# Mutual interactors as a principle for phenotype discovery in molecular interaction networks

Sabri Eyuboglu,[1,*‡] Marinka Zitnik,[2,3,4,*] Jure Leskovec[1]

[1]*Department of Computer Science, Stanford University, Stanford, CA 94305*
[2]*Department of Biomedical Informatics, Harvard University, Boston, MA 02115*
[3]*Broad Institute of MIT and Harvard, Cambridge, MA 02142*
[4]*Harvard Data Science, Cambridge, MA 02138*

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

Biological networks are powerful representations 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 molecules with phenotypes on the basis of network proximity, interacting molecules are not necessarily similar, and molecules with similar properties do not necessarily interact. Here, we show that molecules are more likely to have similar phenotypes, not if they directly interact in a molecular network, but if they interact with the same molecules. 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 is capable of generalizing to phenotypes it has not seen during training. Our work represents a network-based predictive platform for phenotypic characterization of biological molecules.

*Keywords*: Network medicine, Molecular phenotypes, Protein Interactions, Graph neural networks

## 1. Introduction

Molecules in and across living cells are constantly interacting, giving rise to complex biological networks. These networks serve as a powerful resource for the study of human disease, molecular function and drug-target interactions.[1,2] For instance, evidence from multiple sources suggests that causative genes from the same or similar diseases tend to reside in the same neighborhood of protein-protein interaction networks.[3–6] Similarly, proteins associated with the same molecular functions form highly-connected modules within protein-protein interaction networks.[7]

These observations have motivated the development of bioinformatics methods that use molecular networks to infer associations between proteins and molecular phenotypes, including diseases, molecular functions, and drug targets.[8–11] Many of these methods assume that molecular networks obey the organizing principle of homophily: the idea that similarity breeds connection (see Figure 1b).[12] However, while this principle has been well-documented in social networks of many types

(*e.g.* friendship, work, co-membership), it is unclear whether it captures the dynamics of biological networks. If not, existing bioinformatics methods that assume homophily may not realize the full potential of biological networks for scientific discovery.

To better understand the place for homophily in bioinformatics, we consider groups of phenotypically similar molecules (*e.g.* molecules associated with the same disease, involved in the same function, or targeted by the same drug) and study their interactions in large-scale biological networks. We find that most molecules associated with similar phenotypes do not interact directly in molecular networks, a result which puts into question the assumption of homophily, an assumption that is taken for granted by so many bioinformatics methods.

In fact, a different principle better explains how phenotypic similarity relates to network structure in biology. On average, two molecules that interact directly with one another will have less in common than two molecules that share many *mutual interactors*, just as people in a social network may share mutual friends. We call this the mutual interactor principle and validate it empirically on a diverse set of biological networks (see Figure 1c).

Motivated by our findings, we develop a machine learning framework, *Mutual Interactors*, that can predict a molecule's phenotype based on the mutual interactors it shares with other molecules. We demonstrate the power, robustness, and scalability of *Mutual Interactors* on three key prediction tasks: disease protein prediction, drug target identification, and protein function prediction. With experiments across three different kinds of molecular networks (protein-protein interaction, signaling and genetic interaction) and four species (*H. sapiens*, *S. cerevisiae*, *A. thaliana*, *M. musculus*), we find that *Mutual Interactors* substantially outperforms existing methods, with gains in recall up to 61%. Additionally, we show that the weights learned by our method provide insight into the functional properties and druggability of mutual interactors.

*Mutual Interactors* is an approach based on a different network principle than existing bioinformatics methods. That it can outperform state-of-the-art approaches suggests a need to rethink the fundamental assumptions underlying machine learning methods for network biology.

## 2. Network connectivity of molecular phenotypes

One way we measure phenotypic similarity between two molecules is by comparing the set of phenotypes (e.g., diseases or functions) associated with each molecule and quantifying their similarity with the Jaccard index. We find that the average Jaccard index of the 62,084 molecule pairs that interact in the human reference interactome (HuRI) is significantly smaller than the average Jaccard index of the 62,084 molecule pairs with most degree-normalized mutual interactors ($p = 2.00 \times 10^{-59}$, dependent $t$-test).[13] We replicate this finding on three other large-scale interactomes: a PPI network derived from the BioGRID database[14] ($p = 3.56 \times 10^{-26}$) another derived from the STRING database[15] ($p = 1.29 \times 10^{-10}$) and the PPI network compiled by Menche *et al.* ($p = 1.02 \times 10^{-4}$).[16]

To further evaluate these two principles (i.e., homophily and Mutual Interactor), we collect $75{,}744$ disease-protein associations[17] and analyze their interactions in the protein-protein interaction network (see Figure 1d-f and Figure D4). For each disease-protein association we compute the fraction of the protein's direct interactors that are also associated with the disease. In only 17.8% of disease-protein associations is this fraction statistically significant ($P < 0.05$, permutation test). Moreover, in 46.5% of disease-protein associations, the protein does not interact directly with any

