*Bioinformatics*
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

SYSTEMS

# Functional geometry of protein interactomes

**Noël Malod-Dognin**[1] **and Nataša Pržulj**[1,2,*]

[1] Department of Life Sciences, Barcelona Supercomputing Center, 08034 Barcelona, Spain
[2] ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Spain

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

## Abstract

**Motivation:** Protein-protein interactions (PPIs) are usually modelled as networks. These networks have extensively been studied using graphlets, small induced subgraphs capturing the local wiring patterns around nodes in networks. They revealed that proteins involved in similar functions tend to be similarly wired. However, such simple models can only represent pairwise relationships and cannot fully capture the higher-order organization of protein interactomes, including protein complexes.

**Results:** To model the multi-scale organization of these complex biological systems, we utilize simplicial complexes from computational geometry. The question is how to mine these new representations of protein interactomes to reveal additional biological information. To address this, we define *simplets*, a generalization of graphlets to simplicial complexes. By using simplets, we define a sensitive measure of similarity between simplicial complex representations that allows for clustering them according to their data types better than clustering them by using other state-of-the-art measures, e.g., spectral distance, or facet distribution distance.

We model human and baker's yeast protein interactomes as simplicial complexes that capture PPIs and protein complexes as simplices. On these models, we show that our newly introduced simplet-based methods cluster proteins by function better than the clustering methods that use the standard PPI networks, uncovering the new underlying functional organization of the cell. We demonstrate the existence of the functional geometry in the protein interactome data and the superiority of our simplet-based methods to effectively mine for new biological information hidden in the complexity of the higher order organization of protein interactomes.

**Availability:** Codes and datasets are freely available at http://www0.cs.ucl.ac.uk/staff/natasa/Simplets/

**Contact:** natasa@cs.ucl.ac.uk

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

### 1.1 Motivation

Genome is the blueprint of a cell. DNA regions called genes are transcribed into messenger RNAs that are translated into proteins. These proteins interact with each other and with other molecules to perform their biological functions. Deciphering the patterns of molecular interactions (also called topology) is fundamental to understanding the functioning of the cell (Ryan *et al.*, 2013). In system biology, molecular interactions are modeled as various molecular interaction networks, in which nodes represent molecules and edges connect molecules that interact in some way. Examples include the well-known protein-protein interaction (PPI) networks in which nodes represent proteins and edges connect proteins that can physically bind.

Because exact comparison between networks has long been known to be computationally intractable (Cook, 1971), the topological analyses of biological networks use approximate comparisons (heuristics), commonly called network properties, such as the degree distribution, to approximately say whether the structures of networks are similar (Newman, 2010). Advanced network properties that utilize graphlets (small induced subgraphs) (Pržulj *et al.*, 2004) have been successfully used to mine biological network datasets. Graphlet-based properties include measures of topological similarities between nodes and between networks (Pržulj *et al.*, 2004;

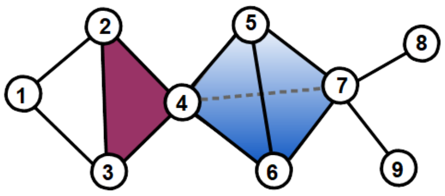

**Fig. 1. Illustration of a 3-dimensional simplicial complex.** In the presented simplicial complex, nodes 1, 2 and 3 are only connected by 1-dimensional simplices (edges, in black). Nodes 2, 3 and 4 are connected by a 2-dimensional simplex (triangle, in magenta). Nodes 4, 5 , 6 and 7 are connected by a 3-dimensional simplex (tetrahedron, in blue).

Pržulj, 2007; Yaveroğlu *et al.*, 2014), as well as between protein 3D structures represented by networks (Malod-Dognin and Pržulj, 2014; Faisal *et al.*, 2017). In particular, graphlets have been used to characterize and compare the local wiring patterns around nodes in a PPI network (Milenković and Pržulj, 2008), which revealed that molecules involved in similar functions tend to be similarly wired (Davis *et al.*, 2015). These topological similarities between nodes have also been used to guide the node mapping process of network alignment methods (Kuchaiev *et al.*, 2010; Kuchaiev and Pržulj, 2011; Malod-Dognin and Pržulj, 2015; Vijayan *et al.*, 2015), which allowed for transferring of biological annotation between nodes in different networks of well-studied species to less studied ones.

Despite significant progress, these simple network (also called graph) models of molecular interaction data can only represent pairwise relationships and cannot fully capture the higher organization of molecular interactions, such as protein complexes and biological pathways (Estrada and Rodriguez-Velazquez, 2005). Hence, we need to model these data by using new mathematical formalisms capable of capturing their multi-scale organization. Furthermore, we need to design new algorithms capable of extracting new biological information hidden in the wiring patterns of the molecular interaction data modeled by using these mathematical formalisms. This paper addresses these issues.

## 1.2 Simplicial complexes basics

A candidate model for capturing higher-order molecular organization is a simplicial complex (Munkres, 1984). A *simplicial complex* is a set of *simplices*, where a 0-dimensional simplex is a node, a 1-dimensional simplex is an edge, a 2-dimensional simplex is a triangle, a 3-dimensional simplex is a tetrahedron and their $n$-dimensional counterparts (illustrated in Figure 1). The dimension of a simplicial complex is the largest dimension of its simplices.

The $(n\text{-}1)$-dimensional sub-simplices of an $n$-dimensional simplex are called its *faces* (e.g., a triangle has three faces, the three edges). A simplicial complex, $K$, is required to satisfy two conditions:

- For any simplex $\delta \in K$, any face $\delta'$ of $\delta$ is also in $K$.
- For any two simplices, $\delta_1, \delta_2 \in K$, $\delta_1 \bigcap \delta_2$ is either $\emptyset$, or a face of both $\delta_1$ and $\delta_2$.

