Interactome-scale comparison of co-immunoprecipitation and yeast two-hybrid assays for protein interaction prediction

Anonymous Authors[1]

Abstract

Protein-protein interactions (PPIs) are fundamental to biological processes, and computational prediction of PPIs is important for supplementing gaps in experimental data coverage. However, the quality of PPI predictions critically relies on the distributional characteristics of the training data. Here we investigate how data obtained from two widely-used experimental methods for detecting PPIs, Co-Immunoprecipitation (Co-IP) and Yeast Two-Hybrid (Y2H), differ in graph-theoretic and functional aspects. We document substantial differences between the two modalities, and find each assay type to be significantly more predictive than the other at specific function and network-associated tasks. We accordingly provide concrete recommendations on assay choice for a range of downstream tasks. Our work emphasizes the need for careful curation of PPI data based on the downstream task and underscores the importance of accounting for subtle but critical variations within biological training data.

1. Introduction

Protein-Protein Interaction (PPI) Networks are known to capture the systems-level biological mechanisms that govern complex cellular activities. However, a pervasive concern in the field is whether the experimental protocols used to obtain the PPI can introduce assay-specific biases, which may affect not only the properties of interactions identified but also the connectivity patterns within the networks. Such biases may portray a distorted understanding of the underlying biological mechanisms, imposing both experimental and computational challenges. Especially in the context of deep learning-based predictive PPI models, poor curation of PPI training datasets that fails to account for the underlying detection protocols and their impacts on network-wide structures/functions risks severely compromising the generalizability and predictive power of the model.

Focusing on the two main assays for PPI detection, co-immunoprecipitation (Co-IP) (Iqbal et al., 2018) and yeast two-hybrid (Y2H) (Striebinger et al., 2013), we present a rigorous comparative analysis of the coverage, functional informativeness, and relative advantages of the human networks derived from Co-IP and Y2H protocols (Figure 1). We broadly sought to test the hypothesis that wide-ranging differences between Co-IP and Y2H protocols contribute to significant variations in the interpretability and applicability of PPI data. We began by curating high-confidence experimental reports of physically-binding PPIs and then deployed a broad array of statistical and computational tests. Our tests aimed to address a few fundamental concepts: network-structure similarities and differences between the two protocols, explicit and implicit overlap between their interactomes, functional consequences of variations in their coverage and connectivity, and lastly, guidance on protocol selection and appropriate computational techniques.

Our analysis systematically confirmed the field folklore that Co-IP assays are more likely to produce clique-rich networks. Going beyond such folklore, our analysis offers novel insights into the network-structural, functional and cross-functional differences between the Co-IP and Y2H networks. We believe our findings are of broad relevance to computational modeling (specifically, task-specific dataset curation) and could also guide biologists about the assay to choose for their specific investigation.

2. Results

Dataset creation We extracted high-confidence connected human PPI networks for both Co-IP and Y2H assays from the IntAct Molecular Interaction Database (Orchard et al., 2013). After a series of preprocessing steps that included thresholding (to obtain high-confidence PPI edges) and largest connected component extraction (a common pre-processing method employed in many recent network-theoretic methods (Li et al., 2024), to ensure that many

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the ICML 2024 Workshop on Accessible and Efficient Foundation Models for Biological Discovery. Do not distribute.

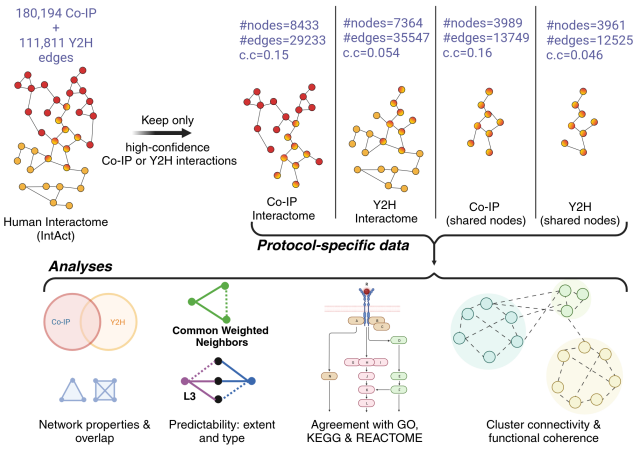

Figure 1. Schematic of different statistical analysis performed on the Co-IP and Y2H networks

