

Combining Bottom-up and top-down approaches through graph learning over interaction networks for drug-target-interaction prediction

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Supplementary information:10264703 Supplementary data are available at *Bioinformatics* online.

© The Author 2015. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

molecular fingerprints, similarity to other drugs (See Bioinf Survey), and other molecular features may be used. On the protein side, secondary structure prediction (CITATION), contact prediction (CITATION), or simply convolution over the amino acid sequences can be used to obtain a feature representation for a given proteins. However, both bottom-up and top-down approaches to drug-target interaction prediction

Don't use "simply".

replace: "contain and share some problems" with something like "have some limitations"

that are not solvable within themselves.

Following is not sufficiently precise; here, you need to clearly state the challenges faced by both approaches, ideally with references.

Thus, bottom-up approaches share the lack of ability to generalize, which we will show in later sections, and usually focus on engineering sophisticated features for the drugs, while neglecting to formulate meaningful features on the protein side. Top-down approaches lack the ability to spot small differences to cope with small differences within the drug structure and rely heavily on given data for the considered drug-target pair. The latter is not suitable for predictions on novel or unseen compounds, as e.g., data on side effects or its impact on diseases is seldom given for novel drugs.

In order to design such a feature for proteins and drugs, respectively, we make use of the interaction networks for both proteins and compounds. Drug-drug interaction networks were introduced and standardized by Ayvaz et al. (2015) and have been used for clinical decision support (Scheife et al., 2015). Drug-drug interaction networks may give a hint on common targeted pathways. As an additional compound feature we will use semantic side effect similarity, which we will discuss later on.

Generally, try to avoid pointers to "later".

Protein-protein interaction networks have shown great results in . . . ((Vazquez et al., 2003), (Ackerman et al., 2019)) in granting context for molecular system biology. However, these contexts were never applied to the problem of drug-target-interaction prediction. Thus we formalized our hypotheses over these interaction graphs and will test them in the following chapters.

2 Methods

2.1 Problem Description

The issue of predicting drug-target interactions can be described quite briefly: For a given drug and a given protein we want to forecast whether those interact or not. Additionally, we presume the closed world assumption, that all drug-protein pairs without an interaction, do not have one eventually. We also do not differentiate between activation and inhibition, and do not erect statements on the strength of the bond, thus leading to a binary classification task.

2.2 Datasets

The data for the different parts of this model were obtained from various sources. Starting with the protein-protein interactions, we fetched ≈ 11000 human proteins with over 170.000 links from STRING (Szklarczyk et al., 2014). For the drug-target interactions themselves, we fetched 137.000 links from STITCH database (Szklarczyk et al., 2015). As both STRING and STITCH provide probability scores for each association, we filtered them as advised by a threshold of 700, thus only obtaining likely interactions.

For the ontology segment we utilized PhenomeNET (Hoehndorf et al., 2011), a collection of various ontologies such as Human Phenotype Ontology (Köhler et al., 2018), Gene Ontology (Ashburner et al. (2000) and Seth Carbon et al. (2020)), Mammalian Phenotype Ontology (Smith and Eppig, 2009) and numerous others. Side effects and their links to drugs were obtained Side Effect Resource (SIDER)(Kuhn et al., 2015)

which ontologies to cite

and structured according MedDRA database (Mozzicato, 2009). They were mapped to PhenomeNET with aid of *Phenomebrowser.net*, which provides a SPARQL query endpoint for the mentioned resources.

For comparative evaluation we used the gold standard dataset introduced by ?, which includes both drug-target interaction pairs and side effects.

Eventually, we only considered proteins that had at least one link in either STITCH or STRING, and drugs with at least one side effect and one existing target. Thus, the intersection between these resources yielded 1160 drugs and 6680 human proteins for the training phase. We provide links to and methods for the necessary data in the provided Github repository.

2.3 Evaluation and metrics

As the number of drug-targets are sparsely given w.r.t. to the number of both drugs and proteins considered, the resulting training, validation and testing datasets are highly imbalanced. As there are only 22.000 links in the considered subset the ratio

$$\frac{\#dt_links}{\#drugs \cdot \#proteins} \approx 360,$$

consequently needing compensation in the computed loss function and appropriate metrics for the evaluation.