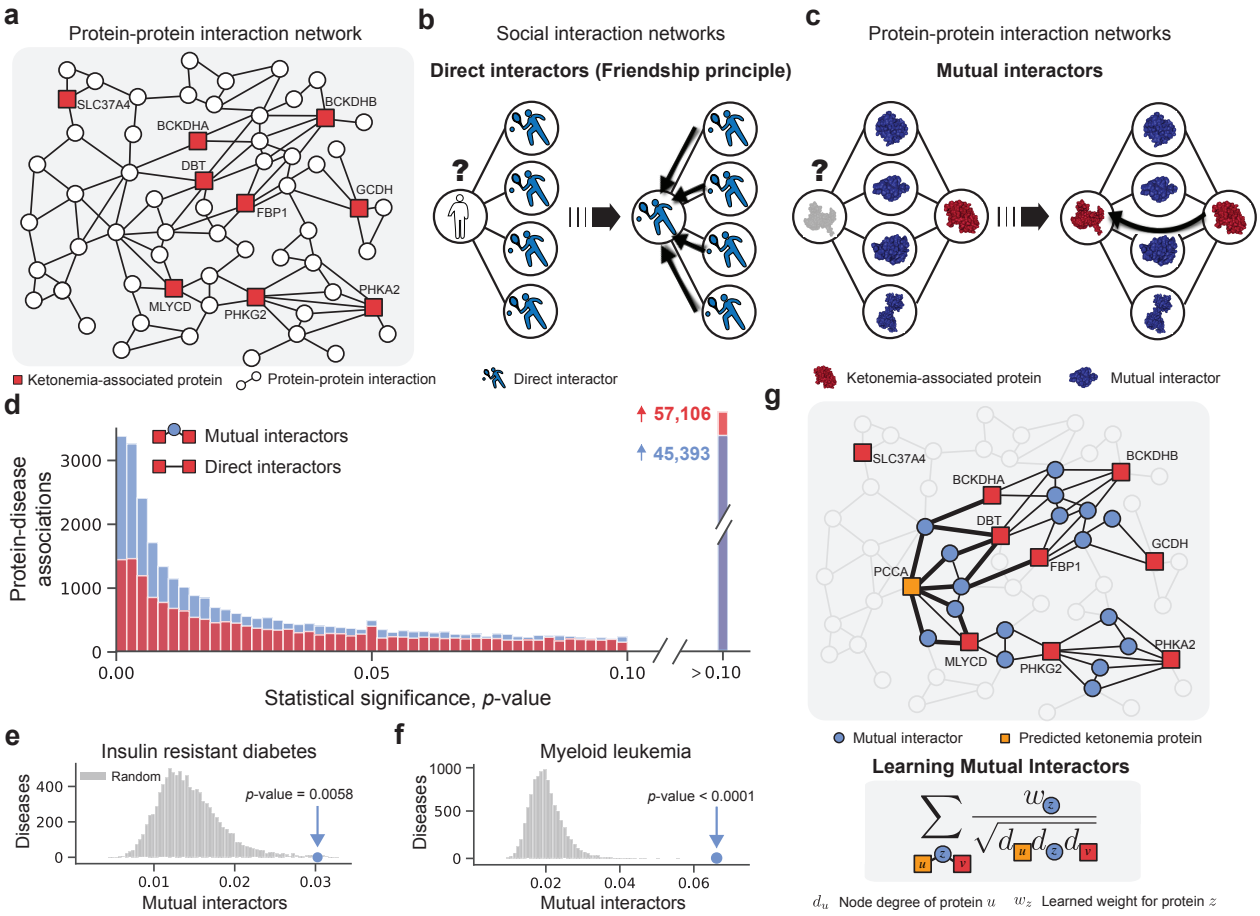


Fig. 1: **The mutual interactor principle. (a)** The human protein-protein interaction (PPI) network with proteins associated with ketonemia highlighted (in red). **(b)** Schematic illustration of the friendship principle (*i.e.*, network homophily[12]) in a social network of five individuals. **(c)** Schematic illustration of the mutual interactor principle in a PPI network. According to the *mutual interactor* principle, the grey protein is likely associated with ketonemia because it interacts with the same proteins as a known ketonemia protein (in red); the two proteins share four mutual interactors (in blue). **(d)** Comparison of mutual interactors and direct interactors 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 (in blue) and the direct interactor score (in red) is computed and plotted for comparison (see Section B.3). We calculate the average mutual interactor score of proteins associated with **(e)** insulin resistant diabetes and **(f)** myeloid leukemia (see Section B.3). **(e-f)** The observed mutual interactor scores (in blue) are significantly larger than random expectation (in grey).

other proteins associated with the same disease. For each disease-protein association, we also compute the degree-normalized count of mutual interactors between the protein and other proteins associated with the disease. We call this the association's *mutual interactor score* (see Section B.3). In 31.0% of disease-protein associations, this score is significant (permutation test, $P < 0.05$). For other molecular phenotypes, we get similar results: proteins targeted by the same drug have a significant direct interactor score 35.1% of the time and a significant mutual interactor score 67.5% of the time (see Figure 3b).[18] In only 31.0% of the protein-function associations in the Gene Ontology is the direct interactor score significant, compared with 56.7% for the mutual interactor score (see Figure D1a).[19] For biological processes in the Gene Ontology, these fractions are 26.7% and 46.3% for the direct and mutual interactor scores, respectively (see Figure D1b). These results suggest that, in biological networks, there is more empirical evidence for the Mutual Interactor principle than there is for the principle of homophily.

## 3. *Mutual Interactors* as a machine learning method for predicting molecular phenotypes

Based on the mutual interactor principle, we develop a machine learning method for inferring associations between molecules and phenotypes. Below, we describe how our method can predict disease-protein associations using the protein-protein interaction network.

In network-based disease protein prediction, the objective is to discover new disease-protein associations by leveraging the network properties of proteins we already know to be involved in the disease. Our method, *Mutual Interactors*, scores candidate disease-protein associations by evaluating the mutual interactors between the candidate protein and other proteins already known to be associated with the disease. Rather than score candidate disease-protein associations according to the raw count of these mutual interactors, our method learns to weight each mutual interactor differently. Intuitively, this makes sense: the significance of a mutual interactor depends on it's profile. For example, that two proteins both interact with the same hub-protein is probably less significant than two proteins both interacting with a low-degree signalling receptor. Rather than hard-code which mutual interactors we deem significant, through training on a large set of disease pathways, *Mutual Interactors* learns which proteins often interact with multiple proteins in the same disease pathway. *Mutual Interactors* maintains a weight $w_z$ for every protein $z$ in the interactome. This allows *Mutual Interactors* to down-weight spurious mutual interactors when evaluating a candidate association.

