched annotations (column 1), the mean SS for the uniquely enriched annotations (column 2) as well as the mean SS for the top 10 largest uniquely enriched annotations (column 3) for each GraCoal used in SAFE. Finally, in column 5, we report the names of the top 10 uniquely enriched annotations and their corresponding size in terms of enriched neighborhoods (column 4). Here we recall to the previous chapter to define what a uniquely enriched annotation is: annotations enriched in a particular GraCoal embedding that are not enriched in any of the other GraCoal embeddings. For the budding yeast, we find that on average, 30.12 GO-BPs are uniquely enriched for each particular GraCoal (Table 5.2). Our first observation for the budding yeast, is that every GraCoal uncovers uniquely enriched GO-BPs (column 1), which validates the claim that graphlet adjacencies capture complementary topology-function relationships in molecular networks (Windels et al., 2019). Next, the mean semantic similarities (column 2) indicate some degree of functional similarity uncovered by GraCoals in the budding yeast. In this regard, the lowest degree of functional similarity is achieved by $GraCoal_8$ (mean SS = 0.13, std=0.03) while the most functional similarity is captured by $GraCoal_4$ (mean SS = 0.38, std=0.05). Finally, when evaluating the top 10 enriched annotations, we observe a larger degree of functional similarity between the enriched GO-BPs, which ranges from 0.22 (Std=0.01) to 0.38 (Std=0.05) $GraCoal_6$ and $GraCoal_4$, respectively. Next, we rank the uniquely enriched annotations according to their size, defined as the total number of neighborhoods they are enriched in, as a measure of how well they are captured by each particular GraCoal. The average size of the uniquely enriched GO-BPs uncovered by GraCoals in the budding yeast is 176.99 (std=85.11). For our other PPI molecular networks, we summarize these statistics in terms of the mean uniquely enriched GO-BPs, mean size of enriched GO-BPs (i.e., in mean number of neighborhoods they are enriched in) and mean semantic similarity in Table 5.3.

In general, we observe that GraCoal embeddings uncover complementary information in all of our PPI molecular networks (i.e., uniquely enriched annotations across all GraCoals, represented by the means in column 2 in Table 5.3). Interestingly, the fewer uniquely enriched GO-BPs captured, the higher the semantic similarity (i.e., more functional relevance of the annotations) and the lower the sizes in terms of enriched neighborhoods (column 3). This indicates that in species such as *E. coli* and fission yeast, uniquely enriched GO-BPs tend to be highly specific (i.e., enriched in fewer neighborhoods) and very closely related (i.e., large semantic similarity).

$\widetilde{A}_{G_i}$	**Total annotations**	**Mean SS**	**Mean SS Top 10**	**EN**	**Annotation**
$\widetilde{A}_{G_0}$	34	0.17 (std=0.03)	0.27 (std=0.05)	365.0	ribosomal small subunit export from nucleus
$\widetilde{A}_{G_0}$	34	0.17 (std=0.03)	0.27 (std=0.05)	199.0	positive regulation of cellular component biogenesis
$\widetilde{A}_{G_0}$	34	0.17 (std=0.03)	0.27 (std=0.05)	180.0	regulation of cellular component biogenesis
$\widetilde{A}_{G_0}$	34	0.17 (std=0.03)	0.27 (std=0.05)	179.0	regulation of protein complex assembly
$\widetilde{A}_{G_0}$	34	0.17 (std=0.03)	0.27 (std=0.05)	165.0	cellular protein modification process
$\widetilde{A}_{G_0}$	34	0.17 (std=0.03)	0.27 (std=0.05)	165.0	protein modification process
$\widetilde{A}_{G_0}$	34	0.17 (std=0.03)	0.27 (std=0.05)	153.0	positive regulation of cytoskeleton organization
$\widetilde{A}_{G_0}$	34	0.17 (std=0.03)	0.27 (std=0.05)	130.0	vesicle budding from membrane
$\widetilde{A}_{G_0}$	34	0.17 (std=0.03)	0.27 (std=0.05)	126.0	karyogamy
$\widetilde{A}_{G_0}$	34	0.17 (std=0.03)	0.27 (std=0.05)	121.0	intralumenal vesicle formation
$\widetilde{A}_{G_1}$	20	0.18 (std=0.03)	0.23 (std=0.01)	400.0	regulation of ribosome biogenesis
$\widetilde{A}_{G_1}$	20	0.18 (std=0.03)	0.23 (std=0.01)	391.0	aggrephagy
$\widetilde{A}_{G_1}$	20	0.18 (std=0.03)	0.23 (std=0.01)	391.0	ribophagy
$\widetilde{A}_{G_1}$	20	0.18 (std=0.03)	0.23 (std=0.01)	345.0	regulation of ribosomal subunit export from nucleus
$\widetilde{A}_{G_1}$	20	0.18 (std=0.03)	0.23 (std=0.01)	265.0	regulation of transcription elongation from RNA polymerase II promoter
$\widetilde{A}_{G_1}$	20	0.18 (std=0.03)	0.23 (std=0.01)	250.0	regulation of intracellular transport
$\widetilde{A}_{G_1}$	20	0.18 (std=0.03)	0.23 (std=0.01)	242.0	regulation of translational initiation
$\widetilde{A}_{G_1}$	20	0.18 (std=0.03)	0.23 (std=0.01)	225.0	regulation of nucleobase-containing compound transport
$\widetilde{A}_{G_1}$	20	0.18 (std=0.03)	0.23 (std=0.01)	169.0	regulation of histone modification
$\widetilde{A}_{G_1}$	20	0.18 (std=0.03)	0.23 (std=0.01)	148.0	catabolic process
$\widetilde{A}_{G_2}$	32	0.17 (std=0.03)	0.27 (std=0.05)	319.0	protein localization
$\widetilde{A}_{G_2}$	32	0.17 (std=0.03)	0.27 (std=0.05)	292.0	regulation of vesicle-mediated transport
$\widetilde{A}_{G_2}$	32	0.17 (std=0.03)	0.27 (std=0.05)	228.0	pyridine-containing compound biosynthetic process
$\widetilde{A}_{G_2}$	32	0.17 (std=0.03)	0.27 (std=0.05)	222.0	vesicle docking
$\widetilde{A}_{G_2}$	32	0.17 (std=0.03)	0.27 (std=0.05)	212.0	vesicle tethering
$\widetilde{A}_{G_2}$	32	0.17 (std=0.03)	0.27 (std=0.05)	209.0	membrane docking
$\widetilde{A}_{G_2}$	32	0.17 (std=0.03)	0.27 (std=0.05)	209.0	organelle localization by membrane tethering
$\widetilde{A}_{G_2}$	32	0.17 (std=0.03)	0.27 (std=0.05)	180.0	regulation of vacuole organization
$\widetilde{A}_{G_2}$	32	0.17 (std=0.03)	0.27 (std=0.05)	178.0	vacuole fusion
$\widetilde{A}_{G_2}$	32	0.17 (std=0.03)	0.27 (std=0.05)	178.0	vacuole fusion, non-autophagic

Table 5.2. Summary of uniquely enriched GO-BPs for Gracoal embeddings, Part 1. We report, for the Budding yeast PPI network, the number of uniquely enriched GO-BPs and the mean semantic similarity (SS) between the uniquely enriched annotations (GO-BPs) (columns 2 and 3) for GraCoals based on all graphlet adjacencies for up to four node graphlets, i.e. $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$ (column 1). In column 4, we report the mean SS for the top ten largest enriched annotations (column 6), i.e., ranking them in descending order according to the number of neighborhoods that the annotations are enriched in (column 5).

$\widetilde{A}_{G_i}$	**Total annotations**	**Mean SS**	**Mean SS Top 10**	**EN**	**Annotation**
$\widetilde{A}_{G_3}$	35	0.17 (std=0.05)	0.25 (std=0.05)	275.0	DNA-templated transcription, termination
$\widetilde{A}_{G_3}$	35	0.17 (std=0.05)	0.25 (std=0.05)	198.0	cellular protein localization
$\widetilde{A}_{G_3}$	35	0.17 (std=0.05)	0.25 (std=0.05)	195.0	protein localization to organelle
$\widetilde{A}_{G_3}$	35	0.17 (std=0.05)	0.25 (std=0.05)	184.0	nucleosome mobilization
$\widetilde{A}_{G_3}$	35	0.17 (std=0.05)	0.25 (std=0.05)	160.0	regulation of cytoskeleton organization
$\widetilde{A}_{G_3}$	35	0.17 (std=0.05)	0.25 (std=0.05)	158.0	leading strand elongation
$\widetilde{A}_{G_3}$	35	0.17 (std=0.05)	0.25 (std=0.05)	158.0	regulation of cell cycle phase transition
$\widetilde{A}_{G_3}$	35	0.17 (std=0.05)	0.25 (std=0.05)	158.0	regulation of mitotic cell cycle phase transition
$\widetilde{A}_{G_3}$	35	0.17 (std=0.05)	0.25 (std=0.05)	145.0	proteasomal protein catabolic process
$\widetilde{A}_{G_3}$	35	0.17 (std=0.05)	0.25 (std=0.05)	139.0	proteasome-mediated ubiquitin-dependent protein catabolic process
$\widetilde{A}_{G_4}$	9	0.38 (std=0.05)	0.38 (std=0.05)	260.0	organonitrogen compound biosynthetic process
$\widetilde{A}_{G_4}$	9	0.38 (std=0.05)	0.38 (std=0.05)	95.0	Golgi vesicle budding
$\widetilde{A}_{G_4}$	9	0.38 (std=0.05)	0.38 (std=0.05)	45.0	nuclear-transcribed mRNA catabolic process, deadenylation-dependent decay
$\widetilde{A}_{G_4}$	9	0.38 (std=0.05)	0.38 (std=0.05)	44.0	negative regulation of cell aging
$\widetilde{A}_{G_4}$	9	0.38 (std=0.05)	0.38 (std=0.05)	34.0	filamentous growth of a population of unicellular organisms
$\widetilde{A}_{G_4}$	9	0.38 (std=0.05)	0.38 (std=0.05)	34.0	invasive growth in response to glucose limitation
$\widetilde{A}_{G_4}$	9	0.38 (std=0.05)	0.38 (std=0.05)	34.0	growth of unicellular organism as a thread of attached cells
$\widetilde{A}_{G_4}$	9	0.38 (std=0.05)	0.38 (std=0.05)	34.0	invasive filamentous growth
$\widetilde{A}_{G_4}$	9	0.38 (std=0.05)	0.38 (std=0.05)	33.0	filamentous growth
$\widetilde{A}_{G_5}$	14	0.20 (std=0.04)	0.23 (std=0.04)	282.0	organelle organization
$\widetilde{A}_{G_5}$	14	0.20 (std=0.04)	0.23 (std=0.04)	209.0	positive regulation of organelle organization
$\widetilde{A}_{G_5}$	14	0.20 (std=0.04)	0.23 (std=0.04)	206.0	rRNA transcription
$\widetilde{A}_{G_5}$	14	0.20 (std=0.04)	0.23 (std=0.04)	170.0	negative regulation of cell cycle phase transition
$\widetilde{A}_{G_5}$	14	0.20 (std=0.04)	0.23 (std=0.04)	170.0	negative regulation of mitotic cell cycle phase transition
$\widetilde{A}_{G_5}$	14	0.20 (std=0.04)	0.23 (std=0.04)	162.0	negative regulation of chromosome organization
$\widetilde{A}_{G_5}$	14	0.20 (std=0.04)	0.23 (std=0.04)	102.0	transfer RNA gene-mediated silencing
$\widetilde{A}_{G_5}$	14	0.20 (std=0.04)	0.23 (std=0.04)	85.0	chromatin silencing at telomere
$\widetilde{A}_{G_5}$	14	0.20 (std=0.04)	0.23 (std=0.04)	57.0	regulation of actin filament bundle assembly
$\widetilde{A}_{G_5}$	14	0.20 (std=0.04)	0.23 (std=0.04)	40.0	regulation of SNARE complex assembly

Table 5.2. Summary of uniquely enriched GO-BPs for Gracoal embeddings, Part 2. We report, for the Budding yeast PPI network, the number of uniquely enriched GO-BPs and the mean semantic similarity (SS) between the uniquely enriched annotations (GO-BPs) (columns 2 and 3) for GraCoals based on all graphlet adjacencies for up to four node graphlets, i.e. $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$ (column 1). In column 4, we report the mean SS for the top ten largest enriched annotations (column 6), i.e., ranking them in descending order according to the number of neighborhoods that the annotations are enriched in (column 5).

$\widetilde{A}_{G_i}$	**Total annotations**	**Mean SS**	**Mean SS Top 10**	**EN**	**Annotation**
$\widetilde{A}_{G_6}$	33	0.15 (std=0.05)	0.22 (std=0.01)	275.0	positive regulation of translational fidelity
$\widetilde{A}_{G_6}$	33	0.15 (std=0.05)	0.22 (std=0.01)	274.0	rRNA methylation
$\widetilde{A}_{G_6}$	33	0.15 (std=0.05)	0.22 (std=0.01)	267.0	nucleolar large rRNA transcription by RNA polymerase I
$\widetilde{A}_{G_6}$	33	0.15 (std=0.05)	0.22 (std=0.01)	244.0	snRNA modification
$\widetilde{A}_{G_6}$	33	0.15 (std=0.05)	0.22 (std=0.01)	228.0	intracellular protein transport
$\widetilde{A}_{G_6}$	33	0.15 (std=0.05)	0.22 (std=0.01)	221.0	snoRNA processing
$\widetilde{A}_{G_6}$	33	0.15 (std=0.05)	0.22 (std=0.01)	214.0	U5 snRNA 3'-end processing
$\widetilde{A}_{G_6}$	33	0.15 (std=0.05)	0.22 (std=0.01)	197.0	cellular catabolic process
$\widetilde{A}_{G_6}$	33	0.15 (std=0.05)	0.22 (std=0.01)	127.0	organic substance transport
$\widetilde{A}_{G_6}$	33	0.15 (std=0.05)	0.22 (std=0.01)	108.0	SRP-dependent cotranslational protein targeting to membrane, translocation
$\widetilde{A}_{G_7}$	12	0.23 (std=0.05)	0.26 (std=0.06)	152.0	nuclear polyadenylation-dependent CUT catabolic process
$\widetilde{A}_{G_7}$	12	0.23 (std=0.05)	0.26 (std=0.06)	143.0	protein-lipid complex assembly
$\widetilde{A}_{G_7}$	12	0.23 (std=0.05)	0.26 (std=0.06)	143.0	lipid tube assembly
$\widetilde{A}_{G_7}$	12	0.23 (std=0.05)	0.26 (std=0.06)	129.0	CUT metabolic process
$\widetilde{A}_{G_7}$	12	0.23 (std=0.05)	0.26 (std=0.06)	129.0	CUT catabolic process
$\widetilde{A}_{G_7}$	12	0.23 (std=0.05)	0.26 (std=0.06)	121.0	protein-lipid complex subunit organization
$\widetilde{A}_{G_7}$	12	0.23 (std=0.05)	0.26 (std=0.06)	104.0	transposition
$\widetilde{A}_{G_7}$	12	0.23 (std=0.05)	0.26 (std=0.06)	44.0	regulation of actin filament organization
$\widetilde{A}_{G_7}$	12	0.23 (std=0.05)	0.26 (std=0.06)	40.0	positive regulation of actin filament polymerization
$\widetilde{A}_{G_7}$	12	0.23 (std=0.05)	0.26 (std=0.06)	27.0	organic cyclic compound biosynthetic process
$\widetilde{A}_{G_8}$	80	0.13 (std=0.03)	0.28 (std=0.05)	253.0	RNA surveillance
$\widetilde{A}_{G_8}$	80	0.13 (std=0.03)	0.28 (std=0.05)	253.0	nuclear RNA surveillance
$\widetilde{A}_{G_8}$	80	0.13 (std=0.03)	0.28 (std=0.05)	199.0	cell cycle checkpoint
$\widetilde{A}_{G_8}$	80	0.13 (std=0.03)	0.28 (std=0.05)	181.0	chromosome organization
$\widetilde{A}_{G_8}$	80	0.13 (std=0.03)	0.28 (std=0.05)	176.0	nuclear polyadenylation-dependent mRNA catabolic process
$\widetilde{A}_{G_8}$	80	0.13 (std=0.03)	0.28 (std=0.05)	176.0	snoRNA 3'-end processing
$\widetilde{A}_{G_8}$	80	0.13 (std=0.03)	0.28 (std=0.05)	176.0	polyadenylation-dependent mRNA catabolic process
$\widetilde{A}_{G_8}$	80	0.13 (std=0.03)	0.28 (std=0.05)	174.0	recombinational repair
$\widetilde{A}_{G_8}$	80	0.13 (std=0.03)	0.28 (std=0.05)	146.0	snRNA processing
$\widetilde{A}_{G_8}$	80	0.13 (std=0.03)	0.28 (std=0.05)	129.0	intracellular protein transmembrane transport

Table 5.2. Summary of uniquely enriched GO-BPs for Gracoal embeddings, Part 3. We report, for the Budding yeast PPI network, the number of uniquely enriched GO-BPs and the mean semantic similarity (SS) between the uniquely enriched annotations (GO-BPs) (columns 2 and 3) for GraCoals based on all graphlet adjacencies for up to four node graphlets, i.e. $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$ (column 1). In column 4, we report the mean SS for the top ten largest enriched annotations (column 6), i.e., ranking them in descending order according to the number of neighborhoods that the annotations are enriched in (column 5).

Organism	**Mean unique GO-BPs**	**Mean enriched neighborhoods**	**Mean SS**
Budding yeast	30.12 (std=20.31)	176.99 (std=85.11)	0.20 (std=0.07)
E. coli	4.25 (std=1.24)	41.09 (std=42.11)	0.55 (std=0.13)
Fission yeast	9.09 (std=4.73)	98.36 (std=74.90)	0.36 (std=0.17)
Fruit fly	40.44 (std=17.82)	237.31 (std=199.47)	0.18 (std=0.03)
House mouse	62.88 (std=27.58)	333.31 (std=266.62)	0.14 (std=0.02)
Human	39.76 (std=22.02)	278.85 (std=195.29)	0.18 (std=0.06)
Roundworm	20.66 (std=15.77)	54.45 (std=81.63)	0.27 (std=0.08)

Table 5.3. Summary statistics of uniquely enriched GO-BPs for Gracoal embeddings on PPI networks. We report, for the our seven PPI networks (column 1), the mean number of uniquely enriched GO-BPs obtained with each GraCoal embedding (column 2), the mean size of the uniquely enriched GO-BPs in terms of the number of neighborhoods that the annotations are enriched in (column 3) and the mean semantic similarity (column 4).

5.2 Results for GraSpring with SAFE

In the previous section we described how GraCoal embeddings are the best approach for uncovering biological information of our PPI molecular networks in terms of genes enriched (Figure 5.1). In this regard, GraSpring embeddings are the worst of the three methods in terms of gene enrichment, outperformed, on average by 14.1% by GraCoal embeddings, which are the best of the three methods. For the percentages of enriched GO-BP annotations, the differences are less noticeable, even though it is still outperformed by the best method, on average by 4.3% by graphlet Spectral embeddings. The only two PPI networks where GraSpring manages to perform best is on the PPI networks of *E. coli* and fruit fly.

GraSpring uncover complementary biological information in PPI networks

When applying GraSpring embeddings with SAFE to our PPI molecular networks, we can uncover biological information in complementary ways, just as we previously discussed for our PPI networks with GraCoal embeddings. In Figure 5.3 we show the percentages of genes enriched (top) and percentages of GO-BPs enriched (bottom) for all GraSprings (i.e., $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$) across our seven PPI networks. In terms of genes, most GraSprings achieve less than 20% genes enriched. The largest values are achieved on the *E. coli* PPI network, in particular for $GraSpring_{0,1,3,4,6}$ with percentages of genes enriched that range from 26.1% to 27.8%. For the fruit fly, the percentages of genes enriched are the second largest, achieving more than 20% in 6 out of 9 GraSprings, $GraSpring0, 1, 3, 4, 6, 7$. Interestingly, the percentages of genes enriched for the roundworm (brown label) are amongst the worst of all, except for $GraSpring_2$, which achieves the best score overall across all species (34.3%). Moreover, $GraSpring_0$ performs relatively well in comparison with the other GraSprings, achieving at least 20% of genes enriched in 5 out of the 7 PPI networks.

Similarly, in terms of annotations, most GraSprings achieve less than 20% enriched GO-BPs. The largest percentages are achieved on the fruit fly PPI network, all GraSprings achieving more than 20% enriched GO-BPs, which is consistent with the percentages of genes enriched, as is the second best in this regard. For *E. coli*, which achieves the best overall percentages of genes enriched, achieves the worst percentages of GO-BPs enriched. Our last general observation is also consistent with the percentages of genes enriched, which is that the percentages of GO-BPs enriched tend to be larger for $GraSpring_0$ than for the other GraSprings. Finally, our conclusion when comparing between our different PPI networks, there is no clear top performing GraSpring, as it varies between each species

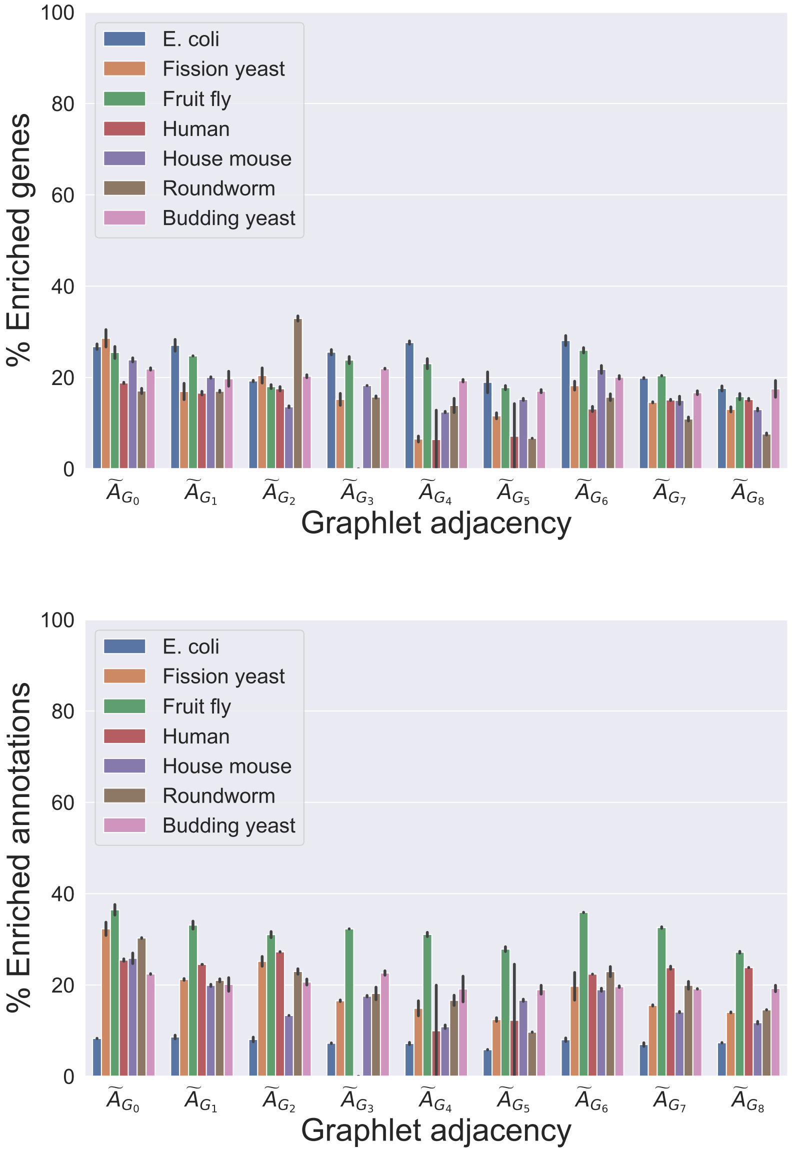


Figure 5.3. SAFE GO-BP enrichment analysis comparing GraSprings in PPI networks. For the PPI networks of our seven species (legend), we show, on the y-axis, the percentage of enriched genes (top) and the percentage of enriched annotations (bottom) for each of the different GraSpring embeddings (x-axis).

Next, we focus on identifying what characterizes each particular GraSpring (i.e., $\widetilde{A}_{G_0}$-$\widetilde{A}_{G_8}$) from a biological perspective, just as we did previously for GraCoal embeddings. In table 5.4 we present the same format as before: we report for the budding yeast, the number of uniquely enriched annotations (column 1), the mean SS for the uniquely enriched annotations (column 2) as well as the mean SS for the top 10 largest uniquely enriched annotations (column 3) for each GraSpring used in SAFE. Finally, in column 5, we report the names of the top 10 uniquely enriched annotations and their corresponding size in terms of enriched neighborhoods (column 4). We find that on average, 12.03 GO-BPs are uniquely enriched for each particular GraSpring. Our first observation for the budding yeast, is that every GraSpring uncovers uniquely enriched GO-BPs (column 1), which is consistent with our previous findings for GI networks and for GraCoal embeddings on PPI networks. That is, graphlet adjacencies capture complementary information from molecular networks (Windels et al., 2019). The mean semantic similarity (column 2) of each particular GraSpring represents some degree of functional relevance of the GO-BPs uncovered by the differeng GraSprings in SAFE. For instance, the lowest mean semantic similarity is achieved by $GraSpring_8$ at 0.21 (std=0.05) and the largest semantic similarity (i.e., of highest functional relevance) is achieved by $GraSpring_4$ at 0.45 (std=0.05). When evaluating the top 10 enriched annotations, we observe a larger degree of functional similarity between the enriched GO-BPs. Interstingly, the lowest and highest semantic similarities, at 0.25 (Std=0.06) and 0.45 (Std=0.05) are achied also for $GraSpring_8$ and $GraSpring_4$, respectively. In the case of $GraSpring_4$ this value does not change, as it only uncovers 6 uniquely enriched GO-BPs. Finally, the average size of the uniquely enriched GO-BPs uncovered by GraSprings in the budding yeast is 171.83 (std=134.03). For our other PPI molecular networks, we summarize these statistics in terms of the mean uniquely enriched GO-BPs, mean size of enriched GO-BPs (i.e., in mean number of neighborhoods they are enriched in) and mean semantic similarity in Table 5.5.

In general, we observe that GraSpring embeddings uncover complementary information in all of our PPI molecular networks (i.e., uniquely enriched annotations across all GraSprings, represented by the means in column 2 in Table 5.5). Except for *E. coli* and fission yeast, we observe that GraSpring embeddigns uncover less unique biological information than GraCoal embeddings. For instance, for budding yeast on average 12.03 unique GO-BPs can be captured with any particular GraSpring embedding, but almost three times as many (30.12 on average) can be captured by any particular GraCoal.

$\widetilde{A}_{G_i}$	**Total annotations**	**Mean SS**	**Mean SS Top 10**	**EN**	**Annotation**
$\widetilde{A}_{G_0}$	10	0.31 (std=0.05)	0.31 (std=0.05)	434.0	negative regulation of cellular amide metabolic process
$\widetilde{A}_{G_0}$	10	0.31 (std=0.05)	0.31 (std=0.05)	403.0	organelle organization
$\widetilde{A}_{G_0}$	10	0.31 (std=0.05)	0.31 (std=0.05)	286.0	positive regulation of cell cycle phase transition
$\widetilde{A}_{G_0}$	10	0.31 (std=0.05)	0.31 (std=0.05)	286.0	positive regulation of mitotic cell cycle phase transition
$\widetilde{A}_{G_0}$	10	0.31 (std=0.05)	0.31 (std=0.05)	219.0	positive regulation of mitotic cell cycle
$\widetilde{A}_{G_0}$	10	0.31 (std=0.05)	0.31 (std=0.05)	146.0	telomere organization
$\widetilde{A}_{G_0}$	10	0.31 (std=0.05)	0.31 (std=0.05)	122.0	ATP-dependent chromatin remodeling
$\widetilde{A}_{G_0}$	10	0.31 (std=0.05)	0.31 (std=0.05)	104.0	chromatin silencing at telomere
$\widetilde{A}_{G_0}$	10	0.31 (std=0.05)	0.31 (std=0.05)	34.0	ribonucleoprotein complex subunit organization
$\widetilde{A}_{G_0}$	10	0.31 (std=0.05)	0.31 (std=0.05)	25.0	ribonucleoprotein complex assembly
$\widetilde{A}_{G_1}$	15	0.23 (std=0.06)	0.26 (std=0.05)	309.0	cellular component assembly
$\widetilde{A}_{G_1}$	15	0.23 (std=0.06)	0.26 (std=0.05)	88.0	chromatin remodeling at centromere
$\widetilde{A}_{G_1}$	15	0.23 (std=0.06)	0.26 (std=0.05)	37.0	ribonucleoprotein complex biogenesis
$\widetilde{A}_{G_1}$	15	0.23 (std=0.06)	0.26 (std=0.05)	17.0	maturation of 5.8S rRNA
$\widetilde{A}_{G_1}$	15	0.23 (std=0.06)	0.26 (std=0.05)	17.0	maturation of 5.8S rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)
$\widetilde{A}_{G_1}$	15	0.23 (std=0.06)	0.26 (std=0.05)	13.0	ribonucleoprotein complex export from nucleus
$\widetilde{A}_{G_1}$	15	0.23 (std=0.06)	0.26 (std=0.05)	13.0	ribosome localization
$\widetilde{A}_{G_1}$	15	0.23 (std=0.06)	0.26 (std=0.05)	13.0	rRNA-containing ribonucleoprotein complex export from nucleus
$\widetilde{A}_{G_1}$	15	0.23 (std=0.06)	0.26 (std=0.05)	13.0	ribosomal subunit export from nucleus
$\widetilde{A}_{G_1}$	15	0.23 (std=0.06)	0.26 (std=0.05)	12.0	protein export from nucleus
$\widetilde{A}_{G_2}$	7	0.35 (std=0.05)	0.35 (std=0.05)	432.0	DNA-dependent DNA replication maintenance of fidelity
$\widetilde{A}_{G_2}$	7	0.35 (std=0.05)	0.35 (std=0.05)	354.0	small nucleolar ribonucleoprotein complex assembly
$\widetilde{A}_{G_2}$	7	0.35 (std=0.05)	0.35 (std=0.05)	344.0	peptidyl-threonine phosphorylation
$\widetilde{A}_{G_2}$	7	0.35 (std=0.05)	0.35 (std=0.05)	344.0	peptidyl-threonine modification
$\widetilde{A}_{G_2}$	7	0.35 (std=0.05)	0.35 (std=0.05)	330.0	box C/D snoRNP assembly
$\widetilde{A}_{G_2}$	7	0.35 (std=0.05)	0.35 (std=0.05)	239.0	protein ubiquitination
$\widetilde{A}_{G_2}$	7	0.35 (std=0.05)	0.35 (std=0.05)	146.0	ncRNA processing

Table 5.4. Summary of uniquely enriched GO-BPs for Graspring embeddings, Part 1. We report, for the Budding yeast PPI network, the number of uniquely enriched GO-BPs and the mean semantic similarity (SS) between the uniquely enriched annotations (GO-BPs) (columns 2 and 3) for GraSprings based on all graphlet adjacencies for up to four node graphlets, i.e. $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$ (column 1). In column 4, we report the mean SS for the top ten largest enriched annotations (column 6), i.e., ranking them in descending order according to the number of neighborhoods that the annotations are enriched in (column 5).

$\widetilde{A}_{G_i}$	**Total annotations**	**Mean SS**	**Mean SS Top 10**	**EN**	**Annotation**
$\widetilde{A}_{G_3}$	7	0.36 (std=0.04)	0.36 (std=0.04)	406.0	regulation of cellular component biogenesis
$\widetilde{A}_{G_3}$	7	0.36 (std=0.04)	0.36 (std=0.04)	219.0	snoRNA processing
$\widetilde{A}_{G_3}$	7	0.36 (std=0.04)	0.36 (std=0.04)	23.0	ribosomal large subunit biogenesis
$\widetilde{A}_{G_3}$	7	0.36 (std=0.04)	0.36 (std=0.04)	20.0	endonucleolytic cleavage in 5'-ETS of tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)
$\widetilde{A}_{G_3}$	7	0.36 (std=0.04)	0.36 (std=0.04)	18.0	endonucleolytic cleavage in ITS1 to separate SSU-rRNA from 5.8S rRNA and LSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)
$\widetilde{A}_{G_3}$	7	0.36 (std=0.04)	0.36 (std=0.04)	16.0	rRNA 5'-end processing
$\widetilde{A}_{G_3}$	7	0.36 (std=0.04)	0.36 (std=0.04)	16.0	endonucleolytic cleavage to generate mature 5'-end of SSU-rRNA from (SSU-rRNA, 5.8S rRNA, LSU-rRNA)
$\widetilde{A}_{G_4}$	6	0.45 (std=0.05)	0.45 (std=0.05)	61.0	RNA transport
$\widetilde{A}_{G_4}$	6	0.45 (std=0.05)	0.45 (std=0.05)	61.0	establishment of RNA localization
$\widetilde{A}_{G_4}$	6	0.45 (std=0.05)	0.45 (std=0.05)	57.0	nucleic acid transport
$\widetilde{A}_{G_4}$	6	0.45 (std=0.05)	0.45 (std=0.05)	56.0	RNA export from nucleus
$\widetilde{A}_{G_4}$	6	0.45 (std=0.05)	0.45 (std=0.05)	55.0	nuclear export
$\widetilde{A}_{G_4}$	6	0.45 (std=0.05)	0.45 (std=0.05)	54.0	proteasome assembly
$\widetilde{A}_{G_5}$	7	0.34 (std=0.05)	0.34 (std=0.05)	318.0	positive regulation of transcription initiation from RNA polymerase II promoter
$\widetilde{A}_{G_5}$	7	0.34 (std=0.05)	0.34 (std=0.05)	318.0	positive regulation of DNA-templated transcription, initiation
$\widetilde{A}_{G_5}$	7	0.34 (std=0.05)	0.34 (std=0.05)	282.0	positive regulation of protein complex assembly
$\widetilde{A}_{G_5}$	7	0.34 (std=0.05)	0.34 (std=0.05)	167.0	positive regulation of cellular component biogenesis
$\widetilde{A}_{G_5}$	7	0.34 (std=0.05)	0.34 (std=0.05)	156.0	DNA strand elongation
$\widetilde{A}_{G_5}$	7	0.34 (std=0.05)	0.34 (std=0.05)	16.0	organelle membrane fusion
$\widetilde{A}_{G_5}$	7	0.34 (std=0.05)	0.34 (std=0.05)	12.0	membrane fusion

Table 5.4. Summary of uniquely enriched GO-BPs for GraSpring embeddings, Part 1. We report, for the Budding yeast PPI network, the number of uniquely enriched GO-BPs and the mean semantic similarity (SS) between the uniquely enriched annotations (GO-BPs) (columns 2 and 3) for GraSprings based on all graphlet adjacencies for up to four node graphlets, i.e. $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$ (column 1). In column 4, we report the mean SS for the top ten largest enriched annotations (column 6), i.e., ranking them in descending order according to the number of neighborhoods that the annotations are enriched in (column 5).

$\widetilde{A}_{G_i}$	**Total annotations**	**Mean SS**	**Mean SS Top 10**	**EN**	**Annotation**
$\widetilde{A}_{G_6}$	27	0.28 (std=0.05)	0.40 (std=0.03)	398.0	nucleobase-containing compound catabolic process
$\widetilde{A}_{G_6}$	27	0.28 (std=0.05)	0.40 (std=0.03)	330.0	heterocycle catabolic process
$\widetilde{A}_{G_6}$	27	0.28 (std=0.05)	0.40 (std=0.03)	326.0	aromatic compound catabolic process
$\widetilde{A}_{G_6}$	27	0.28 (std=0.05)	0.40 (std=0.03)	323.0	cellular nitrogen compound catabolic process
$\widetilde{A}_{G_6}$	27	0.28 (std=0.05)	0.40 (std=0.03)	283.0	organic cyclic compound catabolic process
$\widetilde{A}_{G_6}$	27	0.28 (std=0.05)	0.40 (std=0.03)	186.0	nucleic acid phosphodiester bond hydrolysis
$\widetilde{A}_{G_6}$	27	0.28 (std=0.05)	0.40 (std=0.03)	152.0	RNA phosphodiester bond hydrolysis
$\widetilde{A}_{G_6}$	27	0.28 (std=0.05)	0.40 (std=0.03)	15.0	regulation of mitotic sister chromatid separation
$\widetilde{A}_{G_6}$	27	0.28 (std=0.05)	0.40 (std=0.03)	15.0	mitotic spindle checkpoint
$\widetilde{A}_{G_6}$	27	0.28 (std=0.05)	0.40 (std=0.03)	15.0	regulation of chromosome separation
$\widetilde{A}_{G_7}$	3	0.38 (std=0.01)	0.38 (std=0.01)	192.0	regulation of response to DNA damage stimulus
$\widetilde{A}_{G_7}$	3	0.38 (std=0.01)	0.38 (std=0.01)	170.0	histone deubiquitination
$\widetilde{A}_{G_7}$	3	0.38 (std=0.01)	0.38 (std=0.01)	128.0	regulation of transcription involved in G1/S transition of mitotic cell cycle
$\widetilde{A}_{G_8}$	13	0.21 (std=0.05)	0.25 (std=0.06)	300.0	organic substance biosynthetic process
$\widetilde{A}_{G_8}$	13	0.21 (std=0.05)	0.25 (std=0.06)	288.0	transfer RNA gene-mediated silencing
$\widetilde{A}_{G_8}$	13	0.21 (std=0.05)	0.25 (std=0.06)	282.0	regulation of histone ubiquitination
$\widetilde{A}_{G_8}$	13	0.21 (std=0.05)	0.25 (std=0.06)	280.0	biosynthetic process
$\widetilde{A}_{G_8}$	13	0.21 (std=0.05)	0.25 (std=0.06)	217.0	regulation of phosphorylation of RNA polymerase II C-terminal domain serine 2 residues
$\widetilde{A}_{G_8}$	13	0.21 (std=0.05)	0.25 (std=0.06)	217.0	positive regulation of phosphorylation of RNA polymerase II C-terminal domain
$\widetilde{A}_{G_8}$	13	0.21 (std=0.05)	0.25 (std=0.06)	217.0	positive regulation of phosphorylation of RNA polymerase II C-terminal domain serine 2 residues
$\widetilde{A}_{G_8}$	13	0.21 (std=0.05)	0.25 (std=0.06)	213.0	transcription initiation from RNA polymerase II promoter
$\widetilde{A}_{G_8}$	13	0.21 (std=0.05)	0.25 (std=0.06)	190.0	regulation of histone methylation
$\widetilde{A}_{G_8}$	13	0.21 (std=0.05)	0.25 (std=0.06)	180.0	ncRNA 3'-end processing

Table 5.4. Summary of uniquely enriched GO-BPs for Gracoal embeddings, Part 3. We report, for the Budding yeast PPI network, the number of uniquely enriched GO-BPs and the mean semantic similarity (SS) between the uniquely enriched annotations (GO-BPs) (columns 2 and 3) for GraCoals based on all graphlet adjacencies for up to four node graphlets, i.e. $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$ (column 1). In column 4, we report the mean SS for the top ten largest enriched annotations (column 6), i.e., ranking them in descending order according to the number of neighborhoods that the annotations are enriched in (column 5).

Organism	**Mean unique GO-BPs**	**Mean enriched neighborhoods**	**Mean SS**
Budding yeast	12.03 (std=6.99)	171.83 (std=134.03)	0.31 (std=0.07)
E. coli	15.33 (std=12.29)	47.72 (std=16.60)	0.52 (std=0.31)
Fission yeast	14.59 (std=6.44)	62.86 (std=59.12)	0.28 (std=0.12)
Fruit fly	31.89 (std=23.11)	192.22 (std=155.52)	0.18 (std=0.04)
House mouse	38.41 (std=21.31)	125.68 (std=81.74)	0.18 (std=0.05)
Human	39.51 (std=27.35)	352.71 (std=156.33)	0.17 (std=0.06)
Roundworm	18.36 (std=12.00)	35.24 (std=39.34)	0.30 (std=0.14)

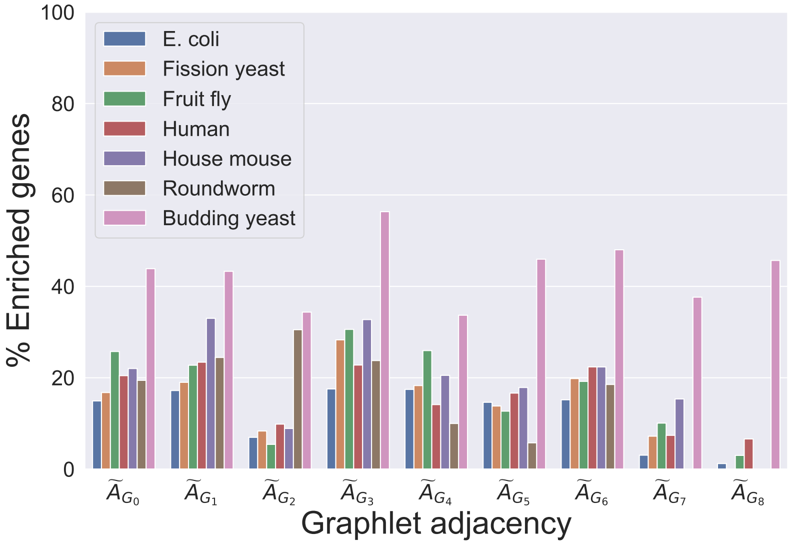
Table 5.5. Summary statistics of uniquely enriched GO-BPs for GraSpring embeddings on PPI networks. We report, for the our seven PPI networks (column 1), the mean number of uniquely enriched GO-BPs obtained with each GraSpring embedding (column 2), the mean size of the uniquely enriched GO-BPs in terms of the number of neighborhoods that the annotations are enriched in (column 3) and the mean semantic similarity (column 4).

5.3 Results for graphlet Spectral with SAFE

In the previous sections we described how GraCoal embeddings and GraSpring embeddings are the best and worst approaches, respectively for uncovering biological information of our PPI molecular networks in terms of genes enriched (Figure 5.1). Thus, GraSpring embeddings perform second best after GraCoals, although the differences are not noticeable. Next, in terms of enriched annotations, graphlet Spectral achieves the best overall percentages of enriched GO-BPs, outperforming the second best, GraCoals, by 1.49% and GraSprings by 4.3%. Lastly, the best overall enrichments in terms of the union of genes enriched and GO-BPs enriched are achieved by graphlet Spectrals on the budding yeast (Figure 5.1 far right).

Graphlet Spectrals uncover complementary biological information in PPI networks

When applying graphlet Spectral embeddings with SAFE to our PPI molecular networks, we can uncover biological information in complementary ways, just as we previously discussed for our PPI networks with GraCoal embeddings and GraSpring embeddings. In Figure 5.4 we show the percentages of genes enriched (top) and percentages of GO-BPs enriched (bottom) for all graphlet Spectral embeddings (i.e., $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$) across our seven PPI networks. In terms of genes enriched, less than half of all graphlet Spectrals across all species achieve greater than 20% enriched genes. In general, the best enrichments are achieved by the budding yeast, with all graphlet Spectrals uncovering between 33.3% ($Spectral_4$) and 57.1% ($Spectral_3$) genes enriched. For the other PPI networks, the enrichments in terms of genes are more balanced and thus there is no clear second best performer. Interestingly, $Spectral_3$ tends to outperform the other Spectrals except in house mouse and round worm, where the best performing are $Spectral_1$ and $Spectral_2$, respectively.



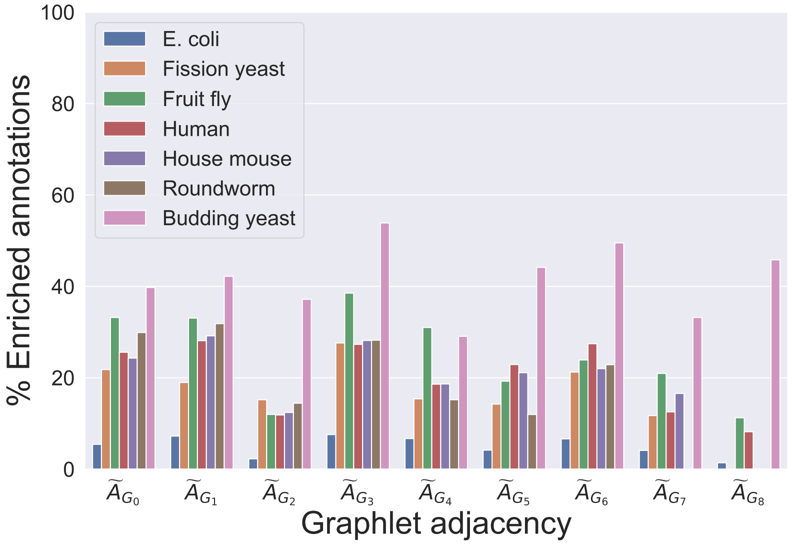


Figure 5.4. SAFE GO-BP enrichment analysis comparing graphlet Spectrals in PPI networks. For the PPI networks of our seven species (legend), we show, on the y-axis, the percentage of enriched genes (top) and the percentage of enriched annotations (bottom) for each of the different graphlet Spectral embeddings (x-axis).

In terms of annotations enriched, most graphlet Spectrals achieve less than 20% enriched GO-BPs on all species, except for budding yeast, which performs the best. In this regard, the percentages of enriched GO-BPs for budding yeast are more or less consistent with the previous results for genes enriched, achieving between 32.9% ($Spectral_4$) and 54.4% ($Spectral_3$) enriched GO-BPs. For *E. coli*, we observe a similar pattern as with GraSpring embeddings, where graphlet Spectrals achieve the worst in terms of enriched GO-BPs.

Next, we focus on identifying what characterizes each particular graphlet Spectral (i.e., $\widetilde{A}_{G_0}$-$\widetilde{A}_{G_8}$) from a biological perspective, just as we did previously for GraCoal embeddings and GraSpring embeddings. In table 5.6 we present the same format as before: we report for the budding yeast, the number of uniquely enriched annotations (column 1), the mean SS for the uniquely enriched annotations (column 2) as well as the mean SS for the top 10 largest uniquely enriched annotations (column 3) for each graphlet Spectral used in SAFE. Finally, in column 5, we report the names of the top 10 uniquely enriched annotations and their corresponding size in terms of enriched neighborhoods (column 4). We find that on average, 43.11 GO-BPs are uniquely enriched for each particular graphlet Spectral. We first note that every graphlet Spectral uncovers unique biological information in the budding yeast, i.e., this is consistent with the claim that complementary information can be captured by different graphlet adjacencies (Windels et al., 2019). Next, the mean semantic similarity (column 2) of each particular graphlet Spectral represents some degree of functional relevance of the GO-BPs uncovered by the different embeddings in SAFE. For instance, the lowest mean semantic similarity is achieved by $GraSpring_3$ at 0.12 (std=0.04) and the largest semantic similarity (i.e., of highest functional relevance) is achieved by $GraSpring_7$ at 0.30 (std=0.05). When evaluating the top 10 enriched annotations, we observe a larger degree of functional similarity between the enriched GO-BPs. The lowest and highest semantic similarities for the top 10 enriched GO-BPs are 0.21 (Std=0.04) for $Spectral_1$ and 0.47 (Std=0.03) for $Spectral_6$. Finally, the average size of the uniquely enriched GO-BPs uncovered by graphlet Spectrals in the budding yeast is 355.67 (std=229.33). For our other PPI molecular networks, we summarize these statistics in terms of the mean uniquely enriched GO-BPs, mean size of enriched GO-BPs (i.e., in mean number of neighborhoods they are enriched in) and mean semantic similarity in Table 5.7.

In general, we observe that graphlet Spectral embeddings uncover complementary information in all of our PPI molecular networks (i.e., uniquely enriched annotations across all graphlet Spectrals, represented by the means in column 2 in Table 5.7). We observe graphlet Spectral embeddings uncover, in general, more unique functional information than GraSpring embeddings, being *E. coli* and fission yeast the exception. With respect to GraCoal embeddings, graphlet Spectrals uncover more unique functional information in yeast, *E. coli* and roundworm, but not in the other four species. For instance, for budding yeast on average 43.11 unique GO-BPs can be captured with any particular graphlet Spectral embedding, which is around 30% more than with GraCoal embeddings (30.12 uniquely enriched GO-BPs on average for budding yeast).

$\widetilde{A}_{G_i}$	**Total annotations**	**Mean SS**	**Mean SS Top 10**	**EN**	**Annotation**
$\widetilde{A}_{G_0}$	46	0.16 (std=0.04)	0.29 (std=0.06)	556.0	organic substance biosynthetic process
$\widetilde{A}_{G_0}$	46	0.16 (std=0.04)	0.29 (std=0.06)	522.0	cellular biosynthetic process
$\widetilde{A}_{G_0}$	46	0.16 (std=0.04)	0.29 (std=0.06)	519.0	biosynthetic process
$\widetilde{A}_{G_0}$	46	0.16 (std=0.04)	0.29 (std=0.06)	355.0	ncRNA transcription
$\widetilde{A}_{G_0}$	46	0.16 (std=0.04)	0.29 (std=0.06)	283.0	mitochondrial membrane organization
$\widetilde{A}_{G_0}$	46	0.16 (std=0.04)	0.29 (std=0.06)	181.0	pre-replicative complex assembly involved in nuclear cell cycle DNA replication
$\widetilde{A}_{G_0}$	46	0.16 (std=0.04)	0.29 (std=0.06)	181.0	pre-replicative complex assembly
$\widetilde{A}_{G_0}$	46	0.16 (std=0.04)	0.29 (std=0.06)	181.0	pre-replicative complex assembly involved in cell cycle DNA replication
$\widetilde{A}_{G_0}$	46	0.16 (std=0.04)	0.29 (std=0.06)	96.0	histone deubiquitination
$\widetilde{A}_{G_0}$	46	0.16 (std=0.04)	0.29 (std=0.06)	75.0	mitochondrial translational initiation
$\widetilde{A}_{G_1}$	50	0.13 (std=0.04)	0.21 (std=0.04)	457.0	cellular localization
$\widetilde{A}_{G_1}$	50	0.13 (std=0.04)	0.21 (std=0.04)	347.0	regulation of cell cycle phase transition
$\widetilde{A}_{G_1}$	50	0.13 (std=0.04)	0.21 (std=0.04)	347.0	regulation of mitotic cell cycle phase transition
$\widetilde{A}_{G_1}$	50	0.13 (std=0.04)	0.21 (std=0.04)	336.0	protein localization to chromosome
$\widetilde{A}_{G_1}$	50	0.13 (std=0.04)	0.21 (std=0.04)	296.0	autophagy of peroxisome
$\widetilde{A}_{G_1}$	50	0.13 (std=0.04)	0.21 (std=0.04)	246.0	negative regulation of chromosome organization
$\widetilde{A}_{G_1}$	50	0.13 (std=0.04)	0.21 (std=0.04)	213.0	microtubule cytoskeleton organization involved in mitosis
$\widetilde{A}_{G_1}$	50	0.13 (std=0.04)	0.21 (std=0.04)	185.0	Golgi vesicle transport
$\widetilde{A}_{G_1}$	50	0.13 (std=0.04)	0.21 (std=0.04)	174.0	translational termination
$\widetilde{A}_{G_1}$	50	0.13 (std=0.04)	0.21 (std=0.04)	123.0	endonucleolytic cleavage to generate mature 3'-end of SSU-rRNA from (SSU-rRNA, 5.8S rRNA, LSU-rRNA)
$\widetilde{A}_{G_2}$	21	0.19 (std=0.05)	0.24 (std=0.03)	461.0	regulation of cell cycle
$\widetilde{A}_{G_2}$	21	0.19 (std=0.05)	0.24 (std=0.03)	308.0	regulation of protein modification process
$\widetilde{A}_{G_2}$	21	0.19 (std=0.05)	0.24 (std=0.03)	249.0	regulation of phosphorus metabolic process
$\widetilde{A}_{G_2}$	21	0.19 (std=0.05)	0.24 (std=0.03)	249.0	regulation of phosphate metabolic process
$\widetilde{A}_{G_2}$	21	0.19 (std=0.05)	0.24 (std=0.03)	183.0	negative regulation of MAPK cascade
$\widetilde{A}_{G_2}$	21	0.19 (std=0.05)	0.24 (std=0.03)	182.0	osmosensory signaling pathway via Sho1 osmosensor
$\widetilde{A}_{G_2}$	21	0.19 (std=0.05)	0.24 (std=0.03)	136.0	negative regulation of protein kinase activity
$\widetilde{A}_{G_2}$	21	0.19 (std=0.05)	0.24 (std=0.03)	133.0	negative regulation of DNA damage checkpoint
$\widetilde{A}_{G_2}$	21	0.19 (std=0.05)	0.24 (std=0.03)	42.0	chromatin remodeling
$\widetilde{A}_{G_2}$	21	0.19 (std=0.05)	0.24 (std=0.03)	41.0	fungal-type cell wall chitin biosynthetic process

Table 5.6. Summary of uniquely enriched GO-BPs for graphlet Spectral embeddings, Part 1. We report, for the Budding yeast PPI network, the number of uniquely enriched GO-BPs and the mean semantic similarity (SS) between the uniquely enriched annotations (GO-BPs) (columns 2 and 3) for graphlet Spectral embeddings based on all graphlet adjacencies for up to four node graphlets, i.e. $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$ (column 1). In column 4, we report the mean SS for the top ten largest enriched annotations (column 6), i.e., ranking them in descending order according to the number of neighborhoods that the annotations are enriched in (column 5).

