Emergence of power-law distributions in protein-protein interaction networks through study bias

Marta Lucchetta[1], Markus List[2,†], David B. Blumenthal[3,†], and Martin H. Schaefer[1,†]

[1]Department of Experimental Oncology, IEO European Institute of Oncology IRCCS, Milan, Italy
[2]Chair of Experimental Bioinformatics, TUM School of Life Sciences, Technical University of Munich, Freising, Germany
[3]Biomedical Network Science Lab, Department Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
[†]Joint senior and corresponding authors: `markus.list@tum.de`, `david.b.blumenthal@fau.de`, `martin.schaefer@ieo.it`

Abstract

Protein-protein interaction (PPI) networks have been found to be power-law-distributed, i. e., in observed PPI networks, the fraction of nodes with degree k often follows a power-law (PL) distribution $k^{-\alpha}$. The emergence of this property is typically explained by evolutionary or functional considerations. However, the experimental procedures used to detect PPIs are known to be heavily affected by technical and study bias. For instance, proteins known to be involved in cancer are often heavily overstudied and proteins used as baits in large-scale experiments tend to have many false-positive interaction partners. This raises the question whether PL distributions in observed PPI networks could be explained by these biases alone. Here, we address this question using statistical analyses of the degree distributions of 1000s of observed PPI networks of controlled provenance as well as simulation studies. Our results indicate that study bias and technical bias can indeed largely explain the fact that observed PPI networks tend to be PL-distributed. This implies that it is problematic to derive hypotheses about the degree distribution and emergence of the true biological interactome from the PL distributions in observed PPI networks.

Introduction

Barabasi and Albert[1] proposed in the late 1990s that naturally occurring networks have a commonality: The distribution of their node degrees k (i. e., the number of interactions each node is participating in) tends to follow a power-law (PL) distribution $P(k) \propto k^{-\alpha}$. For $2 < \alpha < 3$, this distribution is scale-free, as its variance diverges with increasing network size. An important consequence of this assumed long-tail distribution of the node degrees is that it explains the existence of hub nodes with many connections (which are unlikely to occur under other statistical models), contrasting a large number of lowly connected nodes. Another feature of PL-distributed networks is their small world property, where a small network diameter leads to relatively high resilience against random perturbations[2]. This commonality in the topology of real-world networks is considered a universal law, as it seems to describe common features of such diverse networks such as food webs, metabolic networks, the internet, and protein-protein interaction (PPI) networks[3–5].

With respect to PPI networks, the PL property is typically explained with biological considerations: Protein families that are involved in general biological processes such as protein folding, gene regulation, or post-translational modifications are very promiscuous, binding to a large number of other proteins, whereas the majority of proteins show few interactions[6]. Moreover, it is crucial for the emergence of the PL property that, in the evolution of networks, "new vertices attach preferentially to sites that are already well connected"[1]. It has been suggested that, in the evolution of PPI networks, such preferential attachment can be explained via gene duplication and subsequent mutation[7].

The assumption that PPI networks show a PL distribution with hub proteins being characterized by distinctive biological features had important implications on the network biology field: Some studies use PL fittings as quality criteria for their measured networks[8]; others use topological protein properties ex- or implicitly for predicting disease genes[9,10]. Even more, a recent study demonstrated that the node degree is sufficient to predict disease genes even when ignoring the actual connectivity between proteins[11].

While it is thus plausible that PPI networks are PL-distributed, several studies have shown that, for many empirical PPI networks, this is not the case[12–14]. Moreover, in the absence of the ground truth interactome, it is difficult to

assess if the node degree distribution follows a PL due to the properties of the underlying biological system or if this is the result of technical or study bias. We know that biases of the experimental procedures used to infer networks can affect the resulting topology[15] and we expect that this problem in particular affects PPI networks, which are based on techniques with an estimated false positive rate of up to 80%[16]. PPIs are typically detected using yeast-2-hybrid (Y2H) studies, where individually selected protein pairs or libraries can be tested as bait and prey. Alternatively, affinity purification-mass spectrometry (AP-MS) allows one or several bait proteins to be tested against a large number of preys. In particular, the latter type of experiment can be expected to be sensitive to study bias, where already overstudied proteins such as oncogenes or tumor suppressors are tested more frequently than others[17]. Given a considerable false positive rate in detecting interactions, we can expect the degree of overstudied proteins to increase over multiple rounds of testing, possibly leading to the emergence of a PL distribution of the node degrees in the first place.

These considerations lead to the following question: Could it be that preferential attachment which leads to PL-distributed node degrees in aggregated PPI networks arises not because of biological mechanisms such as gene duplication but simply because of the fact that well-studied proteins are likely to be tested more often (Figure 1A)? Or in other words: Might it be the case that PL-distributed node degrees in PPI networks mainly describe our interest in proteins rather than the properties of the underlying biological interactome?

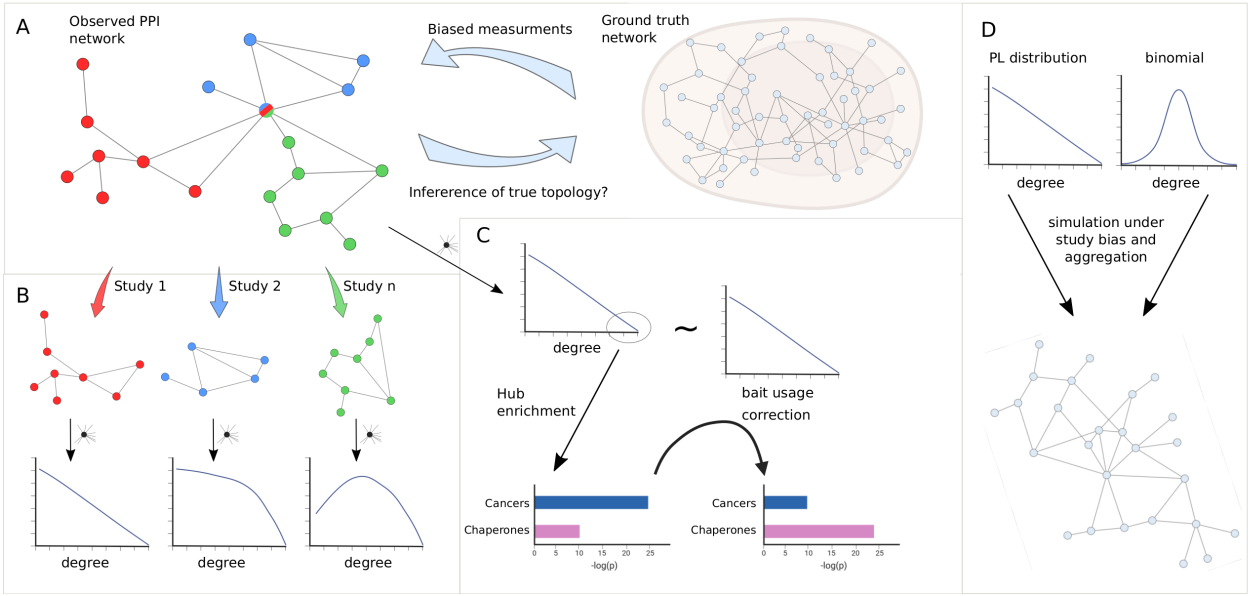

Figure 1. (A) We seek to answer the question of how much we can learn about the topology of ground truth networks from the topology of observed and aggregated PPI networks and how much biased measurements might impact the observed PL degree distribution. (B) To answer this question we decompose aggregated, observed networks into single study-networks and investigate their individual degree distributions. We then ask how much the aggregation process of those single studies into larger networks could explain the PL property of the observed network. (C) We aim to identify true hub proteins by applying different types of normalization strategies, which reveals that disease-associated functions disappear that are likely associated with hub proteins because of their inflated testing frequency due to the study bias. (D) Finally, we simulated the measurement of observed aggregated PPI networks under study bias from ground truth networks with either PL or binomial degree distribution.

In this study, we seek to answer this question. Observing that only a subset of networks show a node degree distribution following a PL, we systematically tested if the PL property arises simply by aggregating studies (Figure 1B), as is common practice in PPI databases. Next, we tested if the node degree distribution still follows a PL if we account for the bias introduced by bait proteins. Further, we test to which extent accounting for such biases changes the functional enrichment of highly promiscuous hub proteins, where we expect that heavily studied disease-related proteins show reduced enrichment whereas functions carried out by proteins known to be promiscuous should show increased enrichment (Figure 1C). Finally, we simulate the measurement process of observed PPI networks under study bias for different false negative and false positive rates, given hypothetical PL-distributed and binomially distributed ground truth interactomes (Figure 1D). Using Bayesian inference, we then quantify to which extent the degree distributions of

observed PPI networks allow to derive conclusions about the topology of the ground truth interactome. Overall, our results indicate that technical and study bias can indeed largely explain the fact that observed PPI networks tend to be PL-distributed. This implies that it is problematic to derive hypotheses about the degree distribution and emergence of the true biological interactome from the fact that node degrees in observed PPI networks tend to be PL-distributed.

