Hyperedge entanglement in high-order multilayer networks

Alessandro Muscoloni[1,2,†], Ilyes Abdelhamid[2-4,†], Julius L. Decano[4], Edwin Souza[4], Enrico Maiorino[5], Masanori Aikawa[4,5], Edwin K. Silverman[5], Amitabh Sharma[4-7,*] & Carlo Vittorio Cannistraci[1,2,*]

[1]Center for Complex Network Intelligence (CCNI) at the Tsinghua Laboratory of Brain and Intelligence (THBI), Department of Bioengineering, Tsinghua University, 160 Chengfu Rd., SanCaiTang Building, Haidian District, 100084, Beijing, China
[2]Biomedical Cybernetics Group, Biotechnology Center (BIOTEC), Center for Molecular and Cellular Bioengineering (CMCB), Center for Systems Biology Dresden (CSBD), Cluster of Excellence Physics of Life (PoL), Department of Physics, Technische Universität Dresden, Tatzberg 47/49, 01307 Dresden, Germany.
[3]Lipotype GmbH, Dresden, Germany.
[4]Center for Interdisciplinary Cardiovascular Sciences, Division of Cardiovascular Medicine, Department of Medicine, Brigham Women's Hospital, Harvard Medical School, Boston, MA, USA.
[5]Channing Division of Network Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA.
[6]Center for Complex Network Research, Department of Physics, Northeastern University, Boston, Massachusetts, USA.
[7]Center for Cancer Systems Biology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA.

[†]Joint first authors
[*]Corresponding authors:
Amitabh Sharma (reash@channing.harvard.edu)
Carlo Vittorio Cannistraci (kalokagathos.agon@gmail.com)

Abstract

Many real complex systems present multilayer structure where high-order metadata on one layer refers to dyadic data on a lower layer. Significant progresses to analyse high-order metadata under the assumption of community organization have been done. However, there are no planted communities in real-world networks, and the necessity of new frameworks to analyze high-order metadata regardless of community organization has been raised.

Here, we propose to adopt hyperedge organization. Predicting ‘entanglements’ between a hyperedge and nodes scattered in the rest of the network might suggest structural or functional liaisons, without assumption of any community organization. We introduce a novel concept: *hyperedge entanglement (HE)*, which associates to each hyperedge an *entangled hyperedge*, by means of a network operator that predicts significant ‘interactions at distance’ between network nodes and existing hyperedges. We also introduce a new challenge termed *hyperedge entanglement prediction (HEP)*, and an algorithm to perform this task. We evaluated HEP performance on social, biological and synthetic data where, given only topology and hyperedges (such as communities or functional modules), the goal is to predict whether nodes not connected to a certain hyperedge might be candidates for a significant entanglement. Finally, as real application in diseasome systems biomedicine, we perform HEP on the human protein interactome to predict unknown gene entanglements with the COPD disease gene hyperedge. HEP predictions are validated by biological experiments, enlarging our understanding of molecular mechanisms behind COPD/aneurysm comorbidity.

1. Introduction

Research on the modular dyadic organization of complex networked systems has become increasingly popular in many scientific disciplines diverse from physics, including socio-behavioral[1–3], biological[3,4] and maritime[5] science. Progress has been achieved also in modeling community organization in artificial networks by random graphs without[6] and with geometry[7–10]. Community detection is an attractive area of investigation in the field of complex networks[11], and recent studies investigated how network embedding can enhance community detection[12–16]. However, how topological community organization should be interpreted and harmonized with node metadata group information - and what are the limitations of such endeavor – is a controversial topic currently under investigation[17]. When Science is stuck in front of a dilemma or conceptual obstacle – let us term it a 'Gordian Knot' – then it is the moment that new ideas emerge. These ideas are not always the final solutions to the problem but represent a detour towards a new definition of the conceptual framework inside which flourishing of new interpretations and methods can bring toward 'unleashing the Gordian Knot'. Here, we are in front of a 'Community Knot'; and this study aims to select a different conceptual and theoretical 'blade' and to sharpen it, in the hope that others in the future might success to 'cut the Community Knot'.

The current conceptual framework is that many real networks are composed of several communities, also referred to as clusters or modules. An important and accepted property to distinguish a community in respect to others, is that its nodes have a higher probability to link to one another than to nodes that belong to a different community[18,19]. These groups of densely connected nodes can have entirely different meanings depending on the complex system under observation. Frequently, communities might be matched with groups of nodes that, according to metadata, share common properties or play similar roles. In a social network, a community can have a certain overlap with a tight group of friends, people sharing the same hobbies or the same job. In a protein-protein interaction (PPI) network it can have a certain correspondence to proteins involved in a specific process, function or pathway. Whereas, in a citation network, it might be to a certain extent associated to a set of scientific papers with a related topic. Yet the matching between the node metadata information and the topological community segregation is not always respected. In some cases, this is due to a multiplayer data structure. For instance, we can have a scientific co-authorship network, where each node is an author, and each link indicates a collaboration in a study between two authors, whereas the metadata information is the research institution of the authors. In other cases, such as for PPI networks, the network topology is incomplete and metadata information too, therefore some modules indicating a cohort of genes whose proteins are involved in a disease are composed by nodes which are scattered in the network without a true community organization between them, that is also one of the cases we will investigate in the present study. In all these scenarios, we can still say that those groups of nodes are involved in a hyperedge. It means that they are linked (all together) at high-order by means of the metadata information without a detailed specification of their dyadic interactions.

In this study we set a conceptual and computational framework that, at the best of our knowledge, has not been investigated in the previous literature. We start from the observation that many real complex systems present multilayer structure where high-order metadata on one layer refers to dyadic data on another layer. For instance, hyperedge organization (in the first layer) might be associated to node functional segregation and integration of dyadic networks (in the second layer). Predicting ‘entanglements’ between a hyperedge and nodes scattered in the rest of the network might suggest either structural (direct) or functional (indirect) liaisons. Hence, we introduce the novel concept of hyperedge entanglement (HE), which associates to each hyperedge an entangled hyperedge, by means of a network operator that predicts significant ‘interactions at distance’ between network nodes and existing hyperedges. To this aim we also introduce a new challenge termed hyperedge entanglement prediction (HEP), and an algorithm to perform this task.

Depending on the area of application, entanglement can be interpreted as missing association due to a lack of observation or as a future connection. However, we would like to clarify that the topic is related, but not equivalent, to the classical link prediction. Indeed, in the classical link prediction problem, given the network topology, the goal is to identify the connections between nodes that have not been observed and that are likely to occur[20]. In other words, the links to predict are dyadic interactions between two nodes. Differently, in the problem that we are going to address in this study, the entanglement to predict is between a node and a hyperedge. In addition, we can also predict an entangled hyperedge which is an ensemble of nodes that are significantly entangled to the ‘seed’ hyperedge. In the rest of the study, we will discriminate these two different tasks respectively as node2hyperedge entanglement (N2HE) and hyperedge entanglement prediction (HEP); being N2HE a subroutine of HEP. The motivation for studying HEP is that connections between a hyperedge and other nodes in the rest of the network are pivotal bridges which might suggest either missed node membership or integration with other functionally overlapping hyperedges.

Examples of applications could be: i) to spot unknown associations between genes and a disease module whose proteins are scattered in the PPI network; ii) to predict contacts between terrorists around the world and an existing terrorist cell; iii) to predict contagions, in a contact tracing network, between a cluster of COVID 19 infected people that emerged in a certain time and location, and unknown people that are susceptible to the disease. Among these three examples, we recovered the data to investigate the first real case scenario, whereas we hope that in future studies we might procure the data to investigate the others. After evaluation of HEP performance on social, biological and synthetic networks, we indeed offer an example of real application in systems biomedicine of the diseasome[21]. The diseasome is the bipartite network where diseases on one layer connect to associated genes on the other layer. Hence a disease can be represented by a hyperedge of genes that can be projected on the PPI network nodes. We compute HEP on the human protein interactome in order to predict the entanglement hyperedge that contains genes significantly entangled with the chronic obstructive pulmonary disease (COPD) hyperedge. We consider this peculiar disease because it impacts large part of the worldwide population, and because we have specific expertise to biologically validate some of the predictions by techniques such as co-immunoprecipitation and gene silencing. The result of this validation confirms HEP computational prediction, and spread light on unknown molecular mechanisms behind COPD/aneurysm comorbidity.

2. Results

2.1 Hyperedge entanglement: definition and algorithm

In this section, we will firstly introduce the theoretical definitions related to the hyperedge entanglement concepts and then we will provide the description of the algorithm for computing the significance (a p-value) of node2hyperedge entanglement. From here forward, we will consider as a working framework a high-order multilayer network (Fig. 1), which is a multiplex composed of two layers: one high-order (Fig. 1A) and one dyadic (Fig. 1B). The first high-order layer is a hypergraph $HG(V,H)$, where V is the set of vertices (or nodes) and H is the set of hyperedges; the second dyadic layer is a graph $G(V,E)$, where V is the same set of vertices as the hypergraph HG and E is the set of edges (or links). In contrast to the dyadic graph, where an edge connects exactly two vertices, in the hypergraph a hyperedge can connect any number of vertices. Finally, in this study we will refer to the terms network and graph as interchangeable, but we warn the reader that this is a simplification that we apply only for easing the readability of the study. Although these two terms are often used in the literature of complex systems interchangeably, the graph is in general a pure mathematical structure, whereas the network might be better interpreted as a graph that emerges from the underlying dynamics of a process in a complex system.

2.1.1 Hyperedge entanglement (HE)

The hyperedge entanglement is an operation that associates to a given hyperedge h_j an *entangled hyperedge* eh_j (Fig. 1A), whose members are the nodes $v_i \in V$ that satisfy both the node2hyperedge entanglement necessary (ENC) and sufficient (ESC) conditions, that we will describe below. Intuitively, the *entangled hyperedge* eh_j includes all the nodes that have a significant *'interaction at distance'* with the group of nodes in a hyperedge from which they are disconnected both in the hypergraph *HG* and the graph *G* (Fig. 1).

The hyperedge entanglement should not be confused with the entanglement concept in quantum physics[22], where the quantum state of each entangled particle of the group cannot be described independently of the state of the others. Indeed, in our scenario, the given hyperedge h_j can be described independently of the entangled hyperedge eh_j, while the opposite is not true. In other words, in the quantum entanglement there is mutual (bidirectional) dependency between the particles, whereas in the here defined hyperedge entanglement there is unidirectional dependency of the entangled hyperedge eh_j on the given hyperedge h_j.

2.1.2 node2hyperedge entanglement necessary condition (ENC)

Given a node v_i and a hyperedge h_j (Fig. 1), the ENC is satisfied and v_i becomes a *candidate node for entanglement* c_{ij} if and only if:

$$(v_i \notin h_j) \wedge \left[(v_i, v_k) \notin E, \forall v_k \in h_j\right] \tag{1}$$

In other words, the ENC guarantees that the *candidate node* c_{ij} is *disconnected* with the other nodes in the hyperedge h_j, in both the hypergraph (by not being a member of the hyperedge) and the graph (by not having links to the nodes of the hyperedge). Note that the condition (1) is a necessary and sufficient condition to become a candidate node for entanglement, whereas being a candidate node is a necessary condition for the entanglement, because we can entangle only a node and a hyperedge which are disconnected both in *HG* and *G*.

2.1.3 node2hyperedge entanglement sufficient condition (ESC)

doi:10.20944/preprints202012.0500.v1

Given (Fig. 1) a hyperedge h_j, a candidate node c_{ij} (which is defined as candidate because it satisfies the ENC for h_j) and a graph G (which can include not only topological information but any meta-information such as the nodes coordinates in a geometrical space), the ESC depends from the type of *entanglement operator EO* (which estimates the entanglement of c_{ij} with h_j) and the statistic S defined on the EO (which estimates the significance of the entanglement). Formally, we can write that the ESC is satisfied and c_{ij} becomes a member of the entangled hyperedge eh_j if and only if:

$$S\left[EO\left(c_{ij}, h_j, G\right)\right] \leq \varepsilon \tag{2}$$

This means that if the statistic is lower than or equal to a certain significance threshold ε, then the ESC is satisfied. Without lack of generality, in our case we define the ESC according to an entanglement operator that is based on topological information in the graph G between the node c_{ij} and the nodes of the hyperedge h_j. In particular, we develop an algorithm that combines a topological entanglement operator *TEO* (based on link prediction in the graph G) and a statistical test (based on a null model), in order to compute a node2hyperedge entanglement p-value. The p-value assesses the extent to which the candidate node c_{ij} is significantly entangled to the hyperedge h_j. If the p-value is significant (according to a significance threshold ε), then the ESC is satisfied, and the node c_{ij} becomes a member of the entangled hyperedge eh_j:

$$S\left[EO\left(c_{ij}, h_j, G\right)\right] = pvalue\left[TEO\left(c_{ij}, h_j, G\right)\right] \leq \varepsilon \tag{3}$$

Obviously, the significance threshold value determines the number of nodes that satisfy the ESC, the lower the ε the smaller the size of the eh_j, hence the significance threshold can be interpreted also as a tuning parameter that allows to restrict the selection of nodes entangled with a hyperedge h_j to a required number.

We stress that the design of the EO is not confined to the mere adoption of topological information. For instance, we define also a second interesting class of entanglement operators that are based on network geometry (geometrical entanglement operator: *GEO*), and that estimate the entanglement of a node c_{ij} with the hyperedge h_j on the basis of their distance in the geometrical space in which a network lies or is embedded[12]:

$$S\left[EO\left(c_{ij}, h_j, G\right)\right] = S\left[GEO\left(c_{ij}, h_j, G\right)\right] \leq \varepsilon \tag{4}$$

Theoretically, also hybrid operators TGEO that integrate topological and geometrical information can be designed. However, here we will focus on TEO and we will leave to a future study the chance to explore GEO and TGEO. The reason to prioritize TEO investigation is that the literature offers already solid evidences that topological link prediction is currently performing better than geometrical-embedding link prediction[4].

2.1.4 Global node2hyperedge entanglement sufficient condition in case of more than two layers

Although the examples discussed in this study will consider only two layers, here for completeness we will discuss also the general case of a high-order multilayer network composed of *n* graph layers and *m* hypergraph layers:
$[G_1(V_1, E_1),\ G_2(V_2, E_2),\ \dots\ ,\ G_n(V_n, E_n)]$ and $[HG_1(V_1, H_1),\ HG_2(V_2, H_2),\ \dots\ ,\ HG_m(V_m, H_m)]$.
We can define a node2hyperedge global entanglement sufficient condition (GESC) when a node v_i respect the ENC necessary condition (1) for h_{jt} (where *t* indicates the hypergraph layer HG_t) in all possible *n* graphs, and becomes a global candidate node gc_{ijt}.
Then, the GESC for the global candidate node gc_{ijt} across the high-order multilayer network is:

$$\bar{S} = \frac{1}{n} \cdot \sum_{l=1}^{n} S\big[EO\big(c_{ijt}, h_{jt}, G_l\big)\big] \leq \varepsilon \tag{5}$$

In brief, the mean statistic $\bar{S}$ computed across all the *n* graphs should be lower than or equal to a significance threshold in order to confirm that the global candidate node gc_{ijt} has a global significant entanglement with h_{jt}. Note that when *n* = 1 and *m* > 1 this formula is equivalent to ESC (2), although it is defined for each h_{jt}.

2.1.5 Algorithm for node2hyperedge entanglement (N2HE) p-value

In this section and Fig. 2, we describe the node2hyperedge entanglement (N2HE) algorithm:
- given in input a hyperedge h_j, a candidate node c_{ij}, a graph G, a link-prediction-based topological entanglement operator $LP(G)$ and a significance threshold ε;
- N2HE computes and offers in output a node2hyperedge entanglement p-value, to assess if the ESC is satisfied.

The link predictor $LP(G)$ is an operator that, exploiting topological information of the graph G, can associate a similarity score to any disconnected node pair, suggesting the likelihood for a link between them to exist. Hence, here we define a topological entanglement operator that estimates an ensemble ‘interaction at distance’ between c_{ij} and h_j nodes by means of link prediction. Although in principle any link prediction algorithm can be adopted, without lack of generality we are going to focus on local approaches, which are able to perform a likelihood prediction of a missing link between two nodes, by exploiting only local information that is related to their neighborhood[4]. The rationale for this choice is that local approaches allow the prediction just for a selected subset of missing links (such as the ones that could connect a node to a community). This can save a significant amount of computational time for application on real networks, where we might be interested to predict exclusively the entanglement of one node with one hyperedge. In contrast, many highly-performing global approaches such as Structural Perturbation Method[23], since they are based on global operations, compute all missing links likelihoods at one time, requiring much more time. Hence, local methods allow to design efficient N2HE predictors that adapt their computational time to the query of the user, while preserving performance. Local approaches are therefore the proper candidates to perform computational demanding tests such as the ones we will present in this study.

One of the simplest and well-known local-based link prediction methods is the common neighbors (CN) index, which states that the higher the number of CN between two nodes, the higher the likelihood to be connected[24]. Similar indexes have been developed introducing a normalization factor to the CN rule, such as in the Resource-Allocation (RA)[25], Jaccard (J) and Adamic-Adar (AA)[26] indexes. In 2013 Cannistraci et al.[4] introduced the local-community paradigm (LCP) theory, suggesting that local-based topological link prediction should complement the information content related with the common neighbours nodes using also the topological knowledge emerging from the cross-interactions between them. Initially detected in brain-network self-organization topology[4] and then extended to any monopartite and bipartite[27] complex network, the LCP theory derives from a purely topological inspired interpretation of a local-learning-rule of neuronal networks named Hebbian learning rule. Based on the LCP theory, Cannistraci et al.[4] designed several local, parameter-free and model-based deterministic rules for topological link prediction in both monopartite and bipartite networks, which in a later study have been also re-interpreted as network automata that participate to learn and form new structures on a network by growing according to some local mechanisms of self-organization known as Cannistraci-Hebb (CH) models[28]. While all the previously mentioned indexes for link prediction on monopartite networks are based on paths of length two (L2), a recent and brilliant study by Kovács et al.[29] proposed an algorithm based on paths of length three (L3) for prediction on protein interactomes. A subsequent study on network automata for link prediction of Muscoloni et al.[28] defined CH network automata models on paths of length n (where n indicates any length), and demonstrated that Kovács et al. similarity is equivalent to RA defined on L3, therefore a subcase of generalized network automata defined on paths of length n. A wide comparative study on local-based link predictors[30] has recently shown that some CH models (both L2 and L3) overcome in prediction performance all the other methods including Kovács et al. similarity.

