GuiltyTargets: Prioritization of Novel Therapeutic Targets With Network Representation Learning

Özlem Muslu, Charles Tapley Hoyt, Mauricio Lacerda, Martin Hofmann-Apitius, and Holger Fröhlich

Abstract—The majority of clinical trials fail due to low efficacy of investigated drugs, often resulting from a poor choice of target protein. Existing computational approaches aim to support target selection either via genetic evidence or by putting potential targets into the context of a disease specific network reconstruction. The purpose of this work was to investigate whether network representation learning techniques could be used to allow for a machine learning based prioritization of putative targets. We propose a novel target prioritization approach, GuiltyTargets, which relies on attributed network representation learning of a genome-wide protein-protein interaction network annotated with disease-specific differential gene expression and uses positive-unlabeled (PU) machine learning for candidate ranking. We evaluated our approach on 12 datasets from six diseases of different type (cancer, metabolic, neurodegenerative) within a 10 times repeated 5-fold stratified cross-validation and achieved AUROC values between 0.92 - 0.97, significantly outperforming previous approaches that relied on manually engineered topological features. Moreover, we showed that GuiltyTargets allows for target repositioning across related disease areas. An application of GuiltyTargets to Alzheimer's disease resulted in a number of highly ranked candidates that are currently discussed as targets in the literature. Interestingly, one (COMT) is also the target of an approved drug (Tolcapone) for Parkinson's disease, highlighting the potential for target repositioning with our method. The GuiltyTargets Python package is available on PyPI and all code used for analysis can be found under the MIT License at https://github.com/GuiltyTargets. Attributed network representation learning techniques provide an interesting approach to effectively leverage the existing knowledge about the molecular mechanisms in different diseases. In this work, the combination with positive-unlabeled learning for target prioritization demonstrated a clear superiority compared to classical feature engineering approaches. Our work highlights the potential of attributed network representation learning for target prioritization. Given the overarching relevance of networks in computational biology we believe that attributed network representation learning techniques could have a broader impact in the future.

Index Terms—Artificial intelligence, neural networks, bioinformatics

1 Introduction

Drug discovery is a time consuming, expensive and complicated process in which there are two major steps prior to *in vivo* pre-clinical and clinical trials [1], [2], [3], [4]. The first is the identification, prioritization, and validation of a target with suitable physical properties whose modulation could affect disease pathways. The second is to identify and optimize compounds which bind to the target and modulate its biological activity (Fig. 1). Even though both have proven crucial for the discovery of efficacious drugs, many still fail in clinical studies due to low efficacy [5], [6], [7]. While computational methods for compound-target interaction prediction have been widely-studied [8], computer based target prioritization remains less so.

Traditionally, scientists identified targets by searching through the relevant literature, following clues from mRNA and protein expression, integrating expression data with pathway analyses, experimenting with knockout mice, investigating somatic mutations, gene fusions, and copy number variations, and using the accumulated knowledge from multiple experimental studies to generate a hypothesis on how proteins or other macromolecules might work as targets [2], [9], [10]. However, manually interpreting many data sources is prone to biased identification of targets as it limits the potential to use all available and helpful data. By computationally integrating multiple biological data sources to analyze prior knowledge, it should be possible to make target identification process faster, less biased, and more informed.

- *Ö. Muslu, C. Tapley Hoyt, and M. Hofmann-Apitius are with the Business Area Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Schloss Birlinghoven, 53754 Sankt Augustin, Germany , and also with the Bonn-Aachen International Center for Information Technology (B-IT), University of Bonn, 53115 Bonn, Germany. E-mail: {ozlemmuslu, cthoyt}@gmail.com, martin.hofmann-apitius@scai.fraunhofer.de.*
- *M. Lacerda is with the Bonn-Aachen International Center for Information Technology (B-IT), University of Bonn, 53115 Bonn, Germany. E-mail: s0mapiod@uni-bonn.de.*
- *H. Fröhlich is with the Business Area Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Schloss Birlinghoven, 53754 Sankt Augustin, Germany, and with the Bonn-Aachen International Center for Information Technology (B-IT), University of Bonn, 53115 Bonn, Germany, and also with the UCB Biosciences GmbHAlfred-Nobel Str. 1040789 Monheim, Germany. E-mail: holger.froehlich@scai.fraunhofer.de.*

Manuscript received 2 Apr. 2019; revised 6 May 2020; accepted 2 June 2020. Date of publication 19 June 2020; date of current version 3 Feb. 2022. (Corresponding author: Holger Fröhlich.)
Digital Object Identifier no. 10.1109/TCBB.2020.3003830

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

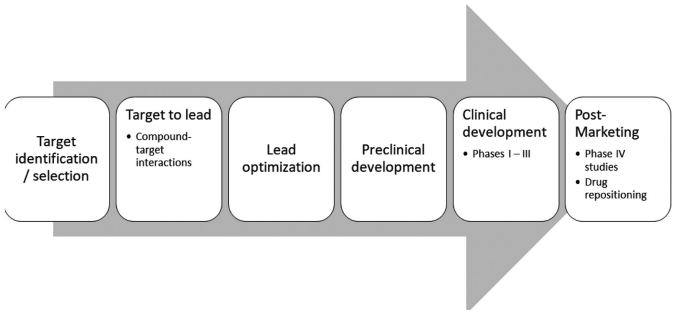

Fig. 1. Overview about the pharmaceutical drug development process: Target prioritization / selection is of relevance before the actual start of compound development. Compound-target interaction prediction focuses on understanding a compound's mode of action. Drug repositioning or repurposing aims for finding new indications for existing drugs on the market.

Computational target prioritization approaches thus aim for improving target identification process by ranking proteins based on their likelihood of being targets in the context of a specific disease [11], [12], [13], [14], [15], [16], [17], [18]. Most of them integrate biological networks with other data sources into a knowledge graph that can help to prioritize targets [9], partially in particular for infectious diseases [12], [13], [14], cancers [15], [16], [17], or neuro-degenerative diseases [18].

In addition to network based approaches, statistical genetic evidence from phenome-wide association studies (PheWAS) has received an increasing attention for identifying targets [19]. However, selecting targets purely based on genetic evidence is likely to narrow the view towards a subset of indications and possible targets. Moreover, such an approach is agnostic against the wealth of known biological mechanisms and existing data.

Another approach is to employ machine learning methods that learn features of known targets within a given disease area or a closely related one in order to prioritize future candidates. Emig *et al.* proposed a method in which for each protein a number of network topological features are combined with proximity to differentially expressed genes in a given disease of interest [11]. All features are subsequently combined into a logistic regression model for ranking proteins as candidate targets. The authors successfully tested their approach with data from 30 different diseases. Ferrero *et al.* used features provided by the Open Targets database [20] and combine them into one ranking score using support vector machines [21].

In this paper, we propose GuiltyTargets, a novel guilt-by-association approach for prioritizing protein targets using a combination of unsupervised attributed network representation learning [22] and PU learning [23], [24], [25], [26]. GuiltyTargets first embeds a genome-wide protein-protein interaction (PPI) network annotated with differential gene expression information in a euclidean space using Gat2Vec, an attributed network representation learning method [27]. It then learns to rank candidate targets, leveraging network representation learning techniques to implicitly learn relevant features to represent a protein-protein interaction network together with mapped data rather than manually engineering various topological network attributes in a time consuming process that might still miss relevant information. Our approach is thus data driven, and to the the best of our knowledge it has not been used for target prioritization so far. The proposed approach is compared to the approaches from [11] and [9] based on 12 datasets from six diseases, demonstrating its superior ranking performance. Finally, we present a case study on Alzheimer's disease (AD), in which we show how GuiltyTargets can be used to reposition known targets from other neurological indications.