Results

Less than one in three study-specific protein-protein interaction networks are power-law-distributed

Mosca et al.[18] recently showed that aggregated observed PPI networks generally show a node degree distribution following the PL. To confirm this, we aggregated a large human PPI network consisting of 41,862 studies and a total of 471,693 unique interactions among 17,865 proteins. We tested if the resulting degree distribution follows a PL by quantifying the plausibility of a goodness-of-fit test as described before[19]. We observed that the resulting degree distribution can be approximated by a PL distribution ($P = 0.35$; where $P \geq 0.1$ is by convention[19] indicative of a PL distribution; Figure 2A).

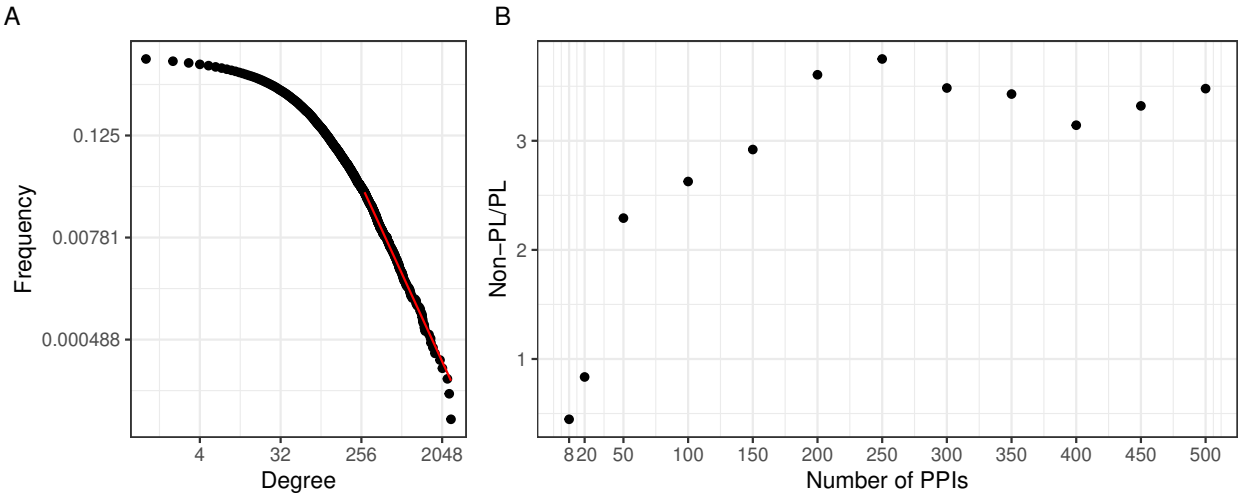

Figure 2. (A) The black dots represent the degree distribution of our aggregated network and the red line corresponds to the fitted PL distribution with parameters $k_{min} = 278$ and $\alpha = 3.3$ in a log-log scale. (B) Plot of the ratio between the number of non-PL and PL studies with more than a certain number of PPIs specified in the x-axis.

An interesting question is if the PL property is inherent to single PPI networks or if it possibly arises through the aggregation process. To investigate this, we next tested for the PL property of the constituting single studies. We observed that when considering networks of size 200 or larger, there were approximately 3.5 times as many non-PL-distributed networks as compared to PL-distributed networks (Figure 2B). The ratio reduces to 1 when also small networks were considered. However, we reasoned that this is likely an artefact of the relatively poor fit of the degree distribution for small networks: The majority of networks has a small size (Supplementary Figure 1) and those small networks that are not filtered out (see Methods), are typically classified as PL-distributed. Eg. 84% of the 739 single-study networks with at most 20 PPIs (a network size which we consider unlikely to lead to reasonable degree distribution fits) are classified as PL-distributed. This suggests that for network sizes where it is possible to reliably fit degree distributions, the non-PL networks largely outnumber PL networks.

The power-law property is associated with bait usage

We next tested systematically if the PL property of networks can emerge from aggregating non-PL networks. We therefore randomly merged non-PL networks (1,000 times 100, 200, and 300 non-PL studies). As shown in Figure 3A, we obtained more than 50% of PL aggregated networks after the aggregation of non-PL studies. In particular, the more

studies we merged, the greater the fraction of PL studies (from 53% to 83%), demonstrating that the PL property can emerge from the aggregation of non-PL networks.

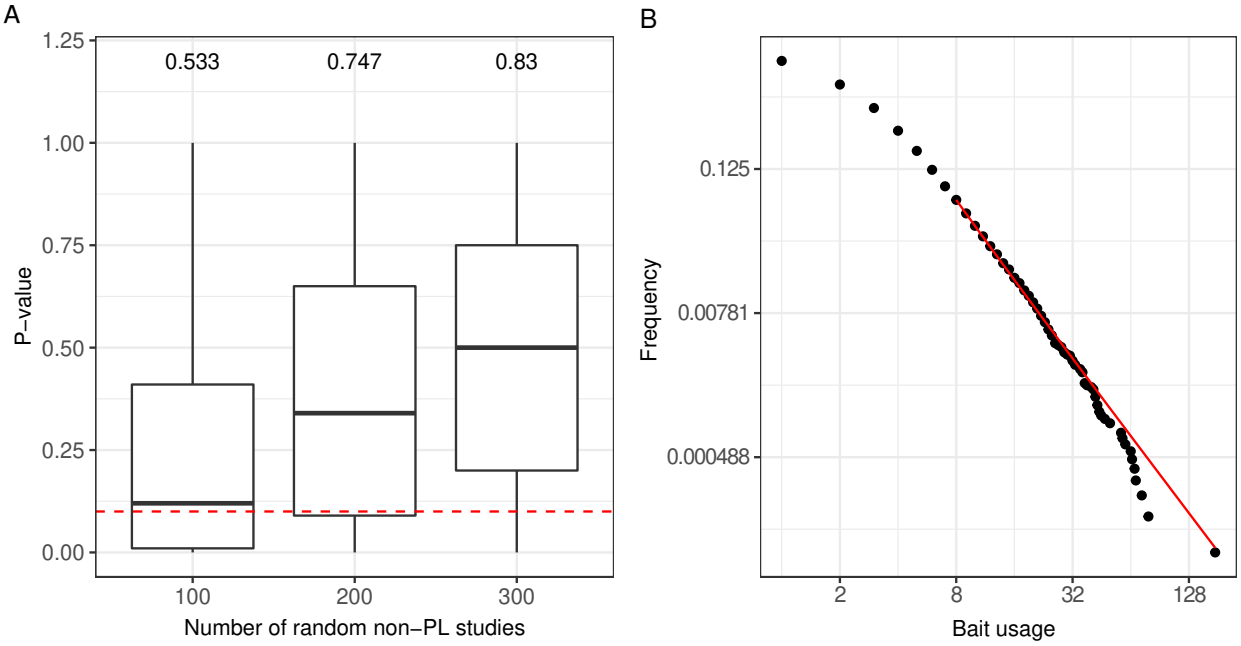

Figure 3. (A) Distribution of P-values obtained through the aggregation of 100, 200, and 300 random non-PL studies. The numbers on the top of each boxplot represent the fraction of PL networks obtained among the 1,000 tests. The dotted red line represents the limit of significance (i. e. 0.1); above the line, the PL hypothesis is plausible. (B) The black points correspond to the bait usage distribution and the red line corresponds to the fitted PL distribution (in a log-log scale) with parameters $k_{min} = 8$ and $\alpha = 3.13$.

We observed a good correlation between the number of times a protein has been tested for interaction partners and its degree ($r = 0.57$, $P < 10^{-16}$; Pearson correlation test) in agreement with what has been previously described[17]. We thus wondered if the PL property of the merged network could have been inherited from a potential bias in the number of times proteins have been tested for interaction partners. Indeed, we observed that the bait usage distribution follows a PL ($P = 0.34$; Figure 3B). To test if the bait usage distribution could impact the observed degree distribution, we randomly subsampled networks for which we have bait information 3,000 times (1,000 times 50, 100, and 150 non-PL studies). For each resulting aggregated network, we fitted PL distributions to both the bait usage distribution and the degree distribution. We observed a significant association between finding that one or the other distribution would follow a PL distribution ($P = 4 \cdot 10^{-2}$; one-sided Fisher's exact test).

