Mutual interactors as a principle for the discovery of phenotypes in molecular networks

Sabri Eyuboglu,[1,*] Marinka Zitnik,[1,*] Jure Leskovec[1,2,‡]
[1] Computer Science Department, Stanford University, Stanford, CA 94305, USA
[2] Chan Zuckerberg Biohub, San Francisco, CA 94158, USA

[*]Equal Contribution. ‡Corresponding author. Email: jure@cs.stanford.edu

Biological networks are powerful resources for the discovery of molecular phenotypes. Fundamental to network analysis is the principle—rooted in social networks—that nodes that interact in the network tend to have similar properties. While this long-standing principle underlies powerful methods in biology that associate proteins with phenotypes on the basis of network proximity, interacting proteins are not necessarily similar, and proteins with similar properties do not necessarily interact. Here, we show that proteins are more likely to have similar phenotypes, not if they directly interact in a molecular network, but if they interact with the same proteins. We call this the mutual interactor principle and show that it holds for several kinds of molecular networks, including protein-protein interaction, genetic interaction, and signaling networks. We then develop a machine learning framework for predicting molecular phenotypes on the basis of mutual interactors. Strikingly, the framework can predict drug targets, disease proteins, and protein functions in different species, and it performs better than much more complex algorithms. The framework is robust to incomplete biological data and capable of generalizing to phenotypes it has not seen during training. Our work represents a network-based predictive platform for phenotypic characterization of proteins.