In a simplicial complex, a *facet* is a simplex that is not a face of any higher dimensional simplex. Because of this property, a simplicial complex can be summarized by its set of facets.

Note that a network is a 1-dimensional simplicial complex and thus, our proposed methodology is directly applicable to both traditional networks and the higher dimensional simplicial complexes.

While simple network statistics, such as degrees, shortest paths and centralities, have been generalized to simplicial complexes (Estrada and Ross, 2018), the lack of more advanced statistics capturing the geometry of simplicial complexes limits their usage in practical applications

## 1.3 Contributions

To comprehensively capture the multi-scale organization of complex molecular networks, we propose to model them by using simplicial complexes. To extract the information hidden in the geometric patterns of these models, we generalize graphlets to simplicial complexes, which we call *simplets*. Our simplets extend the applicability of graphlets to high-dimensional simplicial complexes. When applied to one-dimensional simplicial complexes, i.e., networks, they are identical to graphlets. On large scale real-world and synthetic simplicial complexes, we show that simplets can be used to define a sensitive measure of geometric similarity between simplicial complexes. Then, on simplicial complexes capturing the protein interactomes of human and yeast, we show that simplets can be used to relate the local geometry around proteins in simplicial complexes with their biological functions. Comparison between 1-dimensional protein-protein interaction networks and the higher-dimensional simplicial complex representations of the interactomes formed by protein interactions and protein complexes shows that higher-order modeling enabled by simplicial complexes allows for capturing more biological information, which can efficiently be mined with our proposed simplets.

# 2 Methods

## 2.1 Datasets and their simplicial complex representations

### 2.1.1 Yeast and human protein interactomes

From BioGRID (v. 3.4.156)(Chatr-Aryamontri *et al.*, 2017), we collected the experimentally validated protein-protein interaction (PPI) networks of human (H. sapiens) and of yeast (S. cerevisiae). From CORUM (Ruepp *et al.*, 2010), we collected (on the $2^{nd}$ of July, 2017) the experimentally validated protein complexes of human, and from CYC2008 (v.2.0) (Pu *et al.*, 2009) the experimentally validated protein complexes of yeast. We consider two different models of an organism's interactome.

**The 1-dimensional PPI network:** it is the usual PPI network, in which proteins (nodes) are connected by an edge if they can physically bind. Recall that a network is a 1-dimensional simplicial complexes on which our new simplet methodologies can be applied and are equivalent to the standard graphlet methodologies.

**The higher-dimensional PPI Complex:** starting from the PPI network, we additionally connect by simplices all the proteins that belong to common complexes. I.e., the proteins belonging to a $k$-protein complex are connected by a $(k\text{-}1)$ dimensional simplex.

For human, the PPI network has 16,100 nodes and 212,319 edges. When unifying the lower dimensional protein-protein interaction data and the higher order protein complex data as described above, the resulting PPI Complex is a 140-dimensional simplicial complex having 16,140 nodes (with 40 proteins being part of proteins complexes but not having any reported protein-protein interaction) and 205,192 facets. For yeast, the PPI network has 5,842 nodes and 80,900 edges. When unifying the lower dimensional protein-protein interaction data and the higher order protein complex data as described above, the resulting PPI Complex is a 80-dimensional simplicial complex having 5,842 nodes and 76,790 facets.

### 2.1.2 Other real-world datasets

We collected real-world higher-dimensional datasets from biology and beyond.

- **1,569 simplicial complexes of protein 3D structures:** Proteins are linear arrangements of amino-acids that in the aqueous environment of the cell fold and acquire specific three-dimensional (3D) shapes called tertiary structures. We collected from Astral-40 (SCOPe v.2.06)

(Fox *et al.*, 2013) the 3D structures of 1,569 protein domains that are at-least 100 amino-acid long. Each protein domain is modeled as a simplicial complex in which simplices connect together all the amino-acids (nodes) that are less than 7.5 Å apart (as measured by the distances between their $\alpha$-carbons).

- **132 simplicial complexes of publication authorships:** From the preprint repository arXiv, we collected all the scientific publications in the "computer science" category over eleven years from 2007 to 2017. For each month, we model the scientific collaborations as a simplicial complex in which simplices are formed by all scientists (nodes) that co-authored a scientific publication.

- **60 simplicial complexes of genes' biological annotations:** We collected pathway annotations from Reactome database (v.63) (Fabregat *et al.*, 2017), as well as the experimentally validated Gene Ontology (GO)(Ashburner *et al.*, 2000) annotations from NCBI's entrez webserver (collected in February 2018). For GO, we consider biological process, molecular function, and cellular component annotations separately. For each annotation set, we model the functional annotations of the genes of a given species as a simplicial complex in which simplices are formed by all genes (nodes) that have a common annotation term (restricted to terms annotating up-to 50 genes for computational complexity issues). We only considered simplicial complexes having more than 100 nodes. Following this procedure, we generated 18 pathway simplicial complexes, 13 biological process simplicial complexes, 14 molecular function simplicial complexes and 15 cellular component simplicial complexes.

- **14 simplicial complexes of protein-protein interactions:** We collected the experimentally validated protein-protein interactions (PPIs) from BioGRID database (v. 3.4.156)(Chatr-Aryamontri *et al.*, 2017). These PPIs are first modeled as networks in which proteins (nodes) are connected by edges if they can interact. The corresponding networks are converted into so-called *clique complexes*, by creating a simplex between all nodes belonging to a maximal clique in the network.