On that account, we weighted all positive drug-protein pair samples with this ratio by introducing the following loss function w.r.t. to binary cross-entropy:

$$l(x, y) = -w [y \cdot \log x + (1 - y) \cdot \log(1 - x)] \quad (1)$$

for given prediction x and target y . We average this loss among all drug-protein pairs in the training set, leading to a stable environment for the used optimizing scheme *Adam* (Kingma and Ba, 2014). We implemented a 5-fold cross validation among the proteins as justified in the results section. Furthermore, we used early stopping to detect plateaus in the training process.

To assess each model, we computed the *area under receiver operating characteristic-score* (AUROC), *F1-score* (F1) and *Matthews correlation coefficient* (MCC) for each considered hyperparameter configuration setting.

Avoiding validation overfitting, we eventually tested the best performing model on the Yamanishi dataset as presented and discussed. Furthermore, we compare our results on this very dataset with various other cutting edge approaches to DTI-prediction.

2.4 Model

In order to build a method that incorporates both top-down and bottom-up features, we first created a model for each individually. Hence, we assemble rich molecular structure based features for drugs from *SmilesTransformer*(Honda et al., 2019) and proteins from *DeepGOPlus*(Kulmanov and Hoehndorf, 2019). *SmilesTransformer* introduces an autoencoder, learning over the SMILES strings and thus the molecular organization of each drug in an unsupervised manner. On the other hand, we take advantage of the pretrained models of *DeepGOPlus*, obtaining features from the proteins amino acid sequence, showing significant performance in the field

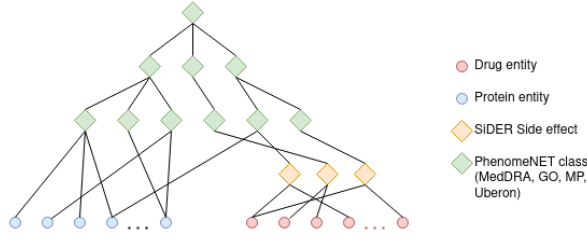


Figure 1. Drugs and proteins with annotations to SIDER and PhenomeNET

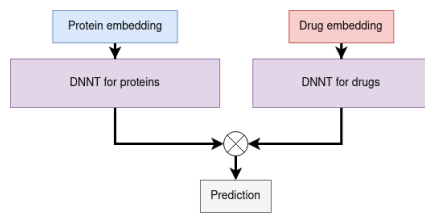


Figure 2. Siamese network applied to molecular and DL2vec features, utilizing deep neural network transformations (DNNT). The similarity function \otimes yields the similarity between both transformed embeddings e.g. by computing the cosine similarity.

of protein function prediction. Thus, both embeddings seem to suitably supplement the following ontology based representations.

In the top-down section, we used *DL2vec*(Chen *et al.*, 2020) to obtain ontology based representations. Hereby, *DL2vec* constructs a graph by introducing vertices and edges for each ontology class and axiom, respectively, followed by random walks starting from each entity. These walks are eventually learned on using a *Word2vec* (Mikolov *et al.*, 2013) model. Thus, we pick up rich, neighbourhood focused representations for each entity, which has shown great results for representing protein function and phenotypes. The overall structure of the ontology can be seen in Figure 1.

As we wanted to learn from the similarity of drug side effects and protein phenotypes we opted for a deep siamese network approach, hence learning a high-dimensional embedding emphasizing this identity by forcing a maximal cosine similarity between these embeddings. On the other hand we built a deep neural network for the molecular structure based features, also benefiting from the siamese network architecture. Therefore, the precomputed embeddings are run through a deep neural network transformation (DNNT) which also reduces the eventual representation size for drugs and proteins separately. An example structure for both types of features can be found in Figure 2.

2.4.1 Graph convolutional layers

These molecular and ontology based sub-models were added to a larger graph convolutional model, based on the protein-protein interaction (PPI) graph. The PPI dataset is represented by a graph $G = (V, E)$, where each protein is represented by a vertex $v \in V$, and each edge $e \in E \subseteq V \times V$ symbolizes an interaction between two proteins. Additionally, we introduce mapping $x : V \rightarrow \mathbb{R}^d$ projecting each vertex v to its node feature $x_v = x(v)$, where d denotes the dimensionality of the node features.

