Improving target-disease association prediction through a graph neural network with credibility information

Chang Liu[1,†], Cuinan Yu[2,†], Yipin Lei[1,†],
Kangbo Lyu[1], Tingzhong Tian[1], Qianhao Li[3],
Dan Zhao[1,*], Fengfeng Zhou[2,*], and Jianyang Zeng[1,*]

[1] *Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing 100084, China*
[*] *E-mail: zhaodan2018@tsinghua.edu.cn (D.Z.), zengjy321@tsinghua.edu.cn (J.Z.)*
[2] *Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, College of Computer Science and Technology, Jilin University, Changchun, Jilin 130012, China*
[*] *E-mail: FengfengZhou@gmail.com (F.Z.)*
[3] *Silexon AI Technology Co., Ltd., Nanjing, Jiangsu Province, China*

Identifying effective target-disease associations (TDAs) can alleviate the tremendous cost incurred by clinical failures of drug development. Although many machine learning models have been proposed to predict potential novel TDAs rapidly, their credibility is not guaranteed, thus requiring extensive experimental validation. In addition, it is generally challenging for current models to predict meaningful associations for entities with less information, hence limiting the application potential of these models in guiding future research. Based on recent advances in utilizing graph neural networks to extract features from heterogeneous biological data, we develop CreaTDA, an end-to-end deep learning-based framework that effectively learns latent feature representations of targets and diseases to facilitate TDA prediction. We also propose a novel way of encoding credibility information obtained from literature to enhance the performance of TDA prediction and predict more novel TDAs with real evidence support from previous studies. Compared with state-of-the-art baseline methods, CreaTDA achieves substantially better prediction performance on the whole TDA network and its sparse sub-networks containing the proteins associated with few known diseases. Our results demonstrate that CreaTDA can provide a powerful and helpful tool for identifying novel target-disease associations, thereby facilitating drug discovery.

Keywords: target-disease association, graph neural network, credibility information, drug discovery.

1. Introduction

The development of a drug generally takes more than five years and costs more than $4.5 billion,[27] with most of the resources sunk into clinical failures that happen at later stages of drug development.[11] To alleviate the massive cost of drug development, it is crucial to determine credible (i.e., to identify plausible drug targets for a specific disease) at the beginning of the drug development process.

Based on the latent feature representations and similarities between targets and diseases

[†] These authors contributed equally.

learned from sufficient data, machine learning (ML) models can "predict" potential target-disease associations (TDAs) useful for future studies. For example, a range of ML classifiers trained based on TDA data from the Open Targets platform have been used to predict novel TDAs.[12] A tensor factorization method has also been proposed to reconstruct a drug-target-disease network by integrating drug-drug, target-target, and disease-disease similarity matrices as multi-view auxiliary networks.[4] However, the underlying Tucker tensor model generally suffers from linearity and data sparsity,[5] thus undermining its prediction capacity.

Graph neural networks (GNNs) are nonlinear ML models that generalize convolutional neural networks (CNNs) to graph/network data,[10] combined with information passing and aggregation techniques.[13] Moreover, recent advances in generalizing GNNs to heterogeneous network (HN) data have brought considerable performance improvement.[15,28,32] Since the relation prediction tasks such as target-disease association (TDA) prediction can be viewed as link prediction on networks of biological data, GNNs can theoretically be utilized as high-capacity models for these tasks. Indeed, NeoDTI, a GNN that predicts DTIs from an HN, outperformed state-of-the-art DTI prediction models under several challenging and realistic scenarios.[30]

Nevertheless, these machine learning methods still have the following two shortcomings:

First, human labor is generally needed to verify the prediction results by searching for supporting evidence from literature or conducting wet-lab experiments. Without a gauge of the credibility of these predictions, the amount of human effort needed in these analyses would be daunting, undermining the level of autonomy of the prediction pipeline and thus failing to address the lengthiness and costliness problem of drug development.