Random Walk-based techniques that requires graph ergodicity as a prerequisite can be applied on the two networks), the finalized product comprised of a Co-IP network with 8433 nodes and 29,233 edges, and a Y2H network with 7364 nodes and 35,547 edges. We additionally constructed shared-node Co-IP and Y2H networks (which we called Co-IP-shared and Y2H-shared), containing only the proteins present in both the finalized networks. These networks consisted of 3989 nodes and 13749 edges for Co-IP-shared and 3961 nodes and 12525 edges for Y2H-shared (some of the nodes were removed to enforce network connectivity, causing slight deviations in the number of nodes between the two networks; Appendix A.1).

2.1. Graph-theoretic comparisons

We now describe the results from several graph-theoretic analysis we did on the Co-IP and Y2H networks.

Co-IP and Y2H networks have different local network properties. We found that the two graphs diverge significantly when considering clustering metrics, with the 3-clustering coefficient for the Co-IP network being approximately 2.75 times higher. This difference persists even in the shared networks, suggesting deeper differences in the underlying network structure. We tested this by checking if an unbiased sub-network sampling procedure causes significant divergence in the triangle-count between the two networks (Appendix A.2.1). We ensured that the resulting Co-IP and Y2H sub-networks sampled had fixed network sizes by controlling the number of nodes and edges. After conducting experiments on a broad range of node and edge combinations, we found that the triangle-counts from Co-IP sub-network samples were significantly larger, even after multiple testing correction. This shows that the two networks have substantial local differences, which is unlikely to have occurred through randomness.

Link prediction analysis shows that Co-IP and Y2H interactomes are connected in different patterns Link prediction, the task of identifying missing edges in a graph, can be repurposed as a test of graph structure: link prediction methods vary in their predictive scope (i.e., local vs. global), and the data-generating process they assume when predicting local links (Appendix A.2.2).

By checking which link prediction method works best for a given network, we can thus decipher its governing network mechanism. For our analysis, we selected two local link-prediction methods with very distinct guiding principles, CWN and L3, and a global method called GLIDE (Appendix A.2.2). CWN, which counts the number of common neighbors between two nodes, would be a good heuristic for finding new PPIs if the network is generated through multiple superpositions of cliques formed by protein complexes. This would however result in a worse performance for L3, as it scores edge likelihoods through a more complicated 'lock-and-key mechanism', drastically different from the clique-based mechanism favored by CWN. Therefore, identifying which method works the best for a given network can clarify the underlying generating principles observed by the network.

The link prediction results are provided in the Figure 2(A1). The results show that, while L3 outperformed CWN in both Co-IP and Y2H networks, it was especially true for Y2H edges: the ratio $\frac{AUPR(L3)}{AUPR(CWN)}$ is 1.82 (1.73) for the full (shared) Co-IP networks, but substantially higher at 3.31 (3.20) for the full (shared) Y2H networks. Biologically, this crucial finding allays the concern that Co-IP is optimized only for detecting protein complexes new cite: (Geva & Sharan, 2011). Had this been true, CWN would have outperformed L3 for Co-IP, based on CWN's governing principle (Appendix A.2.2). Notably, Y2H networks favor L3 over CWN even more, so L3 may still be the superior approach for identifying interactions spanning protein categories.

2.2. Functional associations in Co-IP and Y2H networks

To check if these graph-theoretic differences induce significant changes in the functional relationships implied by the networks, we conducted a broad range of Gene Ontology (GO) based tests to quantify the functional differences between the Co-IP and Y2H networks. We utilized all three GO hierarchies: biological process (BP), molecular function (MF), and cellular component (CC) (Appendix A.3.1).