$\widetilde{A}_{G_i}$	**Total annotations**	**Mean SS**	**Mean SS Top 10**	**EN**	**Annotation**
$\widetilde{A}_{G_3}$	75	0.12 (std=0.03)	0.26 (std=0.04)	863.0	regulation of cellular macromolecule biosynthetic process
$\widetilde{A}_{G_3}$	75	0.12 (std=0.03)	0.26 (std=0.04)	796.0	organic cyclic compound metabolic process
$\widetilde{A}_{G_3}$	75	0.12 (std=0.03)	0.26 (std=0.04)	758.0	DNA metabolic process
$\widetilde{A}_{G_3}$	75	0.12 (std=0.03)	0.26 (std=0.04)	684.0	cellular response to stress
$\widetilde{A}_{G_3}$	75	0.12 (std=0.03)	0.26 (std=0.04)	466.0	telomere organization
$\widetilde{A}_{G_3}$	75	0.12 (std=0.03)	0.26 (std=0.04)	464.0	DNA biosynthetic process
$\widetilde{A}_{G_3}$	75	0.12 (std=0.03)	0.26 (std=0.04)	419.0	negative regulation of mitotic cell cycle
$\widetilde{A}_{G_3}$	75	0.12 (std=0.03)	0.26 (std=0.04)	397.0	positive regulation of cell cycle process
$\widetilde{A}_{G_3}$	75	0.12 (std=0.03)	0.26 (std=0.04)	377.0	RNA-dependent DNA biosynthetic process
$\widetilde{A}_{G_3}$	75	0.12 (std=0.03)	0.26 (std=0.04)	373.0	telomere maintenance via telomerase
$\widetilde{A}_{G_4}$	12	0.25 (std=0.04)	0.27 (std=0.03)	497.0	mRNA metabolic process
$\widetilde{A}_{G_4}$	12	0.25 (std=0.04)	0.27 (std=0.03)	487.0	regulation of nitrogen compound metabolic process
$\widetilde{A}_{G_4}$	12	0.25 (std=0.04)	0.27 (std=0.03)	484.0	regulation of biosynthetic process
$\widetilde{A}_{G_4}$	12	0.25 (std=0.04)	0.27 (std=0.03)	480.0	regulation of cellular biosynthetic process
$\widetilde{A}_{G_4}$	12	0.25 (std=0.04)	0.27 (std=0.03)	458.0	regulation of macromolecule metabolic process
$\widetilde{A}_{G_4}$	12	0.25 (std=0.04)	0.27 (std=0.03)	184.0	organonitrogen compound biosynthetic process
$\widetilde{A}_{G_4}$	12	0.25 (std=0.04)	0.27 (std=0.03)	182.0	regulation of chromosome segregation
$\widetilde{A}_{G_4}$	12	0.25 (std=0.04)	0.27 (std=0.03)	163.0	translation reinitiation
$\widetilde{A}_{G_4}$	12	0.25 (std=0.04)	0.27 (std=0.03)	119.0	cell redox homeostasis
$\widetilde{A}_{G_4}$	12	0.25 (std=0.04)	0.27 (std=0.03)	104.0	recombinational repair
$\widetilde{A}_{G_5}$	52	0.15 (std=0.03)	0.31 (std=0.06)	689.0	cellular macromolecule metabolic process
$\widetilde{A}_{G_5}$	52	0.15 (std=0.03)	0.31 (std=0.06)	499.0	cellular response to DNA damage stimulus
$\widetilde{A}_{G_5}$	52	0.15 (std=0.03)	0.31 (std=0.06)	466.0	response to stimulus
$\widetilde{A}_{G_5}$	52	0.15 (std=0.03)	0.31 (std=0.06)	449.0	response to stress
$\widetilde{A}_{G_5}$	52	0.15 (std=0.03)	0.31 (std=0.06)	428.0	DNA repair
$\widetilde{A}_{G_5}$	52	0.15 (std=0.03)	0.31 (std=0.06)	421.0	regulation of RNA metabolic process
$\widetilde{A}_{G_5}$	52	0.15 (std=0.03)	0.31 (std=0.06)	417.0	regulation of nucleic acid-templated transcription
$\widetilde{A}_{G_5}$	52	0.15 (std=0.03)	0.31 (std=0.06)	417.0	regulation of RNA biosynthetic process
$\widetilde{A}_{G_5}$	52	0.15 (std=0.03)	0.31 (std=0.06)	417.0	regulation of transcription, DNA-templated
$\widetilde{A}_{G_5}$	52	0.15 (std=0.03)	0.31 (std=0.06)	387.0	modification-dependent protein catabolic process

Table 5.6. Summary of uniquely enriched GO-BPs for graphlet Spectral embeddings, Part 2. We report, for the Budding yeast PPI network, the number of uniquely enriched GO-BPs and the mean semantic similarity (SS) between the uniquely enriched annotations (GO-BPs) (columns 2 and 3) for graphlet Spectral embeddings based on all graphlet adjacencies for up to four node graphlets, i.e. $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$ (column 1). In column 4, we report the mean SS for the top ten largest enriched annotations (column 6), i.e., ranking them in descending order according to the number of neighborhoods that the annotations are enriched in (column 5).

$\widetilde{A}_{G_i}$	**Total annotations**	**Mean SS**	**Mean SS Top 10**	**EN**	**Annotation**
$\widetilde{A}_{G_6}$	56	0.15 (std=0.03)	0.47 (std=0.03)	879.0	cellular component organization
$\widetilde{A}_{G_6}$	56	0.15 (std=0.03)	0.47 (std=0.03)	761.0	negative regulation of nucleobase-containing compound metabolic process
$\widetilde{A}_{G_6}$	56	0.15 (std=0.03)	0.47 (std=0.03)	751.0	negative regulation of macromolecule biosynthetic process
$\widetilde{A}_{G_6}$	56	0.15 (std=0.03)	0.47 (std=0.03)	751.0	negative regulation of cellular macromolecule biosynthetic process
$\widetilde{A}_{G_6}$	56	0.15 (std=0.03)	0.47 (std=0.03)	721.0	negative regulation of biosynthetic process
$\widetilde{A}_{G_6}$	56	0.15 (std=0.03)	0.47 (std=0.03)	721.0	negative regulation of cellular biosynthetic process
$\widetilde{A}_{G_6}$	56	0.15 (std=0.03)	0.47 (std=0.03)	712.0	negative regulation of RNA biosynthetic process
$\widetilde{A}_{G_6}$	56	0.15 (std=0.03)	0.47 (std=0.03)	712.0	negative regulation of transcription, DNA-templated
$\widetilde{A}_{G_6}$	56	0.15 (std=0.03)	0.47 (std=0.03)	712.0	negative regulation of nucleic acid-templated transcription
$\widetilde{A}_{G_6}$	56	0.15 (std=0.03)	0.47 (std=0.03)	703.0	negative regulation of RNA metabolic process
$\widetilde{A}_{G_7}$	13	0.30 (std=0.05)	0.38 (std=0.03)	496.0	positive regulation of biological process
$\widetilde{A}_{G_7}$	13	0.30 (std=0.05)	0.38 (std=0.03)	469.0	positive regulation of cellular process
$\widetilde{A}_{G_7}$	13	0.30 (std=0.05)	0.38 (std=0.03)	313.0	regulation of cytoskeleton organization
$\widetilde{A}_{G_7}$	13	0.30 (std=0.05)	0.38 (std=0.03)	35.0	microtubule-based transport
$\widetilde{A}_{G_7}$	13	0.30 (std=0.05)	0.38 (std=0.03)	35.0	transport along microtubule
$\widetilde{A}_{G_7}$	13	0.30 (std=0.05)	0.38 (std=0.03)	35.0	microtubule-based movement
$\widetilde{A}_{G_7}$	13	0.30 (std=0.05)	0.38 (std=0.03)	35.0	nuclear migration along microtubule
$\widetilde{A}_{G_7}$	13	0.30 (std=0.05)	0.38 (std=0.03)	35.0	movement of cell or subcellular component
$\widetilde{A}_{G_7}$	13	0.30 (std=0.05)	0.38 (std=0.03)	35.0	organelle transport along microtubule
$\widetilde{A}_{G_7}$	13	0.30 (std=0.05)	0.38 (std=0.03)	35.0	cytoskeleton-dependent intracellular transport
$\widetilde{A}_{G_8}$	63	0.13 (std=0.05)	0.29 (std=0.04)	342.0	regulation of cell cycle process
$\widetilde{A}_{G_8}$	63	0.13 (std=0.05)	0.29 (std=0.04)	283.0	regulation of signal transduction
$\widetilde{A}_{G_8}$	63	0.13 (std=0.05)	0.29 (std=0.04)	283.0	regulation of signaling
$\widetilde{A}_{G_8}$	63	0.13 (std=0.05)	0.29 (std=0.04)	236.0	process utilizing autophagic mechanism
$\widetilde{A}_{G_8}$	63	0.13 (std=0.05)	0.29 (std=0.04)	236.0	regulation of intracellular signal transduction
$\widetilde{A}_{G_8}$	63	0.13 (std=0.05)	0.29 (std=0.04)	216.0	protein localization by the Cvt pathway
$\widetilde{A}_{G_8}$	63	0.13 (std=0.05)	0.29 (std=0.04)	164.0	endoplasmic reticulum to Golgi vesicle-mediated transport
$\widetilde{A}_{G_8}$	63	0.13 (std=0.05)	0.29 (std=0.04)	147.0	positive regulation of cell communication
$\widetilde{A}_{G_8}$	63	0.13 (std=0.05)	0.29 (std=0.04)	91.0	response to endoplasmic reticulum stress
$\widetilde{A}_{G_8}$	63	0.13 (std=0.05)	0.29 (std=0.04)	64.0	respiratory chain complex II assembly

Table 5.6. Summary of uniquely enriched GO-BPs for graphlet Spectral embeddings, Part 3. We report, for the Budding yeast PPI network, the number of uniquely enriched GO-BPs and the mean semantic similarity (SS) between the uniquely enriched annotations (GO-BPs) (columns 2 and 3) for graphlet Spectral embeddings based on all graphlet adjacencies for up to four node graphlets, i.e. $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$ (column 1). In column 4, we report the mean SS for the top ten largest enriched annotations (column 6), i.e., ranking them in descending order according to the number of neighborhoods that the annotations are enriched in (column 5).

Organism	**Mean unique GO-BPs**	**Mean enriched neighborhoods**	**Mean SS**
Budding yeast	43.11 (std=21.40)	355.67 (std=229.33)	0.18 (std=0.06)
E. coli	4.69 (std=1.32)	63.00 (std=67.45)	0.57 (std=0.16)
Fission yeast	7.58 (std=5.34)	111.16 (std=122.16)	0.45 (std=0.23)
Fruit fly	26.95 (std=14.36)	96.19 (std=118.52)	0.23 (std=0.12)
House mouse	36.61 (std=18.33)	389.73 (std=353.06)	0.18 (std=0.06)
Human	35.29 (std=28.04)	271.39 (std=367.03)	0.22 (std=0.11)
Roundworm	22.20 (std=13.13)	63.36 (std=63.95)	0.25 (std=0.10)

Table 5.7. Summary statistics of uniquely enriched GO-BPs for graphlet Spectral embeddings on PPI networks. We report, for the our seven PPI networks (column 1), the mean number of uniquely enriched GO-BPs obtained with each graphlet Spectral embedding (column 2), the mean size of the uniquely enriched GO-BPs in terms of the number of neighborhoods that the annotations are enriched in (column 3) and the mean semantic similarity (column 4).

Chapter 6

Application 3: Analysis of COEX networks

In this chapter we evaluate the performance of the graphlet-based embeddings (i.e., GraCoal, GraSpring and graphlet based Spectral) with the Spatial Analysis of Functional Enrichment (SAFE) framework on the COEX networks of the following species: *Drosophila melanogaster*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Homo sapiens*, and *Caenorhabditis elegans* which throughout the text we will refer to as fruit fly, budding yeast, fission yeast, human and roundworm, respectively. We present the COEX network statistics in Table 6.1. For more information on how we built these molecular networks please refer to section A.1 in Appendix A. Moreover, we focus mainly on analysing results based on Gene Ontology Biological Processes (GO-BP), as is one of the most complete set of annotations. For detailed results corresponding to our other annotations (e.g., GO molecular functions and GO cellular components), please refer to section A.4 in Appendix A. In the next sections, we summarize the results obtained by our GraCoal embedding, GraSpring embedding and finally graphlet based Spectral embedding. We also perform model fitting experiments for our COEX networks, see section A.5, to show that none of our COEX networks are well fitted by any of the model networks.

		COEX	
	Nodes	**Edges**	**Density**
Budding yeast	5,879	231,979	0.013
Fission Yeast	5,260	213,970	0.015
Fruit fly	12,173	954,450	0.013
Human	16,795	1,801,441	0.013
Roundworm	14,492	1,324,617	0.013

Table 6.1. COEX molecular network data statistics. For each species (row), we report the number of nodes, the number of edges and the density of the corresponding COEX network (columns 1-3).

6.1 Results for GraCoal with SAFE

In general, we observe that GraSpring embeddings appear to be the best performing embedding method both in terms of genes enriched and annotations enriched

(Figure 6.1). For instance, when we consider the union of genes enriched over the different graphlet based embeddings (i.e., over $\widetilde{A}_0$ to $\widetilde{A}_8$) GraSpring outperforms both GraCoal embeddings and graphlet Spectral embeddings on three of the five COEX networks (GraCoal performs best on human and fission yeast). On average, GraSpring outperforms GraCoal and graphlet Spectrals by 5.61% and 10.73%, respectively. Next, when we consider the union of GO-BP annotations enriched, GraSpring embeddings outperform GraCoal embeddings and graphlet Spectral embeddings on every COEX network (except human) by 2.75% and 4.71%, respectively (Figure 6.1). Lastly, the differences between GraCoal embeddings and graphlet Spectral embeddings are not as noticeable, although GraCoal achieves the second best enrichments overall, outperforming graphlet Spectral by 5.12% for enriched genes and 2.1% for enriched GO-BP.

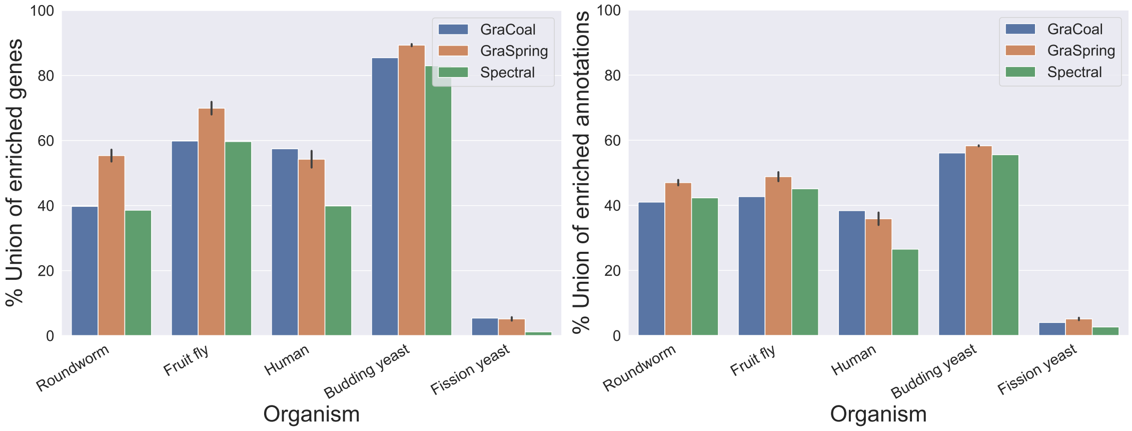


Figure 6.1. SAFE GO-BP enrichment analysis for COEX networks. For the COEX networks of our six species (x-axis), we show the percentage of enriched genes (y-axis) and percentage of enriched annotations for each of the embedding algorithms considered (legend). In the case of GraSpring, we show the average across ten randomised runs and the standard deviation (error-bars).

On the other hand, when we take individual enrichments as opposed to the union over all graphlet based embeddings, we observe that the best scoring GraSpring embeddings (Figure 6.3) outperform the best GraCoal embeddings (Figure 6.2) and graphlet Spectral embeddings (Figure 6.4), on average by 4.89% and 6.77, respectively in terms of genes enriched. This is also the case when we consider GO-BPs enrichment. That is, the best scoring GraSpring embedddings outperform the best scoring GraCoal embeddings and graphlet Spectral embeddings, on average by 2.76% and 5.69%, respectively. We observe similar results for GO-CC annotations on our COEX networks (Section A.4 in Appendix A). For instance for genes enriched in GO-CC annotations, the best GraSpring embeddings outperform the best GraCoal embeddings and graphlet Spectral embeddings by 6.85% and 7.94%, respectively.

GraCoals uncover complementary biological information in COEX networks

When applying GraCoal embeddings with SAFE to our COEX molecular networks, we can uncover biological information in complementary ways, just as we previously discussed for our GI and PPI molecular networks. In Figure 6.2 we show the percentages of genes enriched (top) and percentages of GO-BPs enriched (bottom) for all GraCoals (i.e., $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$) across our six COEX networks. We observe the largest percentages of both genes enriched and annotations enriched for the budding yeast COEX network. In terms of genes, GraCoal embeddings achieve at least 40% genes enriched in GO-BPs across all GraCoals. For GO-BP enrichment, these values range between 33.4% and 40.2% for $GraCoal_4$ and $GraCoal_8$, respectively. For genes enriched, we observe that fruit fly and human achieve very similar enrichment percentrages across all GraCoals (between 18.1% and 23.9%). On the other hand, roundworm (red label) performs half as good as fruit fly and human, achieving between 10.2% and 16.5% enriched genes. Finally, fission yeast is the worst, achieving close to 0% in most cases, except for $GraCoal_{5,8}$ which achieve close to 5% enriched genes. For GO-BPs enrichments, fruit fly, human and roundworm all perform very similar, achieving between 10.2% and 22.3% enriched GO-BPs. Finally, GraCoal embeddings achieve the worst performance in terms of GO-BPs enrichments with fission yeast COEX network, achieving close to 0% enriched GO-BPs across all GraCoals (purple label).

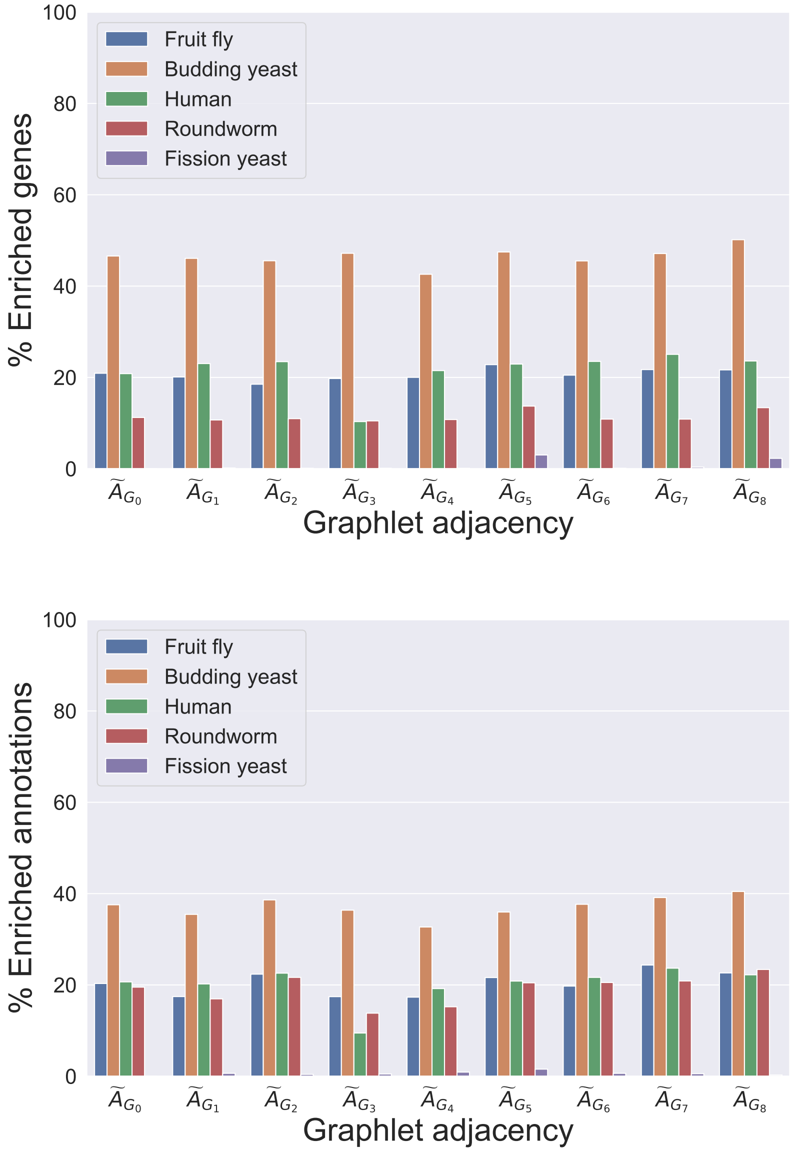


Figure 6.2. SAFE GO-BP enrichment analysis comparing GraCoals in COEX networks. For the COEX networks of six seven species (legend), we show, on the y-axis, the percentage of enriched genes (top) and the percentage of enriched annotations (bottom) for each of the different GraCoal embeddings (x-axis).

Additionally, for each specie, we focus on identifying what characterizes each particular GraCoal (i.e., $\widetilde{A}_{G_0}$-$\widetilde{A}_{G_8}$) from a biological perspective, just as we did previously for our GI and PPI networks with our graphlet based embeddings. In table 6.2 we present the same format as previously presented for the PPI budding yeast network: we report for the budding yeast, the number of uniquely enriched annotations (column 1), the mean SS for the uniquely enriched annotations (column 2) as well as the mean SS for the top 10 largest uniquely enriched annotations (column 3) for each GraCoal used in SAFE. Finally, in column 5, we report the names of the top 10 uniquely enriched annotations and their corresponding size in terms of enriched neighborhoods (column 4). We find that on average, 37.57 GO-BPs are uniquely enriched for each particular GraCoal (Table 6.2). Next, we observe that every GraCoal uncovers unique biological information (i.e., GO-BPs), validating the claim that graphlet adjacencies capture complementary biological information (Windels et al., 2019). Next, we validate if the sets of uniquely enriched GO-BPs are biologically relevant (i.e., are functionally related). The mean semantic similarities for all GraCoal embeddings used with SAFE on the budding yeast COEX network, ranges from 0.13 (std=0.03) for $GraCoal_8$ to 0.89 (std=0.17) for $GraCoal_1$, indicating some degree of functional of functional similarity, with $GraCoal_8$ capturing the most functional relevance between uniquely enriched GO-BPs. Next, we rank the uniquely enriched annotations according to their size, defined as the total number of neighborhoods they are enriched in, as a measure of how well they are captured by each particular GraCoal. The average size of the uniquely enriched GO-BPs uncovered by GraCoals in the budding yeast is 92.64 (std=52.23). For our other COEX molecular networks, we summarize these statistics in terms of the mean uniquely enriched GO-BPs, mean size of enriched GO-BPs (i.e., in mean number of neighborhoods they are enriched in) and mean semantic similarity in Table 6.3.

In general, we observe that GraCoal embeddings uncover complementary information in all of our COEX molecular networks except for fission yeast (i.e., uniquely enriched annotations across all GraCoals, represented by the means in column 2 in Table 6.3). In this regard, fission yeast only achieves 9 uniquely enriched GO-BPs with $GraCoal_5$ and 0 with the other GraCoals.

$\widetilde{A}_{G_i}$	**Total annotations**	**Mean SS**	**Mean SS Top 10**	**EN**	**Annotation**
$\widetilde{A}_{G_0}$	16	0.29 (std=0.05)	0.35 (std=0.05)	50.0	cysteine biosynthetic process
$\widetilde{A}_{G_0}$	16	0.29 (std=0.05)	0.35 (std=0.05)	50.0	cysteine biosynthetic process via cystathionine
$\widetilde{A}_{G_0}$	16	0.29 (std=0.05)	0.35 (std=0.05)	48.0	homocysteine metabolic process
$\widetilde{A}_{G_0}$	16	0.29 (std=0.05)	0.35 (std=0.05)	43.0	leucine biosynthetic process
$\widetilde{A}_{G_0}$	16	0.29 (std=0.05)	0.35 (std=0.05)	38.0	leucine metabolic process
$\widetilde{A}_{G_0}$	16	0.29 (std=0.05)	0.35 (std=0.05)	26.0	pyrimidine deoxyribonucleoside triphosphate catabolic process
$\widetilde{A}_{G_0}$	16	0.29 (std=0.05)	0.35 (std=0.05)	26.0	nucleoside triphosphate catabolic process
$\widetilde{A}_{G_0}$	16	0.29 (std=0.05)	0.35 (std=0.05)	26.0	pyrimidine nucleoside triphosphate catabolic process
$\widetilde{A}_{G_0}$	16	0.29 (std=0.05)	0.35 (std=0.05)	26.0	deoxyribonucleoside triphosphate catabolic process
$\widetilde{A}_{G_0}$	16	0.29 (std=0.05)	0.35 (std=0.05)	26.0	purine deoxyribonucleoside triphosphate metabolic process
$\widetilde{A}_{G_1}$	4	0.89 (std=0.17)	0.89 (std=0.17)	140.0	protein localization to organelle
$\widetilde{A}_{G_1}$	4	0.89 (std=0.17)	0.89 (std=0.17)	112.0	cellular protein localization
$\widetilde{A}_{G_1}$	4	0.89 (std=0.17)	0.89 (std=0.17)	103.0	cellular macromolecule localization
$\widetilde{A}_{G_1}$	4	0.89 (std=0.17)	0.89 (std=0.17)	16.0	drug catabolic process
$\widetilde{A}_{G_2}$	77	0.16 (std=0.04)	0.41 (std=0.05)	99.0	RNA biosynthetic process
$\widetilde{A}_{G_2}$	77	0.16 (std=0.04)	0.41 (std=0.05)	91.0	positive regulation of cellular process
$\widetilde{A}_{G_2}$	77	0.16 (std=0.04)	0.41 (std=0.05)	90.0	positive regulation of biological process
$\widetilde{A}_{G_2}$	77	0.16 (std=0.04)	0.41 (std=0.05)	73.0	positive regulation of metabolic process
$\widetilde{A}_{G_2}$	77	0.16 (std=0.04)	0.41 (std=0.05)	72.0	positive regulation of cellular metabolic process
$\widetilde{A}_{G_2}$	77	0.16 (std=0.04)	0.41 (std=0.05)	69.0	positive regulation of nucleobase-containing compound metabolic process
$\widetilde{A}_{G_2}$	77	0.16 (std=0.04)	0.41 (std=0.05)	68.0	negative regulation of gene expression
$\widetilde{A}_{G_2}$	77	0.16 (std=0.04)	0.41 (std=0.05)	67.0	positive regulation of nucleic acid-templated transcription
$\widetilde{A}_{G_2}$	77	0.16 (std=0.04)	0.41 (std=0.05)	67.0	positive regulation of transcription, DNA-templated
$\widetilde{A}_{G_2}$	77	0.16 (std=0.04)	0.41 (std=0.05)	67.0	positive regulation of RNA biosynthetic process

Table 6.2. Summary of uniquely enriched GO-BPs for GraCoal embeddings, Part 1. We report, for the Budding yeast COEX network, the number of uniquely enriched GO-BPs and the mean semantic similarity (SS) between the uniquely enriched annotations (GO-BPs) (columns 2 and 3) for GraCoals based on all graphlet adjacencies for up to four node graphlets, i.e. $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$ (column 1). In column 4, we report the mean SS for the top ten largest enriched annotations (column 6), i.e., ranking them in descending order according to the number of neighborhoods that the annotations are enriched in (column 5).

$\widetilde{A}_{G_i}$	**Total annotations**	**Mean SS**	**Mean SS Top 10**	**EN**	**Annotation**
$\widetilde{A}_{G_3}$	13	0.25 (std=0.06)	0.33 (std=0.05)	143.0	transcription by RNA polymerase I
$\widetilde{A}_{G_3}$	13	0.25 (std=0.06)	0.33 (std=0.05)	137.0	transcription, DNA-templated
$\widetilde{A}_{G_3}$	13	0.25 (std=0.06)	0.33 (std=0.05)	137.0	nucleic acid-templated transcription
$\widetilde{A}_{G_3}$	13	0.25 (std=0.06)	0.33 (std=0.05)	116.0	nucleolar large rRNA transcription by RNA polymerase I
$\widetilde{A}_{G_3}$	13	0.25 (std=0.06)	0.33 (std=0.05)	108.0	endonucleolytic cleavage in ITS1 upstream of 5.8S rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)
$\widetilde{A}_{G_3}$	13	0.25 (std=0.06)	0.33 (std=0.05)	89.0	rRNA transcription
$\widetilde{A}_{G_3}$	13	0.25 (std=0.06)	0.33 (std=0.05)	84.0	mRNA pseudouridine synthesis
$\widetilde{A}_{G_3}$	13	0.25 (std=0.06)	0.33 (std=0.05)	75.0	protein-heme linkage
$\widetilde{A}_{G_3}$	13	0.25 (std=0.06)	0.33 (std=0.05)	75.0	cytochrome c-heme linkage
$\widetilde{A}_{G_3}$	13	0.25 (std=0.06)	0.33 (std=0.05)	75.0	protein-tetrapyrrole linkage
$\widetilde{A}_{G_4}$	36	0.17 (std=0.05)	0.27 (std=0.04)	384.0	protein-containing complex assembly
$\widetilde{A}_{G_4}$	36	0.17 (std=0.05)	0.27 (std=0.04)	177.0	mitochondrial proton-transporting ATP synthase complex assembly
$\widetilde{A}_{G_4}$	36	0.17 (std=0.05)	0.27 (std=0.04)	177.0	proton-transporting ATP synthase complex assembly
$\widetilde{A}_{G_4}$	36	0.17 (std=0.05)	0.27 (std=0.04)	172.0	translational frameshifting
$\widetilde{A}_{G_4}$	36	0.17 (std=0.05)	0.27 (std=0.04)	141.0	protein maturation by [4Fe-4S] cluster transfer
$\widetilde{A}_{G_4}$	36	0.17 (std=0.05)	0.27 (std=0.04)	134.0	protein maturation by iron-sulfur cluster transfer
$\widetilde{A}_{G_4}$	36	0.17 (std=0.05)	0.27 (std=0.04)	128.0	mitochondrial RNA processing
$\widetilde{A}_{G_4}$	36	0.17 (std=0.05)	0.27 (std=0.04)	128.0	gene expression
$\widetilde{A}_{G_4}$	36	0.17 (std=0.05)	0.27 (std=0.04)	128.0	mitochondrial gene expression
$\widetilde{A}_{G_4}$	36	0.17 (std=0.05)	0.27 (std=0.04)	117.0	positive regulation of translational elongation
$\widetilde{A}_{G_5}$	11	0.29 (std=0.06)	0.31 (std=0.06)	191.0	response to stress
$\widetilde{A}_{G_5}$	11	0.29 (std=0.06)	0.31 (std=0.06)	188.0	cellular response to stress
$\widetilde{A}_{G_5}$	11	0.29 (std=0.06)	0.31 (std=0.06)	173.0	cellular response to stimulus
$\widetilde{A}_{G_5}$	11	0.29 (std=0.06)	0.31 (std=0.06)	165.0	response to stimulus
$\widetilde{A}_{G_5}$	11	0.29 (std=0.06)	0.31 (std=0.06)	90.0	tricarboxylic acid metabolic process
$\widetilde{A}_{G_5}$	11	0.29 (std=0.06)	0.31 (std=0.06)	86.0	organelle localization
$\widetilde{A}_{G_5}$	11	0.29 (std=0.06)	0.31 (std=0.06)	74.0	protein-DNA complex assembly
$\widetilde{A}_{G_5}$	11	0.29 (std=0.06)	0.31 (std=0.06)	52.0	nucleoside metabolic process
$\widetilde{A}_{G_5}$	11	0.29 (std=0.06)	0.31 (std=0.06)	27.0	homologous recombination
$\widetilde{A}_{G_5}$	11	0.29 (std=0.06)	0.31 (std=0.06)	27.0	reciprocal meiotic recombination

Table 6.2. Summary of uniquely enriched GO-BPs for GraCoal embeddings, Part 2. We report, for the Budding yeast COEX network, the number of uniquely enriched GO-BPs and the mean semantic similarity (SS) between the uniquely enriched annotations (GO-BPs) (columns 2 and 3) for GraCoals based on all graphlet adjacencies for up to four node graphlets, i.e. $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$ (column 1). In column 4, we report the mean SS for the top ten largest enriched annotations (column 6), i.e., ranking them in descending order according to the number of neighborhoods that the annotations are enriched in (column 5).

$\widetilde{A}_{G_i}$	**Total annotations**	**Mean SS**	**Mean SS Top 10**	**EN**	**Annotation**
$\widetilde{A}_{G_6}$	26	0.20 (std=0.04)	0.27 (std=0.05)	91.0	mRNA metabolic process
$\widetilde{A}_{G_6}$	26	0.20 (std=0.04)	0.27 (std=0.05)	88.0	mitochondrial fusion
$\widetilde{A}_{G_6}$	26	0.20 (std=0.04)	0.27 (std=0.05)	73.0	polyadenylation-dependent mRNA catabolic process
$\widetilde{A}_{G_6}$	26	0.20 (std=0.04)	0.27 (std=0.05)	73.0	nuclear polyadenylation-dependent mRNA catabolic process
$\widetilde{A}_{G_6}$	26	0.20 (std=0.04)	0.27 (std=0.05)	64.0	cellular metabolic compound salvage
$\widetilde{A}_{G_6}$	26	0.20 (std=0.04)	0.27 (std=0.05)	54.0	beta-alanine biosynthetic process
$\widetilde{A}_{G_6}$	26	0.20 (std=0.04)	0.27 (std=0.05)	54.0	beta-alanine metabolic process
$\widetilde{A}_{G_6}$	26	0.20 (std=0.04)	0.27 (std=0.05)	51.0	regulation of phosphorus metabolic process
$\widetilde{A}_{G_6}$	26	0.20 (std=0.04)	0.27 (std=0.05)	51.0	regulation of phosphate metabolic process
$\widetilde{A}_{G_6}$	26	0.20 (std=0.04)	0.27 (std=0.05)	49.0	chromatin assembly or disassembly
$\widetilde{A}_{G_7}$	29	0.16 (std=0.04)	0.26 (std=0.05)	154.0	regulation of organelle organization
$\widetilde{A}_{G_7}$	29	0.16 (std=0.04)	0.26 (std=0.05)	123.0	regulation of cellular component organization
$\widetilde{A}_{G_7}$	29	0.16 (std=0.04)	0.26 (std=0.05)	95.0	negative regulation of metabolic process
$\widetilde{A}_{G_7}$	29	0.16 (std=0.04)	0.26 (std=0.05)	91.0	protein import into mitochondrial intermembrane space
$\widetilde{A}_{G_7}$	29	0.16 (std=0.04)	0.26 (std=0.05)	81.0	regulation of gene expression
$\widetilde{A}_{G_7}$	29	0.16 (std=0.04)	0.26 (std=0.05)	76.0	nuclear DNA replication
$\widetilde{A}_{G_7}$	29	0.16 (std=0.04)	0.26 (std=0.05)	75.0	2-oxoglutarate metabolic process
$\widetilde{A}_{G_7}$	29	0.16 (std=0.04)	0.26 (std=0.05)	74.0	aging
$\widetilde{A}_{G_7}$	29	0.16 (std=0.04)	0.26 (std=0.05)	74.0	cell aging
$\widetilde{A}_{G_7}$	29	0.16 (std=0.04)	0.26 (std=0.05)	72.0	dicarboxylic acid metabolic process
$\widetilde{A}_{G_8}$	106	0.13 (std=0.03)	0.29 (std=0.05)	103.0	cellular carbohydrate catabolic process
$\widetilde{A}_{G_8}$	106	0.13 (std=0.03)	0.29 (std=0.05)	103.0	tubulin complex assembly
$\widetilde{A}_{G_8}$	106	0.13 (std=0.03)	0.29 (std=0.05)	101.0	protein refolding
$\widetilde{A}_{G_8}$	106	0.13 (std=0.03)	0.29 (std=0.05)	100.0	disaccharide catabolic process
$\widetilde{A}_{G_8}$	106	0.13 (std=0.03)	0.29 (std=0.05)	99.0	oligosaccharide catabolic process
$\widetilde{A}_{G_8}$	106	0.13 (std=0.03)	0.29 (std=0.05)	89.0	organic acid catabolic process
$\widetilde{A}_{G_8}$	106	0.13 (std=0.03)	0.29 (std=0.05)	89.0	carboxylic acid catabolic process
$\widetilde{A}_{G_8}$	106	0.13 (std=0.03)	0.29 (std=0.05)	86.0	serine family amino acid metabolic process
$\widetilde{A}_{G_8}$	106	0.13 (std=0.03)	0.29 (std=0.05)	82.0	methionine biosynthetic process
$\widetilde{A}_{G_8}$	106	0.13 (std=0.03)	0.29 (std=0.05)	76.0	sulfate reduction

Table 6.2. Summary of uniquely enriched GO-BPs for GraCoal embeddings, Part 3. We report, for the Budding yeast COEX network, the number of uniquely enriched GO-BPs and the mean semantic similarity (SS) between the uniquely enriched annotations (GO-BPs) (columns 2 and 3) for GraCoals based on all graphlet adjacencies for up to four node graphlets, i.e. $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$ (column 1). In column 4, we report the mean SS for the top ten largest enriched annotations (column 6), i.e., ranking them in descending order according to the number of neighborhoods that the annotations are enriched in (column 5).

Organism	**Mean unique GO-BPs**	**Mean enriched neighborhoods**	**Mean SS**
Budding yeast	37.57 (std=32.33)	92.64 (std=52.23)	0.24 (std=0.16)
Fission yeast	1.00 (std=3.00)	23.89 (std=41.67)	0.62 (std=0.00)
Fruit fly	27.91 (std=27.41)	53.40 (std=45.80)	0.23 (std=0.07)
Human	57.22 (std=43.06)	190.02 (std=94.23)	0.15 (std=0.03)
Roundworm	26.53 (std=13.58)	43.97 (std=32.53)	0.25 (std=0.11)

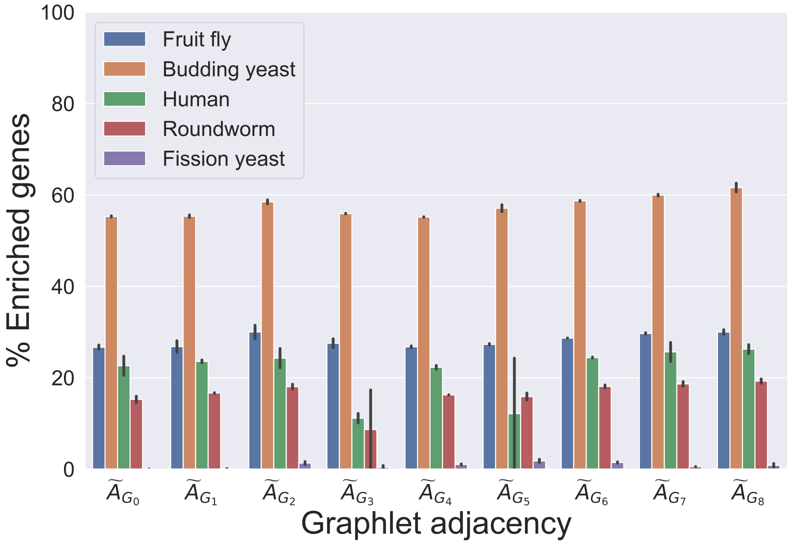
Table 6.3. Summary statistics of uniquely enriched GO-BPs for Gracoal embeddings on COEX networks. We report, for the our six COEX networks (column 1), the mean number of uniquely enriched GO-BPs obtained with each GraCoal embedding (column 2), the mean size of the uniquely enriched GO-BPs in terms of the number of neighborhoods that the annotations are enriched in (column 3) and the mean semantic similarity (column 4).

6.2 Results for GraSpring with SAFE

GraSprings uncover complementary biological information in COEX networks

When applying GraCoal embeddings with SAFE to our COEX molecular networks, we can uncover biological information in complementary ways, just as we previously discussed for our GI and PPI molecular networks. In Figure 6.3 we show the percentages of genes enriched (top) and percentages of GO-BPs enriched (bottom) for all Grasprings (i.e., $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$) across our six COEX networks. We observe the same pattern as when applying GraCoal embeddings on our COEX molecular networks. That is, the largest percentages of both genes enriched and annotations enriched are achieved for the budding yeast COEX network. In terms of genes, all GraSpring embeddings perform better than as previously shown for GraCoals, achieving on average more than 50% of genes enriched. Similarly, we observe that for fruit fly and for human, GraSprings achieve very similar percentages of genes enriched as with GraCoal embeddings across all GraSprings (except for $GraSpring_3$), which range (between 10.1% and 23.7%). Next, the enrichments for roundworm in terms of genes are better with respect to GraCoals, achieving between 8.7% and 19.2%. Finally, fission yeast is the worst, achieving close to 0% in most cases, except for $GraCoal_{2,4,6}$ which achieve close to 3% enriched genes. For GO-BPs enrichments, fruit fly, human and roundworm all perform very similar, achieving between 10.2% and 22.3% enriched GO-BPs. Finally, GraCoal embeddings achieve the worst performance in terms of GO-BPs enrichments with fission yeast COEX network, achieving close to 0% enriched GO-BPs across all GraCoals (purple label).

We observe the same pattern for GO-BP enrichment for the budding yeast (i.e., all GraSprings outperforming GraCoals), achieving between 41.4% and 45.2% for $GraSpring_4$ and $GraSpring_8$, respectively, which interestingly, are the best performers in GraCoal embeddings. As for the other COEX networks, we observe that GraSprings perform second best on fruitfly, followed by roundworm and then human. Finally, the worst enrichments achieved by GraSpring embeddings correspond to the fission yeast COEX network, which is consistent with the low performance in terms of genes enriched.



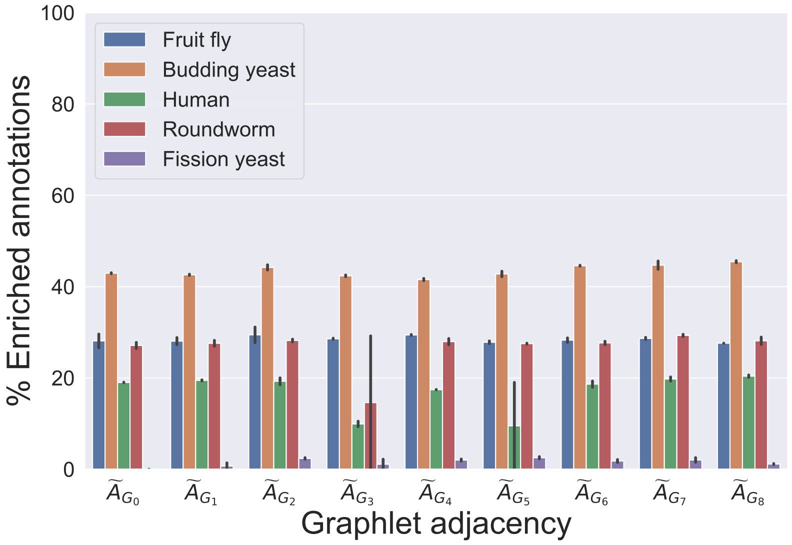


Figure 6.3. SAFE GO-BP enrichment analysis comparing graphlet GraSprings in COEX networks. For the COEX networks of our seven species (legend), we show, on the y-axis, the percentage of enriched genes (top) and the percentage of enriched annotations (bottom) for each of the different GraSpring embeddings (x-axis).

Additionally, for each specie, we focus on identifying what characterizes each particular GraSpring (i.e., $\widetilde{A}_{G_0}$-$\widetilde{A}_{G_8}$) from a biological perspective, just as we did previously for GraCoal embeddings on our COEX molecular networks. In table 6.4 we present the same format as previously presented for the budding yeast COEX network: we report for the budding yeast, the number of uniquely enriched annotations (column 1), the mean SS for the uniquely enriched annotations (column 2) as well as the mean SS for the top 10 largest uniquely enriched annotations (column 3) for each GraSpring used in SAFE. Finally, in column 5, we report the names of the top 10 uniquely enriched annotations and their corresponding size in terms of enriched neighborhoods (column 4). We find that on average, 29.71 GO-BPs are uniquely enriched for each particular GraCoal (Table 6.4). Next, we observe that every GraSpring uncovers unique GO-BPs, validating the claim that graphlet adjacencies capture complementary biological information (Windels et al., 2019). Next, we validate if the sets of uniquely enriched GO-BPs are biologically relevant (i.e., are functionally related). The mean semantic similarities for all GraSpring embeddings used with SAFE on the budding yeast COEX network, ranges from 0.11 (std=0.03) for $GraCoal_8$ to 0.87 (std=0.05) for $GraCoal_5$, indicating some degree of functional of functional similarity, with $GraCoal_5$ capturing the most functional relevance between uniquely enriched GO-BPs. Next, we rank the uniquely enriched annotations according to their size, defined as the total number of neighborhoods they are enriched in, as a measure of how well they are captured by each particular GraSpring. The average size of the uniquely enriched GO-BPs uncovered by GraSprings in the budding yeast is 89.71 (std=45.76). For our other COEX molecular networks, we summarize these statistics in terms of the mean uniquely enriched GO-BPs, mean size of enriched GO-BPs (i.e., in mean number of neighborhoods they are enriched in) and mean semantic similarity in Table 6.5.

In general, we observe that GraSpring embeddings uncover complementary information in all of our COEX molecular networks except for fission yeast (i.e., uniquely enriched annotations across all GraCoals, represented by the means in column 2 in Table 6.3).

$\widetilde{A}_{G_i}$	**Total annotations**	**Mean SS**	**Mean SS Top 10**	**EN**	**Annotation**
$\widetilde{A}_{G_0}$	19	0.34 (std=0.05)	0.55 (std=0.03)	89.0	cellular response to biotic stimulus
$\widetilde{A}_{G_0}$	19	0.34 (std=0.05)	0.55 (std=0.03)	89.0	response to biotic stimulus
$\widetilde{A}_{G_0}$	19	0.34 (std=0.05)	0.55 (std=0.03)	89.0	response to cell cycle checkpoint signaling
$\widetilde{A}_{G_0}$	19	0.34 (std=0.05)	0.55 (std=0.03)	83.0	response to mitotic cell cycle spindle assembly checkpoint signaling
$\widetilde{A}_{G_0}$	19	0.34 (std=0.05)	0.55 (std=0.03)	83.0	negative regulation of protein import into nucleus during spindle assembly checkpoint
$\widetilde{A}_{G_0}$	19	0.34 (std=0.05)	0.55 (std=0.03)	83.0	response to mitotic cell cycle checkpoint signaling
$\widetilde{A}_{G_0}$	19	0.34 (std=0.05)	0.55 (std=0.03)	83.0	response to mitotic spindle checkpoint signaling
$\widetilde{A}_{G_0}$	19	0.34 (std=0.05)	0.55 (std=0.03)	83.0	response to spindle assembly checkpoint signaling
$\widetilde{A}_{G_0}$	19	0.34 (std=0.05)	0.55 (std=0.03)	77.0	response to spindle checkpoint signaling
$\widetilde{A}_{G_0}$	19	0.34 (std=0.05)	0.55 (std=0.03)	67.0	response to endogenous stimulus
$\widetilde{A}_{G_1}$	8	0.26 (std=0.05)	0.26 (std=0.05)	93.0	negative regulation of chromosome organization
$\widetilde{A}_{G_1}$	8	0.26 (std=0.05)	0.26 (std=0.05)	81.0	transmembrane transport
$\widetilde{A}_{G_1}$	8	0.26 (std=0.05)	0.26 (std=0.05)	53.0	carbohydrate derivative biosynthetic process
$\widetilde{A}_{G_1}$	8	0.26 (std=0.05)	0.26 (std=0.05)	41.0	protein folding
$\widetilde{A}_{G_1}$	8	0.26 (std=0.05)	0.26 (std=0.05)	28.0	amino acid transport
$\widetilde{A}_{G_1}$	8	0.26 (std=0.05)	0.26 (std=0.05)	21.0	protein mannosylation
$\widetilde{A}_{G_1}$	8	0.26 (std=0.05)	0.26 (std=0.05)	21.0	mannosylation
$\widetilde{A}_{G_1}$	8	0.26 (std=0.05)	0.26 (std=0.05)	21.0	protein O-linked mannosylation
$\widetilde{A}_{G_2}$	19	0.18 (std=0.04)	0.26 (std=0.05)	160.0	protein metabolic process
$\widetilde{A}_{G_2}$	19	0.18 (std=0.04)	0.26 (std=0.05)	143.0	regulation of exit from mitosis
$\widetilde{A}_{G_2}$	19	0.18 (std=0.04)	0.26 (std=0.05)	137.0	establishment of cell polarity
$\widetilde{A}_{G_2}$	19	0.18 (std=0.04)	0.26 (std=0.05)	117.0	regulation of transcription involved in G1/S transition of mitotic cell cycle
$\widetilde{A}_{G_2}$	19	0.18 (std=0.04)	0.26 (std=0.05)	111.0	cellular bud site selection
$\widetilde{A}_{G_2}$	19	0.18 (std=0.04)	0.26 (std=0.05)	107.0	establishment or maintenance of cell polarity
$\widetilde{A}_{G_2}$	19	0.18 (std=0.04)	0.26 (std=0.05)	95.0	positive regulation of mitotic cell cycle
$\widetilde{A}_{G_2}$	19	0.18 (std=0.04)	0.26 (std=0.05)	86.0	protein import
$\widetilde{A}_{G_2}$	19	0.18 (std=0.04)	0.26 (std=0.05)	54.0	carboxylic acid catabolic process
$\widetilde{A}_{G_2}$	19	0.18 (std=0.04)	0.26 (std=0.05)	54.0	organic acid catabolic process

Table 6.4. Summary of uniquely enriched GO-BPs for GraSpring embeddings, Part 1. We report, for the Budding yeast COEX network, the number of uniquely enriched GO-BPs and the mean semantic similarity (SS) between the uniquely enriched annotations (GO-BPs) (columns 2 and 3) for GraSprings based on all graphlet adjacencies for up to four node graphlets, i.e. $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$ (column 1). In column 4, we report the mean SS for the top ten largest enriched annotations (column 6), i.e., ranking them in descending order according to the number of neighborhoods that the annotations are enriched in (column 5).

$\widetilde{A}_{G_i}$	**Total annotations**	**Mean SS**	**Mean SS Top 10**	**EN**	**Annotation**
$\widetilde{A}_{G_3}$	22	0.23 (std=0.05)	0.25 (std=0.06)	119.0	ribosomal large subunit export from nucleus
$\widetilde{A}_{G_3}$	22	0.23 (std=0.05)	0.25 (std=0.06)	105.0	replication fork reversal
$\widetilde{A}_{G_3}$	22	0.23 (std=0.05)	0.25 (std=0.06)	105.0	G-quadruplex DNA unwinding
$\widetilde{A}_{G_3}$	22	0.23 (std=0.05)	0.25 (std=0.06)	60.0	RNA 3'-end processing
$\widetilde{A}_{G_3}$	22	0.23 (std=0.05)	0.25 (std=0.06)	55.0	response to cold
$\widetilde{A}_{G_3}$	22	0.23 (std=0.05)	0.25 (std=0.06)	55.0	cellular response to cold
$\widetilde{A}_{G_3}$	22	0.23 (std=0.05)	0.25 (std=0.06)	38.0	ketone biosynthetic process
$\widetilde{A}_{G_3}$	22	0.23 (std=0.05)	0.25 (std=0.06)	38.0	amide transport
$\widetilde{A}_{G_3}$	22	0.23 (std=0.05)	0.25 (std=0.06)	38.0	quinone biosynthetic process
$\widetilde{A}_{G_3}$	22	0.23 (std=0.05)	0.25 (std=0.06)	38.0	ubiquinone biosynthetic process
$\widetilde{A}_{G_4}$	3	0.75 (std=0.00)	0.75 (std=0.00)	20.0	regulation of translation
$\widetilde{A}_{G_4}$	3	0.75 (std=0.00)	0.75 (std=0.00)	20.0	regulation of cellular amide metabolic process
$\widetilde{A}_{G_4}$	3	0.75 (std=0.00)	0.75 (std=0.00)	19.0	posttranscriptional regulation of gene expression
$\widetilde{A}_{G_5}$	5	0.87 (std=0.05)	0.87 (std=0.05)	81.0	cofactor metabolic process
$\widetilde{A}_{G_5}$	5	0.87 (std=0.05)	0.87 (std=0.05)	49.0	cofactor biosynthetic process
$\widetilde{A}_{G_5}$	5	0.87 (std=0.05)	0.87 (std=0.05)	44.0	coenzyme metabolic process
$\widetilde{A}_{G_5}$	5	0.87 (std=0.05)	0.87 (std=0.05)	31.0	regulation of small molecule metabolic process
$\widetilde{A}_{G_5}$	5	0.87 (std=0.05)	0.87 (std=0.05)	29.0	coenzyme biosynthetic process

Table 6.4. Summary of uniquely enriched GO-BPs for GraSpring embeddings, Part 2. We report, for the Budding yeast COEX network, the number of uniquely enriched GO-BPs and the mean semantic similarity (SS) between the uniquely enriched annotations (GO-BPs) (columns 2 and 3) for GraSprings based on all graphlet adjacencies for up to four node graphlets, i.e. $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$ (column 1). In column 4, we report the mean SS for the top ten largest enriched annotations (column 6), i.e., ranking them in descending order according to the number of neighborhoods that the annotations are enriched in (column 5).