The algorithmic steps of N2HE, visually shown in part of Fig. 2, are reported below:

a. Fig. 2b,c: using the link predictor $LP(G)$, compute the similarity score s_{v_i,v_k} between node v_i and every node $v_k \in V$ that is not connected to v_i in the graph G, i.e. $(v_i, v_k) \notin E$. (Note: all the disconnected pairs are considered, because this will return useful in the implementation of the hyperedge entanglement predictor described in the next section that computes the entangled hyperedge of any hyperedge h_j).

b. Fig. 2d,e: compute the node2hyperedge entanglement score, as the average similarity score between candidate node c_{ij} and the nodes of the hyperedge h_j:

$$\sigma_{ij} = \operatorname*{avg}_{v_k \in h_j} s_{c_{ij},v_k}$$

where avg is an average operator, such as mean, median or mode.

c. Fig. 2f: generate a set of random hyperedges $h_1^r \dots h_M^r$ with the same size as the hyperedge h_j. Here, we set $M = 1000$, but in relation to available computational resources this number can be also increased. The members of each hyperedge are sampled uniformly at random among all the nodes except the node c_{ij}, its neighbors in the graph G and the members of the hyperedge h_j. More formally, they are sampled from the set:

$$V' = V - \left(\{c_{ij}\} \cup \{v_k \in V: (c_{ij}, v_k) \in E\} \cup \{v_k \in h_j\}\right)$$

d. Fig. 2f: for each random hyperedge h_j^r, compute the node2hyperedge entanglement score σ_{ij}^r with node c_{ij}, resulting in a null-distribution of node2hyperedge entanglement scores $\sigma_{i1}^r \dots \sigma_{iM}^r$.

e. Fig. 2f: compute an empirical p-value p_{ij} representing the node2hyperedge entanglement p-value between the node c_{ij} and the hyperedge h_j:

$$p_{ij} = \frac{m+1}{M+1}$$

where m is the number of node2hyperedge entanglement scores from the null-distribution $\sigma_{i1}^r \dots \sigma_{iM}^r$ that are greater than or equal to the observed one σ_{ij}.

In this study we set a significance threshold of $\varepsilon = 0.05$, therefore if $p_{ij} \le 0.05$ then the candidate node c_{ij} and the hyperedge h_j satisfy the ESC, hence node c_{ij} becomes a member of the entangled hyperedge eh_j.

2.1.6 Hyperedge entanglement predictor (HEP)

In this section and Fig. 2, we describe the algorithm of the *hyperedge entanglement predictor (HEP)*:

- given in input a graph $G(V,E)$, a hypergraph $HG(V,H)$ and an algorithm for node2hyperedge entanglement N2HE (such as the one based on link-prediction and proposed in the previous section);
- HEP is an algorithm that finds the entangled hyperedge eh_j associated to every hyperedge $h_j \in H$.

More in details, for each pair $(c_{ij} \in V, h_j \in H)$, where c_{ij} is a node $v_i \in V$ that satisfies the necessary condition ENC (1) for $h_j \in H$, the HEP computes the node2hyperedge entanglement p-value p_{ij}, assessing whether the sufficient condition ESC (2) for $c_{ij} \in eh_j$ is also satisfied. Then, for each hyperedge $h_j \in H$, the set of nodes $v_i \in V$ that satisfy both ENC and ESC are the members of the entangled hyperedge eh_j. This could be also simplified as: for each hyperedge $h_j \in H$, the candidate nodes $c_{ij} \in V$ that satisfy the ESC are the members of the entangled hyperedge eh_j. Within the algorithmic pipeline of the HEP, and therefore of the N2HE algorithm for the node2hyperedge entanglement p-value (Fig. 2), we have considered several variants, which are summarized below:

- Fig. 2b: 12 different link predictors $LP(G)$: 8 of them are L2-based link prediction methods (CAR, CAA, LCL, CJC, RA-L2, CH1-L2, CH2-L2, CH3-L2) and 4 of them are L3-based (RA-L3, CH1-L3, CH2-L3, CH3-L3). The details about the mathematical formula of the link predictors are described in the Methods section. The rationale to select specifically these predictors is given in the previous section 2.1.5.
- Fig. 2d: 3 different average operators (avg) to compute the node2hyperedge entanglement score: mean, median, mode. Since the similarity scores between a node and the members of a hyperedge can be all different, the mode is not always appropriate to use. For this reason, instead of the mode we compute a probability density estimate of the similarity scores based on a kernel smoothing function ('ksdensity' function in MATLAB), evaluated at 100 linearly spaced points between the minimum and the maximum similarity score, and we select as mode the point with the highest probability.
- Fig. 2g: 2 options for p-values correction, when the node2hyperedge entanglement p-value is computed between a certain node and several hyperedges, we consider as an option a Benjamini-Hochberg correction (C) to adjust for multiple hypothesis testing, whereas the alternative option is to not perform any correction.

Note that when the node2hyperedge entanglement p-value is computed between a certain node and several hyperedges, the computation of the similarity scores at step (a) of the N2HE algorithm only needs to be performed once. In addition, we clarify that in this study we set a significance threshold of 0.05 for the p-values, but the user of the HEP algorithm is free to choose it. If the members of the entangled hyperedge are too many with a significance threshold of 0.05, the downstream analysis can be focused on the most important ones by setting a stricter threshold.

The computational time complexity of the HEP algorithm is the maximum between $O(N^2)$ and the complexity of the chosen link predictor. Considering the link predictors adopted in this study, in case of sparse networks, the complexity is $O(N^2)$. Please, refer to Suppl. Information for a detailed analysis. The MATLAB implementation of the HEP algorithm is available at: https://github.com/biomedical-cybernetics/hyperedge_entanglement.

2.2 node2hyperedge entanglement in real networks

In order to test the algorithmic variants on real data, we collected 5 real networks for which metadata representing the nodes membership to a certain group are available. Nodes belonging to the same group share a common feature and can be considered as a hyperedge. Having a network and a set of hyperedges (a hypergraph) on the same nodes of the network, we fall within our working framework of a high-order multilayer network. Three out of five analysed networks are social and have non-overlapping hyperedges, representing the membership of the nodes to a social community. The remaining two networks are biological and have overlapping hyperedges, representing annotations related to biological metadata. This ensures that we can test our algorithm on different types of complex networked systems. Please refer to the Methods section for a detailed description of the networks and to Suppl. Table 1 for a summary of their topological properties.

The basic idea of our evaluation framework is to remove a test node v_{ij} which is member of a hyperedge h_j - as well as the direct network connections between this test node and the other h_j hyperedge members - and to test whether, once removed from h_j, v_{ij} appears in the entangled hyperedge eh_j and not in other entangled hyperedges. According to these concepts, we build a positive and negative set of node-hyperedge pairs. The positive set is made by all the node-hyperedge pairs such that each test node has at least one link (that is removed) to the other members of the hyperedge and at least one link to nodes that are not members. For each node-hyperedge pair in the positive set we define a related negative pair, made by the same test node and by another hyperedge whose members have the highest average shortest path to the node (considering only pairs that satisfy the ENC). All the possible pairs $(v_{ij} \in h_j, h_j)$ in a hypergraph are tested and a measure of performance based on the correct positive and negative entanglements predicted by HEP is applied. More in details, once obtained predictions for all the pairs in the positive and negative set, we evaluate the performance using the Matthews correlation coefficient (MCC) because it is a measure that accounts for effect size and unbalance in data[31]. The precise algorithm implemented for the evaluation procedure is provided in the Method section.

During the implementation of this evaluation framework, for each node-hyperedge pair evaluated, after a test node is removed from a hyperedge, the network connectivity changes, therefore the computation of the link predictor similarity scores in the N2HE algorithm needs to be performed again regardless of the fact that the same node has been already evaluated for another hyperedge. For this reason, local link predictors are specifically useful in this scenario, since they allow to efficiently compute the similarity scores only for those specific disconnected node pairs that are needed, saving a significant amount of computational time.

On the contrary, global link predictors would require to compute at each round of the evaluation the similarity scores of all the missing links in the network, making this evaluation unfeasible on middle and large networks.

Fig. 3a-c reports the results on 3 real social networks (see section 4.2 for detailed information) with non-overlapping hyperedges (Football, Opsahl_10, Polbooks). Football is a network that presents games between division IA colleges during regular season fall 2000. Hyperedges associate teams belonging to the same conference. The Opsahl_10 network comes from the research team of a manufacturing company, nodes represent employees and edges indicate frequent work interactions. The hyperedges group the employees by the company locations (Paris, Frankfurt, Warsaw, and Geneva). Polbooks is a network that represents frequent co-purchases of books concerning US politics on amazon.com. Hyperedges associate the books with the same political orientation such as conservative, neutral or liberal.

For each link predictor, Fig. 3a-c shows the MCC of the best algorithmic variant (regardless of average operator and p-value correction), whereas in the Suppl. Tables 2-4 all the variants are reported. Note that since there is only one MCC value for each network, no standard error is shown. The figure presents the results also for a random link predictor, obtained by assigning random similarity scores to disconnected node pairs. In Football (Fig. 3a) the overall performance is below MCC=0.4 and all the L3-based predictors outperform the L2-based predictors. Also in Opsahl_10 (Fig. 3b) the L3-based predictors obtain the best MCC (which is around 0.4), while most of the L2-based predictors have performance close to random. In Polbooks (Fig. 3c) the overall MCC of the methods is higher (below 0.6). This seems the only network in which the specific rule of the link predictor is more important than the path length, indeed the two CH2-based link predictors (based on the Cannistraci-Hebb modelling with L2 and L3) are the top performing. In general, for Polbooks the L3-based predictors have similar performance to the one of some L2-predictors (CH2-L2, CH3-L2 and RA-L2), while the other L2-predictors are lower.

The link predictors CH2-L2, CH3-L2 and RA-L2 differ from the other L2-predictors because they perform an explicit and unconditioned minimization of the links external to the local community. Instead, for instance, in CH1-L2 this minimization is conditioned by the existence of the links internal to the local community (see Methods for details). For this reason CH2-L2, CH3-L2 and RA-L2 perform differently from the other L2 methods and, since they are considered forming a homogenous group of methods, in this study they will be represented using the same color (light green) in the figures. Previously, LCP-based methods were focused on the maximization of links internal to the local community[4], instead the recently proposed Cannistraci-Hebb (CH) models emphasize the minimization of the external links, while retaining or not the maximization of the internal links[28]. In particular, CH3 is a new model introduced in this study that is exclusively based on the minimization of the external links (see Methods for details).

Fig. 3d-e reports the results for 2 real biological networks with overlapping hyperedges. S. Cerevisiae PPI stems from the union of 3 recent protein-protein interaction networks of *Saccharomyces cerevisiae*, the hyperedges associate proteins with the same significantly enriched GO terms. E. Coli metabolic is one of the most elaborate metabolic network reconstructions currently available (see section 4.2 for more details), the hyperedges group metabolites with the same metabolic pathway annotation. In these networks the overall MCC reaches much higher levels, up to around 0.9 (in Suppl. Tables 5-6 all the algorithmic variants are reported). In S. Cerevisiae PPI (Fig. 3d) the ranking of the methods is similar to the one seen in Football (Fig. 3a), with L3 methods as best performing (in agreement with link prediction results of Muscoloni et al.[28]), followed by CH-L2 methods (CH2-L2, CH3-L2 and RA-L2), and then by the other L2-based predictors (in agreement with the results of Cannistraci[32], discussing the importance of minimizing external links in PPI networks). We confirm that also in our tests L3-methods overcome L2-methods as reported in previous literature by Kovács et al.[29]. The E. Coli metabolic network (Fig. 3e) seems instead to represent an exception with respect to the other trends. Indeed, RA-L2, CH3-L2 and CH2-L2 obtain the best performances, followed by the other L2-based and then L3-based predictors. The molecular network of the *Escherichia coli* K-12 MG1655 strain accounts for many different reactions where other pairs of metabolites help to take place (exchange of a proton or a phosphate moiety, for example), playing similar role to ATP or ADP. Therefore, the co-occurrence of these currency metabolites (ATP, ADP, water, and so on) in many reactions leads to high clustering, and therefore a better fit with the L2-based models.

The last panel (Fig. 3f) reports the mean and standard error of the MCC over the 5 real networks considered. It highlights that the L3-based methods represent overall the most robust choice, followed by the subset of L2-based methods that explicitly minimize the external links in the CH model (CH2-L2, CH3-L2 and RA-L2), and then by the other L2-based predictors.

Altogether, from the results one might notice that even for social networks, typically associated to a L2-based structural organization, the L3-based methods obtain overall higher performance. In these social datasets the hyperedges represent community-related metadata, groups of nodes that tend to have more connections between themselves than with the rest of the network. Therefore, the majority of the L2 paths are made by links between the members of the hyperedge. In this evaluation framework, for each node-hyperedge pair in the positive set, the network links between the node and the other members of the hyperedge are removed, therefore most of existing L2 paths are eliminated. This is likely the reason why L2-based methods are outperformed by L3-based methods in HEP on social networks.

2.3 Latent geometry plays a role in node2hyperedge entanglement

A point of weakness of the evaluation framework for real networks is that we have to build a positive set by leave-one-out node dismantling of hyperedges and modifying the graph connectivity. In addition, the negative set is based on a node-hyperedge distance that is approximated using topological shortest paths. In this section we want to integrate the previous results by introducing a second (and radically different) evaluation framework which exploits a random geometric graph generative model for realistic artificial complex networks. In this context, ‘realistic’ means that we can control several topological features that are typical of real networks, therefore we can investigate how HEP changes in relation to fundamental modifications of the network topology. Indeed, having the latent geometry that is behind the observable topology allows to build the positive and negative set based on node-hyperedge geometrical distances, leaving unperturbed (by any dismantling) the graph and hypergraph structure on which we apply the HEP.

In particular, in the computational evaluations of this section we adopt the nonuniform popularity-similarity-optimization (nPSO) model[7,8], a generative model recently introduced in order to grow random geometric graphs in the hyperbolic space, reproducing networks that have realistic features such as high clustering, small-worldness, scale-freeness and rich-clubness[12,33], with the additional possibility to control the community organization (for instance by setting a predefined number of communities, their internal distribution and their mixing). In this framework, the hyperedges are associations between the nodes that are members of the same community. The nodes of the network have coordinates in the hyperbolic disk, and the lower the hyperbolic distance between two nodes the higher the likelihood to be connected.

Based on this knowledge we build a positive and negative set of node-hyperedge pairs. After computing all the pairwise hyperbolic distances between the nodes, for each node-hyperedge pair that satisfies the ENC we assign a node-hyperedge distance equal to the minimum hyperbolic distance between the node and the members of the hyperedge. The positive set is composed by the 5% of pairs with the lowest node-hyperedge distance, while the negative set by the 5% of pairs with the highest distance. This is a very conservative definition of a positive and negative set, because only the pairs that are significantly different from the central part of the distribution are tested, making sure that the evaluation is reliable.

For each node-hyperedge pair $\left(c_{ij}, h_j\right)$ in the positive and negative set, we run the algorithm for node2hyperedge entanglement p-value p_{ij}. By considering a significance threshold of 0.05 we obtain a prediction on whether the node c_{ij} becomes a member of the entangled hyperedge eh_j. After obtaining predictions for all the pairs in the positive and negative set, we evaluate the performance using the Matthews correlation coefficient (MCC). Note that if the option of Benjamini-Hochberg correction is considered, then for each pair $\left(c_{ij}, h_j\right)$ the p-value should also be computed between the node c_{ij} and every other hyperedge for which the ENC is satisfied.

In order to test the overall performance of the methods, we generated synthetic networks with the nPSO model[7,8] ranging over several parameter combinations. In particular, we fixed the size of the networks to N = 1000 nodes, we tuned the exponent of the power-law degree distribution to γ = [2, 3], half of the average node degree to m = [4, 8], the temperature (inversely related to the clustering) to T = [0.1, 0.3, 0.5, 0.7] and the number of communities (in this case hyperedges) to C = [10, 20]. For each parameter combination, we generated 10 network realizations. We selected these specific values for the parameters because they are close to the ones of real networks and therefore this allows a type of evaluation that is approximating as much as possible the real scenario.

Fig. 4 reports the MCC evaluation in four panels, each analyzing the variation in performance by changing a certain nPSO parameter. In panel 4a, the barplot shows for each link predictor the average MCC over all the networks with $\gamma = 2$ (white bars) compared with the average MCC over all the networks with $\gamma = 3$ (colored bars). The same applies in panel 4b for $m = 4$ versus $m = 8$; in panel 4c for $T = 0.1$ versus $T = 0.7$; and in panel 4d for $C = 10$ versus $C = 20$. Since we generated networks with four different values of T, in panel 4c we compared the two extreme ones. Note that for each link predictor the figure reports only the best variant, meaning the one with average operator and p-values correction option that provides the highest average performance across all the networks. Please refer to Suppl. Table 7 for the complete overview of all the algorithmic variants. For visual clarity, the standard error of the MCC is not shown in Fig. 4, but it is reported in Suppl. Fig. 1, where the performance related to the two values of each nPSO parameter is shown in separated barplots.

The first message that is evident from the figure is that the L3-based link predictors (average MCC around 0.5) reach overall higher performance than the L2-based links predictors (average MCC around 0.15). In addition, the L2-based methods that explicitly minimize the external links in the CH model (CH2-L2, CH3-L2 and RA-L2), in this evaluation do not show a clear advantage with respect to the other L2-based predictors.