The edges of the Co-IP network link more functionally-similar proteins than Y2H. To assess per-edge functional

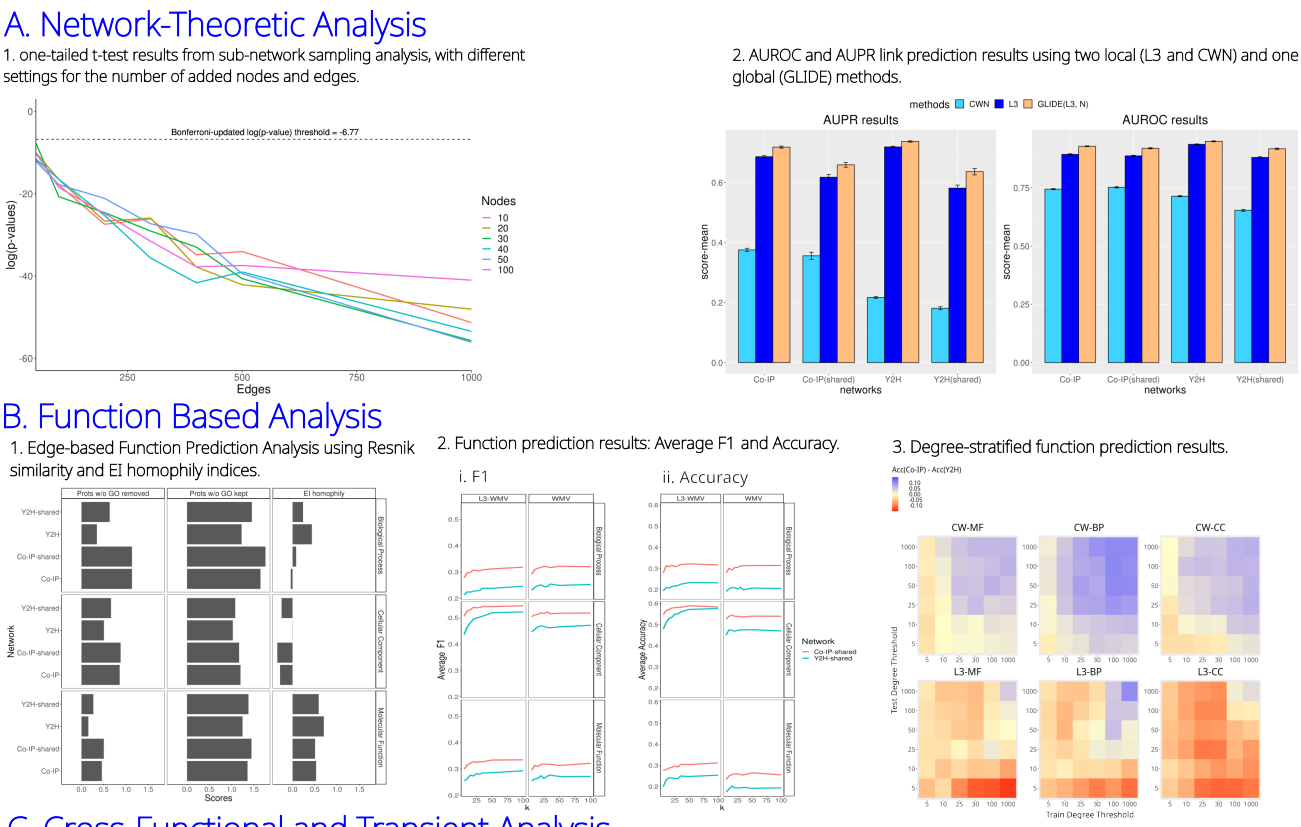

Figure 2. Network-based, function-based and cross-functional differences between Co-IP and Y2H networks

relationship between nodes in Co-IP and Y2H networks, we used two variations of the Resnik score and the EI homophily index (for Resnik score, higher is better; for EI index, lower is better; Appendix A.3.2). These results show that the edge-wide similarities are substantially higher in the full and shared Co-IP networks. For the shared networks in BP and MF classes, Co-IP-shared's Resnik scores were respectively 44% and 46% larger than Y2H-shared. The difference persisted, albeit weaker, even after we discarded nodes with no GO labels (17% and 5%, respectively). Co-IP edges are thus more functionally enriched and the local connectivity differences between the two networks are indeed reflected in the per-edge functional relationships.

Going beyond per-edge measures, we employed two local-neighborhood assessments, majority-vote based WMV and its L3-based variant, to predict GO labels for the Co-IP-shared and Y2H-shared networks (Appendix A.3.3). The results in Figure 2(B2{i-ii}) show that Co-IP-shared's top-1 accuracy and F1-max scores were noticeably larger for both WMV and L3-WMV based predictions across all thee GO branches. The gain was the strongest in GO:BP (32% F1-max improvement with WMV; 29% with L3-WMV), and the weakest in GO:CC (12% F1-max improvement with

WMV, 6% with L3-WMV).