To further ground our method, we consider its application to a specific disease pathway. Ketonemia is a condition wherein the concentration of ketone bodies in the blood is abnormally high.[20,21] In Figure 1a, we show the Ketonemia pathway in the human protein-protein interaction network. In red are the proteins known to be associated with Ketonemia, including MLYCD and BCKDHA.[22,23] We see that Ketonemia-associated proteins rarely interact with one another. In Figure 1g, we show the same network and disease pathway, but now we've highlighted in blue the mutual interactors between Ketonemia-associated proteins. Of all 21,557 proteins in the human protein-protein interaction network, *Mutual Interactors* predicts that PCCA, shown in orange, is the most likely to be associated with Ketonemia. PCCA is a protein involved in the breakdown of fatty acids, a process which produces ketone bodies as a byproduct. PCCA shares mutual interactors with four proteins known to be associated with Ketonemia: BCKDHA, DBT, FBP1, and MLCYD. Further, two of these mutual interactors, MCEE and PCCB, are of very low degree (with 7 and 21 interactions respectively) and are weighted highly by *Mutual Interactors*.

### 3.1. *Problem Formulation*

Though *Mutual Interactors* was motivated by the molecular phenotype prediction problem, it is a general model that can be applied in any setting where we'd like to group nodes on a graph. Suppose we have a graph $G = \{V, E\}$ and a set of node sets $\mathbf{S} = \{S_1, S_2, ..., S_k\}$ where each set $S_i$ is a subset of the full node set $S_i \subseteq V$. Note that the node sets need not be disjoint. For example, $G$ could be a PPI network and each $S_i$ could be the set of proteins associated with a different phenotype. We can split each node set $S_i$ into a set of training nodes $\tilde{S}_i \subset S_i$ and a set of test nodes $S_i - \tilde{S}_i$. Given $\tilde{S}_i$ and the network $G$, we're interested in uncovering the full set of nodes $S_i$. Formally, this means computing a probability $Pr(u \in S|\tilde{S})$ for each node $u \in V$.

### 3.2. *The Mutual Interactors model*

The mutual interactors of two nodes $u$ and $v$ are given by the set $M_{u,v} = N(u) \cap N(v)$, where $N(u)$ is the set of $u$'s one-hop neighbors. For each node $z \in V$, *Mutual Interactors* maintains a weight $w_z$. As we discussed above, these weights are meant to capture the degree to which each node in the graph acts as a mutual interactor in the node sets of $\mathbf{S}$. With a weight $w_z$ for every possible mutual interactor in the network, we model the probability that a query node $u$ is in a full node set $S$ given the training set $\tilde{S} \subseteq S$ as

$$Pr(u \in S|\tilde{S}) = \sigma\left(a\left(\sum_{v \in \tilde{S}} \frac{1}{\sqrt{d_v d_u}} \sum_{z \in M_{v,u}} \frac{w_z}{\sqrt{d_z}}\right) + b\right) \tag{1}$$

where $d_u$ is the degree of node $u$, $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function, $a$ is a scale parameter, $b$ is a bias parameter, and $w_z$ is a learned weight for node $z$. With sparse matrix multiplication we can efficiently compute the probability for every node in the network with respect to a batch of $k$ training sets $\{\tilde{S}_1, ..., \tilde{S}_k\}$. Let's encode training sets with a binary matrix $\mathbf{X} \in \{0,1\}^{k \times n}$, where $x_{ij} = 1$ if and only if $j \in \tilde{S}_i$. With $\mathbf{X}$, we can efficiently compute the probability matrix $\mathbf{P}$ where $P_{ij} = Pr(j \in S_i|\tilde{S}_i)$ with

$$\mathbf{P} = \sigma(a(\mathbf{X}\mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{W}\mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}) + b) \tag{2}$$

where $\mathbf{A}$ is the adjacency matrix, $\mathbf{D}$ is the diagonal degree matrix and $\mathbf{W}$ is a diagonal matrix with the weights $w_z$ on the diagonal. Note this formulation ignores any edge weights in the graph, future work should explore simple extensions of this formulation that incorporate edge weights.

### 3.3. *Training the Mutual Interactors model*

Given a meta-training set of $k$ node sets $\mathbf{S} = \{S_1, ..., S_k\}$, we can learn the model's weights $\mathbf{W}$, $a$, and $b$ that maximize the likelihood of observing the node sets in the meta-training set. During meta-training we simulate node set expansion by splitting each set $S_i$ into a training set $\tilde{S}_i$ encoded by $\mathbf{X} \in \{0,1\}^{m \times n}$ and a target set $S_i - \tilde{S}_i$ encoded by $\mathbf{Y} \in \{0,1\}^{m \times n}$. For each epoch, we randomly sample 90% of associations for the training set and use the remaining 10% for the test set. The input associations $\mathbf{X}$ are fed through our model to produce association probabilities $\mathbf{P}$. We update model weights by minimizing weighted binary cross-entropy loss

$$\ell(\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^{m} \sum_{j=1}^{n} -[\alpha_p Y_{ij} \log P_{ij} + (1 - Y_{ij}) \log(1 - P_{ij})] \tag{3}$$

where $\alpha_p$ is the weight given to positive examples. Since there are far more positive examples than negative examples, we set $\alpha_p = \frac{\#\text{ negative examples}}{\#\text{ positive examples}}$.

We can minimize the loss using a gradient-based optimizer. First, we compute the gradient of the loss with respect to model parameters via backpropagation. Then, we use ADAM with a learning rate of 1.0. We train *Mutual Interactors* with weight decay $10^{-5}$ and a batch size of 200.[24] We train for five epochs and use $\frac{1}{9}$ of the training labels as a validation set for early stopping.