The power-law property often vanishes when correcting for bait usage

We focused on the 27 single-study networks with PL distribution that consisted of more than 200 PPIs (as our initial analysis suggested that the ratio between non-PL and PL studies converges from this value) and for which we had bait and prey information. We observed that the ratio between baits and preys used varied largely across those studies (Figure 4B) resulting in some studies with a symmetric design (i. e. relatively similar number of baits and preys) and some with rather asymmetric design (i. e. big difference in the number of baits and preys). We hypothesized that strongly uneven bait vs. prey usage might contribute to the PL property by inflating the degree of a few proteins, effectively favouring them to become hub proteins. To test this hypothesis, we attempted to correct this bias with the expectation that we could turn PL networks into non-PL networks. To this end, we recomputed the degree distribution by only considering the number of interactions formed by the larger set (either baits or preys, see Methods for details and Figure 4A for a graphical visualization).

We observed that in 9 out of the 27 cases, we turn PL degree distributions into non-PL degree distributions by applying this correction. We observed that symmetry scores for networks that changed from PL to non-PL were

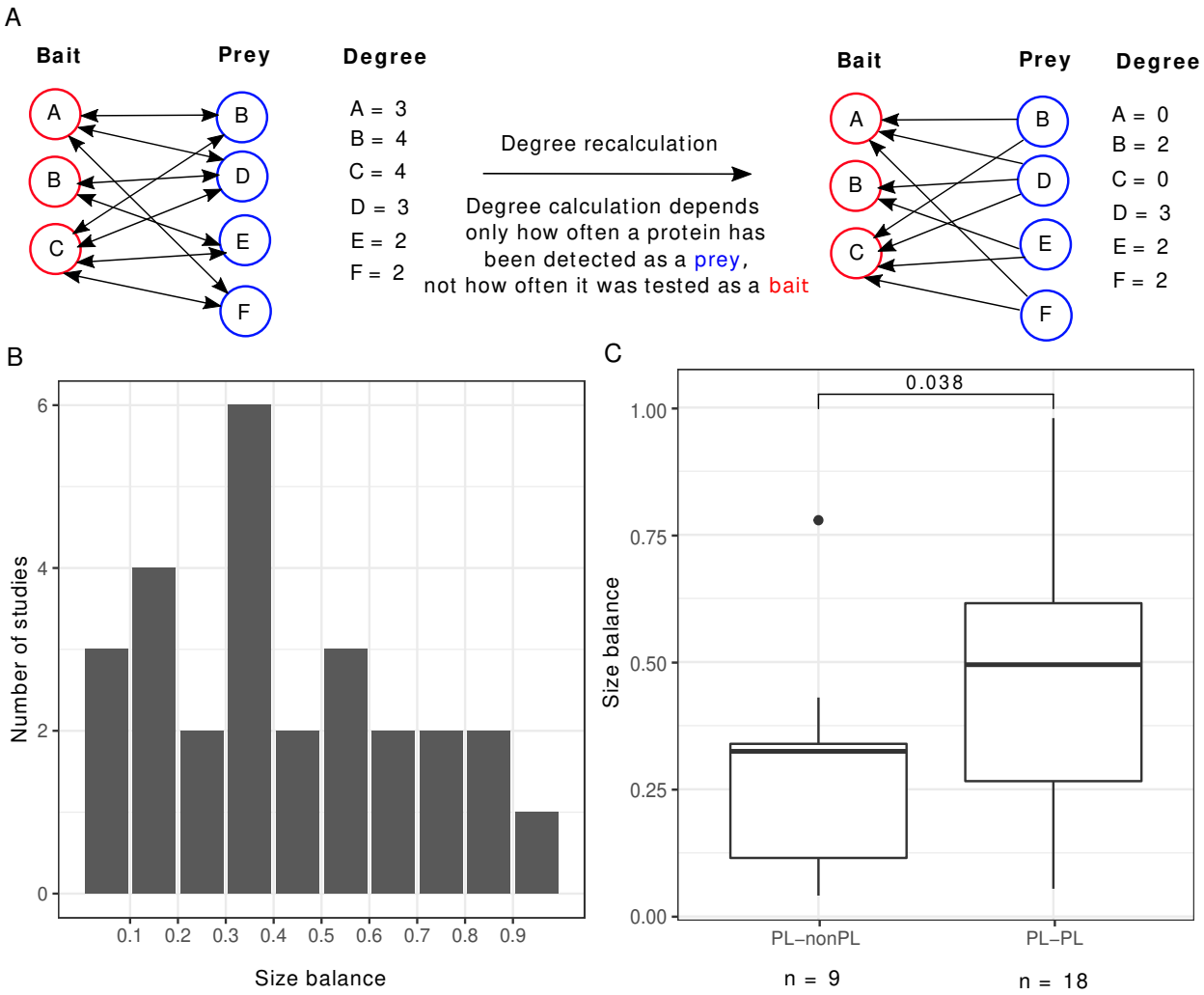

Figure 4. (A) Scheme to illustrate how the degree is recalculated when the number of baits is smaller than the number of preys. (B) Distribution of the size balance (ratio between the number of baits and preys, see eq. (2) for details) among the 27 PL studies. (C) Distribution of the size balance in the 9 studies that switch from PL to non-PL and the 18 studies whose degree distributions remain PL-distributed after the correction.

significantly smaller ($P = 0.038$, one-sided Wilcoxon test; Figure 4C), demonstrating that the bait to prey ratio has a considerable influence on the PL property.

Accounting for study bias reveals functionally meaningful hub proteins

We hypothesize that the degree distribution of proteins might be inflated by study bias, i. e., proteins with a high degree in the aggregated PPI networks might not be proteins with a higher number of interactions but might merely be tested more often due to their relevance in disease or other assumed importance in cellular systems. We hence asked if we can reveal the true identity of hub proteins. To do so, we came up with three strategies:

- We computed the degree using only interactions formed by preys in AP-MS studies and identified those with the largest degree (similar as in the previous section and visualized in Figure 4A; prey hubs).
- We normalized the degree in our initial aggregated network by the number of times the proteins have been used as bait and identified the proteins with the highest normalized degree (normalized hubs).

- We computed the degree distribution within one single study (HuRI[20]) that aims to provide a study-bias-free, near-proteome-scale map of the human interactome. We refer to the proteins with the highest degree in this network as Y2H hubs.

We then tested for functional (Gene Ontology) and disease gene (Disease Ontology) enrichment. Interestingly, we observed that the prey hubs are most strongly enriched for "protein folding" and "chaperone-mediated protein folding" (Figure 5A, Supplementary Table 1). The majority of genes in these categories are chaperones whose function is to mediate protein folding. Since the majority of human proteins requires assistance in folding by chaperones[21], they might indeed be true hub proteins. As chaperones compose 10% of the cellular proteome mass in humans[22], we were concerned that the enrichment of chaperones among the prey hubs could be an artifact of a detection bias of AP-MS towards highly abundant proteins. To rule out this possibility, we retrieved MS quantifications of human proteins in different tissues[23] and performed a Gene Ontology enrichment analysis on sets of the most abundant protein (of different sizes). None had chaperones among the top enriched terms (Supplementary Figure 3) suggesting that the enrichment among prey hubs was not simply an artifact of protein abundance. Similarly, pathway enrichment analysis (Supplementary Figure 2) showed protein folding among the top enriched pathways.

The disease gene enrichment analysis confirms the previous observation that uncorrected hubs are associated with many different types of diseases (Supplementary table 1), in particular with cancer (Figure 5B). Prey hubs exhibit an enrichment of diseases related to the nervous system (though much weaker as compared to the enrichment of cancer among the uncorrected hubs). In contrast, normalized and Y2H hubs do not show any significant enrichment in diseases, challenging the idea that disease genes *per se* have a higher connectivity in PPI networks.

We were surprised to find several nervous system diseases enriched among the prey hubs. Many of the prey hubs related to nervous system diseases were in fact chaperones. To test if the enrichment of nervous system diseases was caused by the chaperones, we retrieved the proteins of the most strongly enriched disease classes (schizophrenia and psychotic disorder) and tested if chaperones were enriched among those proteins (Supplementary Figure 4). Indeed, we found a significant enrichment ($P < 0.05$, one-sided Fisher test) in both cases, suggesting that chaperones might cause the observed disease enrichment among true hubs towards nervous system diseases. This is likely because protein misfolding is a hallmark of many nervous system diseases[24,25] and indeed chaperones play a role in prevention of misfolding.

The power-law property is consistent with a binomially distributed ground truth interactome

The descriptive findings summarized in the previous paragraphs indicate that the PL property in aggregated PPI networks might be due to biases in the PPI measurement process instead of reflecting the topology of the ground truth interactome. To further assess this possibility, we simulated the measurement of observed aggregated PPI networks under study bias (see Methods for details). We parameterized our simulator with four hyper-parameters: The test method (AP-MS or Y2H testing), the false positive and false negative rates of the test method, and the acceptance threshold $\gamma \in [0,1]$ (our simulator includes a PPI (u,v) into the simulated aggregated network if it has been detected at least once and the fraction of positive simulated experiments that test for (u,v) exceeds γ).