Co-IP network's better function-prediction performance is largely due to the high-degree hub nodes. Both Co-IP and Y2H networks have a scale-free characteristic, contain a limited number of high-degree hubs. However, the biological relevance and validity of these hubs has been debated. To explore this, we repeated our WMV and L3-WMV analysis on subsets of proteins that satisfy some degree constraints. Our majority-vote prediction framework consists of a training and a test set, and we used two degree-based thresholds, d_{tr} and d_{te}, respectively, to parameterize our degree-based analysis. The resulting scores for every (d_{te}, d_{tr}) pairs are presented as a heat-map in Figure 2(B3). Higher d_{tr} and d_{te} consistently result in greater Co-IP-shared outperformance over Y2H-shared. Conversely, discarding the contributions from high-degree hub nodes resulted in Y2H-shared's L3 accuracy beating Co-IP-shared. This suggests strongly that the functional edge of Co-IP over Y2H is mainly due to the high degree hub proteins. That Y2H hubs are functionally less informative suggests they may be the consequence of promiscuous binding (Sontag et al., 2007).

2.3. Analysis of PPIs within functionally-related genesets

Categorization of proteins based on their functional properties and their affinity to disorders, relevant in both basic sciences and translational research, still suffers from the lack of experimental data with good coverage. To solve this, many network-based computational approaches have been developed for predicting missing proteins, and in some cases, discover entirely new functional groups. Our assumption is that the performance of the network-based module prediction/detection problems depend on the localized graph connectivity of the geneset in the graph, which can be measured by a graph-theoretic metric called graph conductance.

The graph conductance metric computes the probability that an edge originating from the set of nodes $\mathcal{S}$ terminates at a node outside the set ($V \setminus \mathcal{S}$, where V is the set of all nodes). Therefore, a large graph conductance would imply that the set is well-connected with the rest of the network. We used this conductance metric to analyze the cross-connectivities of KEGG and disease functional genesets.

Functional modules having low network conductance are more predictive. To check the effectiveness of conductance as a measure, we took 107 functional genesets obtained from KEGG (Kanehisa et al., 2007) and isolated 10 groups (5 with the highest conductance, the other 5 having the lowest conductance) for both Co-IP and Y2H shared networks. For these groups, we randomly removed 25% of their member proteins, and used a conductance-based approach described in Appendix A.4.1 to recover the removed members. Finally, we evaluated each functional group by measuring the percentage of removed proteins that were recovered by our approach.

The conductance plots are provided in Figure 2(C1). Applying our methods to both Co-IP and Y2H shared networks, we consistently found that for genesets with low network conductance, the ability to recover removed proteins is significantly higher than those with high conductance (Figure 2(C2)). On average, for Co-IP (and Y2H) networks, we were able to recover 74% (51% for Y2H) of removed proteins for the 5 genesets with low conductance. Compared to this, we were only able to recover 4.6 % (for Co-IP) and 1.3 % (for Y2H) of the removed proteins when the geneset conductance was very high.

Similarly, when we replaced KEGG with the Jensen diseasesets (Grissa et al., 2022), the average retrieval rate corresponding to the low conductance disease genesets were 70% and 33% for the Co-IP and Y2H networks respectively, significantly higher than the scores of high conductance genesets (1.3% for Co-IP and 10% for Y2H), which is in agreement with the KEGG results.

3. Discussion

We have systematically probed some long-held assumptions about differences between Co-IP and Y2H networks. We found that substantial differences in the network-theoretic and functional relationships do indeed exist between these interactomes. We found that the Co-IP networks have a considerably larger number of triangles than Y2H, manifesting in greater common weighted (CWN) link prediction accuracy. However, a simple replacement of CWN with the more biologically-meaningful L3 heuristic mitigates this disparity, indicating that the changes in predictability caused by the differences in triangle motifs frequency can be easily avoided by changing the analytical approach.

The local functional relationships of the two networks were also substantially different from each other. Here we recommend the use of Co-IP networks: it significantly outperformed Y2H consistently, across a broad range of function prediction tests. Furthermore, we observed that the improvement was especially strong in its high degree nodes. Thus, if a deep study of a specific functional system is desired and substantial experimental coverage (of Co-IP or Y2H) can be obtained, we recommend the choice of Co-IP over Y2H.