Second, *exposure bias* may heavily influence model performance. Exposure bias is a phenomenon in recommendation systems where users are only exposed to a part of specific items so that the unobserved interactions do not always represent the negative preferences.[6] In such a scenario, models are inclined to predict more relations between entities with more available information. However, the failure to produce meaningful predictions for entities with less information restricts the application potential of the models in guiding future research. Moreover, it is generally more difficult for the models to learn the latent feature representations of entities with less information, hence undermining their overall prediction performance.

In this paper, we propose CreaTDA (CRedibility-Encoding grAph neural network for TDA prediction), an end-to-end deep learning-based framework, to perform TDA prediction. In addition to exploiting the structured heterogeneous data in the form of biological networks, CreaTDA fully takes advantage of unstructured data in the form of entity co-occurrence in the literature, which encodes the credibility of the interactions/associations between entities. We showed that CreaTDA (i) achieved superior performance over baseline models on the TDA prediction task and (ii) generated novel predictions with higher credibility and more literature support, and (iii) exhibited robustness to the effect of exposure bias. These results suggested that CreaTDA can provide a helpful tool for drug target identification and benefit the whole drug development process.

2. Methods

2.1. *The heteroneneous network data*

CreaTDA uses heterogeneous network (HN) data as input. We first give a formal definition of an HN:

Definition 1 (Heterogeneous Network) An HN is a directed/undirected graph $G = (V, E)$, where each node $v \in V$ is of a node type from a node type set O, and each edge $e \in E, E \subset V \times V \times R$ is of an edge type from an edge type set R.

The HN used in our framework is an undirected graph with the node type set $O =${*drug*, *target (protein)*, *side effect*, *disease*} and the edge type set $R =${*drug-drug-structure-similarity*, *protein-protein-sequence-similarity*, *drug-drug-interaction*, *drug-side-effect-association*, *drug-protein-interaction*, *drug-disease-association*, *protein-disease-association*, *protein-protein-interaction*}. Note that we will use the terms "protein" and "target" interchangeably in the remaining parts of this paper.

Here, our individual networks (defined by specific edge types) are adopted from Luo et al.,[20] including:

- A drug-protein interaction network and a drug-drug interaction network, derived from Drugbank Version 3.0;[17]
- A protein-protein interaction network, extracted from the HPRD database Release 9;[16]
- A drug-disease association network and a protein-disease association (TDA) network, derived from the Comparative Toxicogenomics Database;[8]
- A drug-side-effect association network, derived from the SIDER database Version 2;[18]
- A drug-drug-structure-similarity network, computed using RDKit (`rdkit.org`) according to the Dice similarity of the Morgan fingerprints with radius 2;[24]
- A protein-protein-sequence-similarity network, computed according to the Smith-Waterman scores.[29]

The *association* and *interaction* networks have 0/1 binary edge weights. The 1 values indicate that the entailed associations/interactions exist in the corresponding database. The 0 values indicate either (i) the entailed associations/interactions are established not to exist or (ii) evidence supporting the associations/interactions is lacking. The edges of the *similarity* networks are weighted with real values. With all the networks stored as adjacency matrices, the final HN hosts 12015 nodes, including 1512 targets, 5603 diseases, 708 drugs, and 4102 side effects.

2.2. *The CreaTDA pipeline*

CreaTDA first computes node embeddings that encode the topology of the HN, then uses these embeddings to reconstruct individual networks that encode credibility (Fig. 1), imputing the original 0 values. We describe these two components of CreaTDA below.

2.2.1. *Obtaining node embeddings*

In our framework (Fig. 1), node embeddings are computed via a GNN through two steps: (i) passing and aggregating information for each node through edge-type-specific neighbors and (ii) updating node embeddings. These steps are formally defined as follows:

Definition 2 (Neighborhood information passing and aggregation) Given an HN G, an initial