### 2.1.3 Random simplicial complexes

To test our methods, we considered randomly generated simplicial complexes, which we generate according to ten random models (detailed in Supplementary material, section 1).

The first five models are based on randomly generated graphs, which are converted into so-called *clique complexes*, in which simplices connect nodes that belong to a clique in the graph.

- A *random clique complex* (RCC) is the clique complex of an Erdös-Rènyi random graph (Erdös and Rényi, 1959).
- A *Vietoris-Rips complex* (VRC) (Hausmann *et al.*, 1995) is the clique complex of a geometric random graph (Penrose, 2003).
- A *scale-free complex* (SFC) is the clique complex of a Barabàsi-Albert scale-free graph (Barabási and Albert, 1999).
- A *Watts-Strogatz complex* (WSC) is the clique complex of a small-world graph (Watts and Strogatz, 1998).
- An *nPSO complex* (nPSOC) is the clique complex of a non-uniform Popularity Similarity Optimization graph (Muscoloni and Cannistraci, 2018).

The five other models are extensions of the *Linial-Meshulam* model (Linial and Meshulam, 2006; Meshulam and Wallach, 2009), which originally consists in randomly connecting nodes with $k$-dimensional facets. We extended this model to randomly connect nodes with facets while following the facet distribution of an input simplicial complex. In this way, we can create Linial-Meshulam variant of the five clique complex-based models presented above.

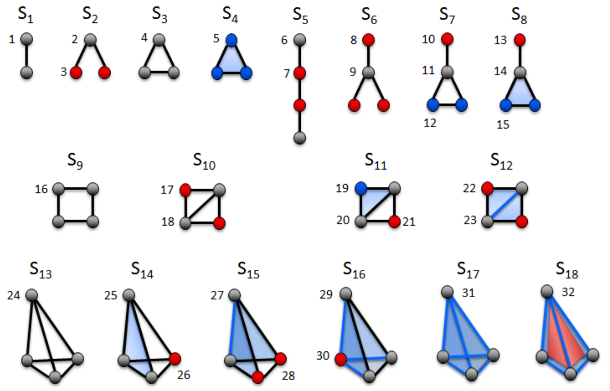

**Fig. 2. Illustration of 2- to 4-nodes simplets.** The 18 2- to 4-nodes simplets are denoted by $S_1$ to $S_{18}$. Within each simplet, geometrically interchangeable nodes, belonging to the same orbit, have the same color. These simplets have 32 different orbits, denoted from 1 to 32. Note that simplets $S_4$, $S_8$, $S_{11}$ and $S_{14}$ have only one 2D face (triangle, in blue), while $S_{12}$ and $S_{15}$ have two triangles, $S_{16}$ has 3 triangles and $S_{17}$ has four triangles. $S_{18}$ has four triangles and one 3D face (tetrahedron, in red).

- A *Linial-Meshulam random clique complex* (LM- RCC) is a Linial-Meshulam complex that follows the facet distribution of an input random clique complex.
- A *Linial-Meshulam Vietoris-Rips complex* (LM- VRC) is a Linial-Meshulam complex that follows the facet distribution of an input Vietoris-Rips complex.
- A *Linial-Meshulam scale-free complex* (LM-SFC) is a Linial-Meshulam complex that follows the facet distribution of an input scale-free complex.
- A *Linial-Meshulam Watts-Strogatz complex* (LM-WSC) is a Linial-Meshulam complex that follows the facet distribution of an input Watts-Strogatz complex.
- A *Linial-Meshulam nPSO complex* (LM-nPSOC) is a Linial-Meshulam complex that follows the facet distribution of an input nPSO complex.

For each model we choose three node sizes, 1,000, 2,000, and 3,000 nodes, and three edge densities, 0.5%, 0.75% and 1%. We generated 25 random simplicial complexes for each model and each of these node sizes and edge densities. Hence, in total, we generated $10 \times 3 \times 3 \times 25 = 2{,}250$ random simplicial complexes. We chose these node sizes and edge densities to roughtly mimic the sizes and densities of real-world data detailed above.

## 2.2 Capturing the local geometry around nodes in a simplicial complex with simplets

We define *simplets* as small, connected, non-isomorphic, induced simplicial complexes of a larger simplicial complex. Figure 2 shows the eighteen 2- to 4-node simplets (denoted by $S_1$ to $S_{18}$). Within each simplet, because of symmetries, some nodes can have identical geometries. Analogous to automorphism orbits in graphlets (Pržulj, 2007), we say that such nodes belong to a common *simplet orbit group*, or *orbit* for brevity. Figure 2 shows the thirty-two orbits of the 2- to 4-node simplets (denoted from 1 to 32). Similar to graphlets, we use simplets to generalize the notion of the node degree: the $i^{th}$ *simplet degree* of node $v$, denoted by $v_i$, is the number of times node $v$ touches a simplet at orbit $i$.

We define the *simplet degree vector* (SDV) of a node as the 32 dimensional vector containing the simplet degrees of the node in the simplicial complex as its coordinates. Hence, the SDV of a node describes the local geometry around the node in the simplicial complex and comparing the SDVs of two nodes provides a measure of local geometric similarity between them.