For module-based analysis, the choice between Co-IP and Y2H is largely dependent on a special graph-theoretic parameter called conductance. For analysis involving a particular functional group (i.e. finding missing members of that functional group), we found that selecting the network with the lowest graph conductance will produce the best performance, as outlined in Section 2.3.

References

Geva, G. and Sharan, R. Identification of protein complexes from co-immunoprecipitation data. *Bioinformatics*, 27 (1):111–117, January 2011.

Grissa, D., Junge, A., Oprea, T. I., and Jensen, L. J. Diseases 2.0: a weekly updated database of disease-gene associations from text mining and data integration. *Database (Oxford)*, 2022, March 2022.

Iqbal, H., Akins, D. R., and Kenedy, M. R. Co-immunoprecipitation for identifying Protein-Protein interactions in borrelia burgdorferi. *Methods Mol. Biol.*, 1690: 47–55, 2018.

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., and Yamanishi, Y. KEGG for linking genomes to life and the environment. *Nucleic Acids Research*, 36, 12 2007.

Li, M. M., Huang, Y., Sumathipala, M., Liang, M. Q., Valdeolivas, A., Ananthakrishnan, A. N., Liao, K., Marbach, D., and Zitnik, M. Contextual AI models for single-cell protein biology. *bioRxivorg*, June 2024.

Orchard, S., Ammari, M., et al. The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res*, 42(Database issue): D358–63, November 2013.

Sontag, D., Singh, R., and Berger, B. Probabilistic modeling of systematic errors in two-hybrid experiments. In *Biocomputing 2007*, pp. 445–457. World Scientific, 2007.

Striebinger, H., Koegl, M., and Bailer, S. M. A high-throughput yeast two-hybrid protocol to determine virus-host protein interactions. *Methods Mol. Biol.*, 1064:1–15, 2013.

A. Results

A.1. Dataset Construction

We extracted human PPI networks for both Co-IP and Y2H assays from the IntAct Molecular Interaction Database (Orchard et al., 2013), which contains 180,194 unique Co-IP and 111,811 Y2H interactions; each interaction was associated with a confidence score between 0 and 1. We focused on human PPIs since they have the widest coverage under both protocols. In order to only allow for high-confidence edges from both networks and to control for network size differences, we selected edges with score ≥ 0.56 and ≥ 0.35, for Co-IP and Y2H, respectively. The choice of these exact thresholds for Co-IP and Y2H networks were primarily motivated by our discovery that the IntAct scores were quantized with large spacing between two consecutive scores, and any deviation for our selected thresholds corresponded to a significantly larger deviation in the size of Co-IP and Y2H networks, leading to biased comparisons.

Since many graph-based algorithms (e.g., random walk with restart) implicitly assume that the input graph is a single connected component, we next ensured graph connectedness by limiting ourselves to the largest connected component of both networks, which resulted in the final Co-IP network having 8433 nodes and 29,233 edges, and Y2H network having 7364 nodes and 35,547 edges. There were 1754 shared edges (i.e. PPIs) between these networks. While this overlap is statistically significant, as more than 5% of edges in one network is also present in another, it is nonetheless fairly modest and indicates that incomplete coverage as well as measurement biases remain a concern. Therefore, we defined overlap more leniently, focusing just on the nodes (i.e., proteins) that are present in both Y2H and Co-IP networks. For this shared subset, we extracted the largest connected sub-graph in each network, naming these "Co-IP-shared" (with 3989 nodes and 13749 edges) and "Y2H-shared" (with 3961 nodes and 12525 edges). We emphasize that only the nodes are identical in these "shared" networks, while the edges are sourced from the respective complete networks.