The four panels highlight that the performance improves by: (a) increasing the power-law exponent γ (at least for L3-based methods); (b) increasing half of average node degree m; (c) decreasing the temperature T (i.e. increasing the clustering); (d) increasing the number of communities C. The change in performance seems overall minor for the variation of m, T and C, highlighting that the predictors are quite robust across different topological organizations. On the other hand, the performance has a considerable boost going from $\gamma = 2$ to $\gamma = 3$ for L3 link predictors. The reason might be that with a decrease in power-lawness (from $\gamma = 2$ to $\gamma = 3$) - and therefore a decrease of hubs geometrical centrality in the hyperbolic disk - the geometrical distances between the hyperedges increase, and this helps to boost the performance in predicting to which entangled hyperedges belongs a candidate node.

2.4 Hyperedge entanglement in network medicine: a case study on COPD

The two different computational evaluations provided in the previous sections were crucial to assess the behaviour of the proposed HEP on data with gold standard and ground-truth hyperedge metadata. In this section we aim to evaluate the impact of hyperedge entanglement theory and the associated HEP on a real case scenario in complex systems biomedicine, where the metadata are incomplete to the point that nodes in the same hyperedge are not topologically connected between them. This is a challenging problem. Most algorithms currently available in network science for community detection or network generative modelling assume that a social group or a disease module should display high dyadic modularity for which nodes inside the hyperedge should have higher connectivity between them than with the rest of the network. For HEP instead, as we will investigate in this section, this might not be a problem and does not represent an obstacle to provide meaningful predictions.

Scientific evidence suggests that proteins of genes associated with complex diseases tend to connect in protein-protein interaction (PPI) networks, participating in the same biological pathways[21,34], and this is also reflected in the proteome[35] and the diseasome[21]. We remind that the diseasome consists of a bipartite network where diseases on one layer connects to associated genes on the other layer[21]. Bipartite networks allow interactions only across the two layers, this means that a disease imposes a hyperedge on its associated genes, which (by definition) do not interact between each other in the gene layer. Hence, we can translate and integrate this information in a high-order multilayer network model: one layer contains the diseasome-derived hypergraph (Fig. 5a, where each hyperedge links together a cohort of genes associated to a disease), another layer contains the PPI network (Fig. 5b).

Previous analysis of the diseasome showed that genes that contribute to common disorders: (i) show an increased tendency for their proteins to interact with each other via protein-protein interactions[36]; (ii) tend to be co-expressed in general or in specific tissues[36]; (iii) tend to share Gene Ontology (GO) terms[36,37]. These biological evidences are at support of a diseasome-based high-order multilayer network model such as the one we propose here. Indeed, pathways in PPI networks are often composed of group of interconnected proteins responsible for specific biological functions[21], and a disease represents a variation-induced perturbation of a specific PPI disease hyperedge that might contain different biological pathways, producing pathophysiological abnormalities.

Given any disease hyperedge and any candidate protein (any protein node that satisfies the ENC, see Results section 2.1.2) in a PPI network, HEP assesses the extent to which that candidate protein (and its gene) is significantly entangled to that disease. Hence, the entangled hyperedge (of a disease hyperedge) can provide insights on proteins whose genes might be significantly associated to a certain disease and its biological pathways. This has relevance because, despite major efforts in high-throughput mapping, the missing human PPIs exceed the experimentally documented interactions[38,39]. Our ability to predict previously undetected associations between parts of the diseasome using network science tools offers the possibility to gain insights about the mechanism underlying a complex disease[40].

To illustrate the benefits of such a modelling approach, we apply HEP algorithm to predict the genes that are members of the entangled hyperedge of COPD (Fig. 5a). COPD is a chronic inflammatory lung disease generally associated with cigarette smoking (with a smaller number of cases due to factors such as air pollution and mere genetics) and leading to obstructed airflow in bronchi. Although genetic studies have identified several risk loci for COPD[41–43], the mechanisms and molecular interactions that rule its pathophysiology need substantial investigation. We based our analysis on a PPI network which is a reference in systems biology and it is compiled from 15 different sources (16,418 nodes and 235,566 edges)[36]. We considered a total of 300 disease hyperedges: the COPD-associated disease hyperedge counts 30 COPD GWAS genes[44,45] (29 of which are present in the PPI network, therefore we do not consider the missing one that is ADGRG6). Note that the COPD disease module genes are not connected between them in the PPI network topology (there is only one link between the 29 genes), which is an important peculiarity not present in the previous evaluations. The remaining 299 disease[36] hyperedges are defined by the Medical Subject Headings (MeSH) ontology that have at least 20 associated genes in the current Online Mendelian Inheritance in Man (OMIM) and genome-wide association study (GWAS) databases[46,47]. Given the PPI network and the 300 disease hyperedges, we apply the HEP algorithm, which provides a p-value for each association between candidate genes (genes that in the PPI network are candidate nodes for entanglement to a given hyperedge) and a disease hyperedge.

At the end of this procedure, given the p-values for a certain gene and all the disease hyperedges of which it is candidate, a Benjamini-Hochberg adjustment is performed in order to correct for multiple hypothesis testing. Hence, in this context, the other 299 disease hyperedges are used only for multiple hypothesis testing adjustment and to obtain a robust estimation of the gene-hyperedge that is significantly entangled to COPD disease hyperedge, which is the focus of this case study.

In Fig. 5b, we display the COPD disease hyperedge (29 genes on the right side) and the respective entangled hyperedge (32 genes on the left side) predicted by at least one of the HEP algorithmic variants. Their pathway enrichment analysis using DAVID[48] (version 6.8 with the following pathways options: BBID, BIOCARTA, KEGG, Reactome, EC number) detects two significantly enriched pathways: platelet degranulation (7/61 genes included) and molecules associated with elastic fibres (5/61 genes included). This pathway analysis was executed considering as background all the PPI network genes, Benjamini adjusted p-values and a significance threshold of 0.05. Fig. 5b also reports part of the PPI network that connects COPD disease hyperedge and its entangled hyperedge by means of the two significant pathways. This is obtained with the following procedure. At first, we identified all the intermediate genes belonging to paths of length two (L2) or length three (L3) that connect COPD disease hyperedge and its entangled hyperedge. Then, we selected the subset of those intermediate genes that are included in at least one of the two enriched pathways. The colours of nodes and links highlight their association to the enriched pathways. From the biological point of view, the fact that the COPD entangled hyperedge is L2/L3 connected to the original COPD hyperedge via PPIs whose nodes are in those two pathways represent an essential confirmation of the biological relevance of the entanglement. Indeed, previous studies reported increased platelet activation in patients with stable and acute exacerbation of COPD[49], association of thrombocytosis with COPD morbidity[50], changes in elastic fibres in the small airways and alveoli in COPD[51].

In Fig. 5b, MFAP5 is the only gene exclusively involved with elastic fibres pathway and for which a significant entanglement to COPD has been predicted by HEP algorithm. An earlier study suggested that MFAP5 might be implicated in the extracellular matrix pathway with COPD proteins[45]. Moreover, MFAP5 has been previously shown to be differentially expressed in severe emphysema and bronchitis in lung tissue[52], two hallmarks of COPD. Hence, we moved forward to perform wet lab experiments that could support and clarify the reason of this significantly predicted entanglement of MFAP5 with the COPD hyperedge.

Fig. 6a reports for each link predictor variant of the HEP algorithm (the mean has been used as average operator in this specific application), the adjusted p-values of possible entanglement between MFAP5 and COPD disease hyperedge. The entanglement is significant ($p\text{-value} < 0.05$) according to three L3-based methods (CH2-L3, CH3-L3 and RA-L3). These L3-based predictors assigned a positive (nonzero) likelihood score to 10 of the 29 potential interactions between MFAP5 and the 29 proteins belonging to the COPD hyperedge. Out of these 10 potential candidates to interact with MFAP5, we had available resources to test 2 for co-immunoprecipitation (TGFB2, ELN) and 7 in a gene silencing experiment (TGFB2, MMP12, EGLN2, FBLN5, SFTPD, CHRNA5, TUFM).

Fig. 6b shows that ELN and TGFB2 were separately detected in the same IP western blot that MFAP5 was detected, indicating that they co-immunoprecipitate with MFAP5. This is an experimental evidence at support of the physical binding between MFAP5-ELN and between MFAP5-TGFB2, which is an important experimental result at validation and support of the computational prediction obtained by means of the HEP algorithm.

Finally, we performed siRNA knockdown assays in bronchial epithelial cells (16HBE cells), showing that MFAP5 suppresses the expression of TGFB2, EGLN2, FBLN5, SFTPD, CHRNA5 and TUFM, and enhances the expression of MMP12 at mRNA level (Fig. 6c), supporting the validity of the entangled suggested by the HEP algorithmic variants. Fig. 6d provides a summary of the new information gained from the biological experiments discussed in this section, in comparison to knowledge already in literature. It is evident the disruptive impact of the HEP algorithm in the process to gain new knowledge.

Since previous studies reported a level of comorbidity between COPD and aneurysm[53,54], we performed an additional analysis in order to spot a possible relation between the genes associated to aneurysm and the COPD entanglement subnetwork genes (Fig. 5b) that were enriched in the two significantly detected pathways (platelet degranulation and molecules associated with elastic fibres). In particular, for each of the two pathways, we performed a hypergeometric test to verify whether the genes associated to aneurysm are significantly enriched in the respective COPD-pathway subnetwork (see Methods for details). The p-values are strongly significant for both pathways (platelet degranulation p=8.35e-07, molecules associated with elastic fibres p=1.67e-07), consolidating and enlarging knowledge on the molecular mechanisms for comorbidity of the two diseases.

3. Discussion

The first important achievement of this study is that we have introduced a new concept and a novel challenge in network science, which take respectively the names of entangled hyperedge and of hyperedge entanglement prediction (HEP). The formal definition of this prediction problem is fundamental to investigate how elementary modelling at the minimal scale of dyadic node-node link formation can impact modelling of hyperedge formation at larger scale (higher order) in the network organization. In addition, the same prediction problem can be employed to address practical questions such as: how to forecast the possible relationship between a member of a social network and a consolidated social group; or how to spot in biomedicine an unknown association between a certain gene and a disease module.

The second important attainment is that we have proposed the first algorithm for hyperedge entanglement prediction which consists of many variants that allow to investigate the multifaceted local rules of self-organization and link formation of a complex network, and their L2 or L3 organization.

The third important accomplishment is that we have introduced two different evaluation frameworks for the prediction performance of HEP algorithms. The first evaluation scheme is on real networks, where we build a positive set by dismantling existing hyperedges and a negative set by exploiting topological information. The second evaluation scheme is on synthetic networks generated with the nPSO model[7,8], where the positive and negative sets are constructed based on node-hyperedge geometrical distances.

In brief, the first important result that we achieve is that HEP is feasible and performs much better than random prediction. This is not trivial, because it represents a necessary and fundamental evidence at support of the fact that HEP is a well-posed problem in network science, for which algorithms can provide meaningful computational solutions.

The second important result is that local link prediction variants - adopted within the pipeline of the HEP algorithm - are enough (in the sense that local topological information is enough) to allow efficient (fast and meaningful) prediction. Furthermore, in line with some recent studies[28–30] the results from our analysis confirm that the Cannistraci-Hebb (CH) models are robust link predictors across different network topologies, and that the adoption of models based on paths of length three (L3) seems more reliable than paths of length two (L2) on most network structural organizations.

The third landmark result of this study is that we successfully exploited the proposed HEP algorithm in systems biomedicine in order to predict and enhance the understanding of the genes involved in the pathogenesis and comorbidity of COPD. To run HEP we simply need very basic knowledge, only two information to provide as inputs in the algorithm: the module of known genes (disease module) associated with the considered complex disease (in this case COPD) and the updated human interactome topology. Some of the associations suggested by the HEP algorithm for COPD have been experimentally tested in this study by co-immunoprecipitation and gene silencing, confirming the validity of the prediction. In addition, we detected two significantly enriched pathways involved in the COPD entanglement (platelet degranulation and molecules associated with elastic fibres) and their related subnetworks are in turn significantly enriched with genes associated to aneurysm. This corroborates previous studies[54,55], reporting COPD as the strongest independently associated disease with aortic aneurysm, with a prevalence of COPD up to 44% among aneurysm patients. Moreover, it has been demonstrated that fibrillinopathies such as aortic aneurysm arises from mutations in genes that encode important molecules associated to elastic fibres. Elastic fibres are insoluble components of the extracellular matrix (ECM) of dynamic connective tissues such as skin, arteries, lungs and ligaments. Mutations in several ECM genes, such as FBN1, MFAP5 and

TGFB2 (shown in Fig. 5b), predispose to aortic aneurysmal disease by affecting aortic stiffness and elasticity[56].

A final important remark is that in our HEP framework, a social group or disease module do not need to be characterized by a high dyadic modularity for which nodes inside the hyperedge should have higher connectivity between them than with the rest of the network. In practice, we do not constrain the hyperedge to be a block module or a segregated community in the network. This is particularly important in predictions in presence of incomplete metadata such as the ones of the COPD disease module, whose genes have only one connection between them in the PPI network topology.

To conclude, there is also a last and remarkable result of HEP that is coming from its application in cardiovascular molecular biomedicine. HEP was recently used with success to identify SDC4, a heparan sulfate proteoglycan, as a potential target of PCSK9 that mediates pro-inflammatory responses in macrophages. This result, now available in a study on a murine animal model of Katsuki et al. [submitted 2020], is relevant because circulating PCSK9 may induce macrophage activation and contribute to vein graft lesion development via mechanisms independent of LDLR degradation and blood cholesterol levels. PCSK9 suppression may thus prevent vein graft failure, which is a major clinical problem with no effective medical therapies up today.

4. Methods

4.1 Link prediction methods

Resource Allocation (RA)

The formula of the RA-L2 model is[25]:

$$RA\_L2(u,v) = \sum_{z \in L2} \frac{1}{d_z}$$

where: *u* and *v* are the two seed nodes of the candidate interaction; *z* is the intermediate node on the considered path of length two; d_z is the respective node degree; and the summation is executed over all the paths of length two.

The formula of the RA-L3 model is[29]:

$$RA\_L3(u,v) = \sum_{z_1, z_2 \in L3} \frac{1}{\sqrt{d_{z_1} * d_{z_2}}}$$

where: *u* and *v* are the two seed nodes of the candidate interaction; z_1, z_2 are the intermediate nodes on the considered path of length three; d_{z_1}, d_{z_2} are the respective node degrees; and the summation is executed over all the paths of length three.

Cannistraci-Hebb (CH)

The formulas of the CH1-L2, CH2-L2 and CH3-L2 models are[28]:

$$CH1\_L2(u,v) = \sum_{z \in L2} \frac{di_z}{d_z}$$

$$CH2\_L2(u,v) = \sum_{z \in L2} \left(\frac{di_z^*}{de_z^*} = \frac{1 + di_z}{1 + de_z}\right)$$

$$CH3\_L2(u,v) = \sum_{z \in L2} \left(\frac{1}{de_z^*} = \frac{1}{1 + de_z}\right)$$

where: *u* and *v* are the two seed nodes of the candidate interaction; *z* is the intermediate node on the considered path of length two; d_z is the respective node degree; di_z is the respective internal node degree; de_z is the respective external node degree; and the summation is executed over all the paths of length two. The asterisk on a degree variable indicates that a unitary term is added, in order to avoid the saturation of the value.

The formulas of the CH1-L3, CH2-L3 and CH3-L3 models are[28]:

$$CH1\_L3(u,v) = \sum_{z_1, z_2 \in L3} \frac{\sqrt{di_{z_1} * di_{z_2}}}{\sqrt{d_{z_1} * d_{z_2}}}$$

$$CH2\_L3(u,v) = \sum_{z_1,z_2 \in L3} \frac{\sqrt{di^*_{z_1} * di^*_{z_2}}}{\sqrt{de^*_{z_1} * de^*_{z_2}}}$$

$$CH3\_L3(u,v) = \sum_{z_1,z_2 \in L3} \frac{1}{\sqrt{de^*_{z_1} * de^*_{z_2}}}$$

where: *u* and *v* are the two seed nodes of the candidate interaction; z_1, z_2 are the intermediate nodes on the considered path of length three; d_{z_1}, d_{z_2} are the respective node degrees; di_{z_1}, di_{z_2} are the respective internal node degreed; de_{z_1}, de_{z_2} are the respective external node degrees; and the summation is executed over all the paths of length three. The asterisk on a degree variable indicates that a unitary term is added, in order to avoid the saturation of the value. Note that the CH3 model, based only on the minimization of the external node degrees, has been introduced in this study.

Cannistraci-Alanis-Ravasi (CAR)

The formula of the CAR model is[4]:

$$CAR(u,v) = CN(u,v) \times LCL(u,v)$$

where: *u* and *v* are the two seed nodes of the candidate interaction; $CN(u,v)$ is the number of common neighbours between them and $LCL(u,v)$ is the number of local community links (links between the common neighbours).

Cannistraci-Adamic-Adar (CAA)

The formula of the CAA model is[4]:

$$CAA(u,v) = \sum_{z \in L2} \frac{di_z}{\log_2 d_z}$$

where: *u* and *v* are the two seed nodes of the candidate interaction; z is the intermediate node on the considered path of length two; d_z is the respective node degree; di_z is the respective internal node degree; and the summation is executed over all the paths of length two.

Cannistraci-Jaccard (CJC)

The formula of the CAA model is[4]:

$$CJC(u,v) = \frac{CAR(u,v)}{|N(u) \cup N(v)|}$$

where: u and v are the two seed nodes of the candidate interaction; $CAR(u,v)$ is the value of the Cannistraci-Alanis-Ravasi model, $|N(u) \cup N(v)|$ is the cardinality of the set composed of the union of the neighbours of u and v.

4.2 Real networks dataset

The methods have been tested on three small real networks and on two large biological networks. For such networks metadata are available, representing the membership of the nodes to a certain group. Nodes belonging to the same group share a common feature. We consider such association between multiple nodes sharing a common feature as a hyperedge. The networks have been transformed into undirected, unweighted, without self-loops and only the largest connected component has been considered.

The Football network presents games between division IA colleges during regular season fall 2000. Hyperedges associate teams belonging to the same conference.