We define the *SDV similarity* between two nodes as an extension of the graphlet degree similarity (Milenković and Pržulj, 2008). It is computed as follows. The distance, $D_i(u,v)$, between the $i^{th}$ simplet orbits of nodes $u$ and $v$ is defined as:

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

where $w_i$ is the weight of orbit $i$ that accounts for dependencies between orbits. Weight, $w_i$, is computed as $w_i = 1 - \dfrac{\log(o_i)}{\log(32)}$, where $o_i$ is the number of orbits that orbit $i$ depends on, including itself. For instance, the count of orbit 2 (the middle of a three node path) of a node depends on its count of orbit 0 (i.e. its node degree) and on itself, so $o_2 = 2$. For orbit 9, $o_9$ = 3, since it is affected by orbits 0, 2, and itself. The values of $o_i$ for all 2- to 4-nodes simplet orbits are listed in Supplementary Table 1. Finally, the SDV similarity, $S(u,v)$, between nodes $u$ and $v$ is defined as:

$$S(u,v) = 1 - \frac{\sum_{i|(u_i \neq 0) \text{ or } (v_i \neq 0)} D_i(u,v)}{\sum_{i|(u_i \neq 0) \text{ or } (v_i \neq 0)} w_i}. \tag{2}$$

$S(u,v)$ is in (0, 1], where similarity 1 means that the SDVs of nodes $u$ and $v$ are identical.

## 2.3 Capturing the global geometry of a simplicial complex with simplets

To the best of our knowledge, researchers from computational geometry have not considered the problem of comparing two simplicial complexes. However, the comparison of biological networks is a foundational problem of system biology. Instead, computational geometry focus on the comparison of two spaces, each represented by a collection of simplicial complexes, e.g. (Collins *et al.*, 2004). Thus, we build upon network analysis and extend graphlet and non-graphlet based network distance measures to directly compare simplicial complexes as follows.

### 2.3.1 Simplet correlation distance

Simplets are like Lego pieces that assemble with each other to build larger simplicial complexes. We exploit this property to summarize the complex structures of simplicial complexes and to compare them, by generalizing Graphlet Correlation Distance (Yaveroğlu *et al.*, 2014), which is a sensitive measure of topological similarity between networks.

Analogous to graphlets, the statistics of different simplet orbits are not independent of each other. The reason behind this is the fact that smaller simplets are induced sub-simplicial complexes of larger simplets. In Supplementary material, section 2, we present the four, non-redundant dependency equations between the simplet degrees of a given node $u$ that we used to assess the correctness of our exhaustive simplet counter.

In addition to these redundancies there also exist dependencies between simplets, which are dataset dependent. We use these dataset dependent simplet orbit dependencies to characterize the global geometry of simplicial complexes. We capture the dependencies between simplet orbits by the simplicial complex's *Simplet Correlation Matrix* (SCM), which we define as follows. We construct a matrix whose rows are the simplet degree vectors of all nodes of the simplicial complex. We calculate the Spearman's correlation between each two pairs of columns in the resulting matrix, i.e., correlations between the orbits over all nodes of the simplicial complex. We present these correlations in a $32 \times 32$ dimensional Simplet Correlation Matrix (SCM): it is symmetric and contains Spearman's correlation values in [-1,1] range. As presented in Supplementary Figure 1, the SCMs of simplicial complexes from different random simplicial complex models are indeed very different. We exploit these differences in SCMs to compare simplicial complexes.

We define the *Simplet Correlation Distance* (SCD) to measure the distance between two simplicial complexes, $K_1$ and $K_2$, by the Euclidean distance between the upper-triangles of their SCMs:

$$SCD(K_1,K_2) = \sqrt{\sum_{i=1}^{32}\sum_{j=i+1}^{32}(SCM_{K_1}[i][j] - SCM_{K_2}[i][j])^2}, \tag{3}$$

where $SCM_{K_1}[i][j]$ is the $(i,j)^{th}$ entry in the SCM of $K_1$ (similar for $K_2$). The ability of SCD to group together simplicial complexes according to their underlying models is demonstrated in section 3.2.

### 2.3.2 Facet distribution distance

In analogy to degree distribution and graphlet degree distribution (Pržulj, 2007), we define the measure of connectivity of a $k$-dimensional simplicial complex, $K$, as the distribution of its facets, $d_K$: it is a $k$-dimensional *facet distribution vector* whose $i^{th}$ entry is the percentage of the facets in $K$ having dimension $i$. The *Facet Distribution Distance* (FDD) measures the distance between two simplicial complexes, $K_1$ and $K_2$, by the Euclidean distance between their facet distribution vectors, $d_{K_1}$ and $d_{K_2}$:

$$FDD(K_1,K_2) = \sqrt{\sum_i (d_{K_1}[i] - d_{K_2}[i])^2}. \tag{4}$$

### 2.3.3 Spectral distance

Spectral theory captures the topology of networks and simplicial complexes by using the eigen-values and eigen-vectors of matrices representing them, such as the adjacency matrix, or Laplacian matrix (Wilson and Zhu, 2008). Let $H$ be the incidence matrix of a simplicial complex, $K$, having $n$ nodes and $f$ facets: $H$ is a $n \times f$ matrix in which entry $H[i][j] = 1$ if node $i$ is in facet $j$, and 0 otherwise. The corresponding degree matrix, $D$, is a $n \times n$ diagonal matrix in which entry $D[i][i]$ is the number of facets containing node $i$. The adjacency matrix, $A$, of a simplicial complex is: $A = HH^T - D$, where $H^T$ is the transpose of $H$ (Zhou *et al.*, 2007). The corresponding Laplacian matrix, $L$, is: $L = \frac{1}{2}D^{-1/2}AD^{-1/2}$.