For each hyper-parameter setup, we simulated 50 hypothetical ground truth networks generated with the Barabasi-Albert (BA) model[1] and 50 hypothetical ground truth networks generated with the Erdős-Rényi (ER) model[26]. In BA networks, node degrees are PL-distributed; in ER networks, they follow a binomial distribution. Subsequently, we simulated the detection of the PPIs in the hypothetical ground truth networks via aggregated PPI testing under study bias. We then computed earth mover's distances between the degree distribution of an empirical aggregated PPI network G_{IntAct} obtained from IntAct and the degree distributions of the simulated networks. Using these distances, we computed signed relative sum of differences ΔSOD (see eq. (8) for details) to quantify whether the degree distribution of G_{IntAct} is more similar to the degree distributions of simulated networks that emerged from PL-distributed or from binomially distributed ground truth interacomes. Moreover, we estimated the posterior probabilities of G_{IntAct} belonging to the two classes of networks, using K-nearest neighbors (K-NN) classification (see eqs. (9) and (10) for details).

The results of our simulation studies for AP-MS testing are shown in Figure 6. When comparing the empirical network G_{IntAct} to hypothetical PL-distributed and binomially distributed ground truth networks, we observe that G_{IntAct}'s degree distribution is much more similar to the degree distributions of the PL-distributed networks (Figure 6A). This is not surprising, given that G_{IntAct} is itself PL-distributed (Supplementary Figure 5A). However, the picture changes when looking at the sums of distances between G_{IntAct} and the simulated aggregated networks: Already for very small false positive rates, the gain in similarity between G_{IntAct} and networks emerging from PL-distributed ground truth networks vanishes. For $\gamma = 0$ (each PPI detected by at least one simulated study is included in the aggregated

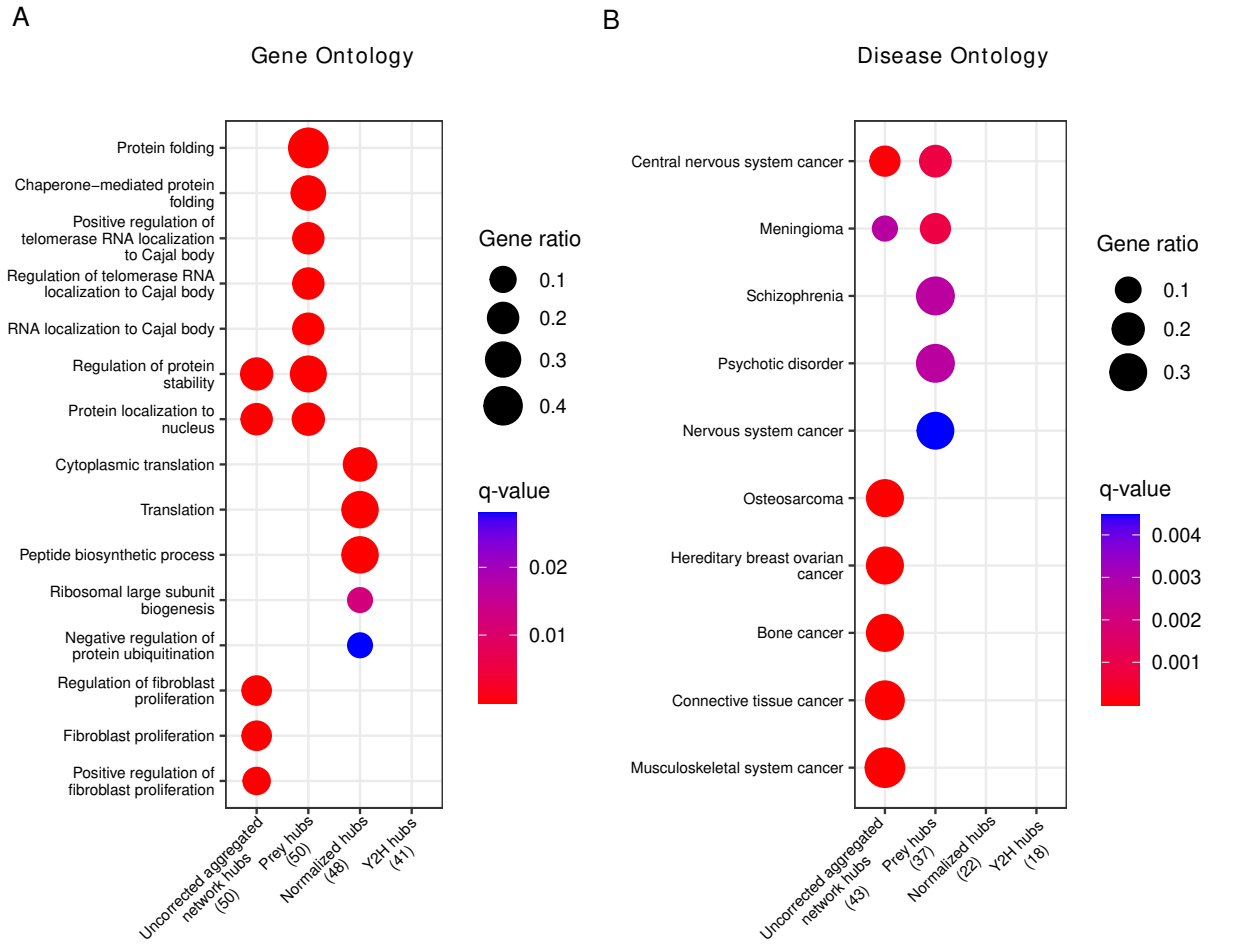

Figure 5. (A) Gene ontology enrichment analysis of the top 50 corrected (prey hubs, normalized hubs, and Y2H hubs) and non-corrected hubs (uncorrected aggregated network hubs). (B) Disease ontology enrichment analysis of the top 50 corrected and non-corrected hubs. The numbers in brackets represent the number of hubs included in the reference databases, and the "Gene ratio" represents the fraction of hubs included in the corresponding (gene or disease ontology) term. If a column is empty, it means there are no significant terms.

network), the tipping point lies between $FPR = 0.00625$ and $FPR = 0.0125$ (Figure 6B); for $\gamma = 0.5$ (a PPI is included in the aggregated network only if it is detected by the majority of the simulated studies that test for it), it lies between $FPR = 0.0125$ and $FPR = 0.025$ (Figure 6C). By increasing γ and keeping only consensus PPIs in our simulated networks, we can hence slightly improve the robustness of our simulated PPI network measurement process w. r. t. the false positive rate of the PPI detection method.

Estimated posterior probabilities for G_{IntAct} having emerged from a PL-distributed or a binomially distributed ground truth interactome for false positive rates just below and just above the tipping points are shown in Figure 6D to Figure 6G: For false positive rates below the tipping points, a PL-distributed ground truth interactome is clearly the more likely origin of G_{IntAct}, independently of the parameter K used for the K-NN-based estimation of the posteriors. For false positive rates above the tipping points and $K \geq 18$, binomially distributed and PL-distributed interactomes are roughly equally probable origins of G_{IntAct}. With smaller K, the estimated probabilities are actually larger for binomially distributed ground truth networks.

If the estimated AP-MS false positive rates of 10% to 40 %[27] are only remotely realistic, they clearly exceed the tipping points between 0.625% and 2.5% uncovered by our simulation study. The results summarized above hence indicate that the observed PL behavior of empirical PPI networks obtained via aggregation of AP-MS studies tells us