$\widetilde{A}_{G_i}$	**Total annotations**	**Mean SS**	**Mean SS Top 10**	**EN**	**Annotation**
$\widetilde{A}_{G_6}$	25	0.15 (std=0.04)	0.24 (std=0.06)	193.0	tRNA modification
$\widetilde{A}_{G_6}$	25	0.15 (std=0.04)	0.24 (std=0.06)	183.0	tRNA processing
$\widetilde{A}_{G_6}$	25	0.15 (std=0.04)	0.24 (std=0.06)	161.0	vesicle-mediated transport
$\widetilde{A}_{G_6}$	25	0.15 (std=0.04)	0.24 (std=0.06)	143.0	protein-DNA complex disassembly
$\widetilde{A}_{G_6}$	25	0.15 (std=0.04)	0.24 (std=0.06)	120.0	protein-containing complex localization
$\widetilde{A}_{G_6}$	25	0.15 (std=0.04)	0.24 (std=0.06)	112.0	nucleosome disassembly
$\widetilde{A}_{G_6}$	25	0.15 (std=0.04)	0.24 (std=0.06)	110.0	protein-containing complex disassembly
$\widetilde{A}_{G_6}$	25	0.15 (std=0.04)	0.24 (std=0.06)	93.0	positive regulation of chromatin organization
$\widetilde{A}_{G_6}$	25	0.15 (std=0.04)	0.24 (std=0.06)	90.0	protein maturation
$\widetilde{A}_{G_6}$	25	0.15 (std=0.04)	0.24 (std=0.06)	89.0	positive regulation of histone acetylation
$\widetilde{A}_{G_7}$	33	0.16 (std=0.03)	0.31 (std=0.05)	145.0	monocarboxylic acid metabolic process
$\widetilde{A}_{G_7}$	33	0.16 (std=0.03)	0.31 (std=0.05)	111.0	chromatin assembly or disassembly
$\widetilde{A}_{G_7}$	33	0.16 (std=0.03)	0.31 (std=0.05)	99.0	DNA-dependent DNA replication
$\widetilde{A}_{G_7}$	33	0.16 (std=0.03)	0.31 (std=0.05)	96.0	small-subunit processome assembly
$\widetilde{A}_{G_7}$	33	0.16 (std=0.03)	0.31 (std=0.05)	80.0	intra-S DNA damage checkpoint
$\widetilde{A}_{G_7}$	33	0.16 (std=0.03)	0.31 (std=0.05)	78.0	response to organonitrogen compound
$\widetilde{A}_{G_7}$	33	0.16 (std=0.03)	0.31 (std=0.05)	75.0	response to nitrogen compound
$\widetilde{A}_{G_7}$	33	0.16 (std=0.03)	0.31 (std=0.05)	71.0	regulation of meiotic cell cycle
$\widetilde{A}_{G_7}$	33	0.16 (std=0.03)	0.31 (std=0.05)	70.0	attachment of spindle microtubules to kinetochore
$\widetilde{A}_{G_7}$	33	0.16 (std=0.03)	0.31 (std=0.05)	69.0	mitotic DNA damage checkpoint
$\widetilde{A}_{G_8}$	98	0.11 (std=0.03)	0.22 (std=0.04)	229.0	generation of precursor metabolites and energy
$\widetilde{A}_{G_8}$	98	0.11 (std=0.03)	0.22 (std=0.04)	213.0	biosynthetic process
$\widetilde{A}_{G_8}$	98	0.11 (std=0.03)	0.22 (std=0.04)	133.0	cytoplasmic translation
$\widetilde{A}_{G_8}$	98	0.11 (std=0.03)	0.22 (std=0.04)	132.0	rRNA transport
$\widetilde{A}_{G_8}$	98	0.11 (std=0.03)	0.22 (std=0.04)	132.0	rRNA export from nucleus
$\widetilde{A}_{G_8}$	98	0.11 (std=0.03)	0.22 (std=0.04)	130.0	ncRNA export from nucleus
$\widetilde{A}_{G_8}$	98	0.11 (std=0.03)	0.22 (std=0.04)	124.0	chromosome organization involved in meiotic cell cycle
$\widetilde{A}_{G_8}$	98	0.11 (std=0.03)	0.22 (std=0.04)	124.0	ribosomal small subunit assembly
$\widetilde{A}_{G_8}$	98	0.11 (std=0.03)	0.22 (std=0.04)	112.0	regulation of cytoskeleton organization
$\widetilde{A}_{G_8}$	98	0.11 (std=0.03)	0.22 (std=0.04)	108.0	protein localization to chromosome

Table 6.4. Summary of uniquely enriched GO-BPs for GraSpring embeddings, Part 3. We report, for the Budding yeast COEX network, the number of uniquely enriched GO-BPs and the mean semantic similarity (SS) between the uniquely enriched annotations (GO-BPs) (columns 2 and 3) for GraSprings based on all graphlet adjacencies for up to four node graphlets, i.e. $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$ (column 1). In column 4, we report the mean SS for the top ten largest enriched annotations (column 6), i.e., ranking them in descending order according to the number of neighborhoods that the annotations are enriched in (column 5).

Organism	**Mean unique GO-BPs**	**Mean enriched neighborhoods**	**Mean SS**
Budding yeast	29.71 (std=28.00)	89.71 (std=45.76)	0.28 (std=0.23)
Fission yeast	1.10 (std=3.33)	17.10 (std=1.45)	0.38 (std=0.00)
Fruit fly	37.88 (std=19.74)	55.64 (std=37.16)	0.18 (std=0.04)
Human	52.67 (std=32.45)	102.37 (std=48.61)	0.16 (std=0.05)
Roundworm	29.33 (std=12.44)	53.74 (std=35.64)	0.19 (std=0.03)

Table 6.5. Summary statistics of uniquely enriched GO-BPs for GraSpring embeddings on COEX networks. We report, for the our six COEX networks (column 1), the mean number of uniquely enriched GO-BPs obtained with each GraSpring embedding (column 2), the mean size of the uniquely enriched GO-BPs in terms of the number of neighborhoods that the annotations are enriched in (column 3) and the mean semantic similarity (column 4).

6.3 Results for graphlet Spectral with SAFE

Graphlet Spectrals uncover complementary biological information in COEX networks

Similar to applying GraCoal embeddings or GraSpring embeddings to our COEX networks, applying graphlet Spectral embeddings with SAFE on our COEX networks also allow for uncovering complementary biological information. We show, in Figure 6.4 the percentages of genes enriched (top) and percentages of GO-BPs enriched (bottom) for all graphlet Spectrals (i.e., $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$) across our six COEX networks. Our first observation is that the top performing graphlet Spectrals correspond to the same COEX network as with GraSpring embeddings or GraCoal embeddings (budding yeast). However, not every graphlet Spectral for the budding yeast outperforms the other graphlet Spectrals for the other species in terms of genes. For instance, $Spectral_8$ achieves 16.7% for yeast, but is slightly higher for the fruit fly COEX network (17.4%). Moreover, graphlet Spectrals based on densely connected graphlets such as $Spectral_{2,7,8}$ perform the worst for budding yeast, achieving between 16.7% and 25.6%. This is consistent for the other COEX networks. That is, the worst performing graphlet Spectrals tend to be those based on densely connected graphlets, in particular the two cliques ($Spectral_{2,8}$). We observe that for fruit fly, the percentages of genes enriched are the second best, achieving on average 19.8%. Next, human and roundworm are third and fourth best, respectively achieving percentages of genes enriched that range from 4.9% to 17.8%. Finally, the worst enrichments achieved by graphlet Spectral embeddings correspond to fission yeast, which is not surprising, considering this is also the case when we apply GraSpring embeddings or GraCoal embeddings on this molecular network.

In terms of GO-BPs enriched, we observe similar patterns as with genes enriched. The best performing graphlet Spectrals correspond to the budding yeast COEX network, achieving between 20.3% for $Spectral_8$ and 38.5% for $Spectral_3$. Next, the second best enrichments are achieved for the fruit fly COEX network (between 19.7% and 22.6% for $Spectral_8$ and $Spectral_0$, respectively). Interestingly, the third and fourth best performing Spectrals, which in terms of genes correspond to human and roundworm, respectively, correspond to roundworm and human, respectively in terms of enriched GO-BPs. Finally, the worst performing Spectrals correspond to fission yeast, achieving close to 0% enriched GO-BPs across all Spectrals, except for $Spectral_5$ (1.8%).

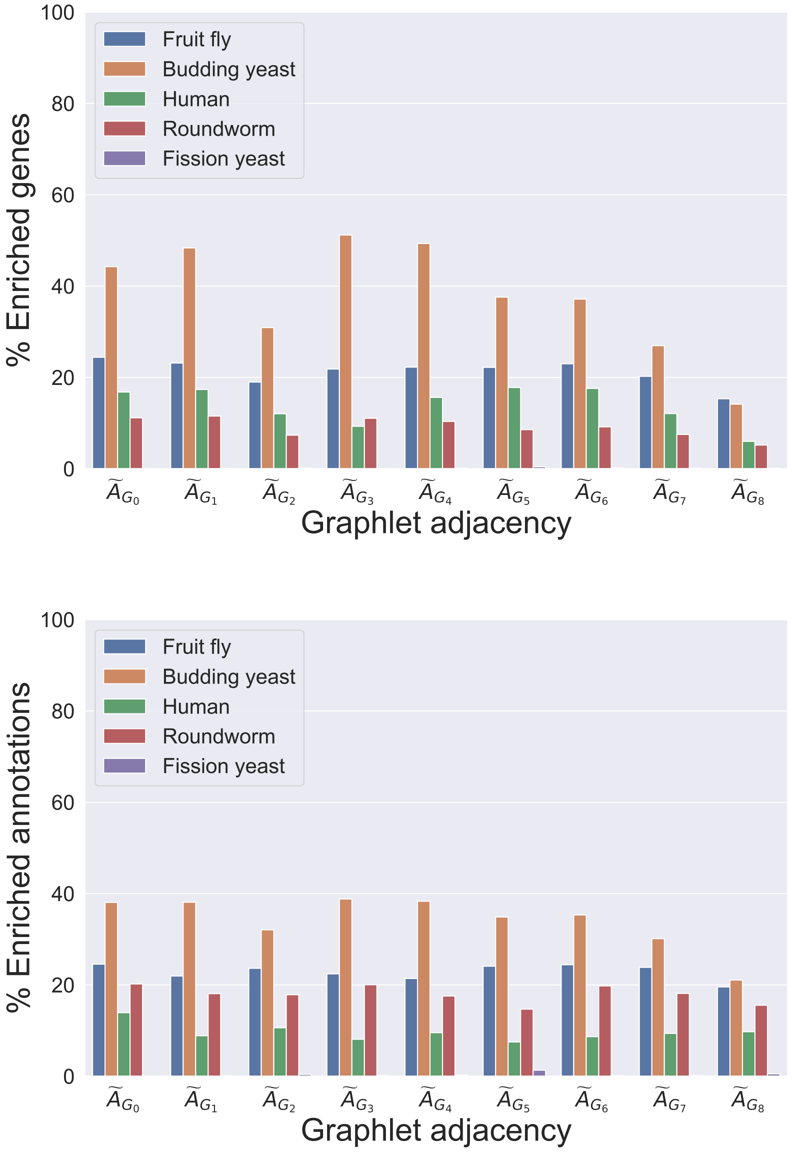


Figure 6.4. SAFE GO-BP enrichment analysis comparing graphlet Spectrals in COEX networks. For the COEX networks of our seven species (legend), we show, on the y-axis, the percentage of enriched genes (top) and the percentage of enriched annotations (bottom) for each of the different graphlet Spectral embeddings (x-axis).

Next, for each specie, we focus on identifying what characterizes each particular graphlet Spectral (i.e., $\widetilde{A}_{G_0}$-$\widetilde{A}_{G_8}$) from a biological perspective, just as we did previously for GraSpring embeddings and GraCoal embeddings on the budding yeast COEX network. In table 6.6, we report for the budding yeast, the number of uniquely enriched annotations (column 1), the mean SS for the uniquely enriched annotations (column 2) as well as the mean SS for the top 10 largest uniquely enriched annotations (column 3) for each graphlet Spectral used in SAFE. Finally, in column 5, we report the names of the top 10 uniquely enriched annotations and their corresponding size in terms of enriched neighborhoods (column 4). We find that on average, 29.67 GO-BPs are uniquely enriched for each particular graphlet Spectral. Next, we observe that every graphlet Spectral uncovers unique biological information (i.e., GO-BPs), validating the claim that graphlet adjacencies capture complementary biological information (Windels et al., 2019). Next, we validate if the sets of uniquely enriched GO-BPs are biologically relevant (i.e., are functionally related). The mean semantic similarity for all graphlet Spectral embeddings used with SAFE on the budding yeast COEX network, ranges from 0.12 (std=0.03) for $GraCoal_6$ to 0.26 (std=0.05) for $GraCoal_8$. Hence, the uniquely enriched GO-BPs uncovered by graphlet Spectral embeddings uncover some degree of functional relevance, with $GraCoal_8$ capturing the most. Next, we rank the uniquely enriched annotations according to their size, defined as the total number of neighborhoods they are enriched in, as a measure of how well they are captured by each particular graphlet Spectral. The average size of the uniquely enriched GO-BPs uncovered by graphlet Spectrals in the budding yeast is 70.49 (std=80.95). For our other COEX networks, we summarize these statistics in terms of the mean uniquely enriched GO-BPs when using graphlet Spectrals with SAFE, mean size of enriched GO-BPs (i.e., in mean number of neighborhoods they are enriched in) and mean semantic similarity in Table 6.7.

Our first observation is that graphlet Spectral embeddings are unable to uncover unique GO-BPs for the fission yeast, which is not surprising considering the method performed the worst on this network and only a couple of graphlet Spectrals achieved really low percentages of both genes and GO-BPs enriched. Besides fission yeast, we observe, for our other COEX networks, graphlet Spectral embeddings uncover some degree of functional relevance across all graphlet Spectrals, as evidenced by the mean semantic similarities in column 4 of Table 6.7).

$\widetilde{A}_{G_i}$	**Total annotations**	**Mean SS**	**Mean SS Top 10**	**EN**	**Annotation**
$\widetilde{A}_{G_0}$	23	0.21 (std=0.06)	0.31 (std=0.06)	192.0	protein metabolic process
$\widetilde{A}_{G_0}$	23	0.21 (std=0.06)	0.31 (std=0.06)	124.0	cytoskeleton organization
$\widetilde{A}_{G_0}$	23	0.21 (std=0.06)	0.31 (std=0.06)	115.0	piecemeal microautophagy of the nucleus
$\widetilde{A}_{G_0}$	23	0.21 (std=0.06)	0.31 (std=0.06)	85.0	organelle disassembly
$\widetilde{A}_{G_0}$	23	0.21 (std=0.06)	0.31 (std=0.06)	45.0	negative regulation of protein import into nucleus during spindle assembly checkpoint
$\widetilde{A}_{G_0}$	23	0.21 (std=0.06)	0.31 (std=0.06)	45.0	response to mitotic cell cycle checkpoint signaling
$\widetilde{A}_{G_0}$	23	0.21 (std=0.06)	0.31 (std=0.06)	45.0	response to mitotic spindle checkpoint signaling
$\widetilde{A}_{G_0}$	23	0.21 (std=0.06)	0.31 (std=0.06)	45.0	response to spindle assembly checkpoint signaling
$\widetilde{A}_{G_0}$	23	0.21 (std=0.06)	0.31 (std=0.06)	45.0	response to mitotic cell cycle spindle assembly checkpoint signaling
$\widetilde{A}_{G_0}$	23	0.21 (std=0.06)	0.31 (std=0.06)	44.0	regulation of exit from mitosis
$\widetilde{A}_{G_1}$	22	0.18 (std=0.04)	0.24 (std=0.04)	187.0	organonitrogen compound metabolic process
$\widetilde{A}_{G_1}$	22	0.18 (std=0.04)	0.24 (std=0.04)	70.0	organic acid transport
$\widetilde{A}_{G_1}$	22	0.18 (std=0.04)	0.24 (std=0.04)	69.0	carboxylic acid transport
$\widetilde{A}_{G_1}$	22	0.18 (std=0.04)	0.24 (std=0.04)	45.0	meiotic chromosome separation
$\widetilde{A}_{G_1}$	22	0.18 (std=0.04)	0.24 (std=0.04)	40.0	regulation of meiotic cell cycle
$\widetilde{A}_{G_1}$	22	0.18 (std=0.04)	0.24 (std=0.04)	38.0	nucleoside phosphate biosynthetic process
$\widetilde{A}_{G_1}$	22	0.18 (std=0.04)	0.24 (std=0.04)	35.0	purine ribonucleotide metabolic process
$\widetilde{A}_{G_1}$	22	0.18 (std=0.04)	0.24 (std=0.04)	34.0	ribose phosphate metabolic process
$\widetilde{A}_{G_1}$	22	0.18 (std=0.04)	0.24 (std=0.04)	33.0	secretion
$\widetilde{A}_{G_1}$	22	0.18 (std=0.04)	0.24 (std=0.04)	33.0	secretion by cell
$\widetilde{A}_{G_2}$	10	0.25 (std=0.04)	0.25 (std=0.04)	41.0	RNA modification
$\widetilde{A}_{G_2}$	10	0.25 (std=0.04)	0.25 (std=0.04)	23.0	negative regulation of cytoskeleton organization
$\widetilde{A}_{G_2}$	10	0.25 (std=0.04)	0.25 (std=0.04)	22.0	cellular localization
$\widetilde{A}_{G_2}$	10	0.25 (std=0.04)	0.25 (std=0.04)	21.0	sulfur compound biosynthetic process
$\widetilde{A}_{G_2}$	10	0.25 (std=0.04)	0.25 (std=0.04)	20.0	sulfur compound transport
$\widetilde{A}_{G_2}$	10	0.25 (std=0.04)	0.25 (std=0.04)	11.0	regulation of transcription involved in G1/S transition of mitotic cell cycle
$\widetilde{A}_{G_2}$	10	0.25 (std=0.04)	0.25 (std=0.04)	11.0	mRNA export from nucleus
$\widetilde{A}_{G_2}$	10	0.25 (std=0.04)	0.25 (std=0.04)	11.0	mRNA transport
$\widetilde{A}_{G_2}$	10	0.25 (std=0.04)	0.25 (std=0.04)	10.0	nucleocytoplasmic transport
$\widetilde{A}_{G_2}$	10	0.25 (std=0.04)	0.25 (std=0.04)	10.0	nuclear transport

Table 6.6. Summary of uniquely enriched GO-BPs for graphlet Spectral embeddings, Part 1. We report, for the Budding yeast COEX network, the number of uniquely enriched GO-BPs and the mean semantic similarity (SS) between the uniquely enriched annotations (GO-BPs) (columns 2 and 3) for graphlet Spectral embeddings based on all graphlet adjacencies for up to four node graphlets, i.e. $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$ (column 1). In column 4, we report the mean SS for the top ten largest enriched annotations (column 6), i.e., ranking them in descending order according to the number of neighborhoods that the annotations are enriched in (column 5).

$\widetilde{A}_{G_i}$	**Total annotations**	**Mean SS**	**Mean SS Top 10**	**EN**	**Annotation**
$\widetilde{A}_{G_3}$	29	0.20 (std=0.04)	0.24 (std=0.05)	87.0	negative regulation of cellular component organization
$\widetilde{A}_{G_3}$	29	0.20 (std=0.04)	0.24 (std=0.05)	82.0	negative regulation of organelle organization
$\widetilde{A}_{G_3}$	29	0.20 (std=0.04)	0.24 (std=0.05)	61.0	regulation of cellular protein metabolic process
$\widetilde{A}_{G_3}$	29	0.20 (std=0.04)	0.24 (std=0.05)	60.0	regulation of protein metabolic process
$\widetilde{A}_{G_3}$	29	0.20 (std=0.04)	0.24 (std=0.05)	52.0	organic acid catabolic process
$\widetilde{A}_{G_3}$	29	0.20 (std=0.04)	0.24 (std=0.05)	52.0	carboxylic acid catabolic process
$\widetilde{A}_{G_3}$	29	0.20 (std=0.04)	0.24 (std=0.05)	40.0	lipid modification
$\widetilde{A}_{G_3}$	29	0.20 (std=0.04)	0.24 (std=0.05)	37.0	mitochondrial RNA metabolic process
$\widetilde{A}_{G_3}$	29	0.20 (std=0.04)	0.24 (std=0.05)	37.0	tRNA aminoacylation for mitochondrial protein translation
$\widetilde{A}_{G_3}$	29	0.20 (std=0.04)	0.24 (std=0.05)	36.0	mitochondrial transport
$\widetilde{A}_{G_4}$	43	0.13 (std=0.03)	0.28 (std=0.06)	424.0	cellular nitrogen compound metabolic process
$\widetilde{A}_{G_4}$	43	0.13 (std=0.03)	0.28 (std=0.06)	412.0	macromolecule metabolic process
$\widetilde{A}_{G_4}$	43	0.13 (std=0.03)	0.28 (std=0.06)	270.0	macromolecule biosynthetic process
$\widetilde{A}_{G_4}$	43	0.13 (std=0.03)	0.28 (std=0.06)	251.0	ribosomal large subunit assembly
$\widetilde{A}_{G_4}$	43	0.13 (std=0.03)	0.28 (std=0.06)	245.0	maturation of LSU-rRNA
$\widetilde{A}_{G_4}$	43	0.13 (std=0.03)	0.28 (std=0.06)	234.0	ribonucleoprotein complex subunit organization
$\widetilde{A}_{G_4}$	43	0.13 (std=0.03)	0.28 (std=0.06)	233.0	ribonucleoprotein complex assembly
$\widetilde{A}_{G_4}$	43	0.13 (std=0.03)	0.28 (std=0.06)	233.0	RNA export from nucleus
$\widetilde{A}_{G_4}$	43	0.13 (std=0.03)	0.28 (std=0.06)	194.0	response to stimulus
$\widetilde{A}_{G_4}$	43	0.13 (std=0.03)	0.28 (std=0.06)	185.0	nucleobase-containing compound transport
$\widetilde{A}_{G_5}$	30	0.19 (std=0.05)	0.26 (std=0.04)	50.0	isopentenyl diphosphate metabolic process
$\widetilde{A}_{G_5}$	30	0.19 (std=0.05)	0.26 (std=0.04)	50.0	farnesyl diphosphate biosynthetic process, mevalonate pathway
$\widetilde{A}_{G_5}$	30	0.19 (std=0.05)	0.26 (std=0.04)	50.0	isopentenyl diphosphate biosynthetic process, mevalonate pathway
$\widetilde{A}_{G_5}$	30	0.19 (std=0.05)	0.26 (std=0.04)	50.0	isopentenyl diphosphate biosynthetic process
$\widetilde{A}_{G_5}$	30	0.19 (std=0.05)	0.26 (std=0.04)	50.0	isoprenoid biosynthetic process via mevalonate
$\widetilde{A}_{G_5}$	30	0.19 (std=0.05)	0.26 (std=0.04)	38.0	membrane lipid biosynthetic process
$\widetilde{A}_{G_5}$	30	0.19 (std=0.05)	0.26 (std=0.04)	35.0	cellular response to acid chemical
$\widetilde{A}_{G_5}$	30	0.19 (std=0.05)	0.26 (std=0.04)	35.0	meiotic mismatch repair
$\widetilde{A}_{G_5}$	30	0.19 (std=0.05)	0.26 (std=0.04)	35.0	mismatch repair
$\widetilde{A}_{G_5}$	30	0.19 (std=0.05)	0.26 (std=0.04)	34.0	cellular response to oxygen-containing compound

Table 6.6. Summary of uniquely enriched GO-BPs for graphlet Spectral embeddings, Part 2. We report, for the Budding yeast COEX network, the number of uniquely enriched GO-BPs and the mean semantic similarity (SS) between the uniquely enriched annotations (GO-BPs) (columns 2 and 3) for graphlet Spectral embeddings based on all graphlet adjacencies for up to four node graphlets, i.e. $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$ (column 1). In column 4, we report the mean SS for the top ten largest enriched annotations (column 6), i.e., ranking them in descending order according to the number of neighborhoods that the annotations are enriched in (column 5).

$\widetilde{A}_{G_i}$	**Total annotations**	**Mean SS**	**Mean SS Top 10**	**EN**	**Annotation**
$\widetilde{A}_{G_6}$	53	0.12 (std=0.03)	0.25 (std=0.06)	92.0	inorganic cation import across plasma membrane
$\widetilde{A}_{G_6}$	53	0.12 (std=0.03)	0.25 (std=0.06)	92.0	inorganic ion import across plasma membrane
$\widetilde{A}_{G_6}$	53	0.12 (std=0.03)	0.25 (std=0.06)	52.0	protein modification by small protein conjugation or removal
$\widetilde{A}_{G_6}$	53	0.12 (std=0.03)	0.25 (std=0.06)	45.0	protein modification by small protein conjugation
$\widetilde{A}_{G_6}$	53	0.12 (std=0.03)	0.25 (std=0.06)	41.0	protein ubiquitination
$\widetilde{A}_{G_6}$	53	0.12 (std=0.03)	0.25 (std=0.06)	38.0	amide biosynthetic process
$\widetilde{A}_{G_6}$	53	0.12 (std=0.03)	0.25 (std=0.06)	36.0	phosphatidylinositol dephosphorylation
$\widetilde{A}_{G_6}$	53	0.12 (std=0.03)	0.25 (std=0.06)	35.0	phospholipid dephosphorylation
$\widetilde{A}_{G_6}$	53	0.12 (std=0.03)	0.25 (std=0.06)	35.0	galactose catabolic process via UDP-galactose
$\widetilde{A}_{G_6}$	53	0.12 (std=0.03)	0.25 (std=0.06)	32.0	nonfunctional rRNA decay
$\widetilde{A}_{G_7}$	47	0.18 (std=0.05)	0.28 (std=0.06)	54.0	cellular protein modification process
$\widetilde{A}_{G_7}$	47	0.18 (std=0.05)	0.28 (std=0.06)	54.0	protein modification process
$\widetilde{A}_{G_7}$	47	0.18 (std=0.05)	0.28 (std=0.06)	50.0	Golgi to plasma membrane transport
$\widetilde{A}_{G_7}$	47	0.18 (std=0.05)	0.28 (std=0.06)	50.0	vesicle-mediated transport to the plasma membrane
$\widetilde{A}_{G_7}$	47	0.18 (std=0.05)	0.28 (std=0.06)	48.0	pre-replicative complex assembly involved in cell cycle DNA replication
$\widetilde{A}_{G_7}$	47	0.18 (std=0.05)	0.28 (std=0.06)	48.0	pre-replicative complex assembly
$\widetilde{A}_{G_7}$	47	0.18 (std=0.05)	0.28 (std=0.06)	48.0	pre-replicative complex assembly involved in nuclear cell cycle DNA replication
$\widetilde{A}_{G_7}$	47	0.18 (std=0.05)	0.28 (std=0.06)	47.0	DNA unwinding involved in DNA replication
$\widetilde{A}_{G_7}$	47	0.18 (std=0.05)	0.28 (std=0.06)	36.0	ion transmembrane transport
$\widetilde{A}_{G_7}$	47	0.18 (std=0.05)	0.28 (std=0.06)	30.0	regulation of protein complex assembly
$\widetilde{A}_{G_8}$	10	0.26 (std=0.05)	0.26 (std=0.05)	30.0	polyphosphate metabolic process
$\widetilde{A}_{G_8}$	10	0.26 (std=0.05)	0.26 (std=0.05)	21.0	regulation of cytoskeleton organization
$\widetilde{A}_{G_8}$	10	0.26 (std=0.05)	0.26 (std=0.05)	20.0	glycerophospholipid biosynthetic process
$\widetilde{A}_{G_8}$	10	0.26 (std=0.05)	0.26 (std=0.05)	20.0	glycerolipid biosynthetic process
$\widetilde{A}_{G_8}$	10	0.26 (std=0.05)	0.26 (std=0.05)	18.0	positive regulation of Arp2/3 complex-mediated actin nucleation
$\widetilde{A}_{G_8}$	10	0.26 (std=0.05)	0.26 (std=0.05)	18.0	positive regulation of actin nucleation
$\widetilde{A}_{G_8}$	10	0.26 (std=0.05)	0.26 (std=0.05)	14.0	sulfate assimilation, phosphoadenylyl sulfate reduction by phosphoadenylyl-sulfate reductase (thioredoxin)
$\widetilde{A}_{G_8}$	10	0.26 (std=0.05)	0.26 (std=0.05)	14.0	sulfate reduction
$\widetilde{A}_{G_8}$	10	0.26 (std=0.05)	0.26 (std=0.05)	13.0	positive regulation of protein polymerization
$\widetilde{A}_{G_8}$	10	0.26 (std=0.05)	0.26 (std=0.05)	12.0	positive regulation of supramolecular fiber organization

Table 6.6. Summary of uniquely enriched GO-BPs for graphlet Spectral embeddings, Part 3. We report, for the Budding yeast COEX network, the number of uniquely enriched GO-BPs and the mean semantic similarity (SS) between the uniquely enriched annotations (GO-BPs) (columns 2 and 3) for graphlet Spectral embeddings based on all graphlet adjacencies for up to four node graphlets, i.e. $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$ (column 1). In column 4, we report the mean SS for the top ten largest enriched annotations (column 6), i.e., ranking them in descending order according to the number of neighborhoods that the annotations are enriched in (column 5).

Organism	**Mean unique GO-BPs**	**Mean enriched neighborhoods**	**Mean SS**
Budding yeast	29.67 (std=14.61)	70.49 (std=80.95)	0.19 (std=0.04)
Fission yeast	nan (std=nan)	nan (std=nan)	nan (std=nan)
Fruit fly	31.89 (std=11.74)	67.97 (std=54.79)	0.17 (std=0.02)
Human	30.35 (std=26.02)	68.28 (std=53.67)	0.21 (std=0.04)
Roundworm	24.18 (std=11.36)	35.00 (std=26.24)	0.21 (std=0.05)

Table 6.7. Summary statistics of uniquely enriched GO-BPs for graphlet Spectral embeddings on COEX networks. We report, for the our six COEX networks (column 1), the mean number of uniquely enriched GO-BPs obtained with each graphlet Spectral embedding (column 2), the mean size of the uniquely enriched GO-BPs in terms of the number of neighborhoods that the annotations are enriched in (column 3) and the mean semantic similarity (column 4).

Chapter 7

Conclusions

In this chapter we summarize the methodological contributions and results presented in this thesis as well as an overview of the general conclusions. Moreover, we discuss some future steps to consider for further applications and methodological improvements of our methods.

7.1 Summary of the Thesis

In **Chapter 3** we present new methods to embed molecular interaction networks based on graphlet topology. First, we generalise the Coalescent embedding algorithm (Muscoloni et al., 2017) based on Laplacian Eigenmaps (Belkin & Niyogi, 2001, 2003) to take as input the graphlet Laplacian matrix of a network (based on any graphlet with up to 4 nodes) as opposed to the standard Laplacian matrix. We find that when using input data based on graphlets, the original equation for computing the radial coordinates of the nodes is not well suited, as it assumes the node degree distribution to follow a power-law, which is not always the case for graphlet node degrees, as shown in Figures 3.1 to 3.4. Next, we also generalise the Spring embedding based on the Fruchterman-Reingold force-directed algorithm (Fruchterman & Reingold, 1991), i.e., the same algorithm used in the original Spatial Analysis of Functional Enrichment framework (Baryshnikova, 2016). For this, we modify the input data to be the normalized graphlet adjacency matrix of a network (based on any graphlet with up to 4 nodes), instead of the standard adjacency matrix. We extend the SAFE framework to include these embedding methods based on graphlets as optional parameters for the user, as well as the already established graphlet based Spectral embedding (Windels et al., 2019). In addition, we also include a new hyperparameter, the neighborhood size, which the user can choose beforehand. Finally, we perform enrichment experiments to determine the optimal value of this new hyperparameter (neighborhood size) to allow for a fair comparison across our graphlet based embedding methods.

In **Chapter 4**, we use SAFE to apply the graphlet based embedding methods, which we name GraCoal embedding, GraSpring embedding and graphlet based Spectral embedding, on the genetic interaction (GI) networks corresponding to various model organisms. We find that GraCoal embeddings lead to the best enrichment results for our GI networks. A possible explanation for this, is the fact that GraCoal embedding spreads the nodes better in the embedding space with respect to graphlet-based Spring embedding or graphlet-based Spectral embedding, which leads to well separated functional domains, which we show in Figures 4.2 to 4.4. We also find that some graphlet topologies lead to better enrichments than others, and thus we try to explain this by performing experiments to explore the structural organization of our GI networks. To this end, we perform model fitting experiments to compare our GI networks to eight model networks commonly used in biology. We find that the Scale-Free with gene duplication is the best fit for our GI networks, which could explain why GraCoal embeddings based on triangle topology, such as $GraCoal_2$ or $GraCoal_8$ tend to outperform the other GraCoals. We validate this by determining, for each species, a set of paralogous genes and and we find that our GI networks (except for fruit fly) are characterized by the pressence of these duplicated genes. Next, we show how these sets of genes are more likely to interact with each other in the network, share more neighbors than expected by chance (as shown by the GDV similarities), are closer in the embedding space than expected by chance and participate in more triangles than expected by chance. Thus, we conclude how there is a strong topology-function relationship that is captured very well by the triangle based topology in these GI networks due to the pressence of many paralogs. Next, we provide biological insights at the annotation level to validate that different grahplet topologies capture complementary biological information and that the information being captured is functionally relevant (as evidenced by the semantic similarity of the sets of enriched annotations). Finaly, we provide biological insight at the functional domain level and identify the most characteristic domains for each graphlet based embedding. We find that unique functional domains corresponding to the triangle topology (i.e., $\widetilde{A}_{G_2}$ or $\widetilde{A}_{G_8}$) tend to be capturing more gene-paralog relationships in GraCoal embeddings than in GraSpring or graphlet Spectral embeddings. We conclude that GraCoal embeddings are the best method overall for uncovering the functional organization of the cell in GI networks, in particular when there is gene duplication involved.

In **Chapter 5**, we present our results when we use SAFE to apply the graphlet based embeddings on the protein-protein interaction (PPI) networks corresponding to various model organisms. We find that GraCoal embeddings tend to perform the best when we evaluate the performance in terms of genes enriched in annotations. But this is not the case when we evaluate the performance in terms of annotations enriched, as it varies from species to species. Furthermore, we provide biological insights at an annotation level to validate if graphlet based embeddings uncover complementary information in PPI networks. We find this is the case for the three embedding methods accross all our PPI networks, as shown by the semantic similarites of the enriched annotations. Additionally, we perform model fitting experiment to compare our PPI networks to eight model networks (same as in chapter 4). However, none of our PPI networks were well fitted by any of the model networks. We conclude that either GraCoal or graphlet based Spectral could be best suited for uncovering biological information from PPI networks using SAFE, as they are the best and second best in terms of genes enriched.

In **Chapter 6**, we present our results when we use SAFE to apply the graphlet based embeddings on the co-expression (COEX) networks corresponding to various model organisms. In general, the results with these networks are very poor in some species for all the three methods. However, we conclude the best embedding method to be GraSpring, as it achieves the best enrichments overall. Because our COEX networks correspond to the most dense networks, this could indicate that GraSpring embeddings are the best suited method for uncovering functional information in very dense networks. Finally, for fission yeast, the results were the worst, achieving between 0% and 5% for the most part of genes enriched and annotations enriched. This can easily be explained by the fact that annotation data for fission yeast is the worst of all across all species and network types, as shown in Section A.3 in Appendix A.

7.2 Conclusions

In this section we present the general conclusions of the Thesis.

1. Generalised the Spring embedding to graphlet-based Spring embedding.
2. Generalised the Coalescent embedding to graphlet-based Coalescent embedding.
3. Graphlet degree distributions do not follower a scale-free distribution.
4. Improved the SAFE framework by integrating the newly proposed methods into the framework.
5. GraCoal embeddings lead to the best enrichments for genetic interaction networks at the node level and annotation level.
6. Some graphlet topologies lead to better results than others. For genetic interaction, the best topologies are based on triangles when there are paralogous genes in the network.
7. Strong topology-function relationship between triangle topology and presence of paralogs in the network.
8. Unique functional domains uncovered by SAFE when using triangle topology tend to capture more paralogous genes when using GraCoal embedding.
9. Biological information captured by graphlet-based embeddings is functionally coherent in all network types.

7.3 Future directions

In this section we discuss future work that could be done to further uncover topology-function relationships in molecular interaction networks.

For our GI networks, we demonstrate how the general structure of these networks is best fitted by the SF-GD model, which allow us to uncover the strong topology-function relationship between triangle topology ($\widetilde{A}_{2,8}$ and the presence of paralogous genes in the network. However, for our PPI and COEX networks none of the model networks contributed to a better understanding of their global structural organization. For this reason, something that remains to be done is the further exploration of the global structure of our PPI and COEX networks. This will allow for a better understanding of why one embedding might work better in terms of genes but not in terms of enriched annotations (i.e., GraCoal embeddings). Moreover, it may the case that there is no global pattern in PPI or COEX networks (as there is for GI) and thus, individual exploration of each particular molecular network could be a better approach.

In terms of applications, our future work will mainly be focused in applying graphlet based embeddings to tissue specific and disease related molecular interaction networks. For instance, to compare between healthy lung tissue versus cancer lung tissue we would build the corresponding GI/PPI/COEX networks according to gene expression profiles of such tissues under both conditions. Our graphlet based embeddings would be a valuable method for uncovering the functional organization of such data, which we would expect the data corresponding to cancer to be disrupted in some way. This usually leads to changes in the wiring patterns of the nodes in the network, which in turn alters the functional organization with respect to the healthy tissue, which could translate into a poor performance of our methods in terms of enrichments, or to a completely different organization (i.e., functional domains that do not make sense and thus are not functionally coherent).

Finally, in terms of methodological improvements there is a lot to be explored. For instance, all of our experiments were based on the default shortest weighted path length distance metric for defining the node neighborhoods, which relies exclusively on node connectivity (i.e., along the paths). However, other distance metrics for defining the node neighborhoods, such as the angular distance or cosine distance could be better suited for uncovering the functional organization of molecular networks using the GraCoal embeddings. Moreover, because the radial coordinate represents the topological importance of a particular node in the network, such as how well connected the node is (i.e. its degree, or graphlet node degree in the case of graphlets), other centrality measures not based on degree remain to be explored, such as betweenness centrality or eigencentrality.

Bibliography

Adamic, L. A., & Glance, N. (2005). The political blogosphere and the 2004 us election: Divided they blog. *Proceedings of the 3rd international workshop on Link discovery*, 36–43.

Alanis-Lobato, G., Mier, P., & Andrade-Navarro, M. A. (2016a). Efficient embedding of complex networks to hyperbolic space via their Laplacian [Number: 1 Publisher: Nature Publishing Group]. *Scientific Reports*, *6*(1), 1–10. https://doi.org/10.1038/srep30108

Alanis-Lobato, G., Mier, P., & Andrade-Navarro, M. A. (2016b). Manifold learning and maximum likelihood estimation for hyperbolic network embedding [Number: 1 Publisher: SpringerOpen]. *Applied Network Science*, *1*(1), 1–14. https://doi.org/10.1007/s41109-016-0013-0

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, *215*(3), 403–410.

Anton, C., Taubas, J. V., & Roncero, C. (2018). The functional specialization of exomer as a cargo adaptor during the evolution of fungi. *Genetics*, *208*(4), 1483–1498.

Aparicio, D., Ribeiro, P., & Silva, F. (2017). Extending the Applicability of Graphlets to Directed Networks. Retrieved April 9, 2020, from https://doi.org/10.1109/TCBB.2016.2586046

Aparício, D., Ribeiro, P., & Silva, F. (2015). Network comparison using directed graphlets [arXiv: 1511.01964]. *arXiv:1511.01964 [physics, q-bio]*. Retrieved April 9, 2020, from http://arxiv.org/abs/1511.01964

Arsov, N., & Mirceva, G. (2019). Network Embedding: An Overview [arXiv: 1911.11726]. *arXiv:1911.11726 [cs, stat]*. Retrieved March 25, 2020, from http://arxiv.org/abs/1911.11726

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., & Sherlock, G. (2000). Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, *25*(1), 25–29. https://doi.org/10.1038/75556

Athanasios, A., Charalampos, V., Vasileios, T., et al. (2017). Protein-protein interaction (ppi) network: Recent advances in drug discovery. *Current drug metabolism*, *18*(1), 5–10.

Ayala, F. J. (1987). Biological Reductionism. In F. E. Yates, A. Garfinkel, D. O. Walter, & G. B. Yates (Eds.), *Self-Organizing Systems: The Emergence of Order* (pp. 315–324). Springer US. https://doi.org/10.1007/978-1-4613-0883-6_17

Barabasi, A.-L., & Oltvai, Z. N. (2004). Network biology: Understanding the cell's functional organization. *Nature reviews genetics*, *5*(2), 101–113.

Barabási, A.-L., & Albert, R. (1999). Emergence of Scaling in Random Networks [Publisher: American Association for the Advancement of Science Section: Report]. *Science*, *286*(5439), 509–512. https://doi.org/10.1126/science.286.5439.509

Baryshnikova, A. (2016). Systematic Functional Annotation and Visualization of Biological Networks. *Cell Systems*, *2*(6), 412–421. https://doi.org/10.1016/j.cels.2016.04.014

Baryshnikova, A., Costanzo, M., Kim, Y., Ding, H., Koh, J., Toufighi, K., Youn, J.-Y., Ou, J., San Luis, B.-J., Bandyopadhyay, S., et al. (2010). Quantitative analysis of fitness and genetic interactions in yeast on a genome scale. *Nature methods*, *7*(12), 1017–1024.

Belkin, M., & Niyogi, P. (2001). Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering. *NIPS*. https://doi.org/10.7551/mitpress/1120.003.0080

Belkin, M., & Niyogi, P. (2003). Laplacian Eigenmaps for dimensionality reduction and data representation. Retrieved March 25, 2020, from https://doi.org/10.1162/089976603321780317

Bellman, R. (1958). On a routing problem. *Quarterly of applied mathematics*, *16*(1), 87–90.

Bender, A., & Pringle, J. R. (1991). Use of a screen for synthetic lethal and multicopy suppressee mutants to identify two new genes involved in morphogenesis in saccharomyces cerevisiae. *Molecular and cellular biology*, *11*(3), 1295–1305.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, *57*(1), 289–300.

Bianconi, G., & Rahmede, C. (2017). Emergent Hyperbolic Network Geometry [Number: 1 Publisher: Nature Publishing Group]. *Scientific Reports*, *7*(1), 1–9. https://doi.org/10.1038/srep41974

Bläsius, T., Friedrich, T., & Katzmann, M. (2021). Force-directed embedding of scale-free networks in the hyperbolic plane. *19th International Symposium on Experimental Algorithms (SEA 2021)*.

Blattner, F. R., Plunkett III, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., et al. (1997). The complete genome sequence of escherichia coli k-12. *science*, *277*(5331), 1453–1462.

Boguna, M., Krioukov, D., & Claffy, K. C. (2009). Navigability of complex networks. *Nature Physics*, *5*(1), 74–80.

Borg, I., & Groenen, P. (1997). *Modern Multidimensional Scaling: Theory and Applications*. Springer-Verlag. https://doi.org/10.1007/978-1-4757-2711-1

Boucher, B., & Jenna, S. (2013). Genetic interaction networks: Better understand to better predict. *Frontiers in genetics*, *4*, 290.

Bratlie, M. S., Johansen, J., Sherman, B. T., Huang, D. W., Lempicki, R. A., & Drabløs, F. (2010). Gene duplications in prokaryotes can be associated with environmental adaptation. *BMC genomics*, *11*(1), 1–17.

Broido, A. D., & Clauset, A. (2019). Scale-free networks are rare. *Nature communications*, *10*(1), 1017.

Brückner, A., Polge, C., Lentze, N., Auerbach, D., & Schlattner, U. (2009). Yeast two-hybrid, a powerful tool for systems biology. *International journal of molecular sciences*, *10*(6), 2763–2788.

Cahan, P., Li, H., Morris, S. A., Da Rocha, E. L., Daley, G. Q., & Collins, J. J. (2014). Cellnet: Network biology applied to stem cell engineering. *Cell*, *158*(4), 903–915.

Cai, H., Zheng, V. W., & Chang, K. C.-C. (2018). A Comprehensive Survey of Graph Embedding: Problems, Techniques and Applications [arXiv: 1709.07604]. *arXiv:1709.07604 [cs]*. Retrieved March 25, 2020, from http://arxiv.org/abs/1709.07604

Cannistraci, C. V., Alanis-Lobato, G., & Ravasi, T. (2013). Minimum curvilinearity to enhance topological prediction of protein interactions by network embedding. *Bioinformatics (Oxford, England)*, *29*(13), i199–209. https://doi.org/10.1093/bioinformatics/btt208

Cannistraci, C. V., & Muscoloni, A. (2018). Latent geometry inspired graph dissimilarities enhance affinity propagation community detection in complex networks. *arXiv preprint arXiv:1804.04566*.

Cannistraci, C. V., Ravasi, T., Montevecchi, F. M., Ideker, T., & Alessio, M. (2010). Nonlinear dimension reduction and clustering by Minimum Curvilinearity unfold neuropathic pain and tissue embryological classes. *Bioinformatics (Oxford, England)*, *26*(18), i531–539. https://doi.org/10.1093/bioinformatics/btq376

Cannoodt, R., Ruyssinck, J., Ramon, J., Preter, K. D., & Saeys, Y. (2018). IncGraph: Incremental graphlet counting for topology optimisation [Publisher: Public Library of Science]. *PLOS ONE*, *13*(4), e0195997. https://doi.org/10.1371/journal.pone.0195997

Chakraborty, C., Priya, D., Chen, L., Zhu, H., et al. (2014). Evaluating protein-protein interaction (ppi) networks for diseases pathway, target discovery, and drug-design using 'in silico pharmacology'. *Current Protein and Peptide Science*, *15*(6), 561–571.

Chalfie, M. (1995). Green fluorescent protein. *Photochemistry and photobiology*, *62*(4), 651–656.

Chen, H., Perozzi, B., Al-Rfou, R., & Skiena, S. (2018). A Tutorial on Network Embeddings [arXiv: 1808.02590]. *arXiv:1808.02590 [cs]*. Retrieved March 25, 2020, from http://arxiv.org/abs/1808.02590

Cheung, V. G., Morley, M., Aguilar, F., Massimi, A., Kucherlapati, R., & Childs, G. (1999). Making and reading microarrays. *Nature genetics*, *21*(1), 15–19.

Chua, H. N., Ning, K., Sung, W.-K., Leong, H. W., & Wong, L. (2008). Using indirect protein–protein interactions for protein complex prediction. *Journal of bioinformatics and computational biology*, *6*(03), 435–466.

Clauset, A., Shalizi, C. R., & Newman, M. E. (2009). Power-law distributions in empirical data. *SIAM review*, *51*(4), 661–703.

Collins, S. R., Kemmeren, P., Zhao, X.-C., Greenblatt, J. F., Spencer, F., Holstege, F. C., Weissman, J. S., & Krogan, N. J. (2007). Toward a comprehensive atlas of the physical interactome of saccharomyces cerevisiae. *Molecular & Cellular Proteomics*, *6*(3), 439–450.

Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E. D., Sevier, C. S., Ding, H., Koh, J. L. Y., 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., . . . Boone, C. (2010). The genetic landscape of a cell. *Science (New York, N.Y.)*, *327*(5964), 425–431. https://doi.org/10.1126/science.1180823

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., . . . Boone, C. (2016). A global genetic interaction network maps a wiring diagram of cellular function. *Science (New York, N.Y.)*, *353*(6306). https://doi.org/10.1126/science.aaf1420

Cross, R., Laseter, T., Parker, A., & Velasquez, G. (2006). Using social network analysis to improve communities of practice. *California Management Review*, *49*(1), 32–60.

Cunningham, F., Allen, J. E., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Austine-Orimoloye, O., Azov, A. G., Barnes, I., Bennett, R., et al. (2022). Ensembl 2022. *Nucleic acids research*, *50*(D1), D988–D995.

Dale, M. R. T. (2017). Graph Structure and System Function: Graphlet Methods [Library Catalog: www.cambridge.org Pages: 252-270 Publisher: Cambridge University Press]. https://doi.org/10.1017/9781316105450.012

Danon, L., Diaz-Guilera, A., Duch, J., & Arenas, A. (2005). Comparing community structure identification. *Journal of statistical mechanics: Theory and experiment*, *2005*(09), P09008.

Davis, D., Yaveroğlu, Ö. N., Malod-Dognin, N., Stojmirovic, A., & Pržulj, N. (2015). Topology-function conservation in protein–protein interaction networks. *Bioinformatics*, *31*(10), 1632–1639.

Dessimoz, C., & Škunca, N. (2017). *The gene ontology handbook*. Springer Nature.

Dhingra, S., Sowdhamini, R., Cadet, F., & Offmann, B. (2020). A glance into the evolution of template-free protein structure prediction methodologies. *Biochimie*, *175*, 85–92. https://doi.org/https://doi.org/10.1016/j.biochi.2020.04.026

Dijkstra, E., Beauguitte, L., & Maisonobe, M. (2021). Ew dijkstra, 1959, a note on two problems in connexion with graphs. numerische mathematik 1, p. 269271 version bilingue et commentée.

Dobson, R. J., Munroe, P. B., Caulfield, M. J., & Saqi, M. A. (2014). Protein interaction networks associated with cardiovascular disease and cancer: Exploring the effect of bias on shared network properties. *International journal of data mining and bioinformatics*, *9*(4), 339–357.

Emmert-Streib, F., & Glazko, G. V. (2011). Network biology: A direct approach to study biological function. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, *3*(4), 379–391.

Erdös, P., & Rényi, A. (1959). On random graphs i. *Publicationes Mathematicae Debrecen*, *6*, 290.

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. https://doi.org/https://doi.org/10.1016/j.laa.2011.11.032

Estrada, E., Hameed, E., Hatano, N., & Langer, M. (2017). Path laplacian operators and superdiffusive processes on graphs. i. one-dimensional case. *Linear Algebra and its Applications*, *523*, 307–334. https://doi.org/https://doi.org/10.1016/j.laa.2017.02.027

Favila, M. E., & Halffter, G. (1997). The use of indicator groups for measuring biodiversity as related to community structure and function. *Acta Zoológica Mexicana (ns)*, (72), 1–25.

Fruchterman, T. M., & Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and experience*, *21*(11), 1129–1164.

Galarneau, A., Primeau, M., Trudeau, L.-E., & Michnick, S. W. (2002). β-lactamase protein fragment complementation assays as in vivo and in vitro sensors of protein–protein interactions. *Nature biotechnology*, *20*(6), 619–622.

García-Díaz, P., Sánchez-Berriel, I., Martínez-Rojas, J. A., & Diez-Pascual, A. M. (2020). Unsupervised feature selection algorithm for multiclass cancer classification of gene expression RNA-Seq data. *Genomics*, *112*(2), 1916–1925. https://doi.org/https://doi.org/10.1016/j.ygeno.2019.11.004

García-Pérez, G., Allard, A., Serrano, M. Á., & Boguñá, M. (2019). Mercator: Uncovering faithful hyperbolic embeddings of complex networks [arXiv: 1904.10814]. *arXiv:1904.10814 [physics]*. Retrieved April 9, 2020, from http://arxiv.org/abs/1904.10814

Gavin, A.-C., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A.-M., Cruciat, C.-M., et al. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, *415*(6868), 141–147.

Giot, L., Bader, J. S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y., Ooi, C., Godwin, B., Vitols, E., et al. (2003). A protein interaction map of drosophila melanogaster. *science*, *302*(5651), 1727–1736.

Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the national academy of sciences*, *99*(12), 7821–7826.

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

Glish, G. L., & Vachet, R. W. (2003). The basics of mass spectrometry in the twenty-first century. *Nature reviews drug discovery*, *2*(2), 140–150.

Grover, A., & Leskovec, J. (2016). Node2vec: Scalable Feature Learning for Networks [arXiv: 1607.00653]. *arXiv:1607.00653 [cs, stat]*. Retrieved March 25, 2020, from http://arxiv.org/abs/1607.00653

Gutiérrez-Gómez, L., & Delvenne, J.-C. (2019). Unsupervised network embeddings with node identity awareness. *Applied Network Science*, *4*(1), 82. https://doi.org/10.1007/s41109-019-0197-1

Ho, B., Baryshnikova, A., & Brown, G. (2018). Unification of protein abundance datasets yields a quantitative saccharomyces cerevisiae proteome. cell syst 6: 192–205. e3.

Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S.-L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M., Muskat, B., … Tyers, M. (2002). Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry [Number: 6868 Publisher: Nature Publishing Group]. *Nature*, *415*(6868), 180–183. https://doi.org/10.1038/415180a

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

Huang, X.-T., Zhu, Y., Chan, L. L. H., Zhao, Z., & Yan, H. (2016). An integrative C. elegans protein-protein interaction network with reliability assessment based on a probabilistic graphical model. *Molecular bioSystems*, *12*(1), 85–92. https://doi.org/10.1039/c5mb00417a

Hulovatyy, Y., Chen, H., & Milenković, T. (2015). Exploring the structure and function of temporal networks with dynamic graphlets [Publisher: Oxford Academic]. *Bioinformatics*, *31*(12), i171–i180. https://doi.org/10.1093/bioinformatics/btv227

Ideker, T., Galitski, T., & Hood, L. (2001). A NEW APPROACH TO DECODING LIFE: Systems Biology [Publisher: Annual Reviews]. *Annual Review of Genomics and Human Genetics*, *2*(1), 343–372. https://doi.org/10.1146/annurev.genom.2.1.343

Ideker, T., & Krogan, N. J. (2012). Differential network biology. *Molecular systems biology*, *8*(1), 565.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., & Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences*, *98*(8), 4569–4574.

Ito, T., Tashiro, K., Muta, S., Ozawa, R., Chiba, T., Nishizawa, M., Yamamoto, K., Kuhara, S., & Sakaki, Y. (2000). Toward a protein–protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proceedings of the National Academy of Sciences*, *97*(3), 1143–1147.