A.2. Link Prediction Analysis

A.2.1. Sub-network sampling procedure to find local 3-clustering differences between Co-IP and Y2H networks

To quantify the significance of greater triadic closure in Co-IP networks, we sampled sub-graphs from the interactomes and counted the number of triangular motifs in each. However, network attributes like the overall number of nodes and edges in the samples could induce biases in the computed differences. To mitigate these size-related effects, we fixed the number of nodes and edges within the Co-IP (or Y2H) samples. We first extracted the common sub-graph (comprising all nodes and edges present in both Co-IP and Y2H), and for various settings of m (number of new nodes) and n (number of new edges), we added exactly m new nodes and n new edges from the original Co-IP (or Y2H) network to this common graph. This measure effectively neutralized the size-related biases in assessing network attributes. After generating 50 samples for each (m, n) pair, we performed a two-sample one-tailed t-test on the closed triangle counts of Co-IP and Y2H subgraphs. For all possible (m, n) configurations appraised, where $m \in \{10, 20, 30, 40, 50, 100\}$ and $n \in \{50, 100, 200, 300, 400, 500, 1000\}$, all the observed *p* values were found to be less than 10^{-10}. Even after the application of multiple testing correction (using the Bonferroni method), all (m, n) parameters yielded corrected p-values well below the 0.05 threshold. These findings support our assertion that the differences in the local network organization of Co-IP and Y2H networks, particularly as defined by the presence of triangular motifs, are unlikely to occur due to randomness.

A.2.2. Link prediction methods employed and their governing principles

We selected three link prediction methods that are commonly used and whose core concepts are seen also in other approaches. The first two, common weight normalized (CWN) and length-3 paths (L3), are localized techniques that predict an edge (p, q) based on the neighborhoods of nodes p and q. The CWN approach gives priority to edges between nodes, p and q, with numerous shared neighbors. Meanwhile, L3 employs a "lock-and-key" paradigm, emphasizing edge predictions that involve diverse protein categories. These approaches encapsulate distinct underlying mechanisms: CWN capitalizes on a "friends-of-friends" clustering model akin to triadic closure, while L3 focuses on interactions between "key" and "lock" protein categories. Notably, these lock and key categories are locally, not universally, applicable, i.e., there may be many types of keys as well as locks.

The third technique is a global link prediction method called GLIDE, which internally uses diffusion based distances (DSD) to characterize and evaluate prospective links. GLIDE scores are composed of both the global (informed by this DSD metric) and local components, where different choices of local heuristics can be selected based on the underlying network principles.

For the link prediction results in the main paper, we chose the GLIDE variant which used the un-normalized version of DSD as the global and the L3 score as the local method, as this combination performed the best in our evaluations.

We used both full and shared-node versions of the Co-IP and Y2H networks in our experiments, and created the test set for each network by randomly choosing 25% of its edges (while ensuring network connectivity) as our hold-out set. The predictive performance on this set was assessed using the Area Under the Precision Recall Curve (AUPRC) and Area Under the Receiver Operator Characteristics (AUROC) scores.

A.3. Function Prediction Analysis

A.3.1. Selecting functional labels

We used Gene Ontology (GO) information to assess the functional similarity between adjacent nodes in each network. We assessed all three GO domains: Molecular Function (MF), Biological Process (BP) and Cellular Component (CC). To remove overly broad labels, we selected only those at level 5 or deeper in the GO hierarchy; to remove overly specific labels, we required that each label annotate at least 50 proteins in the network.

A.3.2. Metrics for evaluating functional enrichment

Resnik similarity score: The hierarchical nature of GO as a Directed Acyclic Graph (DAG) allows distinct GO terms to be functionally related with each other based on the position of their shared ancestors. It is essential for an effective measure to exhibit the hierarchical nature of GO while evaluating protein functional similarities, which is why we used a popular function-based scoring method called Resnik similarity that preserves this property.

For a given GO-term ℓ, let $\mathcal{P}(\ell)$ represent the set of immediate ancestor of ℓ and let $\mathcal{A}(\ell)$ represent all its ancestors going back to the root of the GO hierarchy. The information content of ℓ (i.e. $i(\ell)$) is defined as

$$i(\ell) = -\log(Pr(\mathcal{A}(\ell))) = -\log \prod_{v \in \mathcal{A}(\ell)} Pr(v|\mathcal{P}(v)) \tag{1}$$

$$= -\sum_{v \in L} \log Pr(v|\mathcal{P}(v)) \tag{2}$$

Where the probability $Pr(v|\mathcal{P})$ denotes how likely it is to reach v from its immediate ancestors in $\mathcal{P}(v)$. Then, the Resnik similarity between two GO terms, x and y becomes:

$$res(x, y) = i(lca(x, y)) \tag{3}$$

Where $lca(x, y)$ is the lowest common ancestor of x and y.