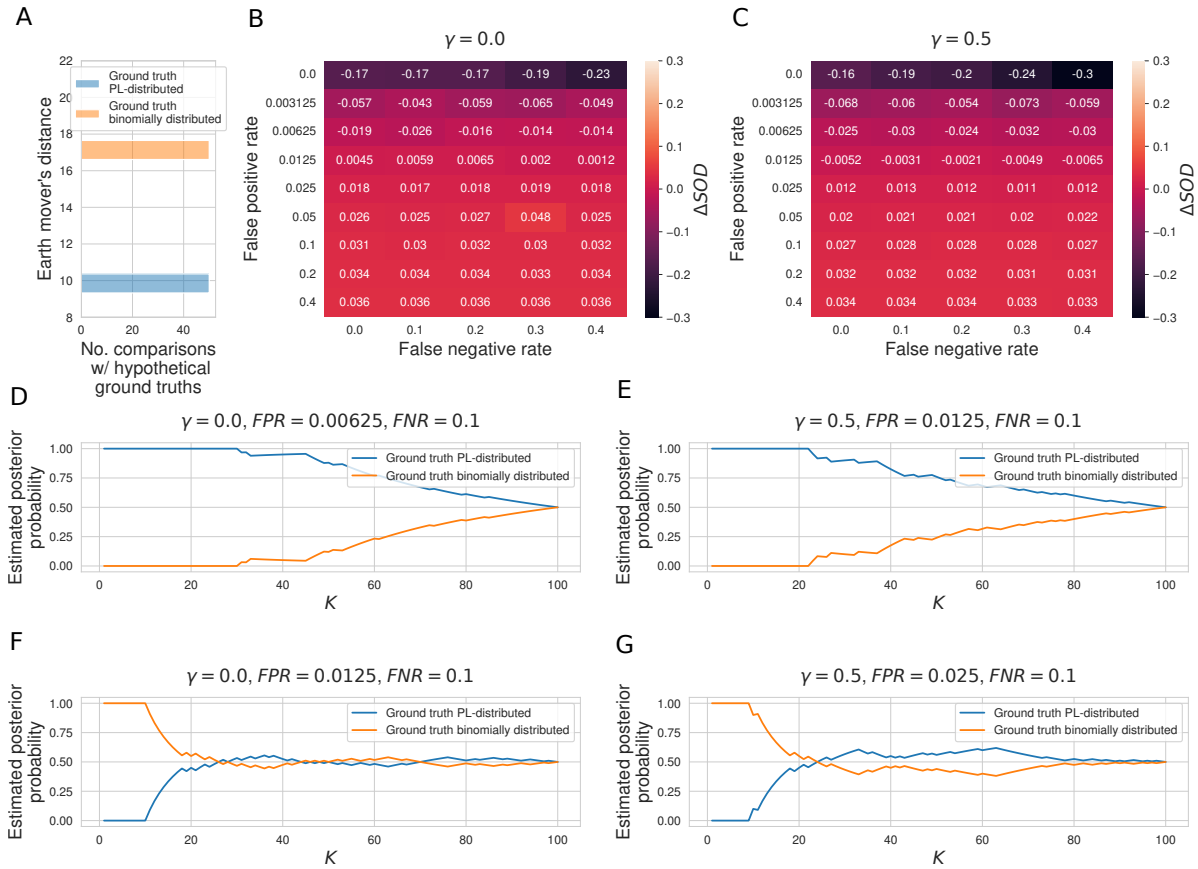

Figure 6. (A) Histogram of earth mover's distances between the degree distribution of the observed PPI network G_{IntAct} obtained via aggregation of all AP-MS studies annotated in IntAct and the degree distributions of 50 PL-distributed and 50 binomially distributed hypothetical ground truth networks. (B, C) Signed relative differences $\Delta SOD \in [-1, 1]$ between sum of distances between degree distribution of G_{IntAct} and degree distributions of networks simulated from, respectively, PL-distributed and binomially distributed hypothetical ground truth networks, given different choices of the hyper-parameters *FPR*, *FNR*, and γ. Negative values of ΔSOD indicate that G_{IntAct} is more similar to simulated networks emerging from PL-distributed hypothetical ground truths; positive values are indicative of the opposite scenario. (D–G) Posterior probabilities that G_{IntAct} emerged from a PL-distributed or from a binomially distributed ground truth interactome, estimated via K-NN classification. (D, E) Estimated posterior probabilities just before the tipping points in the false positive rate. (F, G) Estimated posterior probabilities just after the tipping points.

very little about the topology of the ground truth interactome and is even compatible with binomially distributed node degrees in the ground truth interactome.

The results for simulated Y2H testing (Supplementary Figure 6) are very similar to the ones for AP-MS testing. The only difference is that we observe even smaller tipping points. A likely explanation for this is that, unlike the PPI network obtained by aggregating all AP-MS studies annotated in IntAct, the PPI network obtained by aggregating all Y2H studies is itself not PL-distributed (Supplementary Figure 5B). For both simulated AP-MS and simulated Y2H testing, the false negative rate did not have a strong effect on the results (see small row-wise variances in the heatmaps shown in Figure 6 and Supplementary Figure 6).

Discussion

It is widely believed that the PL behavior of PPI networks arose through evolution, where frequent gene duplication events have led to protein copies that retain the original interaction partners. As one can mathematically prove, such a model eventually leads to a scale-free network[28]. Recently, doubts have emerged that PPI networks are truly scale-free[14].

Furthermore, it was shown that active module discovery methods perform equally well on real and random networks in which the node degree is preserved[11]. Such methods, which are typically applied to PPI networks to extract disease modules in the form of subnetworks, thus do not benefit from the interactions of the network but merely learn from the node degree, suggesting that study bias may be driving these analyses.

Here, we offer an alternative explanation, suggesting that the PL behavior of PPI networks emerges through a combination of biases. Firstly, we show that typically used experimental designs where we find an asymmetry between bait and prey proteins contributes to the PL property. Secondly, we find that the current practice of aggregating study-based PPI networks tends to introduce a PL behavior of the node degree distribution that is not found in the individual studies. We suspect that aggregating studies emphasizes study bias, i. e., the over-representation of proteins used as baits in such experiments. We show that removing such biases leads to the emergence of alternative hub proteins that drive the network. Thirdly, we show through simulation that, already for very small false positive rates, binomially distributed ground truth networks generated with the ER model are equally likely origins for aggregated observed PPI networks as PL-distributed ground truth networks generated with the BA model. This finding is robust across different parameters for the false positive, the false negative, and the study acceptance rate.

It is important to note that the design of our simulation study is based on two major simplifications: Firstly, we only consider ER and BA networks as possible models for the ground truth interactome, although most likely none of the two models fully captures the topology of the unknown true biological interactome. Here, this simplification serves as a conceptual framework which allows us to address the question on the origin of the PL behavior of observed PPI networks from a Bayesian perspective. Secondly, our simulator assumes that study bias affects the emergence of aggregated observed PPI networks via a direct feedback loop (high-degree proteins are preferentially sampled for experimental testing). In reality, the feedback loop is much less direct: Study bias in the emergence of aggregated PPI networks is not only mediated via studies reporting PPIs but also and primarily via more indirect pathways such as over-representation of genes encoding highly studied proteins in gene annotation databases[29]. It is hence likely that real-world repeated PPI testing is slightly less sensitive to the experimental false positive rates than suggested by our simulations. However, in view of the huge margin between the uncovered tipping points (0.625% to 2.5%) and the estimated false positive rates in AP-MS and Y2H testing (10% to 40%[27]), our conclusion remains valid that the topologies of observed PPI networks have little inferential value w. r. t. the unknown ground truth interactome.

We point out several strategies that could help to reduce biases of PPI networks. Employing techniques that can detect PPIs with an FDR as low as 1%[30] would considerably reduce the technical bias in detecting PPIs. Our study suggests a tipping point in the *FPR* at which study bias can no longer be tolerated, but this may in reality be higher or lower than what we anticipate here. It is thus not clear how robust techniques need to be to entirely avoid that study bias distorts the topology of observed PPI networks. However, even the use of an error-free technique would not mitigate pre-existing study bias, which is not just ingrained into existing PPI networks but also indirectly influences the choice of bait and prey proteins used in future studies. An alternative strategy is thus to systematically and objectively study PPIs without prior evidence for the relevance of a protein, a strategy currently followed by the HuRI project[20]. Our results indicate that also the aggregation of non-PL studies tends to lead to networks with PL property, possibly because study bias present in individual studies is magnified in this process. In view of this, an interesting question for the future will be if the aggregation of study-bias-free studies such as HuRI will still favor the emergence of the PL property.

Finally, there are also cost-effective ways to assess and address biases in PPI networks. For instance, we could show that the problem of study bias can be partially mitigated by relying on the information of prey proteins alone. An interesting observation we made was that accounting for this bias revealed a different set of hub proteins enriched for protein folding rather than disease genes. Further work will be needed to establish if true hub proteins exist in the PPI network and what their role is. Importantly, we encourage the field to report negative interactions, since these could be used to define a reasonable study acceptance rate (ratio of positive and negative interactions, λ in our simulation) to limit the distorting effect of the *FPR*, whereas, in the current practice, even unique false positive interactions may be added to the aggregated PPI network.

In conclusion, our analysis supports the alternative hypothesis that the PL behavior observed in aggregated observed PPI networks cannot be treated *per se* as biologically motivated as the gene duplication model suggests. We face the issue that we currently have no means to reliably disentangle study and experimental bias in the node degree distribution. Our attempts to remove this bias led to differing results depending on the type of normalization we used. In all three cases, disease-associated proteins were demoted. Only the prey hub normalization revealed a significant functional enrichment where proteins such as chaperones that are involved in protein folding have been significantly enriched. While these results seem plausible, we cannot prove that this normalization indeed corrects for all conceivable forms of bias. Our results hence suggest that further work is needed to either perform additional studies that avoid known

sources of bias or to develop a robust normalization that removes known biases from existing networks.