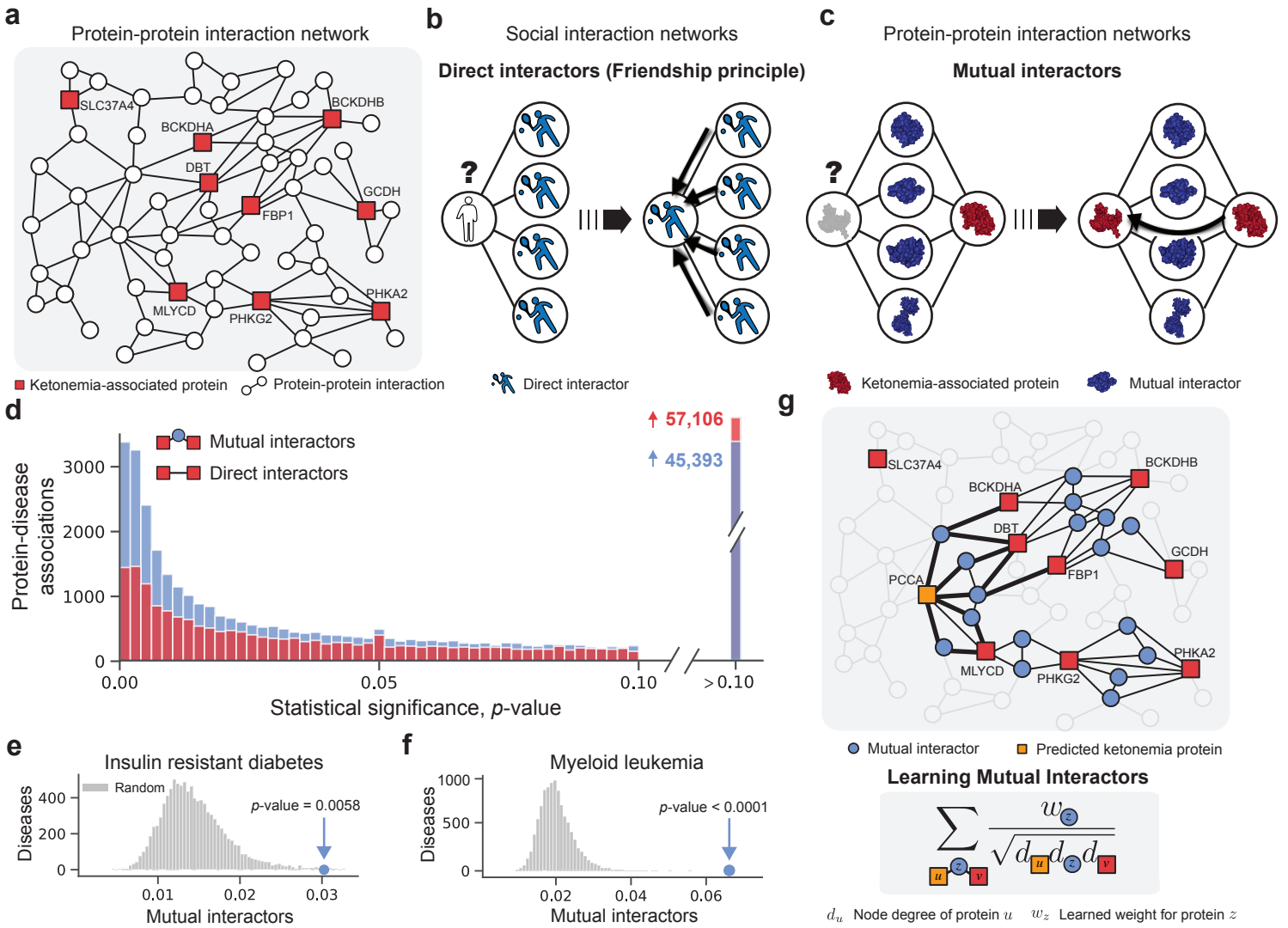

a
Protein-protein interaction network
SLC37A4
BCKDHB
BCKDHA
DBT
GCDH
FBP1
PHKA2
MLYCD
PHKG2
Ketonemia-associated protein
Protein-protein interaction
b
Social interaction networks
Direct interactors (Friendship principle)
?
Direct interactor
c
Protein-protein interaction networks
Mutual interactors
?
Ketonemia-associated protein
Mutual interactor
d
Mutual interactors
Direct interactors
57,106
45,393
Protein-disease associations
3000
2000
1000
0
0.00
0.05
0.10
>0.10
Statistical significance, p-value
e
Insulin resistant diabetes
Random
p-value = 0.0058
Diseases
400
200
0
0.01
0.02
0.03
Mutual interactors
f
Myeloid leukemia
p-value < 0.0001
Diseases
1000
500
0
0.02
0.04
0.06
Mutual interactors
g
SLC37A4
BCKDHB
BCKDHA
DBT
GCDH
PCCA
FBP1
PHKA2
MLYCD
PHKG2
Mutual interactor
Predicted ketonemia protein
Learning Mutual Interactors
$\sum \frac{w_z}{\sqrt{d_u d_z d_v}}$
d_u Node degree of protein u
w_z Learned weight for protein z

Figure 1 *(preceding page)*: The mutual interactor principle and our machine-learning framework for predicting phenotypes in molecular networks. **(a)** Ketonemia proteins in the human protein-protein interaction (PPI) network. Nodes represent proteins and edges indicate physical protein-protein interactions. Shown are 9 proteins associated with ketonemia (as red squares). **(b)** Schematic illustration of the friendship principle (*i.e.*, network homophily [1]) in a social network of five individuals. The principle predicts the favorite sport of a person (in white) based on the interests of her friends (in blue). According to the friendship principle, she is likely to enjoy tennis because many of her friends also like tennis. Bioinformatics methods borrow the friendship principle and use it to predict protein phenotypes (*e.g.*, disease proteins, protein functions, and drug targets) in molecular networks. The friendship principle has led to numerous discoveries in the social sciences, however, it is unclear whether the friendship principle captures the structural and evolutionary forces of biology. **(c)** Schematic illustration of the mutual interactor principle in a PPI network. Two nodes share a *mutual interactor* if they both interact with the same protein. The *mutual interactor* principle posits that the grey protein is likely associated with ketonemia because it interacts with the same proteins as another ketonemia protein (in red); the two proteins share four mutual interactors (in blue). Biological evidence supports the mutual interactor principle (*e.g.*, in structural biology [2, 3], two proteins with similar interfaces interact with the same proteins, but they do not necessarily interact with each other; in evolutionary biology [4], two proteins arising from a gene duplication share mutual interactors, but they do not necessarily interact). In contrast, the friendship principle would not associate the grey protein with ketonemia because it does not directly interact with any ketonemia proteins. **(d)** Comparison of mutual interactors (in blue) and direct interactors (in red) as principles of disease protein connectivity in a human PPI network. For 75,744 disease-protein associations, the statistical significance (p-value) of the mutual interactor score is computed. The mutual interactor score of a disease-protein association is the degree-normalized count of mutual interactors between the protein and other proteins associated with the disease (see Methods). Also computed for each disease-protein association is the statistical significance of the protein's direct interactions with other proteins associated with the same disease. The plot shows that proteins associated with a particular disease tend to interact with the same proteins (*i.e.*, those proteins share mutual interactors) significantly more often than they directly interact with each other. **(e-f)** For each of 1,811 diseases [5] (see Methods), we calculate the average mutual interactor score of proteins associated with a disease. Shown are mutual interactor scores for **(e)** insulin resistant diabetes and **(f)** myeloid leukemia. The observed mutual interactor scores (in blue) are significantly larger than random expectation (in grey), providing empirical evidence for the mutual interactor principle in disease pathways. **(g)** The PPI network from (a), where ketonemia proteins (in red) and their mutual interactors (in blue) are highlighted. As per the mutual interactor principle, PCCA (in orange) is the top candidate ketonemia protein. To predict proteins that are associated with a particular phenotype (*e.g.*, ketonemia) we develop a machine-learning approach that operates on the basis of mutual interactors and mathematically formalizes the principle that proteins are likely to have similar phenotypes if they share mutual interactors. By training on a large dataset of disease pathways [5], the approach automatically learns a mutual interactor strength w_z for each protein z in the network. The learned strength w_z indicates z's propensity to interact with proteins that are in the same disease pathway. The approach then calculates the probability that PCCA is associated with ketonemia by aggregating the strengths w_z of all mutual interactors between PCCA and currently known ketonemia proteins.

