OXFORD

Briefings in Bioinformatics, 22(5), 2021, 1–11

https://doi.org/10.1093/bib/bbab066
Method Review

On the limits of active module identification

Olga Lazareva, Jan Baumbach, Markus List† and David B. Blumenthal†

Corresponding author: David B. Blumenthal, Chair of Experimental Bioinformatics, Technical University of Munich, Maximus-von-Imhof-Forum 3, 85354 Freising, Germany, phone: +49 8161 71 2712; E-mail: david.blumenthal@wzw.tum.de
†Joint senior authors.

Abstract

In network and systems medicine, active module identification methods (AMIMs) are widely used for discovering candidate molecular disease mechanisms. To this end, AMIMs combine network analysis algorithms with molecular profiling data, most commonly, by projecting gene expression data onto generic protein–protein interaction (PPI) networks. Although active module identification has led to various novel insights into complex diseases, there is increasing awareness in the field that the combination of gene expression data and PPI network is problematic because up-to-date PPI networks have a very small diameter and are subject to both technical and literature bias. In this paper, we report the results of an extensive study where we analyzed for the first time whether widely used AMIMs really benefit from using PPI networks. Our results clearly show that, except for the recently proposed AMIM DOMINO, the tested AMIMs do not produce biologically more meaningful candidate disease modules on widely used PPI networks than on random networks with the same node degrees. AMIMs hence mainly learn from the node degrees and mostly fail to exploit the biological knowledge encoded in the edges of the PPI networks. This has far-reaching consequences for the field of active module identification. In particular, we suggest that novel algorithms are needed which overcome the degree bias of most existing AMIMs and/or work with customized, context-specific networks instead of generic PPI networks.

Key words: active module identification; de novo network enrichment; network and systems medicine; systems biology

Introduction

Because of massive advances in high-throughput technologies, large amounts of gene expression data have become available over the past decades. This has raised hopes to identify new molecular mechanisms that might provide valuable insights into cellular function and the pathobiology of diseases [1–3]. However, gene expression data tend to be overdetermined and noisy and, as a result, the discovery of disease genes via purely statistical means is often unstable, since the reported genes are often just surrogates of the actual disease genes and hence functionally not necessarily related to the disease of interest [4, 5].

To mitigate these problems, active module identification methods (AMIMs) leverage additional biological knowledge encoded in protein–protein interaction (PPI) networks [6–9]. These methods project gene expression data onto PPI networks and then use network algorithms to identify disease modules consisting of small subnetworks. This dramatically decreases the size of the search space and prioritizes disease modules consisting of functionally related genes, which, in turn, positively affects both stability and functional relevance of the discovered modules [10]. AMIMs have been successfully used for providing novel pathobiological insights into complex diseases such as pulmonary arterial hypertension [11], coronary heart

Olga Lazareva is doctoral fellow at the Bavarian Research Institute for Digital Transformation and a PhD candidate at the Technical University of Munich.
Jan Baumbach is professor and chair of Computational Systems Biology at the University of Hamburg. He obtained his PhD in Computer Science from Bielefeld University.
Markus List obtained his PhD at the University of Southern Denmark and worked as a postdoctoral fellow at the Max Planck Institute for Informatics before starting his group Big Data in BioMedicine at the Technical University of Munich.
David B. Blumenthal is postdoctoral fellow at the Technical University of Munich. He obtained his PhD in Computer Science from the Free University of Bozen-Bolzano.

Submitted: 11 January 2021; **Received (in revised form):** 29 January 2021

disease [12], diabetes mellitus [13], liver fibrosis [14], chronic obstructive pulmonary disease and idiopathic pulmonary fibrosis [15], as well as asthma [16].

Despite these impressive results, there is increasing awareness in the field that the combination of gene expression data and PPI networks is subject to technical and literature bias. PPI networks suffer from technical bias [17] e. g. since the 'bait' proteins used for measuring new interactions often have significantly more interactions. Moreover, literature bias [18], where research focuses on proteins with already known characteristics (e. g. biological function), leads to a strong correlation between the number of studies conducted on a protein and the protein's degree in the PPI network.

The node degree distribution of PPI networks typically follows a power law. As a consequence, perturbances of cellular programs (e. g. via mutations or other mechanisms) typically have a cascading effect when observed on the level of gene expression. As a result, differential gene expression analysis often reveals hundreds or thousands of genes to be disease-associated. By projecting these noisy gene expression data on PPI networks with a small diameter, disease-associated genes can easily be combined into subnetworks or disease modules, most of which may not contain a single disease-causative gene. Although such network modules may be well suited as robust biomarkers for a disease, they may be less suited to pinpoint a disease mechanism.

To account for network-related biases, some recently proposed methods such as Hierarchical HotNet [19] and NetCore [20] integrate data and network randomization steps into their workflows. These permutation-based methods extract subnetworks whose associations with the disease are significantly stronger in the original PPI networks than in the randomized counterparts. Levi et al. further reported that gene ontology enrichment of several state-of-the-art AMIMs on randomly permuted input data produced similar results, questioning the context-specificity of existing AMIMs. To address this issue, Levi et al. [21] propose a new method DOMINO.

Although the effect of random permutations of the input omics data was systematically tested by Levi et al. [21], the question if AMIMs also benefit from the biological knowledge captured in PPI networks remains unanswered (cf. Figure 1). In this study, we close this gap. For this, we developed a test suite for AMIMs, which studies the effect of different types of network randomization on the results. Our test suite, which is openly available at https://github.com/dbblumenthal/amim-test-suite/, expects a network and expression data (or input that can be derived from expression data) as input and produces a set of candidate disease modules as output. These modules are then evaluated using mutual information (MI) and gene set enrichment analysis (GSEA) with known disease signatures (see 'Methods' for details). Since further AMIMs can easily be integrated by implementing a well-defined interface, our test suite can be used not only to reproduce the results reported in this paper, but also to objectively test novel AMIMs with respect to their robustness against network randomization.

In a large-scale empirical evaluation on gene expression data for five different diseases, we ran eight classical and two permutation-based AMIMs on five different widely used PPI networks as well as on randomized counterparts generated by five different random network generators (more than 10 000 runs in total). The most striking result of our analysis is that all except one of the tested AMIMs did not yield significantly more meaningful subnetworks if run on the original PPI networks than if run on random networks with matching node degrees. Most AMIMs hence pick up on the number of interactions a protein is involved in, but do not benefit from the biological knowledge captured in the PPIs themselves.

The remainder of this paper is organized as follows: In the 'Results' section, we briefly describe the protocol implemented by our test suite and present the results of our analyses. In the 'Discussion' section, we discuss the implications of our findings for the field of active module identification. In the 'Methods' section, we provide a more detailed description of our test protocol and also elaborate on how developers of new AMIMs can use our test suite to evaluate their methods.

Results

Test protocol

Figure 2 visualizes our protocols for method evaluation (cf. 'Methods' section for details). We selected eight classical AMIMs, also referred to as *de novo* network enrichment tools in the literature [6] (ClustEx2 [22], COSINE [23], DIAMOnD [24], DOMINO [21], GiGA [25], GXNA [26], KeyPathwayMiner [27–29] and GrandForest [30]) and two permutation-based methods (Hierarchical HotNet [19] and NetCore [20]). Although the classical methods were run with the full protocol (Figure 2A), we used a subset of the protocol for the two permutation-based methods (Figure 2B) since their runtime prohibits large-scale evaluation.

For the full protocol, we compared five widely used PPI networks (BioGRID [31], APID [32, 33], STRING [34], HPRD [35] and IID [36]), as well as gene expression and case/control data for five complex diseases: amyotrophic lateral sclerosis (ALS), non-small cell lung cancer (LC), ulcerative colitis (UC), Chron's disease (CD) and Huntington's disease (HD). For ALS and LC, we had access to survival data that we used for an additional evaluation. Moreover, we used five different random network generators, which produce randomized networks that preserve selected properties of the original PPI networks. For each PPI network, we generated 10 randomized counterparts with each generator. We then ran each classical AMIM on each of the 1275 network-disease pairs.