## 4. Predicting disease-associated proteins with *Mutual Interactors*

We systematically evaluate our method by simulating disease protein discovery on 1,811 different disease pathways. In ten-fold cross-validation, we find that *Mutual Interactors* recovers a larger frac-

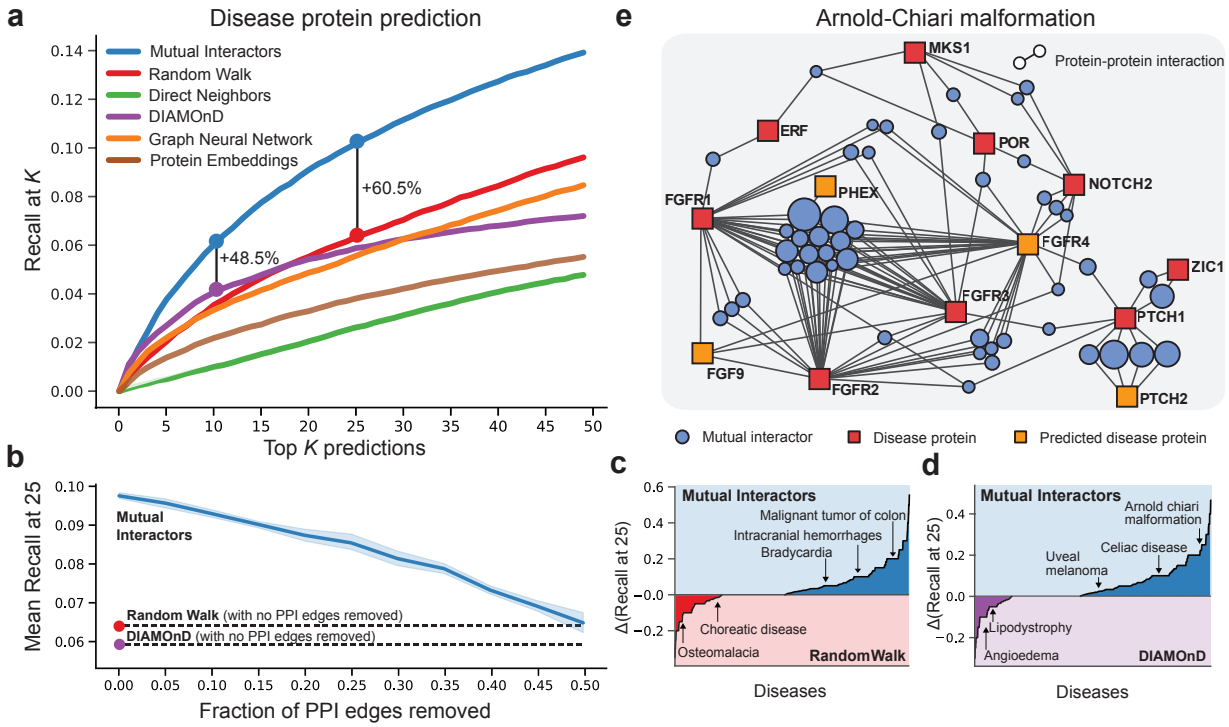


Fig. 2: **Uncovering disease proteins with the mutual interactor principle. (a)** Overall performance evaluation. The plot shows the fraction of disease proteins recovered within the top $k$ predictions for $k = 1$ to $k = 50$ (recall-at-$k$). The dotted lines at $k = 10$ and $k = 25$ show the percent-increase in recall over the next best performing method. **(b)** Effect of data incompleteness on performance. Shown is Mutual Interactors' recall-at-25 as a function of the fraction of protein-protein interactions randomly removed from the network. Dotted lines indicate performance of random walks and DIAMOnD on a full PPI network with no PPIs removed. **(c-d)** Comparison of Mutual Interactors and baseline methods across diseases. For each disease in our dataset (x-axis), we plot the difference in recall-at-25 (y-axis) between Mutual Interactors and two baseline methods: (**c**) random walks, (**d**) DIAMOnD.[25] **(f)** Comparison of the degree-normalized Mutual Interactor weights of drug targets and non-targets. Shown is the distribution of degree-normalized Mutual Interactor weights for 2,212 drug targets[18] (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). **(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, $w_z$.

tion of held-out proteins than do existing disease protein discovery methods. Specifically, for 10.2% of disease-protein associations our method ranks the held-out protein within the first 25 proteins in the network (recall-at-25 = 0.102). *Mutual Interactors*'s performance represents an improvement of 60.9% in recall-at-25 over the next best performing method, random walks. Other network-based methods of disease protein discovery including DIAMOnD[10] (recall-at-25= 0.059), random walks[26] (recall-at-25 = 0.063), and graph convolutional neural networks[25] (recall-at-25 = 0.057) recover considerably fewer disease-protein associations (see Figure 2a,c-d). Moreover, *Mutual Interactors* maintains its advantage over existing methods across disease categories: in all seventeen that we considered *Mutual Interactors*'s mean recall-at-100 exceeds random walks' (see Section C.3 and Figure C3). We also study whether *Mutual Interactors* can generalize to a new disease that is unrelated to the diseases it was trained on. To do so, we train *Mutual Interactors* in the more challenging setting where similar diseases are kept from straddling the train-test divide (see Section C.2 and Figure C2). In this setting, *Mutual Interactors* achieves a recall-at-25 of 0.096, a 50.7% increase in performance over the next best method, random walks. *Mutual Interactors* can naturally be extended to incorporate other sources of protein data.[27] In Section C.4, we describe a parametric *Mutual Interactors* model that incorporates functional profiles from the Gene Ontology when evaluating mutual interactors. Instead of learning a weight $w_z$ for every protein $z$, this model learns one scalar-valued

function mapping gene ontology embeddings to mutual interactor weights. We show that parametric *Mutual Interactors* performs on par with the original *Mutual Interactors* model, outperforming baseline methods by $45.5\%$ in recall-at-25 (see Figure C4).

The experimental data we use to construct molecular interaction networks is often incomplete or noisy: it is estimated that state-of-the-art interactomes are missing $80\%$ of all the interactions in human cells.[16] In light of this, we test if our method is tolerant of data incomplete networks. We find that *Mutual Interactors* exhibits stable performance up to the removal of 50% of known PPI interactions. *Mutual Interactors*'s performance with only half of all known interactions exceeds the performance of existing methods that use all known interactions (Figure 2b).