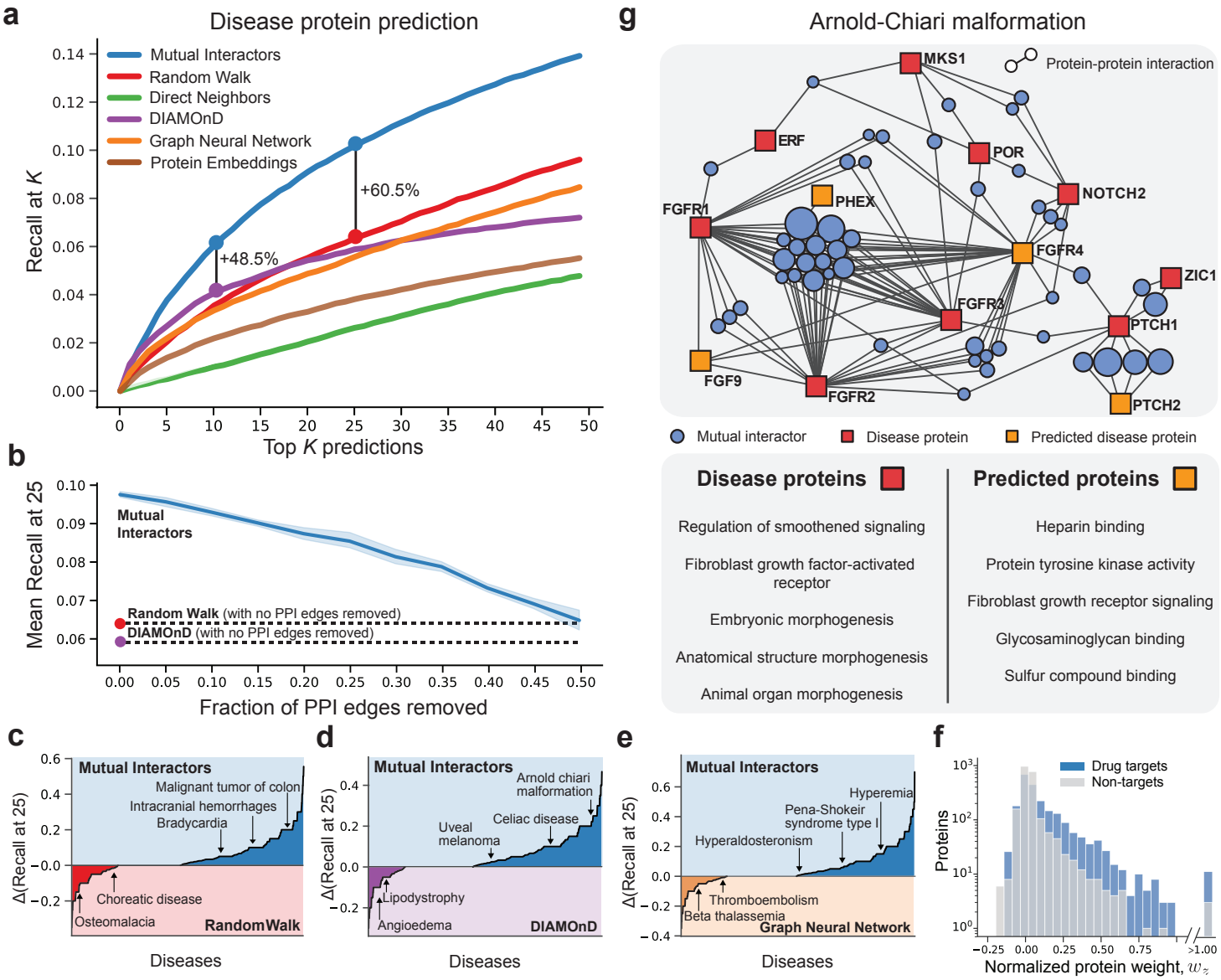

a
Disease protein prediction
Mutual Interactors
Random Walk
Direct Neighbors
DIAMOnD
Graph Neural Network
Protein Embeddings
+60.5%
+48.5%
Recall at K
Top K predictions
b
Mean Recall at 25
Mutual Interactors
Random Walk (with no PPI edges removed)
DIAMOnD (with no PPI edges removed)
Fraction of PPI edges removed
c
Mutual Interactors
Malignant tumor of colon
Intracranial hemorrhages
Bradycardia
Choreatic disease
Osteomalacia
RandomWalk
Δ(Recall at 25)
Diseases
d
Mutual Interactors
Arnold chiari malformation
Celiac disease
Uveal melanoma
Lipodystrophy
Angioedema
DIAMOnD
Δ(Recall at 25)
Diseases
e
Mutual Interactors
Hyperemia
Pena-Shokeir syndrome type I
Hyperaldosteronism
Thromboembolism
Beta thalassemia
Graph Neural Network
Δ(Recall at 25)
Diseases
f
Drug targets
Non-targets
Proteins
Normalized protein weight, w_z
g
Arnold-Chiari malformation
Protein-protein interaction
MKS1
ERF
POR
PHEX
NOTCH2
FGFR1
FGFR4
ZIC1
FGFR3
PTCH1
FGF9
FGFR2
PTCH2
Mutual interactor
Disease protein
Predicted disease protein
Disease proteins
Regulation of smoothened signaling
Fibroblast growth factor-activated receptor
Embryonic morphogenesis
Anatomical structure morphogenesis
Animal organ morphogenesis
Predicted proteins
Heparin binding
Protein tyrosine kinase activity
Fibroblast growth receptor signaling
Glycosaminoglycan binding
Sulfur compound binding