For each subnetwork produced for a network-disease pair, we measured two dimensions of meaningfulness: Firstly, predictive power quantified as mean MI [37] with (i) the phenotype and (ii) the survival data. Secondly, functional relevance quantified via (i) GSEA [38] w. r. t. Kyoto Encyclopedia of Genes and Genomes (KEGG) [39] pathways associated with the disease of interest and (ii) overlap coefficient w. r. t. disease-associated DisGeNET [40] gene sets. Finally, we used the one-sided Mann–Whitney U-test to assess whether the results obtained for the original PPI networks were significantly better than the results obtained for the randomized counterparts. Note that since AMIMs are intended for discovering yet unknown disease modules, the four meaningfulness scores employed in this paper should not be viewed as direct measures of performance but rather as proxy indicators for biological plausibility of the results.

For the slower permutation-based methods, we employed a restricted protocol using only the smallest PPI network (HPRD), the two smallest gene expression datasets (CD and HD) and the degree preserving network generator REWIRED. We selected this generator, because it produces the randomized networks that are most similar to the original PPI networks. We ran both permutation-based methods on each network-disease pair (in total 22 runs per method) and used the one-sided one-sample t-test to assess whether the subnetworks obtained for the original PPI networks were significantly more meaningful than the

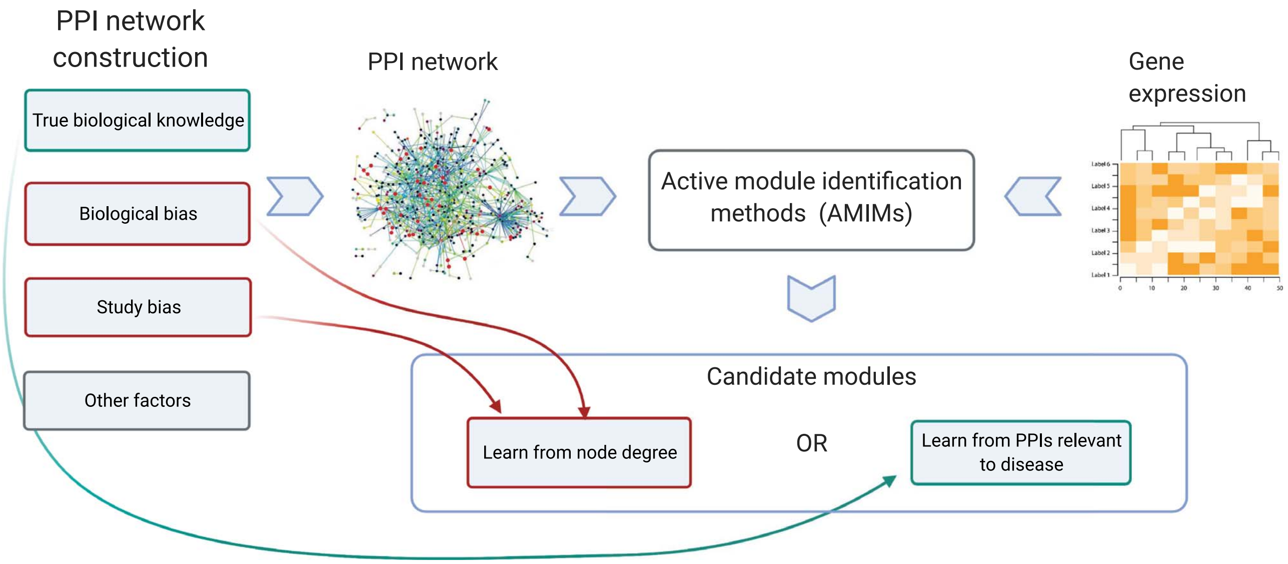

Figure 1. The limitation of AMIMs motivating this study. Since PPI networks suffer from technical and study bias, they usually contain hub-nodes with very high node degrees. In this study, we test the hypothesis whether AMIMs merely learn from the node degrees instead of exploiting the PPIs relevant to the disease of interest.

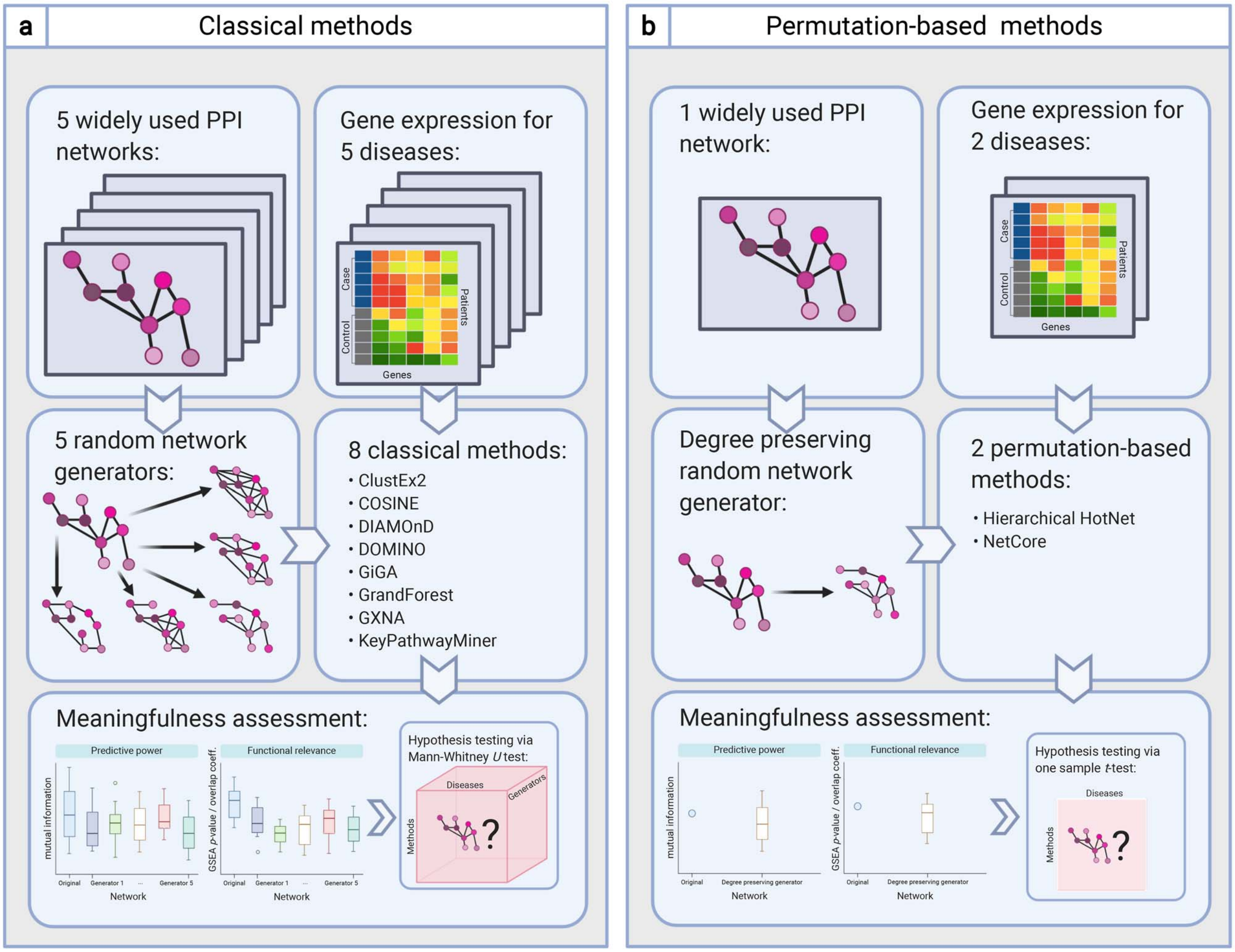
a
Classical methods
5 widely used PPI networks:
Gene expression for 5 diseases:
Case
Control
Patients
Genes
5 random network generators:
8 classical methods:
• ClustEx2
• COSINE
• DIAMOnD
• DOMINO
• GiGA
• GrandForest
• GXNA
• KeyPathwayMiner
Meaningfulness assessment:
Predictive power
Functional relevance
mutual information
GSEA p-value / overlap coeff.
Original
Generator 1
...
Generator 5
Network
Hypothesis testing via Mann-Whitney U test:
Diseases
Generators
Methods
?
b
Permutation-based methods
1 widely used PPI network:
Gene expression for 2 diseases:
Degree preserving random network generator:
2 permutation-based methods:
• Hierarchical HotNet
• NetCore
Degree preserving generator
Hypothesis testing via one sample t-test:

Figure 2. Test protocols employed in this study. (A) Large-scale protocol for classical methods. (B) Restricted protocol for slow permutation-based methods.