The Opsahl_10 network comes from the research team of a manufacturing company and nodes represent employees. The hyperedges group the employees by the company locations (Paris, Frankfurt, Warsaw, and Geneva). The researchers were asked to indicate the extent to which their co-workers provide them with information they use to accomplish their work. The answers were on the following scale: 0—I do not know this person/I never met this person; 1—Very infrequently; 2—Infrequently; 3—Somewhat frequently; 4— Frequently; 5—Very frequently. We set an undirected link when there was at least a weight of 4.

The Polbooks network represents frequent co-purchases of books concerning US politics on amazon.com. Hyperedges associate the books with the same political orientation such as conservative, neutral or liberal. The network is unpublished but can be downloaded at http://www-personal.umich.edu/~mejn/netdata/.

The protein-protein interaction (PPI) network of S. cerevisiae is one of the most studied PPI networks that has been used. It comes from a published dataset of PPI networks compiled into three genome-scale networks: yeast two-hybrid (Y2H), affinity purification followed by mass spectrometry (AP/MS), and literature curated (LC)[57]. This PPI network was preprocessed by Ahn et al.[58] by using the union of these three networks and taking only the largest component of each network. The hyperedges associate proteins with the same significantly enriched GO terms.

The metabolic network reconstruction of E. coli K-12 MG1655 strain (iAF1260) is one of the most elaborate metabolic network reconstructions currently available[59]. From the metabolic network reconstruction iAF1260, Ahn et al.[58] retained only cellular reactions, ignored

doi:10.20944/preprints202012.0500.v1

information regarding the compartments (cytoplasm and periplasm), and projected the network into metabolite space (two metabolites are connected if they share a reaction). For instance, if an enzyme catalyzes the reaction where metabolites A and B are transformed into C and D, the resulting network would contain a clique of A, B, C, and D. The hyperedges group metabolites with the same metabolic pathway annotation from the KEGG database.

Some topological properties of the networks used in this study are summarized in Suppl. Table 1.

4.3 Algorithm for the evaluation procedure of HEP in real networks

In general, after removal from the hyperedge h_j, the test node becomes a test candidate node c_{ij} for entanglement with h_j. Then, we compute by HEP whether this test candidate node is predicted as a member of the eh_j entangled hyperedge. At the same time, we want to test that the test candidate node c_{ij} is not predicted by HEP as member of the entangled hyperedge eh_k of another hyperedge h_k which by definition is considered a wrong assignment because topologically very far from the test candidate node c_{ij}.

In details, the algorithm for the evaluation procedure of HEP in real networks is the following. For each node-hyperedge pair $(v_{ij} \in h_j, h_j)$ in hypergraph:

a. If the node v_{ij} has at least one link to the other members of the hyperedge h_j and at least one link to nodes that are not members, we remove the node v_{ij} from h_j, as well as the network links between v_{ij} and the other members of h_j. After the removal v_{ij} satisfies the ENC and therefore it becomes a candidate node for entanglement c_{ij}. This generates a pair (c_{ij}, h_j) to test for entanglement significance by the ESC.

b. We run the algorithm for node2hyperedge entanglement p-value between the node c_{ij} and the hyperedge h_j, which represents a test on the positive set (here a correct prediction is if the p-value will be significant). Meanwhile for c_{ij} we define the 'negative' hyperedge h_k and we run the algorithm for node2hyperedge entanglement p-value between the node c_{ij} and the hyperedge h_k, which represents a test on the negative set (here a correct prediction is if the p-value will be not significant). We obtain the p-values p_{ij} and p_{ik}.

c. By considering a significance threshold of 0.05 for the p-values p_{ij} and p_{ik}, we obtain a prediction on whether (c_{ij}, h_j) and (c_{ij}, h_k) also satisfy the ESC, indicating if the node c_{ij} becomes a member of the entangled hyperedges eh_j and eh_k respectively.

Once obtained predictions for all the pairs in the positive and negative set, we evaluate the performance using the Matthews correlation coefficient (MCC). Note that if the option of Benjamini-Hochberg correction is considered, then in step (b) the p-value should be computed between the node c_{ij} and every hyperedge for which ENC is satisfied.

4.4 Generation of synthetic networks using the nPSO model

The Popularity-Similarity-Optimization (PSO) model[7] is a generative network model recently introduced in order to describe how random geometric graphs grow in the hyperbolic space. In this model the networks evolve optimizing a trade-off between node popularity, abstracted by the radial coordinate, and similarity, represented by the angular distance. The PSO model can reproduce many structural properties of real networks: clustering, small-worldness (concurrent low characteristic path length and high clustering), node degree heterogeneity with power-law degree distribution and rich-clubness. However, being the nodes uniformly distributed over the angular coordinate, the model lacks a non-trivial community structure.

The nonuniform PSO (nPSO) model[7,8] is a variation of the PSO model that exploits a nonuniform distribution of nodes over the angular coordinate in order to generate networks characterized by communities, with the possibility to tune their number, size and mixing property. We adopted a Gaussian mixture distribution of angular coordinates, with communities that emerge in correspondence of the different Gaussians, and the parameter setting suggested in the original study[7,8]. Given the number of components C, they have means equidistantly arranged over the angular space, $\mu_i = \frac{2\pi}{C} \cdot (i-1)$, the same standard deviation fixed to 1/6 of the distance between two adjacent means, $\sigma_i = \frac{1}{6} \cdot \frac{2\pi}{C}$, and equal mixing proportions, $\rho_i = \frac{1}{C}$ $(i = 1 \dots C)$. The community memberships are assigned considering for each node the component whose mean is the closest in the angular space. We consider such association between multiple nodes in the same community as a hyperedge. The other parameters of the model are the number of nodes N, half of the average node degree m, the network temperature T (inversely related to the clustering) and the exponent γ of the power-law degree distribution. Given the parameters (N, m, T, γ, C), for details on the generative procedure please refer to the original study[7,8].

4.5 Cell culture maintenance

The cell lines used to validate the link predictions with biological experiments were bronchial epithelial cells BEAS-2B (Cat # 95102433, Millipore) and 16-HBE/16HBE (human bronchial epithelium, Cat# SCC150, Millipore). These cells were propagated *in vitro* at 37^{O}C in a humidified incubator with 5% CO_2. The culture media used for *in vitro* experiments are the following: (1) For the BEAS-2B cells, growth media (GM) is Dulbecco's modified eagle medium, DMEM (Cat# 10569-010, Gibco, Thermo Fisher) with 1% penicillin/streptomycin (P/S) and 10% of fetal bovine serum (FBS). For all washing steps, if not indicated otherwise, autoclaved Ca^{++}- and Mg^{++}-free phosphate buffer saline (PBS) was used.

4.6 Co-Immunoprecipitation (Co-IP)

Cells in 150 cm culture flasks were harvested and lysed with 600µl lysis buffer (20 mM HEPES (pH7.9), 0.1 M KCl, 0.2 mM EDTA, 10% Glycerol, 0.1% NP-40 and cOmplete™, EDTA-free Protease Inhibitor Cocktail (Sigma Aldrich, Cat# S8830) and PhosSTOP Phosphatase Inhibitor Cocktail Tablets (Sigma Aldrich, Cat# 4906845001). Cells were incubated in the lysis buffer for ten minutes at 4^{O}C on a rotating shaker. The resulting lysates were then centrifuged at 10,000 rpm for 10 minutes at 4°C. We then collected the supernatant (soluble) fraction for immunoprecipitation. The protein concentration of the lysate was measured using the bicinchoninic acid (BCA) method (Thermo Scientific, PI-Cat#23225). After obtaining protein concentration, 1.0 mg of protein of the 16HBE lysate was incubated overnight at 4°C, with 5.0 µg of the following antibody in PBS: mouse monoclonal anti-human MFAP5 (Sigma-Aldrich, Cat# SAB1406642) which detects the full length of the MFAP5 protein and 25µl per sample of Dynabeads Protein G (Thermo Scientific, Cat# 10004D). Dynabeads were washed once with lysis buffer before addition to the lysate-antibody mix. Washing Dynabeads involved placing the tube containing 25 uL of Dynabeads on the vendor-provided magnet (DynaMag, Thermo Fisher) until the solution clears. The clear solution is discarded, and the tube is removed from the magnet. Then, the Dynabeads are redissolved with the lysate-antibody mix. Each sample replicate was left to incubate for 1 hour at 4°C on a rotating shaker. Subsequently, the lysate was washed five times, with 1.0 ml of lysis buffer using the magnet as above. At this step, any unbound protein from the lysate to the Dynabeads are washed away by the changing washes of lysate buffer. The target protein, along with other proteins bound to that target protein, will be bound by the target antibody (MFAP5), which in turn is bound by the beads. The pellet is then redissolved in 250 □L RIPA buffer (Cat# BP-407, Boston BioProducts). Protein yield from the co-IP supernatant was determined using a protein assay kit (Thermo Fisher Scientific, Cat# 23225). The samples (replicate experiments, n=3) were dissolved in 50.0 □L of Laemmli sample buffer (Boston BioProducts, Cat# BP-111R) and then boiled for 5 minutes at 95°C. At this point, the antibody and bound proteins are eluted or separated from the Dynabeads and the target antibody. The remaining Dynabeads are removed by magnetic isolation as above. The resulting supernatant is then processed for Western blotting.

Standard Western blotting procedure was done. The denatured samples were run in lanes and separated by SDS polyacrylamide electrophoresis. The protein lanes were (1) "input," which refers to protein lysate before IP, and (2) "IP" which is the immunoprecipitated proteins only (Fig. 6b). In some gels/blots (not shown), we added a lane that contained pure recombinant human MFAP5 protein as a reference band to confirm the MFAP5 band in the input and IP lanes. SDS-PAGE used 8.0% acrylamide (Boston BioProducts, Cat# BAC-30PA), separating buffer (Boston BioProducts, Cat# BP-90), stacking buffer (Boston BioProducts, Cat# BP-95), N,N,N',N'-tetramethylethylenediamine (TEMED) (Sigma-Aldrich, Cat# 1610801), and ammonium persulfate (Sigma-Aldrich, Cat# A3678-25G)]. They were subsequently transferred to nitrocellulose membranes (Bio-Rad Laboratories, Hercules, CA, Cat# 1620112). The membranes were blocked with 2.5% non-fat dry milk (Santa Cruz Biotechnology, TX, Cat# sc-2325) in 1X tris-buffered saline with 0.1% Tween 20 (TBST) (Boston BioProducts, MA, Cat# IBB-181). The primary antibody used was a rabbit polyclonal anti-human MFAP5 antibody (1:1000, Cat# abx103845, Abbexa). The secondary detection antibody used was anti-rabbit IgG with peroxidase conjugate (1:5000, Sigma-Aldrich, Cat# A0545-1ML). The proteins of interest were visualized using an ECL blotting substrate (Bio-Rad Laboratories, Cat# 1705060) and an imager (GE Healthcare, ImageQuant LAS 4000).

After detecting MFAP5 in the IP blot, we "stripped" the blot using stripping buffer, Thermo Scientific, Restore Western Blot Stripping Buffer (Cat# 21059, ThermoFisher) to remove the detected bands by MFAP5. Western blot was repeated for other co-immunoprecipitated proteins (TGFB2 and ELN) after the blot-stripping process. We used mouse monoclonal antibody against human TGFB2 (Cat# ab36495, Abcam) or rabbit monoclonal ELN (Cat# ab213720 Abcam) in the same manner as above. The secondary detection antibody used this time was anti-mouse IgG peroxidase conjugate (1:5000, Sigma-Aldrich, Cat# A4416-1ML) for TGFB2 and anti-rabbit IgG peroxidase conjugate (1:5000, Sigma-Aldrich, Cat# A0545-1ML) for ELN. The proteins of interest were visualized using an ECL blotting substrate (Bio-Rad Laboratories, Cat# 1705060) and an imager (GE Healthcare, ImageQuant LAS 4000). Detection of TGFB2 or ELN bands from the same blot where MFAP5 was detected, demonstrates that these proteins were co-immunoprecipitated with MFAP5.

4.7 RNA purification and cDNA synthesis

RNA from cultured 16HBE cells (at passage # 5) was isolated using the Illustra$^{\text{TM}}$ RNAspin mini kit (GE Healthcare, Cat# 25-0500-72). We experimented on nine replicates. For the adherent cells, RNA Lyse solution from the Illustra RNAspin Mini Kit (GE Life Science) mixed with 1% 2-mercaptoethanol (Sigma) was added. Each lysate was then frozen at -30□C for later RNA purification. The purification was done following the manufacturer's protocol for GE Healthcare Illustra™ RNAspin Mini Isolation Kit (lot 1711/001). To get a concentrated RNA yield, the purified RNA was eluted in 16 □L RNase-free H2O. Each sample concentration was measured using a NanoDrop Microvolume Spectrophotometer 2009 (ThermoFisher). To normalize the amount of RNA between samples, the volume needed for each sample to have 500 ng of RNA was calculated. The qScript cDNA Synthesis Kit (Quantabio) was used to make Complementary DNA (cDNA). The volume of RNA was then added to a strip tube and diluted with nuclease-free water to a total of 15 □L. For each tube, 4.0 □L of reaction mix and 1.0 □L of reverse transcriptase (RT). After centrifugation, the prepared strip tube plate was processed in Biosystems 2720 Thermal Cycler for cDNA synthesis. Samples were stored at 4.0□C.

4.8 Pre-Amplification

For the pre-Amplification of cDNA samples, it was performed with the Perfecta PreAmp SuperMix (5X) kit (Quantabio). It was made an assay primers pool using 2.0 µl of each primer (see Suppl. Table 8). The reaction solution was raised to 200.0 µL total volume, additional TE buffer (10mM Tris-HCl (pH 8.0), 0.1ml EDTA) was added. 4.0 µL of PerfeCta PreAmpl SuperMix (5X), 2.5 µl of Taqman Assay Pool, 5.0 µl of cDNA and 8.5 µl of nuclease-free water were mixed to make 20 µl of the pre-amplification reaction mixture for TaqMan assays. Each sample was, then, incubated in a thermal cycle, following these steps: initial denaturation (95°C, 2 minutes), PreAmpl cycling (14 cycles of 95°C, 10 seconds; 60°C, 3 minutes) and hold (4°C). Finally, the concurrent samples were used for qPCR.

4.9 qPCR

For qPCR analysis, it was used a 384 well plate with 2□l of PreAmp cDNA and 8□l of primer cocktail per well. The primer cocktail was made with 5□l of Taqman Fast Universal PCR Master Mix (2X), 2.75□l of nuclease-free water (QuantaBio), and 0.25□l of respective primer (see Suppl. Table 8). PCR was done using 79020HT Fast Real-Time PCR Systems, and the data were analyzed using Prism GraphPad 8. Statistical significance was determined after testing between Control and MFAP5 silenced sample conditions by student's t-test, two-tailed, and unpaired.

4.10 Hypergeometric test

Consider to randomly draw n objects, without replacement, from a finite population of size N that contains exactly K objects with a desired feature. The hypergeometric distribution is a discrete probability distribution that describes the probability of randomly drawing k objects with that feature (successes). The probability mass function of a random variable X that follows the hypergeometric distribution is:

$$\Pr(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

The function is positive when $\max(0, n + K - N) \leq k \leq \min\,(n, K)$.

In the hypergeometric test for over-representation of the desired feature in a sample, the p-value is computed using the hypergeometric distribution as the probability of randomly drawing k or more successes. If in your sample you observe k^* objects with the desired feature, the p-value of the hypergeometric test for over-representation is given by:

$$p = \sum_{k=\max(k^*, n+K-N)}^{\min\,(n,K)} \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

In our simulations related to the Results section 2.4, for each of the two biological pathways considered (platelet degranulation, molecules associated with elastic fibres), we performed the hypergeometric test for over-representation of genes associated to aneurysm in the COPD-pathway subnetwork. Therefore, we have defined the parameters of the hypergeometric distribution as follows: N is equal to the number of nodes in the PPI network; K is equal to the number of nodes associated to aneurysm, meaning that they are either members of the aneurysm hyperedge or they are neighbors of its members in the PPI; n is the number of nodes in the COPD-pathway subnetwork, meaning that they are in the pathway considered and they are either in the COPD hyperedge, in its entangled hyperedge or in a L2/L3 path between them (see Fig. 5b); k is equal to the number of nodes in the COPD-pathway subnetwork and associated to aneurysm.

The actual values of the parameters and the resulting p-values are the following:

- platelet degranulation: N=16418, K=918, n=39, k=12, p=8.35e-07;
- molecules associated with elastic fibres: N=16418, K=918, n=23, k=10, p=1.67e-07.

Authors contributions

C.V.C. conceived the hyperedge entanglement framework and A.S. conceived the application in biomedicine and diseasome analysis. I.A. conceived the analysis of comorbidity between COPD and aneurysm. C.V.C invented the N2EH and HEP algorithms and their evaluation with the help of A.M and I.A. A.M and I.A wrote the computational code of the N2EH and HEP algorithms under the guidance of C.V.C. A.M. wrote the computational code for the evaluation. The biological validation and experiments were conceived by A.S. with the help of J.L.D and E.S. Fig. 1-5, their respective content and analysis was designed by C.V.C with the help of A.M and I.A. Fig. 6 was designed by A.S. and I.A with the help of J.L.D. The manuscript was drafted as follow: C.V.C. wrote abstract, introduction, result section 2.4 and discussion with minor revision of A.M.; A.M and C.V.C. wrote together the remaining result sections; A.M. wrote the method and supplementary section with the help of C.V.C. and I.A. All the manuscript was checked and amended by the other authors. A.S. and M.A provided the funding for the biological experiments and major part of the PhD salary of I.A. The overall project was supervised and directed by C.V.C and A.S.

Acknowledgments

We thank the BIOTEC System Administrators for their IT support, Gloria Marchesi and Sabine Zeissig for the administrative assistance, and the Centre for Information Services and High Performance Computing (ZIH) of the TUD. The work in C.V.C. research group was mainly supported by: the independent group leader starting grant of the Technische Universität Dresden (TUD); by the Klaus Tschira Stiftung (KTS) gGmbH, Germany (Grant number: 00.285.2016). I..A. was supported during the early stage of his doctorate at Channing Division of Network Medicine, Brigham and Women’s Hospital, Boston, USA, by A.S.’s grant title: Systems Biology of Airway Disease–Program Project-PS#115960. I.A. was also supported by: Lipotype GmbH; the scholarship awarded by the Friends of Dresden during the completion and wrap-up phase of his doctorate, based on the recommendation of the board of the Graduate Academy of TU Dresden.