We perform an ablation study to assess the benefits of meta-learning mutual interactor weights $w_z$ (see Figure D8 ). In the study, we compare our model with *Constant Mutual Interactors* where $w_z = 1 \ \forall z$. On tasks for which we have a large dataset of phenotypes (*i.e.* disease protein prediction and molecular function prediction in humans), meta-learning $w_z$ improves performance by up to $16.6\%$ in recall-at-25. However, on tasks for which data is scarce (i.e. drug-target prediction and non-human molecular function prediction) learning $w_z$ does not provide a significant benefit. For these tasks, we report performance on *constant Mutual Interactors* where $w_z = 1 \ \forall z$.

Learned weights provide insight into the function and druggability of mutual interactors. Next we analyze the mutual interactor weights learned by our method. Recall that *Mutual Interactors* learns a weight $w_z$ for every protein $z$ in the interactome. This allows *Mutual Interactors* to down-weight spurious mutual interactors when evaluating a candidate disease-protein association. Here, we study what insights into biological mechanisms these weights reveal. We find that normalized *Mutual Interactors* weight $\frac{w_z}{\sqrt{d_z}}$ is correlated with neither degree ($r = 0.0359$) nor triangle clustering coefficient ($r = 0.0127$) (see Figure D9). However, we do find that proteins with high weight are often involved in cell-cell signaling. We perform a functional enrichment analysis on the $75$ proteins with the highest normalized weight $\frac{w_z}{\sqrt{d_z}}$. Of the fifteen functional classes most enriched in these proteins, ten including *signaling receptor activity* and *cell surface receptor signaling pathway* are directly related to transmembrane signaling and the other five including *plasma membrane part* are tangentially related to signaling (see Figure D6). Further, we find that highly-weighted proteins are often targeted by drugs. Among the 500 proteins with the highest degree-normalized weight, $33.6\%$ are targeted by a drug in the DrugBank database.[18] By contrast only $10.9\%$ of proteins in the wider protein-protein interaction network are targeted by those drugs. This represents a significant increase ($p \leq 6.43 \times 10^{-24}$, Kolmogorov-Smirnov test). Although no drug-target interaction data was used, training our method to predict disease proteins gives us insights into which proteins are druggable.

## 5. Identifying drug targets with *Mutual Interactors*

The development of methods that can identify drug targets is an important area of research,[30–33] in this section we show how our method can also be used for this task. Recall that mutual interactors between proteins targeted by the same drug are statistically overrepresented in the protein-protein interaction network (see Figure 3a). Like with disease-protein associations, *Mutual Interactors* can score candidate drug-target interactions by evaluating the mutual interactors between the candidate target protein and other proteins already known to be targeted by the drug (see Section 3.1 for a technical description of the approach). When we simulate drug-target identification with ten-fold cross