Iyer, V. R., Horak, C. E., Scafe, C. S., Botstein, D., Snyder, M., & Brown, P. O. (2001). Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF [Number: 6819 Publisher: Nature Publishing Group]. *Nature*, *409*(6819), 533–538. https://doi.org/10.1038/35054095

Jaccard, P. (1912). The distribution of the flora in the alpine zone. 1. *New phytologist*, *11*(2), 37–50.

Jeong, H., Mason, S. P., Barabási, A.-L., & Oltvai, Z. N. (2001a). Lethality and centrality in protein networks [Number: 6833 Publisher: Nature Publishing Group]. *Nature*, *411*(6833), 41–42. https://doi.org/10.1038/35075138

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

Jiang, L., Zhong, M., Chen, T., Zhu, X., Yang, H., & Lv, K. (2020). Gene regulation network analysis reveals core genes associated with survival in glioblastoma multiforme. *Journal of Cellular and Molecular Medicine*, *24*(17), 10075–10087.

Johnston, M. (1998). Gene chips: Array of hope for understanding gene regulation. *Current Biology*, *8*(5), R171–R174.

Kadonaga, J. T., & Tjian, R. (1986). Affinity purification of sequence-specific dna binding proteins. *Proceedings of the National Academy of Sciences*, *83*(16), 5889–5893.

Kamada, T., Kawai, S., et al. (1989). An algorithm for drawing general undirected graphs. *Information processing letters*, *31*(1), 7–15.

Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M., & Ishiguro-Watanabe, M. (2023). Kegg for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Research*, *51*(D1), D587–D592.

Kellis, M., Birren, B. W., & Lander, E. S. (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast saccharomyces cerevisiae. *Nature*, *428*(6983), 617–624.

Kirschner, M. W. (2005). The Meaning of Systems Biology [Publisher: Elsevier]. *Cell*, *121*(4), 503–504. https://doi.org/10.1016/j.cell.2005.05.005

Knabe, J. F. (2013). Topological Network Analysis. In J. F. Knabe (Ed.), *Computational Genetic Regulatory Networks: Evolvable, Self-organizing Systems* (pp. 71–81). Springer. https://doi.org/10.1007/978-3-642-30296-1_5

Kobourov, S. G. (2012). Spring Embedders and Force Directed Graph Drawing Algorithms [arXiv: 1201.3011]. *arXiv:1201.3011 [cs]*. Retrieved March 25, 2020, from http://arxiv.org/abs/1201.3011

Krioukov, D., Papadopoulos, F., Kitsak, M., Vahdat, A., & Boguna, M. (2010). Hyperbolic Geometry of Complex Networks [arXiv: 1006.5169]. *Physical Review E*, *82*(3), 036106. https://doi.org/10.1103/PhysRevE.82.036106

Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A. P., et al. (2006). Global landscape of protein complexes in the yeast saccharomyces cerevisiae. *Nature*, *440*(7084), 637–643.

Kulmanov, M., Khan, M. A., & Hoehndorf, R. (2018). DeepGO: Predicting protein functions from sequence and interactions using a deep ontology-aware classifier [Publisher: Oxford Academic]. *Bioinformatics*, *34*(4), 660–668. https://doi.org/10.1093/bioinformatics/btx624

Kuzmanov, U., & Emili, A. (2013). Protein-protein interaction networks: Probing disease mechanisms using model systems. *Genome medicine*, *5*(4), 1–12.

Kuzmin, E., VanderSluis, B., Ba, A. N. N., Wang, W., Koch, E. N., Usaj, M., Khmelinskii, A., Usaj, M. M., Leeuwen, J. v., Kraus, O., Tresenrider, A., Pryszlak, M., Hu, M.-C., Varriano, B., Costanzo, M., Knop, M., Moses, A., Myers, C. L., Andrews, B. J., & Boone, C. (2020). Exploring whole-genome duplicate gene retention with complex genetic interaction analysis [Publisher: American Association for the Advancement of Science Section: Research Article]. *Science*, *368*(6498). https://doi.org/10.1126/science.aaz5667

Kuzmin, E., VanderSluis, B., Wang, W., Tan, G., Deshpande, R., Chen, Y., Usaj, M., Balint, A., Mattiazzi Usaj, M., Van Leeuwen, J., et al. (2018). Systematic analysis of complex genetic interactions. *Science*, *360*(6386), eaao1729.

Lehner, B., Tischler, J., & Fraser, A. G. (2006). Rnai screens in caenorhabditis elegans in a 96-well liquid format and their application to the systematic identification of genetic interactions. *Nature protocols*, *1*(3), 1617–1620.

Li, G., Luo, J., Xiao, Q., Liang, C., Ding, P., & Cao, B. (2017). Predicting MicroRNA-Disease Associations Using Network Topological Similarity Based on DeepWalk [Conference Name: IEEE Access]. *IEEE Access*, *5*, 24032–24039. https://doi.org/10.1109/ACCESS.2017.2766758

Li, M. M., Huang, K., & Zitnik, M. (2022). Graph representation learning in biomedicine and healthcare. *Nature Biomedical Engineering*, 1–17.

Li, S., Armstrong, C. M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P.-O., Han, J.-D. J., Chesneau, A., Hao, T., et al. (2004). A map of the interactome network of the metazoan c. elegans. *Science*, *303*(5657), 540–543.

Li, X., Cai, H., Xu, J., Ying, S., & Zhang, Y. (2010). A mouse protein interactome through combined literature mining with multiple sources of interaction evidence. *Amino Acids*, *38*(4), 1237–1252. https://doi.org/10.1007/s00726-009-0335-7

Luck, K., Kim, D.-K., Lambourne, L., Spirohn, K., Begg, B. E., Bian, W., Brignall, R., Cafarelli, T., Campos-Laborie, F. J., Charloteaux, B., Choi, D., Cote, A. G., Daley, M., Deimling, S., Desbuleux, A., Dricot, A., Gebbia, M., Hardy, M. F., Kishore, N., . . . Calderwood, M. A. (2019). A reference map of the human protein interactome [Publisher: Cold Spring Harbor Laboratory Section: New Results]. *bioRxiv*, 605451. https://doi.org/10.1101/605451

Luck, K., Kim, D.-K., Lambourne, L., Spirohn, K., Begg, B. E., Bian, W., Brignall, R., Cafarelli, T., Campos-Laborie, F. J., Charloteaux, B., Choi, D., Coté, A. G., Daley, M., Deimling, S., Desbuleux, A., Dricot, A., Gebbia, M., Hardy, M. F., Kishore, N., . . . Calderwood, M. A. (2020). A reference map of the human binary protein interactome [Number: 7803 Publisher: Nature Publishing Group]. *Nature*, *580*(7803), 402–408. https://doi.org/10.1038/s41586-020-2188-x

Ma, X., Gao, L., Yong, X., & Fu, L. (2010). Semi-supervised clustering algorithm for community structure detection in complex networks. *Physica A: Statistical Mechanics and its Applications*, *389*(1), 187–197.

Mani, R., St. Onge, R. P., Hartman IV, J. L., Giaever, G., & Roth, F. P. (2008). Defining genetic interaction. *Proceedings of the National Academy of Sciences*, *105*(9), 3461–3466.

Martin, A. J., Contreras-Riquelme, S., Dominguez, C., & Perez-Acle, T. (2017). LoTo: A graphlet based method for the comparison of local topology between gene regulatory networks [Publisher: PeerJ]. *PeerJ*, *2017*(2), e3052. https://doi.org/10.7717/peerj.3052

Mercatelli, D., Scalambra, L., Triboli, L., Ray, F., & Giorgi, F. M. (2020). Gene regulatory network inference resources: A practical overview. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, *1863*(6), 194430.

Milenkoviæ, T., & Pržulj, N. (2008). Uncovering Biological Network Function via Graphlet Degree Signatures. *Cancer Informatics*, *6*, 257–273. Retrieved March 25, 2020, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2623288/

Milenković, T., Memišević, V., Ganesan, A. K., & Pržulj, N. (2010). Systems-level cancer gene identification from protein interaction network topology applied to melanogenesis-related functional genomics data. *Journal of the Royal Society Interface*, *7*(44), 423–437.

Moreira, A. A., Andrade Jr, J. S., Herrmann, H. J., & Indekeu, J. O. (2009). How to make a fragile network robust and vice versa. *Physical review letters*, *102*(1), 018701.

Muscoloni, A., & Cannistraci, C. V. (2018). A nonuniform popularity-similarity optimization (nPSO) model to efficiently generate realistic complex networks with communities [Publisher: IOP Publishing]. *New Journal of Physics*, *20*(5), 052002. https://doi.org/10.1088/1367-2630/aac06f

Muscoloni, A., Thomas, J. M., Ciucci, S., Bianconi, G., & Cannistraci, C. V. (2017). Machine learning meets complex networks via coalescent embedding in the hyperbolic space [Number: 1 Publisher: Nature Publishing Group]. *Nature Communications*, *8*(1), 1–19. https://doi.org/10.1038/s41467-017-01825-5

Nelson, W., Zitnik, M., Wang, B., Leskovec, J., Goldenberg, A., & Sharan, R. (2019). To Embed or Not: Network Embedding as a Paradigm in Computational Biology. *Frontiers in Genetics*, *10*. https://doi.org/10.3389/fgene.2019.00381

Newman, M. E. J. (2010). *Networks: An introduction* [OCLC: ocn456837194]. Oxford University Press.

Newman, M. E. (2004). Detecting community structure in networks. *The European physical journal B*, *38*, 321–330.

Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the national academy of sciences*, *103*(23), 8577–8582.

Ng, A. Y., Jordan, M. I., & Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm. *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, 849–856.

Ning, M., & Lo, E. H. (2010). Opportunities and challenges in omics. *Translational stroke research*, *1*, 233–237.

Niu, H., Wang, J., Zhao, T., Shu, N., & He, Y. (2012). Revealing Topological Organization of Human Brain Functional Networks with Resting-State Functional near Infrared Spectroscopy [Publisher: Public Library of Science]. *PLOS ONE*, *7*(9), e45771. https://doi.org/10.1371/journal.pone.0045771

Novick, P., & Botstein, D. (1985). Phenotypic analysis of temperature-sensitive yeast actin mutants. *Cell*, *40*(2), 405–416.

Obayashi, T., Kagaya, Y., Aoki, Y., Tadaka, S., & 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. https://doi.org/10.1093/nar/gky1155

O'Reilly, F. J., & Rappsilber, J. (2018). Cross-linking mass spectrometry: Methods and applications in structural, molecular and systems biology. *Nature structural & molecular biology*, *25*(11), 1000–1008.

Patterson, C. (1988). Homology in classical and molecular biology. *Molecular biology and evolution*, *5*(6), 603–625.

Pearson, W. R. (2013). An introduction to sequence similarity ("homology") searching. *Current protocols in bioinformatics*, *42*(1), 3–1.

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

Perozzi, B., Al-Rfou, R., & Skiena, S. (2014). DeepWalk: Online learning of social representations. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 701–710. https://doi.org/10.1145/2623330.2623732

Piersimoni, L., Kastritis, P. L., Arlt, C., & Sinz, A. (2021). Cross-linking mass spectrometry for investigating protein conformations and protein–protein interactions a method for all seasons. *Chemical Reviews*, *122*(8), 7500–7531.

Piovesan, D., Giollo, M., Ferrari, C., & Tosatto, S. C. (2015). Protein function prediction using guilty by association from interaction networks. *Amino Acids*, *47*, 2583–2592.

Przulj, N., Corneil, D. G., & Jurisica, I. (2004). Modeling interactome: Scale-free or geometric? *Bioinformatics (Oxford, England)*, *20*(18), 3508–3515. https://doi.org/10.1093/bioinformatics/bth436

Przulj, N. (2007). Biological network comparison using graphlet degree distribution. *Bioinformatics (Oxford, England)*, *23*(2), e177–183. https://doi.org/10.1093/bioinformatics/btl301

Przulj, N., Kuchaiev, O., Stevanovi?, A., & Hayes, W. (2009). Geometric evolutionary dynamics of protein interaction networks. In *Biocomputing 2010* (pp. 178–189). WORLD SCIENTIFIC. https://doi.org/10.1142/9789814295291_0020

Pržulj, N., & Higham, D. J. (2006). Modelling protein–protein interaction networks via a stickiness index. *Journal of the Royal Society Interface*, *3*(10), 711–716. https://doi.org/10.1098/rsif.2006.0147

Puig, O., Caspary, F., Rigaut, G., Rutz, B., Bouveret, E., Bragado-Nilsson, E., Wilm, M., & Séraphin, B. (2001). The tandem affinity purification (tap) method: A general procedure of protein complex purification. *Methods*, *24*(3), 218–229.

Purkanti, R., & Thattai, M. (2022). Genome doubling enabled the expansion of yeast vesicle traffic pathways. *Scientific Reports*, *12*(1), 11213.

Rada, R., Mili, H., Bicknell, E., & Blettner, M. (1989). Development and application of a metric on semantic nets. *IEEE transactions on systems, man, and cybernetics*, *19*(1), 17–30.

Rajesh, A., Chang, Y., Abedalthagafi, M. S., Wong-Beringer, A., Love, M. I., & Mangul, S. (2021). Improving the completeness of public metadata accompanying omics studies.

Ravasz, E., & Barabási, A.-L. (2003). Hierarchical organization in complex networks. *Physical review E*, *67*(2), 026112.

Remy, I., Campbell-Valois, F., & Michnick, S. W. (2007). Detection of protein–protein interactions using a simple survival protein-fragment complementation assay based on the enzyme dihydrofolate reductase. *Nature protocols*, *2*(9), 2120–2125.

Remy, I., & Michnick, S. W. (1999). Clonal selection and in vivo quantitation of protein interactions with protein-fragment complementation assays. *Proceedings of the National Academy of Sciences*, *96*(10), 5394–5399.

Richards, A. L., Eckhardt, M., & Krogan, N. J. (2021). Mass spectrometry-based protein–protein interaction networks for the study of human diseases. *Molecular systems biology*, *17*(1), e8792.

Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M., & Séraphin, B. (1999). A generic protein purification method for protein complex characterization and proteome exploration. *Nature biotechnology*, *17*(10), 1030–1032.

Rizzolo, K., Huen, J., Kumar, A., Phanse, S., Vlasblom, J., Kakihara, Y., Zeineddine, H. A., Minic, Z., Snider, J., Wang, W., Pons, C., Seraphim, T. V., Boczek, E. E., Alberti, S., Costanzo, M., Myers, C. L., Stagljar, I., Boone, C., Babu, M., & Houry, W. A. (2017). Features of the Chaperone Cellular Network Revealed through Systematic Interaction Mapping. *Cell Reports*, *20*(11), 2735–2748. https://doi.org/10.1016/j.celrep.2017.08.074

Rual, J.-F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G. F., Gibbons, F. D., Dreze, M., Ayivi-Guedehoussou, N., Klitgord, N., Simon, C., Boxem, M., Milstein, S., Rosenberg, J., Goldberg, D. S., Zhang, L. V., Wong, S. L., Franklin, G., . . . Vidal, M. (2005). Towards a proteome-scale map of the human protein–protein interaction network [Number: 7062

Publisher: Nature Publishing Group]. *Nature*, *437*(7062), 1173–1178. https://doi.org/10.1038/nature04209

Safari-Alighiarloo, N., Taghizadeh, M., Rezaei-Tavirani, M., Goliaei, B., & Peyvandi, A. A. (2014). Protein-protein interaction networks (ppi) and complex diseases. *Gastroenterology and Hepatology from bed to bench*, *7*(1), 17.

Santolini, M., & Barabási, A.-L. (2018). Predicting perturbation patterns from the topology of biological networks [Publisher: National Academy of Sciences Section: PNAS Plus]. *Proceedings of the National Academy of Sciences*, *115*(27), E6375–E6383. https://doi.org/10.1073/pnas.1720589115

Sarajlić, A., Janjić, V., Stojković, N., Radak, D., & Pržulj, N. (2013). Network topology reveals key cardiovascular disease genes. *PloS one*, *8*(8), e71537.

Sarajlić, A., Malod-Dognin, N., Yaveroğlu, Ö. N., & Pržulj, N. (2016). Graphlet-based Characterization of Directed Networks [Number: 1 Publisher: Nature Publishing Group]. *Scientific Reports*, *6*(1), 1–14. https://doi.org/10.1038/srep35098

Silverman, E. K., Schmidt, H. H., Anastasiadou, E., Altucci, L., Angelini, M., Badimon, L., Balligand, J.-L., Benincasa, G., Capasso, G., Conte, F., et al. (2020). Molecular networks in network medicine: Development and applications. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, *12*(6), e1489.

Simonis, N., Rual, J.-F., Carvunis, A.-R., Tasan, M., Lemmens, I., Hirozane-Kishikawa, T., Hao, T., Sahalie, J. M., Venkatesan, K., Gebreab, F., et al. (2009). Empirically controlled mapping of the caenorhabditis elegans protein-protein interactome network. *Nature methods*, *6*(1), 47–54.

Solava, R. W., Michaels, R. P., & Milenkovic, T. (2012). Identifying edge clusters in networks via edge graphlet degree vectors (edge-gdvs) and edge-gdv-similarities. *arXiv preprint arXiv:1204.2255*.

Song, W., & Wang, S. (2019). Hyperbolic Node Embedding for Signed Networks [arXiv: 1910.13090]. *arXiv:1910.13090 [physics]*. Retrieved April 9, 2020, from http://arxiv.org/abs/1910.13090

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F. H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., Timm, J., Mintzlaff, S., Abraham, C., Bock, N., Kietzmann, S., Goedde, A., Toksöz, E., Droege, A., Krobitsch, S., . . . Wanker, E. E. (2005). A Human Protein-Protein Interaction Network: A Resource for Annotating the Proteome. *Cell*, *122*(6), 957–968. https://doi.org/10.1016/j.cell.2005.08.029

Stoughton, R. B. (2005). Applications of dna microarrays in biology. *Annu. Rev. Biochem.*, *74*, 53–82.

Stuart, J. M., Segal, E., Koller, D., & Kim, S. K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *science*, *302*(5643), 249–255.

Sun, Y., Danila, B., Josić, K., & Bassler, K. E. (2009). Improved community structure detection using a modified fine-tuning strategy. *Europhysics Letters*, *86*(2), 28004.

Tanimoto, T. T. (1958). Elementary mathematical theory of classification and prediction.

Tenenbaum, J. B., de Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science (New York, N.Y.)*, *290*(5500), 2319–2323. https://doi.org/10.1126/science.290.5500.2319

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.

Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., & Rothberg, J. M. (2000). A comprehensive analysis of protein–protein interactions in Saccharomyces cerevisiae [Number: 6770 Publisher: Nature Publishing Group]. *Nature*, *403*(6770), 623–627. https://doi.org/10.1038/35001009

Van Criekinge, W., & Beyaert, R. (1999). Yeast two-hybrid: State of the art. *Biological procedures online*, *2*, 1–38.

Vazquez, A., Flammini, A., Maritan, A., & Vespignani, A. (2001). Modeling of protein interaction networks [arXiv: cond-mat/0108043]. *arXiv:cond-mat/0108043*. Retrieved February 1, 2021, from http://arxiv.org/abs/cond-mat/0108043

Vazquez, A., Flammini, A., Maritan, A., & Vespignani, A. (2003). Global protein function prediction from protein-protein interaction networks. *Nature biotechnology*, *21*(6), 697–700.

Vella, D., Marini, S., Vitali, F., Di Silvestre, D., Mauri, G., & Bellazzi, R. (2018). MTGO: PPI Network Analysis Via Topological and Functional Module Identification [Number: 1 Publisher: Nature Publishing Group]. *Scientific Reports*, *8*(1), 1–13. https://doi.org/10.1038/s41598-018-23672-0

Vissiennon, K., Friederici, A. D., Brauer, J., & Wu, C.-Y. (2017). Functional organization of the language network in three- and six-year-old children. *Neuropsychologia*, *98*, 24–33. https://doi.org/10.1016/j.neuropsychologia.2016.08.014

Wang, B., Tang, H., Guo, C., & Xiu, Z. (2006). Entropy optimization of scale-free networks' robustness to random failures. *Physica A: Statistical Mechanics and its Applications*, *363*(2), 591–596.

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

Wang, J., Rao, S., Chu, J., Shen, X., Levasseur, D. N., Theunissen, T. W., & Orkin, S. H. (2006). A protein interaction network for pluripotency of embryonic stem cells. *Nature*, *444*(7117), 364–368.

Wang, P. (2022). Network biology: Recent advances and challenges. *Gene & Protein in Disease*, *1*(2), 101.

Wang, Z., Gerstein, M., & Snyder, M. (2009). Rna-seq: A revolutionary tool for transcriptomics. *Nature reviews genetics*, *10*(1), 57–63.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world'networks. *nature*, *393*(6684), 440–442.

Windels, S. F. L., Malod-Dognin, N., & Pržulj, N. (2019). Graphlet Laplacians for topology-function and topology-disease relationships. *Bioinformatics (Oxford, England)*, *35*(24), 5226–5234. https://doi.org/10.1093/bioinformatics/btz455

Winterbach, W., Van Mieghem, P., Reinders, M. J. T., Wang, H., & de Ridder, D. (2013). Local Topological Signatures for Network-Based Prediction of Biological Function. In A. Ngom, E. Formenti, J.-K. Hao, X.-M. Zhao, & T.

van Laarhoven (Eds.), *Pattern Recognition in Bioinformatics* (pp. 23–34). Springer. https://doi.org/10.1007/978-3-642-39159-0_3

Wu, Z., Menichetti, G., Rahmede, C., & Bianconi, G. (2015). Emergent complex network geometry. *Scientific Reports*, *5*, 10073. https://doi.org/10.1038/srep10073

Yan, J., Risacher, S. L., Shen, L., & Saykin, A. J. (2018). Network approaches to systems biology analysis of complex disease: Integrative methods for multi-omics data [Publisher: Oxford Academic]. *Briefings in Bioinformatics*, *19*(6), 1370–1381. https://doi.org/10.1093/bib/bbx066

Yan Tong, A. H., & Boone, C. (2006). Synthetic genetic array analysis in saccharomyces cerevisiae. *Yeast Protocol*, 171–191.

Yang, Z., Algesheimer, R., & Tessone, C. J. (2016). A Comparative Analysis of Community Detection Algorithms on Artificial Networks [Number: 1 Publisher: Nature Publishing Group]. *Scientific Reports*, *6*(1), 30750. https://doi.org/10.1038/srep30750

Yaveroğlu, Ö. N., Malod-Dognin, N., Davis, D., Levnajic, Z., Janjic, V., Karapandza, R., Stojmirovic, A., & Pržulj, N. (2014). Revealing the Hidden Language of Complex Networks [Number: 1 Publisher: Nature Publishing Group]. *Scientific Reports*, *4*(1), 4547. https://doi.org/10.1038/srep04547

Youn, J.-Y., Dunham, W. H., Hong, S. J., Knight, J. D. R., Bashkurov, M., Chen, G. I., Bagci, H., Rathod, B., MacLeod, G., Eng, S. W. M., Angers, S., Morris, Q., Fabian, M., Côté, J.-F., & Gingras, A.-C. (2018). High-Density Proximity Mapping Reveals the Subcellular Organization of mRNA-Associated Granules and Bodies. *Molecular Cell*, *69*(3), 517–532.e11. https://doi.org/10.1016/j.molcel.2017.12.020

Zachary, W. W. (1977). An information flow model for conflict and fission in small groups. *Journal of anthropological research*, *33*(4), 452–473.

Zong, N., Kim, H., Ngo, V., & Harismendy, O. (2017). Deep mining heterogeneous networks of biomedical linked data to predict novel drug–target associations [Publisher: Oxford Academic]. *Bioinformatics*, *33*(15), 2337–2344. https://doi.org/10.1093/bioinformatics/btx160

Appendix A

Appendix

A.1 Data

Omics network data

We create genetic interaction (GI), genetic interaction similarity (GIS) and protein-protein interaction networks for different model organisms. For creating the GI and PPI networks, we collect molecular interaction data from the BioGRID database version 3.5.177 (Oughtred et al., 2019) and filter the data to include only 'Genetic' or 'Physical' interactions, respectively. Additionally, for the PPI data, we also filter by the following experimental techniques: 'Two-hybrid', 'Affinity Capture-Luminescence', 'Affinity Capture-MS', 'Affinity Capture-RNA' and 'Affinity Capture-Western'.

To create COEX networks, we collect gene co-expression data from COXPRESdb version 7.3 (Obayashi et al., 2019). For constructing the network, we consider the most co- expressed genes by keeping the top 1% of all mutual ranks in the data. This is done by first calculating the Pearson correlation coefficient between gene expression profiles for all pairs of genes and constructing a Pearson correlation matrix. For each gene, all other genes are ranked according to their correlation values. Finally, a threshold is applied to the ranks, keeping the top 1% to build the network with highly reliable edges.

Finally, to create GIS networks, there is only data available for the budding yeast, which we collect from (Usaj et al., 2017). This dataset contains a matrix with the Pearson correlation coefficients (PCC) between the genetic interaction profiles of the genes. With this matrix we construct a network as previously described by Costanzo et al., (2010, 2016), in which a gene (i.e., node) is linked to another (i.e., connected by an edge) if the PCC between the corresponding profiles is $PCC >= 0.2$.

	GIS		
	Nodes	**Edges**	**Density**
Budding yeast	4,626	30,185	0.003

Table A.1. GIS Budding yeast molecular network data statistics. We report the number of nodes, the number of edges and the density (columns 1-3).

Gene functional annotation data

	Organism	Annotations	Genes
BP	Budding yeast	4,621	5,105
	E. coli	2,773	2,564
	Fission yeast	2,624	739
	Fruit fly	6,317	5,777
	Human	11,368	9,659
	House mouse	12,353	9,933
	Roundworm	4,210	3,060
CC	Budding yeast	960	4,652
	E. coli	221	2,139
	Fission yeast	574	767
	Fruit fly	911	3,762
	Human	1,539	10,648
	House mouse	1,223	7,979
	Roundworm	565	2,115
MF	Budding yeast	2,143	4,124
	E. coli	2,128	2,592
	Fission yeast	879	695
	Fruit fly	1,872	3,227
	Human	3,705	14,270
	House mouse	2,782	8,287
	Roundworm	1,230	2,045

Table A.2. Functional annotation data statistics. For each of the four different annotation types (row), we report the species (column 1), the total number of annotations (column 2) and the total number of genes that are annotated (column 3).

Gene functional annotation coverage statistics

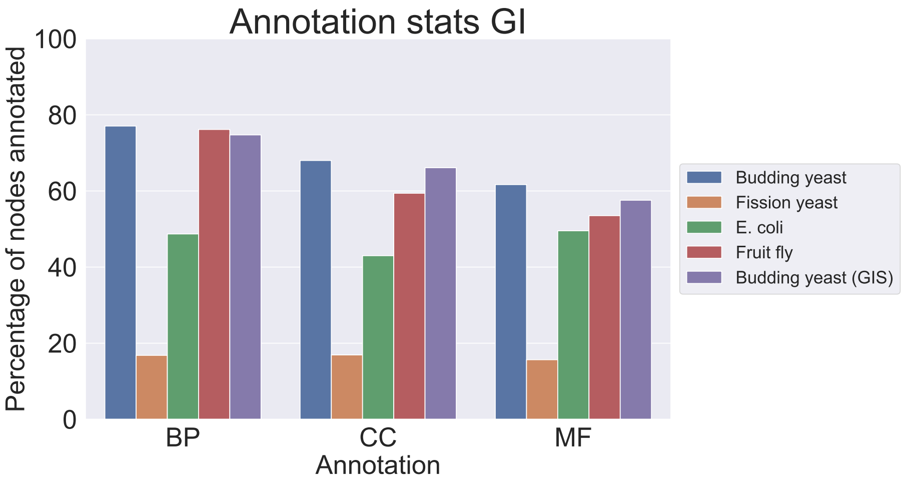


Figure A.1. Gene ontology annotation statistics for GI networks. For our four GI networks (and GIS for budding yeast), we report the percentage of nodes that are annotated by each of the corresponding GO annotation types: GO-BP, GO-CC and GO-MF.

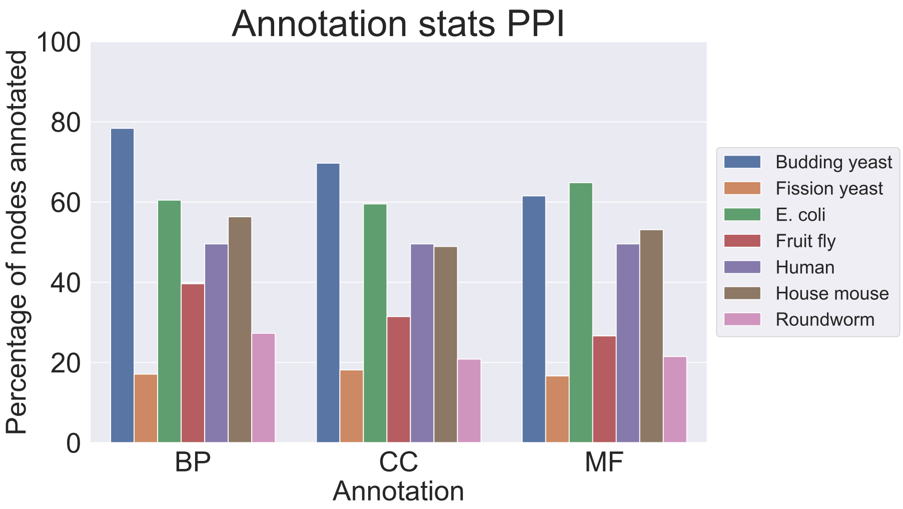


Figure A.2. Gene ontology annotation statistics for PPI networks. For our seven PPI networks, we report the percentage of nodes that are annotated by each of the corresponding GO annotation types: GO-BP, GO-CC and GO-MF.

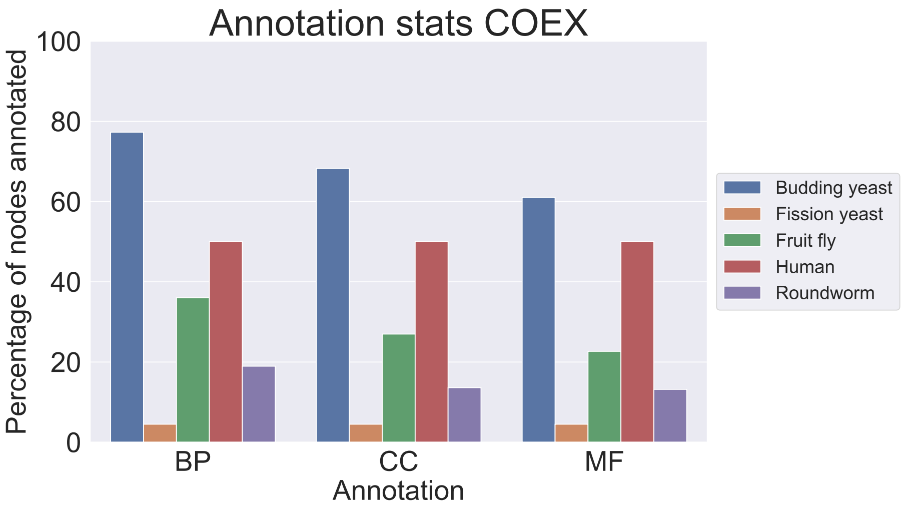


Figure A.3. Gene ontology annotation statistics for COEX networks. For our five COEX networks, we report the percentage of nodes that are annotated by each of the corresponding GO annotation types: GO-BP, GO-CC and GO-MF.

A.2 Enrichment statistics GI networks

In this section, we summarize the results obtained when using SAFE with the different graphlet based embedding algorithms. That is, the percentages of genes that have at least one annotation enriched in their neighborhood and the percentages of enriched annotations for all our GI molecular networks across different annotations.

Gene ontology biological processes

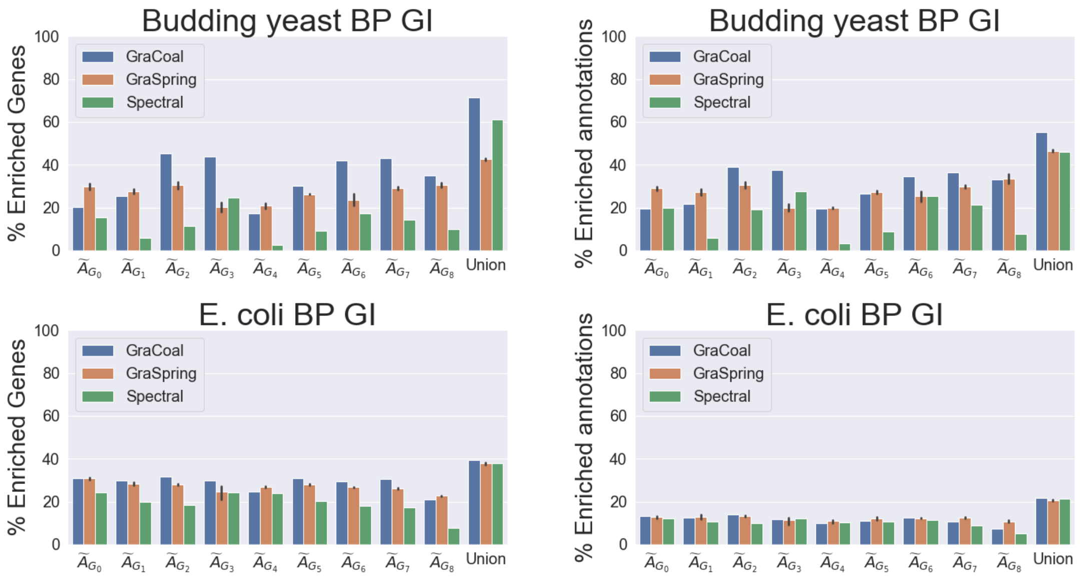


Figure A.4. SAFE GO-BP enrichment analysis for the GI networks, Part 1. On the y-axis, we show the percentages of genes that have at least one annotation enriched in their neighborhood (left) and the percentages of enriched annotations (right). On the x-axis, we show each of the embedding algorithms considered (legend) applied on different types of graphlet adjacencies. Graphlet adjacency 'Union' (x-axis, far right) considers the union of the enriched genes and enriched annotations across all graphlet adjacencies, i.e., $\widetilde{A}_{G0}$ to $\widetilde{A}_{G_8}$. The error bars for Spring embedding indicate the standard deviation across ten runs. From top to bottom: *E. coli*, Fruit fly, Fission yeast and Budding yeast, respectively.

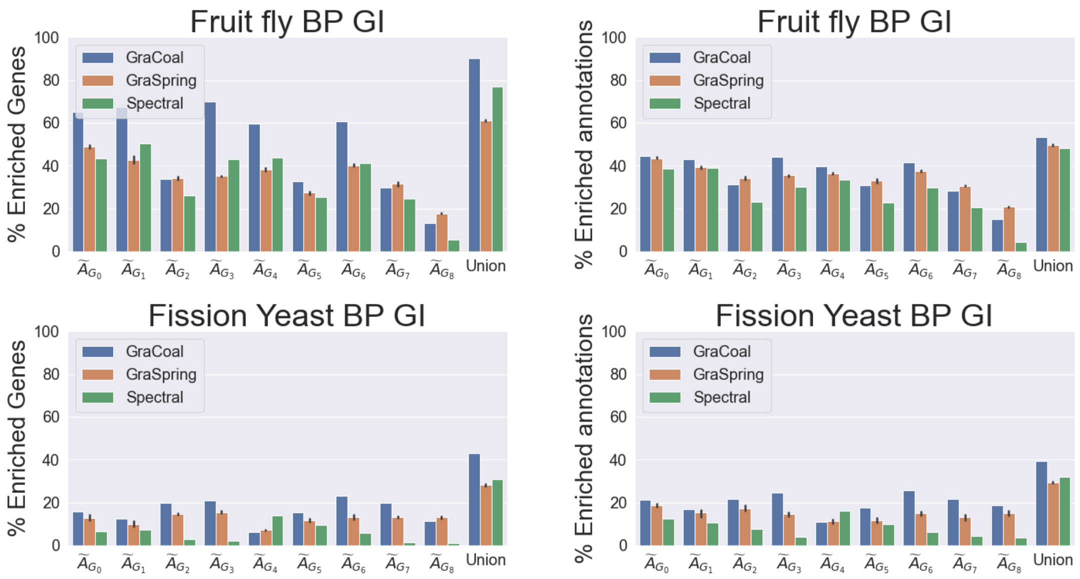


Figure A.4. SAFE GO-BP enrichment analysis for the GI networks, Part 2. On the y-axis, we show the percentages of genes that have at least one annotation enriched in their neighborhood (left) and the percentages of enriched annotations (right). On the x-axis, we show each of the embedding algorithms considered (legend) applied on different types of graphlet adjacencies. Graphlet adjacency 'Union' (x-axis, far right) considers the union of the enriched genes and enriched annotations across all graphlet adjacencies, i.e., $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$. The error bars for Spring embedding indicate the standard deviation across ten runs. From top to bottom: *E. coli*, Fruit fly, Fission yeast and Budding yeast, respectively.

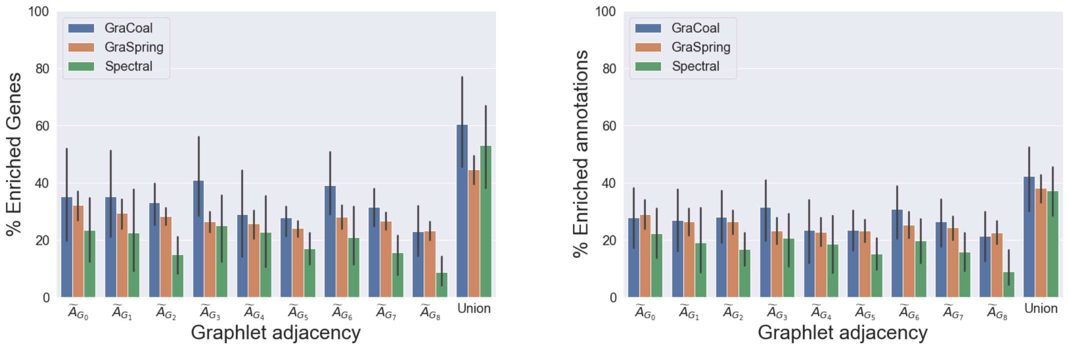


Figure A.5. SAFE GO-BP average enrichment statistics for the GI molecular networks. Average over all GI networks for the different types of underlying graphlet adjacencies, i.e., $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$. On the y-axis, we show the percentages of genes that have at least one annotation enriched in their neighborhood (left) and the percentages of enriched annotations (right). On the x-axis, we show each of the embedding algorithms considered (legend) applied on different types of graphlet adjacencies. Graphlet adjacency 'Union' (x-axis, far right) considers the union of the enriched genes and enriched annotations across all graphlet adjacencies.

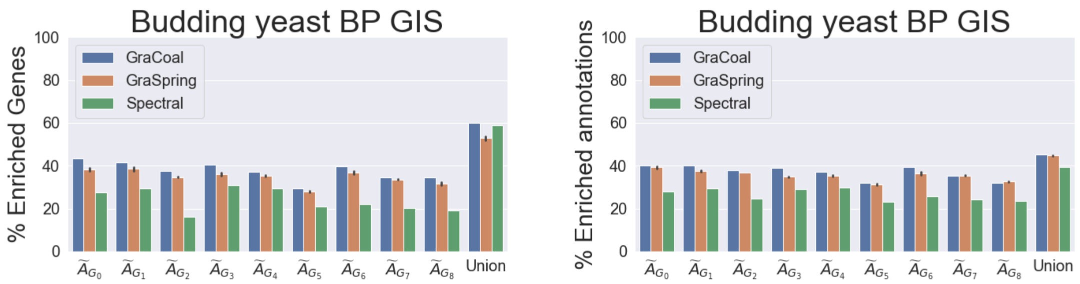


Figure A.6. SAFE GO-BP enrichment analysis for the Budding yeast GIS network. On the y-axis, we show the percentages of genes that have at least one annotation enriched in their neighborhood (left) and the percentages of enriched annotations (right). On the x-axis, we show each of the embedding algorithms considered (legend) applied on different types of graphlet adjacencies. Graphlet adjacency 'Union' (x-axis, far right) considers the union of the enriched genes and enriched annotations across all graphlet adjacencies, i.e., $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$. The error bars for Spring embedding indicate the standard deviation across ten runs.

Gene ontology cellular components

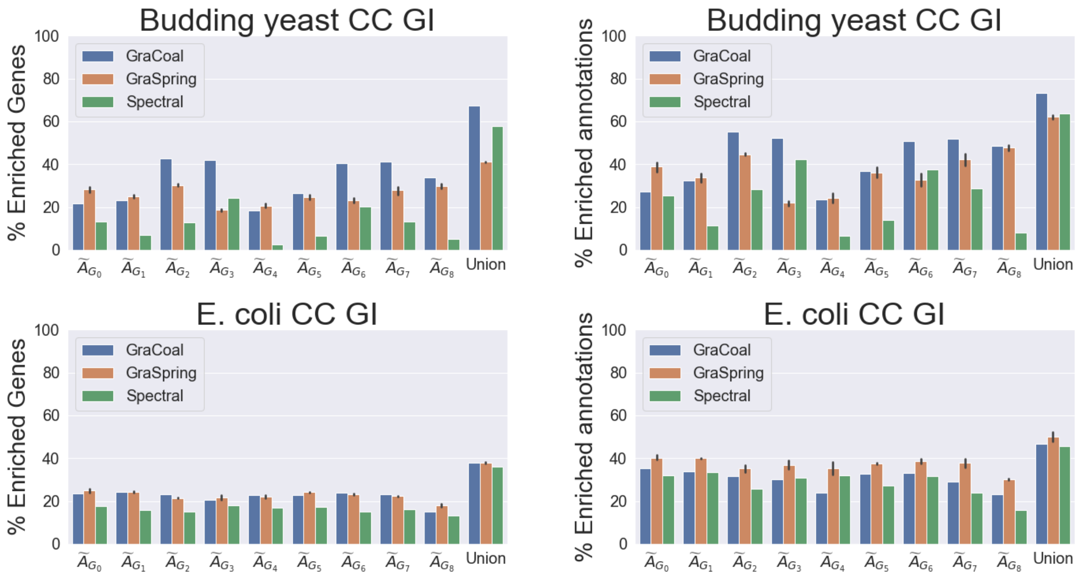


Figure A.7. SAFE GO-CC enrichment analysis for the GI networks, Part 1. On the y-axis, we show the percentages of genes that have at least one annotation enriched in their neighborhood (left) and the percentages of enriched annotations (right). On the x-axis, we show each of the embedding algorithms considered (legend) applied on different types of graphlet adjacencies. Graphlet adjacency 'Union' (x-axis, far right) considers the union of the enriched genes and enriched annotations across all graphlet adjacencies, i.e., $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$. The error bars for Spring embedding indicate the standard deviation across ten runs. From top to bottom: Budding yeast, *E. coli*, Fruit fly and Fission yeast, respectively.

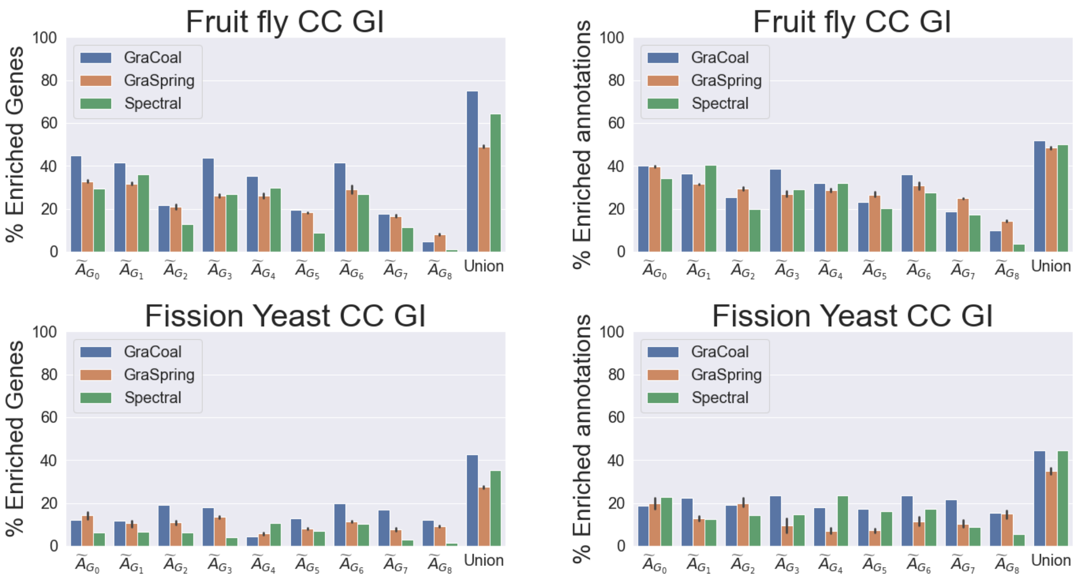


Figure A.7. SAFE GO-CC enrichment analysis for the GI networks, Part 1. On the y-axis, we show the percentages of genes that have at least one annotation enriched in their neighborhood (left) and the percentages of enriched annotations (right). On the x-axis, we show each of the embedding algorithms considered (legend) applied on different types of graphlet adjacencies. Graphlet adjacency 'Union' (x-axis, far right) considers the union of the enriched genes and enriched annotations across all graphlet adjacencies, i.e., $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$. The error bars for Spring embedding indicate the standard deviation across ten runs. From top to bottom: Budding yeast, *E. coli*, Fruit fly and Fission yeast, respectively.

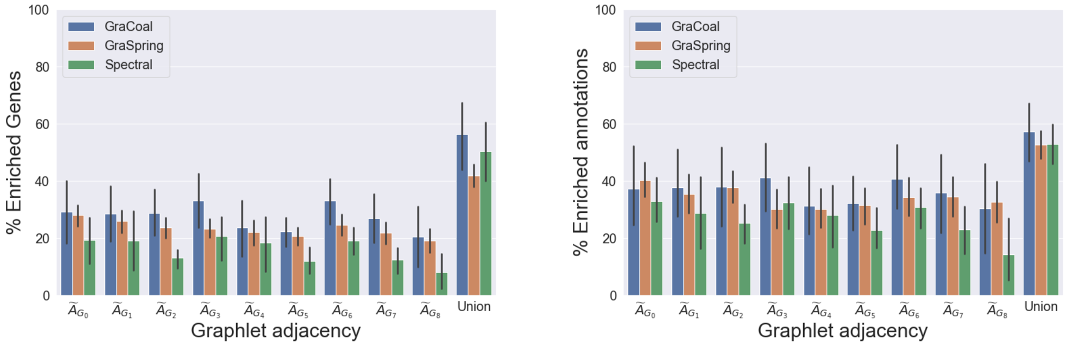


Figure A.8. SAFE GO-CC average enrichment statistics for the GI molecular networks. Average over all GI networks for the different types of underlying graphlet adjacencies, i.e., $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$. On the y-axis, we show the percentages of genes that have at least one annotation enriched in their neighborhood (left) and the percentages of enriched annotations (right). On the x-axis, we show each of the embedding algorithms considered (legend) applied on different types of graphlet adjacencies. Graphlet adjacency 'Union' (x-axis, far right) considers the union of the enriched genes and enriched annotations across all graphlet adjacencies.

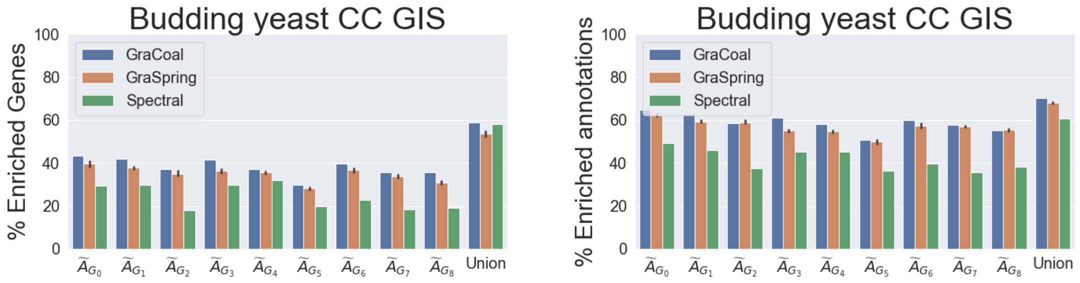


Figure A.9. SAFE GO-CC enrichment analysis for the Budding yeast GIS network. On the y-axis, we show the percentages of genes that have at least one annotation enriched in their neighborhood (left) and the percentages of enriched annotations (right). On the x-axis, we show each of the embedding algorithms considered (legend) applied on different types of graphlet adjacencies. Graphlet adjacency 'Union' (x-axis, far right) considers the union of the enriched genes and enriched annotations across all graphlet adjacencies, i.e., $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$. The error bars for Spring embedding indicate the standard deviation across ten runs.

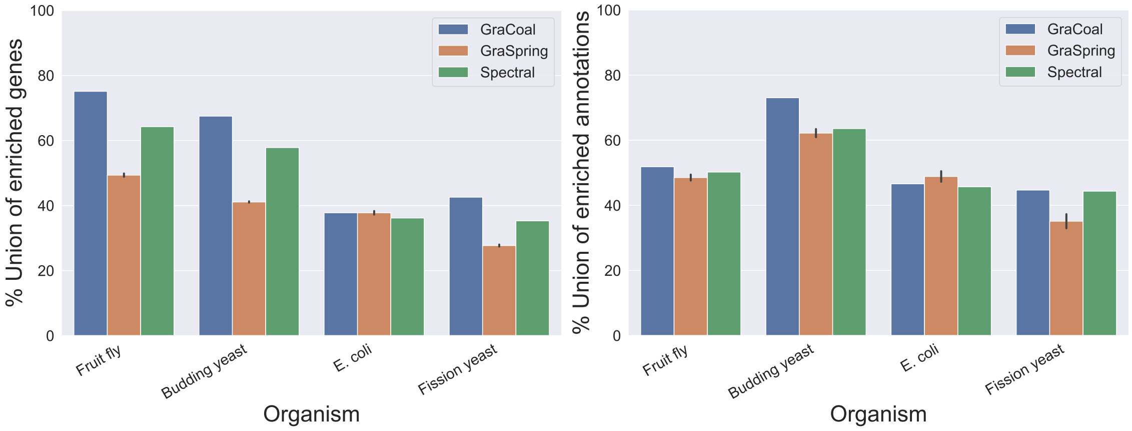


Figure A.10. SAFE GO-CC enrichment analysis for GI networks. For the GI networks of our four species (x-axis), we show the percentage of enriched genes (y-axis) and percentage of enriched annotations for each of the embedding algorithms considered (legend). The error bars in the case of GraSpring embedding indicate the standard deviation across the ten randomised runs.

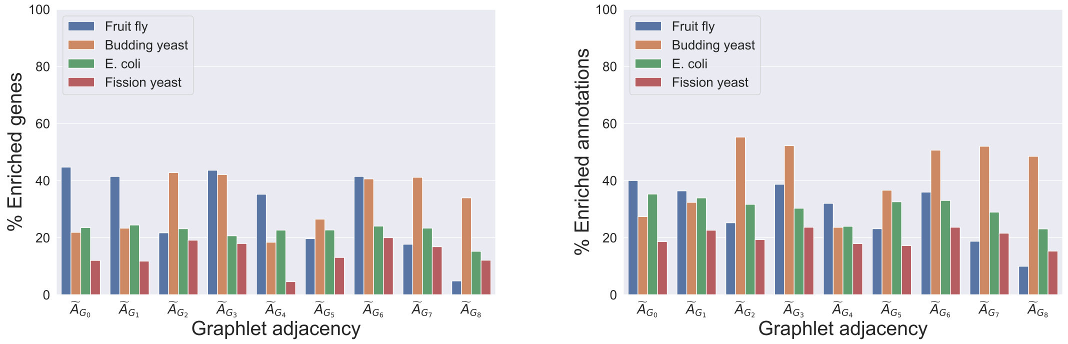


Figure A.11. SAFE GO-CC enrichment analysis comparing GraCoals in GI networks. For the GI networks of our four species (legend), we show, on the y-axis, the percentage of enriched genes (left) and the percentage of enriched annotations (right) for each of the different GraCoal embeddings (x-axis).

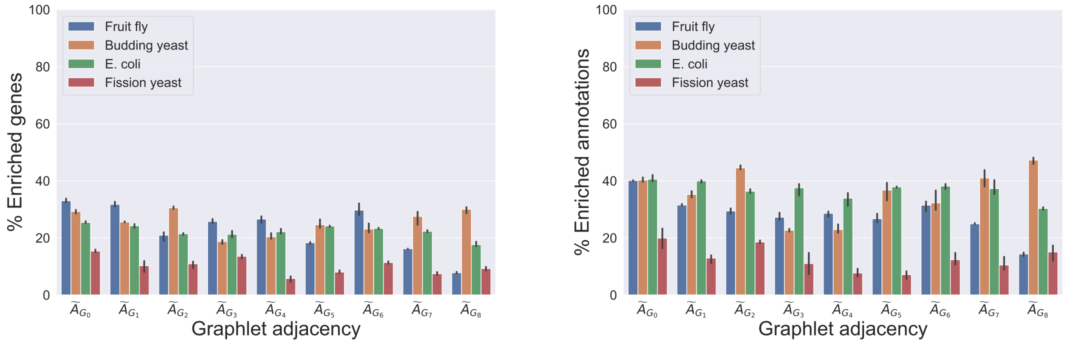


Figure A.12. SAFE GO-CC enrichment analysis comparing GraSprings in GI networks. For the GI networks of our four species (legend), we show, on the y-axis, the percentage of enriched genes (left) and the percentage of enriched annotations (right) for each of the different GraSpring embeddings (x-axis).