The eigen-decomposition of the Laplacian matrix, $L$, of simplicial complex, $K$, is $L = \phi\lambda_K\phi^T$, where $\lambda_K = diag(\lambda_K^1, \lambda_K^2, ..., \lambda_K^n)$ is the diagonal matrix with the ordered eigen-values, $\lambda_K^i$ as elements and $\phi = (\phi_1|\phi_2|...|\phi_n)$ is the matrix with the ordered eigen-vectors as columns. The spectrum of simplicial complex, $K$, is the set of its eigenvalues $S_K = \{\lambda_K^1, \lambda_K^2, ..., \lambda_K^n\}$, which are reordered so that $\lambda_K^1 \geq \lambda_K^2 \geq ... \geq \lambda_K^n$.

We define the *spectral distance* (SD) between two simplicial complexes, $K_1$ and $K_2$, as the Euclidean distance between their spectra (Wilson and Zhu, 2008):

$$SD(K_1,K_2) = \sqrt{\sum_i (\lambda_{K_1}^i - \lambda_{K_2}^i)^2}. \tag{5}$$

When the two spectra are of different sizes, 0 valued eigen-values are added at the end of the smaller spectrum.

# 3 Results and discussion

## 3.1 Comparing simplicial complexes

A sensitive measure of simplicial complex similarity should find smaller distances between simplicial complexes from the same model than between simplicial complexes from different models.

We visually assess if our simplet correlation distance (SCD, presented in section 2.3.1) has this property by embedding simplicial complexes as points in 3-dimensional (3D) space, so that the Euclidean distances between the points in the 3D space best approximate the SCD distances between the corresponding simplicial complexes. We do by using multidimensional scaling, MDS (Borg and Groenen, 2005). As presented in Figure 3, when using the 2,250 model simplicial complexes described in Section 2.1.2, we observe that simplicial complexes from the same models are grouped together (i.e., they have small SCD distances), while simplicial complexes from different models are well separated (i.e., they have larger SCD distances).

We apply SCD and of the two other distances measures described in Section 2.3 on the 2,250 model simplicial complexes described in Section 2.1.2, which results for each distance measure between all 2,530,125 pairs of these 2,250 simplicial complexes. We formally assess the ability of the distance measure to group together the simplicial complexes from the same model by using the standard Precision-Recall and Receiver Operating Characteristic (ROC) curves analyses. The resulting Precision-Recall and ROC curves, which are presented in Supplementary Figures 2 and 3, confirm our visual illustration of the ability of SCD to classify simplicial complexes. We find that SCD achieves the highest classification performance with average precision (AP) of 97.58% and an area under the ROC curve (AUC) of 84.93%. It is followed by the facet distribution distance (AP of 96.00% and AUC of 78.73%) and by the spectral distance (AP of 91.42% and AUC of 60.52%).

We further validate our methodology by assessing its ability to correctly group our 1,775 real-world simplicial complexes. We calculate the distances between all pairs of the 1,775 real-world simplicial complexes, which results in distances between $\binom{1,775}{2}$ = 1,574,425 pairs for each of the three distance measures presented in Section 2.3. As illustrated in Figure 4, when the real-world simplicial complexes are embedded into 3D space based on their SCD distances by using multi-dimensional scaling, the simplicial complexes from the same data type group well together. Out of the four types of real-world simplicial complexes, the ones capturing protein-protein interactions are less well clustered, i.e., these simplicial complexes are more variable than the other ones. This could be due to the incompleteness and noisiness of protein-protein interaction data (Sprinzak *et al.*, 2003), as well as to evolutionary differences in the wiring patterns of the species' interactomes, as our dataset includes diverse species, such as Arabidopsis thaliana (a plant), Homo sapiens (a mammal), and Saccharomyces cerevisiae (a fungus).

Nevertheless, the precision-recall curves presented in Supplementary Figures 4 and 5 show that SCD achieves the highest classification performances (AP of 98.72% and AUC of 99.58%), followed by spectral distance (AP of 94.93% and AUC of 98.64%) and by facet distribution distance (AP of 76.10% and AUC of 93.11%). Taken altogether, our results demonstrate that SCD is a very sensitive measure of simplicial complex similarity.

## 3.2 Uncovering biological information from PPI Complexes

In the experiments presented above, we measured the ability of simplets to capture global geometric features of simplicial complexes. In this section, we focus on the local geometry around nodes in simplicial complexes. We assess if the local geometries of proteins in PPI Complexes (which we capture with simplet degree vectors, see section 2.2) relate to their functional annotations using two different methodologies: clustering and enrichment analysis of the resulting clusters, and canonical correlation analysis.

### 3.2.1 Clustering and enrichment analysis

In system biology, studies such as Davis *et al.* (2015) have shown that proteins having similar local wiring patterns in PPI networks tend to have similar biological functions. This suggests that specific protein functions are performed through specific patterns of protein-protein interactions, and that the biological functions of unnanotated proteins can be predicted from their wiring patterns in the PPI network (Milenković and Pržulj, 2008). This may be explained by evolutionary processes, as genomes are believed to have evolved through gene (and sometimes entire genome) duplication and mutation events. Genes with the same origin have similar sequences and their protein product structures, resulting in similarities in the wiring patterns of their PPIs.