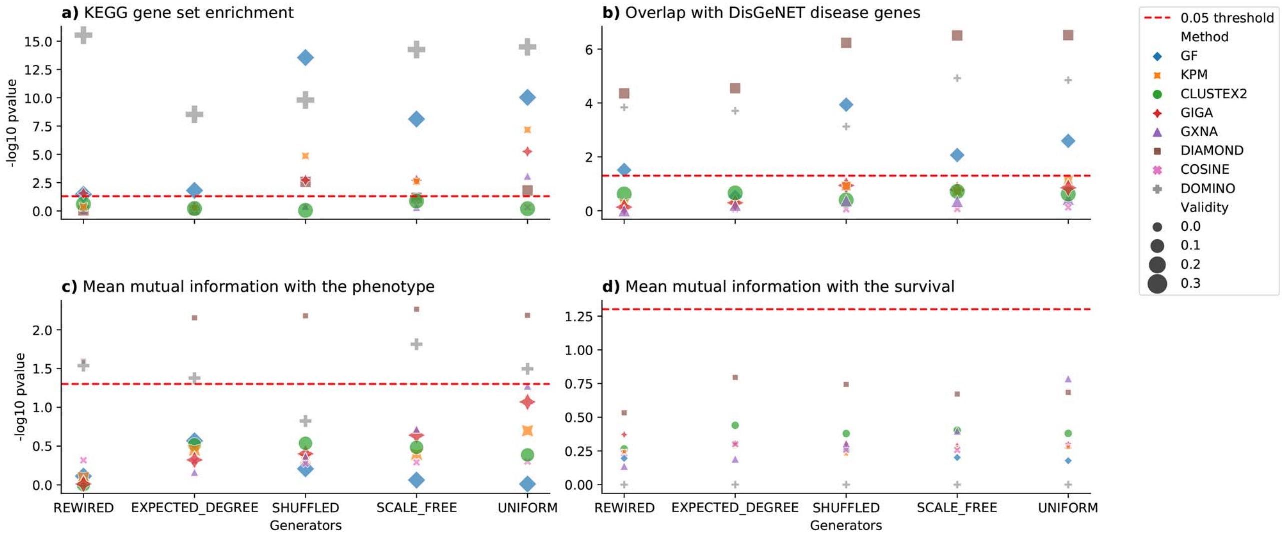

Figure 3. Log-transformed *P*-values for all classical AMIMs and all random network generators computed with the one-sided Mann–Whitney U-test. For each AMIM and each meaningfulness score, we computed a validity score (range from 0 to 1) as the fraction of the original network/condition pairs where the AMIM yielded a score $\geq \tau$ on the original PPI network. For the log-transformed GSEA *P*-values, we employed the cutoff $\tau = -\log_{10} 0.05$; for all other scores, we used the cutoff $\tau = 0.2$. The larger the validity scores, the larger the corresponding semi-transparent shapes.

subnetworks obtained for the random networks with prescribed node degrees.

Results for classical methods

Figure 3 visualizes the *P*-values obtained when comparing the results on the original PPI network to those on randomly generated networks for eight classical AMIMs separately (cf. Supplementary Figure 1 for visualizations of the distributions of the meaningfulness scores).

For the two scores quantifying predictive power (i. e. mean MI w. r. t. phenotype and survival times), we observe that, for most AMIMs, the scores of the candidate disease modules obtained on the original PPI networks are not significantly better than the scores obtained when using random graphs generated by any of the generators. This is the case even for the UNIFORM generator that produces networks that are structurally very different from the original PPI networks. For mean MI w. r. t. survival times (Figure 3D), no AMIM reaches the significance threshold of 0.05. For mean MI w. r. t. the disease phenotypes (Figure 3C), DIAMOnD produces significant results compared with all random network generators but its solution on the original PPI receives a validity score of 0.0 (i. e. there was not a single original network-disease pair for which DIAMOnD computed a candidate module whose mean MI with the phenotype reached 0.2). Notably, DOMINO produced significantly better solutions compared with all random network generators but SHUFFLED. DOMINO results are also slightly more meaningful, as they have a validity score $>$ 0.0. Most of the tested classical AMIMs hence fail to exploit the biological knowledge encoded in generic PPI networks for mining disease modules with high predictive power. DOMINO is the only tool to show potential w. r. t. the phenotype albeit with very low predictive power where the validity score does not exceed 0.1. All tools fail to produce disease modules that are predictive of survival time.

The two scores quantifying functional relevance (GSEA *P*-values w. r. t. disease-associated KEGG pathways and overlap coefficients w. r. t. disease-associated DisGeNET gene sets) present a different picture. Here, we observe that most methods produce significantly more meaningful results on the original network compared with the SHUFFLED, SCALE_FREE and UNIFORM generators. However, when compared with structurally similar networks generated by the REWIRED and the EXPECTED_DEGREE generators, only DOMINO shows good performance. For KEGG gene set enrichment (Figure 3A), GrandForest and DOMINO reach the significance threshold, whereas DIAMOnD and DOMINO do so for DisGeNET enrichment (Figure 3B) when compared with the two degree-preserving generators REWIRED and EXPECTED_DEGREE. Notably, DOMINO is the only tool to produce very significant results on degree-preserving random network generators. However, the validity scores are low in all cases and never exceed 0.3. Our results hence indicate that although most AMIMs are guided toward functionally relevant disease modules, the interactions themselves seem to be largely irrelevant.

To evaluate the effect of the five original PPI networks and the gene expression datasets for the five diseases, we also split the results along the PPI network dimension and along the disease dimension (cf. Supplementary Figures 2 and 3 for visualizations of the distributions of the meaningfulness scores). Figures 4 and 5 visualize the obtained *P*-values. These results suggest that HPRD is the best performing network in terms of KEGG gene set enrichment and DisGeNET overlap. This finding may be explained by the fact that HPRD is the smallest and least frequently updated network and contains mostly well-studied proteins that are more likely to overlap with KEGG pathways or DisGeNET genes.

We also observe that, in terms of functional relevance (especially DisGeNET overlap), the results for the CD dataset were much better than for the other datasets. This may be due to the fact that inflammation is a well-understood process and the DisGeNET disease gene annotation for CD is therefore better suited compared with other diseases. Note that the same argument does not apply to the UC dataset, since DisGeNET only reports on the more general inflammatory bowel disease as a proxy (cf. 'Methods' for details).

The results reported above suggest that, except for DOMINO, the tested AMIMs largely learn from the degree distributions rather than exploiting the biological knowledge encoded in the

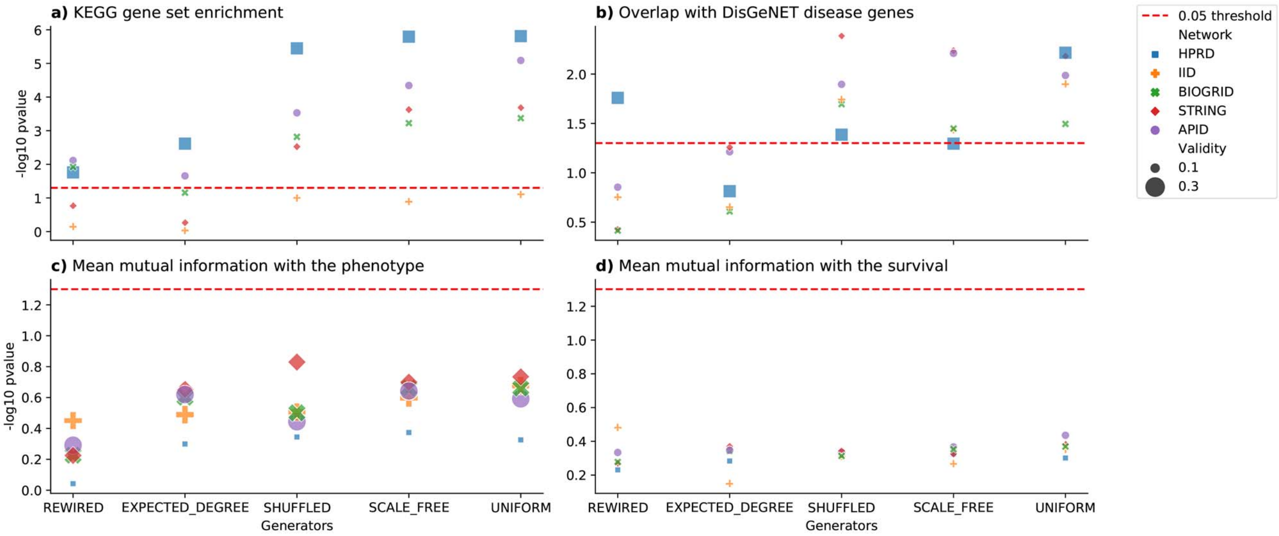

Figure 4. Log-transformed P-values for all PPI networks and all random network generators computed with the one-sided Mann–Whitney U-test.