As described before, graph convolution has shown significant performance increase in a variety of tasks. While there are various methods out there we will only introduce the most basic one here. A graph convolutional layer w.r.t. Kipf and Welling (2016) hereby consists of a learnable weight matrix followed by an aggregation step, formalized by

$$\mathbf{X}' = \hat{\mathbf{D}}^{-1/2} \hat{\mathbf{A}} \hat{\mathbf{D}}^{-1/2} \mathbf{X} \Theta \quad (2)$$

where for a given graph $G = (V, E)$, $\hat{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ denotes the adjacency matrix with added self-loops for each vertex, $\hat{\mathbf{D}}$ is described by $\hat{D}_{ii} = \sum_{j=0} \hat{A}_{ij}$, a diagonal matrix displaying the degree of each node, and Θ denotes the learnable weight matrix. Added self-loops enforce that each node representation is directly dependent on its own preceding one. Notably, the number of graph convolutional layers stacked equals the radius of relevant nodes for each vertex within the graph.

The update rule for each individual node is denoted by

$$\mathbf{x}'_i = \Theta \sum_j^N \frac{1}{\sqrt{\hat{d}_j \hat{d}_i}} \mathbf{x}_j \quad (3)$$

where both \hat{d}_i, \hat{d}_j are dependent on the edge weights e_{ij} of the graph. With simple, single valued edge weights such as $e_{ij} = 1 \forall (i, j) \in E$, all \hat{d}_i reduce to d_i , i.e. the degree of each vertex i . We will denote this type of graph convolutional neural layers with GCNConv.

While in this initial formulation the node-wise update step is defined by the sum over all neighbouring node representations, we are able to alter this formulation for a more sophisticated message passing scheme. As described we are able to rearrange the order of activation function σ , aggregation AGG and linear neural layer MLP with this formulation as proposed by Li *et al.* (2020a):

$$\mathbf{x}'_i = \text{MLP}(\mathbf{x}_i + \text{AGG}(\{\sigma(\mathbf{x}_j + \mathbf{e}_{ji}) + \epsilon : j \in \mathcal{N}(i)\})) \quad (4)$$

where we will generally only consider $\sigma \in \{\text{ReLU}, \text{LeakyReLU}\}$. We will denote this generalized layer type as GENConv, following the notation of Fey and Lenssen (2019). While the reordering is merely import for numerical stability, the authors claim that this alteration of the original formulation eases the vanishing gradient problem for deeper convolutional networks. Additionally, the authors revise the aggregation function by generalizing it, allowing more sophisticated weighting functions such as learnable SoftMax or Power for the incoming signals for each vertex, substituting the previously mentioned averaging step. Hence, while GCNConv suffers from both vanishing gradients and signal fading for large scale, highly connected graphs, each propagation step in GENConv emphasizes strong signals with values close to 0 and 1.

We employ another mechanism to avoid redundancy and fading signals in stacked graph convolutional networks, which was introduced by Li *et al.* (2019) and refined in Li *et al.* (2020b). The authors propose a residual network architecture and normalization scheme, solving these issues in a variety of graph predicting tasks. This structure is depicted in Figure 3 resulting in reusable residual blocks, which can be stacked multiple times, thereby not losing focus of each node neighbourhood.

2.4.2 Hyperparameter tuning

To find the best hyperparameter configuration for the proposed model we performed a grid search, to find the most expressive and non-redundant representation. Therefor, we pretrained the bottom-up and the top-down

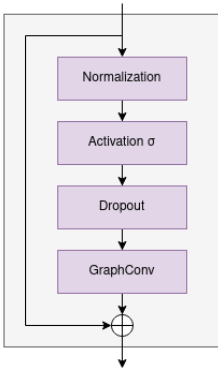


Figure 3. Residual architecture built by Li et al. (2019) and Li et al. (2020b) enabling deeper graph convolutional models

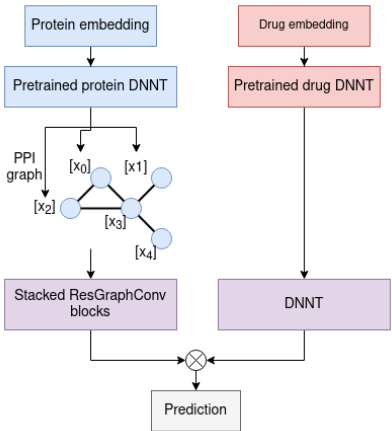


Figure 4. Full DTI prediction model based on the pretrained deep neural network transformers (DNNT) for either molecular structure or ontology based features. The transformed protein representations are added to each corresponded protein as node features for the graph convolutional steps.

model separately and aimed at best performing models w.r.t. previously described metrics. Embedding sizes, depth of the neural network, optimizer, learning rate and layer types were found from an extensive, manual grid search.