2 Method

2.1 Overview of GuiltyTargets

An overview of the GuiltyTargets workflow is presented in Fig. 2. First, a disease-specific differential gene expression

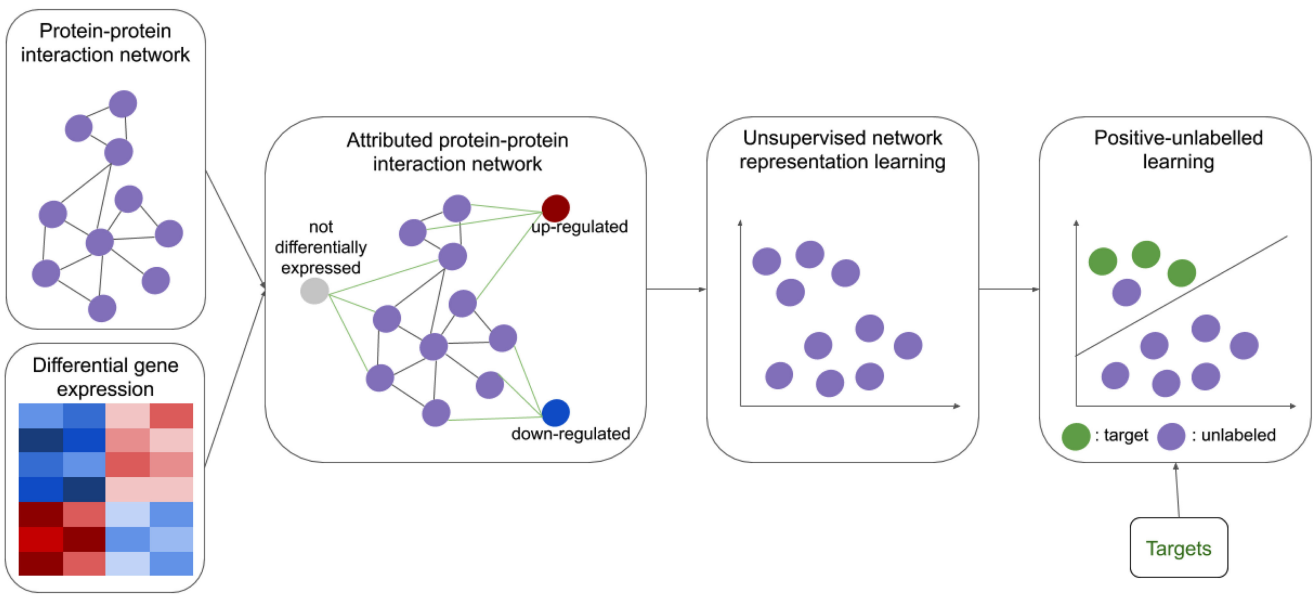

Fig. 2. GuiltyTargets approach: A genome-wide protein-protein interaction network is labeled with information about known drug targets and with differential gene expression (up-regulated, down-regulated, not differentially expressed). The attributed network is subsequently embedded into an euclidean space, in which a (penalized) logistic regression is trained via PU learning. The class conditional probabilities that are calculated by the classifier are then used for ranking of candidate targets.

profile is discretized such that down-regulated genes are assigned a label of -1, up-regulated genes are assigned 1, and unregulated genes are assigned 0. Then, a genome-wide protein-protein interaction (PPI) network is annotated using these labels and input to Gat2Vec, which embeds the nodes in the attributed network into a euclidean space [27] for downstream machine learning tasks.

Following a positive-unlabeled (PU) learning scheme, known disease specific protein targets are assigned positive labels, and the remaining proteins are regarded as pseudo-negatives to train a classifier that ranks a candidate protein according to its similarity to known targets for the given disease. More details are described as follows.

2.2 Network Representation Learning

GuiltyTargets relies on network representation learning of an annotated PPI network via Gat2Vec, where node attributes represent discretized gene expression $\log_2$ fold changes. In the first step, two separate graphs, the structural and the attribute graph, are constructed from the original labeled PPI network, where the structural graph corresponds to the PPI network, and the attribute graph is a bipartite graph between protein nodes and discretized $\log_2$ fold changes. For each node in each of these networks, a low dimensional embedding is calculated using Gat2Vec algorithm using the parameters given in Table S1. In Gat2Vec algorithm, first, random walks are generated in both structural and attribute networks. These random walks are interpreted as a sentence that can be embedded into an euclidean space using a SkipGram neural network, which is an essential part of Word2Vec method [28]. When calculating the low dimensional embedding, Gat2Vec model accordingly aims to maximize the probability of a node v's structural and attribute contexts R and W within a contextual window of length $2c$:

$$L(v) = \sum_{r \in R} \sum_{i=1}^{|r|} \sum_{\substack{-c \le j \le c \\ j \neq i}} \log p(r_j \mid r_i) + \sum_{w \in W} \sum_{i=1}^{|w|} \sum_{\substack{-c \le t \le c \\ t \neq i}} \log p(w_t \mid w_i), \tag{1}$$

Here r_i and w_i denote the ith word (i.e., node or mixed node / attribute sequence) generated by a random walk. $p(r_j \mid r_i)$ is the output of the SkipGram neural network that is defined with a softmax function

$$p(r_j \mid r_i) = \frac{\exp -(\langle \mathbf{v_i}, \mathbf{v_j} \rangle)}{\sum_{-c \le j \le c} \exp -(\langle \mathbf{v_i}, \mathbf{v_j} \rangle)}, \tag{2}$$

where $\mathbf{v_i}, \mathbf{v_j}$ are vector representations of words r_i and r_j in the hidden layer. An equivalent definition holds for $p(w_t \mid w_i)$. Notably, the SkipGram neural network is trained with one-hot vector encoding of word pairs as input. The network aims for learning the probability of observing word r_j in the context (i.e., in the "neighborhood") of r_i (the same holds true for w_t and w_i) by maximizing $\sum_v L(v)$ over all nodes v in the original PPI network. We refer to [28] for more details about SkipGram.

2.3 Target Candidate Ranking

The features retrieved using network representation learning of the annotated PPI networks are used together with the labeling of proteins as known disease-specific targets to train an ℓ_2 penalized logistic regression classifier. Following a PU learning scheme known targets are assigned positive labels, and the remaining proteins are treated as if they were negatives. The model is then used to rank unlabeled proteins via the conditional probability (likelihood score)

$$P(y = 1 \mid \mathbf{x}) = \frac{1}{1 + exp(-\langle \mathbf{x}, w \rangle)}, \tag{3}$$

with w being the learned parameters, y an indicator for a protein being a target and $\mathbf{x}$ the feature vector obtained from deep network representation learning. It is worthwhile mentioning that due to the fact that unlabeled samples contain an unknown fraction of positive samples the class conditional probability shown in the equation is most likely biased compared to the Bayes optimal one. Some authors therefore proposed to down-weight unlabeled samples accordingly [29]. Alternatively, other authors suggested to abstain from a probabilistic classifier and train a biased Support Vector Machine instead [30], or sub-select from the set of all unlabeled samples those, which are most likely negatives, see [31] for a review. According to [29] the effects of these methods on pure ranking performance (which is our primary objective here) are small, and hence we chose to rely on our comparably simpler approach with a conventional ℓ_2 penalized logistic regression classifier here.

For our implementation we used the LogisticRegression class from linear_model module and OneVsRestClassifier class from multiclass module in Python library scikit-learn [32].

2.4 Evaluation and Comparison Against Existing Methods

We compared GuiltyTargets against two alternative methods: a) the machine learning approach by Emig *et al.* and b) the network based Local Radiality (LR) score suggested by Isik *et al.*, which does not employ machine learning. More specifically, the LR score of node n in graph G (here: the PPI network) is defined as:

$$LR(n) = \frac{\sum_{dg \in G} |sp(n, dg)|}{DG}, \tag{4}$$

where $|sp(n, dg)|$ denotes the length of the shortest path connecting node n with differentially expressed gene dg, and $|DG|$ is the total number of differentially expressed genes.