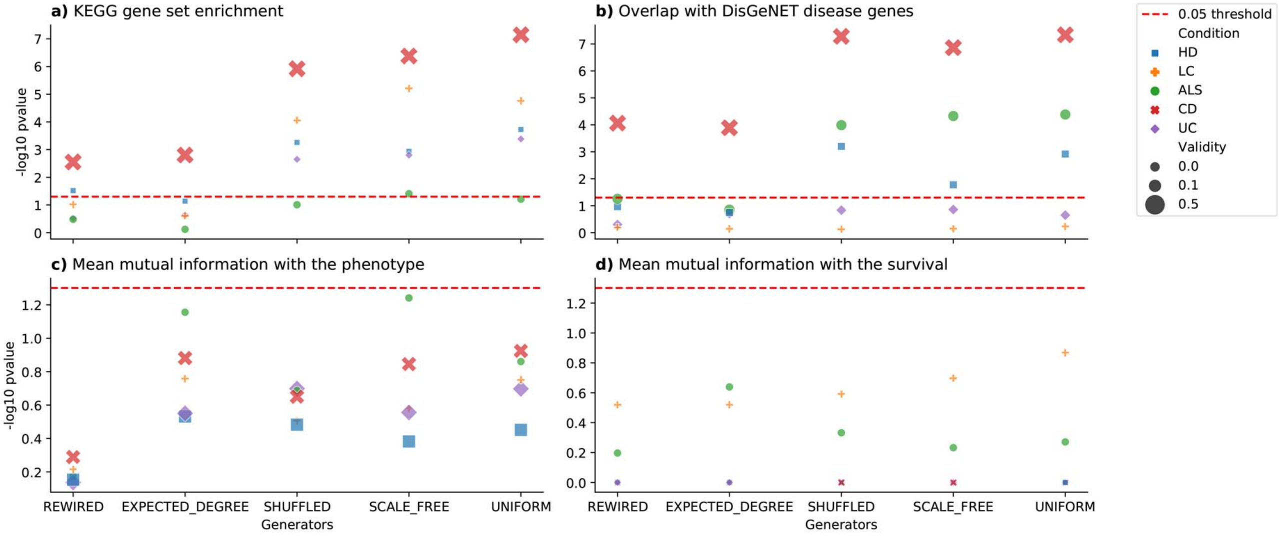

Figure 5. Log-transformed P-values for all diseases and all random network generators computed with the one-sided Mann–Whitney U-test.

interactions themselves. Figure 6 shows the outcomes of further analyses we carried out to find possible explanations for these results. The first interesting finding is that the topologies of the active modules DOMINO and COSINE computed on the original PPI networks are different from the topologies of the other AMIMs' modules (Figure 6A): DOMINO and COSINE's modules tend to have larger maximum pairwise distances, i. e. they tend to include fewer hub-nodes that would ensure a high connectivity. This is reflected by the fact that the mean degrees of the result sets and the two scores quantifying functional relevance are less strongly correlated for DOMINO and COSINE than for the other AMIMs (Figure 6E). These observations indicate that DOMINO and COSINE are less influenced by the node degrees than the other AMIMs. Although we expected this finding for DOMINO, it is somewhat surprising for COSINE. One possible explanation is that COSINE performed poorly even on the original PPI networks and hence neither learned from the node degrees nor from the interactions effectively.

We also observe several global trends in the results of the full protocol, which indicate that when aggregating across all tested AMIMs, the degrees on the genes contained in the result sets are predictive of KEGG gene set enrichment P-value and DisGeNET overlap: Firstly, the mean degrees drop very significantly only for the SHUFFLED, the SCALE_FREE and the UNIFORM generators (Figure 6B). This reflects the results visualized in Figures 3–5, where significant drops in performance compared with the original PPI networks where observed mostly for these generators. Secondly, both the negative log-transformed KEGG gene set enrichment P-value and the DisGeNET overlap coefficient increase with increasing mean degrees (Figure 6C and D). Thirdly, we observe a very strong global correlation between the Mann–Whitney U-test P-values for the mean degrees, on the one side, and for two measures quantifying functional relevance, on the other side (last column in heat map in Figure 6E).

Results for permutation-based methods

Figure 7 shows the results for the two permutation-based AMIMs NetCore and Hierarchical HotNet (cf. Supplementary Figure 4 for visualizations of the distributions of the meaningfulness scores).

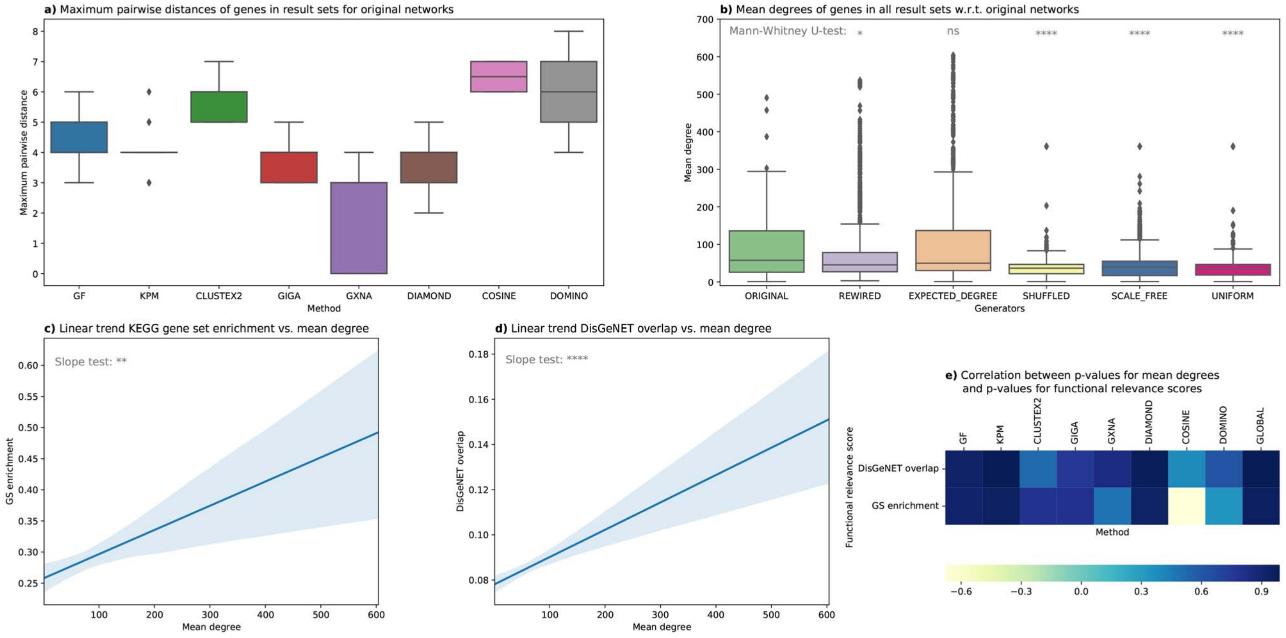

Figure 6. Detailed analyses explaining the results for the functional relevance scores. (**A**) Maximum pairwise distances of genes contained in result sets for original PPI networks for each AMIM. (**B**) Mean degrees in original PPI networks of genes contained in result sets for each generator. (**C**) Linear trend of KEGG gene set enrichment *P*-values versus mean degrees in original PPI networks aggregated across all generators and AMIMs. (**D**) Linear trend of DisGeNET overlap coefficient versus mean degrees in original PPI networks aggregated across all generators and AMIMs. (**E**) AMIM-specific and global correlation coefficients between Mann–Whitney U-test *P*-values for mean degrees in original PPI networks, on the one side, and the two functional relevance scores, on the other side (cf. 'Methods' for details).