Here, we investigate if a similar property holds in our higher dimensional representations of interactomes, i.e., if proteins with similar local geometries (as captured by simplets) also tend to have similar biological functions. To this aim, we cluster proteins having similar local geometries (i.e., having similar simplet degree vectors) and assess if the obtained clusters are functionally enriched in biological functions as follows. For both human and yeast, we computed the simplet degree similarity of the proteins in each of the two models of their interactomes (PPI network and PPI Complex, see section 2.1.1). We used these pairwise similarities as input for spectral clustering (Von Luxburg, 2007), which performs $k$-means clustering on the eigen-vectors of the matrix encoding the pairwise simplet degree similarities between the nodes. Spectral clustering is favored over

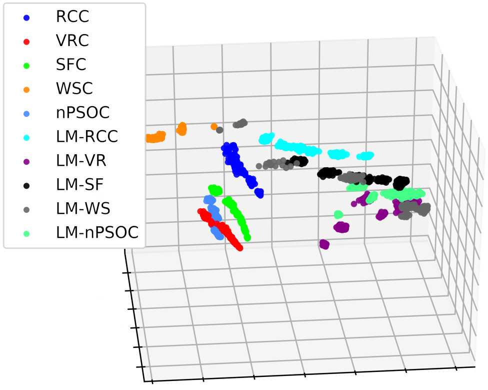

**Fig. 3. Illustration of MDS-based embedding of simplicial complexes from ten random models.** The randomly generated simplicial complexes (color-coded) are embedded into 3D space according to their pairwise SCD distances using multi-dimensional scaling (MDS). The ten models and simplicial complex sizes and densities are described in Section 2.1.2. As described in Section 2.1.2, 25 simplicial complexes are generated for each model and each of its sizes and densities. The grouping of the same colored nodes correspond to simplicial complexes from the same model, which may be of different sizes and densities.

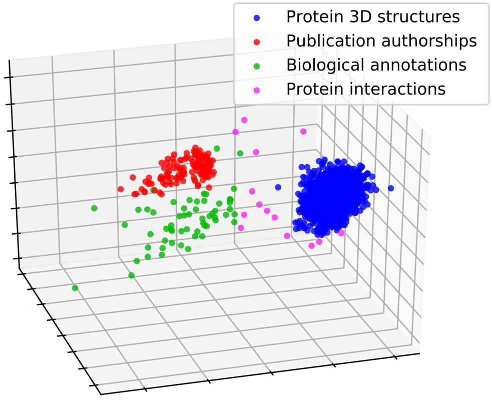

**Fig. 4. Illustration of MDS-based embedding of real-world simplicial complexes based on their SCDs.** The real-world simplicial complexes (color-coded) are embedded into 3D space according to their pairwise SCD distances using multi-dimensional scaling.

traditional $k$-means as it does not make strong assumptions on the shape of the clusters. While $k$-means produces clusters corresponding to convex sets, spectral clustering can solve a more general problem such as intertwined spirals (Von Luxburg, 2007). To account for the randomness of the underlying $k$-means, each clustering experiments is repeated 10 times. As there is no gold-standard way of setting the number of clusters, $k$, we chose the frequently used rule of thumb (Kodinariya and Makwana, 2013), $k = \sqrt{\frac{n}{2}}$, where $n$ in the number of nodes in the simplicial complex. To further motivate our choice of $k$, we performed 10 spectral clusterings for each $k$ from 10 to 150 in steps of 10. For each value of $k$, we measured the consistency of the obtained clusterings by using both their sum of square error and their normalized mutual information scores. We observe that the rule of thumb leads to stable clusterings, as $k = \sqrt{\frac{n}{2}}$ is after the elbows of the two consistency scores. I.e., we set $k = 90$ for human and $k = 54$ for yeast data-set. For comparison purposes, we also generated for both human and yeast one hundred random clusterings having same cluster sizes as the ones obtained by spectral clustering on the PPI Complexes.

Then, we measure the biological coherence of the obtained clustering by the percentage of clusters that are statistically significantly enriched in at least one Gene Ontology (GO) annotation (Ashburner *et al.*, 2000). To this aim, we collected the experimentally validated GO annotations of genes from NCBI's entrez web portal (collected on the $8^{th}$ of March, 2018). We considered GO biological process (GO-BP), GO molecular function (GO-MF), and GO cellular component (GO-CC) annotations separately. A cluster is statistically significantly enriched in a given annotation if the corresponding enrichment $p$-value is lower than or equal to 5% after Benjamini-Hochberg (Benjamini and Hochberg, 1995) correction for multiple hypothesis testing.

As presented in Figure 5, over all ten runs, for both species and for the three GO annotation types, the biological coherence in terms of enriched clusters is larger for the PPI Complexes than for the PPI networks. On average, 79.5% of the clusters from the PPI Complexes are significantly enriched in GO biological process annotations, versus 50.7% for the clusters from the PPI networks. Similarly, 69.8% of the clusters from the PPI Complexes are significantly enriched in GO molecular function annotations, versus 44.8% for the clusters from the PPI networks. Finally, 74.2% of the clusters from the PPI Complexes are significantly enriched in GO biological process annotations, versus 53.1% for the clusters from the PPI networks. These results are all statistically significant (with empirical $p$-values $\leq 1\%$), as the randomly generated clusters are never observed to be as enriched in biological functions than the clusters obtained from the PPI Complex (the random clusterings have, on average, less than 1% of their clusters with at least one enriched function).

These results demonstrate that proteins having similar geometries in PPI Complexes, i.e., that form complex interactions in similar ways, indeed tend to have similar biological functions. This may be due to duplications and divergence of the genome regions coding for these molecular machines. Also, our results show that PPI Complex representation captures more biological annotations than simple PPI network representation of these complex data. This illustrates the importance of modelling and wiring of protein interactomes.