References

1. Strogatz, S. H. Exploring complex networks. **410,** (2001).
2. Wasserman, S. & Faust, K. Social Network Analysis in the Social and Behavioral Sciences. in *Social Network Analysis* (2012). doi:10.1017/cbo9780511815478.002
3. Girvan, M. & Newman, M. E. J. Community structure in social and biological networks. *PNAS* **99,** 7821–7826 (2002).
4. Cannistraci, C. V., Alanis-Lobato, G. & Ravasi, T. From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks. *Sci. Rep.* **3,** 1–13 (2013).
5. Xu, M., Pan, Q., Muscoloni, A., Xia, H. & Cannistraci, C. V. Modular gateway-ness connectivity and structural core organization in maritime network science. *Nat. Commun.* (2020). doi:10.1038/s41467-020-16619-5
6. Lancichinetti, A., Fortunato, S. & Radicchi, F. Benchmark graphs for testing community detection algorithms. *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.* **78,** (2008).
7. Muscoloni, A. & Cannistraci, C. V. A nonuniform popularity-similarity optimization (nPSO) model to efficiently generate realistic complex networks with communities. *New J. Phys.* **20,** (2018).
8. Muscoloni, A. & Cannistraci, C. V. Leveraging the nonuniform PSO network model as a benchmark for performance evaluation in community detection and link prediction. *New J. Phys.* **20,** (2018).
9. García-Pérez, G., Serrano, M. Á. & Boguñá, M. Soft Communities in Similarity Space. *J. Stat. Phys.* (2018). doi:10.1007/s10955-018-2084-z
10. Bianconi, G. & Rahmede, C. Emergent Hyperbolic Network Geometry. *Sci. Rep.* **7,** 41974 (2017).
11. Fortunato, S. & Hric, D. Community detection in networks: A user guide. *Phys. Rep.* **659,** 1–44 (2016).
12. Muscoloni, A., Thomas, J. M., Ciucci, S., Bianconi, G. & Cannistraci, C. V. Machine learning meets complex networks via coalescent embedding in the hyperbolic space. *Nat. Commun.* **8,** (2017).
13. Jin, D. *et al.* Incorporating Network Embedding into Markov Random Field for Better Community Detection. *Proc. AAAI Conf. Artif. Intell.* **33,** 160–167 (2019).
14. Hu, B., Wang, H., Yu, X., Yuan, W. & He, T. Sparse network embedding for community detection and sign prediction in signed social networks. *J. Ambient Intell. Humaniz. Comput.* **10,** 175–186 (2019).
15. Sun, H. *et al.* Network Embedding for Community Detection in Attributed Networks. *ACM Trans. Knowl. Discov. Data* **14,** (2020).
16. Kumar, S., Panda, B. S. & Aggarwal, D. Community detection in complex networks using network embedding and gravitational search algorithm. *J. Intell. Inf. Syst.* (2020). doi:10.1007/s10844-020-00625-6
17. Peel, L., Larremore, D. B. & Clauset, A. The ground truth about metadata and community detection in networks. *Sci. Adv.* (2017). doi:10.1126/sciadv.1602548
18. Fortunato, S. Community detection in graphs. *Physics Reports* (2010). doi:10.1016/j.physrep.2009.11.002
19. Schaeffer, S. E. Graph clustering. *Comput. Sci. Rev.* (2007). doi:10.1016/j.cosrev.2007.05.001
20. Martínez, V., Berzal, F. & Cubero, J. C. A survey of link prediction in complex networks. *ACM Comput. Surv.* (2016). doi:10.1145/3012704
21. Goh, K. Il *et al.* The human disease network. *Proc. Natl. Acad. Sci. U. S. A.* (2007). doi:10.1073/pnas.0701361104
22. Einstein, A., Podolsky, B. & Rosen, N. Can quantum-mechanical description of physical reality be considered complete? *Phys. Rev.* (1935). doi:10.1103/PhysRev.47.777
23. Lü, L., Pan, L., Zhou, T., Zhang, Y.-C. & Stanley, H. E. Toward link predictability of complex networks. *Proc. Natl. Acad. Sci.* **112,** 2325–2330 (2015).
24. Newman, M. E. J. Clustering and preferential attachment in growing networks. **64,** 1–4 (2001).
25. Zhou, T. & Zhang, Y. Predicting Missing Links via Local Information.

26. Adamic, L. A. & Adar, E. Friends and neighbors on the Web. **25,** 211–230 (2003).
27. Daminelli, S., Thomas, J. M., Durán, C. & Cannistraci, C. V. Common neighbours and the local-community-paradigm for topological link prediction in bipartite networks. *New J. Phys.* **17,** 113037 (2015).
28. Muscoloni, A., Abdelhamid, I. & Cannistraci, C. V. Local-community network automata modelling based on length-three-paths for prediction of complex network structures in protein interactomes, food webs and more. *bioRxiv* (2018).
29. Kovács, I. A. *et al.* Network-based prediction of protein interactions. *Nat. Commun.* (2019). doi:10.1038/s41467-019-09177-y
30. Tao, Z., Yan-Li, L. & Guannan, W. Experimental analyses on 2-hop-based and 3-hop-based link prediction algorithms. *arXiv:1909.00174 [physics.soc-ph]* (2019).
31. Chicco, D. & Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **21,** 6 (2020).
32. Cannistraci, C. V. Modelling Self-Organization in Complex Networks Via a Brain-Inspired Network Automata Theory Improves Link Reliability in Protein Interactomes. *Sci. Rep.* **8,** 15760 (2018).
33. Muscoloni, A. & Cannistraci, C. V. Rich-clubness test: how to determine whether a complex network has or doesn’t have a rich-club? *arXiv:1704.03526v1 [physics.soc-ph]* (2017).
34. Barabási, A. L., Gulbahce, N. & Loscalzo, J. Network medicine: A network-based approach to human disease. *Nat. Rev. Genet.* **12,** 56–68 (2011).
35. Corti, V. *et al.* Protein fingerprints of cultured CA3-CA1 hippocampal neurons: comparative analysis of the distribution of synaptosomal and cytosolic proteins. *BMC Neurosci.* **9,** 36 (2008).
36. Menche, J. *et al.* Uncovering disease-disease relationships through the incomplete interactome. *Science* (2015). doi:10.1126/science.1257601
37. Ashburner, M. *et al.* Gene ontology: Tool for the unification of biology. *Nature Genetics* (2000). doi:10.1038/75556
38. Luck, K., Sheynkman, G. M., Zhang, I. & Vidal, M. Proteome-Scale Human Interactomics. *Trends in Biochemical Sciences* (2017). doi:10.1016/j.tibs.2017.02.006
39. Hart, G. T., Ramani, A. K. & Marcotte, E. M. How complete are current yeast and human protein-interaction networks? *Genome Biol.* **7,** 120 (2006).
40. Cannistraci, C. V., Alanis-Lobato, G. & Ravasi, T. Minimum curvilinearity to enhance topological prediction of protein interactions by network embedding. *Bioinformatics* **29,** 199–209 (2013).
41. Cho, M. H. *et al.* Risk loci for chronic obstructive pulmonary disease: A genome-wide association study and meta-analysis. *Lancet Respir. Med.* (2014). doi:10.1016/S2213-2600(14)70002-5
42. Hobbs, B. D. *et al.* Genetic loci associated with chronic obstructive pulmonary disease overlap with loci for lung function and pulmonary fibrosis. *Nat. Genet.* (2017). doi:10.1038/ng.3752
43. Cho, M. H. *et al.* A genome-wide association study of emphysema and airway quantitative imaging phenotypes. *Am. J. Respir. Crit. Care Med.* (2015). doi:10.1164/rccm.201501-0148OC
44. Halu, A. *et al.* Exploring the cross-phenotype network region of disease modules reveals concordant and discordant pathways between chronic obstructive pulmonary disease and idiopathic pulmonary fibrosis. *Hum. Mol. Genet.* **28,** 2352–2364 (2019).
45. Sharma, A. *et al.* Integration of Molecular Interactome and Targeted Interaction Analysis to Identify a COPD Disease Network Module. *Sci. Rep.* **8,** (2018).
46. Mottaz, A., Yip, Y. L., Ruch, P. & Veuthey, A. L. Mapping proteins to disease terminologies: From UniProt to MeSH. in *BMC Bioinformatics* (2008). doi:10.1186/1471-2105-9-S5-S3
47. Ramos, E. M. *et al.* Phenotype-genotype integrator (PheGenI): Synthesizing genome-wide association study (GWAS) data with existing genomic resources. *Eur. J. Hum. Genet.* (2014). doi:10.1038/ejhg.2013.96
48. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* (2009). doi:10.1038/nprot.2008.211
49. Maclay, J. D. *et al.* Increased platelet activation in patients with stable and acute exacerbation of COPD. *Thorax* (2011). doi:10.1136/thx.2010.157529

50. Fawzy, A. *et al.* Association of thrombocytosis with COPD morbidity: The SPIROMICS and COPDGene cohorts. *Respir. Res.* (2018). doi:10.1186/s12931-018-0717-z
51. Black, P. N. *et al.* Changes in elastic fibres in the small airways and alveoli in COPD. *Eur. Respir. J.* (2008). doi:10.1183/09031936.00017207
52. Faner, R. *et al.* Network analysis of lung transcriptomics reveals a distinct b-cell signature in emphysema. *Am. J. Respir. Crit. Care Med.* (2016). doi:10.1164/rccm.201507-1311OC
53. Ando, K., Kaneko, N., Doi, T., Aoshima, M. & Takahashi, K. Prevalence and risk factors of aortic aneurysm in patients with chronic obstructive pulmonary disease. *J. Thorac. Dis.* **6,** 1388–1395 (2014).
54. Meijer, C. A. *et al.* An association between chronic obstructive pulmonary disease and abdominal aortic aneurysm beyond smoking: Results from a case-control study. *Eur. J. Vasc. Endovasc. Surg.* (2012). doi:10.1016/j.ejvs.2012.05.016
55. Ando, K., Kaneko, N., Doi, T., Aoshima, M. & Takahashi, K. Prevalence and risk factors of aortic aneurysm in patients with chronic obstructive pulmonary disease. *J. Thorac. Dis.* **6,** 1388–1395 (2014).
56. Verstraeten, A., Luyckx, I. & Loeys, B. Aetiology and management of hereditary aortopathy. *Nature Reviews Cardiology* (2017). doi:10.1038/nrcardio.2016.211
57. Yu, H. *et al.* High-quality binary protein interaction map of the yeast interactome network. *Science* (2008). doi:10.1126/science.1158684
58. Ahn, Y. Y., Bagrow, J. P. & Lehmann, S. Link communities reveal multiscale complexity in networks. *Nature* **466,** 761–764 (2010).
59. Feist, A. M. *et al.* A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol.* **3,** 1–18 (2007).
60. Fabregat, A. *et al.* The Reactome Pathway Knowledgebase. *Nucleic Acids Res.* (2018). doi:10.1093/nar/gkx1132
61. Florian-Rodriguez, M. *et al.* Effect of Protease Inhibitors in Healing of the Vaginal Wall. *Sci. Rep.* (2019). doi:10.1038/s41598-019-48527-0
62. Dabovic, B. *et al.* Function of latent TGFβ binding protein 4 and fibulin 5 in elastogenesis and lung development. *J. Cell. Physiol.* (2015). doi:10.1002/jcp.24704
63. Raykar, V. C., Duraiswami, R. & Zhao, L. H. Fast computation of kernel estimators. *J. Comput. Graph. Stat.* (2010). doi:10.1198/jcgs.2010.09046

Figures and tables

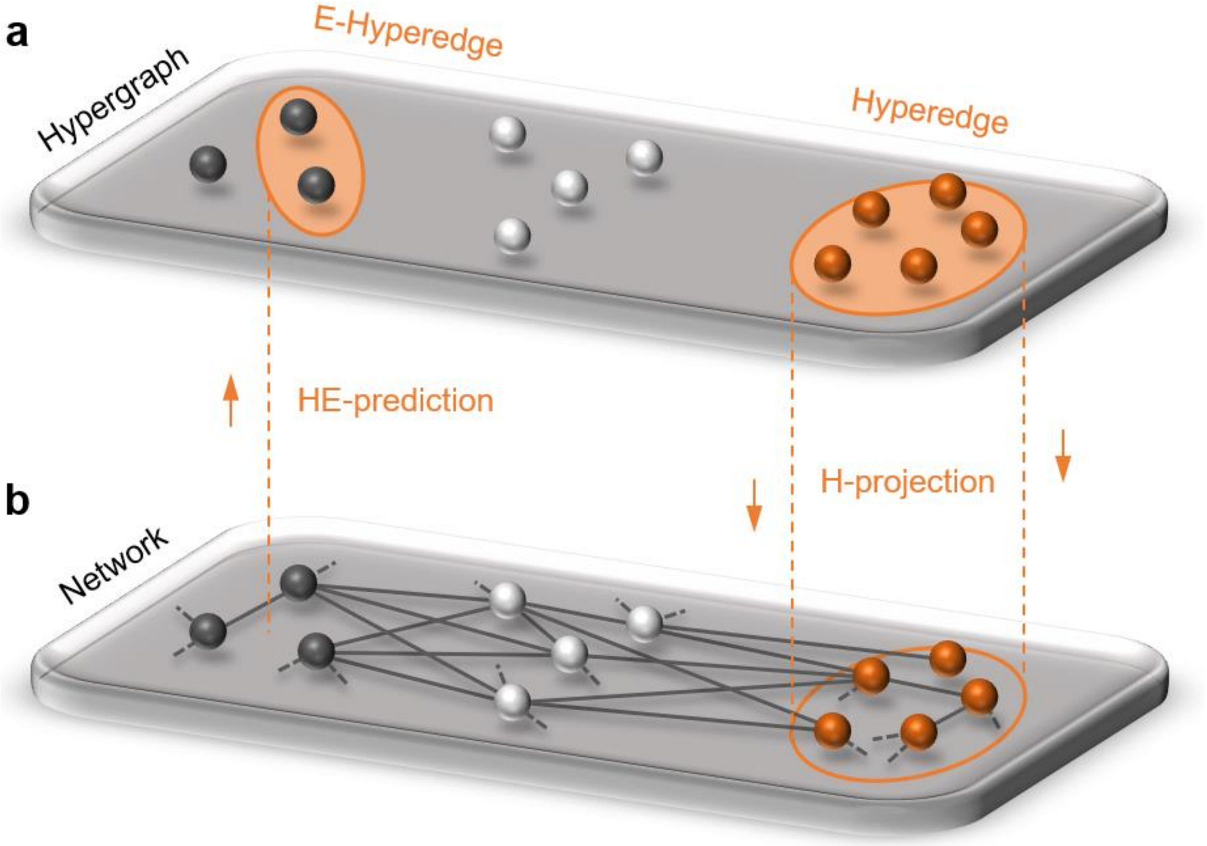
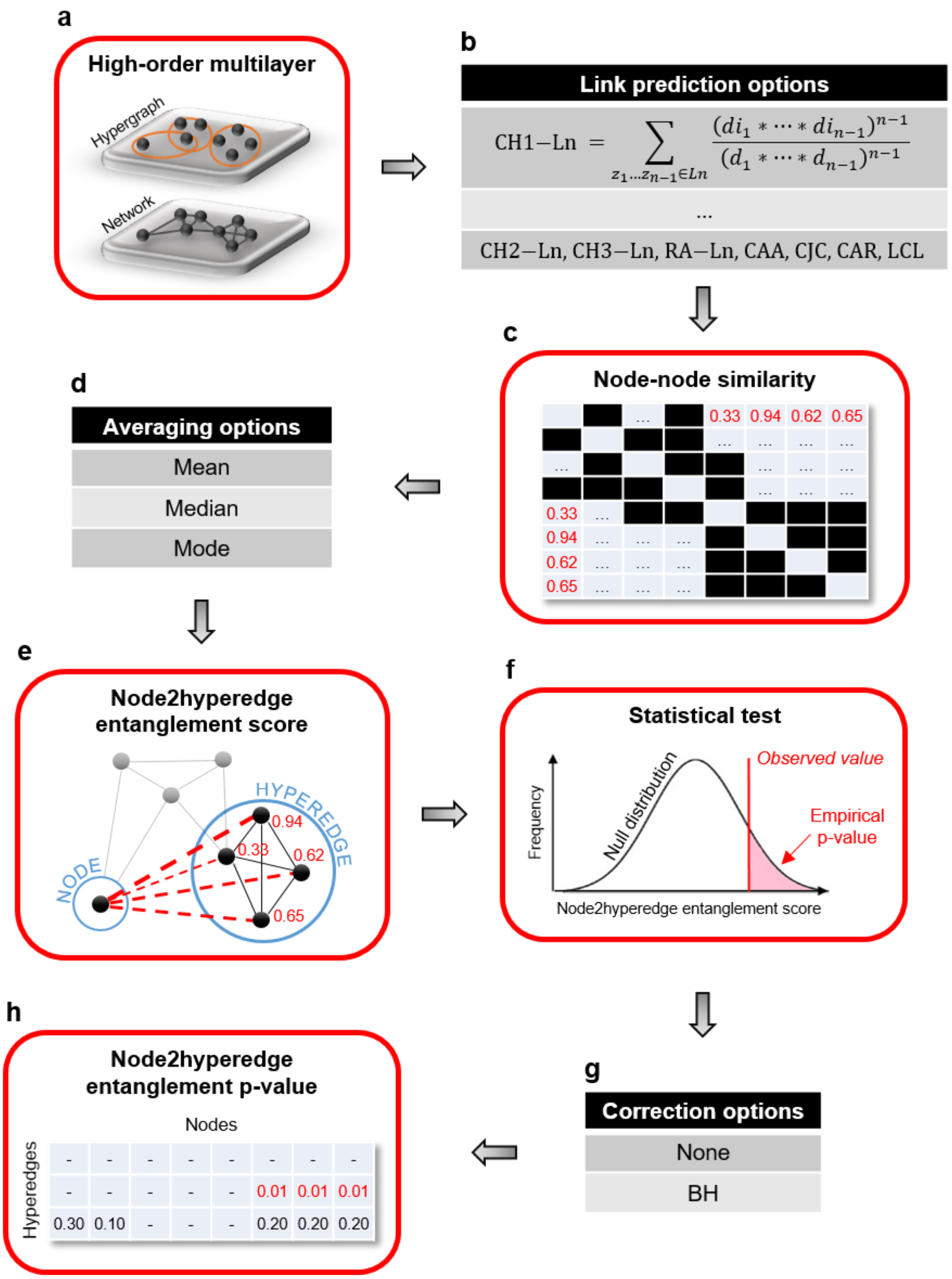