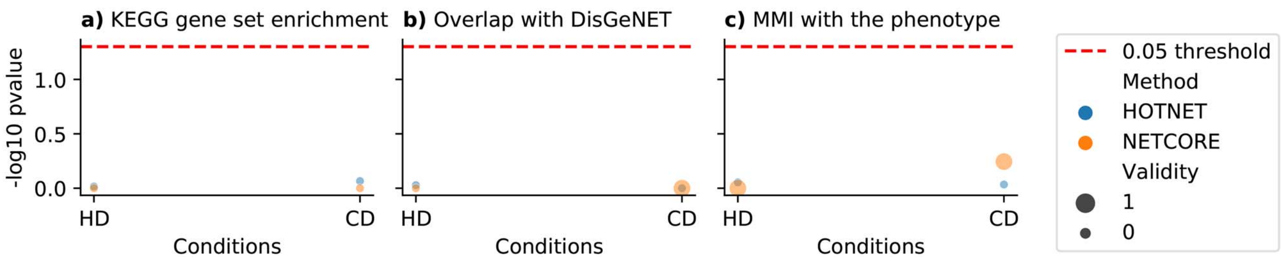

Figure 7. Log-transformed *P*-values for permutation-based AMIMs computed with the one-sided one-sample *t*-test. Mean MI w. r. t. survival times is not reported, because no survival data are available for the HD and CD datasets employed by the restricted protocol. The validity scores are binary here, because the restricted protocol uses only one original PPI network and the *P*-values are computed separately for the two diseases.

Recall that, because of their high computational costs, these methods were run with the restricted protocol visualized in Figure 2B, which only uses one PPI network (HPRD), two diseases (CD and HD), and one random network generator (REWIRED). Surprisingly, the results for the permutation-based methods are not better than the results for the classical AMIMs reported in the previous subsection: Both NetCore and Hierarchical HotNet clearly fail to reach the significance threshold of 0.05 for all three meaningfulness scores.

Discussion

It is commonly believed that prior biological knowledge captured in PPI networks can be leveraged for extracting functionally and mechanistically interpretable disease modules by AMIMs. However, an open question in the field is what characteristic of a PPI network makes these methods successful. Since PPI networks are known to suffer from a considerably node degree bias, we hypothesize here that AMIMs may use the node degree as prior information rather than the connectivity of the network i. e. the biological knowledge captured in the interactions themselves. To test this hypothesis, we compared 10 state-of-the-art AMIMs on original as well as randomly generated networks. Although a few methods produced meaningful results w. r. t. functional enrichment, none of the methods produced disease modules with appreciable predictive power w. r. t. to both phenotype as well as survival. This demonstrates that results of AMIMs are not directly suited for such tasks without further refinement through e. g. supervised machine learning as shown by Alcaraz et al., where disease modules were used successfully as features for disease subtyping in a random forest classifier [10].

To investigate which network properties are exploited by AMIMs, we compared the results against different types of random network generators. Our results clearly show that most methods do not yield more meaningful candidate disease modules on randomized networks if these are constructed such that the (expected) node degrees match the node degrees of the original networks. Unexpectedly, permutation-based methods that include steps to correct for PPI network characteristics in their workflow did not produce more meaningful results on the original PPIs.

Only one tested tool benefits from the PPI networks

The only tool to produce more meaningful results on the original network was the recently proposed method DOMINO [21]. Interestingly, DOMINO's development was motivated by the observations that existing methods are not sensitive to permutations of the input data. Our finding that most existing methods are also not sensitive to network randomization suggests that these two issues are related. Considering the small diameter of a PPI network, AMIMs sensitive to high-degree nodes are likely to produce subnetworks that are enriched for similar biological functions. Given this, we see several possibilities to advance the field, namely (i) further algorithmic improvements to overcome PPI network biases, (ii) the integrative use of complementary omics data that increase the signal to noise ratio, which is inherently low in gene expression data and (iii) the use of more fine-grained tissue-specific, condition-specific or even personalized networks.

Further algorithmic improvements are needed

The encouraging results of DOMINO indicate that algorithmic improvements to overcome the node bias of PPI networks are possible. From an algorithmic point of view, DOMINO differs from all other tested AMIMs in that it discards some of the disease-associated genes in a partially unsupervised manner. We hypothesize that this is the key to DOMINO's success, because it makes hub-genes other AMIMs include into their modules to connect the disease-associated genes less attractive for DOMINO (cf. Figure 6A). Consequently, we expect algorithmically improved AMIMs to be either partially unsupervised such as DOMINO or even fully unsupervised [41]. Importantly, newly developed AMIMs need to be tested with respect to their sensitivity to network randomization. One way to do this systematically is to evaluate them with the test suite presented in this paper.

Different types of omics data and context-specific networks should be considered

For this study, we evaluated the performances of AMIMs when run on gene expression data and PPI networks, which is currently the most common use case. Consequently, our findings are restricted to this use and cannot be generalized to other types of omics data and biological networks. In fact, we expect that using different types of omics data and context-specific networks introduces new opportunities for AMIM users and developers. For instance, promising directions for future research include using microbiome data in combination with metabolic networks, using DNA methylation data with gene regulatory networks [42], or inferring condition- or tissue-specific gene regulatory networks from expression data [43]. Next-generation AMIMs might even integrate the inference of context-specific networks with disease module mining. Although all of these strategies come with their own challenges and limitations, we believe that they could help to overcome some of the biases of PPI networks (especially, the literature bias).

Quantitative measures of functional relevance need to be used carefully

The most widely used method for quantitatively assessing the functional relevance of candidate disease modules is to compare them against known disease-associated genes. In fact, we also follow this strategy in our test suite (recall that we use KEGG GSEA *P*-value and DisGeNET overlap as our measures of functional relevance). However, this approach is severely limited and biased by our current knowledge. In particular in the light of the results shown here, it must hence always be kept in mind that such quantitative measures are at best proxy indicators for functional relevance. Alternatively, the simulation of synthetic gold standard datasets could be considered, but this approach is limited by our understanding and assumptions on network and disease module characteristics [6].

Cross-disciplinary research is key to success

Since quantitative measures of functional relevance are biased, it is unlikely that simply reporting on disease modules will yield novel insights into complex diseases. Interestingly, studies that report successful applications of active module identification are usually co-authored by cross-disciplinary teams of researchers that include not only bioinformaticians but also domain experts for the disease of interest [11–16]. We argue that this is no coincidence and promote cross-disciplinary research. To this end, AMIM developers should follow best practices for developing usable software [44], allowing domain experts without a background in computer science to run the tools on their data and to leverage their domain knowledge in the interpretation of the results. Ideally, such interfaces should follow the expert-in-the-loop paradigm and provide functionality for all three steps of active module identification (data integration, network construction, disease module mining). To the best of our knowledge, such an integrated active module identification platform is available only for COVID-19 [45].

Conclusions

A plethora of tools for identifying disease modules via the integration of gene expression data and PPI networks have been developed over the years. Here, we could show conclusively that most AMIMs do not produce more meaningful results on the original compared with randomized PPI networks in which the (expected) node degrees do not change. Our results indicate that classical but also supposedly bias-aware AMIMs extract disease modules based on the node degree rather than benefiting from the interactions of the nodes. Only a single recently proposed method, DOMINO, showed significantly better results on the original PPI network, suggesting that the development of better algorithmic approaches as well as less biased, context-specific networks are urgently needed to provide the biomedical community with the necessary tools to deliver on the promises that the field of active (disease) module identification and *de novo* network enrichment made almost two decades ago.