As a protein can have more than one GO label, for proteins p and q, we compute the average of the pairwise Resnik scores between their GO labels as our final scoring metric $R(p, q)$; which can be written as:

$$R(p, q) = \frac{1}{|GO(x)||GO(y)|} \sum_{x \in GO(p), y \in GO(q)} res(x, q) \tag{4}$$

Where $GO(m)$ represents the set containing all the GO labels of a protein m.

EI homophily score: We additionally used the Resnik scores between protein neighborhoods to evaluate the degree of network homophily in a network. Network Homophily refers to the extent by which similar nodes (i.e. nodes with similar labels) form direct connections with each other, compared to the dissimilar nodes. We used a popular metric called the "EI homophily" to quantify this, which, given a node p and its neighbors N_p, can be described as:

$$EI(p) = \frac{|D_p| - |S_p|}{|N_p|} \tag{5}$$

Where D_p and S_p represents the set of dissimilar and similar neighbors of p respectively. Note that, the value of EI can range from -1 to 1 and the closer $EI(p)$ is to -1, the more similar p is to its neighborood proteins.

For each protein p, we generated S_p by selecting all neighbors with non-zero R-scores; the remaining neighbors were added to D_p. The scores for all the proteins were averaged to return a network-wide average EI-score.

A.3.3. FUNCTION PREDICTION METHODS

We evaluated the neighborhood functional enrichment of the shared Co-IP and Y2H networks through the use of two simple function prediction algorithms; (a) Weighted Majority Vote (WMV), and (b) L3-based Weighted Majority Vote (L3-WMV), both of which are described below:

Weighted Majority Vote (WMV) Let p be a protein in a weighted graph $G = (V, E, w)$ and let N_p denote the set of its neighbors. We can rank the proteins in N_p by their edge weights with p in a descending order, and select the top k proteins from the ranked list to produce a new set $N_{p,k} \subset N_p$, for a positive integer $k < |N_p|$ (if $k \geq |N_p|$, set $N_{p,k} = N_p$). Then, we can let each protein $q \in N_{p,k}$ to give a weighted vote on functional labels of p, where the weight is decided by the edge weight $w(p, q)$. The label associated with the highest vote is then selected as the predicted label of p.

L3-based Weighted Majority Vote (L3-WMV) Given a network $G = (V, E, w)$, and its corresponding weighted adjacency matrix A, we use the L3 pairwise scores to produce a new adjacency matrix A'. In fact, A' can be conveniently represented by the matrix product $A' = AD^{-1/2}AD^{-1/2}A$, where D is the diagonalized degree matrix of A. Then, we repeat the same procedure for WMV to predict functional labels, while replacing the original edge weights with the edge-weights from the newly computed A'.

We evaluated both algorithms in a 5-fold cross validation setting, where we randomly separated the network's proteins into 5 uniform non-overlapping blocks. For each fold, a particular block has its GO labels masked (i.e. "testing block") and the remaining blocks are used to predict the masked labels (i.e. "training block"). We repeated this process 5 times, with the k values set to $k = \{5, 10, 15, 20, 25, 30, 35, 50, 100\}$, and reported the average Top-1 Accuracy and F1-max scores for the shared networks, and the results are available in the main document (Figure 2(B2)).

A.4. Cross-functional analysis

A.4.1. CONDUCTANCE-BASED METHOD FOR PREDICTING PROTEINS RELATED TO A FUNCTIONAL GENESET

For a functionally related geneset S, we evaluated the predictive capacity of the network G by performing the following operation:

1. Let n be the number of proteins in S. Construct S_{remove} by randomly sampling 25% nodes from S. Let $S_{pos} = S \setminus S_{remove}$.
2. Additionally, generate S_{neg} by randomly choosing $10|S_{remove}|$ nodes that are not also present in S.
3. Initialize $S_c = S_{neg} \cup S_{remove}$, and $S_w = S_{pos}$.
4. For $\lfloor 0.25n \rfloor$ iterations, do
   (a) For each protein p_i in S_c, compute $c_i = Conductance(G, S_w \cup \{p_i\})$
   (b) Select p^* corresponding to the lowest conductance c_i. Update $S_w = S_w \cup \{p^*\}, S_c = S_c - \{p^*\}$.
5. Return retrieval ratio as $\frac{|S_w \cap S_{remove}|}{|S_{remove}|}$