### 3.2.2 Canonical correlation analysis

To quantify the relationships between the local geometry around proteins in simplicial complexes and their biological functions, i.e., to measure how well the simplet degrees of the proteins are predictive of their GO biological process annotations, we adapt the canonical correlation analysis (CCA) methodologies from Yaveroğlu *et al.* (2014). The local geometry around $n$ proteins in a simplicial complex is captured in an $n \times 32$ matrix, $R$, whose entry $R[v][i]$ is the $i^{th}$ simplet degree of node $v$. Similarly, the biological functions of the proteins is captured in an $n \times f$ matrix, $A$, whose entry $A[v][i]$ is 1 if protein $v$ is annotated by term $i$, and 0 otherwise. For both matrices, we excluded the genes that do not have any GO biological process annotations. CCA is an iterative process that identifies linear relationships between the 32 simplet degrees and the $f$ GO biological process annotations. First, CCA outputs two weight vectors, called *canonical variates*, so that the weighted sum of $R$ is maximally correlated with the weighted sum of $A$. The correlation between the two weighted sums is called their *canonical correlation*. After finding the first canonical variates, CCA iterates $min\{32, f\}$ times to find more weight vectors, such that the resulting canonical variates are not correlated with any of the previous canonical variates. We refer the interested reader to Weenink (2003) for the mathematical aspects of CCA.

As presented in Figure 6 and Supplementary Figure 6, the PPI Complex allows for uncovering a larger number of linear relationships that the PPI network model. This is because only 15 out of the 32 simplets can appear in a 1-dimensional simplicial complexes, i.e., a PPI network, which correspond to the 15 2- to 4-node graphlets. Hence, CCA can only produce up-to 15 variates for a PPI network and up-to 32 variates for the PPI Complex. Moreover, these linear relationships have higher canonical

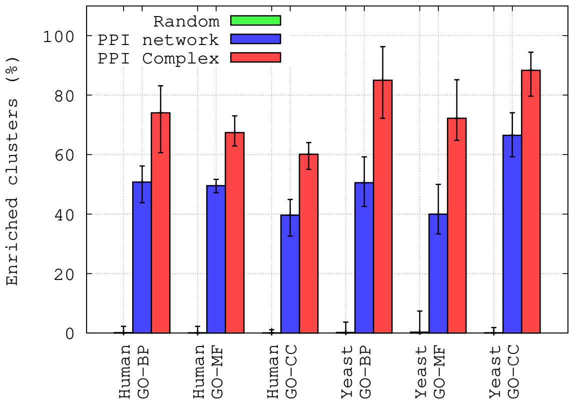

**Fig. 5. Biological relevance of clusters of genes,** as measured by the percentage of clusters having at least one enriched GO annotation. For PPI networks and PPI Complexes, the error bars present minimum, average and maximum enrichment values over 10 runs of spectral clustering, while for Random the error bars present minimum, average and maximum enrichment values over 100 random clustering having same cluster sizes as the ones obtained for PPI Complexes.

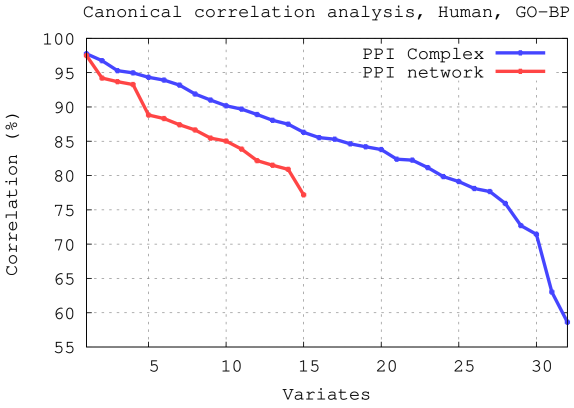

**Fig. 6. Canonical correlation analysis for human.** For a given simplicial complex, canonical correlation produces variates, which are linear combinations of go annotations and linear combinations of simplet degrees that best correlate over the nodes of the simplicial complex. For both models of human interactomes (PPI network and PPI Complex), we plotted for each variate the corresponding correlation value (only statistically significantly correlated variates are presented, with canonical correlation $p$-value $\leq 5\%$).

correlations. This means that by using simplets on the PPI Complexes we can capture more and better quality relationships between local geometry around nodes in simplicial complexes and their biological functions than if we use PPI networks. The same is observed when using GO cellular component and GO molecular function annotations (not shown due to space limitations).

## 4 Conclusion

We demonstrate that by the new way of accounting for multi-scale organization of PPI data both through modeling and new algorithms that we propose, we can uncover substantially more biological information than can be obtained by considering only pairwise interactions between proteins in PPI networks. This pioneering observation can further be utilized to predict biological functions of unnanotated genes, which is a subject of further research.

We demonstrate the existence of the functional geometry in the PPI data by capturing the higher-order organization of these molecular networks by using simplicial complexes. To mine the geometry of simplicial complexes, we propose simplets, which generalize graphlets to simplicial complexes. On randomly generated and real-world datasets, we define a sensitive measure of global geometrical similarity between simplicial complexes. Also, we propose a higher-dimensional, simplicial complex-based model of a species' interactome that we call PPI Complex, which combines protein-protein-interaction and protein complex data. On human and yeast interactomes, by using clustering based on our new simplet-based measures of geometric similarity and cluster enrichment analysis, we show that our PPI Complexes are more biologically coherent than protein-protein interaction networks and that our simplets can efficiently mine PPI Complexes as a new source of biological knowledge. Furthermore, while we focus on simplicial complexes emerging from molecular network organization, our methodology is generic and can be applied to multi-scale datasets from any scientific field, such as the multi-scale network data from physics, social sciences and economy.