Methods

PPI networks and random network generators

We ran our test protocol on five widely used PPI networks: BioGRID [31], APID [32, 33], STRING [34] with high confidence interactions only (score $\geq$ 0.7), HPRD [35] and IID [36] with experimentally validated interactions only. Key properties of the

PPI networks are summarized in Supplementary Table 1. All networks have one giant largest connected component with a very small diameter. Note that although BioGRID, APID, STRING and IID are continuously updated, HPRD is no longer maintained and has not been updated since 2010. However, HPRD is still useful for our study, because it is smaller and focuses on well-studied interactions. Moreover, some of the tested AMIMs were designed with HPRD in mind, which was the largest network available at the time of implementation.

We used five different random network generators, which were chosen to produce randomized networks that preserve selected properties of the original PPI networks:

REWIRED: degree preserving generator. Repeatedly swaps pairs of edges and non-edges to produce random networks whose degree sequences are identical to the degree sequences of the original PPI networks [46, 47]. Preserves the individual node degrees and hence the hub-genes.

EXPECTED_DEGREE: expected degree preserving generator. Creates networks with randomly sampled edges where the sampling probabilities are chosen such that the expected node degrees correspond to the node degrees in the original PPI networks [48, 49]. Preserves individual node degrees and hub-genes in expectation.

SHUFFLED: topology preserving generator. Shuffles the gene IDs. Preserves the degree sequence and the topology but not the individual node degrees and the hub-genes.

SCALE_FREE: scale-free generator. Produces scale-free networks using the Barabási–Albert model [50]. The parameters are chosen such that the numbers of nodes and edges in the random networks match the numbers of nodes and edges in the original PPI network. Preserves neither the topology nor the individual node degrees or the hub-genes, but produces networks that are structurally similar to the original PPI networks, since PPI networks are usually scale-free [51, 52].

UNIFORM: uniform generator. Produces random graphs using the Erdős–Rényi model [53]. The parameters are chosen such that the numbers of nodes and edges in the random networks matches the numbers of nodes and edges in the original PPI network. The produced networks are very different from the original PPI networks. In particular, their degrees are binomially distributed, whereas PPI networks tend to have power law degree distributions [51, 52].

Expression, phenotype and survival data

For testing we considered gene expression datasets for five different diseases: ALS, non-small cell LC, UC, CD and HD. For all datasets, case/control phenotype data are available, whereas for ALS, LC and HD, survival data are also reported. Gene probes were mapped to Entrez gene IDs, and if multiple probes corresponded to a single gene, the median value was used. Key properties of the expression datasets are summarized in Supplementary Table 3.

In the LC dataset, we only considered non-small cell LC patients due to their significant biological difference from small cell LC and the larger number of available samples. For the HD dataset, we preselected samples such that the most distinct gene expression difference is present. To achieve this, we only used samples from caudate nucleus, since this region has been reported to have the largest change in gene expression [54]. As a case group, only patients with Vonsattel grades 2–4 were considered, whereas samples with Vonsattel grade 0–1 were discarded.

AMIMs and method-specific preprocessing

In the past years, various AMIMs have been presented (cf. Batra et al. [6] for a benchmarking paper and Lazareva et al. [9] for a systematic review). Here, we selected 10 tools, namely ClustEx2 [22], COSINE [23], DIAMOnD [24], DOMINO [21], GiGA [25], GXNA [26], KeyPathwayMiner [27–29], GrandForest [30], Hierarchical HotNet [19] and NetCore [20] (cf. Supplementary Table 4 for details). These tools were selected for three reasons:

- They require expression data and phenotypes or input formats that can be derived from these data.
- They return a gene set representing a candidate disease module.
- They are available online and sufficiently bug-free and documented to allow integration in our test suite.

Hierarchical HotNet and NetCore are permutation-based methods, i. e. they include data or network randomization steps in their workflows to correct for typical PPI network biases. All other tools use the PPI networks without applying any corrections.

To set the hyper-parameters of the AMIMs, we used default values whenever available. For parameters where no default values are provided in the implementations, we used the values chosen in the tutorials, READMEs, or original publications. For tools that return several candidate disease modules, we always used the union of all reported subnetworks. We hence did not carry out hyper-parameter tuning. The reason for this is 3-fold: Firstly, hyper-parameter tuning would have been computationally infeasible, since already without our protocol required more than 10 000 AMIM runs. Secondly, our aim is not to obtain the optimal results but to test if equally good results can be obtained using a random network. Thirdly, because of the large number of AMIM runs, small changes in the results for a specific AMIM have little effect on the overall conclusions. Note, however, that since we did not optimize the tools, our findings should not be interpreted as a benchmark but rather as an evaluation of the effect of network biases on AMIMs.

Although COSINE, GXNA and GrandForest can be run directly on the normalized expression data, the other tools require different input formats. More specifically, ClustEx2, DIAMOnD and DOMINO expect a list of disease-associated seed genes, Hierarchical HotNet and NetCore expect gene scores, GiGA expects a sorted list of genes, and KeyPathwayMiner expects an indicator matrix of genes that are differentially expressed in the case samples.

For each gene g, let $\mathbf{x}_g^1$ and $\mathbf{x}_g^0$ be the vectors of expression values for all case and control samples, and $x_{g,s}$ be the expression value for sample s. Furthermore, let n be the number of genes contained in the expression dataset and m be the number of case samples. To derive gene scores, seed genes and sorted gene lists from the expression data, we evaluated the two-sided Mann–Whitney U-test on $\mathbf{x}_g^1$ and $\mathbf{x}_g^0$ to obtain P-values P_g of differential expression for all genes g. We then defined gene scores as $-\log_{10}(P_g)$, used all genes g with $P_g < 0.001/n$ as seed genes, and obtained sorted lists of genes by sorting the genes in non-decreasing order of p_g. The indicator matrix $M = (m_{g,s}) \in \{0,1\}^{n\times m}$ required by KeyPathwayMiner was defined as $m_{g,s} = [|x_{g,s} - \text{mean}(\mathbf{x}_g^0)| > 1.5 \cdot \text{std}(\mathbf{x}_g^0)]$, where s is a case sample, $[\cdot]$ is the Iverson bracket (i. e. $[\texttt{true}] = 1$ and $[\texttt{false}] = 0$), and the operators mean(·) and std(·) denote mean and standard deviation, respectively.

Evaluation metrics

Quantitative measures are needed to evaluate how well AMIMs perform on the original and on the randomized PPI networks. That is, we need to quantify the meaningfulness of the gene sets S returned by the tools. For this, we distinguish two dimensions of meaningfulness: predictive power w. r. t. the phenotype and survival time, and functional relevance for the disease of interest.

For quantifying predictive power, we employed MI, which is widely used for selecting features with high predictive power. More precisely, let $\mathbf{y}$ be the vector of case/control disease phenotypes and $\mathbf{x}_g$ be the vector of expression values of all samples for a gene $g \in S$. We computed the mean MI w. r. t. the phenotype $\sum_{g\in S} \mathrm{MI}(\mathbf{x}_g, \mathbf{y})/|S|$ between $\mathbf{y}$ and $\mathbf{x}_g$ across all genes $g \in S$. Analogously, the mean MI w. r. t. the survival times was computed as $\sum_{g\in S} \mathrm{MI}(\mathbf{x}_g, \mathbf{t})/|S|$, where $\mathbf{t}$ denotes the vector of survival times. The larger the mean MI, the stronger the association between the expression data for the genes contained in S and, respectively, the disease phenotypes and the survival times.

To quantify functional relevance, we computed the mean negative log-transformed GSEA P-values between the result sets S and the KEGG [39] pathways related to the disease of interest. The disease-to-pathway mappings are shown in the Supplementary Table 2. Moreover, we computed the overlap coefficients $|S \cap D|/\min\{|S|, |D|\}$ between the results sets S and the disease-associated DisGeNET [40] gene sets D. These gene sets were obtained by taking all genes connected to the condition of interest in DisGeNET. Only for the UC dataset there was no exact match. Therefore, we used genes associated with inflammatory bowel disease of which UC is a subtype. The full DisGeNET diseases IDs mapping to the conditions is shown in the Supplementary Table 2. Note that, for all four meaningfulness scores, larger means better.