3 Results

3.1 High level description

To obtain a prediction for a given drug-protein pair, we first build DL2vec representations for both drugs and proteins over PhenomeNET, while also preparing structural, molecular representations. Hence, we pretrain the siamese networks for both methods and ontology features, yielding the deep neural network transformers for both methods and both drugs and proteins, individually. These embeddings are now added to the graph neural network as node features, where we start an end-to-end learning and training process. Notably, like in the pretraining phase, we rely on siamese neural networks, enforcing similar representations for fitting drug-protein pairs. The overall architecture can be seen in Figure 4.

4 Findings

4.1 Deification of our method

- we built protein function and ontology based features based on DL2vec
- Ontology derived protein function focused features are highly predictive for dtis
- We built a versatile template for various features to test localization on the PPI graph
- normal GCNs don't work on PPI graph, as it is highly connected → needs stronger more expressive aggregation function → GENConv in residual blocks for better numerical stability
- protein functions localize on the PPI graph, while molecular features don't

all AUROC in % AUROC score on STITCH		
DNNT model	Without graph model	With graph model
MolPred	69	69
PhenomeNETPred	88	92
MolPred + PhenomeNETPred	89	93

- microAUC for MolPred + PhenomeNETPred on graph is about 93+ –
- and on yamanishi dataset

DNNT model	Without graph model	With graph model
PhenomeNETPred	83	84
MolPred + PhenomeNETPred	83	84.5

- MicroAUC is about 83

4.2 How to insult other methods

- Only few other methods perform their split over proteins (Wang and Kurgan, 2018), DTI-CDF does it
- Running split over proteins is harder than, drug and drug protein pair split (see below table)
- this applies for both DTI prediction and drug target affinity prediction (and Saras gene-disease association)
- using indications is like cheating, as not applicable for searching new drugs
- drug indications are highly predictive for downstream tasks, but lack capability to differentiate highly related drugs/proteins
- Stratified Cross validation is suitable for training, but **NOT** for validating and testing (Uselessly high AUPRC)
- microAUC is a superior and more intuitive metric for drug repurposing → why for each protein and not for each drug

Approach	Splitting scheme	Original AUROC score	Protein split AUROC	MicroAUC
DTINet	DP pairs	91	84.1	67.2
DTIGEMS+	DP pairs	93	72.2	67.8
DTI-CDF	Proteins	83	83	79
GNN-DTI				
MolTrans			In this work we are testing the following hypotheses:	
GraphData				

- A naive predictor (ranking proteins) and predict each drug similarly achieves cutting edge performance (87.5 AUROC for whole dataset, 85.5 for 5-fold cross validation in drug-split) → No prot focused microAUC possible. → hub proteins
- Yamanishi Dataset is only partially suitable, if everybody just derives a suitable subset (DTIGEMS)

4.3 Choosing a train-test splitting scheme

In general, drug-target interaction prediction is the task of accurately predicting, whether for a given drug and a given protein there is a biological interaction within the target organism. Hereby, different training and prediction schemes lead to divergent expressiveness of the resulting model. However, when building the train-test split over compound-protein pairs for building the actual model, there are the following three options:

- Build split over drugs
- Build split over drug-target pairs
- Build split over proteins

In general, recent works do perform their split over the drugs or drug-target pairs (Wang and Kurgan, 2018), CITATION). As there are hopefully many more drugs to discover, the drug split scheme both emphasizes the drug re-purposing idea, by applying unseen compounds to existing targets, but also benefits from more complicated drug representations, leading to tremendous results. This performance gain is based on minor variations among large groups of pharmaceuticals, that are easy to acquire. The second scheme has knowledge on all drugs and all proteins, and is thus prone to overfitting and the same development bias. Eventually, as there only limited drug-targets (Overington *et al.*, 2006), predicting per protein is rather counter-intuitive. As it is hard to generalize over proteins representations, we aim at reaching similar performances for both drug and protein splitting schemes.