In agreement with Emig *et al.* the performance of our approach and both competing methods were compared within a 10 times repeated stratified 5-fold cross validation scheme with the area under ROC curve (AUROC) as the evaluation criterion. This assessed the probability of each method to rank in an independent test set (comprising known targets as well as unknown candidates) a true known target higher than an unknown protein. Since there are far less known targets than unknown candidates in our data, stratified cross-validation was used. More specifically, stratification ensured that each independent test set inside the repeated cross-validation procedure contained approximately the same number of known targets. Importantly, tuning of the ℓ_2 penalty for

the logistic regression classifier was performed within the cross-validation procedure. This was done via a grid search over different regularization strengths (0.01, 0.1, 1, 10), where each candidate value was evaluated via an inner 5-fold stratified cross-validation.

Due to the imbalance between positive and unlabeled examples we also considered the possibility to weigh samples from each class differently. More specifically, we considered the following options: no class weighting, weighting of samples inversely to the class size (i.e., samples from smaller class are up-weighted), class weight for smaller class 17, 100, 333, 2000 fold the one of the larger class. Once again, each of these candidate options was evaluated within an inner 5-fold stratified cross-validation.

3 Data and Resources

3.1 Gene Expression Data

Gene expression data for acute myeloid leukemia (GSE30029), hepatocellular carcinoma (GSE36411), idiopathic pulmonary fibrosis (GSE24206), liver cirrhosis (GSE36411) and multiple sclerosis (GSE32988) was obtained from Gene Expression Omnibus (GEO) [33], and differential gene expression was assessed via GEO2R [34], Biobase [35], GEOquery [36] and limma [37] using multiple testing correction via the false discovery rate [38]. Only disease status was considered as predictor in the linear model.

For AD, RNASeq data from the AM-PAD Knowledge Portal (AM-PAD) was used [39]. In particular, MSBB, ROSMAP, and MayoRNASeq studies were utilized. Differential gene expression was assessed by applying DESeq2 to the normalized RNAseq data for each brain region. Table S2 shows more detailed information about AM-PAD data, including the number of subject samples and the potentially confounding factors that were considered in the differential gene expression analysis as additional covariates (e.g., age, gender, tissue source).

In every case differential gene expression was declared below a false discovery rate threshold of 5 percent plus an additional $\log_2$ fold change cutoff, which we varied in our analysis.

3.2 Protein-Protein Interaction Networks

As PPI networks, HIPPIE v2.0 [40] and STRING v10.5[41] were used since both of these networks are created by combining multiple sources of PPIs and provide confidence scores. HIPPIE and STRING differ in the type of interactions they contain (Table S3): HIPPIE relies on physical protein-protein interactions, whereas STRING captures more broadly functional interactions. Hence, STRING has a much larger size than HIPPIE. The analyses on this paper only included the interactions between human proteins. STRING locus identifiers were mapped to Entrez identifiers using the mappings provided by STRING.

The accompanying Excel sheet shows information about the number of differentially expressed genes (at different $\log_2$ fold change cutoffs and falsed discovery threshold of 5 percent), which could be mapped to the STRING and HIPPIE network, respectively.

3.3 Target Databases

Information about disease-specific known targets of compounds that are approved drugs or are currently tested in clinical trials were obtained from two databases: The Therapeutic Target Database (TTD) [42] and Open Targets [20] (see: Table S4). In general, the number of targets found in Open Targets was significantly larger than in TTD. For liver cirrhosis and idiopathic pulmonary fibrosis we only found 1 and 6 targets in TTD, respectively. Therefore, for these diseases we only considered Open Targets.

Target identifiers in TTD database were mapped to UniProt identifiers using the conversion file provided by TTD. These identifiers were then mapped to Entrez gene IDs using R packages AnnotationDBI and org.Hs.eg.db. In addition to TTD, known protein targets were retrieved from Open Targets, by filtering by proteins that have known connections to drugs. HGNC symbols were converted to Entrez identifiers using R packages AnnotationDBI and org.Hs.eg.db.

3.4 Validation Approach

We performed target prioritization analyses for six different diseases using corresponding gene expression data for acute myeloid leukemia, hepatocellular carcinoma, idiopathic pulmonary fibrosis, liver cirrhosis, multiple sclerosis and AD. The choice was made based on the following criteria: First, five of these diseases have also been evaluated in the publication by Emig *et al.*, which we used for comparison here. Second, the number of available known targets for each disease was expected to be relatively high for a statistically meaningful validation. Finally, we added AD to investigate the applicability of our approach to a highly challenging disease, in which so far most attempts to establish new drugs have failed [43]. Notably, for AD we investigated brain region specific RNASeq data from different cohorts (MSBB [44], MayoRNASeq [45], ROSMAP [46]).

To investigate the prediction performance of GuiltyTargets we employed two protein-protein interaction networks (STRING[41], HIPPIE [40]), two target databases (Open Targets [20] and Therapeutic Target Database[42]) and different cutoffs to discretize differential gene expression via $\log_2$ fold change thresholds (0, 0.5, 1.0, 1.5) while requiring a false discovery rate of less than 5 percent. Moreover, we tested the situation that no gene expression data was employed at all (technically realized by setting the $\log_2$ fold change threshold to ∞).

4 Results

4.1 GuiltyTargets Outperforms Existing Methods

The approaches by Emig *et al.* and Isik *et al.* were re-implemented using the same PPI network resources and target databases as used by GuiltyTargets. Comparisons were initially only performed with $\log_2$ fold change cutoffs 0.5, 1.0, 1.5 for differential gene expression, but additional thresholds 0 and ∞ were investigated separately for GuiltyTargets in Section 4.2.

Results shown in Table 1 and the accompanying Excel sheet demonstrate a dramatic performance increase of up to 41 and 36 percent by GuiltyTargets compared to the methods by Emig *et al.* and Isik *et al.* respectively. Notably, AUROC values found by our re-implementation of were not identical (but typically close) to the ones reported in the original paper. This was likely due to the fact that not the same PPI network and target database resources have been

TABLE 1
Performance of GuiltyTargets, Emig *et al.* Method, and Isik *et al.* Method, in Terms of Cross-Validated AUROC (± Standard Error)

Disease	Emig *et al.* (original)	Emig *et al.*	Isik *et al.*	GuiltyTargets
Acute myeloid leukemia	0.8195	0.8356 ± 0.0001	0.8422 ± 0.0000	0.9247 ± 0.0002
Alzheimer's disease (MSBB -BM10)	-	0.5905 ± 0.0009	0.5904 ± 0.0002	0.9380 ± 0.0002
Alzheimer's disease (MSBB -BM22)	-	0.6585 ± 0.0002	0.6127 ± 0.0000	0.9682 ± 0.0002
Alzheimer's disease (MSBB -BM36)	-	0.6609 ± 0.0002	0.6398 ± 0.0000	0.9378 ± 0.0003
Alzheimer's disease (MSBB -BM44)	-	0.6575 ± 0.0007	0.5810 ± 0.0000	0.9386 ± 0.0002
Alzheimer's disease (MayoRNAseq - CBE)	-	0.5862 ± 0.0016	0.6582 ± 0.0000	0.9380 ± 0.0004
Alzheimer's disease (MayoRNAseq - TCX)	-	0.6109 ± 0.0013	0.6649 ± 0.0000	0.9461 ± 0.0002
Alzheimer's disease (ROSMAP)	-	0.6340 ± 0.0007	0.5866 ± 0.0000	0.9429 ± 0.0003
Hepatocellular carcinoma	0.8019	0.7314 ± 0.0002	0.7595 ± 0.0000	0.9438 ± 0.0002
Idiopathic pulmonary fibrosis	0.8826	0.8306 ± 0.0018	0.8393 ± 0.0000	0.9280 ± 0.0007
Liver cirrhosis	0.6747	0.5338 ± 0.0019	0.5971 ± 0.0000	0.9458 ± 0.0002
Multiple sclerosis	0.7151	0.6755 ± 0.0002	0.7018 ± 0.0000	0.9430 ± 0.0003

The results on this table has been obtained using the following parameters: Target DB: Open Targets, PPI Network: STRING, Confidence Threshold:0, Differential Gene Expression: FDR $<$ 0.05 and Log2 Fold Change Cutoff: 1.5 (in agreement to Emig et al.*). Further results can be found in Table S6. The column "Emig* et al. *(original)" shows the AUROC values reported by Emig* et al. *and the column "Emig* et al.*" shows the AUROC values obtained by the reimplementation of their method. CBE = cerebellum; TCX = temporal cortex; BM = Broadmann area*

used. More specifically, Emig *et al*. employed the commercial MetaBase database as PPI network and proprietary target database Integrity, whereas here we purely rely on public resources.