Let O be a batch of meaningfulness scores obtained for one of the original PPI networks and R be a batch of scores obtained for randomized counterparts generated by one of the random network generators described above. In the large-scale protocol used for the classical AMIMs (Figure 2A), we used the one-sided Mann–Whitney U-test to assess whether the scores contained in O are significantly larger than the scores contained in R. In the restricted protocol used for the permutation-based methods (Figure 2B), the Mann–Whitney U-test is not applicable, because we have $|O| \leq 4$ for each partitioning of the results (there are only four runs on the original PPI networks). Consequently, we partitioned the results along the methods and disease dimensions to ensure $|O| = 1$ and instead used the one-sided one-sample t-test.

Although the P-values from the one-sided Mann–Whitney U-test and the one-sided one-sample t-test tell us whether the candidate disease modules computed for the original PPI networks are significantly more meaningful than those obtained for the randomized counterparts, they are oblivious to the question if the candidate disease modules for the original PPI networks are sufficiently meaningful in absolute terms. Assume, for instance, that O and R contain negative log-transformed GSEA P-values, that the values contained in O fall into the range $[0.5, 1]$ and that the values contained in R fall into the range $[0.2, 0.5]$. Then the one-sided Mann–Whitney U-test will return a significant P-value, which, however, should be treated with extreme caution because the scores in O are themselves not significant. To account for this fact, we computed a validity score $|\{o \in O \mid o \geq \tau\}|/|O|$ for each P-value computed by the one-sided Mann–Whitney U-test and the one-sided one-sample t-test. For the negative log-transformed GSEA P-values, the threshold was set to $\tau = -\log_{10} 0.05$; for all other scores, we used $\tau = 0.2$. In Figures 3 and 7 and Supplementary Figures 1 and 2, the validity scores are visualized as the sizes of the shapes corresponding to the P-values.

Let O be a batch of result sets obtained for one of the original PPI networks, R be a batch of result sets obtained for randomized counterparts, and avdeg(S) denote the mean degree of a gene set S, computed w. r. t. the original PPI network. For further analyzing the results of the full protocol, we used the one-sided Mann–Whitney U-test to asses whether the mean degrees $\{\mathrm{avdeg}(S) \mid S \in O\}$ of the gene sets for the original networks are significantly larger than the mean degrees $\{\mathrm{avdeg}(S) \mid S \in R\}$ obtained for the randomized counterparts (cf. Figure 6B and E). By splitting along the AMIM dimension, we obtain an array of P-values for each AMIM with entries for each network generator. The correlation coefficients of these arrays with the arrays of AMIM-specific P-values obtained for the meaningfulness scores visualized in Figure 3 indicate to which extent the AMIMs merely learn from the degree distributions of the PPI networks. The larger the correlation coefficient, the stronger the impact of the degrees of the genes contained in the result sets on the meaningfulness scores (cf. Figure 6E).

Implementation

The overall architecture of our test suite is implemented in Python 3 and schematically visualized in Supplementary Figure 5. Each tested AMIM is wrapped into an implementation of an abstract `AlgorithmWrapper` interface. The wrappers run the AMIMs via system calls to the original executables. Graph operations and random network generators are implemented with NetworkX [55] and graph-tools [56]. GSEA is carried out via the GSEApy interface of the Enrichr API [57], and statistical tests are implemented with SciPy [58].

To reproduce the results reported in this paper, it suffices to execute the top-level Python script `run_tests.py`, which is shipped with our test suite. If developers of new AMIMs would like to use our test suite for evaluating their methods, they can provide a custom implementation of the `AlgorithmWrapper` interface. Our test suite can hence be used to easily benchmark new AMIMs against the 10 pre-implemented existing methods. Our test suite is available at https://github.com/dbblumenthal/amim-test-suite/, along with a detailed README and all data needed to reproduce the experiments.

Key Points

- Most AMIMs only learn from the node degrees but not from the biological knowledge encoded in the edges of PPI networks.
- Only the recently presented AMIM DOMINO yields significantly more meaningful disease modules if run on original PPI networks rather than on randomized counterparts with preserved node degrees.
- Better algorithmic approaches and less biased, context-specific networks are urgently needed in the field of active module identification and *de novo* network enrichment.

Availability

The KEGG pathways were obtained from KEGG: https://www.genome.jp/kegg/disease/. BioGRID (v3.2.149), APID (v1.0), STRING (version 11.0) and HPRD (release 9) as well as DisGeNET (v7.0) were obtained using nDEx [59–61]. The IID network (v2018-11) was downloaded from http://iid.ophid.utoronto.ca/. All gene expression datasets and corresponding metadata were retrieved from Gene Expression Omnibus [62], using the GEO2R R interface (https://www.ncbi.nlm.nih.gov/geo/geo2r/). The associated GEO accession codes are shown in Supplementary Table 2. The entire test-suite (Python environment, tool executables, PPI networks, expression datasets) is available at https://github.com/dbblumenthal/amim-test-suite/.

Supplementary data

Supplementary data are available online at *Briefings in Bioinformatics*.

Funding

J.B. received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program (grant no. 777111 and grant no. 826078). This publication reflects only the authors' view and the European Commission is not responsible for any use that may be made of the information it contains. J.B. and M.L. were supported by the German Federal Ministry of Education and Research (BMBF) within the e:Med framework (grant no. 01ZX1908A). J.B. was supported by the German Federal Ministry of Education and Research (BMBF) within the e:Med framework (grant no. 01ZX1910D) and within the CLINSPECT-M framework (grant no. 031L0214A). Contributions by O.L. are funded by the Bavarian State Ministry of Science and the Arts within the framework coordinated by the Bavarian Research Institute for Digital Transformation (bidt, Doctoral Fellow). Figures 1 and 2 were created with BioRender.com.

References

1. Perou CM, Srlie T, Eisen MB, *et al*. Molecular portraits of human breast tumours. *Nature* 2000; **406**(6797): 747–52.
2. Collisson EA, Campbell JD, Brooks AN, *et al*. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* 2014; **511**(7511): 543–50.
3. Guinney J, Dienstmann R, Wang X, *et al*. The consensus molecular subtypes of colorectal cancer. *Nat Med* 2015; **21**(11): 1350–6.
4. van Vliet MH, Reyal F, Horlings HM, *et al*. Pooling breast cancer datasets has a synergetic effect on classification performance and improves signature stability. *BMC Genomics* 2008; **9**:375.
5. Venet D, Dumont JE, Detours V. Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput Biol* 2011; **7**(10):e1002240.
6. Batra R, Alcaraz N, Gitzhofer K, *et al*. On the performance of de novo pathway enrichment. *NPJ Syst Biol Appl* 2017; **3**:6.
7. Silverman EK, Schmidt HHHW, Anastasiadou E, *et al*. Molecular networks in network medicine: development and applications. *Wiley Interdiscip Rev Syst Biol Med* 2020; **12**(6):e1489.
8. Maron BA, Altucci L, Balligand J-L, *et al*. A global network for network medicine. *NPJ Syst. Biol. Appl*. 2020; **6**(1): 29.
9. Lazareva O, Lautizi M, Fenn A, *et al*. Multi-omics analysis in a network context. In Olaf Wolkenhauer. In: *Systems Medicine*. Oxford: Academic Press, 2021, 224–33.
10. Alcaraz N, List M, Batra R, *et al*. De novo pathway-based biomarker identification. *Nucleic Acids Res* 2017; **45**(16): e151.
11. Samokhin AO, Stephens T, Wertheim BM, *et al*. NEDD9 targets COL3A1 to promote endothelial fibrosis and pulmonary arterial hypertension. *Sci Transl Med* 2018; **10**(445):eaap7294.
12. Wang R-S, Loscalzo J. Network-based disease module discovery by a novel seed connector algorithm with pathobiological implications. *J Mol Biol* 2018; **430**(18, Part A): 2939–50.
13. Amitabh Sharma, Arda Halu, Julius L Decano, *et al*. Controllability in an islet specific regulatory network identifies the transcriptional factor NFATC4, which regulates type 2 diabetes associated genes. *NPJ Syst Biol Appl*, **4**:25, 2018.
14. AbdulHameed MDM, Tawa GJ, Kumar K, *et al*. Systems level analysis and identification of pathways and networks associated with liver fibrosis. *PLoS One* 2014; **9**(11):e112193.
15. Halu A, Liu S, Baek SH, *et al*. Exploring the cross-phenotype network region of disease modules reveals concordant and discordant pathways between chronic obstructive pulmonary disease and idiopathic pulmonary fibrosis. *Hum Mol Genet* 2019; **28**(14): 2352–64.
16. Sharma A, Menche J, Chris Huang C, *et al*. A disease module in the interactome explains disease heterogeneity, drug response and captures novel pathways and genes in asthma. *Hum Mol Genet* 2015; **24**(11): 3005–20.
17. Stibius KB, Sneppen K. Modeling the two-hybrid detector: experimental bias on protein interaction networks. *Biophys J* 2007; **93**(7): 2562–2.
18. Schaefer MH, Serrano L, Andrade-Navarro MA. Correcting for the study bias associated with protein-protein interaction measurements reveals differences between protein degree distributions from different cancer types. *Front Genet* 2015; **6**:260.
19. Reyna MA, Leiserson MDM, Raphael BJ. Hierarchical HotNet: identifying hierarchies of altered subnetworks. *Bioinformatics* 2018; **34**(17): i972–80.
20. Barel G, Herwig R. NetCore: a network propagation approach using node coreness. *Nucleic Acids Res* 2020; **48**(17): e98.
21. Levi H, Elkon R, Shamir R. DOMINO: a network-based active module identification algorithm with reduced rate of false calls. *Mol Syst Biol* 2021; **17**(1): e9593.
22. Ding Z, Guo W, Gu J. ClustEx2: gene module identification using density-based network hierarchical clustering. *In CAC* 2018; **2018**:2407–12.
23. Ma H, Schadt EE, Kaplan LM, *et al*. COSINE: COndition-specific sub-NEtwork identification using a global optimization method. *Bioinformatics* 2011; **27**(9): 1290–8.
24. Ghiassian SD, 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 Comput Biol* 2015; **11**(4).
25. Breitling R, Amtmann A, Herzyk P. Graph-based iterative group analysis enhances microarray interpretation. *BMC Bioinform* 2004; **5**:100.
26. Nacu S, Critchley-Thorne R, Lee P, *et al*. Gene expression network analysis and applications to immunology. *Bioinformatics* 2007; **23**(7): 850–8.
27. Nicolas Alcaraz, Hande Kücük, Jochen Weile, *et al*. KeyPathwayMiner: detecting case-specific biological pathways using expression data. *Internet Mathematics*, **7**(4): 299–313, 2011.