Methods

Analyzed protein-protein interaction networks

We retrieved human PPIs from IntAct[31](version of 2022-02-03 on `https://ftp.ebi.ac.uk/pub/databases/intact/2022-02-03/psimitab/`) and HIPPIE[32] (version 2.2 on `http://cbdm-01.zdv.uni-mainz.de/~mschaefer/hippie/download.php`). Most of the interactions in IntAct are annotated with the information of which protein within the pair was used as a bait and which as a prey during the experimental determination of the interaction. To increase the total number of studies, we expanded the IntAct interactions by merging with HIPPIE. After that, we downloaded a list of all 20,401 human proteins (only reviewed entries from Swiss-Prot / Uniprot) from `https://www.uniprot.org/uniprotkb?facets=model_organism%3A9606&query=reviewed%3Atrue` (version from December 13, 2022). We kept only interactions where both the proteins are in this list resulting in a network consisting of 471,693 interactions and 17,865 proteins detected by 41,862 studies.

Testing the power-law property of empirical distributions

In order to test if sequences of the proteins' node degrees or bait usages (numbers of times the proteins have been tested as a bait) are PL-distributed, we used the poweRlaw R package[33] (version 0.70.6). The package implements methods proposed by Clauset *et al*[19]. It estimates the best-fitting PL distribution to the data of the form

$$p(k) \propto k^{-\alpha}, \tag{1}$$

where $\alpha > 1$ is the scaling exponent, $k \geq k_{min}$ is the degree or the bait usage sequence, and $k_{min} \geq 1$ is the cutoff above which the PL distribution is fit to the data. The package estimates the k_{min} via a minimization of the Kolmogorov–Smirnov (KS) statistic and uses a maximum likelihood estimator to choose α. Subsequently, it carries out a goodness-of-fit test between the empirical data and the fitted PL model. Here, the KS statistic between the fitted model and the empirical distributions is compared to KS statistics between the fitted model and synthetic distributions sampled from the fitted model. Then, a P-value can be computed as the fraction of distances between the fitted model and the synthetic distributions that exceed the distance between the fitted model and the empirical distribution. Following the convention introduced by Clauset *et al*[19], we consider the PL distribution a plausible model for the empirical data if the P-value of the goodness-of-fit test exceeds 0.1. In the poweRlaw R package, the P-value can be computed with the `bootstrap_p` function, which we ran with 100 (default parameter) bootstrap simulations.

We tested the PL property for each single study included in our aggregated network. We discarded studies where the used method failed to estimate the k_{min} and hence could not test the PL hypothesis. Moreover, we filtered out studies for which more than 10% of the 100 bootstrapping simulations failed to produce meaningful results (as pointed out in the documentation of the poweRlaw package, this can occasionally happen if all values in the synthetically sampled distribution are below k_{min}). We applied those exclusion criteria to any analysis that required a PL computation. After these two filtering steps, the remaining studies are 1,427 in total, of which 986 are PL-distributed ($P \geq 0.1$, goodness-of-fit test).

Aggregation of study-specific protein-protein interaction networks

In order to investigate if the PL property arises through the aggregation process, we randomly aggregated 100, 200, and 300 non-PL studies (of 441 in total) 1,000 times and we tested the PL hypothesis of the degree distribution after the aggregation. We used a similar randomization strategy to test if there is an association between the degree and the bait usage distribution: We considered only non-PL studies with bait annotations (184 in total) and we randomly merged 50, 100 and 150 studies 1,000 times. For each aggregated network, we tested the PL property of the degree and bait usage distribution. We used the one-sided Fisher's exact test to analyse any significant association between the two distributions.

Computing the degree distributions based on baits or preys only

To assess if the asymmetry in experimental design (i. e., number of baits and preys) affects the PL property, we focused on the 27 single-study networks with PL distribution, having more than 200 interactions (we removed one study with less than 10 bait-prey-annotated interactions) and for which we had bait and prey information. For each of them, we recalculated the degree distribution as follows: If, in the study under consideration, the number of baits is smaller than the number of preys, we only counted those interactions (u,v) for the degree of u where u was tested as a prey. Like this, the degree of a protein only depends on interactions where it was tested as a prey not where it was tested as a bait. If a protein has been tested only as prey, its degree does not change. For studies with less preys than baits, we proceeded conversely and only counted (u,v) for the degree of u if u has been tested as a bait. In other words, we recomputed the degree as the prey-degree for studies with fewer baits than preys and as the bait-degree for studies with fewer preys than baits.

After the degree recalculation, we computed the size balance between the number of baits and preys, which is defined as follows:

$$\text{Size balance} = \begin{cases} n^{\text{bait}}/n^{\text{prey}} & \text{if } n^{\text{bait}} \leq n^{\text{prey}} \\ n^{\text{prey}}/n^{\text{bait}} & \text{if } n^{\text{bait}} > n^{\text{prey}} \end{cases} \tag{2}$$

In order to test if the asymmetric design has an effect on PL property, we compared the size balances of studies that switch from PL to non-PL with the size balances of studies for which also the recomputed degree distributions are PL-distributed, using the one-sided Wilcoxon test.

Testing for functional and disease gene enrichment

We performed functional and disease enrichment analyses of the top 50 hubs detected by the three strategies proposed to reveal the true hub proteins (prey hubs, normalized hubs, and Y2H hubs) and top 50 hubs of our aggregated network. We used the *enrichGO* function of the clusterProfiler R package[34] (version 4.4.4) and the *enrichDO* function of the DOSE R package[35] (version 3.22.1) to perform the Gene and Disease Ontology analyses, respectively. We also performed pathway enrichment analyses (Reactome-based) using the *enrichPathway* function of the ReactomePA R package[36] (version 1.40). We used the FDR method to correct P-values and we took into account only terms with a q-value < 0.05. For each enrichment analysis, we used the entire lists of genes from which we retrieved our hypothetical true hubs and all the genes in our aggregated network as background genes.

To investigate the biological functions of the most abundant human proteins, we retrieved protein abundance data from GTEx[23] (`https://gtexportal.org/home/datasets`), consisting of 201 samples from 32 normal human tissues. We removed proteins with more than 50% of NA values across all the samples (resulting in 8,104 proteins), and we calculated the median abundance for each protein. We ordered the proteins according to the median (descending order) to perform the Gene Ontology enrichment analysis of the most abundant proteins (of different set sizes). We used the FDR method to correct P-values and we took into account only terms with a q-value < 0.05. To test if there is a significant enrichment of chaperones among nervous system disease genes (in particular for schizophrenia and psychotic disorder), we retrieved the chaperone classification from UniProt and nervous system disease-related genes from Disease Ontology[37] database using the DOSE R package[35].

Design of simulation study

We simulated observed aggregated PPI networks $G' = (V, E')$ under study bias and different false negative rates *FNR* false positive rates *FPR* and from hypothetical ground truth networks $G = (V, E)$. The hypothetical ground truth networks were generated using the BA model, parameterized with the number of nodes n and the number m_{BA} of edges added per iteration, and the ER model, parameterized with the number of nodes n and the number of edges m_{ER}. The degree distributions of BA graphs are known to follow the power law, while node degrees in ER graphs are binomially distributed. Details on choices of n, m_{BA}, and m_{ER} are provided at the end of this subsection.

We start the simulation of G' with a network on the nodes V without any edges. Throughout the simulation, we add edges to the network by iteratively sampling lists of protein pairs $L_i \subset V \times V$ and then simulating an experiment that tests all $(u,v) \in L_i$ for interaction. The experiment returns a binary flag $\text{result}(u,v) \in \{0,1\}$, where 1 encodes "involved in interaction" and 0 encodes "not involved in interaction". The result probabilities depend on whether (u,v) is an edge

in the ground truth network G, as well as on the false negative and false positive rates:

$$Pr(\text{result}(u,v)=1) = \begin{cases} 1-FNR & \text{if } (u,v) \in E \\ FPR & \text{if } (u,v) \notin E \end{cases} \tag{3}$$

To simulate G', we maintain matrices $\mathbf{A} = (a_{u,v}) \in \mathbb{N}^{V \times V}$ and $\mathbf{B} = (b_{u,v}) \in \mathbb{N}^{V \times V}$. The entry $a_{u,v}$ of the matrix $\mathbf{A}$ counts the number of times the pair (u,v) has been tested for interaction, while $b_{u,v} = \sum_{i=0}^{a_{u,v}} \text{result}(u,v)$ counts the number of times the experiments have returned that u and v interact. Both $a_{u,v}$ and $b_{u,v}$ are initially set to 0 and increase during simulation. Note that we always have $b_{u,v} \leq a_{u,v}$. After each simulated experiment, $\mathbf{A}$ and $\mathbf{B}$ are updated. Subsequently, we update the edge set of the simulated network as

$$E' = \{(u,v) \mid b_{u,v} > 0 \wedge b_{u,v}/a_{u,v} > \gamma\}, \tag{4}$$

where $\gamma \in [0,1)$ is the minimum required fraction of experiments with positive result. The simulation stops once we have carried out N simulated experiments (see end of this subsection for details on choice of N).