We employed Wilcoxon signed rank tests (comparing GuiltyTargets against each of the two competing methods) to assess the statistical significance of our findings in each scenario. This confirmed a significant improvement of GuiltyTargets compared to both competing methods in almost every dataset and tested scenario ($p < 0.05$ after Holm's correction for multiple testing , see Supplementary Excel sheet). Fig. 3 shows the AUROC of all three compared methods averaged over all tested scenarios. The overall averaged AUROC improvement by GuiltyTargets was 19 percent compared to Emig *et al*. and 21 percent compared to Isik *et al*..

4.2 In-Depth Analysis of Influence Factors on GuiltyTargets Performance

We wanted to better understand the dependency of the performance of GuiltyTargets on the different tested influence

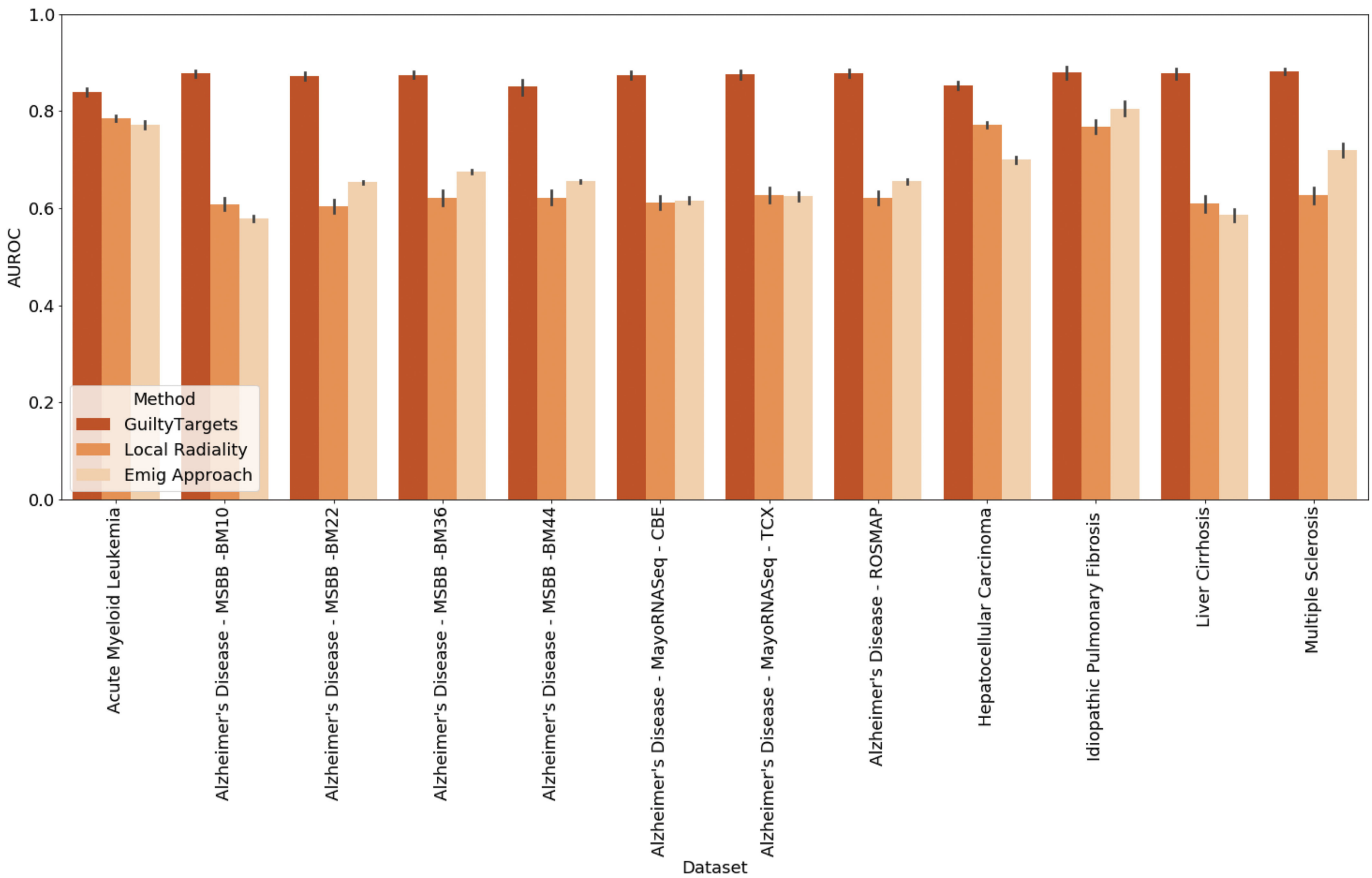

Fig. 3. Comparison of GuiltyTargets versus approach by Emig *et al.* and Isik *et al.* (Local Radiality).: The barplots show AUROC values averaged over possible hyper-parameter choices (log2 fold change cutoff (0.5, 1.0, 1.5), target database (Open Targets, TTD), PPI network (HIPPIE, STRING), PPI network confidence cutoff (0.0, 0.63).

TABLE 2
In-Depth Analysis of Different Influence Factors on the Performance of GuiltyTargets

Influence factor	Comparison	Difference	p-value
PPI network	STRING vs. HIPPIE	0.076	<2E-16
PPI confidence threshold	default vs 0.63	0.021	<2E-16
Target database	Open Targets vs. TTD	0.02	<2E-16
Fold change thresh.	all diseases	overall eff.	0.006
Fold change thresh.	AD	overall eff.	1.633e-09
Dataset	AD	overall eff.	1.125e-11

The table shows the result of contrasts extracted from coefficients of a robust linear model fit and a robust ANOVA, respectively. The difference in AUROC is shown in column 3 together with the corresponding p-value in column 4.

factors, which we had varied individually in our cross-validation analysis:

- PPI network (STRING, HIPPIE), including different confidence level thresholds
- Target database (Open Targets, Therapeutic Targets Database)
- Thresholds for $\log_2$ fold changes

For this purpose we fitted a robust linear model with all possible influence factors (i.e., PPI network, target database, fold change threshold, PPI confidence level cutoff) and the dataset as a further factor. We used R-package "robust" for this purpose. Table 2 demonstrates that the employed PPI network is the most relevant influence factor for GuiltyTargets: Using STRING significantly increased the AUROC compared to using HIPPIE by 7.6 percent on average. A more conservative confidence threshold for the STRING network yielded a drop in prediction performance by 2.1 percent. The use of the Open Targets versus the Therapeutic Target Database significantly increased the ranking performance of GuiltyTargets by 2 percent, hence underlining the relevance of a larger number of known targets for learning the ranking model in the embedded network space.

When averaging over all previously mentioned influence factors (i.e., network resource, target database, disease), the chosen $\log_2$ fold change threshold seemed to have no significant influence on AUROC ($p = 0.36$, robust ANOVA). However, fixing the PPI network to STRING with default confidence threshold and the target database to Open Targets showed a clearly significant positive effect ($p = 0.0061$). To further investigate this fact we analyzed the performance of our GuiltyTargets only across our tested AD gene expression datasets, which again confirmed a highly significant influence of fold change cutoff on the AUROC ($p = 1.633e-09$). In addition, the dataset factor in the robust linear model has a clearly significant effect ($p = 1.125e-11$). Both findings together imply that gene expression data does have a statistical effect on the ranking performance of GuiltyTargets, but indeed effect sizes are small and highly dataset dependent (Fig. S1). Based on the observation that with a sufficiently conservative $\log_2$ fold change threshold prediction performances on average improved, we altogether recommend the use of gene expression data with our methods.