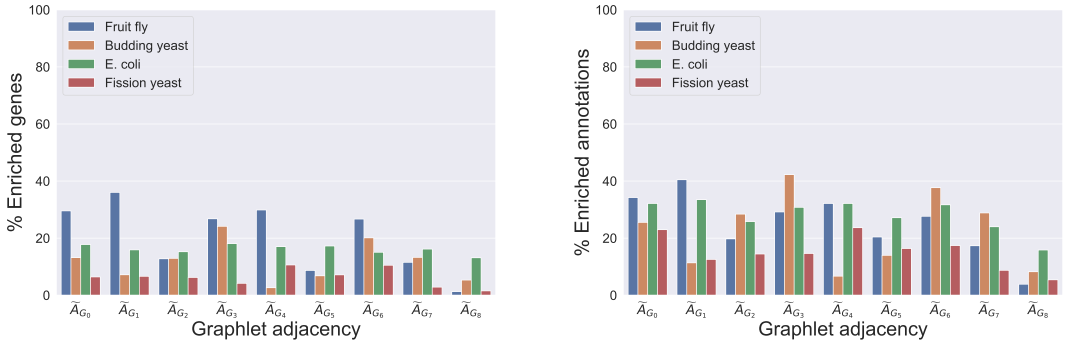


Figure A.13. SAFE GO-CC enrichment analysis comparing Spectrals in GI networks. For the GI networks of our four species (legend), we show, on the y-axis, the percentage of enriched genes (left) and the percentage of enriched annotations (right) for each of the different Spectral embeddings (x-axis).

Gene ontology molecular functions

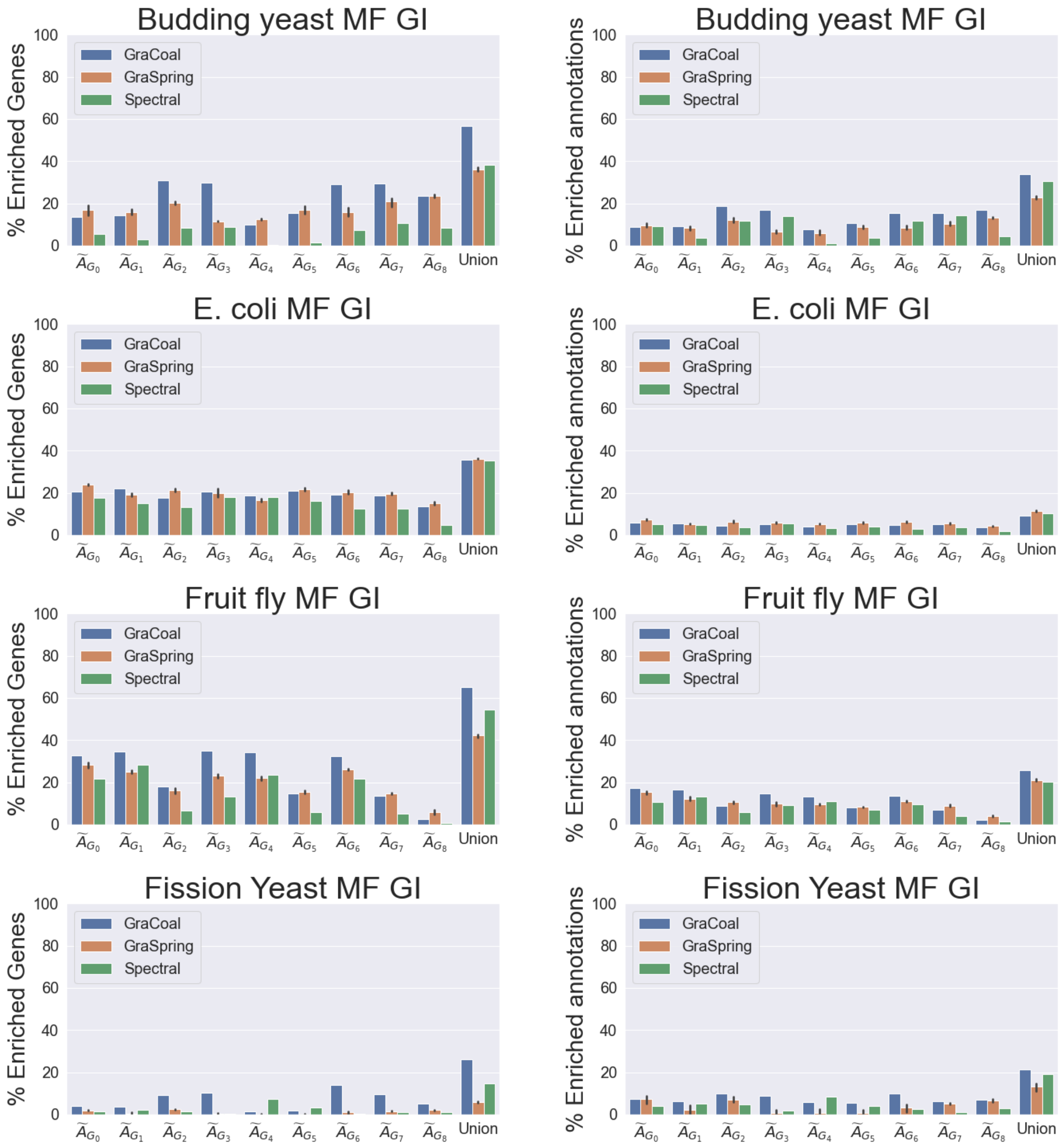


Figure A.14. SAFE GO-MF enrichment analysis for the GI networks. On the y-axis, we show the percentages of genes that have at least one annotation enriched in their neighborhood (left) and the percentages of enriched annotations (right). On the x-axis, we show each of the embedding algorithms considered (legend) applied on different types of graphlet adjacencies. Graphlet adjacency 'Union' (x-axis, far right) considers the union of the enriched genes and enriched annotations across all graphlet adjacencies, i.e., $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$. The error bars for Spring embedding indicate the standard deviation across ten runs. From top to bottom: Budding yeast, *E. coli*, Fruit fly and Fission yeast, respectively.

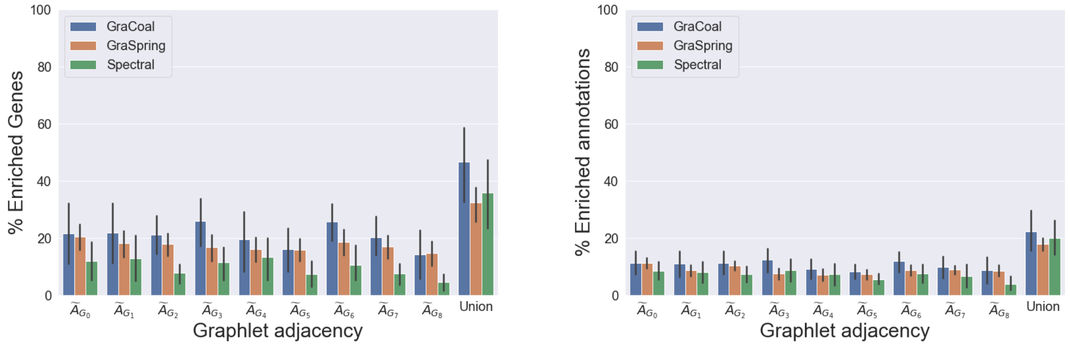


Figure A.15. SAFE GO-MF average enrichment statistics for the GI molecular networks. Average over all GI networks for the different types of underlying graphlet adjacencies, i.e., $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$. On the y-axis, we show the percentages of genes that have at least one annotation enriched in their neighborhood (left) and the percentages of enriched annotations (right). On the x-axis, we show each of the embedding algorithms considered (legend) applied on different types of graphlet adjacencies. Graphlet adjacency 'Union' (x-axis, far right) considers the union of the enriched genes and enriched annotations across all graphlet adjacencies.

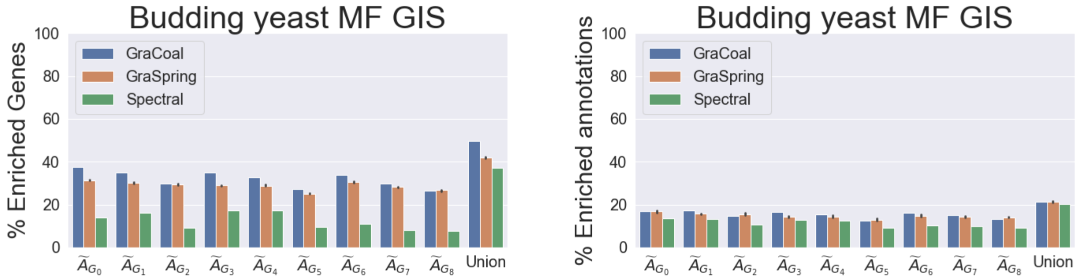


Figure A.16. SAFE GO-MF enrichment analysis for the Budding yeast GIS network. On the y-axis, we show the percentages of genes that have at least one annotation enriched in their neighborhood (left) and the percentages of enriched annotations (right). On the x-axis, we show each of the embedding algorithms considered (legend) applied on different types of graphlet adjacencies. Graphlet adjacency 'Union' (x-axis, far right) considers the union of the enriched genes and enriched annotations across all graphlet adjacencies, i.e., $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$. The error bars for Spring embedding indicate the standard deviation across ten runs.

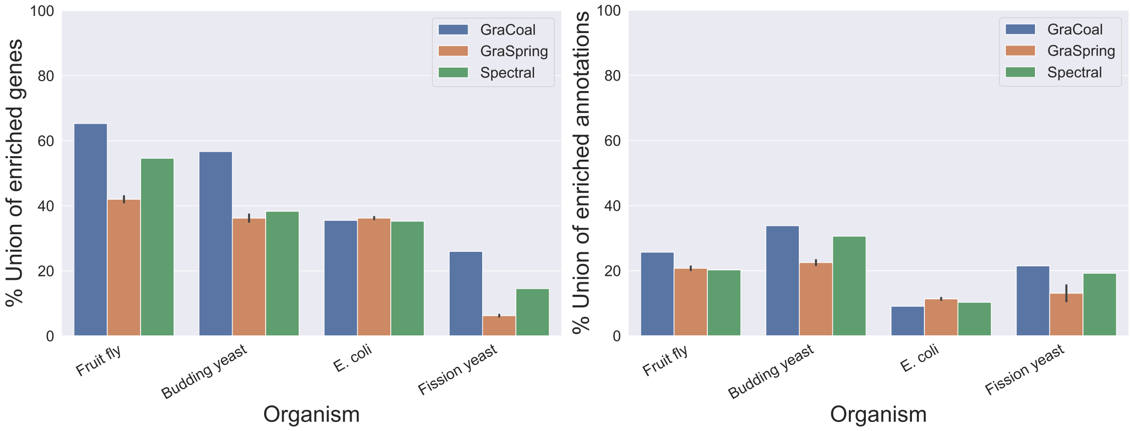


Figure A.17. SAFE GO-MF enrichment analysis for GI networks. For the GI networks of our four species (x-axis), we show the percentage of enriched genes (y-axis) and percentage of enriched annotations for each of the embedding algorithms considered (legend). The error bars in the case of GraSpring embedding indicate the standard deviation across the ten randomised runs.

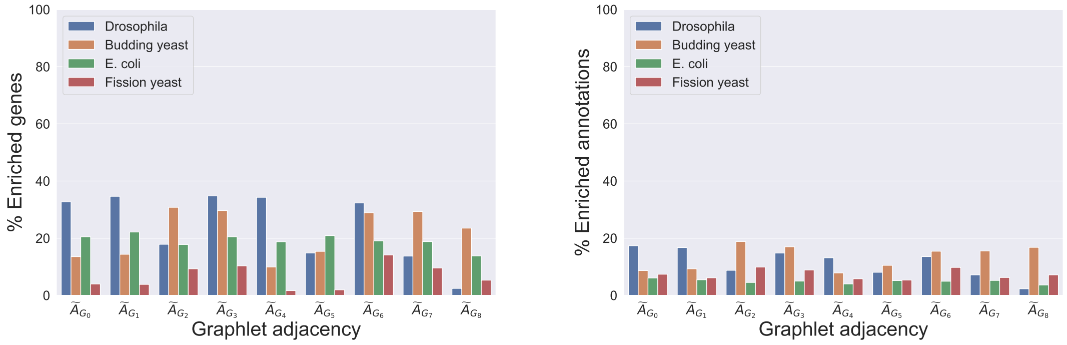


Figure A.18. SAFE GO-MF enrichment analysis comparing GraCoals in GI networks. For the GI networks of our four species (legend), we show, on the y-axis, the percentage of enriched genes (left) and the percentage of enriched annotations (right) for each of the different GraCoal embeddings (x-axis).

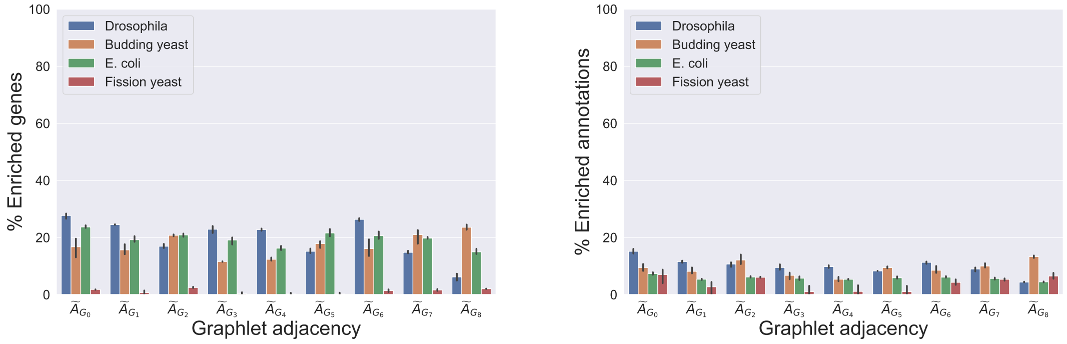


Figure A.19. SAFE GO-MF enrichment analysis comparing GraSprings in GI networks. For the GI networks of our four species (legend), we show, on the y-axis, the percentage of enriched genes (left) and the percentage of enriched annotations (right) for each of the different GraSpring embeddings (x-axis).

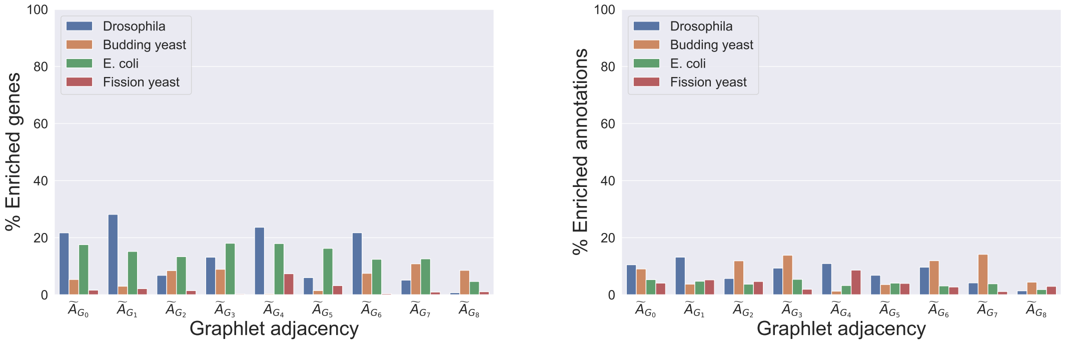


Figure A.20. SAFE GO-MF enrichment analysis comparing Spectrals in GI networks. For the GI networks of our four species (legend), we show, on the y-axis, the percentage of enriched genes (left) and the percentage of enriched annotations (right) for each of the different Spectral embeddings (x-axis).

GO-BPs enrichment summary for GI networks

GraCoals enrichment summary for *E. coli*

	Total annotations	Mean SS	Mean SS Top 10	EN	Annotation
$\widetilde{A}_{G_0}$	18	0.21 (std=0.04)	0.27 (std=0.04)	142.0	peptide metabolic process
$\widetilde{A}_{G_0}$	18	0.21 (std=0.04)	0.27 (std=0.04)	121.0	rRNA modification
$\widetilde{A}_{G_0}$	18	0.21 (std=0.04)	0.27 (std=0.04)	119.0	ribosomal large subunit assembly
$\widetilde{A}_{G_0}$	18	0.21 (std=0.04)	0.27 (std=0.04)	119.0	pseudouridine synthesis
$\widetilde{A}_{G_0}$	18	0.21 (std=0.04)	0.27 (std=0.04)	118.0	N-terminal protein amino acid modification
$\widetilde{A}_{G_0}$	18	0.21 (std=0.04)	0.27 (std=0.04)	115.0	RNA methylation
$\widetilde{A}_{G_0}$	18	0.21 (std=0.04)	0.27 (std=0.04)	114.0	metabolic process
$\widetilde{A}_{G_0}$	18	0.21 (std=0.04)	0.27 (std=0.04)	111.0	organic substance metabolic process
$\widetilde{A}_{G_0}$	18	0.21 (std=0.04)	0.27 (std=0.04)	102.0	tRNA methylation
$\widetilde{A}_{G_0}$	18	0.21 (std=0.04)	0.27 (std=0.04)	100.0	peptide catabolic process
$\widetilde{A}_{G_1}$	1	1.00 (std=nan)	1.00 (std=nan)	82.0	intracellular protein transmembrane transport
$\widetilde{A}_{G_2}$	10	0.32 (std=0.07)	0.32 (std=0.07)	141.0	negative regulation of DNA-templated DNA replication
$\widetilde{A}_{G_2}$	10	0.32 (std=0.07)	0.32 (std=0.07)	138.0	negative regulation of DNA replication
$\widetilde{A}_{G_2}$	10	0.32 (std=0.07)	0.32 (std=0.07)	138.0	negative regulation of DNA metabolic process
$\widetilde{A}_{G_2}$	10	0.32 (std=0.07)	0.32 (std=0.07)	129.0	cell communication
$\widetilde{A}_{G_2}$	10	0.32 (std=0.07)	0.32 (std=0.07)	110.0	regulation of DNA replication
$\widetilde{A}_{G_2}$	10	0.32 (std=0.07)	0.32 (std=0.07)	101.0	response to extracellular stimulus
$\widetilde{A}_{G_2}$	10	0.32 (std=0.07)	0.32 (std=0.07)	47.0	isopentenyl diphosphate metabolic process
$\widetilde{A}_{G_2}$	10	0.32 (std=0.07)	0.32 (std=0.07)	47.0	glyceraldehyde-3-phosphate metabolic process
$\widetilde{A}_{G_2}$	10	0.32 (std=0.07)	0.32 (std=0.07)	47.0	isopentenyl diphosphate biosynthetic process
$\widetilde{A}_{G_2}$	10	0.32 (std=0.07)	0.32 (std=0.07)	47.0	isopentenyl diphosphate biosynthetic process, methylerythritol 4-phosphate pathway
$\widetilde{A}_{G_3}$	7	0.24 (std=0.01)	0.24 (std=0.01)	185.0	DNA topological change
$\widetilde{A}_{G_3}$	7	0.24 (std=0.01)	0.24 (std=0.01)	102.0	viral process
$\widetilde{A}_{G_3}$	7	0.24 (std=0.01)	0.24 (std=0.01)	93.0	regulation of DNA recombination
$\widetilde{A}_{G_3}$	7	0.24 (std=0.01)	0.24 (std=0.01)	93.0	translesion synthesis
$\widetilde{A}_{G_3}$	7	0.24 (std=0.01)	0.24 (std=0.01)	49.0	division septum assembly
$\widetilde{A}_{G_3}$	7	0.24 (std=0.01)	0.24 (std=0.01)	49.0	bile acid and bile salt transport
$\widetilde{A}_{G_3}$	7	0.24 (std=0.01)	0.24 (std=0.01)	32.0	organic anion transport

Table A.3. Summary of uniquely enriched GO-BPs for Gracoal embeddings, part 1. We report, for the *E. coli* GI network, the number of uniquely enriched GO-BPs and the mean semantic similarity (SS) between the uniquely enriched annotations (GO-BPs) (columns 1 and 2). In column 3, we report the mean SS for the top ten largest enriched annotations (column 5), i.e., ranking them in descending order according to the number of neighborhoods that the annotations are enriched in (column 4).

	Total annotations	**Mean SS**	**Mean SS Top 10**	**EN**	**Annotation**
$\widetilde{A}_{G_4}$	19	0.25 (std=0.05)	0.38 (std=0.05)	159.0	enterobacterial common antigen biosynthetic process
$\widetilde{A}_{G_4}$	19	0.25 (std=0.05)	0.38 (std=0.05)	159.0	enterobacterial common antigen metabolic process
$\widetilde{A}_{G_4}$	19	0.25 (std=0.05)	0.38 (std=0.05)	150.0	glycerophospholipid biosynthetic process
$\widetilde{A}_{G_4}$	19	0.25 (std=0.05)	0.38 (std=0.05)	150.0	glycerolipid biosynthetic process
$\widetilde{A}_{G_4}$	19	0.25 (std=0.05)	0.38 (std=0.05)	136.0	glucan biosynthetic process
$\widetilde{A}_{G_4}$	19	0.25 (std=0.05)	0.38 (std=0.05)	131.0	glycogen biosynthetic process
$\widetilde{A}_{G_4}$	19	0.25 (std=0.05)	0.38 (std=0.05)	112.0	glycerolipid metabolic process
$\widetilde{A}_{G_4}$	19	0.25 (std=0.05)	0.38 (std=0.05)	112.0	glycerophospholipid metabolic process
$\widetilde{A}_{G_4}$	19	0.25 (std=0.05)	0.38 (std=0.05)	103.0	cellular macromolecule catabolic process
$\widetilde{A}_{G_4}$	19	0.25 (std=0.05)	0.38 (std=0.05)	96.0	lipid modification
$\widetilde{A}_{G_5}$	2	0.55 (std=0.00)	0.55 (std=0.00)	73.0	tRNA modification
$\widetilde{A}_{G_5}$	2	0.55 (std=0.00)	0.55 (std=0.00)	31.0	proteolysis involved in cellular protein catabolic process
$\widetilde{A}_{G_6}$	1	1.00 (std=nan)	1.00 (std=nan)	25.0	amide biosynthetic process
$\widetilde{A}_{G_7}$	3	0.64 (std=0.01)	0.64 (std=0.01)	15.0	copper ion transport
$\widetilde{A}_{G_7}$	3	0.64 (std=0.01)	0.64 (std=0.01)	15.0	copper ion transmembrane transport
$\widetilde{A}_{G_7}$	3	0.64 (std=0.01)	0.64 (std=0.01)	15.0	copper ion export
$\widetilde{A}_{G_8}$	6	0.33 (std=0.05)	0.33 (std=0.05)	52.0	dipeptide transport
$\widetilde{A}_{G_8}$	6	0.33 (std=0.05)	0.33 (std=0.05)	49.0	dipeptide transmembrane transport
$\widetilde{A}_{G_8}$	6	0.33 (std=0.05)	0.33 (std=0.05)	38.0	organophosphate ester transport
$\widetilde{A}_{G_8}$	6	0.33 (std=0.05)	0.33 (std=0.05)	37.0	heme transport
$\widetilde{A}_{G_8}$	6	0.33 (std=0.05)	0.33 (std=0.05)	37.0	aerobic electron transport chain
$\widetilde{A}_{G_8}$	6	0.33 (std=0.05)	0.33 (std=0.05)	29.0	glycerol-3-phosphate transmembrane transport

Table A.3. Summary of uniquely enriched GO-BPs for Gracoal embeddings, part 2. We report, for the *E. coli* GI network, the number of uniquely enriched GO-BPs and the mean semantic similarity (SS) between the uniquely enriched annotations (GO-BPs) (columns 1 and 2). In column 3, we report the mean SS for the top ten largest enriched annotations (column 5), i.e., ranking them in descending order according to the number of neighborhoods that the annotations are enriched in (column 4).

GraCoals enrichment summary for Fission yeast

	Total annotations	Mean SS	Mean SS Top 10	EN	Annotation
$\widetilde{A}_{G_0}$	5	0.45 (std=0.05)	0.45 (std=0.05)	167.0	regulation of biological process
$\widetilde{A}_{G_0}$	5	0.45 (std=0.05)	0.45 (std=0.05)	167.0	regulation of cell cycle switching, mitotic to meiotic cell cycle
$\widetilde{A}_{G_0}$	5	0.45 (std=0.05)	0.45 (std=0.05)	120.0	negative regulation of conjugation with cellular fusion
$\widetilde{A}_{G_0}$	5	0.45 (std=0.05)	0.45 (std=0.05)	86.0	regulation of cell cycle G1/S phase transition
$\widetilde{A}_{G_0}$	5	0.45 (std=0.05)	0.45 (std=0.05)	86.0	regulation of G1/S transition of mitotic cell cycle
$\widetilde{A}_{G_2}$	24	0.28 (std=0.04)	0.36 (std=0.05)	95.0	regulation of Ras protein signal transduction
$\widetilde{A}_{G_2}$	24	0.28 (std=0.04)	0.36 (std=0.05)	95.0	regulation of small GTPase mediated signal transduction
$\widetilde{A}_{G_2}$	24	0.28 (std=0.04)	0.36 (std=0.05)	90.0	regulation of cell wall macromolecule metabolic process
$\widetilde{A}_{G_2}$	24	0.28 (std=0.04)	0.36 (std=0.05)	90.0	regulation of polysaccharide biosynthetic process
$\widetilde{A}_{G_2}$	24	0.28 (std=0.04)	0.36 (std=0.05)	90.0	regulation of glucan biosynthetic process
$\widetilde{A}_{G_2}$	24	0.28 (std=0.04)	0.36 (std=0.05)	90.0	regulation of polysaccharide metabolic process
$\widetilde{A}_{G_2}$	24	0.28 (std=0.04)	0.36 (std=0.05)	88.0	regulation of septation initiation signaling
$\widetilde{A}_{G_2}$	24	0.28 (std=0.04)	0.36 (std=0.05)	87.0	regulation of cell wall (1->3)-beta-D-glucan biosynthetic process
$\widetilde{A}_{G_2}$	24	0.28 (std=0.04)	0.36 (std=0.05)	87.0	regulation of (1->3)-beta-D-glucan metabolic process
$\widetilde{A}_{G_2}$	24	0.28 (std=0.04)	0.36 (std=0.05)	87.0	regulation of (1->3)-beta-D-glucan biosynthetic process
$\widetilde{A}_{G_3}$	3	0.74 (std=0.05)	0.74 (std=0.05)	91.0	cellular component assembly
$\widetilde{A}_{G_3}$	3	0.74 (std=0.05)	0.74 (std=0.05)	67.0	protein-DNA complex subunit organization
$\widetilde{A}_{G_3}$	3	0.74 (std=0.05)	0.74 (std=0.05)	40.0	protein-DNA complex assembly
$\widetilde{A}_{G_4}$	7	0.61 (std=0.02)	0.61 (std=0.02)	15.0	RNA splicing, via transesterification reactions
$\widetilde{A}_{G_4}$	7	0.61 (std=0.02)	0.61 (std=0.02)	15.0	RNA splicing
$\widetilde{A}_{G_4}$	7	0.61 (std=0.02)	0.61 (std=0.02)	15.0	mRNA processing
$\widetilde{A}_{G_4}$	7	0.61 (std=0.02)	0.61 (std=0.02)	15.0	mRNA cis splicing, via spliceosome
$\widetilde{A}_{G_4}$	7	0.61 (std=0.02)	0.61 (std=0.02)	15.0	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile
$\widetilde{A}_{G_4}$	7	0.61 (std=0.02)	0.61 (std=0.02)	15.0	mRNA splicing, via spliceosome
$\widetilde{A}_{G_4}$	7	0.61 (std=0.02)	0.61 (std=0.02)	13.0	RNA processing

Table A.4. Summary of uniquely enriched GO-BPs for Gracoal embeddings, Part 1. We report, for the Fission yeast GI network, the number of uniquely enriched GO-BPs and the mean semantic similarity (SS) between the uniquely enriched annotations (GO-BPs) (columns 1 and 2). In column 3, we report the mean SS for the top ten largest enriched annotations (column 5), i.e., ranking them in descending order according to the number of neighborhoods that the annotations are enriched in (column 4).

	Total annotations	Mean SS	Mean SS Top 10	EN	Annotation
$\widetilde{A}_{G_5}$	17	0.31 (std=0.05)	0.36 (std=0.06)	237.0	ubiquitin-dependent protein catabolic process
$\widetilde{A}_{G_5}$	17	0.31 (std=0.05)	0.36 (std=0.06)	237.0	modification-dependent protein catabolic process
$\widetilde{A}_{G_5}$	17	0.31 (std=0.05)	0.36 (std=0.06)	236.0	proteolysis involved in cellular protein catabolic process
$\widetilde{A}_{G_5}$	17	0.31 (std=0.05)	0.36 (std=0.06)	235.0	modification-dependent macromolecule catabolic process
$\widetilde{A}_{G_5}$	17	0.31 (std=0.05)	0.36 (std=0.06)	227.0	protein metabolic process
$\widetilde{A}_{G_5}$	17	0.31 (std=0.05)	0.36 (std=0.06)	211.0	macromolecule catabolic process
$\widetilde{A}_{G_5}$	17	0.31 (std=0.05)	0.36 (std=0.06)	180.0	regulation of primary metabolic process
$\widetilde{A}_{G_5}$	17	0.31 (std=0.05)	0.36 (std=0.06)	169.0	regulation of macromolecule metabolic process
$\widetilde{A}_{G_5}$	17	0.31 (std=0.05)	0.36 (std=0.06)	162.0	regulation of metabolic process
$\widetilde{A}_{G_5}$	17	0.31 (std=0.05)	0.36 (std=0.06)	82.0	positive regulation of cellular component organization
$\widetilde{A}_{G_6}$	27	0.20 (std=0.04)	0.36 (std=0.04)	226.0	cell cycle DNA replication maintenance of fidelity
$\widetilde{A}_{G_6}$	27	0.20 (std=0.04)	0.36 (std=0.04)	226.0	mitotic recombination-dependent replication fork processing
$\widetilde{A}_{G_6}$	27	0.20 (std=0.04)	0.36 (std=0.04)	226.0	mitotic DNA replication maintenance of fidelity
$\widetilde{A}_{G_6}$	27	0.20 (std=0.04)	0.36 (std=0.04)	188.0	regulation of cytokinetic process
$\widetilde{A}_{G_6}$	27	0.20 (std=0.04)	0.36 (std=0.04)	152.0	UV-damage excision repair
$\widetilde{A}_{G_6}$	27	0.20 (std=0.04)	0.36 (std=0.04)	152.0	response to radiation
$\widetilde{A}_{G_6}$	27	0.20 (std=0.04)	0.36 (std=0.04)	152.0	cellular response to light stimulus
$\widetilde{A}_{G_6}$	27	0.20 (std=0.04)	0.36 (std=0.04)	152.0	cellular response to UV
$\widetilde{A}_{G_6}$	27	0.20 (std=0.04)	0.36 (std=0.04)	152.0	cellular response to radiation
$\widetilde{A}_{G_6}$	27	0.20 (std=0.04)	0.36 (std=0.04)	152.0	response to light stimulus
$\widetilde{A}_{G_7}$	10	0.28 (std=0.05)	0.28 (std=0.05)	133.0	regulation of mitotic cytokinetic process
$\widetilde{A}_{G_7}$	10	0.28 (std=0.05)	0.28 (std=0.05)	103.0	cellular glucan metabolic process
$\widetilde{A}_{G_7}$	10	0.28 (std=0.05)	0.28 (std=0.05)	103.0	glucan metabolic process
$\widetilde{A}_{G_7}$	10	0.28 (std=0.05)	0.28 (std=0.05)	103.0	glucan biosynthetic process
$\widetilde{A}_{G_7}$	10	0.28 (std=0.05)	0.28 (std=0.05)	99.0	DNA biosynthetic process
$\widetilde{A}_{G_7}$	10	0.28 (std=0.05)	0.28 (std=0.05)	85.0	gene conversion
$\widetilde{A}_{G_7}$	10	0.28 (std=0.05)	0.28 (std=0.05)	83.0	telomere organization
$\widetilde{A}_{G_7}$	10	0.28 (std=0.05)	0.28 (std=0.05)	83.0	telomere maintenance
$\widetilde{A}_{G_7}$	10	0.28 (std=0.05)	0.28 (std=0.05)	77.0	protein localization to cell periphery
$\widetilde{A}_{G_7}$	10	0.28 (std=0.05)	0.28 (std=0.05)	48.0	nucleotide-excision repair

Table A.4. Summary of uniquely enriched GO-BPs for Gracoal embeddings, Part 2. We report, for the Fission yeast GI network, the number of uniquely enriched GO-BPs and the mean semantic similarity (SS) between the uniquely enriched annotations (GO-BPs) (columns 1 and 2). In column 3, we report the mean SS for the top ten largest enriched annotations (column 5), i.e., ranking them in descending order according to the number of neighborhoods that the annotations are enriched in (column 4).

	Total annotations	Mean SS	Mean SS Top 10	EN	Annotation
$\widetilde{A}_{G_8}$	40	0.20 (std=0.05)	0.43 (std=0.03)	90.0	regulation of reproductive process
$\widetilde{A}_{G_8}$	40	0.20 (std=0.05)	0.43 (std=0.03)	79.0	positive regulation of protein catabolic process
$\widetilde{A}_{G_8}$	40	0.20 (std=0.05)	0.43 (std=0.03)	79.0	positive regulation of cellular protein catabolic process
$\widetilde{A}_{G_8}$	40	0.20 (std=0.05)	0.43 (std=0.03)	79.0	regulation of protein catabolic process
$\widetilde{A}_{G_8}$	40	0.20 (std=0.05)	0.43 (std=0.03)	79.0	regulation of cellular protein catabolic process
$\widetilde{A}_{G_8}$	40	0.20 (std=0.05)	0.43 (std=0.03)	73.0	positive regulation of mitotic cell cycle phase transition
$\widetilde{A}_{G_8}$	40	0.20 (std=0.05)	0.43 (std=0.03)	72.0	regulation of proteasomal protein catabolic process
$\widetilde{A}_{G_8}$	40	0.20 (std=0.05)	0.43 (std=0.03)	72.0	regulation of ubiquitin-dependent protein catabolic process
$\widetilde{A}_{G_8}$	40	0.20 (std=0.05)	0.43 (std=0.03)	72.0	positive regulation of ubiquitin-dependent protein catabolic process
$\widetilde{A}_{G_8}$	40	0.20 (std=0.05)	0.43 (std=0.03)	72.0	positive regulation of proteasomal protein catabolic process

Table A.4. Summary of uniquely enriched GO-BPs for Gracoal embeddings, Part 3. We report, for the Fission yeast GI network, the number of uniquely enriched GO-BPs and the mean semantic similarity (SS) between the uniquely enriched annotations (GO-BPs) (columns 1 and 2). In column 3, we report the mean SS for the top ten largest enriched annotations (column 5), i.e., ranking them in descending order according to the number of neighborhoods that the annotations are enriched in (column 4).

GraCoals enrichment summary for Fruit fly

	Total annotations	Mean SS	Mean SS Top 10	EN	Annotation
$\widetilde{A}_{G_0}$	67	0.14 (std=0.02)	0.29 (std=0.03)	555.0	cellular component organization
$\widetilde{A}_{G_0}$	67	0.14 (std=0.02)	0.29 (std=0.03)	554.0	cellular component organization or biogenesis
$\widetilde{A}_{G_0}$	67	0.14 (std=0.02)	0.29 (std=0.03)	403.0	cell division
$\widetilde{A}_{G_0}$	67	0.14 (std=0.02)	0.29 (std=0.03)	327.0	regulation of metabolic process
$\widetilde{A}_{G_0}$	67	0.14 (std=0.02)	0.29 (std=0.03)	303.0	organelle localization
$\widetilde{A}_{G_0}$	67	0.14 (std=0.02)	0.29 (std=0.03)	276.0	response to radiation
$\widetilde{A}_{G_0}$	67	0.14 (std=0.02)	0.29 (std=0.03)	247.0	detection of stimulus involved in sensory perception
$\widetilde{A}_{G_0}$	67	0.14 (std=0.02)	0.29 (std=0.03)	241.0	adult behavior
$\widetilde{A}_{G_0}$	67	0.14 (std=0.02)	0.29 (std=0.03)	234.0	regulation of membrane potential
$\widetilde{A}_{G_0}$	67	0.14 (std=0.02)	0.29 (std=0.03)	222.0	calcium ion transport
$\widetilde{A}_{G_1}$	59	0.14 (std=0.03)	0.27 (std=0.04)	492.0	response to stimulus
$\widetilde{A}_{G_1}$	59	0.14 (std=0.03)	0.27 (std=0.04)	347.0	system process
$\widetilde{A}_{G_1}$	59	0.14 (std=0.03)	0.27 (std=0.04)	341.0	gland development
$\widetilde{A}_{G_1}$	59	0.14 (std=0.03)	0.27 (std=0.04)	315.0	behavior
$\widetilde{A}_{G_1}$	59	0.14 (std=0.03)	0.27 (std=0.04)	311.0	animal organ formation
$\widetilde{A}_{G_1}$	59	0.14 (std=0.03)	0.27 (std=0.04)	311.0	heart formation
$\widetilde{A}_{G_1}$	59	0.14 (std=0.03)	0.27 (std=0.04)	245.0	wing disc anterior/posterior pattern formation
$\widetilde{A}_{G_1}$	59	0.14 (std=0.03)	0.27 (std=0.04)	236.0	sensory perception of smell
$\widetilde{A}_{G_1}$	59	0.14 (std=0.03)	0.27 (std=0.04)	230.0	wing disc development
$\widetilde{A}_{G_1}$	59	0.14 (std=0.03)	0.27 (std=0.04)	229.0	neuroblast fate determination
$\widetilde{A}_{G_2}$	10	0.22 (std=0.04)	0.22 (std=0.04)	162.0	morphogenesis of a polarized epithelium
$\widetilde{A}_{G_2}$	10	0.22 (std=0.04)	0.22 (std=0.04)	140.0	cellular response to stimulus
$\widetilde{A}_{G_2}$	10	0.22 (std=0.04)	0.22 (std=0.04)	85.0	establishment of proximal/distal cell polarity
$\widetilde{A}_{G_2}$	10	0.22 (std=0.04)	0.22 (std=0.04)	85.0	imaginal disc-derived wing hair site selection
$\widetilde{A}_{G_2}$	10	0.22 (std=0.04)	0.22 (std=0.04)	72.0	negative regulation of cellular response to growth factor stimulus
$\widetilde{A}_{G_2}$	10	0.22 (std=0.04)	0.22 (std=0.04)	72.0	asymmetric protein localization involved in cell fate determination
$\widetilde{A}_{G_2}$	10	0.22 (std=0.04)	0.22 (std=0.04)	60.0	cell-cell junction organization
$\widetilde{A}_{G_2}$	10	0.22 (std=0.04)	0.22 (std=0.04)	45.0	positive regulation of protein kinase B signaling
$\widetilde{A}_{G_2}$	10	0.22 (std=0.04)	0.22 (std=0.04)	31.0	cellular homeostasis
$\widetilde{A}_{G_2}$	10	0.22 (std=0.04)	0.22 (std=0.04)	16.0	piRNA biosynthetic process

Table A.5. Summary of uniquely enriched GO-BPs for Gracoal embeddings, Part 1. We report, for the Fruit Fruit fly GI network, the number of uniquely enriched GO-BPs and the mean semantic similarity (SS) between the uniquely enriched annotations (GO-BPs) (columns 1 and 2). In column 3, we report the mean SS for the top ten largest enriched annotations (column 5), i.e., ranking them in descending order according to the number of neighborhoods that the annotations are enriched in (column 4).

	Total annotations	**Mean SS**	**Mean SS Top 10**	**EN**	**Annotation**
$\widetilde{A}_{G_3}$	40	0.15 (std=0.03)	0.25 (std=0.04)	390.0	regulation of trans-synaptic signaling
$\widetilde{A}_{G_3}$	40	0.15 (std=0.03)	0.25 (std=0.04)	390.0	modulation of chemical synaptic transmission
$\widetilde{A}_{G_3}$	40	0.15 (std=0.03)	0.25 (std=0.04)	325.0	macromolecule modification
$\widetilde{A}_{G_3}$	40	0.15 (std=0.03)	0.25 (std=0.04)	229.0	regulation of actin filament bundle assembly
$\widetilde{A}_{G_3}$	40	0.15 (std=0.03)	0.25 (std=0.04)	192.0	organic substance metabolic process
$\widetilde{A}_{G_3}$	40	0.15 (std=0.03)	0.25 (std=0.04)	188.0	gonad development
$\widetilde{A}_{G_3}$	40	0.15 (std=0.03)	0.25 (std=0.04)	174.0	regulation of circadian sleep/wake cycle, sleep
$\widetilde{A}_{G_3}$	40	0.15 (std=0.03)	0.25 (std=0.04)	173.0	larval midgut cell programmed cell death
$\widetilde{A}_{G_3}$	40	0.15 (std=0.03)	0.25 (std=0.04)	173.0	regulation of circadian sleep/wake cycle
$\widetilde{A}_{G_3}$	40	0.15 (std=0.03)	0.25 (std=0.04)	166.0	synapse assembly
$\widetilde{A}_{G_4}$	30	0.18 (std=0.05)	0.31 (std=0.03)	363.0	cell fate determination
$\widetilde{A}_{G_4}$	30	0.18 (std=0.05)	0.31 (std=0.03)	263.0	heart development
$\widetilde{A}_{G_4}$	30	0.18 (std=0.05)	0.31 (std=0.03)	245.0	pericardial nephrocyte differentiation
$\widetilde{A}_{G_4}$	30	0.18 (std=0.05)	0.31 (std=0.03)	233.0	response to mechanical stimulus
$\widetilde{A}_{G_4}$	30	0.18 (std=0.05)	0.31 (std=0.03)	193.0	regulation of multi-organism process
$\widetilde{A}_{G_4}$	30	0.18 (std=0.05)	0.31 (std=0.03)	188.0	neuronal stem cell population maintenance
$\widetilde{A}_{G_4}$	30	0.18 (std=0.05)	0.31 (std=0.03)	168.0	defense response
$\widetilde{A}_{G_4}$	30	0.18 (std=0.05)	0.31 (std=0.03)	168.0	response to biotic stimulus
$\widetilde{A}_{G_4}$	30	0.18 (std=0.05)	0.31 (std=0.03)	168.0	response to external biotic stimulus
$\widetilde{A}_{G_4}$	30	0.18 (std=0.05)	0.31 (std=0.03)	164.0	defense response to other organism
$\widetilde{A}_{G_5}$	47	0.16 (std=0.04)	0.27 (std=0.06)	247.0	cell cycle comprising mitosis without cytokinesis
$\widetilde{A}_{G_5}$	47	0.16 (std=0.04)	0.27 (std=0.06)	247.0	syncytial blastoderm mitotic cell cycle
$\widetilde{A}_{G_5}$	47	0.16 (std=0.04)	0.27 (std=0.06)	233.0	mitotic cell cycle, embryonic
$\widetilde{A}_{G_5}$	47	0.16 (std=0.04)	0.27 (std=0.06)	232.0	anterior/posterior axis specification
$\widetilde{A}_{G_5}$	47	0.16 (std=0.04)	0.27 (std=0.06)	193.0	regulation of biological quality
$\widetilde{A}_{G_5}$	47	0.16 (std=0.04)	0.27 (std=0.06)	191.0	DNA conformation change
$\widetilde{A}_{G_5}$	47	0.16 (std=0.04)	0.27 (std=0.06)	190.0	regulation of mitotic cell cycle phase transition
$\widetilde{A}_{G_5}$	47	0.16 (std=0.04)	0.27 (std=0.06)	189.0	regulation of cell cycle phase transition
$\widetilde{A}_{G_5}$	47	0.16 (std=0.04)	0.27 (std=0.06)	177.0	meiotic chromosome segregation
$\widetilde{A}_{G_5}$	47	0.16 (std=0.04)	0.27 (std=0.06)	167.0	organelle Fission

Table A.5. Summary of uniquely enriched GO-BPs for Gracoal embeddings, Part 2. We report, for the Fruit Fruit fly GI network, the number of uniquely enriched GO-BPs and the mean semantic similarity (SS) between the uniquely enriched annotations (GO-BPs) (columns 1 and 2). In column 3, we report the mean SS for the top ten largest enriched annotations (column 5), i.e., ranking them in descending order according to the number of neighborhoods that the annotations are enriched in (column 4).

	Total annotations	Mean SS	Mean SS Top 10	EN	Annotation
$\widetilde{A}_{G_6}$	40	0.15 (std=0.04)	0.26 (std=0.05)	263.0	regulation of chromatin organization
$\widetilde{A}_{G_6}$	40	0.15 (std=0.04)	0.26 (std=0.05)	247.0	histone modification
$\widetilde{A}_{G_6}$	40	0.15 (std=0.04)	0.26 (std=0.05)	247.0	covalent chromatin modification
$\widetilde{A}_{G_6}$	40	0.15 (std=0.04)	0.26 (std=0.05)	208.0	appendage segmentation
$\widetilde{A}_{G_6}$	40	0.15 (std=0.04)	0.26 (std=0.05)	208.0	imaginal disc-derived leg segmentation
$\widetilde{A}_{G_6}$	40	0.15 (std=0.04)	0.26 (std=0.05)	195.0	peptidyl-lysine modification
$\widetilde{A}_{G_6}$	40	0.15 (std=0.04)	0.26 (std=0.05)	162.0	attachment of spindle microtubules to kinetochore
$\widetilde{A}_{G_6}$	40	0.15 (std=0.04)	0.26 (std=0.05)	152.0	protein metabolic process
$\widetilde{A}_{G_6}$	40	0.15 (std=0.04)	0.26 (std=0.05)	151.0	embryonic anterior midgut (ectodermal) morphogenesis
$\widetilde{A}_{G_6}$	40	0.15 (std=0.04)	0.26 (std=0.05)	146.0	multi-organism metabolic process
$\widetilde{A}_{G_7}$	15	0.22 (std=0.04)	0.24 (std=0.05)	223.0	regulation of cell projection organization
$\widetilde{A}_{G_7}$	15	0.22 (std=0.04)	0.24 (std=0.05)	223.0	regulation of plasma membrane bounded cell projection organization
$\widetilde{A}_{G_7}$	15	0.22 (std=0.04)	0.24 (std=0.05)	151.0	cellular component assembly
$\widetilde{A}_{G_7}$	15	0.22 (std=0.04)	0.24 (std=0.05)	132.0	transport
$\widetilde{A}_{G_7}$	15	0.22 (std=0.04)	0.24 (std=0.05)	97.0	Rho protein signal transduction
$\widetilde{A}_{G_7}$	15	0.22 (std=0.04)	0.24 (std=0.05)	92.0	determination of adult lifespan
$\widetilde{A}_{G_7}$	15	0.22 (std=0.04)	0.24 (std=0.05)	85.0	imaginal disc-derived appendage development
$\widetilde{A}_{G_7}$	15	0.22 (std=0.04)	0.24 (std=0.05)	79.0	appendage development
$\widetilde{A}_{G_7}$	15	0.22 (std=0.04)	0.24 (std=0.05)	72.0	positive regulation of transmembrane receptor protein serine/threonine kinase signaling pathway
$\widetilde{A}_{G_7}$	15	0.22 (std=0.04)	0.24 (std=0.05)	72.0	negative regulation of cell cycle G1/S phase transition
$\widetilde{A}_{G_8}$	5	0.42 (std=0.06)	0.42 (std=0.06)	45.0	regulation of lipid storage
$\widetilde{A}_{G_8}$	5	0.42 (std=0.06)	0.42 (std=0.06)	36.0	positive regulation of immune system process
$\widetilde{A}_{G_8}$	5	0.42 (std=0.06)	0.42 (std=0.06)	27.0	regulation of immune response
$\widetilde{A}_{G_8}$	5	0.42 (std=0.06)	0.42 (std=0.06)	24.0	negative regulation of cell cycle phase transition
$\widetilde{A}_{G_8}$	5	0.42 (std=0.06)	0.42 (std=0.06)	24.0	negative regulation of mitotic cell cycle phase transition

Table A.5. Summary of uniquely enriched GO-BPs for Gracoal embeddings, Part 3. We report, for the Fruit Fruit fly GI network, the number of uniquely enriched GO-BPs and the mean semantic similarity (SS) between the uniquely enriched annotations (GO-BPs) (columns 1 and 2). In column 3, we report the mean SS for the top ten largest enriched annotations (column 5), i.e., ranking them in descending order according to the number of neighborhoods that the annotations are enriched in (column 4).

GraSprings enrichment summary for *E. coli*

	Total annotations	Mean SS	Mean SS Top 10	EN	Annotation
$\widetilde{A}_{G_0}$	8	0.32 (std=0.06)	0.32 (std=0.06)	88.0	regulation of anatomical structure morphogenesis
$\widetilde{A}_{G_0}$	8	0.32 (std=0.06)	0.32 (std=0.06)	88.0	regulation of cell morphogenesis
$\widetilde{A}_{G_0}$	8	0.32 (std=0.06)	0.32 (std=0.06)	88.0	regulation of developmental process
$\widetilde{A}_{G_0}$	8	0.32 (std=0.06)	0.32 (std=0.06)	81.0	heme transport
$\widetilde{A}_{G_0}$	8	0.32 (std=0.06)	0.32 (std=0.06)	46.0	regulation of cell shape
$\widetilde{A}_{G_0}$	8	0.32 (std=0.06)	0.32 (std=0.06)	42.0	macromolecule modification
$\widetilde{A}_{G_0}$	8	0.32 (std=0.06)	0.32 (std=0.06)	42.0	division septum assembly
$\widetilde{A}_{G_0}$	8	0.32 (std=0.06)	0.32 (std=0.06)	28.0	tRNA modification
$\widetilde{A}_{G_1}$	4	0.44 (std=0.04)	0.44 (std=0.04)	44.0	SRP-dependent cotranslational protein targeting to membrane, translocation
$\widetilde{A}_{G_1}$	4	0.44 (std=0.04)	0.44 (std=0.04)	44.0	protein insertion into membrane from inner side
$\widetilde{A}_{G_1}$	4	0.44 (std=0.04)	0.44 (std=0.04)	37.0	protein insertion into membrane
$\widetilde{A}_{G_1}$	4	0.44 (std=0.04)	0.44 (std=0.04)	10.0	organonitrogen compound biosynthetic process
$\widetilde{A}_{G_2}$	10	0.25 (std=0.06)	0.25 (std=0.06)	204.0	macromolecule biosynthetic process
$\widetilde{A}_{G_2}$	10	0.25 (std=0.06)	0.25 (std=0.06)	99.0	DNA-templated DNA replication
$\widetilde{A}_{G_2}$	10	0.25 (std=0.06)	0.25 (std=0.06)	76.0	organelle organization
$\widetilde{A}_{G_2}$	10	0.25 (std=0.06)	0.25 (std=0.06)	35.0	response to cold
$\widetilde{A}_{G_2}$	10	0.25 (std=0.06)	0.25 (std=0.06)	31.0	cellular component disassembly
$\widetilde{A}_{G_2}$	10	0.25 (std=0.06)	0.25 (std=0.06)	27.0	tRNA 3'-end processing
$\widetilde{A}_{G_2}$	10	0.25 (std=0.06)	0.25 (std=0.06)	27.0	RNA 3'-end processing
$\widetilde{A}_{G_2}$	10	0.25 (std=0.06)	0.25 (std=0.06)	27.0	ncRNA 3'-end processing
$\widetilde{A}_{G_2}$	10	0.25 (std=0.06)	0.25 (std=0.06)	27.0	translational termination
$\widetilde{A}_{G_2}$	10	0.25 (std=0.06)	0.25 (std=0.06)	16.0	glutamate biosynthetic process
$\widetilde{A}_{G_3}$	8	0.41 (std=0.06)	0.41 (std=0.06)	146.0	intracellular protein transport
$\widetilde{A}_{G_3}$	8	0.41 (std=0.06)	0.41 (std=0.06)	146.0	intracellular transport
$\widetilde{A}_{G_3}$	8	0.41 (std=0.06)	0.41 (std=0.06)	95.0	glycerolipid biosynthetic process
$\widetilde{A}_{G_3}$	8	0.41 (std=0.06)	0.41 (std=0.06)	95.0	glycerophospholipid biosynthetic process
$\widetilde{A}_{G_3}$	8	0.41 (std=0.06)	0.41 (std=0.06)	93.0	glycerolipid metabolic process
$\widetilde{A}_{G_3}$	8	0.41 (std=0.06)	0.41 (std=0.06)	93.0	glycerophospholipid metabolic process
$\widetilde{A}_{G_3}$	8	0.41 (std=0.06)	0.41 (std=0.06)	64.0	intracellular protein transmembrane transport
$\widetilde{A}_{G_3}$	8	0.41 (std=0.06)	0.41 (std=0.06)	15.0	cellular component assembly

Table A.6. Summary of uniquely enriched GO-BPs for GraSpring embeddings, Part 1. We report, for the *E. coli* GI network, the number of uniquely enriched GO-BPs and the mean semantic similarity (SS) between the uniquely enriched annotations (GO-BPs) (columns 1 and 2). In column 3, we report the mean SS for the top ten largest enriched annotations (column 5), i.e., ranking them in descending order according to the number of neighborhoods that the annotations are enriched in (column 4).