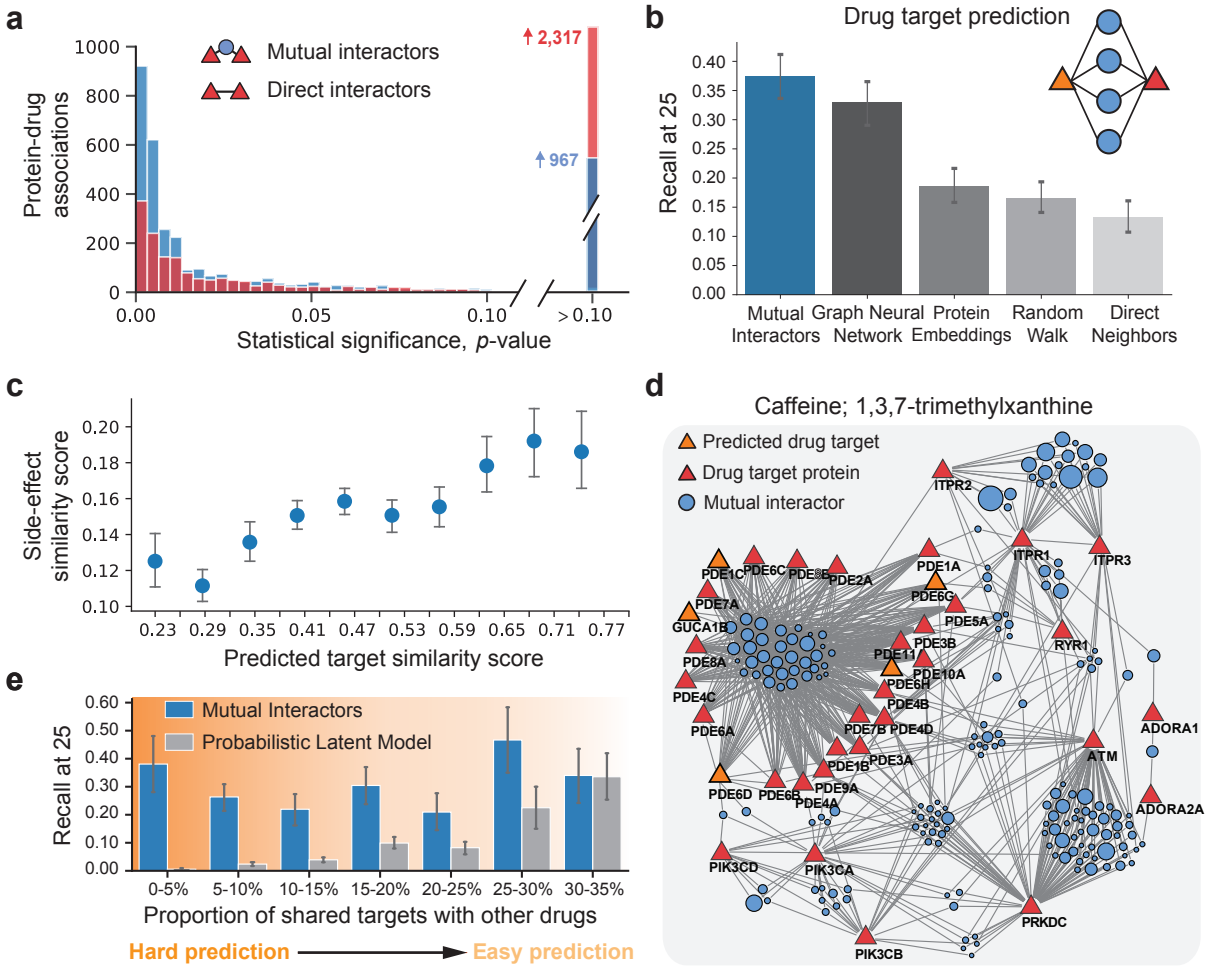


Fig. 3: **Identifying drug targets using the principle of mutual interactors.** Comparison of Mutual Interactors (in blue) and direct interactors (in red) as principles of drug-target connectivity in a human PPI network. For 4,403 drug-target associations,[28] the statistical significance ($p$-value) of the mutual interactor score (in blue) and the direct interactor score (in red) is computed and plotted for comparison (see Section B.3). **(a) (b)** Drug target identification. Shown is mean recall-at-25 across 190 drugs. **(c)** The side-effect similarity of drugs[29] (y-axis) is linearly related to the similarity of Mutual Interactors' predictions for those drugs (x-axis). **(d)** Mutual Interactors neighborhood for proteins targeted by Caffeine. The neighborhood includes caffeine-targeted 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 Interactors weight, $w_z$ (see 3.1). **(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.

validation on the drugs and targets in the DrugBank database,[18] we find that our method outperforms existing network-based methods of drug-target identification (recall-at-25=0.374), including graph neural networks (recall-at-25=0.329) and random walks (recall-at-25=0.166). We also compare *Mutual Interactors* with probabilistic non-negative matrix factorization (NMF).[30–32] On aggregate, our method's performance is comparable to NMF's. However, on the hardest examples, drugs that share few targets with the drugs in the training set, our method (recall-at-25=0.381) significantly outperforms NMF (recall-at-25=0.006) (see Figure 3e). Further, our method provides insight into the side-effects caused by off-target binding. For each drug in DrugBank, we use *Mutual Interactors* to identify potential protein targets that are not already known targets of the drug. Pairs of drugs for which our method makes similar target predictions tend to have similar side effects[34–37] (Figure 3c).

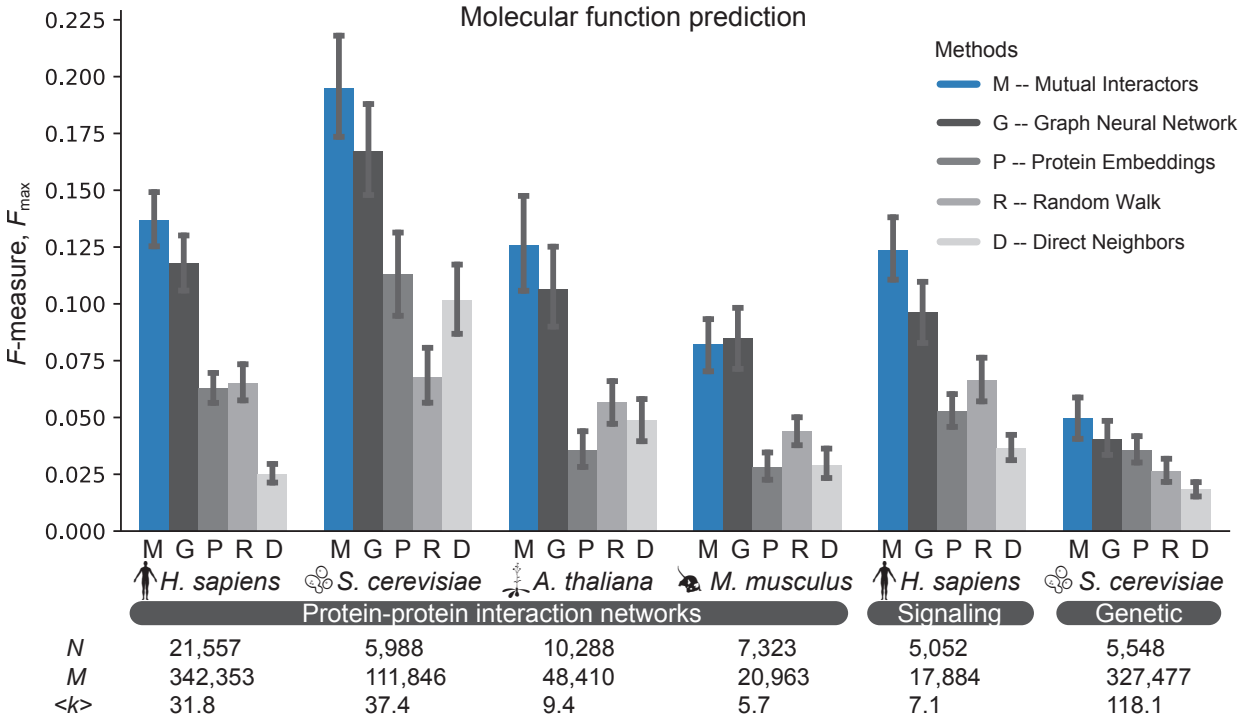


Fig. 4: **Predicting protein functions across species and molecular networks using mutual interactors.** Overall protein function prediction performance across four species and six molecular networks. We predict Molecular Function Ontology[38] terms using PPI, signaling, and genetic interaction networks for human, yeast *S. cerevisiae*, mouse *M. musculus*, and thale cress *A. thaliana*. We show average maximum $F$-measure.[39] A perfect predictor would be characterized by $F_{max} = 1$. Confidence intervals (95%) were determined using bootstrapping with $n = 1{,}000$ iterations. $N$ – number of nodes, $M$ – number of edges, $<k>$ – average node degree.