4.3 GuiltyTargets Learns from Known Targets

We tested whether the performance of GuiltyTargets was dependent on known targets or whether also with a random set of proteins a similar performance could have been achieved. For this purpose we trained GuiltyTargets for each disease with 100 randomly drawn sets of targets of the same size as the actual ones, which we incorrectly labeled as "targets". Prediction performance was evaluated using the same cross-validation procedure as before. Table S6 confirms that the AUROC for random proteins drops to about 50 percent, i.e., chance level. Hence, GuiltyTargets indeed learns properties of known targets.

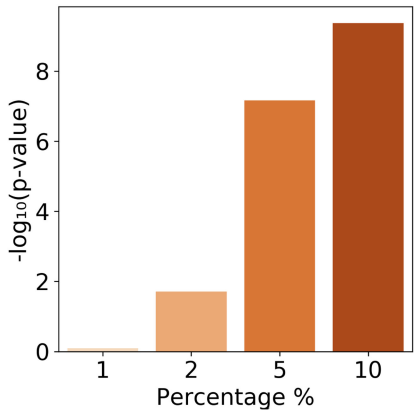

Fig. 4. Target repositioning potential of GuiltyTargets: The barplot shows the result of a hypergeometric test conducted on the top p% of a ranked list of candidate proteins when looking for overrepresentation of known AD targets. GuiltyTargets was trained without any known AD targets here.

4.4 GuiltyTargets Allows for Target Repositioning Across Related Diseases

We explored whether GuiltyTargets could transfer properties learned from known targets in one disease to another one, hence allowing for repositioning of targets. To address this question we trained GuiltyTargets with all known targets of neurodegenerative diseases obtained from Open Targets, while excluding known AD targets. We then ran a hypergeometric test on the resulting prioritization to see if known AD targets where statistically overrepresented at the top of the list. The results were significant when at least 2 percent of the top candidates were considered (Fig. 4). This shows that GuiltyTargets could help for repositioning targets across related disease areas.

4.5 Case Study: GuiltyTargets Predicts New Candidate Targets for Alzheimer's Disease

Despite 179 therapeutic targets listed in the Open Targets database, the AD field urgently requires new and more effective medications that either prevent, mitigate, or reverse its progression. We therefore picked out AD as a test case for GuiltyTargets to prioritize new target candidates. We used post-mortem gene expression data from brain tissue from the ROSMAP study and combined it with STRING network and Open Targets as a resource for known targets. ROSMAP data was chosen because of its comparably large number of samples (495 AD patients and 438 controls). Table 3 shows the top 0.1 percent of a ranked list of novel candidate targets obtained with GuiltyTargets. Enriched GO terms were calculated using DAVID [47], indicating that candidate targets are mostly related to synaptic transmission and ion transport, in line with the neurological characteristics of AD. Furthermore, the GO terms "learning" and "memory" was enriched, which shows that the proposed targets cover the basic properties of AD. Detailed results of our enrichment analysis can be found

TABLE 3
Target Prioritization for AD Using ROSMAP Gene Expression Data

Entrez identifier	HGNC symbol	Likelihood score	Class	Known drugs / druggable
1143	CHRNB4	0.7	Nicotinic acetylcholine receptor	SIB-1553A (Alzheimer, discontinued in Phase 2)
3708	ITPR1	0.689	IP3 receptor	yes
2742	GLRA2	0.619	Ligand gated chloride channel	yes
1312	COMT	0.587	Catechol-O-Methyltransferase	Tolcapone (Parkinson)
2898	GRIK2	0.587	Ionotropic glutamate receptor	yes
1132	CHRM4	0.586	Muscarinic acetylcholine receptor	yes
89832	CHRFAM7A	0.557	Nicotinic acetylcholine receptor	yes
3363	HTR7	0.532	Serotonin receptor	JNJ-18038683 (Major depressive disorder) ATI-9242 (Schizophrenia, discontinued in Phase 2)
3777	KCNK3	0.523	Potassium channel	yes
2741	GLRA1	0.484	Glycine receptor	D-Serine (Parkinson's disease, Phase 4)
1136	CHRNA3	0.461	Nicotinic acetylcholine receptor	yes
3269	HRH1	0.451	Histamine receptor	Doxepin (Depression) Doxylamine (Anxiety disorder) Propiomazine (Insomnia, Anxiety disorder)] Pyrilamine Maleate (headache)
6543	SLC8A2	0.448	Solute carrier	
2911	GRM1	0.447	Metabotropic glutamate receptor	PF-1913539 (Alzheimer's disease, discontinued in Phase 3) A-841720 (Pain, preclinical) AZD8529 (Scizophrenia, discontinued in Phase 2)
2913	GRM3	0.445	Metabotropic glutamate receptor	BCI-632 (Alzheimer disease, Major depressive disorder, Phase 1) Pomaglumetad (Schizophrenia, Phase 1) LY-544344 (Anxiety disorder, Discontinued in Phase 3) LY354740 (Anxiety disorder, Discontinued in Phase 2) R-1578 (Mood disorder, Discontinued in Phase 2) RO-4995819 (Major depressive disorder, Discontinued in Phase 2)
3361	HTR5A	0.436	Serotonin receptor	yes
8001	GLRA3	0.412	Ligand gated chloride channel	yes
1360	CPB1	0.411	Protaminase	yes
9456	HOMER1	0.407	Neuronal immediate-early gene	

This list shows the top 0.1 percent candidate proteins. The last column shows either known drugs (including indications) against the respective target or the classification as "druggable" using the information from DGIdb [49] and TTD [42].