Figure 2 *(preceding page)*: Uncovering disease proteins with the mutual interactor principle and predicting protein functions across species and different types of molecular networks. (a) Overall performance evaluation. Fraction of disease proteins recovered within top k predictions for $k = 1$ to $k = 50$ (recall-at-k). At $k = 10$ and $k = 25$, we compute the percent-increase in Mutual Interactors' recall over the next best performing method. Mutual Interactors correctly predicts 48.5% more disease proteins within the first 10 proteins than DIAMOnD [6] and it identifies 60.5% more disease proteins within the first 25 proteins than random walks [7]. Mutual Interactors performs equally well when predicting protein functions (see Extended Figure 1) and biological processes (see Extended Figure 2). Further, the principle of mutual interactors holds for different types of molecular networks, and it is a consistently powerful predictor of protein functions in different species (Extended Figures 1-2). **(b)** Effect of data incompleteness on performance. A fraction of PPIs is randomly sampled and removed from the PPI network (x-axis in the plot). The Mutual Interactor method is then trained on the incomplete PPI network and method performance is evaluated (y-axis). Shown is Recall-at-25 as a function of the fraction of PPIs removed from the network. Dotted lines indicate performance of random walks and DIAMOnD on a full PPI network with no PPIs removed. With as many as 50% of the PPIs removed, Mutual Interactors still outperforms other methods, even though they have access to the full PPI network. **(c-e)** Comparison of Mutual Interactors and baseline methods. For each disease, the baseline's recall-at-25 is subtracted from Mutual Interactors'. The diseases are plotted in order of increasing difference and a few diseases are annotated. Shown are results for three methods: (**c**) random walks, (**d**) DIAMOnD, and (**e**) graph neural networks [8,9]. See Supplementary Figure 10 for direct neighbors [10] and protein embeddings [11]. **(f)** Comparison of degree-normalized Mutual Interactor weights of drug targets versus non-targets. The Mutual Interactor weight of protein z is the weight w_z that Mutual Interactors learns for protein z (see Methods and Figure 1). Shown is the distribution of degree-normalized Mutual Interactor weights for 2,212 drug targets [12] (in blue), and, for comparison, the distribution of degree-normalized Mutual interactor weights for 2,212 random proteins that are not targets of any drug (in grey). Results show that Mutual Interactors, when optimized to predict disease proteins, automatically learns to assign higher weights w_z to proteins z that are druggable targets, even though information on targets is not given to the machine learning method during optimization ($p = 5.43\text{e}{-177}$; two-sided Kolmogorov-Smirnov test). Further, proteins with highest weights w_z are enriched in transmembrane signaling receptors ($p = 1.10\text{e}{-6}$), molecular transducer activity ($p = 9.68\text{e}{-6}$), signal transduction ($p = 9.68\text{e}{-6}$), and G protein-coupled receptors ($p = 5.86\text{e}{-5}$). **(g)** Mutual Interactor neighborhood for Arnold-Chiari (AC) malformation. The neighborhood includes known disease proteins (red squares), Mutual Interactors' top predictions (orange squares), and the mutual interactors between them (blue circles). Mutual interactors are sized proportional to their learned Mutual Interactor weight, w_z. Listed in the tables are the cellular functions most significantly enriched in known AC proteins (left) and Mutual Interactors' predictions (right).

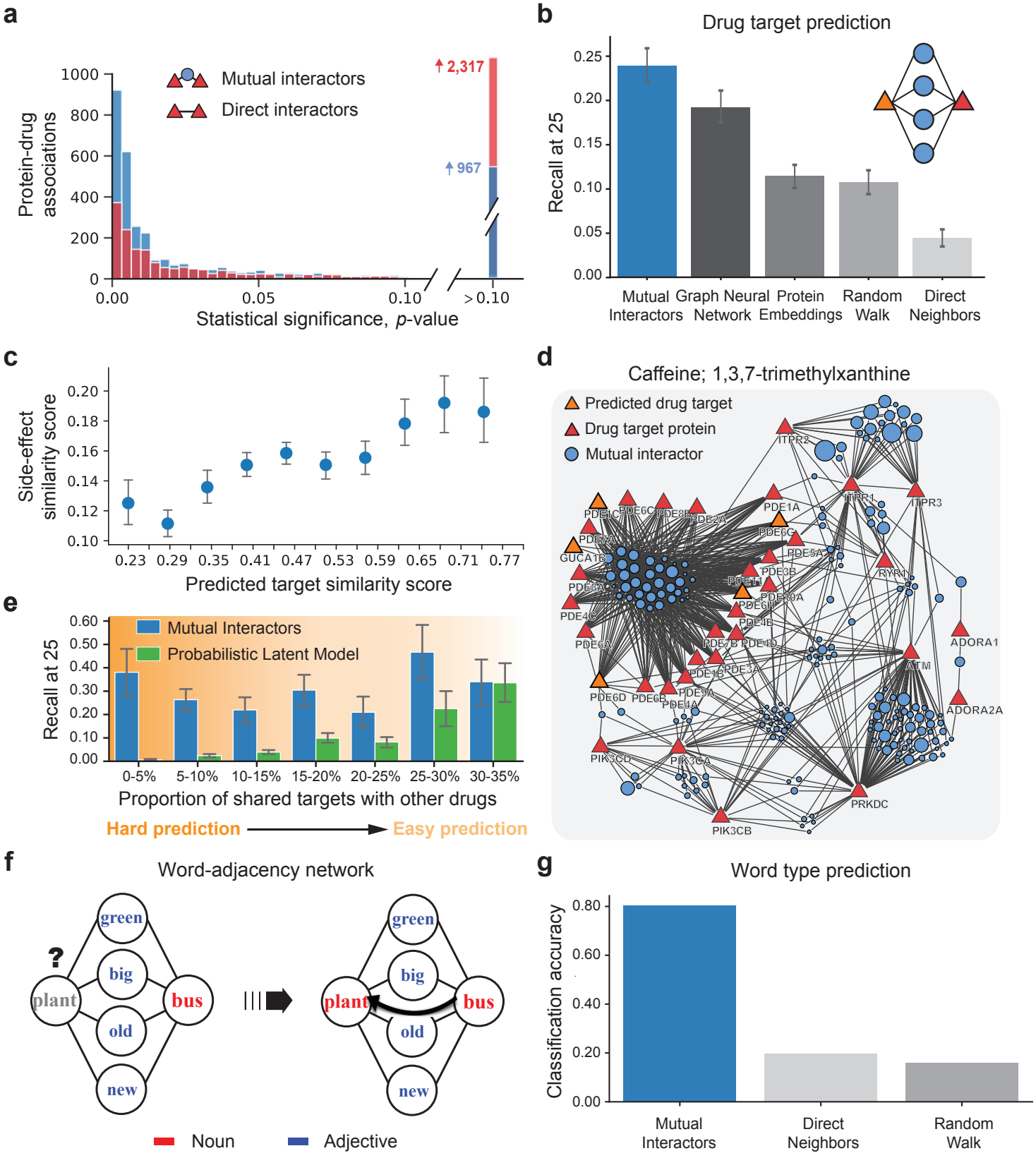

a
Protein-drug associations
Mutual interactors
Direct interactors
2,317
967
Statistical significance, p-value
b
Drug target prediction
Recall at 25
Mutual Interactors
Graph Neural Network
Protein Embeddings
Random Walk
Direct Neighbors
c
Side-effect similarity score
Predicted target similarity score
d
Caffeine; 1,3,7-trimethylxanthine
Predicted drug target
Drug target protein
Mutual interactor
e
Recall at 25
Mutual Interactors
Probabilistic Latent Model
Proportion of shared targets with other drugs
Hard prediction
Easy prediction
f
Word-adjacency network
green
big
plant
bus
old
new
Noun
Adjective
g
Word type prediction
Classification accuracy
Mutual Interactors
Direct Neighbors
Random Walk