28. Alcaraz N, Pauling J, Batra R, *et al.* KeyPathwayMiner 4.0: condition-specific pathway analysis by combining multiple omics studies and networks with cytoscape. *BMC Syst Biol* 2014; **8**(99).
29. List M, Alcaraz N, Dissing-Hansen M, *et al.* KeyPathwayMiner-Web: online multi-omics network enrichment. *Nucleic Acids Res* 2016; **44**(Webserver-Issue): W98–104.
30. Larsen SJ, Schmidt HHHW, Baumbach J. De novo and supervised endophenotyping using network-guided ensemble learning. *Systems Medicine* 2020; **3**(1): 8–21.
31. Oughtred R, Stark C, Breitkreutz B-J, *et al.* The BioGRID interaction database: 2019 update. *Nucleic Acids Res* 2019; **47**(D1): D529–41.
32. Alonso-Lpez D, Gutirrez MA, Lopes KP, *et al.* APID interactomes: providing proteome-based interactomes with controlled quality for multiple species and derived networks. *Nucleic Acids Res* 2016; **44**(W1): W529–35.
33. Alonso-Lpez D, Campos-Laborie FJ, Gutirrez MA, *et al.* APID database: redefining protein-protein interaction experimental evidences and binary interactomes. *Database* 2019; **2019**.
34. Szklarczyk D, Gable AL, Lyon D, *et al.* STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* 2019; **47**(D1): D607–13.
35. Keshava Prasad TS, Goel R, Kandasamy K, *et al.* Human protein reference database–2009 update. *Nucleic Acids Res* 2009; **37**(Database issue): D767–72.
36. Kotlyar M, Pastrello C, Malik Z, *et al.* IID 2018 update: context-specific physical protein-protein interactions in human, model organisms and domesticated species. *Nucleic Acids Res* 2019; **47**(D1): D581–9.
37. Ross BC. Mutual information between discrete and continuous data sets. *PLoS ONE* 2014; **9**(2):e87357.
38. Subramanian A, Tamayo P, Mootha VK, *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 2005; **102**(43): 15545–50.
39. Kanehisa M, Sato Y, Kawashima M, *et al.* KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* 2016; **44**(D1): D457–62.
40. Piero J, Ram-rez-Anguita JM, Sach-Pitarch J, *et al.* The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res* 2020; **48**(D1): D845–55.
41. Lazareva O, Canzar S, Yuan K, *et al.* BiCoN: network-constrained biclustering of patients and omics data. *Bioinformatics* 2020.
42. Wu J, Gu Y, Xiao Y, *et al.* Characterization of DNA methylation associated gene regulatory networks during stomach cancer progression. *Front Genet* 2018; **9**:711.
43. Selber-Hnatiw S, Sultana T, Tse W, *et al.* Metabolic networks of the human gut microbiota. *Microbiology* 2020; **166**(2): 96–119.
44. List M, Ebert P, Albrecht F. Ten simple rules for developing usable software in computational biology. *PLoS Comput Biol* 2017; **13**(1):e1005265.
45. Sadegh S, Matschinske J, Blumenthal DB, *et al.* Exploring the SARS-CoV-2 virus-host-drug interactome for drug repurposing. *Nat Commun* 2020; **11**(1): 3518.
46. Gkantsidis C, Mihail M, Zegura EW. The markov chain simulation method for generating connected power law random graphs. In: Ladner RE (ed). *ALENEX 2003*. SIAM, 2003, 16–25.
47. Viger F, Latapy M. Efficient and simple generation of random simple connected graphs with prescribed degree sequence. *J Complex Networks* 2016; **4**(1): 15–37.
48. Chung F, Lu L. Connected components in random graphs with given expected degree sequences. *Ann Combinatorics* 2002; **6**(2): 125–45.
49. Joel C. Miller and Aric A. Hagberg. Efficient generation of networks with given expected degrees. In Alan M. Frieze, Paul Horn, and Pawel Pralat, editors, *WAW 2011*, volume **6732** of *LNCS*, pages 115–26, Berlin, Heidelberg, 2011. Springer.
50. Barabsi A-L, Albert R. Emergence of scaling in random networks. *Science* 1999; **286**(5439): 509–12.
51. Jeong H, Mason SP, Barabsi AL, *et al.* Lethality and centrality in protein networks. *Nature* 2001; **411**(6833): 41–2.
52. Barabsi A-L, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat. Rev Genet* 2004; **5**(2): 101–13.
53. Erdoős P, Rényi A. On random graphs I. *Publ Math Debrecen* 1959; **6**:290.
54. Hodges A, Strand AD, Aragaki AK, *et al.* Regional and cellular gene expression changes in human Huntington's disease brain. *Hum Mol Genet* 2006; **15**(6): 965–77.
55. Hagberg AA, Schult DA, Swart PJ. Exploring network structure, dynamics, and function using networkx. In: Varoquaux G, Vaught T, Millman J (eds). *SciPy 2008*. Pasadena, 2008, 11–5.
56. Peixoto TP. The graph-tool python library. *figshare* 2014.
57. Kuleshov MV, Jones MR, Rouillard AD, *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* 2016; **44**(W1): W90–7.
58. Virtanen P, Gommers R, Oliphant TE, *et al.* SciPy 1.0: fundamental algorithms for scientific computing in python. *Nat Methods* 2020; **17**:261–72.
59. Pratt D, Chen J, Welker D, *et al.* NDEx, the network data exchange. *Cell Syst* 2015; **1**(4): 302–5.
60. Pratt D, Chen J, Pillich R, *et al.* NDEx 2.0: a clearinghouse for research on cancer pathways. *Cancer Res* 2017; **77**(21): e58–61.
61. Pillich RT, Chen J, Rynkov V, *et al.* NDEx: a community resource for sharing and publishing of biological networks. *Methods Mol Biol* 2017; **1558**:271–301.
62. Barrett T, Wilhite SE, Ledoux P, *et al.* NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res* 2013; **41**(Database issue): D991–5.