	Total annotations	Mean SS	Mean SS Top 10	EN	Annotation
$\widetilde{A}_{G_4}$	6	0.33 (std=0.06)	0.33 (std=0.06)	95.0	response to stress
$\widetilde{A}_{G_4}$	6	0.33 (std=0.06)	0.33 (std=0.06)	83.0	response to antibiotic
$\widetilde{A}_{G_4}$	6	0.33 (std=0.06)	0.33 (std=0.06)	75.0	response to ionizing radiation
$\widetilde{A}_{G_4}$	6	0.33 (std=0.06)	0.33 (std=0.06)	49.0	organic hydroxy compound transport
$\widetilde{A}_{G_4}$	6	0.33 (std=0.06)	0.33 (std=0.06)	47.0	lipopolysaccharide metabolic process
$\widetilde{A}_{G_4}$	6	0.33 (std=0.06)	0.33 (std=0.06)	47.0	lipopolysaccharide biosynthetic process
$\widetilde{A}_{G_6}$	3	0.44 (std=0.01)	0.44 (std=0.01)	123.0	regulation of protein stability
$\widetilde{A}_{G_6}$	3	0.44 (std=0.01)	0.44 (std=0.01)	74.0	chaperone-mediated protein folding
$\widetilde{A}_{G_6}$	3	0.44 (std=0.01)	0.44 (std=0.01)	41.0	protein-containing complex assembly
$\widetilde{A}_{G_7}$	10	0.28 (std=0.05)	0.28 (std=0.05)	123.0	ion transport
$\widetilde{A}_{G_7}$	10	0.28 (std=0.05)	0.28 (std=0.05)	100.0	bacteriocin transport
$\widetilde{A}_{G_7}$	10	0.28 (std=0.05)	0.28 (std=0.05)	93.0	cation transport
$\widetilde{A}_{G_7}$	10	0.28 (std=0.05)	0.28 (std=0.05)	79.0	septum digestion after cytokinesis
$\widetilde{A}_{G_7}$	10	0.28 (std=0.05)	0.28 (std=0.05)	71.0	monocarboxylic acid metabolic process
$\widetilde{A}_{G_7}$	10	0.28 (std=0.05)	0.28 (std=0.05)	58.0	amide biosynthetic process
$\widetilde{A}_{G_7}$	10	0.28 (std=0.05)	0.28 (std=0.05)	52.0	isoprenoid biosynthetic process
$\widetilde{A}_{G_7}$	10	0.28 (std=0.05)	0.28 (std=0.05)	52.0	isoprenoid metabolic process
$\widetilde{A}_{G_7}$	10	0.28 (std=0.05)	0.28 (std=0.05)	49.0	small molecule biosynthetic process
$\widetilde{A}_{G_7}$	10	0.28 (std=0.05)	0.28 (std=0.05)	41.0	cellular amide metabolic process
$\widetilde{A}_{G_8}$	9	0.54 (std=0.03)	0.54 (std=0.03)	125.0	lipooligosaccharide metabolic process
$\widetilde{A}_{G_8}$	9	0.54 (std=0.03)	0.54 (std=0.03)	125.0	lipooligosaccharide biosynthetic process
$\widetilde{A}_{G_8}$	9	0.54 (std=0.03)	0.54 (std=0.03)	125.0	membrane lipid biosynthetic process
$\widetilde{A}_{G_8}$	9	0.54 (std=0.03)	0.54 (std=0.03)	125.0	glycolipid biosynthetic process
$\widetilde{A}_{G_8}$	9	0.54 (std=0.03)	0.54 (std=0.03)	125.0	glycolipid metabolic process
$\widetilde{A}_{G_8}$	9	0.54 (std=0.03)	0.54 (std=0.03)	125.0	membrane lipid metabolic process
$\widetilde{A}_{G_8}$	9	0.54 (std=0.03)	0.54 (std=0.03)	102.0	oligosaccharide metabolic process
$\widetilde{A}_{G_8}$	9	0.54 (std=0.03)	0.54 (std=0.03)	64.0	lipid A metabolic process
$\widetilde{A}_{G_8}$	9	0.54 (std=0.03)	0.54 (std=0.03)	64.0	lipid A biosynthetic process

Table A.6. Summary of uniquely enriched GO-BPs for GraSpring embeddings, Part 2. We report, for the *E. coli* GI network, the number of uniquely enriched GO-BPs and the mean semantic similarity (SS) between the uniquely enriched annotations (GO-BPs) (columns 1 and 2). In column 3, we report the mean SS for the top ten largest enriched annotations (column 5), i.e., ranking them in descending order according to the number of neighborhoods that the annotations are enriched in (column 4).

GraSprings enrichment summary for Fission yeast

	Total annotations	Mean SS	Mean SS Top 10	EN	Annotation
$\widetilde{A}_{G_1}$	7	0.63 (std=0.03)	0.63 (std=0.03)	23.0	positive regulation of mitotic cell cycle
$\widetilde{A}_{G_1}$	7	0.63 (std=0.03)	0.63 (std=0.03)	22.0	positive regulation of mitotic cell cycle phase transition
$\widetilde{A}_{G_1}$	7	0.63 (std=0.03)	0.63 (std=0.03)	22.0	positive regulation of cell cycle
$\widetilde{A}_{G_1}$	7	0.63 (std=0.03)	0.63 (std=0.03)	22.0	positive regulation of cell cycle process
$\widetilde{A}_{G_1}$	7	0.63 (std=0.03)	0.63 (std=0.03)	21.0	positive regulation of cell cycle phase transition
$\widetilde{A}_{G_1}$	7	0.63 (std=0.03)	0.63 (std=0.03)	10.0	positive regulation of biological process
$\widetilde{A}_{G_1}$	7	0.63 (std=0.03)	0.63 (std=0.03)	10.0	positive regulation of cellular process
$\widetilde{A}_{G_2}$	8	0.44 (std=0.05)	0.44 (std=0.05)	15.0	nucleobase-containing compound biosynthetic process
$\widetilde{A}_{G_2}$	8	0.44 (std=0.05)	0.44 (std=0.05)	14.0	aromatic compound biosynthetic process
$\widetilde{A}_{G_2}$	8	0.44 (std=0.05)	0.44 (std=0.05)	13.0	heterocycle biosynthetic process
$\widetilde{A}_{G_2}$	8	0.44 (std=0.05)	0.44 (std=0.05)	13.0	DNA biosynthetic process
$\widetilde{A}_{G_2}$	8	0.44 (std=0.05)	0.44 (std=0.05)	10.0	cellular nitrogen compound biosynthetic process
$\widetilde{A}_{G_2}$	8	0.44 (std=0.05)	0.44 (std=0.05)	10.0	organic cyclic compound biosynthetic process
$\widetilde{A}_{G_2}$	8	0.44 (std=0.05)	0.44 (std=0.05)	10.0	regulation of cytosolic calcium ion concentration
$\widetilde{A}_{G_2}$	8	0.44 (std=0.05)	0.44 (std=0.05)	10.0	positive regulation of cytosolic calcium ion concentration
$\widetilde{A}_{G_3}$	6	0.37 (std=0.04)	0.37 (std=0.04)	43.0	chromosome organization
$\widetilde{A}_{G_3}$	6	0.37 (std=0.04)	0.37 (std=0.04)	36.0	mitotic recombination
$\widetilde{A}_{G_3}$	6	0.37 (std=0.04)	0.37 (std=0.04)	30.0	gene conversion
$\widetilde{A}_{G_3}$	6	0.37 (std=0.04)	0.37 (std=0.04)	26.0	cellular developmental process
$\widetilde{A}_{G_3}$	6	0.37 (std=0.04)	0.37 (std=0.04)	24.0	reciprocal homologous recombination
$\widetilde{A}_{G_3}$	6	0.37 (std=0.04)	0.37 (std=0.04)	24.0	reciprocal meiotic recombination

Table A.7. Summary of uniquely enriched GO-BPs for GraSpring embeddings, Part 1. We report, for the Fission yeast GI network, the number of uniquely enriched GO-BPs and the mean semantic similarity (SS) between the uniquely enriched annotations (GO-BPs) (columns 1 and 2). In column 3, we report the mean SS for the top ten largest enriched annotations (column 5), i.e., ranking them in descending order according to the number of neighborhoods that the annotations are enriched in (column 4).

	Total annotations	Mean SS	Mean SS Top 10	EN	Annotation
$\widetilde{A}_{G_6}$	7	0.48 (std=0.04)	0.48 (std=0.04)	147.0	subtelomeric heterochromatin assembly
$\widetilde{A}_{G_6}$	7	0.48 (std=0.04)	0.48 (std=0.04)	107.0	negative regulation of macromolecule metabolic process
$\widetilde{A}_{G_6}$	7	0.48 (std=0.04)	0.48 (std=0.04)	78.0	regulation of nitrogen compound metabolic process
$\widetilde{A}_{G_6}$	7	0.48 (std=0.04)	0.48 (std=0.04)	72.0	regulation of nucleobase-containing compound metabolic process
$\widetilde{A}_{G_6}$	7	0.48 (std=0.04)	0.48 (std=0.04)	68.0	regulation of cellular metabolic process
$\widetilde{A}_{G_6}$	7	0.48 (std=0.04)	0.48 (std=0.04)	67.0	regulation of primary metabolic process
$\widetilde{A}_{G_6}$	7	0.48 (std=0.04)	0.48 (std=0.04)	31.0	positive regulation of RNA metabolic process
$\widetilde{A}_{G_7}$	24	0.29 (std=0.06)	0.29 (std=0.06)	42.0	signal transduction
$\widetilde{A}_{G_7}$	24	0.29 (std=0.06)	0.29 (std=0.06)	35.0	regulation of DNA recombination
$\widetilde{A}_{G_7}$	24	0.29 (std=0.06)	0.29 (std=0.06)	30.0	negative regulation of cellular process
$\widetilde{A}_{G_7}$	24	0.29 (std=0.06)	0.29 (std=0.06)	23.0	cell cycle G1/S phase transition
$\widetilde{A}_{G_7}$	24	0.29 (std=0.06)	0.29 (std=0.06)	22.0	positive regulation of mitotic cytokinetic process
$\widetilde{A}_{G_7}$	24	0.29 (std=0.06)	0.29 (std=0.06)	22.0	positive regulation of mitotic division septum assembly
$\widetilde{A}_{G_7}$	24	0.29 (std=0.06)	0.29 (std=0.06)	21.0	cell cycle phase transition
$\widetilde{A}_{G_7}$	24	0.29 (std=0.06)	0.29 (std=0.06)	21.0	septation initiation signaling
$\widetilde{A}_{G_7}$	24	0.29 (std=0.06)	0.29 (std=0.06)	20.0	positive regulation of cytokinesis
$\widetilde{A}_{G_7}$	24	0.29 (std=0.06)	0.29 (std=0.06)	20.0	G1/S transition of mitotic cell cycle
$\widetilde{A}_{G_8}$	18	0.29 (std=0.05)	0.41 (std=0.04)	30.0	mitotic cytokinetic process
$\widetilde{A}_{G_8}$	18	0.29 (std=0.05)	0.41 (std=0.04)	24.0	actomyosin structure organization
$\widetilde{A}_{G_8}$	18	0.29 (std=0.05)	0.41 (std=0.04)	24.0	assembly of actomyosin apparatus involved in cytokinesis
$\widetilde{A}_{G_8}$	18	0.29 (std=0.05)	0.41 (std=0.04)	24.0	cortical actin cytoskeleton organization
$\widetilde{A}_{G_8}$	18	0.29 (std=0.05)	0.41 (std=0.04)	24.0	actomyosin contractile ring organization
$\widetilde{A}_{G_8}$	18	0.29 (std=0.05)	0.41 (std=0.04)	24.0	actomyosin contractile ring assembly
$\widetilde{A}_{G_8}$	18	0.29 (std=0.05)	0.41 (std=0.04)	24.0	assembly of actomyosin apparatus involved in mitotic cytokinesis
$\widetilde{A}_{G_8}$	18	0.29 (std=0.05)	0.41 (std=0.04)	24.0	mitotic actomyosin contractile ring assembly
$\widetilde{A}_{G_8}$	18	0.29 (std=0.05)	0.41 (std=0.04)	23.0	cortical cytoskeleton organization
$\widetilde{A}_{G_8}$	18	0.29 (std=0.05)	0.41 (std=0.04)	23.0	mitotic actomyosin contractile ring contraction

Table A.7. Summary of uniquely enriched GO-BPs for GraSpring embeddings, Part 2. We report, for the Fission yeast GI network, the number of uniquely enriched GO-BPs and the mean semantic similarity (SS) between the uniquely enriched annotations (GO-BPs) (columns 1 and 2). In column 3, we report the mean SS for the top ten largest enriched annotations (column 5), i.e., ranking them in descending order according to the number of neighborhoods that the annotations are enriched in (column 4).

GraSprings enrichment summary for Fruit fly

	Total annotations	Mean SS	Mean SS Top 10	EN	Annotation
$\widetilde{A}_{G_0}$	38	0.15 (std=0.04)	0.22 (std=0.04)	299.0	melanin metabolic process
$\widetilde{A}_{G_0}$	38	0.15 (std=0.04)	0.22 (std=0.04)	285.0	genital disc anterior/posterior pattern formation
$\widetilde{A}_{G_0}$	38	0.15 (std=0.04)	0.22 (std=0.04)	285.0	genital disc pattern formation
$\widetilde{A}_{G_0}$	38	0.15 (std=0.04)	0.22 (std=0.04)	277.0	terminal branching, open tracheal system
$\widetilde{A}_{G_0}$	38	0.15 (std=0.04)	0.22 (std=0.04)	261.0	lymph gland plasmatocyte differentiation
$\widetilde{A}_{G_0}$	38	0.15 (std=0.04)	0.22 (std=0.04)	261.0	determination of adult lifespan
$\widetilde{A}_{G_0}$	38	0.15 (std=0.04)	0.22 (std=0.04)	236.0	neurogenesis
$\widetilde{A}_{G_0}$	38	0.15 (std=0.04)	0.22 (std=0.04)	208.0	positive regulation of histone modification
$\widetilde{A}_{G_0}$	38	0.15 (std=0.04)	0.22 (std=0.04)	196.0	negative regulation of cellular component organization
$\widetilde{A}_{G_0}$	38	0.15 (std=0.04)	0.22 (std=0.04)	189.0	cardioblast differentiation
$\widetilde{A}_{G_1}$	10	0.23 (std=0.00)	0.23 (std=0.00)	327.0	mitotic cell cycle
$\widetilde{A}_{G_1}$	10	0.23 (std=0.00)	0.23 (std=0.00)	313.0	photoreceptor cell fate specification
$\widetilde{A}_{G_1}$	10	0.23 (std=0.00)	0.23 (std=0.00)	294.0	negative regulation of hemocyte differentiation
$\widetilde{A}_{G_1}$	10	0.23 (std=0.00)	0.23 (std=0.00)	278.0	axis specification
$\widetilde{A}_{G_1}$	10	0.23 (std=0.00)	0.23 (std=0.00)	275.0	regulation of cell division
$\widetilde{A}_{G_1}$	10	0.23 (std=0.00)	0.23 (std=0.00)	250.0	regulation of vesicle-mediated transport
$\widetilde{A}_{G_1}$	10	0.23 (std=0.00)	0.23 (std=0.00)	229.0	regulation of smoothened signaling pathway
$\widetilde{A}_{G_1}$	10	0.23 (std=0.00)	0.23 (std=0.00)	226.0	positive regulation of canonical Wnt signaling pathway
$\widetilde{A}_{G_1}$	10	0.23 (std=0.00)	0.23 (std=0.00)	222.0	organ growth
$\widetilde{A}_{G_1}$	10	0.23 (std=0.00)	0.23 (std=0.00)	11.0	cellular response to radiation
$\widetilde{A}_{G_2}$	26	0.20 (std=0.05)	0.22 (std=0.04)	133.0	brain development
$\widetilde{A}_{G_2}$	26	0.20 (std=0.05)	0.22 (std=0.04)	111.0	olfactory behavior
$\widetilde{A}_{G_2}$	26	0.20 (std=0.05)	0.22 (std=0.04)	102.0	protein localization
$\widetilde{A}_{G_2}$	26	0.20 (std=0.05)	0.22 (std=0.04)	97.0	regulation of protein kinase activity
$\widetilde{A}_{G_2}$	26	0.20 (std=0.05)	0.22 (std=0.04)	96.0	regulation of transferase activity
$\widetilde{A}_{G_2}$	26	0.20 (std=0.05)	0.22 (std=0.04)	77.0	regulation of chromosome organization
$\widetilde{A}_{G_2}$	26	0.20 (std=0.05)	0.22 (std=0.04)	67.0	negative regulation of chromatin organization
$\widetilde{A}_{G_2}$	26	0.20 (std=0.05)	0.22 (std=0.04)	66.0	BMP signaling pathway
$\widetilde{A}_{G_2}$	26	0.20 (std=0.05)	0.22 (std=0.04)	61.0	nucleosome organization
$\widetilde{A}_{G_2}$	26	0.20 (std=0.05)	0.22 (std=0.04)	60.0	protein-DNA complex subunit organization

Table A.8. Summary of uniquely enriched GO-BPs for GraSpring embeddings, Part 1. We report, for the Fruit fly GI network, the number of uniquely enriched GO-BPs and the mean semantic similarity (SS) between the uniquely enriched annotations (GO-BPs) (columns 1 and 2). In column 3, we report the mean SS for the top ten largest enriched annotations (column 5), i.e., ranking them in descending order according to the number of neighborhoods that the annotations are enriched in (column 4).

	Total annotations	Mean SS	Mean SS Top 10	EN	Annotation
$\widetilde{A}_{G_3}$	17	0.20 (std=0.05)	0.23 (std=0.04)	395.0	positive regulation of organ growth
$\widetilde{A}_{G_3}$	17	0.20 (std=0.05)	0.23 (std=0.04)	323.0	establishment or maintenance of polarity of larval imaginal disc epithelium
$\widetilde{A}_{G_3}$	17	0.20 (std=0.05)	0.23 (std=0.04)	299.0	nephrocyte differentiation
$\widetilde{A}_{G_3}$	17	0.20 (std=0.05)	0.23 (std=0.04)	299.0	renal filtration cell differentiation
$\widetilde{A}_{G_3}$	17	0.20 (std=0.05)	0.23 (std=0.04)	294.0	regulation of axon guidance
$\widetilde{A}_{G_3}$	17	0.20 (std=0.05)	0.23 (std=0.04)	290.0	lymph gland crystal cell differentiation
$\widetilde{A}_{G_3}$	17	0.20 (std=0.05)	0.23 (std=0.04)	255.0	lymph gland development
$\widetilde{A}_{G_3}$	17	0.20 (std=0.05)	0.23 (std=0.04)	254.0	delamination
$\widetilde{A}_{G_3}$	17	0.20 (std=0.05)	0.23 (std=0.04)	252.0	oenocyte differentiation
$\widetilde{A}_{G_3}$	17	0.20 (std=0.05)	0.23 (std=0.04)	248.0	positive regulation of JNK cascade
$\widetilde{A}_{G_4}$	14	0.20 (std=0.01)	0.23 (std=0.01)	265.0	cell division
$\widetilde{A}_{G_4}$	14	0.20 (std=0.01)	0.23 (std=0.01)	259.0	dorsal/ventral axis specification, ovarian follicular epithelium
$\widetilde{A}_{G_4}$	14	0.20 (std=0.01)	0.23 (std=0.01)	256.0	sensory organ boundary specification
$\widetilde{A}_{G_4}$	14	0.20 (std=0.01)	0.23 (std=0.01)	220.0	epithelial cell differentiation
$\widetilde{A}_{G_4}$	14	0.20 (std=0.01)	0.23 (std=0.01)	219.0	regulation of actin filament bundle assembly
$\widetilde{A}_{G_4}$	14	0.20 (std=0.01)	0.23 (std=0.01)	189.0	regulation of cellular protein localization
$\widetilde{A}_{G_4}$	14	0.20 (std=0.01)	0.23 (std=0.01)	186.0	positive regulation of smoothened signaling pathway
$\widetilde{A}_{G_4}$	14	0.20 (std=0.01)	0.23 (std=0.01)	185.0	cell-cell adhesion mediated by cadherin
$\widetilde{A}_{G_4}$	14	0.20 (std=0.01)	0.23 (std=0.01)	158.0	calcium-dependent cell-cell adhesion via plasma membrane cell adhesion molecules
$\widetilde{A}_{G_4}$	14	0.20 (std=0.01)	0.23 (std=0.01)	92.0	female germ-line stem cell population maintenance
$\widetilde{A}_{G_5}$	34	0.19 (std=0.05)	0.33 (std=0.07)	209.0	behavior
$\widetilde{A}_{G_5}$	34	0.19 (std=0.05)	0.33 (std=0.07)	137.0	cell cycle
$\widetilde{A}_{G_5}$	34	0.19 (std=0.05)	0.33 (std=0.07)	127.0	regulation of proteolysis involved in cellular protein catabolic process
$\widetilde{A}_{G_5}$	34	0.19 (std=0.05)	0.33 (std=0.07)	121.0	regulation of proteasomal ubiquitin-dependent protein catabolic process
$\widetilde{A}_{G_5}$	34	0.19 (std=0.05)	0.33 (std=0.07)	121.0	positive regulation of proteasomal ubiquitin-dependent protein catabolic process
$\widetilde{A}_{G_5}$	34	0.19 (std=0.05)	0.33 (std=0.07)	117.0	regulation of proteasomal protein catabolic process
$\widetilde{A}_{G_5}$	34	0.19 (std=0.05)	0.33 (std=0.07)	116.0	positive regulation of proteasomal protein catabolic process
$\widetilde{A}_{G_5}$	34	0.19 (std=0.05)	0.33 (std=0.07)	115.0	response to external stimulus
$\widetilde{A}_{G_5}$	34	0.19 (std=0.05)	0.33 (std=0.07)	114.0	associative learning
$\widetilde{A}_{G_5}$	34	0.19 (std=0.05)	0.33 (std=0.07)	114.0	cellular component assembly

Table A.8. Summary of uniquely enriched GO-BPs for GraSpring embeddings, Part 2. We report, for the Fruit fly GI network, the number of uniquely enriched GO-BPs and the mean semantic similarity (SS) between the uniquely enriched annotations (GO-BPs) (columns 1 and 2). In column 3, we report the mean SS for the top ten largest enriched annotations (column 5), i.e., ranking them in descending order according to the number of neighborhoods that the annotations are enriched in (column 4).

	Total annotations	Mean SS	Mean SS Top 10	EN	Annotation
$\widetilde{A}_{G_6}$	18	0.22 (std=0.04)	0.24 (std=0.04)	277.0	regulation of compound eye retinal cell programmed cell death
$\widetilde{A}_{G_6}$	18	0.22 (std=0.04)	0.24 (std=0.04)	277.0	regulation of retinal cell programmed cell death
$\widetilde{A}_{G_6}$	18	0.22 (std=0.04)	0.24 (std=0.04)	259.0	chaeta development
$\widetilde{A}_{G_6}$	18	0.22 (std=0.04)	0.24 (std=0.04)	182.0	oocyte axis specification
$\widetilde{A}_{G_6}$	18	0.22 (std=0.04)	0.24 (std=0.04)	171.0	intrinsic apoptotic signaling pathway in response to DNA damage by p53 class mediator
$\widetilde{A}_{G_6}$	18	0.22 (std=0.04)	0.24 (std=0.04)	170.0	equator specification
$\widetilde{A}_{G_6}$	18	0.22 (std=0.04)	0.24 (std=0.04)	141.0	cellular component maintenance
$\widetilde{A}_{G_6}$	18	0.22 (std=0.04)	0.24 (std=0.04)	126.0	regulation of embryonic development
$\widetilde{A}_{G_6}$	18	0.22 (std=0.04)	0.24 (std=0.04)	119.0	carbohydrate metabolic process
$\widetilde{A}_{G_6}$	18	0.22 (std=0.04)	0.24 (std=0.04)	52.0	RNA interference
$\widetilde{A}_{G_7}$	16	0.18 (std=0.05)	0.20 (std=0.01)	106.0	determination of digestive tract left/right asymmetry
$\widetilde{A}_{G_7}$	16	0.18 (std=0.05)	0.20 (std=0.01)	68.0	actomyosin structure organization
$\widetilde{A}_{G_7}$	16	0.18 (std=0.05)	0.20 (std=0.01)	37.0	multicellular organism aging
$\widetilde{A}_{G_7}$	16	0.18 (std=0.05)	0.20 (std=0.01)	33.0	negative regulation of histone methylation
$\widetilde{A}_{G_7}$	16	0.18 (std=0.05)	0.20 (std=0.01)	24.0	triglyceride homeostasis
$\widetilde{A}_{G_7}$	16	0.18 (std=0.05)	0.20 (std=0.01)	23.0	larval feeding behavior
$\widetilde{A}_{G_7}$	16	0.18 (std=0.05)	0.20 (std=0.01)	19.0	non-recombinational repair
$\widetilde{A}_{G_7}$	16	0.18 (std=0.05)	0.20 (std=0.01)	19.0	double-strand break repair via single-strand annealing
$\widetilde{A}_{G_7}$	16	0.18 (std=0.05)	0.20 (std=0.01)	18.0	ncRNA metabolic process
$\widetilde{A}_{G_7}$	16	0.18 (std=0.05)	0.20 (std=0.01)	17.0	regulation of membrane potential
$\widetilde{A}_{G_8}$	15	0.21 (std=0.04)	0.25 (std=0.05)	63.0	terminal region determination
$\widetilde{A}_{G_8}$	15	0.21 (std=0.04)	0.25 (std=0.05)	59.0	determination of bilateral symmetry
$\widetilde{A}_{G_8}$	15	0.21 (std=0.04)	0.25 (std=0.05)	59.0	specification of symmetry
$\widetilde{A}_{G_8}$	15	0.21 (std=0.04)	0.25 (std=0.05)	43.0	homeostatic process
$\widetilde{A}_{G_8}$	15	0.21 (std=0.04)	0.25 (std=0.05)	40.0	regulation of TORC2 signaling
$\widetilde{A}_{G_8}$	15	0.21 (std=0.04)	0.25 (std=0.05)	35.0	germline ring canal formation
$\widetilde{A}_{G_8}$	15	0.21 (std=0.04)	0.25 (std=0.05)	33.0	glucose homeostasis
$\widetilde{A}_{G_8}$	15	0.21 (std=0.04)	0.25 (std=0.05)	30.0	negative regulation of cell size
$\widetilde{A}_{G_8}$	15	0.21 (std=0.04)	0.25 (std=0.05)	27.0	cellular response to organic substance
$\widetilde{A}_{G_8}$	15	0.21 (std=0.04)	0.25 (std=0.05)	25.0	cellular response to hormone stimulus

Table A.8. Summary of uniquely enriched GO-BPs for GraSpring embeddings, Part 3. We report, for the Fruit fly GI network, the number of uniquely enriched GO-BPs and the mean semantic similarity (SS) between the uniquely enriched annotations (GO-BPs) (columns 1 and 2). In column 3, we report the mean SS for the top ten largest enriched annotations (column 5), i.e., ranking them in descending order according to the number of neighborhoods that the annotations are enriched in (column 4).

Graphlet Spectrals enrichment summary for *E. coli*

	Total annotations	Mean SS	Mean SS Top 10	EN	Annotation
$\widetilde{A}_{G_0}$	29	0.21 (std=0.05)	0.21 (std=0.04)	221.0	fatty acid elongation
$\widetilde{A}_{G_0}$	29	0.21 (std=0.05)	0.21 (std=0.04)	204.0	siderophore-dependent iron import into cell
$\widetilde{A}_{G_0}$	29	0.21 (std=0.05)	0.21 (std=0.04)	148.0	response to endogenous stimulus
$\widetilde{A}_{G_0}$	29	0.21 (std=0.05)	0.21 (std=0.04)	124.0	protein import
$\widetilde{A}_{G_0}$	29	0.21 (std=0.05)	0.21 (std=0.04)	121.0	glycerolipid biosynthetic process
$\widetilde{A}_{G_0}$	29	0.21 (std=0.05)	0.21 (std=0.04)	121.0	glycerophospholipid biosynthetic process
$\widetilde{A}_{G_0}$	29	0.21 (std=0.05)	0.21 (std=0.04)	100.0	cell septum assembly
$\widetilde{A}_{G_0}$	29	0.21 (std=0.05)	0.21 (std=0.04)	58.0	SOS response
$\widetilde{A}_{G_0}$	29	0.21 (std=0.05)	0.21 (std=0.04)	56.0	anaerobic respiration
$\widetilde{A}_{G_0}$	29	0.21 (std=0.05)	0.21 (std=0.04)	53.0	ribonucleoprotein complex assembly
$\widetilde{A}_{G_1}$	9	0.37 (std=0.03)	0.37 (std=0.03)	164.0	antibiotic biosynthetic process
$\widetilde{A}_{G_1}$	9	0.37 (std=0.03)	0.37 (std=0.03)	164.0	catechol-containing siderophore biosynthetic process
$\widetilde{A}_{G_1}$	9	0.37 (std=0.03)	0.37 (std=0.03)	164.0	polyketide biosynthetic process
$\widetilde{A}_{G_1}$	9	0.37 (std=0.03)	0.37 (std=0.03)	164.0	catechol-containing compound biosynthetic process
$\widetilde{A}_{G_1}$	9	0.37 (std=0.03)	0.37 (std=0.03)	164.0	lactone biosynthetic process
$\widetilde{A}_{G_1}$	9	0.37 (std=0.03)	0.37 (std=0.03)	164.0	enterobactin biosynthetic process
$\widetilde{A}_{G_1}$	9	0.37 (std=0.03)	0.37 (std=0.03)	164.0	macrolide biosynthetic process
$\widetilde{A}_{G_1}$	9	0.37 (std=0.03)	0.37 (std=0.03)	164.0	phenol-containing compound biosynthetic process
$\widetilde{A}_{G_1}$	9	0.37 (std=0.03)	0.37 (std=0.03)	152.0	ferric-enterobactin import into cell
$\widetilde{A}_{G_2}$	6	0.47 (std=0.05)	0.47 (std=0.05)	160.0	dipeptide transmembrane transport
$\widetilde{A}_{G_2}$	6	0.47 (std=0.05)	0.47 (std=0.05)	146.0	xenobiotic export
$\widetilde{A}_{G_2}$	6	0.47 (std=0.05)	0.47 (std=0.05)	146.0	xenobiotic detoxification by transmembrane export across the plasma membrane
$\widetilde{A}_{G_2}$	6	0.47 (std=0.05)	0.47 (std=0.05)	141.0	organophosphate ester transport
$\widetilde{A}_{G_2}$	6	0.47 (std=0.05)	0.47 (std=0.05)	102.0	oligopeptide transmembrane transport
$\widetilde{A}_{G_2}$	6	0.47 (std=0.05)	0.47 (std=0.05)	89.0	oligopeptide transport
$\widetilde{A}_{G_3}$	5	0.41 (std=0.05)	0.41 (std=0.05)	265.0	regulation of phosphorylation
$\widetilde{A}_{G_3}$	5	0.41 (std=0.05)	0.41 (std=0.05)	265.0	regulation of kinase activity
$\widetilde{A}_{G_3}$	5	0.41 (std=0.05)	0.41 (std=0.05)	185.0	regulation of transferase activity
$\widetilde{A}_{G_3}$	5	0.41 (std=0.05)	0.41 (std=0.05)	53.0	primary metabolic process
$\widetilde{A}_{G_3}$	5	0.41 (std=0.05)	0.41 (std=0.05)	33.0	organic substance metabolic process

Table A.9. Summary of uniquely enriched GO-BPs for graphlet Spectral embeddings, Part 1. We report, for the *E. coli* GI network, the number of uniquely enriched GO-BPs and the mean semantic similarity (SS) between the uniquely enriched annotations (GO-BPs) (columns 1 and 2). In column 3, we report the mean SS for the top ten largest enriched annotations (column 5), i.e., ranking them in descending order according to the number of neighborhoods that the annotations are enriched in (column 4).

	Total annotations	**Mean SS**	**Mean SS Top 10**	**EN**	**Annotation**
$\widetilde{A}_{G_4}$	17	0.23 (std=0.05)	0.35 (std=0.05)	199.0	phenylacetate catabolic process
$\widetilde{A}_{G_4}$	17	0.23 (std=0.05)	0.35 (std=0.05)	177.0	nucleotide-sugar biosynthetic process
$\widetilde{A}_{G_4}$	17	0.23 (std=0.05)	0.35 (std=0.05)	172.0	amide biosynthetic process
$\widetilde{A}_{G_4}$	17	0.23 (std=0.05)	0.35 (std=0.05)	172.0	nucleotide-sugar metabolic process
$\widetilde{A}_{G_4}$	17	0.23 (std=0.05)	0.35 (std=0.05)	168.0	cellular amide metabolic process
$\widetilde{A}_{G_4}$	17	0.23 (std=0.05)	0.35 (std=0.05)	164.0	GDP-mannose biosynthetic process
$\widetilde{A}_{G_4}$	17	0.23 (std=0.05)	0.35 (std=0.05)	164.0	GDP-mannose metabolic process
$\widetilde{A}_{G_4}$	17	0.23 (std=0.05)	0.35 (std=0.05)	137.0	DNA replication
$\widetilde{A}_{G_4}$	17	0.23 (std=0.05)	0.35 (std=0.05)	127.0	10-formyltetrahydrofolate metabolic process
$\widetilde{A}_{G_4}$	17	0.23 (std=0.05)	0.35 (std=0.05)	127.0	10-formyltetrahydrofolate biosynthetic process
$\widetilde{A}_{G_6}$	13	0.26 (std=0.05)	0.26 (std=0.05)	170.0	cellular copper ion homeostasis
$\widetilde{A}_{G_6}$	13	0.26 (std=0.05)	0.26 (std=0.05)	148.0	copper ion homeostasis
$\widetilde{A}_{G_6}$	13	0.26 (std=0.05)	0.26 (std=0.05)	129.0	response to silver ion
$\widetilde{A}_{G_6}$	13	0.26 (std=0.05)	0.26 (std=0.05)	129.0	detoxification of inorganic compound
$\widetilde{A}_{G_6}$	13	0.26 (std=0.05)	0.26 (std=0.05)	129.0	detoxification of copper ion
$\widetilde{A}_{G_6}$	13	0.26 (std=0.05)	0.26 (std=0.05)	28.0	keto-3-deoxy-D-manno-octulosonic acid metabolic process
$\widetilde{A}_{G_6}$	13	0.26 (std=0.05)	0.26 (std=0.05)	28.0	keto-3-deoxy-D-manno-octulosonic acid biosynthetic process
$\widetilde{A}_{G_6}$	13	0.26 (std=0.05)	0.26 (std=0.05)	15.0	small molecule biosynthetic process
$\widetilde{A}_{G_6}$	13	0.26 (std=0.05)	0.26 (std=0.05)	14.0	isopentenyl diphosphate metabolic process
$\widetilde{A}_{G_6}$	13	0.26 (std=0.05)	0.26 (std=0.05)	14.0	glyceraldehyde-3-phosphate metabolic process
$\widetilde{A}_{G_8}$	2	0.67 (std=0.00)	0.67 (std=0.00)	34.0	carbohydrate metabolic process
$\widetilde{A}_{G_8}$	2	0.67 (std=0.00)	0.67 (std=0.00)	20.0	cellular metabolic process

Table A.9. Summary of uniquely enriched GO-BPs for graphlet Spectral embeddings, Part 2. We report, for the *E. coli* GI network, the number of uniquely enriched GO-BPs and the mean semantic similarity (SS) between the uniquely enriched annotations (GO-BPs) (columns 1 and 2). In column 3, we report the mean SS for the top ten largest enriched annotations (column 5), i.e., ranking them in descending order according to the number of neighborhoods that the annotations are enriched in (column 4).

Graphlet Spectrals enrichment summary for fission yeast

	Total annotations	Mean SS	Mean SS Top 10	EN	Annotation
$\widetilde{A}_{G_0}$	19	0.28 (std=0.04)	0.40 (std=0.04)	39.0	mitotic spindle pole body localization
$\widetilde{A}_{G_0}$	19	0.28 (std=0.04)	0.40 (std=0.04)	39.0	spindle pole body localization
$\widetilde{A}_{G_0}$	19	0.28 (std=0.04)	0.40 (std=0.04)	39.0	mitotic spindle pole body insertion into the nuclear envelope
$\widetilde{A}_{G_0}$	19	0.28 (std=0.04)	0.40 (std=0.04)	39.0	microtubule organizing center localization
$\widetilde{A}_{G_0}$	19	0.28 (std=0.04)	0.40 (std=0.04)	26.0	actin filament-based process
$\widetilde{A}_{G_0}$	19	0.28 (std=0.04)	0.40 (std=0.04)	25.0	protein-DNA complex subunit organization
$\widetilde{A}_{G_0}$	19	0.28 (std=0.04)	0.40 (std=0.04)	23.0	actin cytoskeleton organization
$\widetilde{A}_{G_0}$	19	0.28 (std=0.04)	0.40 (std=0.04)	22.0	cellular localization
$\widetilde{A}_{G_0}$	19	0.28 (std=0.04)	0.40 (std=0.04)	20.0	cortical actin cytoskeleton organization
$\widetilde{A}_{G_0}$	19	0.28 (std=0.04)	0.40 (std=0.04)	20.0	nucleosome organization
$\widetilde{A}_{G_1}$	3	0.61 (std=0.00)	0.61 (std=0.00)	165.0	mitotic cell cycle process
$\widetilde{A}_{G_1}$	3	0.61 (std=0.00)	0.61 (std=0.00)	54.0	cell cycle DNA replication
$\widetilde{A}_{G_1}$	3	0.61 (std=0.00)	0.61 (std=0.00)	54.0	nuclear DNA replication
$\widetilde{A}_{G_2}$	10	0.44 (std=0.04)	0.44 (std=0.04)	23.0	cellular response to DNA damage stimulus
$\widetilde{A}_{G_2}$	10	0.44 (std=0.04)	0.44 (std=0.04)	22.0	DNA recombination
$\widetilde{A}_{G_2}$	10	0.44 (std=0.04)	0.44 (std=0.04)	21.0	nucleobase-containing compound metabolic process
$\widetilde{A}_{G_2}$	10	0.44 (std=0.04)	0.44 (std=0.04)	21.0	heterocycle metabolic process
$\widetilde{A}_{G_2}$	10	0.44 (std=0.04)	0.44 (std=0.04)	21.0	organic cyclic compound metabolic process
$\widetilde{A}_{G_2}$	10	0.44 (std=0.04)	0.44 (std=0.04)	21.0	cellular aromatic compound metabolic process
$\widetilde{A}_{G_2}$	10	0.44 (std=0.04)	0.44 (std=0.04)	20.0	cellular nitrogen compound metabolic process
$\widetilde{A}_{G_2}$	10	0.44 (std=0.04)	0.44 (std=0.04)	19.0	nucleic acid metabolic process
$\widetilde{A}_{G_2}$	10	0.44 (std=0.04)	0.44 (std=0.04)	18.0	double-strand break repair
$\widetilde{A}_{G_2}$	10	0.44 (std=0.04)	0.44 (std=0.04)	14.0	macromolecule metabolic process
$\widetilde{A}_{G_4}$	11	0.35 (std=0.05)	0.39 (std=0.04)	207.0	regulation of cell cycle phase transition
$\widetilde{A}_{G_4}$	11	0.35 (std=0.05)	0.39 (std=0.04)	182.0	negative regulation of mitotic cell cycle
$\widetilde{A}_{G_4}$	11	0.35 (std=0.05)	0.39 (std=0.04)	48.0	modification-dependent protein catabolic process
$\widetilde{A}_{G_4}$	11	0.35 (std=0.05)	0.39 (std=0.04)	48.0	ubiquitin-dependent protein catabolic process
$\widetilde{A}_{G_4}$	11	0.35 (std=0.05)	0.39 (std=0.04)	46.0	modification-dependent macromolecule catabolic process
$\widetilde{A}_{G_4}$	11	0.35 (std=0.05)	0.39 (std=0.04)	39.0	proteasomal protein catabolic process
$\widetilde{A}_{G_4}$	11	0.35 (std=0.05)	0.39 (std=0.04)	26.0	macromolecule catabolic process
$\widetilde{A}_{G_4}$	11	0.35 (std=0.05)	0.39 (std=0.04)	26.0	proteolysis
$\widetilde{A}_{G_4}$	11	0.35 (std=0.05)	0.39 (std=0.04)	24.0	cellular macromolecule catabolic process
$\widetilde{A}_{G_4}$	11	0.35 (std=0.05)	0.39 (std=0.04)	15.0	organic substance catabolic process
$\widetilde{A}_{G_5}$	1	1.00 (std=nan)	1.00 (std=nan)	14.0	nucleotide-excision repair

Table A.10. Summary of uniquely enriched GO-BPs for graphlet Spectral embeddings. We report, for the fission yeast GI network, the number of uniquely enriched GO-BPs and the mean semantic similarity (SS) between the uniquely enriched annotations (GO-BPs) (columns 1 and 2). In column 3, we report the mean SS for the top ten largest enriched annotations (column 5), i.e., ranking them in descending order according to the number of neighborhoods that the annotations are enriched in (column 4).

Graphlet Spectrals enrichment summary for fruit fly

	Total annotations	Mean SS	Mean SS Top 10	EN	Annotation
$\widetilde{A}_{G_0}$	82	0.14 (std=0.03)	0.30 (std=0.06)	530.0	cellular process
$\widetilde{A}_{G_0}$	82	0.14 (std=0.03)	0.30 (std=0.06)	516.0	regulation of RNA biosynthetic process
$\widetilde{A}_{G_0}$	82	0.14 (std=0.03)	0.30 (std=0.06)	516.0	regulation of nucleic acid-templated transcription
$\widetilde{A}_{G_0}$	82	0.14 (std=0.03)	0.30 (std=0.06)	516.0	regulation of transcription, DNA-templated
$\widetilde{A}_{G_0}$	82	0.14 (std=0.03)	0.30 (std=0.06)	448.0	ameboidal-type cell migration
$\widetilde{A}_{G_0}$	82	0.14 (std=0.03)	0.30 (std=0.06)	418.0	dendrite morphogenesis
$\widetilde{A}_{G_0}$	82	0.14 (std=0.03)	0.30 (std=0.06)	399.0	cellular component organization
$\widetilde{A}_{G_0}$	82	0.14 (std=0.03)	0.30 (std=0.06)	397.0	cellular component organization or biogenesis
$\widetilde{A}_{G_0}$	82	0.14 (std=0.03)	0.30 (std=0.06)	386.0	enzyme linked receptor protein signaling pathway
$\widetilde{A}_{G_0}$	82	0.14 (std=0.03)	0.30 (std=0.06)	385.0	anatomical structure formation involved in morphogenesis
$\widetilde{A}_{G_1}$	71	0.15 (std=0.02)	0.32 (std=0.04)	456.0	regulation of response to stimulus
$\widetilde{A}_{G_1}$	71	0.15 (std=0.02)	0.32 (std=0.04)	406.0	regulation of signaling
$\widetilde{A}_{G_1}$	71	0.15 (std=0.02)	0.32 (std=0.04)	402.0	regulation of cell communication
$\widetilde{A}_{G_1}$	71	0.15 (std=0.02)	0.32 (std=0.04)	379.0	asymmetric cell division
$\widetilde{A}_{G_1}$	71	0.15 (std=0.02)	0.32 (std=0.04)	363.0	regulation of morphogenesis of an epithelium
$\widetilde{A}_{G_1}$	71	0.15 (std=0.02)	0.32 (std=0.04)	353.0	actin cytoskeleton organization
$\widetilde{A}_{G_1}$	71	0.15 (std=0.02)	0.32 (std=0.04)	353.0	positive regulation of signaling
$\widetilde{A}_{G_1}$	71	0.15 (std=0.02)	0.32 (std=0.04)	353.0	positive regulation of cell communication
$\widetilde{A}_{G_1}$	71	0.15 (std=0.02)	0.32 (std=0.04)	350.0	regulation of actin filament-based process
$\widetilde{A}_{G_1}$	71	0.15 (std=0.02)	0.32 (std=0.04)	341.0	supramolecular fiber organization
$\widetilde{A}_{G_2}$	24	0.20 (std=0.05)	0.25 (std=0.05)	133.0	positive regulation of developmental process
$\widetilde{A}_{G_2}$	24	0.20 (std=0.05)	0.25 (std=0.05)	124.0	negative regulation of cell differentiation
$\widetilde{A}_{G_2}$	24	0.20 (std=0.05)	0.25 (std=0.05)	112.0	ovarian follicle cell migration
$\widetilde{A}_{G_2}$	24	0.20 (std=0.05)	0.25 (std=0.05)	107.0	border follicle cell migration
$\widetilde{A}_{G_2}$	24	0.20 (std=0.05)	0.25 (std=0.05)	94.0	positive regulation of hippo signaling
$\widetilde{A}_{G_2}$	24	0.20 (std=0.05)	0.25 (std=0.05)	92.0	positive regulation of intracellular signal transduction
$\widetilde{A}_{G_2}$	24	0.20 (std=0.05)	0.25 (std=0.05)	85.0	positive regulation of nervous system development
$\widetilde{A}_{G_2}$	24	0.20 (std=0.05)	0.25 (std=0.05)	52.0	regulation of response to external stimulus
$\widetilde{A}_{G_2}$	24	0.20 (std=0.05)	0.25 (std=0.05)	26.0	cellular response to DNA damage stimulus
$\widetilde{A}_{G_2}$	24	0.20 (std=0.05)	0.25 (std=0.05)	20.0	DNA synthesis involved in DNA repair

Table A.11. Summary of uniquely enriched GO-BPs for graphlet Spectral embeddings, Part 1. We report, for the fruit fly GI network, the number of uniquely enriched GO-BPs and the mean semantic similarity (SS) between the uniquely enriched annotations (GO-BPs) (columns 1 and 2). In column 3, we report the mean SS for the top ten largest enriched annotations (column 5), i.e., ranking them in descending order according to the number of neighborhoods that the annotations are enriched in (column 4).

	Total annotations	Mean SS	Mean SS Top 10	EN	Annotation
$\widetilde{A}_{G_3}$	29	0.21 (std=0.05)	0.41 (std=0.05)	149.0	germ-line stem cell population maintenance
$\widetilde{A}_{G_3}$	29	0.21 (std=0.05)	0.41 (std=0.05)	138.0	cell cycle checkpoint
$\widetilde{A}_{G_3}$	29	0.21 (std=0.05)	0.41 (std=0.05)	77.0	positive regulation of filopodium assembly
$\widetilde{A}_{G_3}$	29	0.21 (std=0.05)	0.41 (std=0.05)	20.0	inorganic cation transmembrane transport
$\widetilde{A}_{G_3}$	29	0.21 (std=0.05)	0.41 (std=0.05)	20.0	zinc ion transmembrane transport
$\widetilde{A}_{G_3}$	29	0.21 (std=0.05)	0.41 (std=0.05)	20.0	inorganic cation import across plasma membrane
$\widetilde{A}_{G_3}$	29	0.21 (std=0.05)	0.41 (std=0.05)	20.0	cation transmembrane transport
$\widetilde{A}_{G_3}$	29	0.21 (std=0.05)	0.41 (std=0.05)	20.0	inorganic ion transmembrane transport
$\widetilde{A}_{G_3}$	29	0.21 (std=0.05)	0.41 (std=0.05)	20.0	inorganic ion import across plasma membrane
$\widetilde{A}_{G_3}$	29	0.21 (std=0.05)	0.41 (std=0.05)	20.0	import across plasma membrane
$\widetilde{A}_{G_4}$	63	0.15 (std=0.04)	0.35 (std=0.06)	458.0	signal transduction
$\widetilde{A}_{G_4}$	63	0.15 (std=0.04)	0.35 (std=0.06)	451.0	negative regulation of response to stimulus
$\widetilde{A}_{G_4}$	63	0.15 (std=0.04)	0.35 (std=0.06)	287.0	regulation of cell death
$\widetilde{A}_{G_4}$	63	0.15 (std=0.04)	0.35 (std=0.06)	184.0	regulation of phosphorylation
$\widetilde{A}_{G_4}$	63	0.15 (std=0.04)	0.35 (std=0.06)	183.0	regulation of cell size
$\widetilde{A}_{G_4}$	63	0.15 (std=0.04)	0.35 (std=0.06)	182.0	regulation of phosphate metabolic process
$\widetilde{A}_{G_4}$	63	0.15 (std=0.04)	0.35 (std=0.06)	182.0	regulation of phosphorus metabolic process
$\widetilde{A}_{G_4}$	63	0.15 (std=0.04)	0.35 (std=0.06)	181.0	regulation of protein phosphorylation
$\widetilde{A}_{G_4}$	63	0.15 (std=0.04)	0.35 (std=0.06)	175.0	regulation of MAPK cascade
$\widetilde{A}_{G_4}$	63	0.15 (std=0.04)	0.35 (std=0.06)	175.0	regulation of protein modification process
$\widetilde{A}_{G_5}$	53	0.15 (std=0.03)	0.28 (std=0.06)	35.0	regulation of cytoskeleton organization
$\widetilde{A}_{G_5}$	53	0.15 (std=0.03)	0.28 (std=0.06)	35.0	regulation of G protein-coupled receptor signaling pathway
$\widetilde{A}_{G_5}$	53	0.15 (std=0.03)	0.28 (std=0.06)	33.0	deactivation of rhodopsin mediated signaling
$\widetilde{A}_{G_5}$	53	0.15 (std=0.03)	0.28 (std=0.06)	33.0	regulation of rhodopsin mediated signaling pathway
$\widetilde{A}_{G_5}$	53	0.15 (std=0.03)	0.28 (std=0.06)	32.0	neuron fate determination

Table A.11. Summary of uniquely enriched GO-BPs for graphlet Spectral embeddings, Part 2. We report, for the fruit fly GI network, the number of uniquely enriched GO-BPs and the mean semantic similarity (SS) between the uniquely enriched annotations (GO-BPs) (columns 1 and 2). In column 3, we report the mean SS for the top ten largest enriched annotations (column 5), i.e., ranking them in descending order according to the number of neighborhoods that the annotations are enriched in (column 4).

	Total annotations	Mean SS	Mean SS Top 10	EN	Annotation
$\widetilde{A}_{G_5}$	53	0.15 (std=0.03)	0.28 (std=0.06)	31.0	thermotaxis
$\widetilde{A}_{G_5}$	53	0.15 (std=0.03)	0.28 (std=0.06)	31.0	photoreceptor cell fate determination
$\widetilde{A}_{G_5}$	53	0.15 (std=0.03)	0.28 (std=0.06)	28.0	response to other organism
$\widetilde{A}_{G_5}$	53	0.15 (std=0.03)	0.28 (std=0.06)	27.0	multi-organism process
$\widetilde{A}_{G_5}$	53	0.15 (std=0.03)	0.28 (std=0.06)	21.0	unidimensional cell growth
$\widetilde{A}_{G_6}$	19	0.21 (std=0.06)	0.33 (std=0.07)	404.0	transmembrane receptor protein serine/threonine kinase signaling pathway
$\widetilde{A}_{G_6}$	19	0.21 (std=0.06)	0.33 (std=0.07)	339.0	regulation of transmembrane receptor protein serine/threonine kinase signaling pathway
$\widetilde{A}_{G_6}$	19	0.21 (std=0.06)	0.33 (std=0.07)	337.0	regulation of cellular response to growth factor stimulus
$\widetilde{A}_{G_6}$	19	0.21 (std=0.06)	0.33 (std=0.07)	323.0	regulation of BMP signaling pathway
$\widetilde{A}_{G_6}$	19	0.21 (std=0.06)	0.33 (std=0.07)	277.0	negative regulation of BMP signaling pathway
$\widetilde{A}_{G_6}$	19	0.21 (std=0.06)	0.33 (std=0.07)	276.0	negative regulation of transmembrane receptor protein serine/threonine kinase signaling pathway
$\widetilde{A}_{G_6}$	19	0.21 (std=0.06)	0.33 (std=0.07)	230.0	photoreceptor cell fate specification
$\widetilde{A}_{G_6}$	19	0.21 (std=0.06)	0.33 (std=0.07)	150.0	cell-cell fusion
$\widetilde{A}_{G_6}$	19	0.21 (std=0.06)	0.33 (std=0.07)	150.0	syncytium formation by plasma membrane fusion
$\widetilde{A}_{G_6}$	19	0.21 (std=0.06)	0.33 (std=0.07)	150.0	syncytium formation
$\widetilde{A}_{G_7}$	18	0.23 (std=0.05)	0.24 (std=0.04)	150.0	positive regulation of response to biotic stimulus
$\widetilde{A}_{G_7}$	18	0.23 (std=0.05)	0.24 (std=0.04)	138.0	regulation of immune response
$\widetilde{A}_{G_7}$	18	0.23 (std=0.05)	0.24 (std=0.04)	25.0	wing disc dorsal/ventral pattern formation
$\widetilde{A}_{G_7}$	18	0.23 (std=0.05)	0.24 (std=0.04)	23.0	synapse organization
$\widetilde{A}_{G_7}$	18	0.23 (std=0.05)	0.24 (std=0.04)	20.0	negative regulation of macromolecule biosynthetic process
$\widetilde{A}_{G_7}$	18	0.23 (std=0.05)	0.24 (std=0.04)	20.0	negative regulation of cellular macromolecule biosynthetic process
$\widetilde{A}_{G_7}$	18	0.23 (std=0.05)	0.24 (std=0.04)	19.0	sex determination
$\widetilde{A}_{G_7}$	18	0.23 (std=0.05)	0.24 (std=0.04)	16.0	cellular nitrogen compound metabolic process
$\widetilde{A}_{G_7}$	18	0.23 (std=0.05)	0.24 (std=0.04)	16.0	organic cyclic compound metabolic process
$\widetilde{A}_{G_7}$	18	0.23 (std=0.05)	0.24 (std=0.04)	16.0	heterocycle metabolic process

Table A.11. Summary of uniquely enriched GO-BPs for graphlet Spectral embeddings, Part 3. We report, for the fruit fly GI network, the number of uniquely enriched GO-BPs and the mean semantic similarity (SS) between the uniquely enriched annotations (GO-BPs) (columns 1 and 2). In column 3, we report the mean SS for the top ten largest enriched annotations (column 5), i.e., ranking them in descending order according to the number of neighborhoods that the annotations are enriched in (column 4).