Figure 1. Visualization of hyperedge entanglement in high-order multilayer network.
The figure represents the hyperedge entanglement framework in high-order multilayer network proposed in this study. **a.** The first (top) high-order layer is a hypergraph $HG(V,H)$, where V is the set of vertices (or nodes) and H is the set of hyperedges. **b.** The second (bottom) layer is a graph $G(V,E)$, where V is the same set of vertices as the hypergraph HG and E is the set of edges (or links). Orange nodes are the members of a hyperedge and the information is projected on the network. Black nodes represent candidate nodes for entanglement (nodes that satisfy the ENC with the hyperedge). White nodes are not candidate nodes (because they do not satisfy the ENC), they are intermediate nodes in the network paths which enforce the entanglement between black and orange nodes. The HEP algorithm, which exploits link prediction on the network, predict (HE-prediction arrow) which ones of the black nodes are significantly entangled with the orange nodes (Hyperedge) and therefore are members of the entangled hyperedge (E-Hyperedge).


a
High-order multilayer
Hypergraph
Network
b
Link prediction options
$\mathrm{CH1{-}Ln} \;=\; \sum_{z_1 \dots z_{n-1} \in Ln} \frac{(di_1 * \dots * di_{n-1})^{n-1}}{(d_1 * \dots * d_{n-1})^{n-1}}$
...
CH2–Ln, CH3–Ln, RA–Ln, CAA, CJC, CAR, LCL
c
Node-node similarity
d
Averaging options
Mean
Median
Mode
e
Node2hyperedge entanglement score
HYPEREDGE
NODE
0.94
0.33
0.62
0.65
f
Statistical test
Frequency
Null distribution
Observed value
Empirical p-value
Node2hyperedge entanglement score
g
Correction options
None
BH
h
Node2hyperedge entanglement p-value
Nodes
Hyperedges

	■	...	■	0.33	0.94	0.62	0.65
■		■	■	...	...	...	...
...	■		■	■	...	...	...
■	■	■		■	...	...	...
0.33	...	■	■		■	■	■
0.94	...	...	...	■		■	■
0.62	...	...	...	■	■		■
0.65	...	...	...	■	■	■	

-	-	-	-	-	-	-	-
-	-	-	-	-	0.01	0.01	0.01
0.30	0.10	-	-	-	0.20	0.20	0.20

Figure 2. Flow chart of HEP algorithm.

The figure is a flow-chart representation of the HEP algorithmic steps which include also the N2EH algorithm (*b-f* panels). **a.** high-order multilayer network in input to the algorithm; **b**. choice of the link prediction option; **c**. computation of the node-node similarities between disconnected node pairs; **d**. choice of the average operator option for entanglement; **e**. computation of the node2hyperedge entanglement score; **f**. statistical test comparing the observed value of the node2hyperedge entanglement score with a null distribution, resulting in an empirical p-value; **g**. choice of the p-value correction option (if required); **h**. as output of the algorithm, a p-value for every candidate-node (a node that satisfies the ENC) to hyperedge pair, indicating if the pair also satisfy the ESC. For more details please refer to the Results section 2.1.

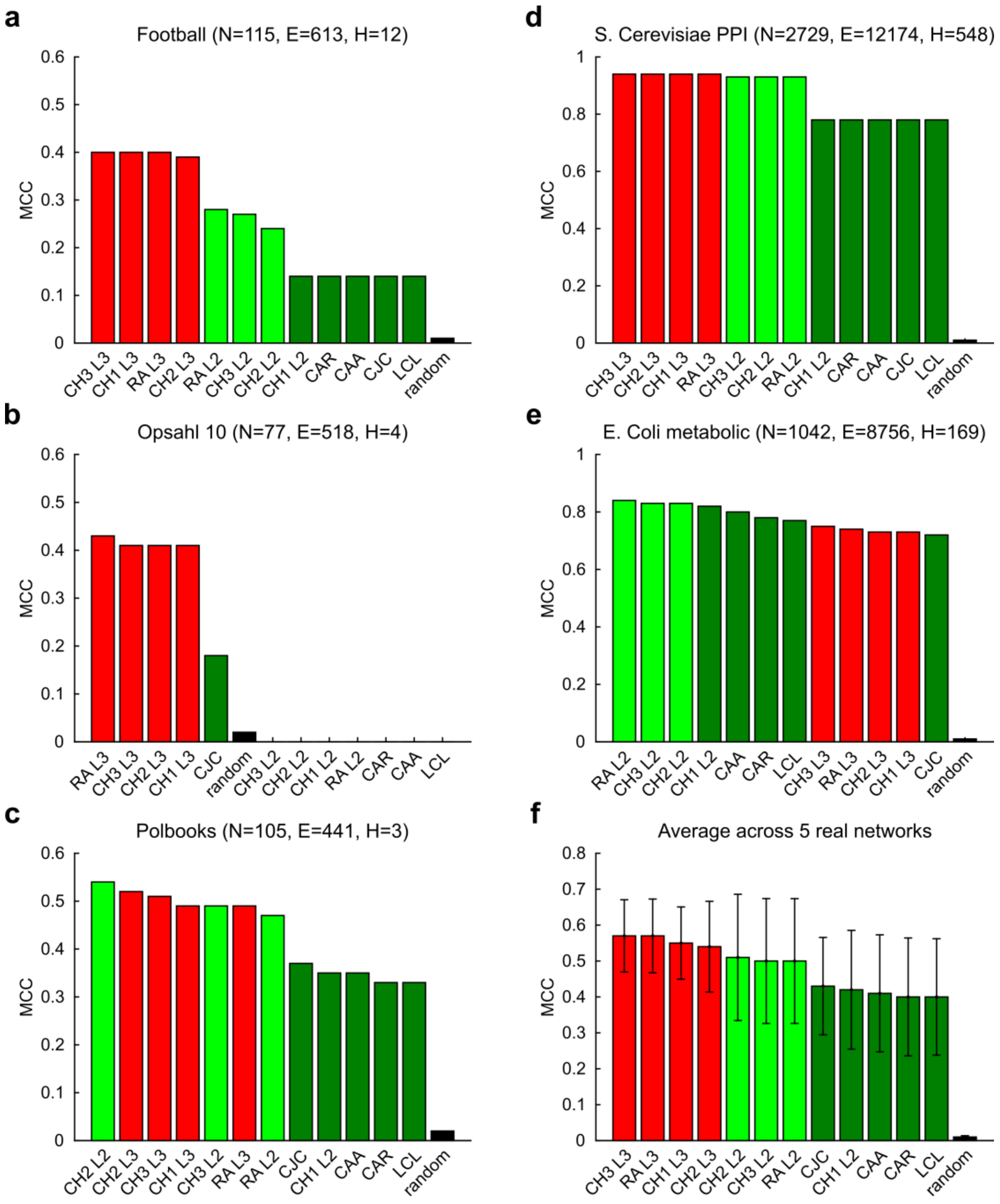


Figure 3. MCC evaluation of node2hyperedge entanglement in real networks.

The figure reports the results of the node2hyperedge entanglement simulation in 5 real networks: **a.** Football; **b.** Opsahl 10; **c.** Polbooks; **d.** S. Cerevisiae PPI; **e.** E. Coli metabolic. For each network, after creating a positive and negative set of node-hyperedge pairs, we run the node2hyperedge entanglement algorithm in order to obtain the predictions, and the performance is evaluated using the Matthews correlation coefficient (MCC). Please refer to section 2.2 for more details. For each link predictor only the best algorithmic variant is shown (regardless of average operator and p-value correction), whereas in the Suppl. Tables 2-6 all the variants are reported. Note that since there is only one MCC value for each network, no standard error is shown. **f.** The last panel reports the mean and standard error of the MCC over the 5 networks.

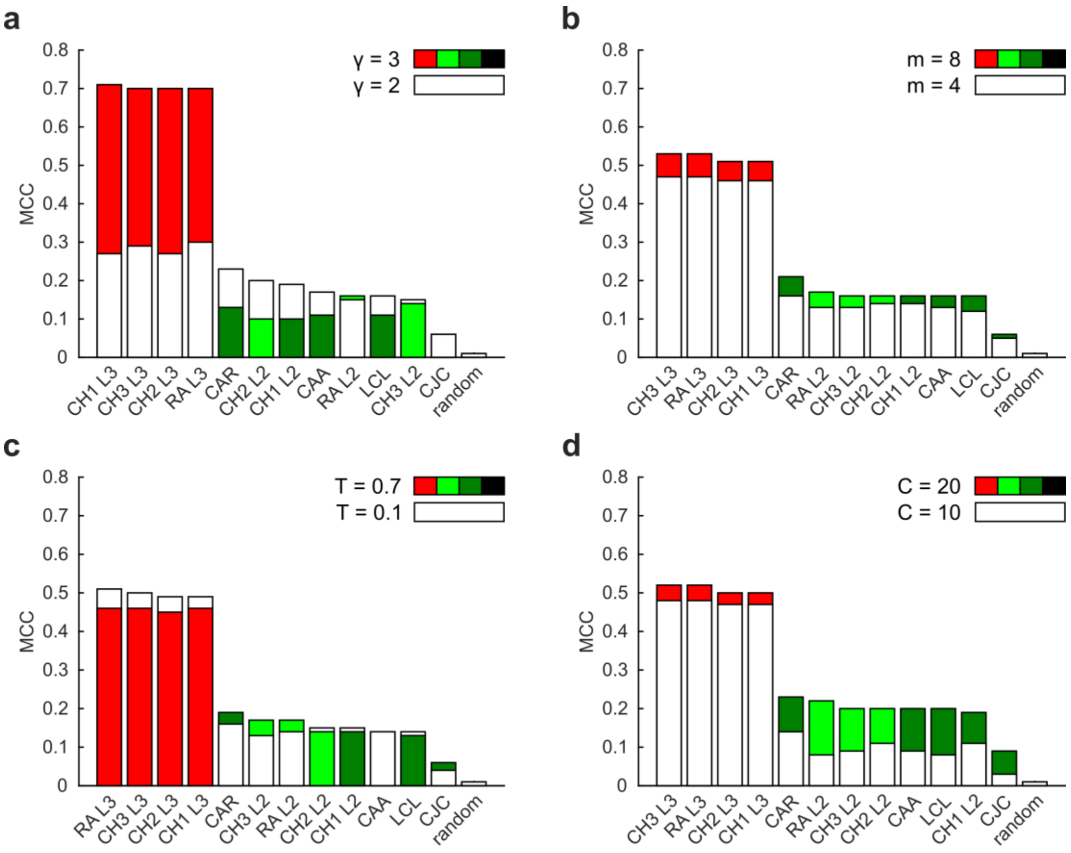


Figure 4. MCC evaluation of node2hyperedge entanglement in synthetic networks.
We generated synthetic networks using the nPSO model[7,8] with parameters N = 1000, γ = [2, 3], m = [4, 8], T = [0.1, 0.3, 0.5, 0.7] and C = [10, 20]. For each network, after creating a positive and negative set of node-hyperedge pairs, we run the node2hyperedge entanglement algorithm in order to obtain the predictions, and the performance is evaluated using the Matthews correlation coefficient (MCC). Please refer to section 2.3 for more details. The figure reports the MCC evaluation in four panels, each analyzing the variation in performance by changing a certain nPSO parameter.
a. The barplot shows for each link predictor the average MCC over all the networks with γ = 2 (white bars) compared with the average MCC over all the networks with γ = 3 (colored bars). The same applies in panels: **b.** m = 4 versus m = 8; **c.** T = 0.1 versus T = 0.7; **d.** C = 10 versus C = 20. Since we generated networks with four different values of T, in panel c we compared the two extreme ones. Note that for each link predictor the figure reports only the best variant, meaning the one with average operator and p-values correction option that provides the highest average performance across all the networks. Please refer to Suppl. Table 7 for the complete overview of all the algorithmic variants. For visual clarity, the standard error of the MCC is not shown, but it is reported in Suppl. Fig. 1.

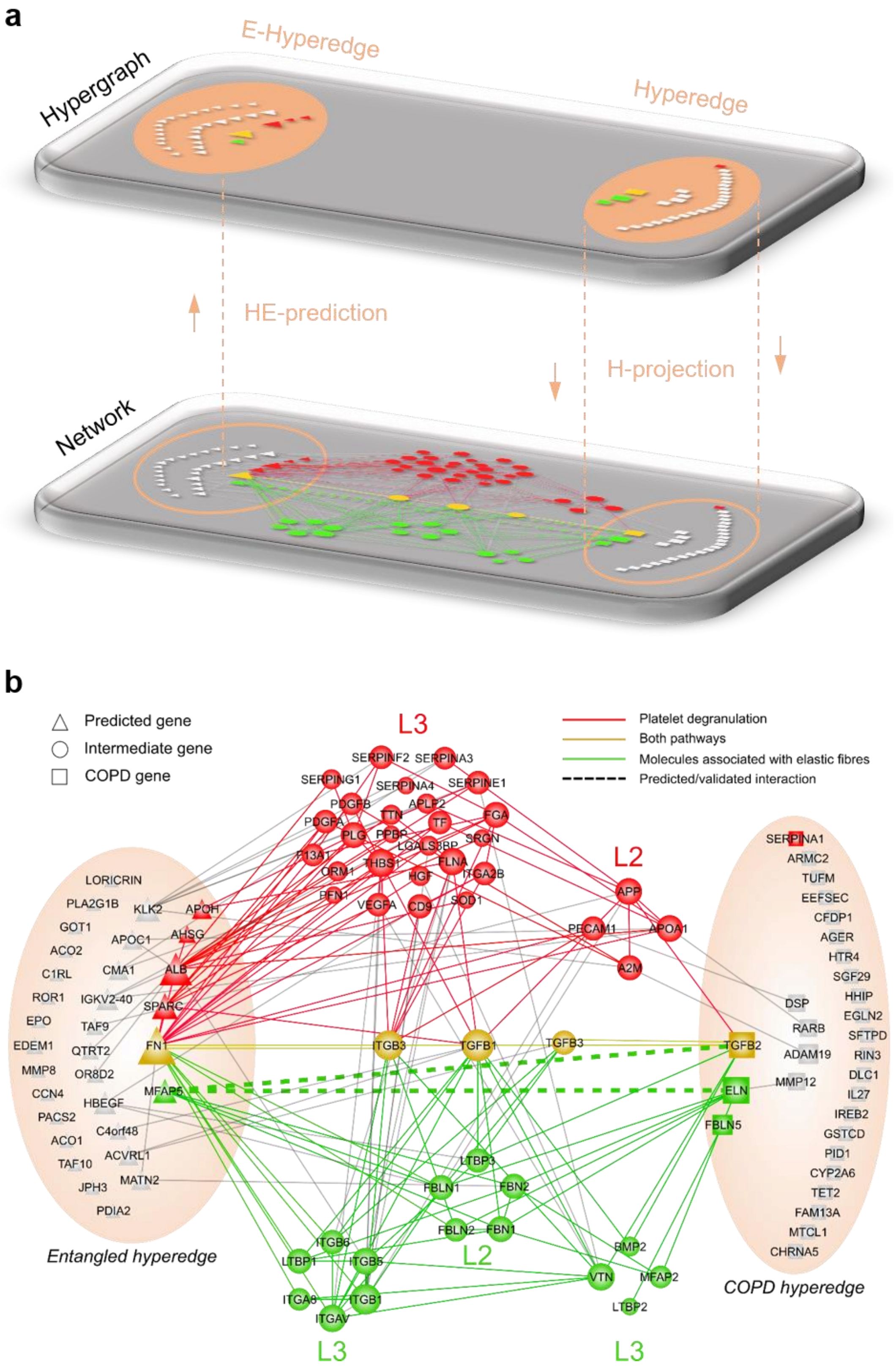

a
E-Hyperedge
Hypergraph
Hyperedge
HE-prediction
H-projection
Network
b
△ Predicted gene
○ Intermediate gene
□ COPD gene
Platelet degranulation
Both pathways
Molecules associated with elastic fibres
Predicted/validated interaction
L3
L2
L2
L3
L3
SERPINF2
SERPINA3
SERPING1
SERPINA4
SERPINE1
PDGFB
APLP2
TTN
PDGFA
TF
FGA
PLG
PPBP
LGALS3BP
SRGN
F13A1
THBS1
FLNA
ORM1
HGF
ITGA2B
PFN1
VEGFA
CD9
SOD1
APP
PECAM1
APOA1
A2M
LORICRIN
PLA2G1B
KLK2
APOH
GOT1
APOC1
AHSG
ACO2
C1RL
CMA1
ALB
ROR1
IGKV2-40
SPARC
EPO
TAF9
EDEM1
QTRT2
FN1
MMP8
OR8D2
MFAP5
CCN4
HBEGF
PACS2
C4orf48
ACO1
ACVRL1
TAF10
MATN2
JPH3
PDIA2
Entangled hyperedge
ITGB3
TGFB1
TGFB3
TGFB2
ELN
FBLN5
LTBP3
FBLN1
FBN2
FBLN2
FBN1
ITGB6
LTBP1
ITGB5
ITGA8
ITGB1
ITGAV
BMP2
VTN
MFAP2
LTBP2
SERPINA1
ARMC2
TUFM
EEFSEC
CFDP1
AGER
HTR4
SGF29
HHIP
DSP
EGLN2
RARB
SFTPD
ADAM19
RIN3
MMP12
DLC1
IL27
IREB2
GSTCD
PID1
CYP2A6
TET2
FAM13A
MTCL1
CHRNA5
COPD hyperedge