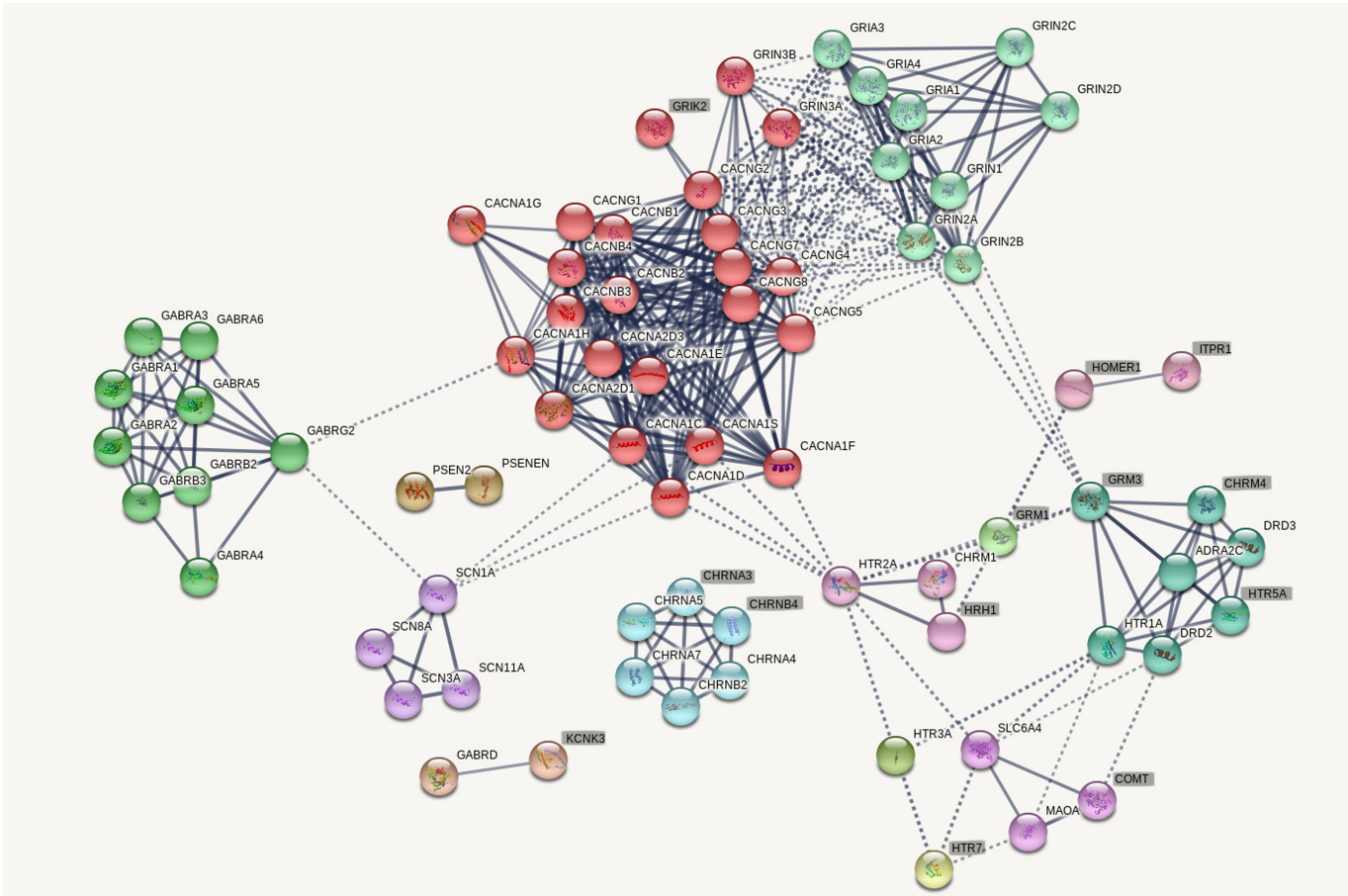

Fig. 5. Interactions between known and candidate targets, with confidence scores higher than 0.7. Clusters of nodes were calculated using MCL clustering [59] with inflation parameter of 3.4 and the nodes were colored based on the clusters they are in. The transparency of the links shows the confidence score of the interaction. If a node has some known or predicted 3D structure, it is filled with a structure image. Highlighted nodes show the proposed candidates, whereas the rest show the known targets. Image generated using STRING [41].

in the Supplementary Excel sheet. Fig. 5 visualizes candidate proteins and their interactions with known target proteins, demonstrating a higher than expected interaction rate (PPI enrichment $p < 1.0E-16$, calculated using STRING web interface) meaning that candidates likely address the same or a similar disease biology as the known targets. According to the Therapeutic Target Database [42] and DGIdb [48] databases, all but two candidates are labeled as "druggable", i.e., they could be used as targets for drugs using the current drug development methods.

Many of the candidate targets are receptors, namely four acetylcholine receptors (three nicotinic, one muscarinic), and three glutamate receptors (two metabotropic, one ionotropic), in agreement with the observation that receptors constitute a large portion of known targets for small molecule drugs [49]. The remaining candidates were identified as ion channels. The top candidate CHRNB4 is the target of the compound SIB-1553A, which has been tested in a phase 2 clinical trial for AD, but discontinued (source: Therapeutic Target Database). Out of the other top candidates we found CHRFAM7A, GRM1, GRM3, ITPR1, HTR7, and COMT particularly interesting: CHRFAM7A is an alpha-7 nicotinic cholinergic receptor subunit interacting with amyloid-β, whose aggregates (i.e., plaques) are one of the hallmarks of AD [50]. CHRFAM7A may promote neuronal survival and function, and subunits are expressed by astrocytes participating in synaptic communication [51].

GRM1 is the target of the compound PF-1913539, which has been discontinued in a phase 3 AD trial [42].

GRM3 (mGlu3) is found in astrocytes as well as neuronal cells, and have been observed to have neuroprotective properties. Its agonists and positive allosteric modulators were reported to be potentially helpful for AD treatment [52]. Glial mGlu3 receptors regulate the production of neurotrophic factors such as nerve growth factor, brain-derived neurotrophic factor and glial-derived neurotrophic factor [52]. BCI-632, a compound that targets GRM3, is currently being tested in a phase 1 AD trial [42].

ITPR1, an intracellular Ca 2+ channel, mediates calcium release from the endoplasmic reticulum, triggering apoptosis, and its deletion has been linked to spinocerebellar ataxia type 15, a neurodegenerative disease [53], [54].

Single nucleotide polymorphisms (SNPs) rs73310256 in HTR7 [55] and rs4680 in COMT have been associated with AD [56]. COMT is currently discussed as a target for AD [57]. It is the target of the anti-Parkinson drug Tolcapone (source: Therapeutic Target Database), supporting our previous finding that GuiltyTargets can re-propose targets from related diseases.

5 Conclusion

We presented a network representation learning based approach for target prioritization, GuiltyTargets. The main advantage of network representation learning over traditional feature engineering is that these methods directly learn useful network features from a combination of network topology and experimental data. To our knowledge such an approach has not been applied to target prioritization so far. Our approach uses a protein-protein interaction network, a differential gene expression profile and a list of known targets to prioritizes proteins as targets for a particular disease. We showed that GuiltyTargets is highly robust and significantly outperforms the methoda by Emig *et al.* and Isik *et al.* in terms of ranking performance. As demonstrated by our validation studies, it is applicable to various types of diseases, including cancers, metabolic and neurodegenerative diseases. We demonstrated that GuiltyTargets can be used to repurpose existing targets from a different (but related) disease area. Application of GuiltyTargets to AD showed that several of the highest ranked candidates are indeed proposed in the literature for AD (but are not included into the Open Targets database), and three of them have been targeted by candidate AD drugs.

We have developed GuiltyTargets with the classical view of targets being proteins in mind. However, it should be mentioned that nowadays also other structures are considered as targets, e.g., non-coding RNA molecules (ncRNAs) [59], [60]. Since ncRNAs, specifically miRNAs, have a regulatory influence on gene expression, these molecules could potentially be integrated into a PPI network, and therefore GuiltyTargets may also be able to rank miRNAs. However, this will require more detailed investigation in future research.

GuiltyTargets as well as other machine learning based target prioritization methods (including the one by Emig *et al.*) learn properties of known targets in the same or a related disease to rank candidate proteins. Hence, these approaches rely on available data and knowledge about known targets [61]. Because of this dependency GuiltyTargets has in principle the same limitations as all machine learning methods: The performance is dependent on the quality of existing data and employed network resource. PPI networks are never complete, and there might exist biases, because some proteins are better functionally characterized than others. Furthermore, it is unlikely to discover completely novel disease and target biology with a machine learning based target prioritization approach. Instead, the rationale is a) to address the incompleteness of any available target database by allowing to fill gaps, and b) to allow for repurposing targets from related disease areas. Despite its limitations GuiltyTargets showed promising results for both use cases, including our case study for AD. Hence, we see GuiltyTargets as a promising tool to support the decision process in the context of target identification in pharmaceutical research in addition to PheWAS based approaches. In particular, GuiltyTargets can provide hints to interesting targets in indication areas, where genetic evidence is (still) missing.

From a broader perspective our work demonstrates the potential of network representation learning in the bioinformatics field. Given the overarching relevance of networks in computational biology we believe that attributed network representation learning techniques could have a broader impact for other applications in the future.

6 Funding

This work was partially supported by the Fraunhofer Gesellschaft.