## 6. Predicting molecular function across species and molecular networks

Molecules associated with the same molecular function (e.g., RNA polymerase I activity) or involved in the same biological process (e.g., nucleosome mobilization) tend to share mutual interactors in molecular networks of various type and species (see Figure D1a-b). For example, the eleven proteins involved in the formation of the secondary messenger cAMP (cyclase activity, GO:0009975) do not interact directly with one another in the protein-protein interaction network, but almost all of them interact with the same group of twenty-five mutual interactors (see Figure D3). Using the Mutual Interactor principle, we can predict the molecular functions and biological processes of molecules. Via ten-fold cross validation, we compare *Mutual Interactors* to existing methods of molecular function prediction, including Graph Neural Networks[40] and Random Walks.[26] Across all four species and in three different molecular networks (protein-protein interaction, signaling, and genetic interaction), we find that *Mutual Interactors* is the strongest predictor of both molecular function (see Figure 4) and biological process (see Figure D2).

## 7. Conclusion

This work demonstrates the importance of rooting biomedical network science methods in principles that are empirically validated in biological data, rather than borrowed from other domains. This need for more domain-specific methodology in biomedical network science is also demonstrated by Kovács *et al.*, who find that social network principles do not apply for link prediction in PPI

networks.[41] This study complements these findings: with experiments across three different kinds of molecular networks (protein-protein interaction, signaling and genetic interaction), and four species (*H. sapiens*, *S. cerevisiae*, *A. thaliana*, *M. musculus*) we show that a method designed specifically for biological data can better predict disease-protein associations, drug-target interactions and molecular function than can general methods of greater complexity. The power of *Mutual Interactors* to predict molecular phenotypes lies not in it's algorithmic complexity—it outperforms far more involved methods—but rather in the simple, yet fundamental, principle that underpins it. Motivated by our findings that molecules with similar phenotypes tend to share mutual interactors, we formalize the Mutual Interactor principle mathematically with machine learning. *Mutual Interactors* is fast, easy to implement, and robust to incomplete network data—its foundational formulation makes it ripe for extension to new domains and problems.

**Supplementary Material and Code.** Supplementary materials are available online at: `https://cs.stanford.edu/people/sabrieyuboglu/psb-mi.pdf`. Code is available online at: `https://github.com/seyuboglu/milieu`.

## References


1. E. E. Schadt, Molecular networks as sensors and drivers of common human diseases, *Nature* **461**, 218 (September 2009).
2. P. Bork *et al.*, Protein interaction networks from yeast to human, *Current Opinion in Structural Biology* **14**, 292 (June 2004).
3. K. Lage *et al.*, A human phenome-interactome network of protein complexes implicated in genetic disorders, *Nature Biotechnology* **25**, p. 309 (2007).
4. A.-L. Barabási *et al.*, Network medicine: A network-based approach to human disease, *Nature Reviews Genetics* **12**, 56 (January 2011).
5. L. D. Wood *et al.*, The genomic landscapes of human breast and colorectal cancers, *Science* **318**, 1108 (2007).
6. J. Lim *et al.*, A protein–protein interaction network for human inherited ataxias and disorders of purkinje cell degeneration, *Cell* **125**, 801 (2006).
7. J. Chen *et al.*, Detecting functional modules in the yeast protein–protein interaction network, *Bioinformatics* **22**, 2283 (September 2006).
8. X. Wu *et al.*, Network-based global inference of human disease genes, *Molecular Systems Biology* **4**, p. 189 (2008).
9. W. Peng *et al.*, Improving protein function prediction using domain and protein complexes in PPI networks, *BMC Systems Biology* **8**, p. 35 (March 2014).
10. S. D. Ghiassian *et al.*, 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**, p. e1004120 (2015).
11. M. Agrawal *et al.*, Large-scale analysis of disease pathways in the human interactome, *Pacific Symposium on Bicomomputing* **1**, 111 (2018).
12. M. McPherson *et al.*, Birds of a feather: Homophily in social networks, *Annual Review of Sociology* **27**, 415 (2001).
13. K. Luck *et al.*, A reference map of the human binary protein interactome, *Nature* **580**, 402 (April 2020).
14. R. Oughtred *et al.*, The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions, *Protein Science* **30**, 187 (2021).
15. D. Szklarczyk *et al.*, The string database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets, *Nucleic Acids Research* **49**, D605 (2021).