Figure 5. Prediction of entanglement hyperedge of COPD disease hyperedge

a. Visual representation of the HEP algorithm performed on the high-order multilayer graph, having the COPD disease hyperedge at the upper layer and the PPI network at the bottom layer.

b. On the left, the 32 genes (triangle symbol) of the entangled hyperedge (predicted information). On the centre, the 47 intermediate genes (circle symbol) that belong to paths of length two (L2) or length three (L3) and that are included in at least one of the two enriched pathways: platelet degranulation (red colour) and molecules associated with elastic fibres (green colour). On the right, the 29 genes (square symbol) of the COPD hyperedge (known information). Grey nodes are COPD genes or entangled genes that are not part of the pathways, but they are still reported for completeness. Solid lines are for interactions between the nodes in the PPI network, dashed lines are for the interactions experimentally validated in this study. Colours of nodes and links highlight their association to the two pathways, as indicated in the legend. Grey links are between nodes that are not associated to the same pathway.

a

CAR	CAA	CJC	LCL	CH1 L2	CH2 L2	CH3 L2	RA L2	CH1 L3	CH2 L3	CH3 L3	RA L3	Rand
0.55	1.00	0.83	1.00	1.00	0.37	0.37	0.37	0.06	0.04	0.02	0.01	0.90

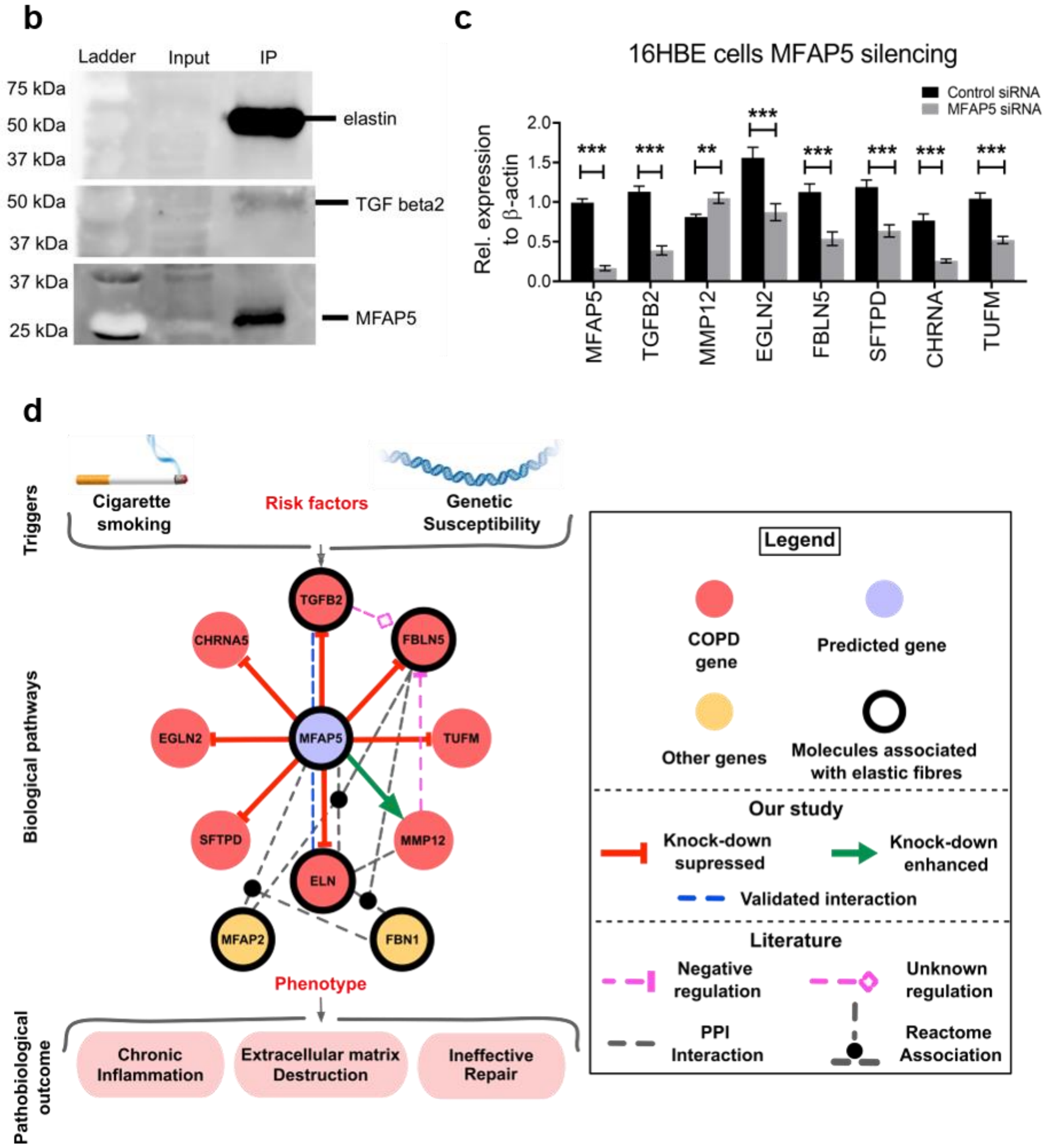


Figure 6. Computational and experimental validation on COPD disease

a. p-values of possible entanglement between MFAP5 and COPD disease hyperedge assessed according to the 12 variants of the HEP algorithm. "Rand" corresponds to the random predictor. The entanglement is significant (p-value < 0.05) according to three L3-based methods (CH2-L3, CH3-L3, RA-L3).

b. Immunoblot of proteins after immunoprecipitation with Microfibril Associated Protein 5 (MFAP5) antibody (IP) and staining with elastin and TGFB2 antibody. MFAP5 interaction with ELN and TGFB2 is confirmed.

c. MFAP5 siRNA suppresses the expression of TGFB2, EGLN2, FBLN5, SFTPD, CHRNA5 and TUFM and enhances the expression of MMP12 at mRNA level.

d. Summary of the known knowledge and new knowledge gained in this study. The new knowledge from our prediction (panel *a*) supported by the experimental validations (panels *b* and *c*) is represented with a red arrow when, as a consequence of MFAP5 regulation, the expression of a COPD-related disease gene is decreased (repression), and with a green arrow when it is increased (enhancement). The dashed blue line indicates the validated interactions. Genes associated to the biological pathway of molecules associated with elastic fibres are highlighted with a black border. References for literature-curated interactions: the complex associations come from Reactome[60]; MMP12 cleaves FBLN5 *in vitro* and may mediate injury-induced loss of FBLN5[61] (negative regulation); TGFB2 may regulate indirectly FBLN5 by the change of its level in the lungs[62] (unknown regulation); the interaction between ELN and MMP12 comes from the PPI network considered in this study[36].

doi:10.20944/preprints202012.0500.v1

Supplementary Information

Computational complexity

- L2-based link predictors

The L2-based algorithms for topological link prediction consist of a main loop exploring all the non-observed links, and evaluating the likelihood of one link at each iteration.

Given the number of nodes *N* and the number of observed links *E*, the number of iterations is:

$$\frac{N(N-1)}{2} - E$$

For a given evaluation of the candidate link between nodes *i* and *j*, the dominant operation is to find the common neighbours of *i* and *j*. In order to do this, for each neighbour of *i*, we need to check if it is connected to *j*. The time complexity of this computation is linear on the degree of *i*, therefore on average it is linear on the average node degree of the network *k*:

$$O(k) = O\left(\frac{E}{N}\right)$$

Given that we perform $\frac{N(N-1)}{2} - E$ iterations with average complexity $O\left(\frac{E}{N}\right)$, the overall complexity is:

$$O\left(\left(\frac{N(N-1)}{2} - E\right) * \frac{E}{N}\right) = O\left(\frac{E(N-1)}{2} - \frac{E^2}{N}\right)$$

Gathering the factor $\frac{E(N-1)}{2}$ we obtain:

$$O\left(\frac{E(N-1)}{2}\left(1 - \frac{2E}{N(N-1)}\right)\right) = O\left(\frac{E(N-1)}{2}(1-D)\right)$$

where D is the network density $D = \frac{2E}{N(N-1)}$.

Removing the multiplicative factor $\frac{1}{2}$ and replacing (*N* – 1) with just *N*, we can rewrite in a more compact form:

$$O\big(EN(1-D)\big)$$

The complexity analysis is remarkable in three particular cases:

(1) Minimum number of links for a connected network (tree): $E = N - 1$

$$O\left(\frac{E(N-1)}{2}\left(1 - \frac{2E}{N(N-1)}\right)\right) = O\left(\frac{(N-1)^2}{2}\left(1 - \frac{2(N-1)}{N(N-1)}\right)\right)$$

$$= O\left(\frac{(N-1)^2}{2} - \frac{(N-1)^2}{N}\right) = O\left(N^2 - \frac{N^2}{N}\right) = O(N^2)$$

(2) Half of the number of possible links: $E = \frac{N(N-1)}{4}$

$$O\left(\frac{E(N-1)}{2}\left(1-\frac{2E}{N(N-1)}\right)\right) = O\left(\frac{N(N-1)^2}{8}\left(1-\frac{1}{2}\right)\right) = O\left(\frac{N(N-1)^2}{16}\right) = O(N^3)$$

(3) Fully connected network (no non-observed links to evaluate): $E = \frac{N(N-1)}{2}$

$$O\left(\frac{E(N-1)}{2}\left(1-\frac{2E}{N(N-1)}\right)\right) = O\left(\frac{N(N-1)^2}{8}(1-1)\right) = 0$$

This suggests that running the algorithm over a sparse network achieves complexity $O(N^2)$, which increases to $O(N^3)$ for intermediate values of network density. However, for densely connected networks the computational cost decreases, ideally reaching zero for a fully connected network without non-observed links to evaluate.

Since reasonable values of density for real networks are generally low or intermediate, we may assert that within the domain of real and practical problems in which topological link prediction is applied, the complexity of L2-based algorithms can be expressed as $O(EN)$, and very often approximated by *O(N*^{2}*)*.

Note that the link likelihoods are computed independently of each other and therefore the implementation can be easily parallelized in order to speed up the running time.

- *L3-based link predictors*

The L3-based algorithms for topological link prediction, analogously to the L2-based, consist of a main loop exploring all the non-observed links, and evaluating the likelihood of one link at each iteration.

In this case, for a given evaluation of the candidate link between nodes *i* and *j*, the dominant operation is to find the L3 paths between *i* and *j*. In order to do this, for each neighbour *l* of *i*, we need to check if each neighbour of *l* is connected to *j*. The time complexity of this computation is dependent on the degree of *i* multiplied by the degree of *l*, therefore on average it is dependent on the squared average node degree of the network *k*:

$$O(k^2) = O\left(\frac{E^2}{N^2}\right)$$

Given that we perform $\frac{N(N-1)}{2} - E$ iterations with average complexity $O\left(\frac{E^2}{N^2}\right)$, the overall complexity is:

$$O\left(\left(\frac{N(N-1)}{2} - E\right) * \frac{E^2}{N^2}\right)$$

The complexity analysis is remarkable in three particular cases:

(1) Minimum number of links for a connected network (tree): $E = N - 1 \cong N$

$$O\left(\left(\frac{N(N-1)}{2} - E\right) * \frac{E^2}{N^2}\right) = O\left(\left(\frac{N(N-1)}{2} - N\right) * \frac{N^2}{N^2}\right)$$

$$= O\left(\frac{N(N-1)}{2} - N\right) = O\left(\frac{N^2}{2} - \frac{N}{2} - N\right) = O(N^2)$$

(2) Half of the number of possible links: $E = \frac{N(N-1)}{4}$

$$O\left(\left(\frac{N(N-1)}{2} - E\right) * \frac{E^2}{N^2}\right) = O\left(\left(\frac{N(N-1)}{2} - \frac{N(N-1)}{4}\right) * \frac{\left(\frac{N(N-1)}{4}\right)^2}{N^2}\right)$$

$$= O\left(\frac{N(N-1)}{4} * \frac{N^2(N-1)^2}{16 * N^2}\right) = O\left(\frac{N(N-1)^3}{64}\right) = O(N^4)$$

(3) Fully connected network: no non-observed links to evaluate.

This suggests that running the algorithm over a sparse network achieves complexity $O(N^2)$, which increases to $O(N^4)$ for intermediate values of network density.

- HEP

For each node-hyperedge pair that satisfies the ENC, the HEP computes the node2hyperedge entanglement p-value (for details see the Results sections 2.1.5 and 2.1.6). Given N nodes and H hyperedges, the node-hyperedges pairs to evaluate are $O(N * H)$.

As preliminary step, we can compute the similarity scores for all the pairs of disconnected nodes in the network, with time complexity dependent on the link predictor, as discussed in the previous sections. Therefore such operation does not have to be repeated for each node-hyperedge pair. Given a node-hyperedge pair and the node-node similarities already computed, the algorithm for the node2hyperedge entanglement p-value requires the following operations (see section 2.1.5):

- Computation of the node2hyperedge entanglement score, using one of the three average operators: mean, median, mode. Mean and median can be executed with time complexity linear to the number of samples to average. The mode option, as described in section 2.1.6, is based on a kernel smoothing function estimate evaluated at 100 points, which can also be implemented with linear time complexity[63]. The number of samples to average is equal to the average size of the hyperedge, that we can here refer with s. The time complexity of this step is $O(s)$.

- Generation of M random hyperedges. In order to sample the members of one random hyperedge the time complexity is equal to the average hyperedge size, and this procedure is repeated M times. The time complexity of this step is $O(M * s)$.
- Computation of the null-distribution of node2hyperedge entanglement scores. This is equivalent to computing M times a node2hyperedge entanglement score, therefore the time complexity is $O(M * s)$.
- Computation of the empirical p-value. This operation requires M comparisons, therefore the time complexity is $O(M)$.

The dominant operation among the four listed is $O(M * s)$. A reasonable assumption is that the average hyperedge size is approximately equal to $s = \frac{N}{H}$.

Since the four operations above are repeated $O(N * H)$ times, we obtain:

$$O\left(N * H * M * \frac{N}{H}\right) = O(N^2 * M)$$

Note that in our simulations the number of repetitions M is fixed to 1000, therefore it can be considered as a constant factor and indicate the time complexity of the statistical test as $O(N^2)$.

At last, for each node, the Benjamini-Hochberg correction can be performed as an option to adjust for multiple hypothesis testing over the hyperedges, with an overall cost of $O(N * HlogH)$. Since generally $N \gg H$, we can consider such complexity lower than or equal to $O(N^2)$.

In summary, the time complexity of the HEP algorithm is the maximum between $O(N^2)$ and the complexity of the link predictor. Considering the link predictors adopted in this study, in case of sparse networks, the complexity is $O(N^2)$.

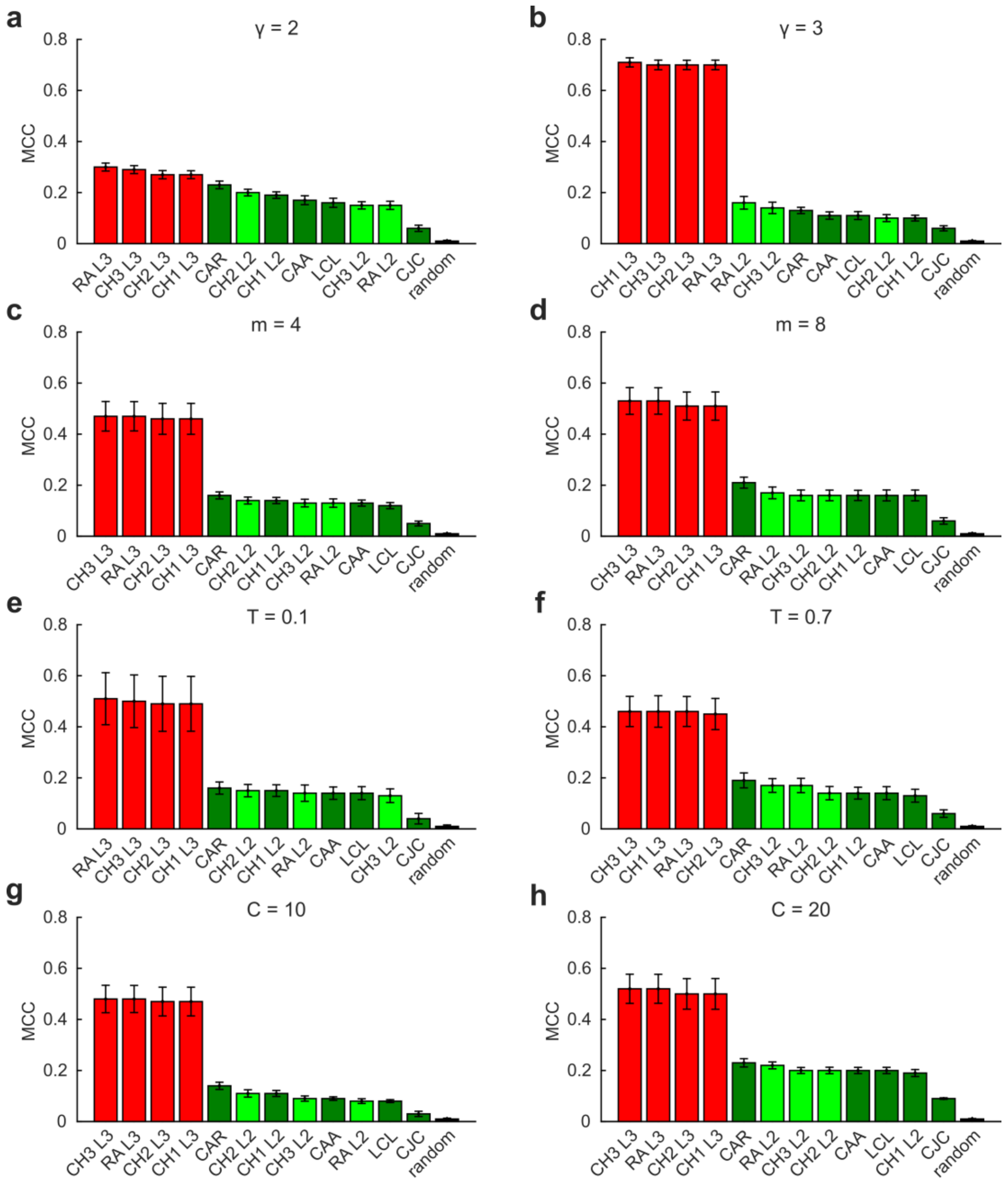


Figure S1. MCC evaluation of node2hyperedge entanglement in synthetic networks. The figure reports the same results as Fig. 4, but the performance related to the two values of each nPSO parameter is shown in separate barplots. In addition, the standard error of the MCC is also reported.