Figure 3 *(preceding page)*: Identifying drug targets using the principle of mutual interactors and the validity of the principle in non-biological networks. **(a)** Comparison of Mutual Interactors (in blue) and direct interactors (in red) as principles of drug-target connectivity in a human PPI network. For $4{,}403$ drug-target associations [13], the statistical significance of the Mutual Interactor score is calculated using a non-parametric permutation test. The Mutual Interactor score of a drug-target association is the degree-normalized count of mutual interactors between the target and other proteins targeted by the same drug (see Methods). Also calculated for each drug-target association is the statistical significance of the target's direct interactions with other proteins targeted by the drug. Shown is the distribution of p-values over all drug-target associations indicating that proteins targeted by the same drug interact with the same proteins (*i.e.*, share mutual interactors) significantly more often than they directly interact with each other. **(b)** Drug target identification. The plot shows average Recall-at-25 across 190 drugs. Higher values indicate better performance. Mutual Interactors outperforms deep graph neural networks [8, 9] by 13% and it improves upon random walk-based approaches [7] by 125%. Confidence intervals (95%) were determined using bootstrapping with $n = 1{,}000$ iterations. **(c)** The side-effect similarity of drugs [14] (y-axis) is linearly related to the similarity of Mutual Interactors' predictions for those drugs (x-axis), implying a direct correlation between side-effect similarity and target binding and hence a possibility for Mutual Interactors to predict off-target binding [15]. **(d)** Mutual Interactor neighborhood for proteins targeted by Caffeine. The neighborhood includes caffeine-targetted proteins (red triangles), Mutual Interactors' top predictions for novel caffeine targets (orange triangles), and the mutual interactors between them (blue circles). Mutual interactors are sized proportional to their learned Mutual Interactor weight, w_z (see Methods). **(e)** The fraction of a drug's targets recovered within the top 25 predictions (recall-at-25) vs. the maximum Jaccard similarity between the drug's targets and targets of other drugs in the training set used for machine learning. Bars indicate average recall-at-25 in each bucket. Confidence intervals (95%) were computed via bootstrapping with $N = 1{,}000$ samples. Results show that the performance of a probabilistic latent-based method [16] is highly dependent on the similarity between the test drug and drugs in the training set. In contrast, Mutual Interactors' performance is largely independent of the similarity between training drugs and the test drug. Even when there are no drugs with similar targets in the training set (far left in the plot), Mutual Interactors recovers 38.1% of drug targets. By contrast, a probabilistic latent method correctly identifies only 0.05% of drug targets, indicating that the method fails to generalize to drugs that differ substantially from drugs on which it was trained. **(f)** A word-adjacency network based on the novel *David Copperfield* by Charles Dickens [17]. To construct the network, we take the 60 most commonly occurring nouns and the 60 most commonly occurring adjectives in the text. The nodes in the network represent words and an edge connects any two words that appear adjacent to one another at any point in the text. Typically adjectives occur next to nouns in English. It is possible for adjectives to occur next to other adjectives (*e.g.* "the big, green bus") or for nouns to occur next to other nouns (*e.g.* "the car keys"), but these positionings are less common. Thus, the principle of Mutual Interactors holds for the word-adjacency network: edges run primarily between words of different types with fewer edges between words of the same type. **(g)** Mutual Interactors correctly predicts part-of-speech (*i.e.*, nouns vs. adjectives) for 80.3% of the words, whereas direct neighbor and random walk approaches make correct predictions only 19.6% and 16.0% of the time, respectively.

Data availability. All data used in the paper, including molecular interaction networks and datasets with phenotype information, will be shared with the community and available from snap.stanford.edu.

Code availability. PyTorch implementation of Mutual Interactors and the associated machine learning framework as well as implementations of all baseline methods and all experiments will be publicly available on Github.

Acknowledgements. S.E., M.Z., and J.L. were supported by NSF, NIH BD2K, Stanford Data Science Initiative, and the Chan Zuckerberg Biohub.

Author contribution. S.E., M.Z., and J.L. designed and performed research, contributed new analytic tools, analyzed data, and wrote the paper.

Author information. The authors declare no conflict of interest. Correspondence should be addressed to J.L. (jure@cs.stanford.edu).

References

1. McPherson, M., Smith-Lovin, L. & Cook, J. M. Birds of a feather: Homophily in social networks. *Annual Review of Sociology* **27**, 415–444 (2001).
2. Consortium, A. I. M. *et al.* Evidence for network evolution in an arabidopsis interactome map. *Science* **333**, 601–607 (2011).
3. Keskin, O., Tuncbag, N. & Gursoy, A. Predicting protein–protein interactions from the molecular to the proteome level. *Chemical Reviews* **116**, 4884–4909 (2016).
4. Zitnik, M., Sosic, J., Feldman, M. W. & Leskovec, J. Evolution of resilience in protein interactomes across the tree of life. *Proceedings of the National Academy of Sciences* **116**, 4426–4433 (2019).
5. Piñero, J. *et al.* DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Research* gkw943 (2016).