16. J. Menche *et al.*, Uncovering disease-disease relationships through the incomplete interactome, *Science* **347**, p. 1257601 (February 2015).
17. J. Piñero *et al.*, The disgenet knowledge platform for disease genomics: 2019 update, *Nucleic acids research* **48**, D845 (2020).
18. D. S. Wishart *et al.*, DrugBank 5.0: a major update to the DrugBank database for 2018, *Nucleic Acids Research* **46**, D1074 (January 2018).
19. T. G. Ontology, The Gene Ontology Resource: 20 years and still GOing strong, *Nucleic Acids Research* **47**, D330 (January 2019).
20. K. Ennis *et al.*, Hyperglycemia accentuates and ketonemia attenuates hypoglycemia-induced neuronal injury in the developing rat brain, *Pediatric Research* **77**, 84 (2015).
21. O. Arisaka *et al.*, Iron, ketone bodies, and brain development, *The Journal of Pediatrics* **222**, 262 (2020).
22. X. Li *et al.*, Eleven novel mutations of the BCKDHA, BCKDHB and DBT genes associated with maple syrup urine disease in the Chinese population: report on eight cases, *European Journal of Medical Genetics* **58**, 617 (2015).
23. S. H. Lee *et al.*, A Korean child diagnosed with malonic aciduria harboring a novel start codon mutation following presentation with dilated cardiomyopathy, *Molecular Genetics & Genomic Medicine* **8**, p. e1379 (2020).
24. D. P. Kingma *et al.*, Adam: A Method for Stochastic Optimization, *arXiv:1412.6980 [cs]* (December 2014), arXiv: 1412.6980.
25. T. N. Kipf *et al.*, Semi-supervised classification with graph convolutional networks, *International Conference on Learning Representations* (2017).
26. *Analysis of protein-protein interaction networks using random walks* 2005.
27. M. Zitnik *et al.*, Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities, *Information Fusion* **50**, 71 (2019).
28. D. S. Wishart *et al.*, Drugbank 5.0: a major update to the drugbank database for 2018, *Nucleic Acids Research* **46**, D1074 (2017).
29. N. P. Tatonetti *et al.*, Data-driven prediction of drug effects and interactions, *Science Translational Medicine* **4**, 125ra31 (2012).
30. *Probabilistic matrix factorization* 2008.
31. D. D. Lee *et al.*, Learning the parts of objects by non-negative matrix factorization, *Nature* **401**, 788 (October 1999).
32. *Structured non-negative matrix factorization with sparsity patterns* October 2008.
33. M. J. Keiser, B. L. Roth, B. N. Armbruster, P. Ernsberger, J. J. Irwin and B. K. Shoichet, Relating protein pharmacology by ligand chemistry, *Nature biotechnology* **25**, 197 (2007).
34. M. Campillos *et al.*, Drug Target Identification Using Side-Effect Similarity, *Science* **321**, 263 (July 2008).
35. R. Santos *et al.*, A comprehensive map of molecular drug targets, *Nature Reviews Drug Discovery* **16**, 19 (2017).
36. L. H. Calabrese *et al.*, Rheumatic immune-related adverse events from cancer immunotherapy, *Nature Reviews Rheumatology* **14**, 569 (2018).
37. F. Cheng *et al.*, Network-based prediction of drug combinations, *Nature Communications* **10**, 1 (2019).
38. M. Ashburner *et al.*, Gene Ontology: tool for the unification of biology, *Nature Genetics* **25**, p. 25 (2000).
39. P. Radivojac *et al.*, A large-scale evaluation of computational protein function prediction, *Nature Methods* **10**, p. 221 (2013).
40. T. N. Kipf *et al.*, Semi-Supervised Classification with Graph Convolutional Networks (September 2016).
41. I. A. Kovács *et al.*, Network-based prediction of protein interactions, *Nature communications* **10**, 1 (2019).
42. V. Matys *et al.*, TRANSFAC ® : transcriptional regulation, from patterns to profiles, *Nucleic Acids Research* **31**, 374 (January 2003).

43. A. Ceol *et al.*, MINT, the molecular interaction database: 2009 update, *Nucleic Acids Research* **38**, D532 (January 2010).
44. B. Aranda *et al.*, The IntAct molecular interaction database in 2010, *Nucleic Acids Research* **38**, D525 (January 2010).
45. M. Giurgiu *et al.*, Corum: the comprehensive resource of mammalian protein complexes—2019, *Nucleic acids research* **47**, D559 (2019).
46. M. Costanzo *et al.*, The Genetic Landscape of a Cell, *Science* **327**, 425 (January 2010).
47. M. Costanzo *et al.*, A global genetic interaction network maps a wiring diagram of cellular function, *Science* **353**, p. aaf1420 (September 2016).
48. D. Türei *et al.*, OmniPath: guidelines and gateway for literature-curated signaling pathway resources, *Nature Methods* **13**, 966 (December 2016).
49. S. Choobdar *et al.*, Assessment of network module identification across complex diseases, *Nature Methods* **16**, 843 (2019).
50. J. Piñero *et al.*, DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes., *Database : the journal of biological databases and curation* **2015**, bav028 (2015).
51. A. P. Davis *et al.*, The Comparative Toxicogenomics Database: update 2013, *Nucleic Acids Research* **41**, D1104 (January 2013).
52. UniProt Consortium, Activities at the Universal Protein Resource (UniProt), *Nucleic Acids Research* **42**, D191 (January 2014).
53. W. A. Kibbe *et al.*, Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data., *Nucleic acids research* **43**, D1071 (January 2015).
54. S. Navlakha *et al.*, The power of protein interaction networks for associating genes with diseases, *Bioinformatics* **26**, 1057 (2010).
55. S. Kohler *et al.*, Walking the Interactome for Prioritization of Candidate Disease Genes, *The American Journal of Human Genetics* **82**, 949 (April 2008).
56. A. Grover *et al.*, node2vec: Scalable Feature Learning for Networks (July 2016).
57. J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan and Q. Mei, Line: Large-scale information network embedding, in *Proceedings of the 24th international conference on world wide web*, 2015.
58. B. Perozzi, R. Al-Rfou and S. Skiena, Deepwalk: Online learning of social representations, in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014.
59. S. D. Ghiassian *et al.*, 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 (April 2015).
60. M. Zitnik *et al.*, NIMFA: A Python Library for Nonnegative Matrix Factorization, *Journal of Machine Learning Research* , p. 5 (2012).
61. E. Guney *et al.*, Network-based in silico drug efficacy screening, *Nature Communications* **7**, 10331 (February 2016).
62. D. V. Klopfenstein *et al.*, GOATOOLS: A Python library for Gene Ontology analyses, *Scientific Reports* **8**, p. 10872 (July 2018).
63. A. Chatr-Aryamontri *et al.*, The BioGRID interaction database: 2015 update., *Nucleic acids research* **43**, D470 (January 2015).
64. D. Szklarczyk *et al.*, STRING v10: protein-protein interaction networks, integrated over the tree of life., *Nucleic acids research* **43**, D447 (January 2015).
65. L. Cowen *et al.*, Network propagation: a universal amplifier of genetic associations, *Nature Reviews Genetics* **18**, p. 551 (2017).
66. A.-L. Barabási *et al.*, Network medicine: a network-based approach to human disease., *Nature reviews. Genetics* **12**, 56 (January 2011).