	Network						Hypergraph	
	N	***E***	***m***	***Cl***	*γ*	***LCP-corr***	***H***	***overlapping***
Football	115	613	5.33	0.40	9.09	0.89	12	No
Opsahl 10	77	518	6.73	0.65	5.06	0.96	4	No
Polbooks	105	441	4.20	0.49	2.62	0.94	3	No
S. cerevisiae PPI	2729	12174	4.46	0.29	3.03	0.96	548	Yes
E. coli metabolic	1042	8756	8.40	0.73	2.24	0.93	169	Yes

Table S1. Main topological properties of real networks and hypergraph information. The table reports the following topological properties of the 5 real networks considered in this study: number of nodes *N*, number of edges *E*, half of average node degree *m*, clustering coefficient *Cl*, exponent *γ* of the fitted power-law degree distribution, *LCP-corr* that is a measure of local community organization of the networks. The table also reports the number of hyperedges *H* in the hypergraph and whether the hyperedges are overlapping or not.

Football							
CH3 L3 mode	0.40	CH2 L3 mode C	0.25	CH3 L2 median C	0.14	CJC mode	0.12
CH1 L3 mode	0.40	CH1 L3 median C	0.25	CH1 L2 mean	0.14	LCL mean	0.12
RA L3 median	0.40	CH2 L2 mode	0.24	CH1 L2 median	0.14	LCL median C	0.12
CH2 L3 mode	0.39	CH2 L2 mean	0.23	CH1 L2 median C	0.14	CH1 L2 mean C	0.10
RA L3 mode	0.38	CH3 L2 median	0.22	CH1 L2 mode	0.14	CAR mean	0.10
CH3 L3 median	0.37	CH2 L2 median	0.22	RA L2 median C	0.14	CAA mean C	0.10
CH1 L3 median	0.37	RA L2 median	0.22	CAR median	0.14	CJC mean C	0.10
CH2 L3 median	0.36	CH1 L3 mean	0.21	CAA mean	0.14	CJC mode C	0.10
RA L3 mean	0.32	RA L3 mean C	0.21	CAA median	0.14	LCL mode C	0.10
CH3 L3 mode C	0.30	CH3 L2 mean C	0.18	CAA median C	0.14	CAR mean C	0.07
RA L3 mode C	0.29	CH3 L2 mode C	0.18	CAA mode	0.14	CAR mode C	0.07
RA L2 mean	0.28	CH2 L2 mean C	0.18	CJC median	0.14	LCL mean C	0.07
CH3 L3 median C	0.27	RA L2 mode C	0.18	CJC median C	0.14	random mode	0.01
CH3 L2 mode	0.27	CH2 L3 mean	0.17	LCL median	0.14	random mean	0.00
RA L2 mode	0.27	RA L2 mean C	0.17	LCL mode	0.14	random mean C	0.00
CH1 L3 mode C	0.26	CH3 L3 mean C	0.15	CH1 L2 mode C	0.12	random median	-0.02
CH3 L2 mean	0.26	CH2 L3 mean C	0.15	CAR median C	0.12	random median C	-0.02
RA L3 median C	0.26	CH1 L3 mean C	0.15	CAR mode	0.12	random mode C	-0.03
CH3 L3 mean	0.25	CH2 L2 median C	0.15	CAA mode C	0.12		
CH2 L3 median C	0.25	CH2 L2 mode C	0.15	CJC mean	0.12		

Table S2. MCC evaluation on Football network for all the algorithmic variants.
For each HEP algorithmic variant (link predictor, average operator and p-value correction), the table reports the MCC evaluated on the Football network.

Opsahl 10							
RA L3 mode	0.43	random mode C	0.02	CH1 L2 mean	0.00	CAA mean C	0.00
CH3 L3 mode	0.41	random median C	0.01	CH1 L2 mean C	0.00	CAA median	0.00
CH2 L3 mode	0.41	random mode	0.01	CH1 L2 median	0.00	CAA median C	0.00
CH1 L3 mode C	0.41	CH3 L3 mean	0.00	CH1 L2 median C	0.00	CAA mode	0.00
RA L3 mode C	0.41	CH3 L3 mean C	0.00	CH1 L2 mode	0.00	CAA mode C	0.00
CH3 L3 mode C	0.39	CH2 L3 mean	0.00	CH1 L2 mode C	0.00	CJC median	0.00
CH1 L3 mode	0.39	CH2 L3 mean C	0.00	RA L3 mean C	0.00	CJC median C	0.00
CH3 L3 median	0.34	CH1 L3 mean C	0.00	RA L2 mean	0.00	CJC mode	0.00
CH3 L3 median C	0.34	CH3 L2 mean	0.00	RA L2 mean C	0.00	CJC mode C	0.00
CH2 L3 median	0.34	CH3 L2 mean C	0.00	RA L2 median	0.00	LCL mean	0.00
CH2 L3 median C	0.34	CH3 L2 median	0.00	RA L2 median C	0.00	LCL mean C	0.00
CH2 L3 mode C	0.34	CH3 L2 median C	0.00	RA L2 mode	0.00	LCL median	0.00
CH1 L3 median	0.34	CH3 L2 mode	0.00	RA L2 mode C	0.00	LCL median C	0.00
CH1 L3 median C	0.34	CH3 L2 mode C	0.00	CAR mean	0.00	LCL mode	0.00
RA L3 median	0.34	CH2 L2 mean	0.00	CAR mean C	0.00	LCL mode C	0.00
RA L3 median C	0.34	CH2 L2 mean C	0.00	CAR median	0.00	random median	0.00
RA L3 mean	0.21	CH2 L2 median	0.00	CAR median C	0.00	random mean	-0.01
CH1 L3 mean	0.18	CH2 L2 median C	0.00	CAR mode	0.00	random mean C	-0.01
CJC mean	0.18	CH2 L2 mode	0.00	CAR mode C	0.00		
CJC mean C	0.10	CH2 L2 mode C	0.00	CAA mean	0.00		

Table S3. MCC evaluation on Opsahl 10 network for all the algorithmic variants. For each HEP algorithmic variant (link predictor, average operator and p-value correction), the table reports the MCC evaluated on the Opsahl 10 network.

Polbooks							
CH2 L2 mean	0.54	RA L2 mean	0.47	CH3 L3 median	0.32	CH3 L2 median C	0.00
CH2 L2 mean C	0.54	RA L2 mean C	0.47	CH3 L3 median C	0.32	CH2 L2 median	0.00
CH2 L3 mean C	0.52	CH2 L3 mode	0.45	CH2 L3 median	0.32	CH2 L2 median C	0.00
CH3 L3 mean	0.51	CH3 L2 mode C	0.45	CH2 L3 median C	0.32	CH1 L2 median	0.00
CH3 L3 mean C	0.51	RA L3 mean	0.45	CH1 L3 median	0.32	CH1 L2 median C	0.00
CH3 L3 mode	0.49	CH2 L3 mode C	0.43	RA L3 median	0.32	RA L2 median	0.00
CH3 L3 mode C	0.49	RA L2 mode	0.43	CH1 L2 mode	0.28	RA L2 median C	0.00
CH2 L3 mean	0.49	RA L2 mode C	0.43	CH1 L2 mode C	0.28	CAR median	0.00
CH1 L3 mean	0.49	CJC mean	0.37	CAR mode	0.28	CAR median C	0.00
CH1 L3 mode	0.49	CJC mean C	0.37	CAR mode C	0.28	CAA median	0.00
CH1 L3 mode C	0.49	CH1 L3 median C	0.36	CAA mode	0.28	CAA median C	0.00
CH3 L2 mean	0.49	RA L3 median C	0.36	CAA mode C	0.28	CJC median	0.00
CH3 L2 mean C	0.49	CH1 L2 mean	0.35	CJC mode	0.28	CJC median C	0.00
CH2 L2 mode	0.49	CH1 L2 mean C	0.35	CJC mode C	0.28	LCL median	0.00
RA L3 mean C	0.49	CAA mean	0.35	LCL mode	0.28	LCL median C	0.00
CH1 L3 mean C	0.47	CAA mean C	0.35	LCL mode C	0.28	random mode	-0.02
CH3 L2 mode	0.47	CAR mean	0.33	random mean C	0.02	random mode C	-0.02
CH2 L2 mode C	0.47	CAR mean C	0.33	random median C	0.02	random median	-0.04
RA L3 mode	0.47	LCL mean	0.33	random mean	0.01		
RA L3 mode C	0.47	LCL mean C	0.33	CH3 L2 median	0.00		

Table S4. MCC evaluation on Polbooks network for all the algorithmic variants.
For each HEP algorithmic variant (link predictor, average operator and p-value correction), the table reports the MCC evaluated on the Polbooks network.

S. Cerevisiae PPI							
CH3 L3 mean	0.94	RA L2 median	0.89	RA L3 mode C	0.67	CAA mode C	0.33
CH3 L3 median	0.94	RA L3 median C	0.83	CH3 L2 median C	0.65	CH3 L2 mean C	0.31
CH3 L3 mode	0.94	CH3 L3 median C	0.82	RA L2 median C	0.65	CH2 L2 mode C	0.30
CH2 L3 median	0.94	CH2 L3 median C	0.81	CH2 L2 median C	0.64	CH1 L2 mode C	0.30
CH1 L3 median	0.94	CH1 L3 median C	0.81	RA L3 mean C	0.58	CAA mean C	0.30
RA L3 mean	0.94	CH1 L2 mean	0.78	CH1 L3 mode C	0.55	LCL mean C	0.30
RA L3 median	0.94	CAR mean	0.78	CH3 L3 mode C	0.51	CH2 L2 mean C	0.29
RA L3 mode	0.94	CAA mean	0.78	CH1 L2 median C	0.50	CH1 L2 mean C	0.28
CH2 L3 mean	0.93	CJC mean	0.78	RA L2 mode C	0.50	CAR mean C	0.27
CH2 L3 mode	0.93	LCL mean	0.78	CAR median C	0.50	CAR mode C	0.25
CH1 L3 mean	0.93	CAA mode	0.77	CAA median C	0.50	CJC mean C	0.23
CH1 L3 mode	0.93	CH1 L2 mode	0.76	CJC median C	0.50	CJC mode C	0.23
CH3 L2 mean	0.93	LCL mode	0.76	LCL median C	0.50	random mean	0.01
CH3 L2 mode	0.93	CAR mode	0.74	CH1 L3 mean C	0.47	random median	0.01
CH2 L2 mean	0.93	CJC mode	0.74	CH2 L3 mode C	0.43	random mean C	0.00
RA L2 mean	0.93	CH1 L2 median	0.72	CH3 L3 mean C	0.39	random median C	0.00
RA L2 mode	0.93	CAR median	0.72	CH3 L2 mode C	0.38	random mode	0.00
CH2 L2 mode	0.91	CAA median	0.72	RA L2 mean C	0.37	random mode C	0.00
CH3 L2 median	0.89	CJC median	0.72	CH2 L3 mean C	0.36		
CH2 L2 median	0.89	LCL median	0.72	LCL mode C	0.34		

Table S5. MCC evaluation on S. Cerevisiae PPI network for all the algorithmic variants. For each HEP algorithmic variant (link predictor, average operator and p-value correction), the table reports the MCC evaluated on the S. Cerevisiae PPI network.

E. Coli metabolic							
RA L2 mean	0.84	CH3 L3 mode	0.61	CAR median	0.55	CJC median C	0.42
CH3 L2 mean	0.83	CH1 L3 median C	0.61	LCL median	0.55	CH3 L3 mean C	0.41
CH2 L2 mean	0.83	CH3 L2 median	0.61	CAR mode	0.54	CH2 L3 mode C	0.41
CH1 L2 mean	0.82	CH2 L2 median	0.61	CJC mode	0.53	CH3 L2 mode C	0.41
CAA mean	0.80	CH1 L2 mean C	0.61	CH2 L3 mode	0.51	CJC mean C	0.41
CAR mean	0.78	CH1 L2 median	0.61	CH1 L3 mean	0.51	CH1 L3 mean C	0.40
LCL mean	0.77	RA L2 median	0.61	RA L3 mode C	0.51	CH3 L2 median C	0.40
CH3 L3 median	0.75	CAA median	0.61	CH3 L3 mode C	0.50	RA L2 median C	0.40
RA L3 median	0.74	CAA mode	0.60	CH2 L3 mean	0.49	CH2 L3 mean C	0.38
CH2 L3 median	0.73	CJC median	0.60	LCL mode C	0.47	CAR median C	0.36
CH1 L3 median	0.73	LCL mode	0.60	RA L3 mean C	0.46	LCL median C	0.35
CJC mean	0.72	CH2 L2 mean C	0.59	CAA mode C	0.46	CJC mode C	0.30
CAA mean C	0.68	RA L2 mean C	0.59	CH1 L3 mode C	0.43	random mean	0.01
LCL mean C	0.67	CH3 L2 mode	0.58	CH2 L2 median C	0.43	random mean C	0.01
RA L3 mode	0.65	RA L3 mean	0.58	RA L2 mode C	0.43	random median	0.01
CAR mean C	0.65	CH3 L3 mean	0.57	CH2 L2 mode C	0.42	random mode	0.01
CH3 L3 median C	0.64	CH1 L3 mode	0.57	CH1 L2 median C	0.42	random median C	0.00
RA L3 median C	0.64	CH3 L2 mean C	0.57	CH1 L2 mode C	0.42	random mode C	0.00
CH2 L3 median C	0.63	CH1 L2 mode	0.57	CAR mode C	0.42		
RA L2 mode	0.62	CH2 L2 mode	0.56	CAA median C	0.42		

Table S6. MCC evaluation on E. Coli metabolic network for all the algorithmic variants. For each HEP algorithmic variant (link predictor, average operator and p-value correction), the table reports the MCC evaluated on the E. Coli metabolic network.

nPSO							
CH3 L3 median	0.50	RA L3 mean C	0.24	RA L2 median	0.11	CAA median	0.04
RA L3 median	0.50	CH2 L3 mean C	0.21	CH3 L2 median C	0.10	CJC mean	0.04
CH2 L3 median	0.49	CH1 L3 mean C	0.21	CH2 L2 median	0.10	CH1 L2 median C	0.03
CH1 L3 median	0.49	CH3 L3 mean C	0.20	CH2 L2 median C	0.10	CAA median C	0.03
CH3 L3 median C	0.48	CAR mean	0.18	RA L2 median C	0.10	CJC median	0.03
RA L3 median C	0.48	CH3 L2 mode	0.15	CH2 L2 mean C	0.09	CJC median C	0.03
CH2 L3 median C	0.47	CH2 L2 mean	0.15	CH2 L2 mode C	0.09	CJC mode C	0.03
CH1 L3 median C	0.47	CH1 L2 mean	0.15	CH3 L2 mode C	0.08	CAR median	0.02
RA L3 mode	0.42	RA L2 mode	0.15	CH1 L2 mean C	0.08	CAR median C	0.02
CH3 L3 mode	0.40	CH3 L2 mean	0.14	RA L2 mode C	0.08	CJC mean C	0.02
RA L3 mode C	0.39	CH2 L2 mode	0.14	CAA mean C	0.08	LCL median	0.02
CH3 L3 mode C	0.37	CAA mean	0.14	CH3 L2 mean C	0.07	LCL median C	0.02
RA L3 mean	0.36	LCL mean	0.14	CAR mode C	0.07	random mean C	0.01
CH1 L3 mean	0.34	RA L2 mean	0.13	LCL mean C	0.07	random median C	0.01
CH1 L3 mode	0.33	CAR mode	0.13	CH1 L2 mode C	0.06	random mode C	0.01
CH3 L3 mean	0.32	CAR mean C	0.12	RA L2 mean C	0.06	random mean	0.00
CH2 L3 mean	0.32	CAA mode	0.12	CAA mode C	0.06	random median	0.00
CH1 L3 mode C	0.31	LCL mode	0.12	CJC mode	0.06	random mode	0.00
CH2 L3 mode	0.30	CH3 L2 median	0.11	LCL mode C	0.06		
CH2 L3 mode C	0.29	CH1 L2 mode	0.11	CH1 L2 median	0.04		

Table S7. MCC evaluation on synthetic nPSO networks for all the algorithmic variants. For each HEP algorithmic variant (link predictor, average operator and p-value correction), the table reports the mean MCC evaluated over all the nPSO networks.

Gene	Description	Ref#	Source	Concentration	Species
MFAP5	microfibrillar associated protein 5	Hs00969608_g1	ThermoFisher Scientific	20x	Human
TGFB2	transforming growth factor beta 2	Hs00234244_m1	ThermoFisher Scientific	20x	Human
MMP12	matrix metallopeptidase 12	Hs00899666_m1	ThermoFisher Scientific	20x	Human
EGLN2	egl-9 family hypoxia-inducible factor 2	Hs00363196_m1	ThermoFisher Scientific	20x	Human
FBLN5	fibulin 5	Hs01056640_m1	ThermoFisher Scientific	20x	Human
SFTPD	surfactant protein D	Hs01108490_m1	ThermoFisher Scientific	20x	Human
CHRNA	cholinergic receptor nicotinic alpha-1 subunit	Hs00909664_m1	ThermoFisher Scientific	20x	Human
TUFM	Tu translation elongation factor, mitochondrial	Hs00607042_gH	ThermoFisher Scientific	20x	Human

Table S8. PCR Primer set.