6. Ghiassian, S. D., Menche, J. & Barabási, A.-L. A DIseAse MOdule Detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. *PLoS Computational Biology* **11**, e1004120 (2015).

7. Cowen, L., Ideker, T., Raphael, B. J. & Sharan, R. Network propagation: a universal amplifier of genetic associations. *Nature Reviews Genetics* **18**, 551 (2017).

8. Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representations* (2017).

9. Hamilton, W., Ying, Z. & Leskovec, J. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, 1024–1034 (2017).

10. Barabási, A.-L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics* **12**, 56 (2011).

11. Grover, A. & Leskovec, J. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 855–864 (2016).

12. Wishart, D. S. *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research* **46**, D1074–D1082 (2018).

13. Wishart, D. S. *et al.* Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic Acids Research* **46**, D1074–D1082 (2017).

14. Tatonetti, N. P., Patrick, P. Y., Daneshjou, R. & Altman, R. B. Data-driven prediction of drug effects and interactions. *Science Translational Medicine* **4**, 125ra31–125ra31 (2012).

15. Campillos, M., Kuhn, M., Gavin, A.-C., Jensen, L. J. & Bork, P. Drug target identification using side-effect similarity. *Science* **321**, 263–266 (2008).

16. Mnih, A. & Salakhutdinov, R. R. Probabilistic matrix factorization. In *Advances in Neural Information Processing Systems*, 1257–1264 (2008).

17. Newman, M. E. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E* **74**, 036104 (2006).

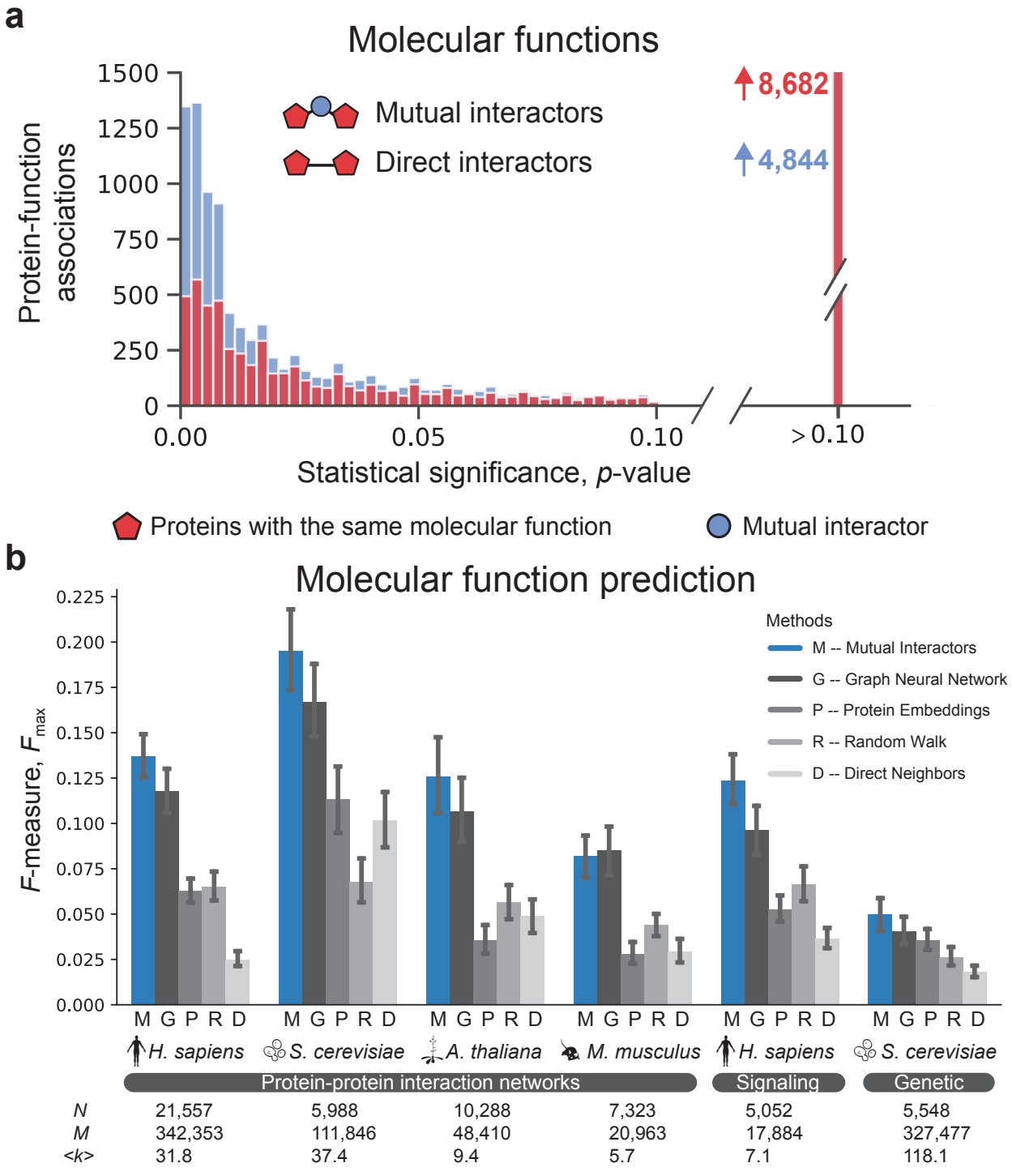
18. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**, 25 (2000).

19. Radivojac, P. *et al.* A large-scale evaluation of computational protein function prediction. *Nature Methods* **10**, 221 (2013).