In general, recent works do perform their split over the drugs or drug-target pairs (Wang and Kurgan, 2018), CITATION). The first is more relevant for novel drugs, as it is much more likely to test a new compound than a innovative protein. However, it lies in the very nature of the used datasets, making the prediction for new drugs much easier. Thus, drugs are often built by minor variations of existing drugs, thus leading to no deviations in the functional group of that very compound (CITATION/EXAMPLE). When distributed over both train and test split, the models do not perform inductive inference and generalize, but rather implement transductive inference by just predicting the recently seen structures. Hence, when entirely new molecules are seen, the models perform much worse. The same applies to splits of drug-target pairs, as all drugs were already seen, and novelty cannot be coped with. As mentioned in the introduction it is quite difficult to learn suitable features from proteins. In general, attempts search for motifs in the protein sequences under usage of convolutional neural networks and filters, which is more suitable for tasks like protein function prediction, than for for drug-target interaction prediction, and lack a more in-depth hypothesis on the protein side, while investing in refined drug features. Thus, building splitting over proteins is the most challenging of the three

4.4 Tested hypotheses

- In this work we are testing the following hypotheses:
- Can we build a model that outperforms state of the art approaches, combining top-down and bottom-up approaches?
 - Are interaction networks sufficient to improve the performance of simple molecular predictors?

We will test the first hypothesis by building a model that takes both top-down and bottom-up features into account. Thus, we propose a novel approach to combine those mutual exclusive attempts, through the usage of interaction networks, similarity and molecular features. Additionally, we test the latter by building a simple molecular DTI predictor and enhance it under usage of the interaction networks.

For the bottom-up approach we build a model that only relies on molecular features, which we will discuss in more detail in the following methods chapter. For the combination of both approaches we now attach the predictions to the protein-protein interaction graph as node features for future graph learning steps. In this graph we tried to find both patterns and regions for each drug that could be of interest through application of different graph convolutional layers, which in return represent the feature for each protein. Representing the drug we take the drug-drug interaction graph and the semantic similarity over side effects which we will explain in the following paragraphs.

5 Methods

5.1 Models

The used model consists of two separate models, that help to fuse together the two methods:

- The molecular predictor
- The interaction network based predictor

We build the molecular predictor by using pretrained, molecular fingerprints models for both drugs and proteins. Regarding proteins, we used the pretrained feature generator from *DeepGoPlus* ((Kulmanov and Hoehndorf, 2019)) that was originally designed for protein function prediction and is regarded as state of the art for this purpose. For drugs we used a pretrained fingerprint model from *SMILES transformer* ((Honda *et al.*, 2019)), that provides a simple and fast method to compute fingerprints through autoencoder models. The encodings from these two models were funneled into a simple deep neural network (see Figure ??) with few fully connected. The results of that prediction flow into the annotation of the protein-protein interaction (PPI) graph as depicted in (IMAGE). Hereby, the predictions of the molecular predictor are used as node features for the graph, with respect to the given drug. Thus, given a compound-target pair, the nodes of the PPI graph now hold bottom-up features, which can now be processed by the graph learning algorithms. The PPI graph is processed by different graph convolutional layers, that may underline the importance of either patterns or regions within the graph, to obtain a feature vector for the wanted node. In contrast to learning over whole graphs we perform node classification within the graph. These layers are either graph convolutional layers, that learn a certain kernel over the graph, or attention based. Different layers of both and other types such as were tested.

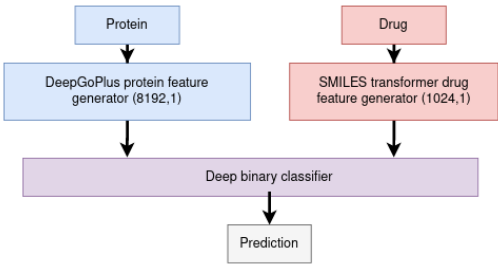


Figure 5. Molecular predictor based on the generated features from DeepGoPlus and SMILES transformer.