Acknowledgments

The authors would like to thank Daniel Domingo-Fernández for his valuable help in interpreting the ranked candidate list for Alzheimer's disease. They would also like to thank Murat

Can Özdemir and Jonas Ibn-Salem for valuable discussions. The results published here are in part based on data obtained from the AMP-AD Knowledge Portal (doi: 10.7303/syn2580853). For MSBB data set, these data were generated from postmortem brain tissue collected through the Mount Sinai VA Medical Center Brain Bank and were provided by Dr. Eric Schadt from Mount Sinai School of Medicine. For MayoRNASeq data set, study data were provided by the following sources: The Mayo Clinic Alzheimer's Disease Genetic Studies, led by Dr. Nilufer Taner and Dr. Steven G. Younkin, Mayo Clinic, Jacksonville, FL using samples from the Mayo Clinic Study of Aging, the Mayo Clinic Alzheimer's Disease Research Center, and the Mayo Clinic Brain Bank. Data collection was supported through funding by NIA grants P50 AG016574, R01 AG032990, U01 AG046139, R01 AG018023, U01 AG006576, U01 AG006786, R01 AG025711, R01 AG017216, R01 AG003949, NINDS grant R01 NS080820, CurePSP Foundation, and support from Mayo Foundation. Study data includes samples collected through the Sun Health Research Institute Brain and Body Donation Program of Sun City, Arizona. The Brain and Body Donation Program is supported by the National Institute of Neurological Disorders and Stroke (U24 NS072026 National Brain and Tissue Resource for Parkinsons Disease and Related Disorders), the National Institute on Aging (P30 AG19610 Arizona Alzheimer's Disease Core Center), the Arizona Department of Health Services (contract 211002, Arizona Alzheimer's Research Center), the Arizona Biomedical Research Commission (contracts 4001, 0011, 05- 901 and 1001 to the Arizona Parkinson's Disease Consortium) and the Michael J. Fox Foundation for Parkinson's Research. For ROSMAP data set, study data were provided by the Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago. Data collection was supported through funding by NIA grants P30AG10161, R01AG15819, R01AG17917, R01AG30146, R01AG36836, U01AG32984, U01AG46152, the Illinois Department of Public Health, and the Translational Genomics Research Institute.

References

[1] P. Csermely, T. Korcsmáros, H. J. Kiss, G. London, and R. Nussinov, "Structure and dynamics of molecular networks: A novel paradigm of drug discovery: a comprehensive review," *Pharmacol. & Herapeutics*, vol. 138, no. 3, pp. 333–408, 2013.

[2] I. Gashaw, P. Ellinghaus, A. Sommer, and K. Asadullah, "What makes a good drug target?" *Drug Discovery Today*, vol. 16, no. 23–24, pp. 1037–1043, 2011.

[3] J. P. Hughes, S. Rees, S. B. Kalindjian, and K. L. Philpott, "Principles of early drug discovery," *British J. Pharmacol.*, vol. 162, no. 6, pp. 1239–1249, 2011.

[4] M. Lotfi Shahreza, N. Ghadiri, S. R. Mousavi, J. Varshosaz, and J. R. Green, "A review of network-based approaches to drug repositioning," *Briefings Bioinf.*, vol. 28, pp. 878–892, 2017.

[5] J. Arrowsmith, "Trial watch: Phase II failures: 2008–2010," *Nature Rev. Drug Discov.*, vol. 10, no. 5, pp. 328–329, 2011.

[6] G. Laenen, L. Thorrez, D. Börnigen, and Y. Moreau, "Finding the targets of a drug by integration of gene expression data with a protein interaction network," *Mol. BioSystems*, vol. 9, no. 7, pp. 1676–1685, 2013.

[7] J. Arrowsmith, "Phase III and submission failures: 2007–2010," *Nature Rev. Drug Discovery*, vol. 10, no. 2, pp. 1–2, 2011.

[8] X. Chen *et al.*, "Drug–target interaction prediction: Databases, web servers and computational models," *Briefings Bioinf.*, vol. 17, no. 4, pp. 696–712, Jul. 2016. [Online]. Available: https://academic.oup.com/bib/article/17/4/696/2240330

[9] Z. Isik, C. Baldow, C. V. Cannistraci, and M. Schroeder, "Drug target prioritization by perturbed gene expression and network information," *Sci. Reports*, vol. 5, Nov. 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4663505/

[10] F. L Moseley, K. Bicknell, M. S Marber, and G. Brooks, "The use of proteomics to identify novel therapeutic targets for the treatment of disease," *J. Pharmacy Pharmacology*, vol. 59, pp. 609–28, 2007.

[11] D. Emig *et al.*, "Drug target prediction and repositioning using an integrated network-based approach," *PLoS One*, vol. 8, no. 4, 2013, Art. no. e60618.

[12] M. A. Doyle, R. B. Gasser, B. J. Woodcroft, R. S. Hall, and S. A. Ralph, "Drug target prediction and prioritization: using orthology to predict essentiality in parasite genomes," *BMC Genomics*, vol. 11, no. 1, 2010, Art. no. 222.

[13] M. S. Paul, A. Kaur, A. Geete, and M. E. Sobhia, "Essential gene identification and drug target prioritization in leishmania species," *Mol. BioSystems*, vol. 10, no. 5, pp. 1184–1195, 2014.

[14] S. K. Gupta, R. Gross, and T. Dandekar, "An antibiotic target ranking and prioritization pipeline combining sequence, structure and network-based approaches exemplified for serratia marcescens," *Gene*, vol. 591, no. 1, pp. 268–278, 2016.

[15] S.-H. Yeh, H.-Y. Yeh, and V.-W. Soo, "A network flow approach to predict drug targets from microarray data, disease genes and interactome network-case study on prostate cancer," *J. Clinical Bioinf.*, vol. 2, no. 1, 2012, Art. no. 1.

[16] F. Vitali *et al.*, "A network-based data integration approach to support drug repurposing and multi-target therapies in triple negative breast cancer," *PloS One*, vol. 11, no. 9, 2016, Art. no. e0162407.

[17] G. Bidkhori *et al.*, "Metabolic network-based identification and prioritization of anticancer targets based on expression data in hepatocellular carcinoma," *Frontiers Physiol.*, vol. 9, 2018, Art, no. 916.

[18] H. Keane, B. J. Ryan, B. Jackson, A. Whitmore, and R. Wade-Martins, "Protein-protein interaction networks identify targets which rescue the mpp+ cellular model of parkinson's disease," *Sci. Reports*, vol. 5, 2015, Art. no. 17004.

[19] D. Diogo *et al.*, "Phenome-wide association studies across large population cohorts support drug target validation," *Nature Commun.*, vol. 9, no. 1, Oct. 2018, Art. no. 4285. [Online]. Available: https://www.nature.com/articles/s41467–018-06540-3/

[20] G. Koscielny *et al.*, "Open targets: a platform for therapeutic target identification and validation," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D985–D994, 2016.

[21] E. Ferrero, I. Dunham, and P. Sanseau, "In silico prediction of novel therapeutic targets using gene–disease association data," *J. Translational Med.*, vol. 15, no. 1, 2017, Art. no. 182.

[22] N. Sheikh, Z. Kefato, and A. Montresor, "gat2vec: Representation learning for attributed graphs," *Computing*, vol. 101, pp. 187–209, 2019.

[23] H. Li *et al.*, "Activation of signal transducer and activator of transcription-5 in prostate cancer predicts early recurrence," *Clin Cancer Res.*, vol. 11, no. 16, pp. 5863–5868, Aug. 2005.

[24] L. Peng *et al.*, "Screening drug-target interactions with positive-unlabeled learning," *Sci. Reports*, vol. 7, no. 1, 2017, Art. no. 8087.

[25] P. N. Hameed, K. Verspoor, S. Kusljic, and S. Halgamuge, "Positive-unlabeled learning for inferring drug interactions based on heterogeneous attributes," *BMC Bioinf.*, vol. 18, no. 1, 2017, Art. no. 140.

[26] R. Yang, B. J. Daigle, L. R. Petzold, and F. J. Doyle, "Core module biomarker identification with network exploration for breast cancer metastasis," *BMC Bioinf.*, vol. 13, no. 1, Jan. 2012, Art. no. 12.

[27] N. Sheikh, Z. Kefato, and A. Montresor, "gat2vec: Representation learning for attributed graphs," *Computing*, Apr. 2018. [Online]. Available: https://doi.org/10.1007/s00607–018-0622-9

[28] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*.

[29] C. Elkan and K. Noto, "Learning classifiers from only positive and unlabeled data," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2008, pp. 213–220.

[30] B. Liu, Y. Dai, X. Li, W. S. Lee, and P. S. Yu, "Building text classifiers using positive and unlabeled examples," in *Proc. 3rd IEEE Int. Conf. Data Mining*, 2003, pp. 179–188.