Functional domain summary GI networks

GraCoals functional domain summary for *E. coli*

	Num functional domains	Mean paralog ratio	Domain paralog ratio	Domain max JI	Domain description
$\widetilde{A}_{G_0}$	12	0.21 (std=0.16)	0.20	0.25	protein, transport, establishment, localization, targeting
$\widetilde{A}_{G_0}$	12	0.21 (std=0.16)	0.22	0.35	process, metabolic, rRNA, tRNA, pseudouridine
$\widetilde{A}_{G_0}$	12	0.21 (std=0.16)	0.24	0.50	localization, membrane, lipoprotein, protein, cellular
$\widetilde{A}_{G_1}$	7	0.23 (std=0.17)	0.18	0.30	metabolic, process, processing, tRNA, rRNA
$\widetilde{A}_{G_1}$	7	0.23 (std=0.17)	0.07	0.50	process, cell, transport, protein, biosynthetic
$\widetilde{A}_{G_1}$	7	0.23 (std=0.17)	0.14	0.57	process, localization, metabolic, biosynthetic, cellular
$\widetilde{A}_{G_2}$	9	0.15 (std=0.13)	0.12	0.10	proteolysis
$\widetilde{A}_{G_2}$	9	0.15 (std=0.13)	0.12	0.35	process, metabolic, biosynthetic, regulation, isopentenyl
$\widetilde{A}_{G_2}$	9	0.15 (std=0.13)	0.33	0.69	localization, transport, protein, cellular, macromolecule
$\widetilde{A}_{G_3}$	8	0.21 (std=0.17)	0.52	0.39	transport, ion, iron, localization, organic
$\widetilde{A}_{G_3}$	8	0.21 (std=0.17)	0.00	0.50	process, biosynthetic, cell, macromolecule, wall
$\widetilde{A}_{G_3}$	8	0.21 (std=0.17)	0.17	0.57	process, biosynthetic, localization, regulation, cellular

Table A.12. Summary of most unique functional domains for Gracoal embeddings, part 1. We report, for each GraCoal embedding used with SAFE with the *E. coli* GI network, the number of functional domains (column 1) the mean paralog ratio (column 2) and the top three most characteristic functional domains (column 5). Lastly, for each functional domain we report the paralog ratio (column 3) and the maximum Jaccard similarity index (JI).

	Num functional domains	Mean paralog ratio	Domain paralog ratio	Domain max JI	Domain description
$\widetilde{A}_{G_4}$	9	0.17 (std=0.09)	0.05	0.42	localization, process, protein, cell, cytokinesis
$\widetilde{A}_{G_4}$	9	0.17 (std=0.09)	0.14	0.43	process, biosynthetic, metabolic, lipid, cellular
$\widetilde{A}_{G_4}$	9	0.17 (std=0.09)	0.17	0.44	localization, protein, membrane, cellular, within
$\widetilde{A}_{G_5}$	11	0.22 (std=0.16)	0.17	0.22	ubiquinone, process, biosynthetic, metabolic
$\widetilde{A}_{G_5}$	11	0.22 (std=0.16)	0.37	0.40	rRNA, metabolic, process, processing
$\widetilde{A}_{G_5}$	11	0.22 (std=0.16)	0.51	0.44	iron, ion, transport, cell, import
$\widetilde{A}_{G_6}$	10	0.21 (std=0.17)	0.14	0.50	process, localization, metabolic, cellular, lipid
$\widetilde{A}_{G_6}$	10	0.21 (std=0.17)	0.17	0.67	processing, ncRNA, RNA
$\widetilde{A}_{G_6}$	10	0.21 (std=0.17)	0.29	0.70	homeostasis, ion, cellular, metal, copper
$\widetilde{A}_{G_7}$	12	0.19 (std=0.16)	0.00	0.22	ion, copper, transport, detoxification, homeostasis
$\widetilde{A}_{G_7}$	12	0.19 (std=0.16)	0.19	0.40	subunit, assembly, rRNA, ribonucleoprotein, complex
$\widetilde{A}_{G_7}$	12	0.19 (std=0.16)	0.16	0.47	processing, macromolecule, methylation, RNA, modification
$\widetilde{A}_{G_8}$	6	0.18 (std=0.09)	0.12	0.00	transport, glycerol, 3, phosphate, transmembrane
$\widetilde{A}_{G_8}$	6	0.18 (std=0.09)	0.36	0.48	transport, localization, protein, establishment, transmembrane
$\widetilde{A}_{G_8}$	6	0.18 (std=0.09)	0.17	0.50	metabolic, process, compound, cellular, DNA

Table A.12. Summary of most unique functional domains for Gracoal embeddings, part 2. We report, for each GraCoal embedding used with SAFE with the *E. coli* GI network, the number of functional domains (column 1) the mean paralog ratio (column 2) and the top three most characteristic functional domains (column 5). Lastly, for each functional domain we report the paralog ratio (column 3) and the maximum Jaccard similarity index (JI).

GraCoals functional domain summary for fission yeast

	Num functional domains	Mean paralog ratio	Domain paralog ratio	Domain max JI	Domain description
$\widetilde{A}_{G_0}$	2	0.12 (std=0.12)	0.00	0.09	DNA, replication, initiation, cell, cycle
$\widetilde{A}_{G_0}$	2	0.12 (std=0.12)	0.24	0.27	regulation, cell, cycle, mitotic, negative
$\widetilde{A}_{G_1}$	3	0.03 (std=0.04)	0.00	0.03	regulation, mitotic, division, septum, assembly
$\widetilde{A}_{G_1}$	3	0.03 (std=0.04)	0.08	0.07	DNA, repair
$\widetilde{A}_{G_1}$	3	0.03 (std=0.04)	0.00	0.50	DNA, replication, independent, chromatin, assembly
$\widetilde{A}_{G_2}$	5	0.12 (std=0.10)	0.00	0.18	homeostasis, ion, cellular, calcium, chemical
$\widetilde{A}_{G_2}$	5	0.12 (std=0.10)	0.06	0.28	heterochromatin, assembly, organization, constitutive, negative
$\widetilde{A}_{G_2}$	5	0.12 (std=0.10)	0.21	0.33	regulation, process, positive, biological, conjugation
$\widetilde{A}_{G_3}$	2	0.08 (std=0.03)	0.10	0.54	DNA, cell, cycle, process, mitotic
$\widetilde{A}_{G_3}$	2	0.08 (std=0.03)	0.05	0.56	assembly, heterochromatin, organization, chromatin, cellular
$\widetilde{A}_{G_4}$	3	0.04 (std=0.05)	0.00	0.00	splicing, RNA, mRNA, transesterification, reactions
$\widetilde{A}_{G_4}$	3	0.04 (std=0.05)	0.11	0.05	organization, cellular, component, or, biogenesis
$\widetilde{A}_{G_4}$	3	0.04 (std=0.05)	0.00	0.50	DNA, replication, independent, chromatin, organization
$\widetilde{A}_{G_5}$	5	0.13 (std=0.10)	0.27	0.00	regulation, biosynthetic, process, transcription, templated
$\widetilde{A}_{G_5}$	5	0.13 (std=0.10)	0.23	0.07	regulation, cellular, component, biogenesis, positive
$\widetilde{A}_{G_5}$	5	0.13 (std=0.10)	0.02	0.08	repair, recombination, double, strand, break
$\widetilde{A}_{G_6}$	6	0.22 (std=0.19)	0.42	0.00	negative, regulation, response, stimulus, cell
$\widetilde{A}_{G_6}$	6	0.22 (std=0.19)	0.07	0.00	transport, anion, transmembrane, ion, organic
$\widetilde{A}_{G_6}$	6	0.22 (std=0.19)	0.53	0.15	positive, regulation, cell, cycle, phase
$\widetilde{A}_{G_7}$	5	0.10 (std=0.10)	0.00	0.18	regulation, cytosolic, calcium, ion, concentration
$\widetilde{A}_{G_7}$	5	0.10 (std=0.10)	0.28	0.33	regulation, process, positive, cell, cycle
$\widetilde{A}_{G_7}$	5	0.10 (std=0.10)	0.12	0.36	regulation, assembly, process, organization, actomyosin
$\widetilde{A}_{G_8}$	6	0.15 (std=0.06)	0.13	0.00	chromosome, segregation, attachment, spindle, microtubules
$\widetilde{A}_{G_8}$	6	0.15 (std=0.06)	0.24	0.03	microtubule, organization, transport, based, cytoskeleton
$\widetilde{A}_{G_8}$	6	0.15 (std=0.06)	0.07	0.24	metabolic, process, compound, cellular, DNA

Table A.13. Summary of most unique functional domains for Gracoal embeddings. We report, for each GraCoal embedding used with SAFE with the Fission yeast GI network, the number of functional domains (column 1) the mean paralog ratio (column 2) and the top three most characteristic functional domains (column 5). Lastly, for each functional domain we report the paralog ratio (column 3) and the maximum Jaccard similarity index (JI).

GraCoals functional domain summary for fruit fly

	Num functional domains	Mean paralog ratio	Domain paralog ratio	Domain max JI	Domain description
$\widetilde{A}_{G_0}$	15	0.06 (std=0.07)	0.22	0.00	response, regulation, cuticle, stress, chitin
$\widetilde{A}_{G_0}$	15	0.06 (std=0.07)	0.00	0.00	intrinsic, apoptotic, signaling, pathway, in
$\widetilde{A}_{G_0}$	15	0.06 (std=0.07)	0.00	0.05	dosage, compensation
$\widetilde{A}_{G_1}$	13	0.06 (std=0.05)	0.00	0.00	mitotic, negative, regulation, spindle, checkpoint
$\widetilde{A}_{G_1}$	13	0.06 (std=0.05)	0.11	0.10	regulation, TORC1, signaling
$\widetilde{A}_{G_1}$	13	0.06 (std=0.05)	0.05	0.15	cell, establishment, polarity, organization, maintenance
$\widetilde{A}_{G_2}$	9	0.06 (std=0.04)	0.09	0.16	regulation, signaling, pathway, growth, neuromuscular
$\widetilde{A}_{G_2}$	9	0.06 (std=0.04)	0.07	0.18	regulation, cell, establishment, polarity, organization
$\widetilde{A}_{G_2}$	9	0.06 (std=0.04)	0.03	0.32	regulation, negative, process, cell, metabolic
$\widetilde{A}_{G_3}$	11	0.05 (std=0.03)	0.07	0.02	synaptic, response, external, stimulus, cell
$\widetilde{A}_{G_3}$	11	0.05 (std=0.03)	0.04	0.16	cell, regulation, signaling, pathway, in
$\widetilde{A}_{G_3}$	11	0.05 (std=0.03)	0.11	0.20	regulation, process, positive, cell, cellular
$\widetilde{A}_{G_4}$	10	0.07 (std=0.06)	0.08	0.00	muscle, cell, cellular, homeostasis
$\widetilde{A}_{G_4}$	10	0.07 (std=0.06)	0.08	0.00	synaptic, signaling, trans, anterograde, chemical
$\widetilde{A}_{G_4}$	10	0.07 (std=0.06)	0.00	0.00	olfactory, behavior
$\widetilde{A}_{G_5}$	9	0.05 (std=0.05)	0.03	0.03	guidance, neuron, projection, axon
$\widetilde{A}_{G_5}$	9	0.05 (std=0.05)	0.07	0.16	regulation, positive, process, cell, negative
$\widetilde{A}_{G_5}$	9	0.05 (std=0.05)	0.19	0.21	detection, stimulus, phototransduction, external, light
$\widetilde{A}_{G_6}$	11	0.06 (std=0.05)	0.04	0.20	regulation, signaling, insulin, receptor, pathway
$\widetilde{A}_{G_6}$	11	0.06 (std=0.05)	0.16	0.22	regulation, cascade, morphogenesis, stress, cell
$\widetilde{A}_{G_6}$	11	0.06 (std=0.05)	0.00	0.25	RNA, 3', end, processing
$\widetilde{A}_{G_7}$	10	0.07 (std=0.05)	0.14	0.00	cofactor, metabolic, process
$\widetilde{A}_{G_7}$	10	0.07 (std=0.05)	0.04	0.02	regulation, Notch, signaling, pathway
$\widetilde{A}_{G_7}$	10	0.07 (std=0.05)	0.09	0.21	regulation, cell, organization, projection, morphogenesis
$\widetilde{A}_{G_8}$	4	0.05 (std=0.02)	0.04	0.19	regulation, response, mitochondrion, organization, cellular
$\widetilde{A}_{G_8}$	4	0.05 (std=0.02)	0.02	0.22	regulation, process, positive, templated, transcription
$\widetilde{A}_{G_8}$	4	0.05 (std=0.02)	0.07	0.32	regulation, cell, morphogenesis, positive, process

Table A.14. Summary of most unique functional domains for Gracoal embeddings. We report, for each GraCoal embedding used with SAFE with the Fruit Fruit fly GI network, the number of functional domains (column 1) the mean paralog ratio (column 2) and the top three most characteristic functional domains (column 5). Lastly, for each functional domain we report the paralog ratio (column 3) and the maximum Jaccard similarity index (JI).

GraSprings functional domain summary for *E. coli*

	Num functional domains	Mean paralog ratio	Domain paralog ratio	Domain max JI	Domain description
$\widetilde{A}_{G_0}$	11	0.18 (std=0.11)	0.24	0.00	regulation, cell, morphogenesis, shape, anatomical
$\widetilde{A}_{G_0}$	11	0.18 (std=0.11)	0.08	0.30	metabolic, process, glycosaminoglycan, aminoglycan, peptidoglycan
$\widetilde{A}_{G_0}$	11	0.18 (std=0.11)	0.29	0.40	transport, protein, localization, peptide, establishment
$\widetilde{A}_{G_1}$	8	0.15 (std=0.09)	0.18	0.00	organonitrogen, compound, biosynthetic, process
$\widetilde{A}_{G_1}$	8	0.15 (std=0.09)	0.10	0.44	localization, protein, membrane, cellular, insertion
$\widetilde{A}_{G_1}$	8	0.15 (std=0.09)	0.24	0.83	processing, rRNA, metabolic, process, RNA
$\widetilde{A}_{G_2}$	6	0.16 (std=0.14)	0.00	0.00	glutamate, biosynthetic, process
$\widetilde{A}_{G_2}$	6	0.16 (std=0.14)	0.19	0.45	processing, RNA, metabolic, process, tRNA
$\widetilde{A}_{G_2}$	6	0.16 (std=0.14)	0.00	0.64	process, biosynthetic, cell, wall, macromolecule
$\widetilde{A}_{G_3}$	9	0.12 (std=0.09)	0.00	0.31	process, cytokinesis, cell, cycle, FtsZ
$\widetilde{A}_{G_3}$	9	0.12 (std=0.09)	0.15	0.40	protein, transport, localization, intracellular, transmembrane
$\widetilde{A}_{G_3}$	9	0.12 (std=0.09)	0.07	0.40	process, localization, biosynthetic, metabolic, cellular
$\widetilde{A}_{G_4}$	8	0.15 (std=0.13)	0.16	0.20	process, biosynthetic, lipopolysaccharide, metabolic, organic
$\widetilde{A}_{G_4}$	8	0.15 (std=0.13)	0.09	0.22	process, biosynthetic, metabolic, polysaccharide, cellular
$\widetilde{A}_{G_4}$	8	0.15 (std=0.13)	0.46	0.53	homeostasis, ion, transport, cellular, metal
$\widetilde{A}_{G_5}$	6	0.13 (std=0.10)	0.14	0.38	process, biosynthetic, metabolic, localization, cellular
$\widetilde{A}_{G_5}$	6	0.13 (std=0.10)	0.04	0.50	response, heat
$\widetilde{A}_{G_5}$	6	0.13 (std=0.10)	0.13	0.67	metabolic, process, processing, RNA, ncRNA
$\widetilde{A}_{G_6}$	5	0.16 (std=0.11)	0.20	0.33	protein, folding, ribonucleoprotein, complex, 'de
$\widetilde{A}_{G_6}$	5	0.16 (std=0.11)	0.08	0.38	process, biosynthetic, metabolic, acid, cell
$\widetilde{A}_{G_6}$	5	0.16 (std=0.11)	0.15	0.65	metabolic, process, response, compound, DNA
$\widetilde{A}_{G_7}$	7	0.14 (std=0.11)	0.16	0.30	process, biosynthetic, metabolic, localization, acid
$\widetilde{A}_{G_7}$	7	0.14 (std=0.11)	0.12	0.45	process, metabolic, biosynthetic, lipid, cellular
$\widetilde{A}_{G_7}$	7	0.14 (std=0.11)	0.00	0.71	process, biosynthetic, cell, macromolecule, wall
$\widetilde{A}_{G_8}$	6	0.18 (std=0.11)	0.25	0.61	process, biosynthetic, metabolic, cellular, homeostasis
$\widetilde{A}_{G_8}$	6	0.18 (std=0.11)	0.13	0.75	processing, ncRNA, tRNA, metabolic, process
$\widetilde{A}_{G_8}$	6	0.18 (std=0.11)	0.17	0.79	metabolic, process, compound, cellular, macromolecule

Table A.15. Summary of most unique functional domains for GraSpring embeddings. We report, for each GraSpring embedding used with SAFE with the *E. coli* GI network, the number of functional domains (column 1) the mean paralog ratio (column 2) and the top three most characteristic functional domains (column 5). Lastly, for each functional domain we report the paralog ratio (column 3) and the maximum Jaccard similarity index (JI).

GraSprings functional domain summary for fission yeast

	Num functional domains	Mean paralog ratio	Domain paralog ratio	Domain max JI	Domain description
$\widetilde{A}_{G_0}$	4	0.12 (std=0.10)	0.00	0.07	DNA, replication, independent, chromatin, organization
$\widetilde{A}_{G_0}$	4	0.12 (std=0.10)	0.22	0.41	cell, cycle, regulation, mitotic, checkpoint
$\widetilde{A}_{G_0}$	4	0.12 (std=0.10)	0.05	0.67	process, metabolic, DNA, cellular, compound
$\widetilde{A}_{G_1}$	5	0.20 (std=0.16)	0.48	0.00	positive, regulation, cell, cycle, process
$\widetilde{A}_{G_1}$	5	0.20 (std=0.16)	0.17	0.33	regulation, gene, expression
$\widetilde{A}_{G_1}$	5	0.20 (std=0.16)	0.05	0.36	checkpoint, signaling, mitotic, DNA, metabolic
$\widetilde{A}_{G_2}$	5	0.06 (std=0.07)	0.00	0.00	regulation, cytosolic, calcium, ion, concentration
$\widetilde{A}_{G_2}$	5	0.06 (std=0.07)	0.02	0.38	DNA, replication, cell, cycle, biosynthetic
$\widetilde{A}_{G_2}$	5	0.06 (std=0.07)	0.02	0.46	heterochromatin, assembly, pericentric, organization, chromatin
$\widetilde{A}_{G_3}$	4	0.11 (std=0.09)	0.03	0.28	process, DNA, metabolic, recombination, checkpoint
$\widetilde{A}_{G_3}$	4	0.11 (std=0.09)	0.07	0.29	process, cell, cycle, DNA, replication
$\widetilde{A}_{G_3}$	4	0.11 (std=0.09)	0.09	0.50	chromatin, regulation, assembly, heterochromatin, organization
$\widetilde{A}_{G_4}$	3	0.12 (std=0.06)	0.09	0.14	cellular, response, DNA, damage, stimulus
$\widetilde{A}_{G_4}$	3	0.12 (std=0.06)	0.06	0.32	metabolic, process, compound, cellular, nitrogen
$\widetilde{A}_{G_4}$	3	0.12 (std=0.06)	0.20	0.80	regulation, process, biological, metabolic, gene
$\widetilde{A}_{G_5}$	5	0.14 (std=0.08)	0.11	0.26	cell, regulation, DNA, cycle, checkpoint
$\widetilde{A}_{G_5}$	5	0.14 (std=0.08)	0.07	0.29	cell, cycle, process, meiotic, meiosis
$\widetilde{A}_{G_5}$	5	0.14 (std=0.08)	0.27	0.46	regulation, biosynthetic, process, transcription, RNA
$\widetilde{A}_{G_6}$	4	0.09 (std=0.10)	0.00	0.38	DNA, replication, initiation, cell, cycle
$\widetilde{A}_{G_6}$	4	0.09 (std=0.10)	0.04	0.50	DNA, checkpoint, signaling, metabolic, process
$\widetilde{A}_{G_6}$	4	0.09 (std=0.10)	0.08	0.59	regulation, heterochromatin, assembly, process, negative
$\widetilde{A}_{G_7}$	4	0.12 (std=0.08)	0.12	0.03	regulation, positive, septum, assembly, mitotic
$\widetilde{A}_{G_7}$	4	0.12 (std=0.08)	0.13	0.33	DNA, cell, cycle, regulation, mitotic
$\widetilde{A}_{G_7}$	4	0.12 (std=0.08)	0.00	0.38	heterochromatin, pericentric, assembly, organization, constitutive
$\widetilde{A}_{G_8}$	4	0.09 (std=0.09)	0.02	0.03	actomyosin, actin, contractile, ring, organization
$\widetilde{A}_{G_8}$	4	0.09 (std=0.09)	0.07	0.59	process, metabolic, DNA, checkpoint, signaling
$\widetilde{A}_{G_8}$	4	0.09 (std=0.09)	0.03	0.67	heterochromatin, assembly, chromatin, organization, regulation

Table A.16. Summary of most unique functional domains for GraSpring embeddings. We report, for each GraSpring embedding used with SAFE with the fission yeast GI network, the number of functional domains (column 1) the mean paralog ratio (column 2) and the top three most characteristic functional domains (column 5). Lastly, for each functional domain we report the paralog ratio (column 3) and the maximum Jaccard similarity index (JI).

GraSprings functional domain summary for fruit fly

	Num functional domains	Mean paralog ratio	Domain paralog ratio	Domain max JI	Domain description
$\widetilde{A}_{G_0}$	12	0.06 (std=0.09)	0.00	0.00	copper, ion, homeostasis, transport, cellular
$\widetilde{A}_{G_0}$	12	0.06 (std=0.09)	0.06	0.17	positive, regulation, establishment, morphogenesis, axon
$\widetilde{A}_{G_0}$	12	0.06 (std=0.09)	0.11	0.21	regulation, cell, actin, organization, filament
$\widetilde{A}_{G_1}$	6	0.08 (std=0.07)	0.03	0.41	regulation, response, signaling, cellular, cell
$\widetilde{A}_{G_1}$	6	0.08 (std=0.07)	0.07	0.67	DNA, meiotic, recombination, repair, double
$\widetilde{A}_{G_1}$	6	0.08 (std=0.07)	0.05	0.71	regulation, cell, process, positive, negative
$\widetilde{A}_{G_2}$	14	0.06 (std=0.06)	0.04	0.00	regulation, signaling, pathway, BMP, negative
$\widetilde{A}_{G_2}$	14	0.06 (std=0.06)	0.01	0.11	regulation, cell, negative, growth, positive
$\widetilde{A}_{G_2}$	14	0.06 (std=0.06)	0.04	0.19	establishment, cell, polarity, hair, regulation
$\widetilde{A}_{G_3}$	6	0.04 (std=0.03)	0.04	0.38	DNA, karyosome, formation, process, repair
$\widetilde{A}_{G_3}$	6	0.04 (std=0.03)	0.00	0.47	process, pigment, biosynthetic, metabolic, ommochrome
$\widetilde{A}_{G_3}$	6	0.04 (std=0.03)	0.04	0.51	process, cellular, response, cell, catabolic
$\widetilde{A}_{G_4}$	7	0.07 (std=0.07)	0.08	0.30	regulation, cell, organization, projection, morphogenesis
$\widetilde{A}_{G_4}$	7	0.07 (std=0.07)	0.03	0.40	response, cellular, regulation, cell, growth
$\widetilde{A}_{G_4}$	7	0.07 (std=0.07)	0.00	0.48	process, pigment, biosynthetic, metabolic, regulation
$\widetilde{A}_{G_5}$	6	0.05 (std=0.03)	0.07	0.37	regulation, response, cellular, process, stress
$\widetilde{A}_{G_5}$	6	0.05 (std=0.03)	0.08	0.38	cell, regulation, organization, establishment, polarity
$\widetilde{A}_{G_5}$	6	0.05 (std=0.03)	0.00	0.44	mitochondrion, organization, cell, ectopic, germ
$\widetilde{A}_{G_6}$	8	0.08 (std=0.10)	0.07	0.21	regulation, organization, cell, actin, maintenance
$\widetilde{A}_{G_6}$	8	0.08 (std=0.10)	0.02	0.36	process, metabolic, response, cellular, cell
$\widetilde{A}_{G_6}$	8	0.08 (std=0.10)	0.01	0.48	process, RNA, pigment, gene, silencing
$\widetilde{A}_{G_7}$	11	0.07 (std=0.07)	0.00	0.00	regulation, membrane, potential
$\widetilde{A}_{G_7}$	11	0.07 (std=0.07)	0.12	0.04	response, oxidative, stress, stimulus
$\widetilde{A}_{G_7}$	11	0.07 (std=0.07)	0.11	0.38	regulation, organization, cell, actin, filament
$\widetilde{A}_{G_8}$	11	0.03 (std=0.04)	0.02	0.02	recognition, axon, guidance, neuron, choice
$\widetilde{A}_{G_8}$	11	0.03 (std=0.04)	0.00	0.11	positive, regulation, apoptotic, process, programmed
$\widetilde{A}_{G_8}$	11	0.03 (std=0.04)	0.00	0.15	cell, death, maturation, negative, regulation

Table A.17. Summary of most unique functional domains for GraSpring embeddings. We report, for each GraSpring embedding used with SAFE with the fruit fly GI network, the number of functional domains (column 1) the mean paralog ratio (column 2) and the top three most characteristic functional domains (column 5). Lastly, for each functional domain we report the paralog ratio (column 3) and the maximum Jaccard similarity index (JI).

Graphlet Spectral functional domain summary for *E. coli*

	Num functional domains	Mean paralog ratio	Domain paralog ratio	Domain max JI	Domain description
$\widetilde{A}_{G_0}$	6	0.19 (std=0.09)	0.15	0.00	sulfur, cluster, assembly, iron, metallo
$\widetilde{A}_{G_0}$	6	0.19 (std=0.09)	0.04	0.05	response, stimulus, cellular, DNA, extracellular
$\widetilde{A}_{G_0}$	6	0.19 (std=0.09)	0.22	0.28	localization, transport, protein, membrane, oxidation
$\widetilde{A}_{G_1}$	5	0.20 (std=0.06)	0.18	0.21	process, biosynthetic, phospholipid, lipid, cellular
$\widetilde{A}_{G_1}$	5	0.20 (std=0.06)	0.19	0.56	RNA, ncRNA, processing, metabolic, process
$\widetilde{A}_{G_1}$	5	0.20 (std=0.06)	0.32	0.63	process, biosynthetic, transport, homeostasis, cellular
$\widetilde{A}_{G_2}$	6	0.24 (std=0.13)	0.18	0.20	ncRNA, metabolic, process
$\widetilde{A}_{G_2}$	6	0.24 (std=0.13)	0.39	0.24	homeostasis, ion, cellular, metal, transition
$\widetilde{A}_{G_2}$	6	0.24 (std=0.13)	0.45	0.28	transport, export, xenobiotic, acid, oxidation
$\widetilde{A}_{G_3}$	4	0.22 (std=0.07)	0.18	0.44	localization, cellular, protein, membrane, macromolecule
$\widetilde{A}_{G_3}$	4	0.22 (std=0.07)	0.14	0.44	metabolic, process, cellular, macromolecule, DNA
$\widetilde{A}_{G_3}$	4	0.22 (std=0.07)	0.33	0.65	metabolic, process, rRNA, processing, modification
$\widetilde{A}_{G_4}$	6	0.17 (std=0.11)	0.21	0.04	transport, cell, response, iron, division
$\widetilde{A}_{G_4}$	6	0.17 (std=0.11)	0.40	0.26	transport, localization, process, establishment, protein
$\widetilde{A}_{G_4}$	6	0.17 (std=0.11)	0.08	0.44	membrane, protein, localization, insertion, cellular
$\widetilde{A}_{G_5}$	5	0.15 (std=0.10)	0.14	0.33	metabolic, process, DNA, cellular, macromolecule
$\widetilde{A}_{G_5}$	5	0.15 (std=0.10)	0.00	0.46	process, biosynthetic, macromolecule, cell, wall
$\widetilde{A}_{G_5}$	5	0.15 (std=0.10)	0.29	0.63	process, transport, biosynthetic, cellular, metabolic
$\widetilde{A}_{G_6}$	4	0.22 (std=0.05)	0.17	0.38	process, biosynthetic, metabolic, isopentenyl, diphosphate
$\widetilde{A}_{G_6}$	4	0.22 (std=0.05)	0.23	0.66	process, homeostasis, cellular, biosynthetic, ion
$\widetilde{A}_{G_6}$	4	0.22 (std=0.05)	0.18	0.71	metabolic, process, cellular, compound, nucleic
$\widetilde{A}_{G_7}$	4	0.18 (std=0.03)	0.21	0.50	processing, rRNA, metabolic, process, tRNA
$\widetilde{A}_{G_7}$	4	0.18 (std=0.03)	0.18	0.67	metabolic, process, compound, cellular, heterocycle
$\widetilde{A}_{G_7}$	4	0.18 (std=0.03)	0.18	0.74	process, biosynthetic, metabolic, cellular, macromolecule
$\widetilde{A}_{G_8}$	4	0.16 (std=0.08)	0.28	0.00	cellular, metabolic, process
$\widetilde{A}_{G_8}$	4	0.16 (std=0.08)	0.17	0.40	ubiquinone, process, biosynthetic, metabolic
$\widetilde{A}_{G_8}$	4	0.16 (std=0.08)	0.07	0.46	process, biosynthetic, macromolecule, metabolic, cell

Table A.18. Summary of most unique functional domains for graphlet Spectral embeddings. We report, for each graphlet Spectral embedding used with SAFE with the *E. coli* GI network, the number of functional domains (column 1) the mean paralog ratio (column 2) and the top three most characteristic functional domains (column 5). Lastly, for each functional domain we report the paralog ratio (column 3) and the maximum Jaccard similarity index (JI).

Graphlet Spectral functional domain summary for fission yeast

	Num functional domains	Mean paralog ratio	Domain paralog ratio	Domain max JI	Domain description
$\widetilde{A}_{G_0}$	5	0.04 (std=0.03)	0.04	0.00	actin, cytoskeleton, organization, cortical, filament
$\widetilde{A}_{G_0}$	5	0.04 (std=0.03)	0.03	0.00	DNA, organization, replication, independent, chromatin
$\widetilde{A}_{G_0}$	5	0.04 (std=0.03)	0.10	0.10	checkpoint, signaling, response, DNA, integrity
$\widetilde{A}_{G_1}$	2	0.05 (std=0.05)	0.11	0.19	process, metabolic, DNA, mitotic, cell
$\widetilde{A}_{G_1}$	2	0.05 (std=0.05)	0.00	0.40	DNA, replication, initiation, cell, cycle
$\widetilde{A}_{G_2}$	2	0.11 (std=0.04)	0.15	0.17	microtubule, based, process
$\widetilde{A}_{G_2}$	2	0.11 (std=0.04)	0.07	0.19	metabolic, process, cellular, DNA, compound
$\widetilde{A}_{G_4}$	3	0.08 (std=0.06)	0.14	0.00	catabolic, process, dependent, protein, macromolecule
$\widetilde{A}_{G_4}$	3	0.08 (std=0.06)	0.10	0.40	cell, cycle, DNA, replication, initiation
$\widetilde{A}_{G_4}$	3	0.08 (std=0.06)	0.00	0.88	splicing, RNA, mRNA, transesterification, reactions
$\widetilde{A}_{G_5}$	2	0.21 (std=0.07)	0.29	0.00	nucleotide, excision, repair
$\widetilde{A}_{G_5}$	2	0.21 (std=0.07)	0.14	0.12	cell, cycle, process

Table A.19. Summary of most unique functional domains for graphlet Spectral embeddings. We report, for each graphlet Spectral embedding used with SAFE with the fission yeast GI network, the number of functional domains (column 1) the mean paralog ratio (column 2) and the top three most characteristic functional domains (column 5). Lastly, for each functional domain we report the paralog ratio (column 3) and the maximum Jaccard similarity index (JI).

Graphlet Spectral functional domain summary for fruit fly

	Num functional domains	Mean paralog ratio	Domain paralog ratio	Domain max JI	Domain description
$\widetilde{A}_{G_0}$	6	0.03 (std=0.03)	0.00	0.00	positive, regulation, feeding, behavior
$\widetilde{A}_{G_0}$	6	0.03 (std=0.03)	0.03	0.12	mating, behavior, cell, male, signaling
$\widetilde{A}_{G_0}$	6	0.03 (std=0.03)	0.03	0.19	response, cellular, cell, stimulus, death
$\widetilde{A}_{G_1}$	11	0.08 (std=0.07)	0.18	0.00	response, fungus, regulation, Toll, signaling
$\widetilde{A}_{G_1}$	11	0.08 (std=0.07)	0.19	0.11	detection, stimulus, abiotic, external, light
$\widetilde{A}_{G_1}$	11	0.08 (std=0.07)	0.05	0.12	synaptic, signaling, male, behavior, trans
$\widetilde{A}_{G_2}$	10	0.11 (std=0.09)	0.08	0.00	regulation, response, external, stimulus
$\widetilde{A}_{G_2}$	10	0.11 (std=0.09)	0.00	0.00	centrosome, cycle
$\widetilde{A}_{G_2}$	10	0.11 (std=0.09)	0.13	0.15	ion, homeostasis, metal, cellular, transition
$\widetilde{A}_{G_3}$	11	0.09 (std=0.10)	0.14	0.00	positive, regulation, filopodium, assembly
$\widetilde{A}_{G_3}$	11	0.09 (std=0.10)	0.00	0.00	negative, regulation, cascade, stress, activated
$\widetilde{A}_{G_3}$	11	0.09 (std=0.10)	0.00	0.00	behavior, rhythmic, circadian, locomotor, rhythm
$\widetilde{A}_{G_4}$	9	0.06 (std=0.06)	0.12	0.00	telomere, maintenance, organization
$\widetilde{A}_{G_4}$	9	0.06 (std=0.06)	0.00	0.00	pole, plasm, mRNA, localization
$\widetilde{A}_{G_4}$	9	0.06 (std=0.06)	0.03	0.07	negative, defense, response, Gram, bacterium
$\widetilde{A}_{G_5}$	12	0.05 (std=0.03)	0.08	0.00	response, immune, mucosal, osmotic, stress
$\widetilde{A}_{G_5}$	12	0.05 (std=0.03)	0.02	0.00	regulation, gene, expression, epigenetic, dosage
$\widetilde{A}_{G_5}$	12	0.05 (std=0.03)	0.01	0.00	anterior, posterior, axis, specification, chromatin
$\widetilde{A}_{G_6}$	8	0.06 (std=0.06)	0.04	0.00	fusion, syncytium, formation, cell, actin
$\widetilde{A}_{G_6}$	8	0.06 (std=0.06)	0.18	0.02	cascade, tumor, necrosis, factor, mediated
$\widetilde{A}_{G_6}$	8	0.06 (std=0.06)	0.12	0.04	morphogenesis, regulation, anatomical, structure, embryonic
$\widetilde{A}_{G_7}$	12	0.04 (std=0.03)	0.00	0.00	metabolic, process, catecholamine, dopamine, ammonium
$\widetilde{A}_{G_7}$	12	0.04 (std=0.03)	0.00	0.00	sex, determination
$\widetilde{A}_{G_7}$	12	0.04 (std=0.03)	0.02	0.01	negative, regulation, process, metabolic, macromolecule
$\widetilde{A}_{G_8}$	1	0.08 (std=0.00)	0.08	0.14	cell, cycle, process

Table A.20. Summary of most unique functional domains for graphlet Spectral embeddings. We report, for each graphlet Spectral embedding used with SAFE with the fruit fly GI network, the number of functional domains (column 1) the mean paralog ratio (column 2) and the top three most characteristic functional domains (column 5). Lastly, for each functional domain we report the paralog ratio (column 3) and the maximum Jaccard similarity index (JI).

A.3 Enrichment statistics PPI networks

In this section, we summarize the results obtained when using SAFE with the different graphlet based embedding algorithms. That is, the percentages of genes that have at least one annotation enriched in their neighborhood and the percentages of enriched annotations for all our PPI molecular networks across different annotations.

Gene ontology biological processes

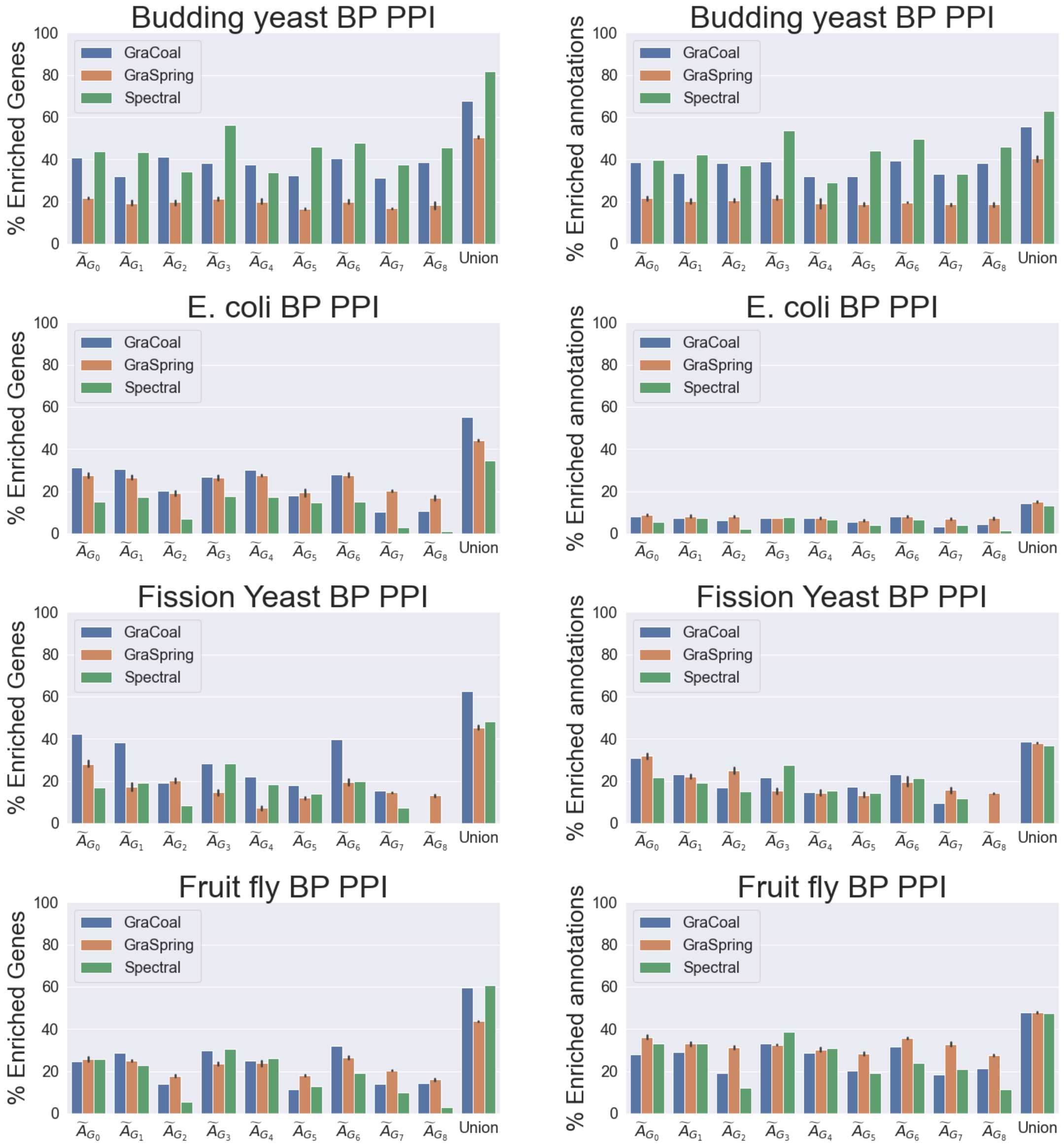


Figure A.21. SAFE GO-BP enrichment analysis for the PPI networks, Part 1. On the y-axis, we show the percentages of genes that have at least one annotation enriched in their neighborhood (left) and the percentages of enriched annotations (right). On the x-axis, we show each of the embedding algorithms considered (legend) applied on different types of graphlet adjacencies. Graphlet adjacency 'Union' (x-axis, far right) considers the union of the enriched genes and enriched annotations across all graphlet adjacencies, i.e., $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$. The error bars for Spring embedding indicate the standard deviation across ten runs.

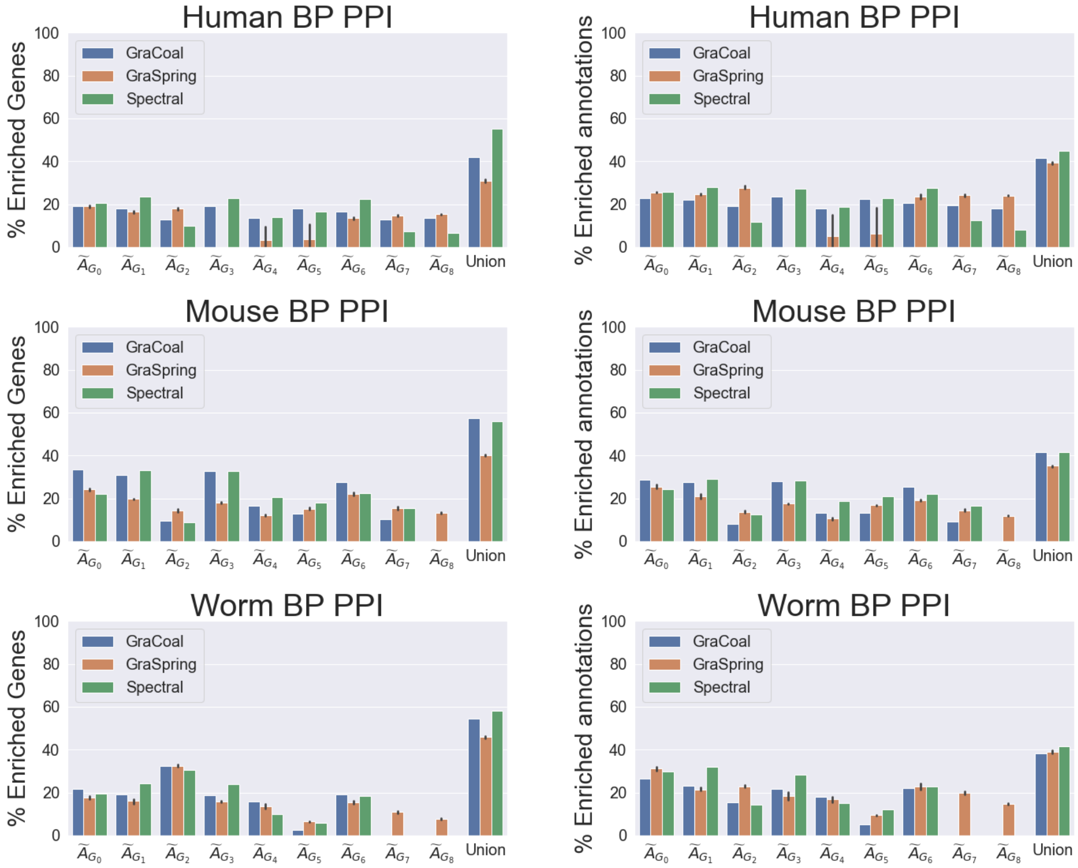


Figure A.21. SAFE GO-BP enrichment analysis for the PPI networks, Part 2. On the y-axis, we show the percentages of genes that have at least one annotation enriched in their neighborhood (left) and the percentages of enriched annotations (right). On the x-axis, we show each of the embedding algorithms considered (legend) applied on different types of graphlet adjacencies. Graphlet adjacency ’Union’ (x-axis, far right) considers the union of the enriched genes and enriched annotations across all graphlet adjacencies, i.e., $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$. The error bars for Spring embedding indicate the standard deviation across ten runs.

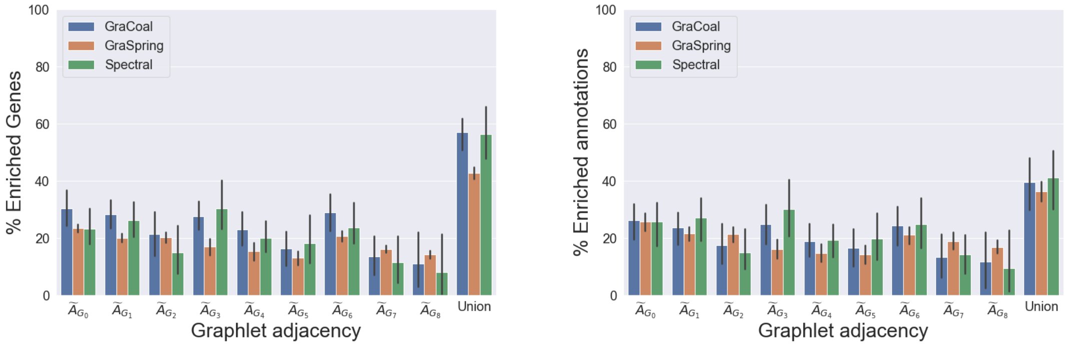


Figure A.22. SAFE GO-BP average enrichment statistics for the PPI molecular networks. Average over all PPI networks for the different types of underlying graphlet adjacencies, i.e., $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$. On the y-axis, we show the percentages of genes that have at least one annotation enriched in their neighborhood (left) and the percentages of enriched annotations (right). On the x-axis, we show each of the embedding algorithms considered (legend) applied on different types of graphlet adjacencies. Graphlet adjacency 'Union' (x-axis, far right) considers the union of the enriched genes and enriched annotations across all graphlet adjacencies.

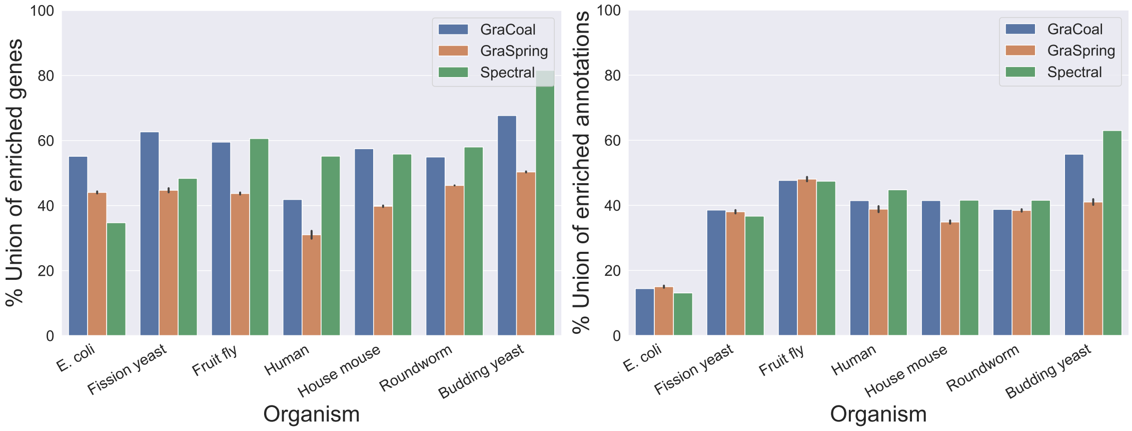


Figure A.23. SAFE GO-BP enrichment analysis for PPI networks. For the PPI networks of our seven species (x-axis), we show the percentage of enriched genes (y-axis) and percentage of enriched annotations for each of the embedding algorithms considered (legend). The error bars in the case of GraSpring embedding indicate the standard deviation across the ten randomised runs.

Gene ontology cellular components

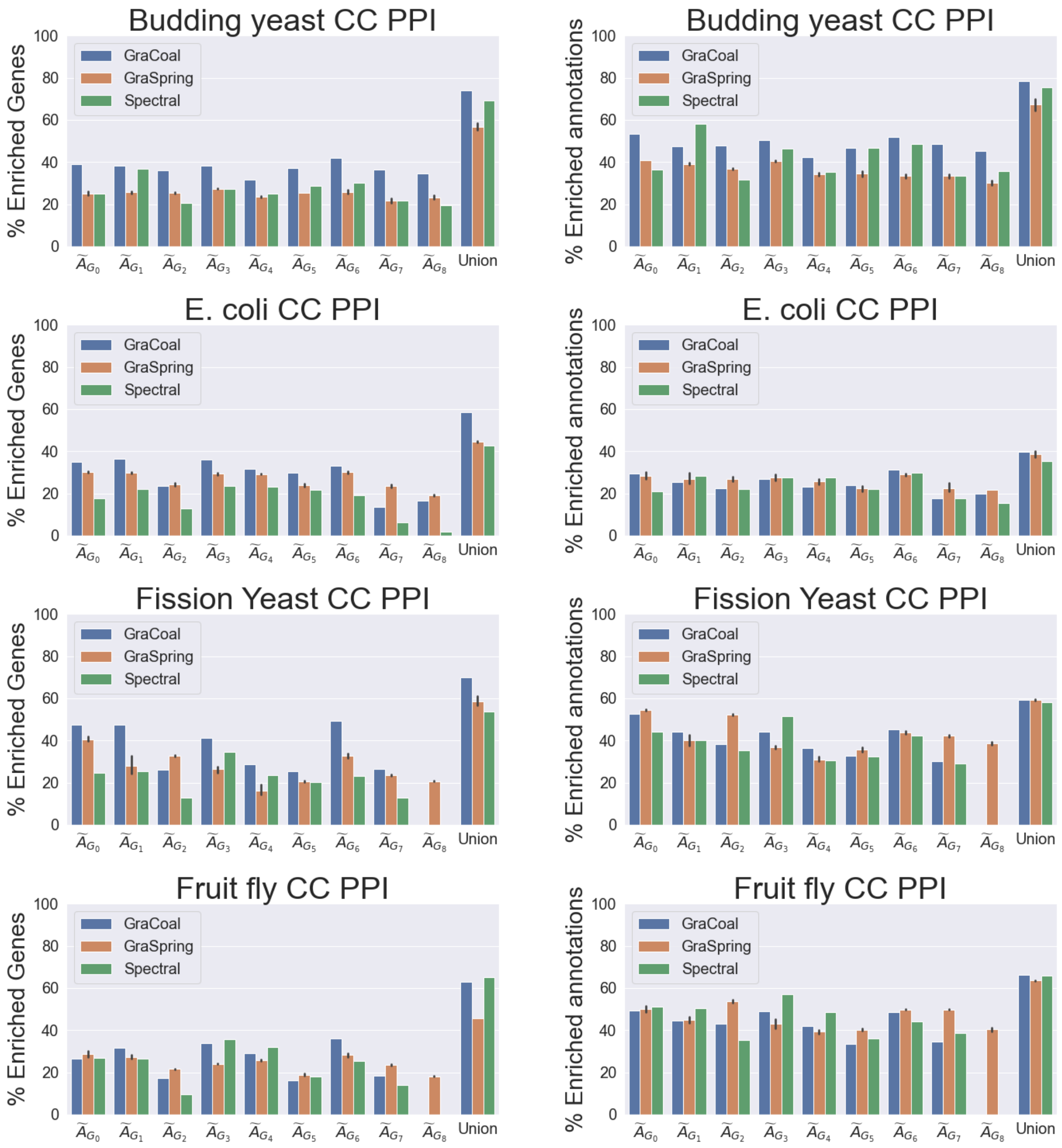


Figure A.24. SAFE GO-CC enrichment analysis for the PPI networks, Part 1. On the y-axis, we show the percentages of genes that have at least one annotation enriched in their neighborhood (left) and the percentages of enriched annotations (right). On the x-axis, we show each of the embedding algorithms considered (legend) applied on different types of graphlet adjacencies. Graphlet adjacency 'Union' (x-axis, far right) considers the union of the enriched genes and enriched annotations across all graphlet adjacencies, i.e., $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$. The error bars for Spring embedding indicate the standard deviation across ten runs. From top to bottom: Budding yeast, *E. coli*, Fission yeast and Fruit fly, respectively.