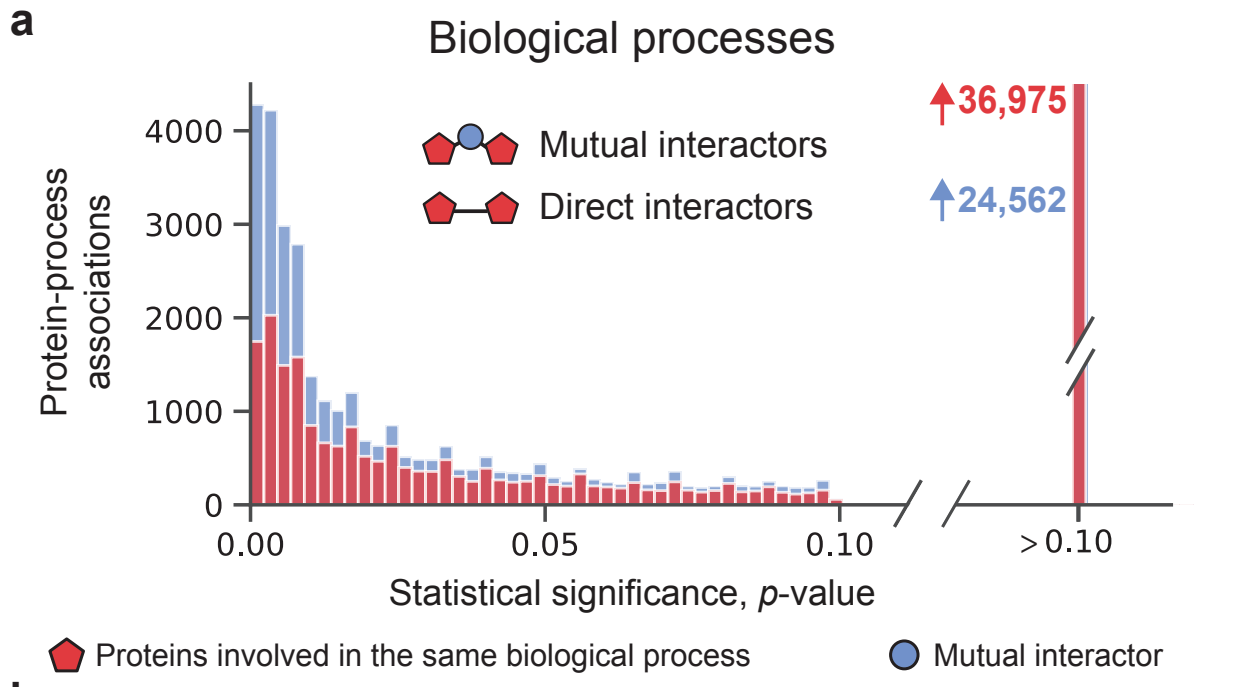
20. Costanzo, M. *et al.* The genetic landscape of a cell. *Science* **327**, 425–431 (2010).

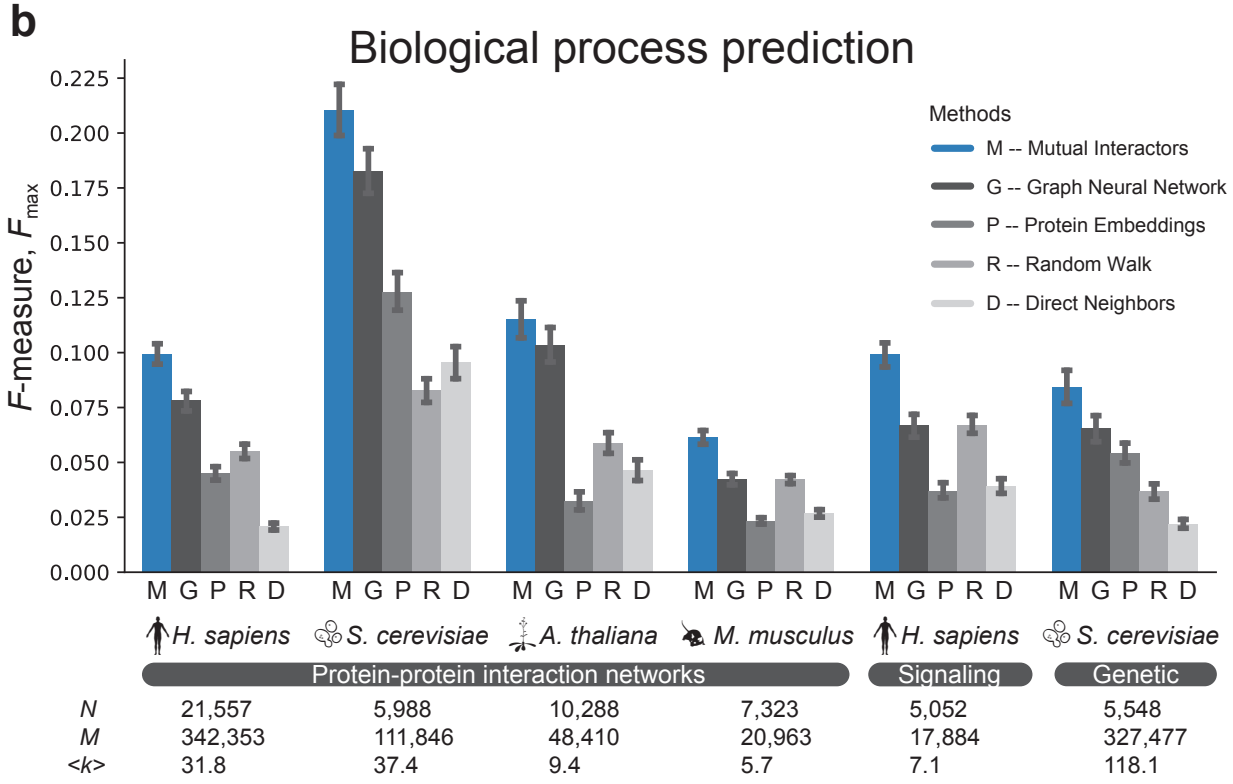
21. Choobdar, S. *et al.* Assessment of network module identification across complex diseases. *Nature Methods* **16**, 843–852 (2019).