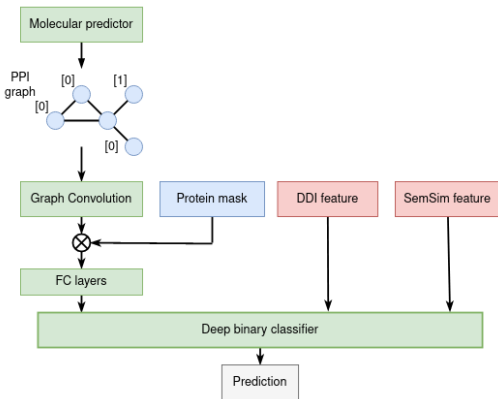


Figure 6. Deep neural network that predicts based on drug-drug interaction features and semantic similarity features over side effects for drugs, and graph convolution over protein-protein interaction networks for proteins. Protein and drug features are represented by blue and red, respectively.

The drug-drug interaction features are retrieved by choosing the corresponding row in the adjacency matrix of the graph, thus leading to quite simple features. For the semantic similarity feature, that once again represents a top-down attribute, we artificially link each drug to its corresponding side effects in the MedDRA hierarchy. Concerning this hierarchy, drug-drug similarity is computed by the Resnik similarity ((Resnik, 1995)). For the given compound we take the corresponding row of this symmetric similarity matrix.

Thereby, we concatenate these three features together and funnel them into another deep neural network as depicted in figure 6. This network finally yields our prediction. We hereby perform splits over both drugs and proteins, in order to test and show the discrepancy and increasing difficulty. Implementation was done in PyTorch ((Paszke et al., 2019)) and is available on Github under github.com/thinnerichs/KAUST-dti-metabol. Graph learning methods were build with help of PyTorch-Geometric ((Fey and Lenssen, 2019)), a geometric deep learning extension library for PyTorch, that recently got a lot of attention in the machine learning community. This library gives the potential to use many state of the art graph learning mechanisms, such as plain but effective graph convolution (Kipf and Welling (2016)), Chebychev kernels ((Defferrard et al., 2016)), ARMA kernels ((Bianchi et al., 2019)), translation-invariant operators ((Verma et al., 2017)), attention mechanisms ((Veličković et al., 2017)), random walks ((Klicpera et al., 2018)) and mixtures of the latter two ((Hamilton et al., 2017)). The performance of these various layer types were tested for this particular problem, as discussed in the results section.

6 Discussion

7 Conclusion

Acknowledgements

This work has been supported by the... Text Text Text Text.

References

Ackerman, E. E., Alcorn, J. F., Hase, T., and Shoemaker, J. E. (2019). A dual controllability analysis of influenza virus-host protein-protein interaction networks for antiviral drug target discovery. *BMC Bioinformatics*, 20(1).