[31] E. Sansone, F. G. B. D. Natale, and Z. Zhou, "Efficient training for positive unlabeled learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2584–2598, Nov. 2019.

[32] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[33] R. Edgar, M. Domrachev, and A. E. Lash, "Gene expression omnibus: NCBI gene expression and hybridization array data repository," *Nucleic Acids Res.*, vol. 30, no. 1, pp. 207–210, 2002.
[34] T. Barrett *et al.*, "Ncbi geo: Archive for functional genomics data sets—update," *Nucleic Acids Res.*, vol. 41, no. D1, pp. D991–D995, 2012.
[35] W. Huber *et al.*, "Orchestrating high-throughput genomic analysis with bioconductor," *Nature Methods*, vol. 12, no. 2, 2015, Art. no. 115.
[36] S. Davis and P. S. Meltzer, "Geoquery: A bridge between the gene expression omnibus (GEO) and bioconductor," *Bioinformatics*, vol. 23, no. 14, pp. 1846–1847, 2007.
[37] M. E. Ritchie *et al.*, "limma powers differential expression analyses for RNA-sequencing and microarray studies," *Nucleic Acids Res.*, vol. 43, no. 7, pp. e47–e47, 2015.
[38] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *J. Royal Statist. Soc. Ser. Methodol.*, vol. 57, pp. 289–300, 1995.
[39] R. J. Hodes and N. Buckholtz, "Accelerating medicines partnership: Alzheimer's disease (AMP-AD) knowledge portal aids alzheimer's drug discovery through open data sharing," *Expert Opinion Ther. Targets*, vol. 19, pp. 878–892, 2016.
[40] G. Alanis-Lobato, M. A. Andrade-Navarro, and M. H. Schaefer, "Hippie v2. 0: Enhancing meaningfulness and reliability of protein–protein interaction networks," *Nucleic Acids Res.*, vol. 45, pp. D408–D414, 2016.
[41] D. Szklarczyk *et al.*, "The STRING database in 2017: Quality-controlled protein–protein association networks, made broadly accessible," *Nucleic Acids Res.*, pp. D362–D368, 2016.
[42] Y. H. Li *et al.*, "Therapeutic target database update 2018: Enriched resource for facilitating bench-to-clinic research of targeted therapeutics," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D1121–D1127, 2017.
[43] D. Mehta, R. Jackson, G. Paul, J. Shi, and M. Sabbagh, "Why do trials for alzheimer's disease drugs keep failing? a discontinued drug perspective for 2010–2015," *Expert Opinion Investigational Drugs*, vol. 26, no. 6, pp. 735–739, 2017.
[44] M. Wang *et al.*, "The mount sinai cohort of large-scale genomic, transcriptomic and proteomic data in alzheimer's disease," *Sci. Data*, vol. 5, 2018, Art. no. 180185.
[45] M. Allen *et al.*, "Human whole genome genotype and transcriptome data for alzheimer's and other neurodegenerative diseases," *Sci. Data*, vol. 3, 2016, Art. no. 160089.
[46] S. Mostafavi *et al.*, "A molecular network of the aging human brain provides insights into the pathology and cognitive decline of alzheimer's disease," *Nature Neurosci.*, vol. 21, no. 6, 2018, Art. no. 811.
[47] G. Dennis *et al.*, "David: Database for annotation, visualization, and integrated discovery," *Genome Biol.*, vol. 4, no. 9, 2003, Art. no. R60.
[48] K. C. Cotto *et al.*, "Dgidb 3.0: A redesign and expansion of the drug–gene interaction database," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D1068–D1073, 2018. [Online]. Available: http://dx.doi.org/10.1093/nar/gkx1143
[49] R. Santos *et al.*, "A comprehensive map of molecular drug targets," *Nature Rev. Drug Discovery*, vol. 16, no. 1, 2017, Art. no. 19.
[50] M. Murpy and H. LeVine III, "Alzheimer's disease and the β-amyloid peptide," *J Alzheimers Dis*, vol. 19, no. 1, pp. 311–323, 2010.
[51] K. T. Dineley, A. A. Pandya, and J. L. Yakel, "Nicotinic ACH receptors as therapeutic targets in CNS disorders," *Trends Pharmacol. Sci.*, vol. 36, no. 2, pp. 96–108, 2015.
[52] F. Caraci *et al.*, "Metabotropic glutamate receptors in neurodegeneration/neuroprotection: Still a hot topic?" *Neurochemistry Int.*, vol. 61, no. 4, pp. 559–565, 2012.
[53] K. Hara *et al.*, "Total deletion and a missense mutation of ITPR1 in japanese SCA15 families," *Neurology*, vol. 71, no. 8, pp. 547–551, 2008.
[54] J. Van de Leemput *et al.*, "Deletion at ITPR1 underlies ataxia in mice and spinocerebellar ataxia 15 in humans," *PLoS Genetics*, vol. 3, no. 6, 2007, Art. no. e108.
[55] C. Herold *et al.*, "Family-based association analyses of imputed genotypes reveal genome-wide significant association of alzheimer's disease with OSBPL6, PTPRG, and PDCL3," *Mol. Psychiatry*, vol. 21, no. 11, 2016, Art. no. 1608.
[56] R. M. Corbo, G. Gambina, E. Broggio, D. Scarabino, and R. Scacchi, "Association study of two steroid biosynthesis genes (COMT and CYP17) with alzheimer's disease in the italian population," *J. Neurological Sci.*, vol. 344, no. 1–2, pp. 149–153, 2014.
[57] M. N. Perkovic, D. S. Strac, L. Tudor, M. Konjevod, G. N. Erjavec, and N. Pivac, "Catechol-o-methyltransferase, cognition and alzheimer's disease," *Current Alzheimer Res.*, vol. 15, no. 5, pp. 408–419, 2018.
[58] A. J. Enright, S. Van Dongen, and C. A. Ouzounis, "An efficient algorithm for large-scale detection of protein families," *Nucleic Acids Res.*, vol. 30, no. 7, pp. 1575–1584, Apr. 2002.
[59] J. Qu, X. Chen, Y.-Z. Sun, J.-Q. Li, and Z. Ming, "Inferring potential small molecule-miRNA association based on triple layer heterogeneous network," *J. Cheminformatics*, vol. 10, no. 1, Jun. 2018, Art. no. 30.
[60] J. Hanna, G. S. Hossain, and J. Kocerha, "The potential for microRNA therapeutics and clinical research," *Frontiers. Genetics*, vol. 10, May 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6532434/
[61] P. Zakeri, J. Simm, A. Arany, S. ElShal, and Y. Moreau, "Gene prioritization using Bayesian matrix factorization with genomic and phenotypic side information," *Bioinformatics*, vol. 34, no. 13, pp. i447–i456, 2018.

Özlem Muslu received the bachelor's degree in computer science and engineering from Sabanci University, and the master's degree in life science informatics from the University of Bonn. She is working toward the PhD degree in bioinformatics and is currently working at TRON Translational Oncology Mainz.

Charles Tapley Hoyt received the PhD degree in computational life sciences from the University of Bonn. His research interests include the applications of knowledge graph embeddings on biomedical knowledge graphs.

Mauricio Lacerda just submitted his master's thesis in life science informatics at the University of Bonn. He contributed to the implementation and data analysis shown in this article.

Martin Hofmann-Apitius received the PhD degree in molecular genetics (biology) from the University of Tübingen. He is currently a professor of applied life science informatics with the Bonn-Aachen International Centre for Information Technology (B-IT), University of Bonn. In addition, he holds a position as a head of the Department of Bioinformatics, Fraunhofer SCAI in St. Augustin, Germany.

Holger Fröhlich (Member, IEEE) received the PhD degree in computer science and heads the AI & Data Science Group at Fraunhofer SCAI. In addition, he is currently a professor with the University of Bonn. His research interests include the development and Application of Data Science and AI Methods in biomedicine and biopharmaceutical research.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/csdl.