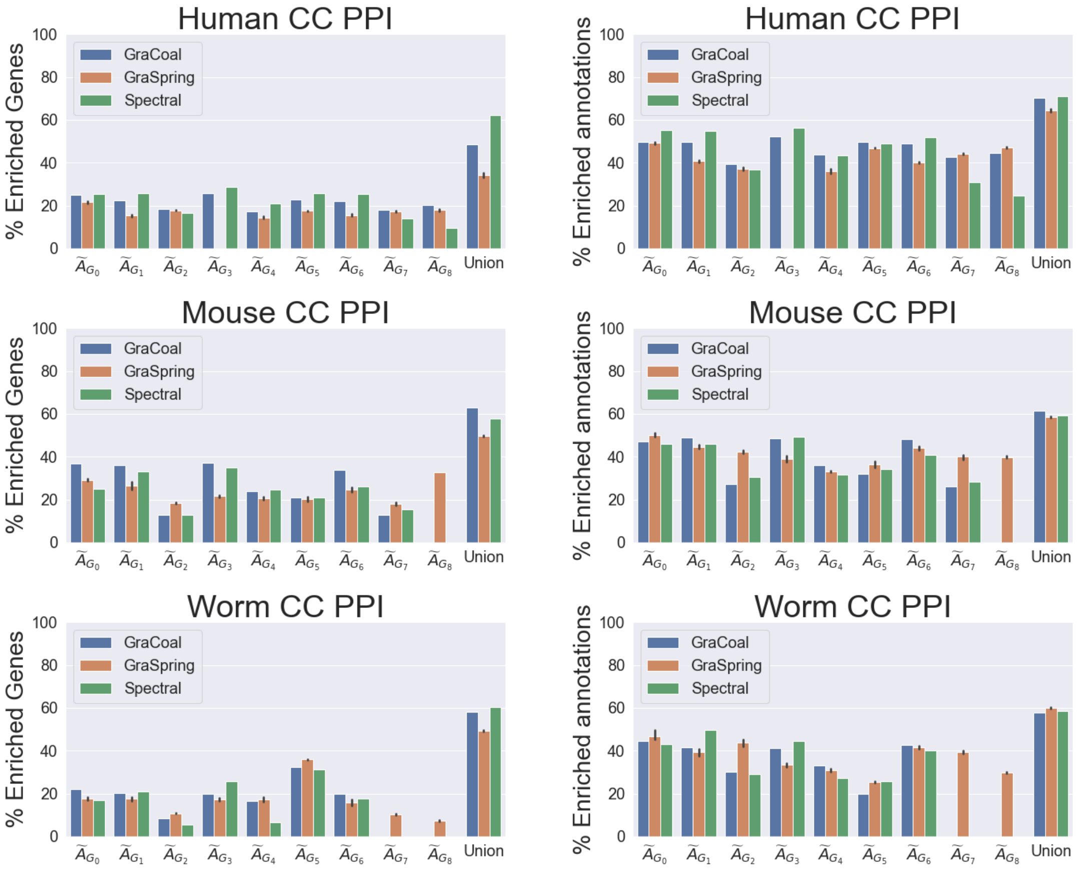


Figure A.24. SAFE GO-CC enrichment analysis for the PPI networks, Part 2. On the y-axis, we show the percentages of genes that have at least one annotation enriched in their neighborhood (left) and the percentages of enriched annotations (right). On the x-axis, we show each of the embedding algorithms considered (legend) applied on different types of graphlet adjacencies. Graphlet adjacency 'Union' (x-axis, far right) considers the union of the enriched genes and enriched annotations across all graphlet adjacencies, i.e., $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$. The error bars for Spring embedding indicate the standard deviation across ten runs. From top to bottom: human, House mouse and Roundworm,.

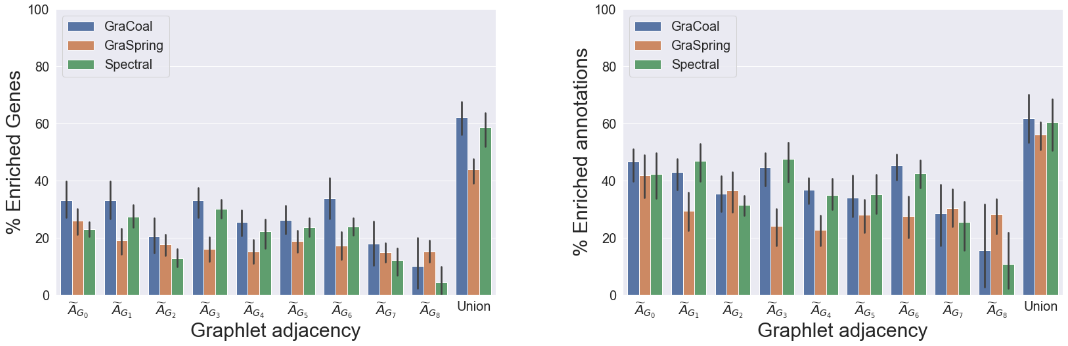


Figure A.25. SAFE GO-CC average enrichment statistics for the PPI molecular networks. Average over all PPI networks for the different types of underlying graphlet adjacencies, i.e., $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$. On the y-axis, we show the percentages of genes that have at least one annotation enriched in their neighborhood (left) and the percentages of enriched annotations (right). On the x-axis, we show each of the embedding algorithms considered (legend) applied on different types of graphlet adjacencies. Graphlet adjacency 'Union' (x-axis, far right) considers the union of the enriched genes and enriched annotations across all graphlet adjacencies.

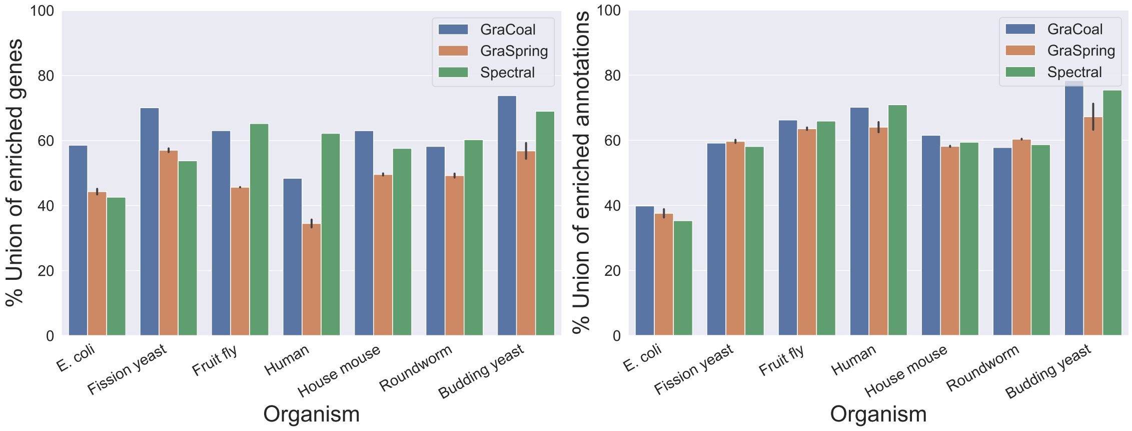


Figure A.26. SAFE GO-CC enrichment analysis for PPI networks. For the PPI networks of our seven species (x-axis), we show the percentage of enriched genes (y-axis) and percentage of enriched annotations for each of the embedding algorithms considered (legend). The error bars in the case of GraSpring embedding indicate the standard deviation across the ten randomised runs.

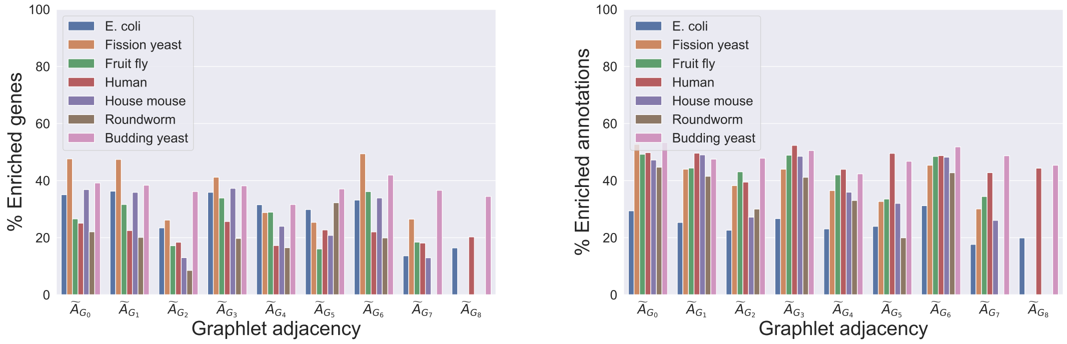


Figure A.27. SAFE GO-CC enrichment analysis comparing GraCoals in PPI networks. For the PPI networks of our seven species (legend), we show, on the y-axis, the percentage of enriched genes (left) and the percentage of enriched annotations (right) for each of the different GraCoal embeddings (x-axis).

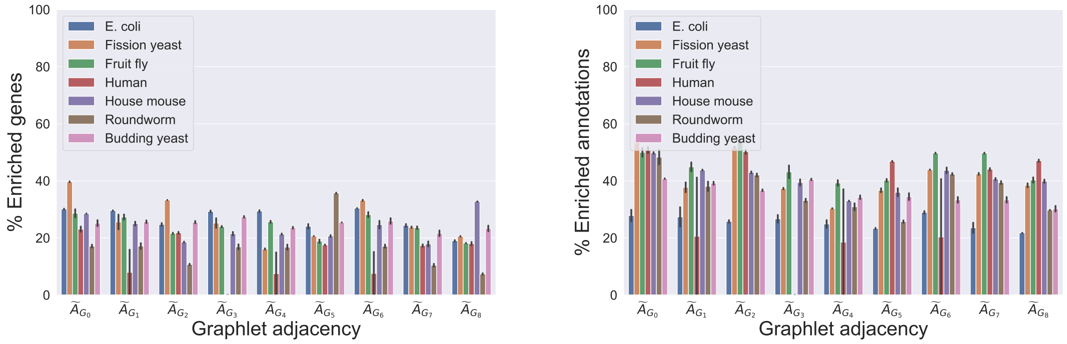


Figure A.28. SAFE GO-CC enrichment analysis comparing GraSprings in PPI networks. For the PPI networks of our seven species (legend), we show, on the y-axis, the percentage of enriched genes (left) and the percentage of enriched annotations (right) for each of the different GraSpring embeddings (x-axis).

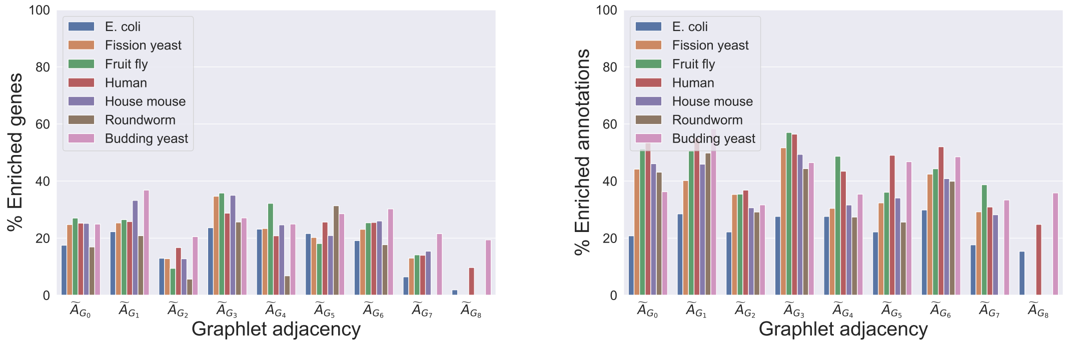


Figure A.29. SAFE GO-CC enrichment analysis comparing Spectrals in PPI networks. For the PPI networks of our seven species (legend), we show, on the y-axis, the percentage of enriched genes (left) and the percentage of enriched annotations (right) for each of the different Spectral embeddings (x-axis).

Gene ontology molecular functions

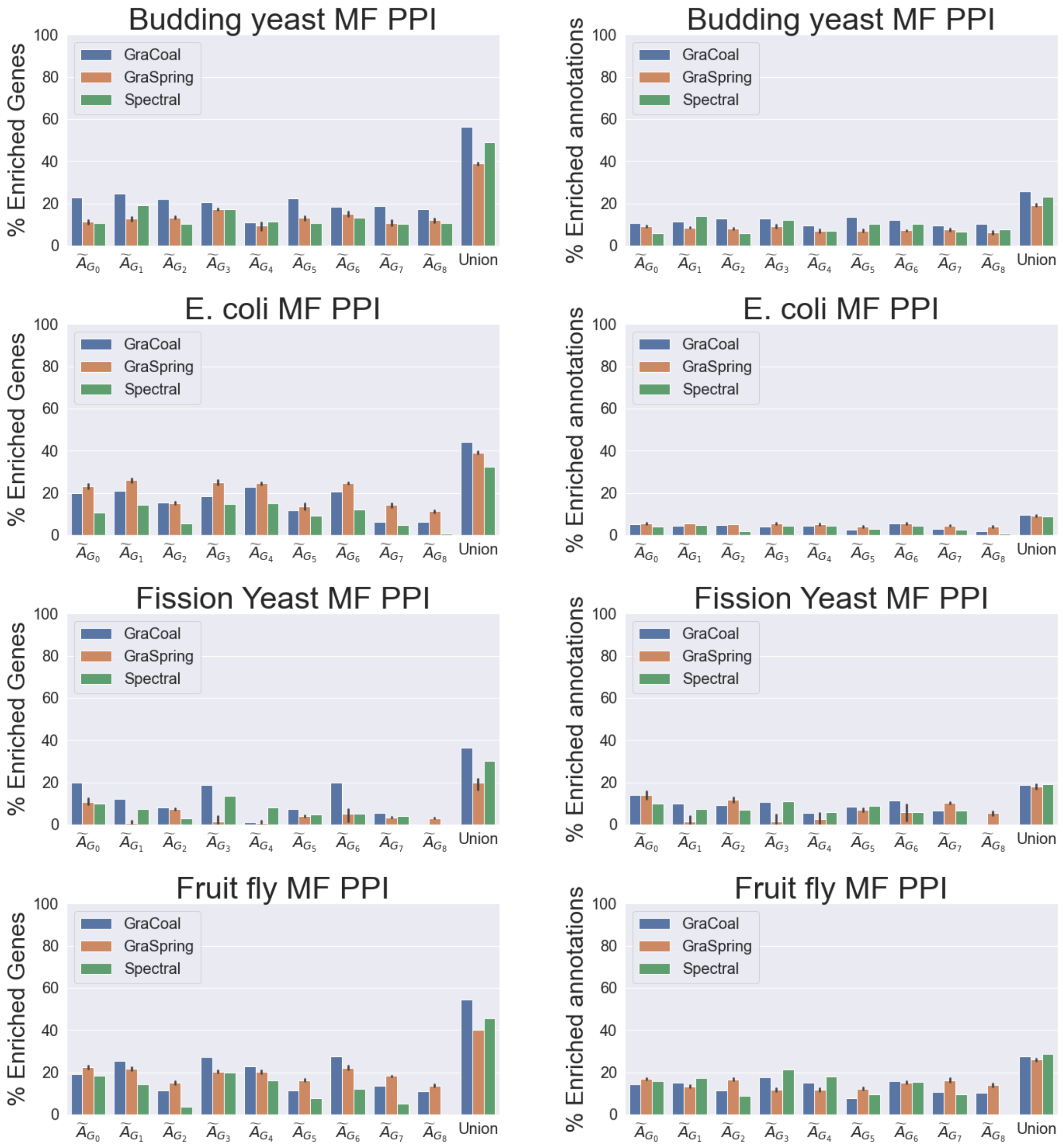


Figure A.30. SAFE GO-MF enrichment analysis for the PPI networks, Part 1. On the y-axis, we show the percentages of genes that have at least one annotation enriched in their neighborhood (left) and the percentages of enriched annotations (right). On the x-axis, we show each of the embedding algorithms considered (legend) applied on different types of graphlet adjacencies. Graphlet adjacency 'Union' (x-axis, far right) considers the union of the enriched genes and enriched annotations across all graphlet adjacencies, i.e., $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$. The error bars for Spring embedding indicate the standard deviation across ten runs.

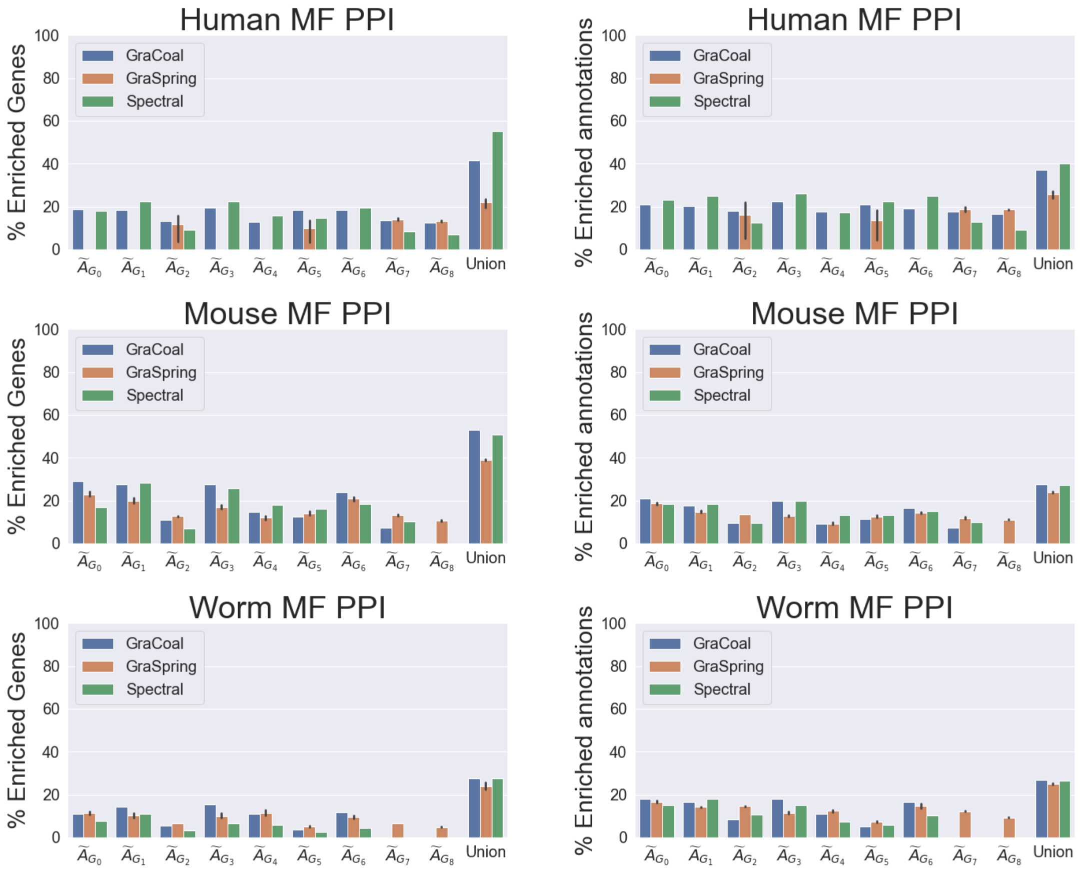


Figure A.30. SAFE GO-MF enrichment analysis for the PPI networks, Part 1. On the y-axis, we show the percentages of genes that have at least one annotation enriched in their neighborhood (left) and the percentages of enriched annotations (right). On the x-axis, we show each of the embedding algorithms considered (legend) applied on different types of graphlet adjacencies. Graphlet adjacency 'Union' (x-axis, far right) considers the union of the enriched genes and enriched annotations across all graphlet adjacencies, i.e., $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$. The error bars for Spring embedding indicate the standard deviation across ten runs.

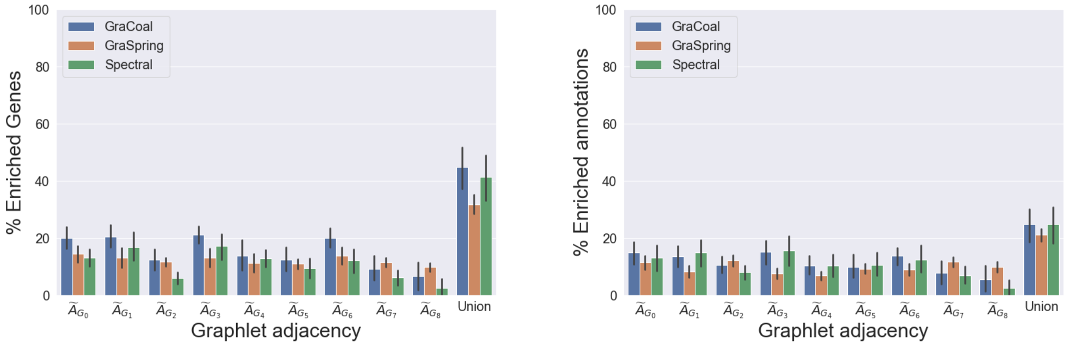


Figure A.31. SAFE GO-MF average enrichment statistics for the PPI molecular networks. Average over all PPI networks for the different types of underlying graphlet adjacencies, i.e., $\widetilde{A}_{G_0}$ to $\widetilde{A}_{G_8}$. On the y-axis, we show the percentages of genes that have at least one annotation enriched in their neighborhood (left) and the percentages of enriched annotations (right). On the x-axis, we show each of the embedding algorithms considered (legend) applied on different types of graphlet adjacencies. Graphlet adjacency 'Union' (x-axis, far right) considers the union of the enriched genes and enriched annotations across all graphlet adjacencies.

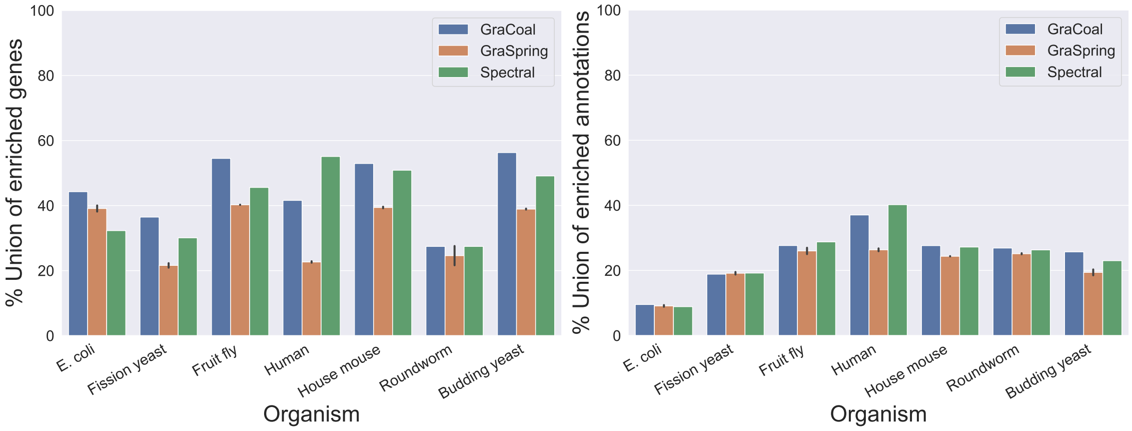


Figure A.32. SAFE GO-MF enrichment analysis for PPI networks. For the PPI networks of our seven species (x-axis), we show the percentage of enriched genes (y-axis) and percentage of enriched annotations for each of the embedding algorithms considered (legend). The error bars in the case of GraSpring embedding indicate the standard deviation across the ten randomised runs.

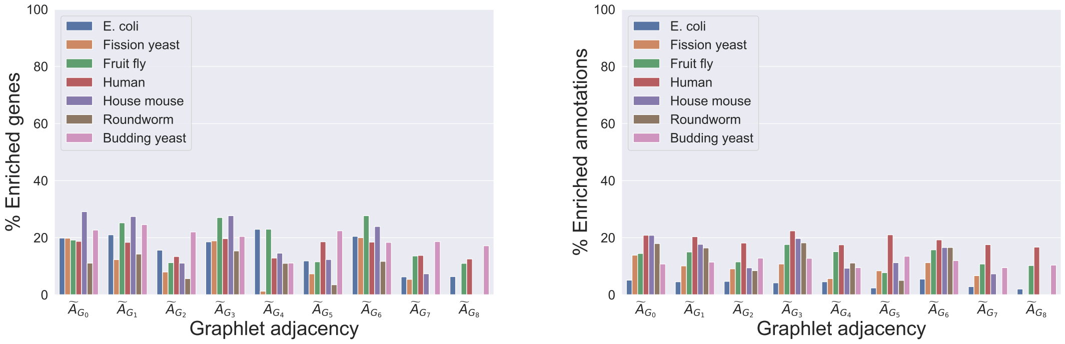


Figure A.33. SAFE GO-MF enrichment analysis comparing GraCoals in PPI networks. For the PPI networks of our seven species (legend), we show, on the y-axis, the percentage of enriched genes (left) and the percentage of enriched annotations (right) for each of the different GraCoal embeddings (x-axis).

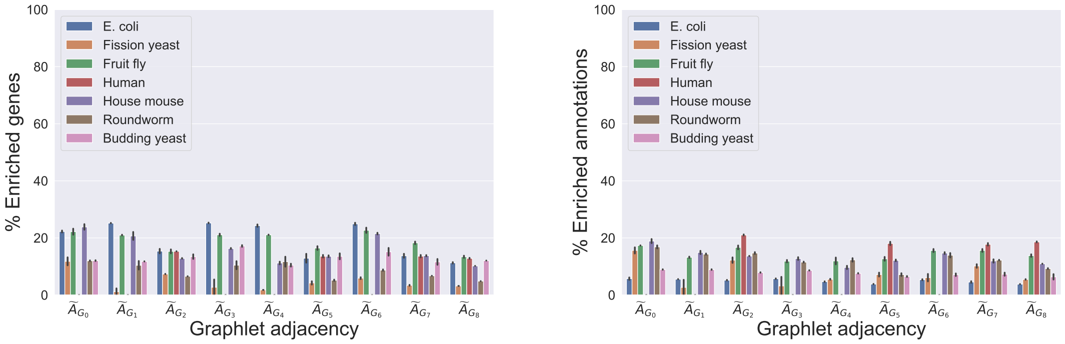


Figure A.34. SAFE GO-MF enrichment analysis comparing GraSprings in PPI networks. For the PPI networks of our seven species (legend), we show, on the y-axis, the percentage of enriched genes (left) and the percentage of enriched annotations (right) for each of the different GraSpring embeddings (x-axis).

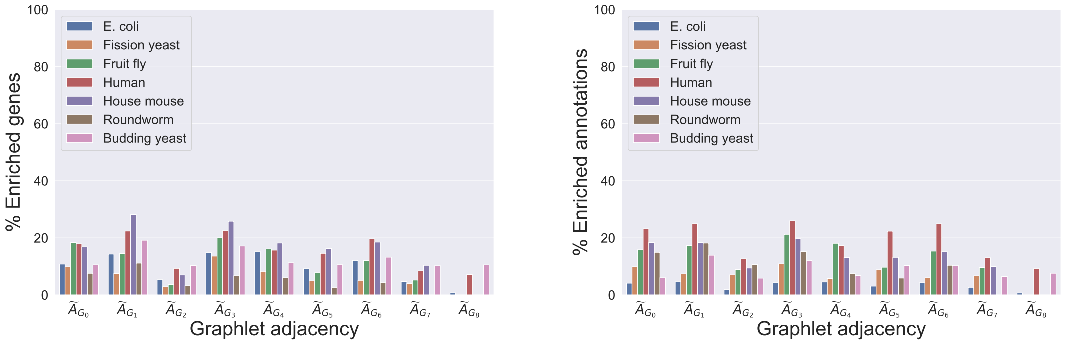


Figure A.35. SAFE GO-MF enrichment analysis comparing Spectrals in PPI networks. For the PPI networks of our seven species (legend), we show, on the y-axis, the percentage of enriched genes (left) and the percentage of enriched annotations (right) for each of the different Spectral embeddings (x-axis).

A.4 Enrichment statistics COEX networks

In this section, we summarize the enrichment results achieved when using SAFE with the different graphlet based embedding algorithms on our COEX molecular networks. That is, the percentages of genes that have at least one annotation enriched in their neighborhood and the percentages of enriched annotations for all our COEX molecular networks across different annotations. We show the summarized statistics for GO-CC and GO-MF, as for GO-BP we already covered in chapter 6.

Gene ontology cellular components

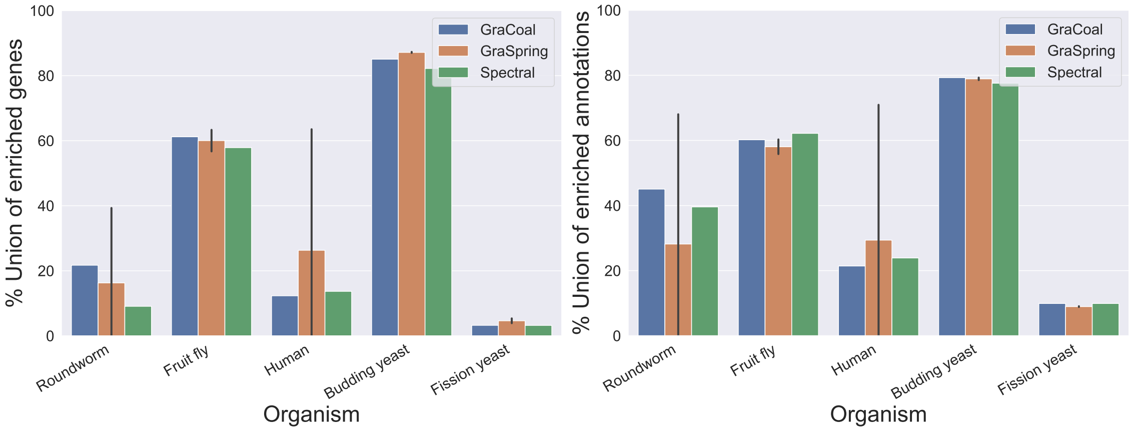


Figure A.36. SAFE GO-CC enrichment analysis for COEX networks. For the COEX networks of our seven species (x-axis), we show the percentage of enriched genes (y-axis) and percentage of enriched annotations for each of the embedding algorithms considered (legend). In the case of GraSpring, we show the average across ten randomised runs and the standard deviation (error-bars).

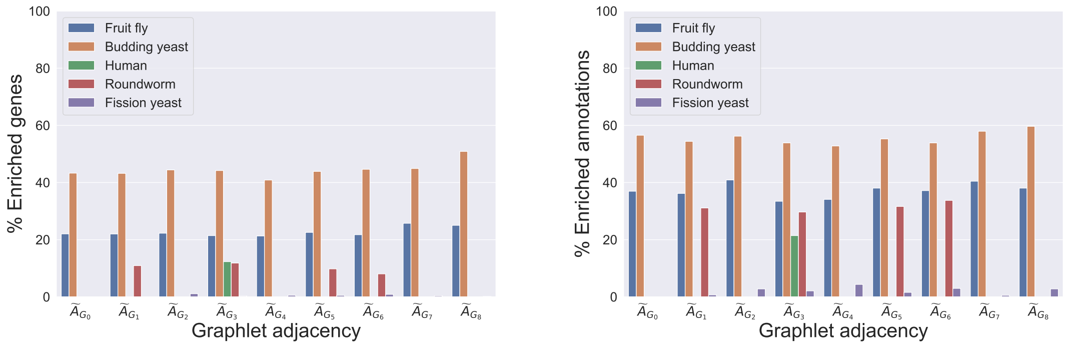


Figure A.37. SAFE GO-CC enrichment analysis comparing GraCoals in COEX networks. For the COEX networks of our six species (legend), we show, on the y-axis, the percentage of enriched genes (left) and the percentage of enriched annotations (right) for each of the different GraCoal embeddings (x-axis).

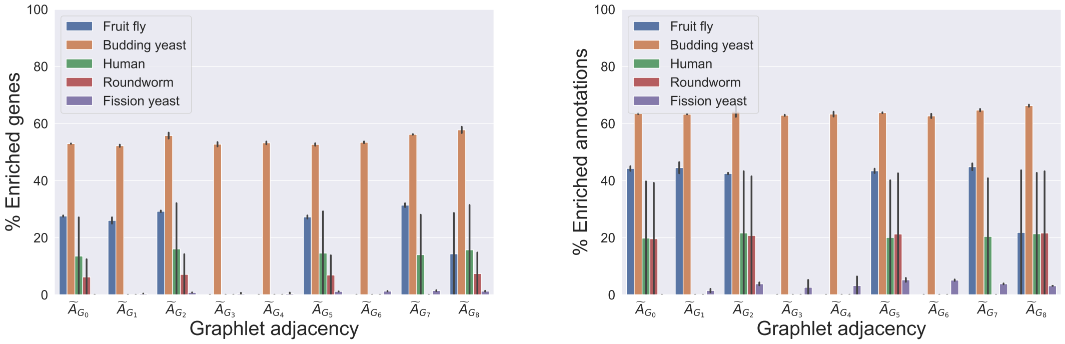


Figure A.38. SAFE GO-CC enrichment analysis comparing GraSprings in COEX networks. For the COEX networks of our six species (legend), we show, on the y-axis, the percentage of enriched genes (left) and the percentage of enriched annotations (right) for each of the different GraSpring embeddings (x-axis).

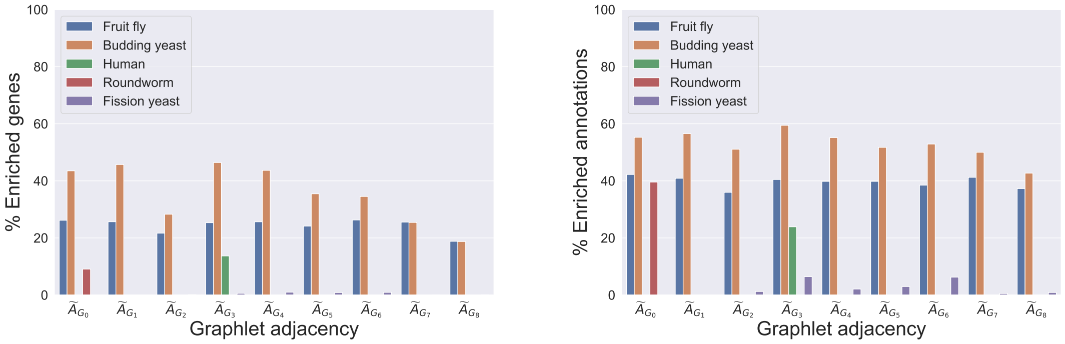


Figure A.39. SAFE GO-CC enrichment analysis comparing Spectrals in COEX networks. For the COEX networks of our six species (legend), we show, on the y-axis, the percentage of enriched genes (left) and the percentage of enriched annotations (right) for each of the different graphlet Spectral embeddings (x-axis).

Gene ontology molecular functions

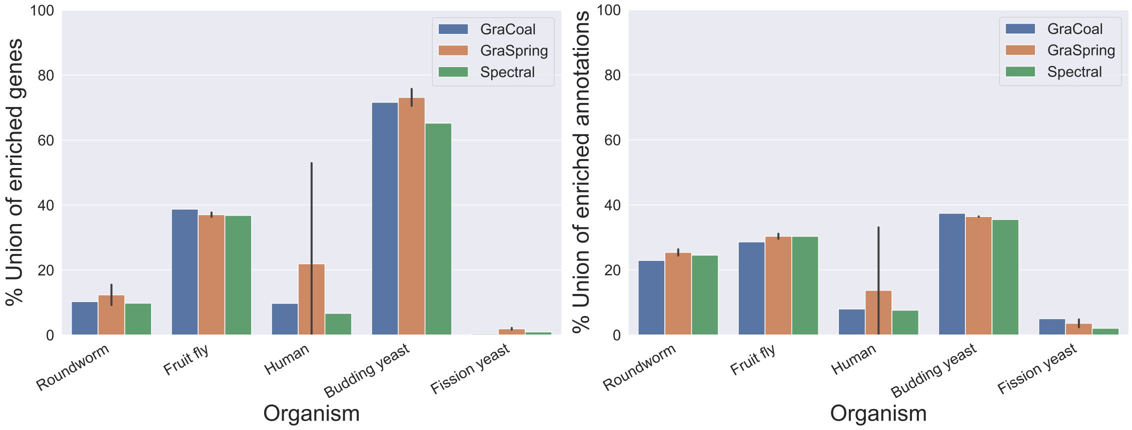


Figure A.40. SAFE GO-MF enrichment analysis for COEX networks. For the COEX networks of our seven species (x-axis), we show the percentage of enriched genes (y-axis) and percentage of enriched annotations for each of the embedding algorithms considered (legend). The error bars in the case of GraSpring embedding indicate the standard deviation across the ten randomised runs.

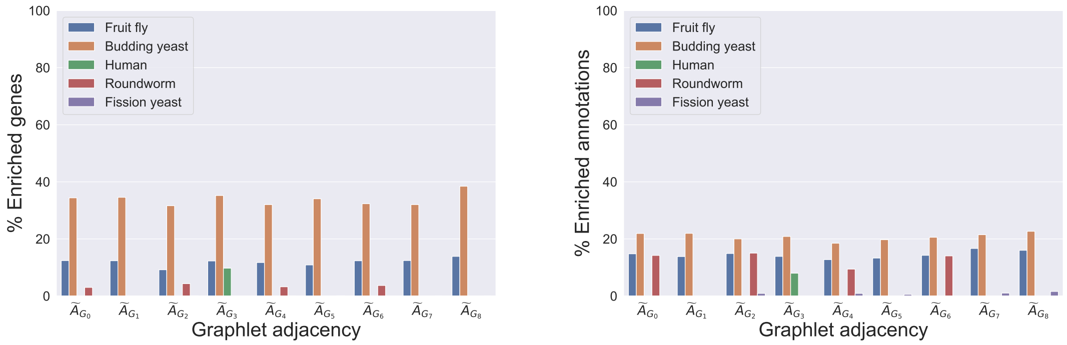


Figure A.41. SAFE GO-MF enrichment analysis comparing GraCoals in COEX networks. For the COEX networks of our six species (legend), we show, on the y-axis, the percentage of enriched genes (left) and the percentage of enriched annotations (right) for each of the different GraCoal embeddings (x-axis).

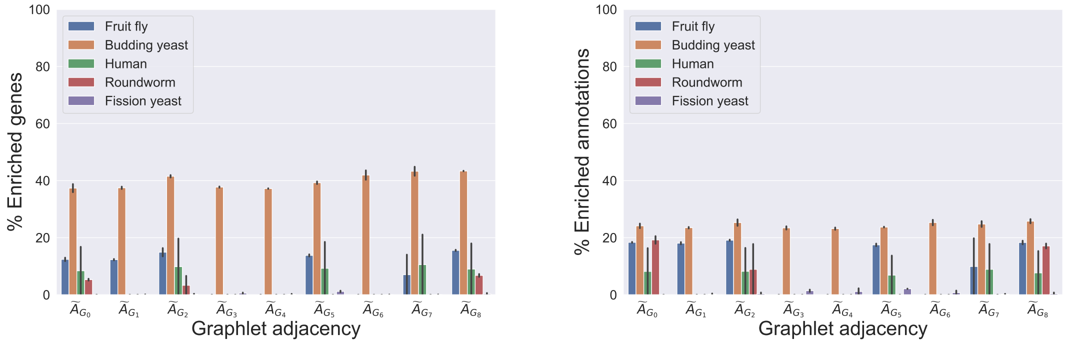


Figure A.42. SAFE GO-MF enrichment analysis comparing GraSprings in COEX networks. For the COEX networks of our six species (legend), we show, on the y-axis, the percentage of enriched genes (left) and the percentage of enriched annotations (right) for each of the different GraSpring embeddings (x-axis).

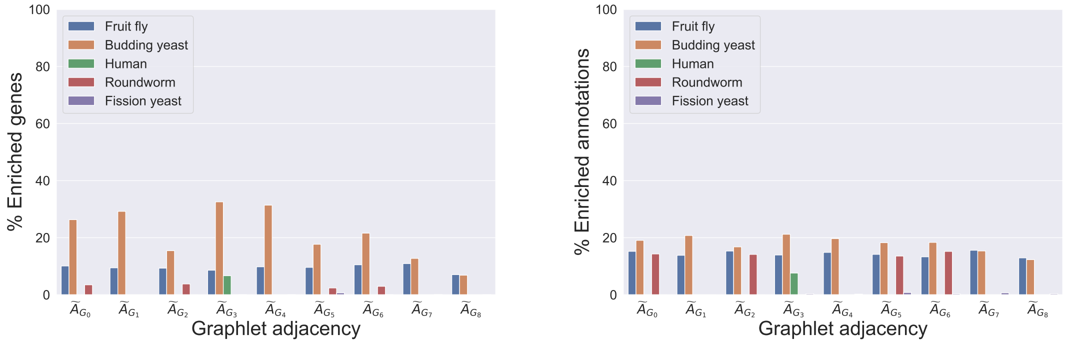


Figure A.43. SAFE GO-MF enrichment analysis comparing Spectrals in COEX networks. For the COEX networks of our six species (legend), we show, on the y-axis, the percentage of enriched genes (left) and the percentage of enriched annotations (right) for each of the different graphlet Spectral embeddings (x-axis).

A.5 Model network fitting

To characterize the structure of the molecular networks, we perform model fitting experiments to compare our real molecular networks to eight different types of random model networks commonly used in biology (see section 2.4 - Model networks). To do this, for a given real molecular network, we generate 15 random networks for each network model. We set the number of nodes, edge density, node degree distribution and number of communities (when needed) to match those of the input data to randomly generate synthetic networks that follow each of the random network models.

Next, to measure the dissimilarity between two networks (i.e., a real network and a model network) we first characterize the global wiring patterns of each network with its Graphlet Correlation Matrix (GCM) (Yaveroğlu et al., 2014), which is an 11 × 11 matrix with the Spearman's correlations between the eleven non-redundant graphlet orbit counts over all nodes in the network. As such, orbit counts, i.e., the numbers of times a node touches each graphlet orbit in a network, generalize the notion of graphlet degrees (Przulj, 2007). Next, we use the graphlet correlation distance-11 (GCD-11) between two networks, which is the Euclidean distance of the upper triangle values of the corresponding GCMs (Yaveroğlu et al., 2014).

Finally, To measure the fit between a real network (e.g., the Budding yeast GI network) and a model network (e.g., the GEO model), we first compute the GCD-11 distances between the real network and the 15 generated networks that correspond to the network model as well as the GCD-11 distances between the 15 generated networks. We measure the overlap between these two distance distributions (real to model and model to model) by means of a Wilcoxon-Mann-Whitney U-test (MWU). We can distinguish the real network form the given model network if the p-value of our MWU test is less than 5%. In Figures A.44, A.46 and A.47 and Tables A.21, A.22 and A.23, we present the fit of network models and the corresponding MWU-test p-values, respectively for the GI, PPI and COEX networks.

None of the real molecular networks were well fitted by the network models, except for the Scale-free with gene duplication model for some of the GI networks. Thus, in Figure A.45, we show the GCD-11 distances between the real GI networks and the 15 randomly generated networks with SFGD properties (blue line) and the GCD-11 distances between the randomly generated networks (orange line). Except for the Fruit fly GI network, we observe the blue and orange lines are very close to each other, almost overlapping.

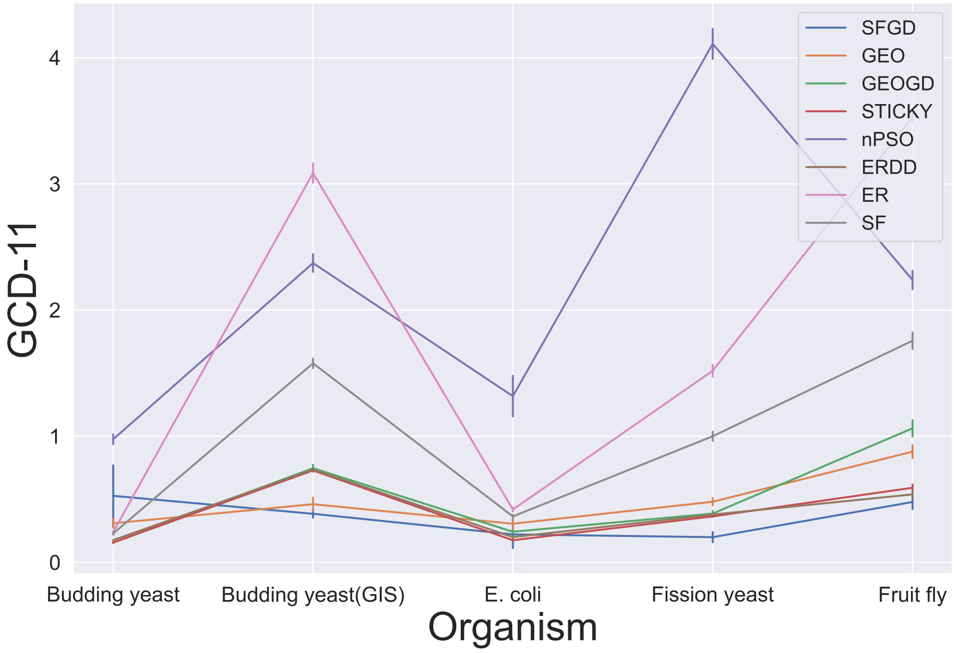


Figure A.44. Fit of network models for the GI networks. Each line shows the fitting of a network model for the different GI networks and the Budding yeast GIS network (x-axis). The error-bars show the averages and standard deviations of the pairwise GCD-11 distances (y-axis) between the GI networks and for each network model, 15 randomly generated networks of the same size as the GI networks.

	Network Model							
Organism	ER	ERDD	GEO	GEOGD	nPSO	SF	SFGD	STICKY
E. coli	9.63E-06	9.63E-06	9.63E-06	9.63E-06	9.63E-06	9.63E-06	**0.5197422**	9.63E-06
Fruit fly	9.63E-06	9.63E-06	9.63E-06	9.63E-06	9.63E-06	9.63E-06	9.63E-06	9.63E-06
Fission yeast	9.63E-06	9.63E-06	9.63E-06	9.63E-06	9.63E-06	9.63E-06	**0.0161461**	9.63E-06
Budding yeast	9.63E-06	9.63E-06	9.63E-06	9.63E-06	9.63E-06	9.63E-06	**0.017908**	9.63E-06

Table A.21. Mann-Whitney-U test p-values. For each random model, we perform a MWU test between two distance distributions to evaluate if there is any statistical difference: GCD-11 between the real data to model network data and model network data to model network data. Non significant p-values (>0.01) indicate no statistical difference between a molecular network and a particular network model, for instance for *E. coli* for the SFGD network model.

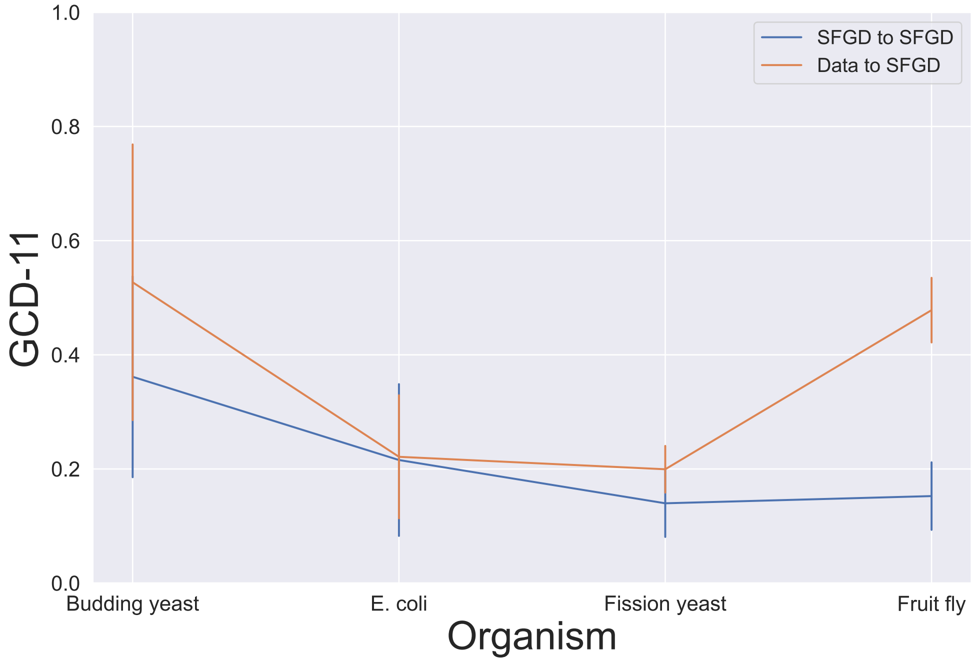


Figure A.45. Fit of the SFGD model for the GI networks. The blue line shows the fitting of the SFGD network model to the real GI networks (x-axis). The blue line and error-bars represent the averages and standard deviations of the pairwise GCD-11 distances (y-axis) between the GI networks and 15 randomly generated networks of the same size as the GI networks with scale-free and gene duplication properties. The orange line and error-bars in the orange line represent the same statistics, but between the randomly generated networks.

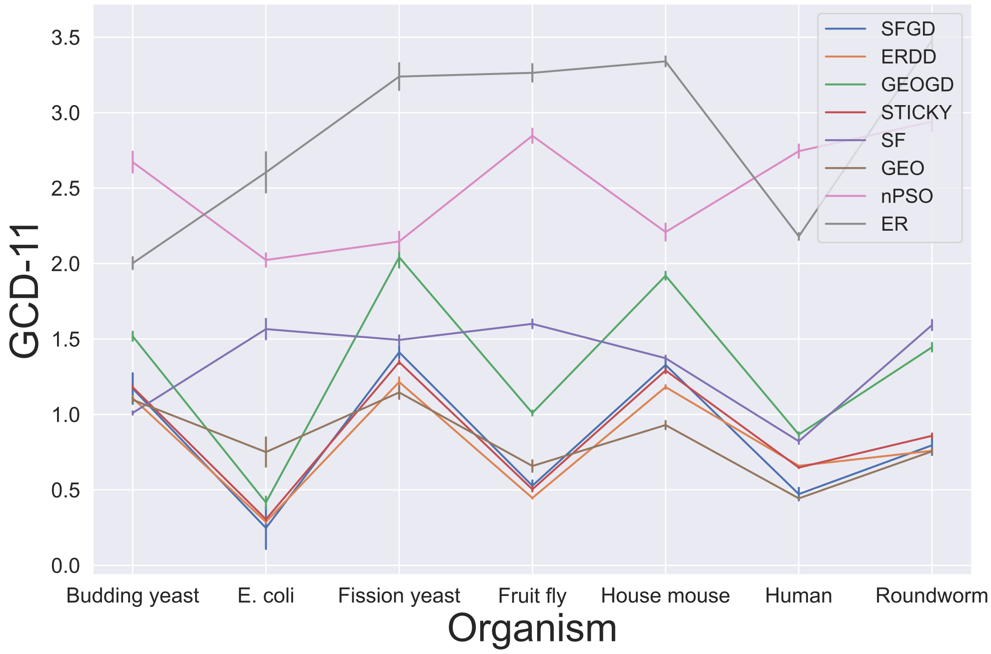


Figure A.46. Fit of network models for the PPI networks. Each line shows the fitting of a network model for the different PPI networks (x-axis). The error-bars show the averages and standard deviations of the pairwise GCD-11 distances (y-axis) between the PPI networks and for each network model, 15 randomly generated networks of the same size as the PP networks.

	Network Model							
Organism	ER	ERDD	GEO	GEOGD	nPSO	SF	SFGD	STICKY
Budding yeast	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07
E. coli	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07
Fruit fly	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07
Fission yeast	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07
Human	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07
House mouse	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07
Roundworm	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07

Table A.22. Mann-Whitney-U test p-values. For each random model, we perform a MWU test between two distance distributions to evaluate if there is any statistical difference: GCD-11 between the real data to model network data and model network data to model network data. All the PPI molecular networks are statistically different than the model networks. Note that the p-values presented are the minimum that can be achieved when comparing two non-overlapping distributions of 15 (data-to-model) and 105 (model-to-model) GCD-11 values.

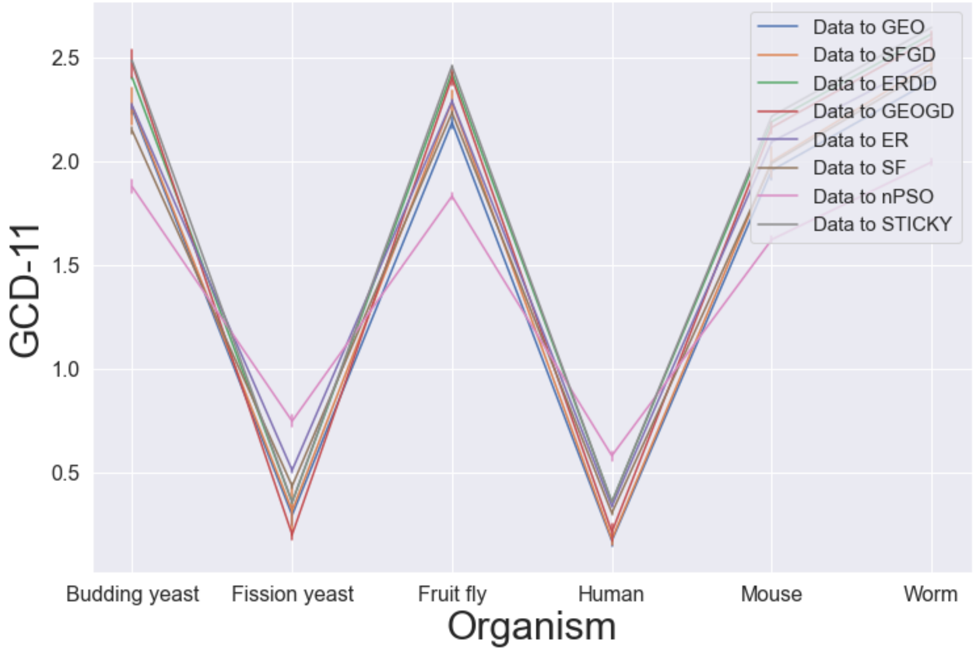


Figure A.47. Fit of network models for the COEX networks. Each line shows the fitting of a network model for the different COEX networks (x-axis). The error-bars show the averages and standard deviations of the pairwise GCD-11 distances (y-axis) between the COEX networks and for each network model, 15 randomly generated networks of the same size as the PP networks.

	Network Model							
Organism	ER	ERDD	GEO	GEOGD	nPSO	SF	SFGD	STICKY
Budding yeast	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07
E. coli	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07
Fruit fly	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07
Fission yeast	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07
Human	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07
Roundworm	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07	9.63E-07

Table A.23. Mann-Whitney-U test p-values. For each random model, we perform a MWU test between two distance distributions to evaluate if there is any statistical difference: GCD-11 between the real data to model network data and model network data to model network data. All the PPI molecular networks are statistically different than the model networks. Note that the p-values presented are the minimum that can be achieved when comparing two non-overlapping distributions of 15 (data-to-model) and 105 (model-to-model) GCD-11 values.

A.6 Gene-paralog statistics for GI networks

Organism	Paralogs	Total genes	Paralog coverage (%)
Budding yeast	1,870	6,000	31.17
E. coli	1,420	4,402	32.26
Fission yeast	798	5,122	15.58
Fruit fly	323	14,000	2.31

Table A.24. Paralog data statistics. For the four GI networks (column 1), we report the total paralogous genes identified with BLASTp and total number of genes known to date for each species according to the UniProt database (columns 2 and 3). On column 4, we report the number of paralogous genes with respect to the total number of genes, in terms of percentage.