and Seth Carbon, Douglass, E., Good, B. M., Unni, D. R., Harris, N. L., Mungall, C. J., Basu, S., Chisholm, R. L., Dodson, R. J., Hartline, E., Fey, P., Thomas, P. D., Albou, L.-P., Ebert, D., Kesling, M. J., Mi, H., Muruganujan, A., Huang, X., Mushayahama, T., LaBonte, S. A., Siegele, D. A., Antonazzo, G., Attrill, H., Brown, N. H., Garapati, P., Marygold, S. J., Trovisco, V., dos Santos, G., Falls, K., Tabone, C., Zhou, P., Goodman, J. L., Strelets, V. B., Thurmond, J., Garmiri, P., Ishtiaq, R., Rodríguez-López, M., Acencio, M. L., Kuiper, M., Lægrend, A., Logie, C., Lovering, R. C., Kramarz, B., Saverimuttu, S. C. C., Pinheiro, S. M., Gunn, H., Su, R., Thurlow, K. E., Chibucos, M., Giglio, M., Nadendla, S., Munro, J., Jackson, R., Duesbury, M. J., Del-Toro, N., Meldal, B. H. M., Paneerselvam, K., Perfetto, L., Porras, P., Orchard, S., Shrivastava, A., Chang, H.-Y., Finn, R. D., Mitchell, A. L., Rawlings, N. D., Richardson, L., Sangrador-Vegas, A., Blake, J. A., Christie, K. R., Dolan, M. E., Drabkin, H. J., Hill, D. P., Ni, L., Sitnikov, D. M., Harris, M. A., Oliver, S. G., Rutherford, K., Wood, V., Hayles, J., Bähler, J., Bolton, E. R., Pons, J. L. D., Dwinell, M. R., Hayman, G. T., Kaldunski, M. L., Kwitek, A. E., Laulederkind, S. J. F., Plasterer, C., Tutaj, M. A., Vedi, M., Wang, S.-J., D'Eustachio, P., Matthews, L., Balhoff, J. P., Aleksander, S. A., Alexander, M. J., Cherry, J. M., Engel, S. R., Gondwe, F., Karra, K., Miyasato, S. R., Nash, R. S., Simison, M., Skrzypek, M. S., Weng, S., Wong, E. D., Feuermann, M., Gaudet, P., Morgat, A., Bakker, E., Berardini, T. Z., Reiser, L., Subramaniam, S., Huala, E., Arighi, C. N., Auchincloss, A., Axelsen, K., Argoud-Puy, G., Bateman, A., Blatter, M.-C., Boutet, E., Bowler, E., Breuza, L., Bridge, A., Britto, R., Bye-A-Jee, H., Casas, C. C., Coudert, E., Denny, P., Estreicher, A., Famiglietti, M. L., Georgioui, G., Gos, A., Gruaz-Gumowski, N., Hatton-Ellis, E., Hulo, C., Ignatchenko, A., Jungo, F., Laiho, K., Mercier, P. L., Lieberherr, D., Lock, A., Lussi, Y., MacDougall, A., Magrane, M., Martin, M. J., Masson, P., Natale, D. A., Hyka-Nouspikel, N., Orchard, S., Pedruzzi, I., Pourcel, L., Poux, S., Pundir, S., Rivoire, C., Speretta, E., Sundaram, S., Tyagi, N., Warner, K., Zaru, R., Wu, C. H., Diehl, A. D., Chan, J. N., Grove, C., Lee, R. Y. N., Muller, H.-M., Raciti, D., Auken, K. V., Sternberg, P. W., Berriman, M., Paulini, M., Howe, K., Gao, S., Wright, A., Stein, L., Howe, D. G., Toro, S., Westerfield, M., Jaiswal, P., Cooper, L., and Elser, J. (2020). The gene ontology resource: enriching a Gold mine. *Nucleic Acids Research*, 49(D1), D325–D334.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1), 25–29.

Ayvaz, S., Horn, J., Hassanzadeh, O., Zhu, Q., Stan, J., Tatonetti, N. P., Vilar, S., Brochhausen, M., Samwald, M., Rastegar-Mojarad, M., Dumontier, M., and Boyce, R. D. (2015). Toward a complete dataset of drug–drug interaction information from publicly available sources. *Journal of Biomedical Informatics*, 55, 206–217.

Bianchi, F. M., Grattarola, D., Alippi, C., and Livi, L. (2019). Graph neural networks with convolutional arma filters.

Chen, J., Althagafi, A., and Hoehndorf, R. (2020). Predicting candidate genes from phenotypes, functions and anatomical site of expression. *Bioinformatics*.

Defferrard, M., Bresson, X., and Vandergheynst, P. (2016). Convolutional neural networks on graphs with fast localized spectral filtering.

Fey, M. and Lenssen, J. E. (2019). Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*.

Hamilton, W. L., Ying, R., and Leskovec, J. (2017). Inductive representation learning on large graphs.

Hoehndorf, R., Schofield, P. N., and Gkoutos, G. V. (2011). PhenomeNET: a whole-phenome approach to disease gene discovery. *Nucleic Acids Research*, 39(18), e119–e119.

Honda, S., Shi, S., and Ueda, H. R. (2019). Smiles transformer: Pre-trained molecular fingerprint for low data drug discovery.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization.

Kipf, T. N. and Welling, M. (2016). Semi-supervised classification with graph convolutional networks.

Klicpera, J., Bojchevski, A., and Günnemann, S. (2018). Predict then propagate: Graph neural networks meet personalized pagerank.