To sample the list L_i of protein pairs to be tested for interaction in the ith experiment, three hyper-parameters are required: the number of baits $n_i^{\text{bait}} \in \mathbb{N}$, the number of preys $n_i^{\text{prey}} \in \mathbb{N}$, and the test method $M \in \{\text{Y2H}, \text{AP-MS}\}$ (which does not depend on i). L_i is constructed as $L_i = B_i \times P_i$, where $B_i \subseteq V$ and $P_i \subseteq V$ are sampled lists of baits and preys, respectively. To construct B_i, n_i^{bait} proteins are sampled without replacement from V. A protein $u \in V$ is included in B_i with probability

$$Pr(u \in B_i) \propto \deg_{i-1}(u) + \delta, \tag{5}$$

where $\deg_{i-1}(u)$ is u's node degree in the version of the simulated network G' after $i-1$ experiments and $\delta > 0$ is a hyper-parameter encoding a baseline probability (set to $\delta = 0.01$ in our simulation study). $Pr(u \in B_i)$ hence increases with increasing node degree in the simulated observed network. This leads to a positive feedback loop in the selection of bait proteins, which models study bias in our simulation study.

Since the selection of bait proteins is influenced by study bias both in AP-MS and in Y2H experiments, we use eq. (5) independently of the test method M. In contrast to AP-MS, prey and bait are selected in Y2H and thus both are subject to study bias. Consequently, we construct P_i by sampling n_i^{prey} proteins without replacement from V, where $u \in V$ is included in P_i with probability

$$Pr(u \in P_i) \propto \begin{cases} 1 & \text{if } M = \text{AP-MS} \\ \deg_{i-1}(u) + \delta & \text{if } M = \text{Y2H} \end{cases}. \tag{6}$$

We carried out our simulations for $M \in \{\text{AP-MS}, \text{Y2H}\}$, $FNR \in \{0.0, 0.1, \ldots, 0.4\}$, $FPR \in \{0.0, 0.4 \cdot 2^{-7}, \ldots, 0.4 \cdot 2^{-1}, 0.4\}$, and $\gamma \in \{0.0, 0.5\}$. The upper bound 0.4 for FNR and FPR was chosen based on estimates for false positive and negative rates in AP-MS and Y2H experiments found in the literature[27]. The values for γ were chosen to mirror a scenario where a PPI is included in the aggregated PPI network as soon as it is reported by at least one study ($\gamma = 0.0$), as well as a scenario where only those PPIs (u,v) are included for which the majority of studies testing (u,v) report an interaction ($\gamma = 0.5$). Overall, we hence carried out simulations for 180 configurations (M, FNR, FPR, γ) of free hyper-parameters.

The remaining hyper-parameters were chosen based on the sizes of observed PPI networks obtained for IntAct. For $M = \text{AP-MS}$, we set the overall number of simulated experiments N to the number of AP-MS studies annotated in IntAct where, for each PPI, information about the roles (bait or prey) of the interacting proteins is available. For each study i, n_i^{bait} and n_i^{prey} are set to the number of unique preys and baits used in the study. To set the hyper-parameters of the hypothetical ground truth networks $G = (V,E)$, we aggregated the PPIs from all N IntAct AP-MS studies and then set n and m_{ER} to the numbers of nodes and edges in the aggregated network G_{IntAct}. To ensure that also the ground truth networks generated with the BA model have approximately the same number of edges as G_{IntAct}, we set

$$m_{\text{BA}} = \text{round}\left(\frac{n}{2} - \sqrt{\frac{n^2}{4} - m_{\text{ER}}}\right) \tag{7}$$

and initialized the generation of the BA graph with the star on $m_{\text{BA}} + 1$ nodes (default in NetworkX). With this initialization, the number of edges in the final BA graph equals $|E| = m_{\text{BA}} + m_{\text{BA}} \cdot (n - (m_{\text{BA}} + 1))$, which implies

$|E| \approx m_{\text{ER}}$ if m_{BA} is chosen as specified in eq. (7). For $M = \text{Y2H}$, the hyper-parameters N, n_i^{bait}, n_i^{prey}, n, m_{ER}, and m_{BA} were chosen analogously.

For each configuration (M, FNR, FPR, γ) of free hyper-parameters, we sought to answer the question whether, given (M, FNR, FPR, γ), the observed PPI network G_{IntAct} is more similar to simulated networks that emerged from a PL-distributed or from a binomially distributed ground truth. For this, we simulated 50 networks G' from BA ground truths (which we collect in the set $\mathscr{G}_{\text{BA}}$) and 50 networks G' from ER ground truths (which we collect in the set $\mathscr{G}_{\text{BA}}$), using the simulator described above. Next, for each $G' \in \mathscr{G}_{\text{BA}} \cup \mathscr{G}_{\text{ER}}$, we computed the earth mover's distance $EMD(G_{\text{IntAct}}, G')$ between the node degree distributions of G_{IntAct} and G', and then computed the normalized signed difference

$$\Delta SOD = \frac{\sum_{G' \in \mathscr{G}_{\text{BA}}} EMD(G_{\text{IntAct}}, G') - \sum_{G' \in \mathscr{G}_{\text{ER}}} EMD(G_{\text{IntAct}}, G')}{\sum_{G' \in \mathscr{G}_{\text{ER}}} EMD(G_{\text{IntAct}}, G')} \tag{8}$$

between the sum of distances between the observed PPI network G_{IntAct} and the simulated networks contained in $\mathscr{G}_{\text{BA}}$ and $\mathscr{G}_{\text{ER}}$, respectively. ΔSOD is negative if G_{IntAct}'s degree distribution is closer to degree distributions of simulated networks which emerged from a PL-distributed ground truth rather than from a binomially distributed ground truth. Positive values of ΔSOD are indicative of the opposite scenario.

We also analyzed our results from a Bayesian perspective, addressing the question whether the observed PPI network G_{IntAct} is more likely to have emerged from an PL-distributed or from a binomially distributed biological interactome. Using Bayesian inference, this can be phrased as the task to estimate the posterior probabilities $Pr(\mathscr{C}_{\text{PL}} \mid G_{\text{IntAct}})$ and $Pr(\mathscr{C}_{\text{BN}} \mid G_{\text{IntAct}})$, where $\mathscr{C}_{\text{PL}}$ and $\mathscr{C}_{\text{BN}}$ denote the classes of observed PPI networks that have emerged from PL-distributed and binomially distributed ground truth networks, respectively. Note that, by construction, we have $\mathscr{G}_{\text{BA}} \subset \mathscr{C}_{\text{PL}}$ and $\mathscr{G}_{\text{ER}} \subset \mathscr{C}_{\text{BN}}$. To estimate the posterioris, we sorted the simulated networks $G' \in \mathscr{G}_{\text{BA}} \cup \mathscr{G}_{\text{ER}}$ in increasing order w. r. t. $EMD(G_{\text{IntAct}}, G')$, leading to a sorted list of networks $(G'_j)_{j=1}^{100}$. For varying $K \in \{1, 2, \ldots, 100\}$, we then used K-NN classification to estimate the posterior probabilities as follows ($[\cdot]$ is the Iverson bracket, i. e., $[\texttt{true}] = 1$ and $[\texttt{false}] = 0$):

$$Pr(\mathscr{C}_{\text{PL}} \mid G_{\text{IntAct}}) \approx \frac{1}{K} \cdot \sum_{j=1}^{K} \left[G'_j \in \mathscr{G}_{\text{BA}} \right] \tag{9}$$

$$Pr(\mathscr{C}_{\text{BN}} \mid G_{\text{IntAct}}) \approx \frac{1}{K} \cdot \sum_{j=1}^{K} \left[G'_j \in \mathscr{G}_{\text{ER}} \right] \tag{10}$$

Acknowledgements

We thank Richard Koll for discussions and his contributions to a preliminary version of the simulator. The work leading to this manuscript was supported by Fondazione AIRC, grant reference number MFAG 21791, and partially supported by the Italian Ministry of Health with Ricerca Corrente and 5x1000 funds. DBB is supported by the German Federal Ministry of Education and Research (BMBF, grant no. 031L0309A). Part of the resources used in this work have been provided by the Cloud infrastructure at GARR, the Italian Research and Education Network. Figure 1 was created with BioRender.com.

Competing interests

The authors declare no competing interests.