## Funding

This work was supported by the European Research Council (ERC) Starting Independent Researcher Grant 278212, the European Research Council (ERC) Consolidator Grant 770827, the Serbian Ministry of Education and Science Project III44006, the Slovenian Research Agency project J1-8155 and the awards to establish the Farr Institute of Health Informatics Research, London, from the Medical Research Council, Arthritis Research UK, British Heart Foundation, Cancer Research UK, Chief Scientist Office, Economic and Social Research Council, Engineering and Physical Sciences Research Council, National Institute for Health Research, National Institute for Social Care and Health Research, and Wellcome Trust (grant MR/K006584/1).

## References

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., *et al.* (2000). Gene ontology: tool for the unification of biology. *Nature Genetics*, **25**(1), 25.

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

Benjamini, Y. and 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)*, pages 289–300.

Borg, I. and Groenen, P. J. (2005). *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media.

Chatr-Aryamontri, A., Oughtred, R., Boucher, L., Rust, J., Chang, C., Kolas, N. K., O'Donnell, L., Oster, S., Theesfeld, C., Sellam, A., *et al.* (2017). The biogrid interaction database: 2017 update. *Nucleic Acids Research*, **45**(D1), D369–D379.

Collins, A., Zomorodian, A., Carlsson, G., and Guibas, L. J. (2004). A barcode shape descriptor for curve point cloud data. *Computers & Graphics*, **28**(6), 881–894.

Cook, S. A. (1971). The complexity of theorem-proving procedures. In *Proceedings of the third annual ACM symposium on Theory of computing*, pages 151–158. ACM.

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

Erdös, P. and Rényi, A. (1959). On random graphs. *Publicationes Mathematicae*, **6**, 290–297.

Estrada, E. and Rodriguez-Velazquez, J. A. (2005). Complex networks as hypergraphs. *arXiv preprint physics/0505137*.

Estrada, E. and Ross, G. J. (2018). Centralities in simplicial complexes. applications to protein interaction networks. *Journal of theoretical biology*, **438**, 46–60.

Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., *et al.* (2017). The reactome pathway knowledgebase. *Nucleic Acids Research*, **46**(D1), D649–D655.

Faisal, F. E., Newaz, K., Chaney, J. L., Li, J., Emrich, S. J., Clark, P. L., and Milenković, T. (2017). Grafene: Graphlet-based alignment-free network approach integrates 3d structural and sequence (residue order) data to improve protein structural comparison. *Scientific Reports*, **7**(1), 14890.

Fox, N. K., Brenner, S. E., and Chandonia, J.-M. (2013). Scope: Structural classification of proteins extended, integrating scop and astral data and classification of new structures. *Nucleic Acids Research*, **42**(D1), D304–D309.

Hausmann, J.-C. *et al.* (1995). On the vietoris-rips complexes and a cohomology theory for metric spaces. *Annals of Mathematics Studies*, **138**, 175–188.

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

Kuchaiev, O. and Pržulj, N. (2011). Integrative network alignment reveals large regions of global network similarity in yeast and human. *Bioinformatics*, **27**(10), 1390–1396.

Kuchaiev, O., Milenković, T., Memišević, V., Hayes, W., and Pržulj, N. (2010). Topological network alignment uncovers biological function and phylogeny. *Journal of the Royal Society Interface*, page rsif20100063.

Linial, N. and Meshulam, R. (2006). Homological connectivity of random 2-complexes. *Combinatorica*, **26**(4), 475–487.

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

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

Meshulam, R. and Wallach, N. (2009). Homological connectivity of random k-dimensional complexes. *Random Structures & Algorithms*, **34**(3), 408–417.

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

Munkres, J. R. (1984). *Elements of algebraic topology*, volume 4586. Addison-Wesley Longman.

Muscoloni, A. and Cannistraci, C. V. (2018). A nonuniform popularity-similarity optimization (npso) model to efficiently generate realistic complex networks with communities. *New Journal of Physics*, **20**(5), 052002.

Newman, M. (2010). *Networks: an introduction*. Oxford university press.

Penrose, M. (2003). Random geometric graphs. *Oxford Studies in Probability*, **5**.

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

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

Pu, S., Wong, J., Turner, B., Cho, E., and Wodak, S. J. (2009). Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Research*, **37**(3), 825–831.

Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., and Mewes, H.-W. (2010). Corum: the comprehensive resource of mammalian protein complexes 2009. *Nucleic Acids Research*, **38**(suppl_1), D497–D501.

Ryan, C. J., Cimermančič, P., Szpiech, Z. A., Sali, A., Hernandez, R. D., and Krogan, N. J. (2013). High-resolution network biology: connecting sequence with function. *Nature Reviews Genetics*, **14**(12), 865.

Sprinzak, E., Sattath, S., and Margalit, H. (2003). How reliable are experimental protein–protein interaction data? *Journal of Molecular Biology*, **327**(5), 919–923.

Vijayan, V., Saraph, V., and Milenković, T. (2015). Magna++: Maximizing accuracy in global network alignment via both node and edge conservation. *Bioinformatics*, **31**(14), 2409–2411.

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

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of "small-world" networks. *nature*, **393**(6684), 440.

Weenink, D. (2003). Canonical correlation analysis. In *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam*, volume 25, pages 81–99. University of Amsterdam.

Wilson, R. C. and Zhu, P. (2008). A study of graph spectra for comparing graphs and trees. *Pattern Recognition*, **41**(9), 2833–2841.

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

Zhou, D., Huang, J., and Schölkopf, B. (2007). Learning with hypergraphs: Clustering, classification, and embedding. In *Advances in neural information processing systems*, pages 1601–1608.