Köhler, S., Carmody, L., Vasilevsky, N., Jacobsen, J. O. B., Danis, D., Gourdine, J.-P., Gargano, M., Harris, N. L., Matentzoglou, N., McMurry, J. A., Osumi-Sutherland, D., Cipriani, V., Balhoff, J. P., Conlin, T., Blau, H., Baynam, G., Palmer, R., Gratian, D., Dawkins, H., Segal, M., Jansen, A. C., Muaz, A., Chang, W. H., Bergerson, J., Laulederkind, S. J. F., Yüksel, Z., Beltran, S., Freeman, A. F., Sergouniotis, P. I., Durkin, D., Storm, A. L., Hanauer, M., Brudno, M., Bello, S. M., Sincan, M., Rageth, K., Wheeler, M. T., Oegema, R., Loughi, H., Rocca, M. G. D., Thompson, R., Castellanos, F., Priest, J., Cunningham-Rundles, C., Hegde, A., Lovering, R. C., Hajek, C., Olry, A., Notarangelo, L., Similuk, M., Zhang, X. A., Gómez-Andrés, D., Lochmüller, H., Dollfus, H., Rosenzweig, S., Marwaha, S., Rath, A., Sullivan, K., Smith, C., Milner, J. D., Leroux, D., Boerkoel, C. F., Klion, A., Carter, M. C., Groza, T., Smedley, D., Haendel, M. A., Mungall, C., and Robinson, P. N. (2018). Expansion of the human phenotype ontology (HPO) knowledge base and resources. *Nucleic Acids Research*, **47**(D1), D1018–D1027.

Kuhn, M., Letunic, I., Jensen, L. J., and Bork, P. (2015). The SIDER database of drugs and side effects. *Nucleic Acids Research*, **44**(D1), D1075–D1079.

Kulmanov, M. and Hoehndorf, R. (2019). DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics*.

Li, G., Müller, M., Thabet, A., and Ghanem, B. (2019). Deepgcns: Can gcns go as deep as cnns?

Li, G., Xiong, C., Thabet, A., and Ghanem, B. (2020a). Deeppercn: All you need to train deeper gcns.

Li, G., Xiong, C., Thabet, A., and Ghanem, B. (2020b). Deeppercn: All you need to train deeper gcns.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space.

Mozzicato, P. (2009). MedDRA. *Pharmaceutical Medicine*, **23**(2), 65–75.

Overington, J. P., Al-Lazikani, B., and Hopkins, A. L. (2006). How many drug targets are there? *Nature Reviews Drug Discovery*, **5**(12), 993–996.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy.

Scheife, R. T., Hines, L. E., Boyce, R. D., Chung, S. P., Momper, J. D., Sommer, C. D., Abernethy, D. R., Horn, J. R., Sklar, S. J., Wong, S. K., Jones, G., Brown, M. L., Grizzle, A. J., Comes, S., Wilkins, T. L., Borst, C., Wittie, M. A., and Malone, D. C. (2015). Consensus recommendations for systematic evaluation of drug–drug interaction evidence for clinical decision support. *Drug Safety*, **38**(2), 197–206.

Smith, C. L. and Eppig, J. T. (2009). The mammalian phenotype ontology: enabling robust annotation and comparative analysis. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, **1**(3), 390–399.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K. P., Kuhn, M., Bork, P., Jensen, L. J., and von Mering, C. (2014). STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, **43**(D1), D447–D452.

Szklarczyk, D., Santos, A., von Mering, C., Jensen, L. J., Bork, P., and Kuhn, M. (2015). STITCH 5: augmenting protein–chemical interaction networks with tissue and affinity data. *Nucleic Acids Research*, **44**(D1), D380–D384.

Vazquez, A., Flammini, A., Maritan, A., and Vespignani, A. (2003). Global protein function prediction from protein-protein interaction networks. *Nature Biotechnology*, **21**(6), 697–700.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. (2017). Graph attention networks.

Verma, N., Boyer, E., and Verbeek, J. (2017). Feastnet: Feature-steered graph convolutions for 3d shape analysis.

Wang, C. and Kurgan, L. (2018). Review and comparative assessment of similarity-based methods for prediction of drug–protein interactions in the druggable human proteome. *Briefings in Bioinformatics*, **20**(6), 2066–2087.