Data availability

Data used to generate the results are available at `https://doi.org/10.5281/zenodo.7695121`

Code availability

Source code to reproduce the results of the simulation study is available at `https://github.com/bionetslab/ppi-network-simulation`. Source code to reproduce all other analyses is available at `https://github.com/`

martaluc/powerlaw-ppi-network

Author contributions

MLi, DBB, and MHS conceived this study and designed it with the help of MLu. DBB developed and implemented the simulator. MLu carried out the experiments. MLu, MLi, DBB, and MHS wrote the manuscript. All authors approved the final manuscript.

References

[1] Barabási, A.-L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–512 (1999). https://doi.org/10.1126/science.286.5439.509.

[2] Cohen, R., Erez, K., ben Avraham, D. & Havlin, S. Resilience of the internet to random breakdowns. *Phys. Rev. Lett.* **85**, 4626–4628 (2000). http://dx.doi.org/10.1103/PhysRevLett.85.4626. https://doi.org/10.1103/PhysRevLett.85.4626.

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

[4] Barabási, A.-L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* **5**, 101–113 (2004). https://doi.org/10.1038/nrg1272.

[5] Yook, S.-H., Oltvai, Z. N. & Barabási, A.-L. Functional and topological characterization of protein interaction networks. *Proteomics* **4**, 928–942 (2004). https://doi.org/10.1002/pmic.200300636.

[6] Nobeli, I., Favia, A. D. & Thornton, J. M. Protein promiscuity and its implications for biotechnology. *Nat. Biotechnol.* **27**, 157–167 (2009). http://dx.doi.org/10.1038/nbt1519. https://doi.org/10.1038/nbt1519.

[7] Pastor-Satorras, R., Smith, E. & Solé, R. V. Evolving protein interaction networks through gene duplication. *J. Theor. Biol.* **222**, 199–210 (2003). https://doi.org/10.1016/s0022-5193(03)00028-6.

[8] Stelzl, U. *et al.* A human protein-protein interaction network: A resource for annotating the proteome. *Cell* **122**, 957–968 (2005). https://www.sciencedirect.com/science/article/pii/S0092867405008664. https://doi.org/https://doi.org/10.1016/j.cell.2005.08.029.

[9] Xu, J. & Li, Y. Discovering disease-genes by topological features in human protein–protein interaction network. *Bioinformatics* **22**, 2800–2805 (2006). https://doi.org/10.1093/bioinformatics/btl467. https://doi.org/10.1093/bioinformatics/btl467. https://academic.oup.com/bioinformatics/article-pdf/22/22/2800/48838835/bioinformatics_22_22_2800.pdf.

[10] Janyasupab, P., Suratanee, A. & Plaimas, K. Network diffusion with centrality measures to identify disease-related genes. *Mathematical Biosciences and Engineering* **18**, 2909–2929 (2021). https://www.aimspress.com/article/doi/10.3934/mbe.2021147. https://doi.org/10.3934/mbe.2021147.

[11] Lazareva, O., Baumbach, J., List, M. & Blumenthal, D. B. On the limits of active module identification. *Brief. Bioinform.* (2021). http://dx.doi.org/10.1093/bib/bbab066. https://doi.org/10.1093/bib/bbab066.

[12] Tanaka, R., Yi, T.-M. & Doyle, J. Some protein interaction data do not exhibit power law statistics. *FEBS Lett.* **579**, 5140–5144 (2005). https://doi.org/10.1016/j.febslet.2005.08.024.

[13] Lima-Mendez, G. & van Helden, J. The powerful law of the power law and other myths in network biology. *Mol. Biosyst.* **5**, 1482–1493 (2009). https://doi.org/10.1039/b908681a.

[14] Broido, A. D. & Clauset, A. Scale-free networks are rare. *Nat. Commun.* **10**, 1017 (2019). http://dx.doi.org/10.1038/s41467-019-08746-5. https://doi.org/10.1038/s41467-019-08746-5.

[15] Peel, L., Peixoto, T. P. & De Domenico, M. Statistical inference links data and theory in network science. *Nat. Commun.* **13**, 6794 (2022). http://dx.doi.org/10.1038/s41467-022-34267-9. https://doi.org/10.1038/s41467-022-34267-9.

[16] Berggård, T., Linse, S. & James, P. Methods for the detection and analysis of protein-protein interactions. *Proteomics* **7**, 2833–2842 (2007). http://dx.doi.org/10.1002/pmic.200700131. https://doi.org/10.1002/pmic.200700131.

[17] Schaefer, M. H., Serrano, L. & Andrade-Navarro, M. A. Correcting for the study bias associated with protein-protein interaction measurements reveals differences between protein degree distributions from different cancer types. *Front. Genet.* **6**, 260 (2015). http://dx.doi.org/10.3389/fgene.2015.00260. https://doi.org/10.3389/fgene.2015.00260.

[18] Mosca, E. *et al.* Characterization and comparison of gene-centered human interactomes. *Briefings in Bioinformatics* **22** (2021). https://doi.org/10.1093/bib/bbab153. https://doi.org/10.1093/bib/bbab153. Bbab153, https://academic.oup.com/bib/article-pdf/22/6/bbab153/41088110/bbab153.pdf.

[19] Clauset, A., Shalizi, C. R. & Newman, M. E. J. Power-Law distributions in empirical data. *SIAM Rev.* **51**, 661–703 (2009). http://www.jstor.org/stable/25662336.

[20] Luck, K. *et al.* A reference map of the human binary protein interactome. *Nature* **580**, 402–408 (2020). http://dx.doi.org/10.1038/s41586-020-2188-x. https://doi.org/10.1038/s41586-020-2188-x.

[21] Fink, A. L. Chaperone-mediated protein folding. *Physiological reviews* **79**, 425–449 (1999).

[22] Finka, A. & Goloubinoff, P. Proteomic data from human cell cultures refine mechanisms of chaperone-mediated protein homeostasis. *Cell Stress and Chaperones* **18**, 591–605 (2013).

[23] Jiang, L. *et al.* A quantitative proteome map of the human body. *Cell* **183**, 269–283 (2020).

[24] Tittelmeier, J., Nachman, E. & Nussbaum-Krammer, C. Molecular chaperones: a double-edged sword in neurodegenerative diseases. *Frontiers in aging neuroscience* **12**, 581374 (2020).

[25] Nucifora, L. G. *et al.* Increased protein insolubility in brains from a subset of patients with schizophrenia. *American Journal of Psychiatry* **176**, 730–743 (2019).

[26] Erdös, P. & Rényi, A. On random graphs I. *Publ. Math. Debrecen* **6**, 290–297 (1959).

[27] Armean, I. M., Lilley, K. S. & Trotter, M. W. Popular computational methods to assess multiprotein complexes derived from label-free affinity purification and mass spectrometry (ap-ms) experiments*. *Molecular & Cellular Proteomics* **12**, 1–13 (2013). https://www.sciencedirect.com/science/article/pii/S1535947620334253. https://doi.org/https://doi.org/10.1074/mcp.R112.019554.

[28] Chung, F., Lu, L., Dewey, T. G. & Galas, D. J. Duplication models for biological networks. *J. Comput. Biol.* **10**, 677–687 (2003). http://dx.doi.org/10.1089/106652703322539024. https://doi.org/10.1089/106652703322539024.

[29] Haynes, W. A., Tomczak, A. & Khatri, P. Gene annotation bias impedes biomedical research. *Sci. Rep.* **8**, 1362 (2018). https://doi.org/10.1038/s41598-018-19333-x.

[30] Lenz, S. *et al.* Reliable identification of protein-protein interactions by crosslinking mass spectrometry. *Nature communications* **12**, 3564 (2021).

[31] Orchard, S. *et al.* The mintact project—intact as a common curation platform for 11 molecular interaction databases. *Nucleic acids research* **42**, D358–D363 (2014).

[32] Alanis-Lobato, G., Andrade-Navarro, M. A. & Schaefer, M. H. Hippie v2. 0: enhancing meaningfulness and reliability of protein–protein interaction networks. *Nucleic acids research* gkw985 (2016).

[33] Gillespie, C. S. Fitting heavy tailed distributions: the powerlaw package. *arXiv preprint arXiv:1407.3492* (2014).

[34] Wu, T. *et al.* clusterprofiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation* **2**, 100141 (2021).

[35] Yu, G., Wang, L.-G., Yan, G.-R. & He, Q.-Y. Dose: an r/bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics* **31**, 608–609 (2015).

[36] Yu, G. & He, Q.-Y. Reactomepa: an r/bioconductor package for reactome pathway analysis and visualization. *Molecular BioSystems* **12**, 477–479 (2016).

[37] Schriml, L. M. *et al.* Human disease ontology 2018 update: classification, content and workflow expansion. *Nucleic acids research* **47**, D955–D962 (2019).