22. Türei, D., Korcsmáros, T. & Saez-Rodriguez, J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nature Methods* **13**, 966 (2016).


a
Molecular functions
Protein-function associations
1500
1250
1000
750
500
250
0
Mutual interactors
Direct interactors
8,682
4,844
0.00
0.05
0.10
>0.10
Statistical significance, p-value
Proteins with the same molecular function
Mutual interactor
b
Molecular function prediction
F-measure, Fmax
0.225
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
Methods
M -- Mutual Interactors
G -- Graph Neural Network
P -- Protein Embeddings
R -- Random Walk
D -- Direct Neighbors
M G P R D
H. sapiens
S. cerevisiae
A. thaliana
M. musculus
H. sapiens
S. cerevisiae
Protein-protein interaction networks
Signaling
Genetic
N 21,557 5,988 10,288 7,323 5,052 5,548
M 342,353 111,846 48,410 20,963 17,884 327,477
<k> 31.8 37.4 9.4 5.7 7.1 118.1

Extended Figure 1 *(preceding page)*: Predicting protein functions across species and different types of molecular networks. (a) Comparison of Mutual Interactors (in blue) and direct interactors as principles of human PPI connectivity for molecular functions [18]. For each of the 13,983 protein-function associations in our dataset, the statistical significance of the Mutual Interactor score is calculated using a non-parametric permutation test (in blue). The Mutual Interactor score of a protein-function association is the degree-normalized count of mutual interactors between the protein and other proteins associated with the same function (see Methods). Also calculated for each protein-function association is the statistical significance of the protein's direct interactions with other proteins associated with the function (in red). Shown is the distribution of p-values over all 13,983 protein-function associations. Results show that proteins with same molecular function tend to interact with the same proteins (*i.e.*, they share mutual interactors) significantly more often than they directly interact with each other. **(b)** Overall protein function prediction performance across four species and six molecular networks. We predict Molecular Function Ontology [18] terms using PPI, signaling, and genetic interaction networks for human, yeast *S. cerevisiae*, mouse *M. musculus*, and a plant *A. thaliana*. We show average maximum F-measure [19]. A perfect predictor would be characterized by $F_{max} = 1$. On the PPI networks across the four species, Mutual Interactors outperforms deep graph neural networks [8, 9], *i.e.*, the second-best performing approach, by 18%. Further, Mutual Interactors outperforms random walks, currently the most popular network-based prediction method [7], by 96%. On the negative genetic interaction network in yeast [20], Mutual Interactors outperforms graph neural networks by 28%. Finally, on the human signaling network [21,22], Mutual Interactors provides an average improvement of 43% over the next-best method. Confidence intervals (95%) were determined using bootstrapping with $n = 1{,}000$ iterations. N – number of nodes, M – number of edges, $<k>$ – average node degree.


a
Biological processes
Protein-process associations
4000
3000
2000
1000
0
Mutual interactors
Direct interactors
36,975
24,562
0.00
0.05
0.10
>0.10
Statistical significance, p-value
Proteins involved in the same biological process
Mutual interactor


b
Biological process prediction
F-measure, Fmax
0.225
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
Methods
M -- Mutual Interactors
G -- Graph Neural Network
P -- Protein Embeddings
R -- Random Walk
D -- Direct Neighbors
M G P R D
H. sapiens
S. cerevisiae
A. thaliana
M. musculus
H. sapiens
S. cerevisiae
Protein-protein interaction networks
Signaling
Genetic
N 21,557 5,988 10,288 7,323 5,052 5,548
M 342,353 111,846 48,410 20,963 17,884 327,477
<k> 31.8 37.4 9.4 5.7 7.1 118.1

Extended Figure 2 *(preceding page)*: Predicting biological processes across species and different types of molecular networks. **(a)** Comparison of Mutual Interactors (in blue) and direct interactors (in red) as principles of human PPI connectivity for biological processes [18]. For each of the $55{,}884$ protein-process associations in our dataset, the statistical significance of the Mutual Interactor score (in blue) is calculated using a non-parametric permutation test. The Mutual Interactor score of a protein-process association is the degree-normalized count of mutual interactors between the protein and other proteins involved in the same process (see Methods). Also calculated for each protein-process association is the statistical significance of the protein's direct interactions with other proteins in the same process (in red). Shown is the distribution of p-values over all $55{,}884$ protein-process associations. Results show that proteins involved in the same biological process tend to interact with the same proteins (*i.e.*, they share mutual interactors) significantly more often than they directly interact with each other. **(b)** Overall performance evaluation across four species and six molecular networks. We predict Biological Process Ontology [18] terms using PPI, signaling, and genetic interaction networks for human, yeast *S. cerevisiae*, mouse *M. musculus*, and a plant *A. thaliana* (see Extended Figure 1 for references to network datasets). We show average maximum F-measure [19]. A perfect predictor would be characterized by $F_{max} = 1$. Results show that Mutual Interactors consistently achieves the best performance across species and different types of molecular networks. Interestingly, Mutual Interactors performs better than much more complex algorithms, including protein embeddings and graph neural networks (see Methods for description of those algorithms). Confidence intervals (95%) were determined using bootstrapping with $n = 1{,}000$ iterations. N – number of nodes, M – number of edges, $<k>$ – average node degree.