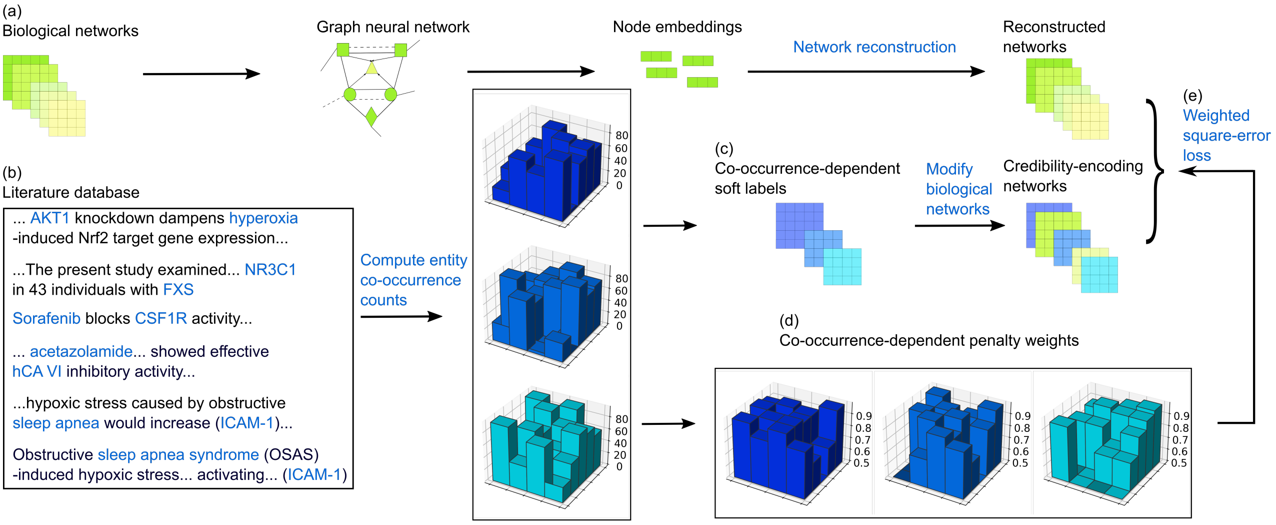

(a)
Biological networks
Graph neural network
Node embeddings
Network reconstruction
Reconstructed networks
(e)
Weighted square-error loss
(b)
Literature database
... AKT1 knockdown dampens hyperoxia -induced Nrf2 target gene expression...
...The present study examined... NR3C1 in 43 individuals with FXS
Sorafenib blocks CSF1R activity...
... acetazolamide... showed effective hCA VI inhibitory activity...
...hypoxic stress caused by obstructive sleep apnea would increase (ICAM-1)...
Obstructive sleep apnea syndrome (OSAS) -induced hypoxic stress... activating... (ICAM-1)
Compute entity co-occurrence counts
80
60
40
20
0
(c)
Co-occurrence-dependent soft labels
Modify biological networks
Credibility-encoding networks
(d)
Co-occurrence-dependent penalty weights
0.9
0.8
0.7
0.6
0.5

Fig. 1. Overview of CreaTDA. CreaTDA uses a graph neural network to **(a)** obtain node embeddings from individual biological networks that encode the network topology. CreaTDA further encodes credibility by **(b)** computing entity co-occurrence counts in the PubMed database and then transforming these raw counts into co-occurrence-dependent **(c)** soft labels (Eq. 3) and **(d)** penalty weights (Eq. 4). **(e)** CreaTDA reconstructs the credibility-encoding networks containing the soft labels by minimizing a weighted square-error loss derived based on the penalty weights (Eq. 5).

node embedding function $f^0 : V \to \mathbb{R}^d$ maps each node to an initial node embedding, and an edge embedding function $m : E \to \mathbb{R}$ maps each edge $e \in E$ to a corresponding value in the network, which can be represented as an adjacency matrix. The information a_v of node $v \in V$ is then aggregated from its neighborhood as follows:

$$a_v = \sum_{\substack{r \in R, u \in N_r(v) \\ e=(u,v,r) \in E}} \frac{m(e)}{Z_{v,r}} (W_r f^0(u) + b_r), \tag{1}$$

where $N_r(v) = \{u | u \in V, u \neq v, (u, v, r) \in E\}$ denotes the nodes connected to $v \in V$ via an edge of type $r \in R$, which are also defined as the "r-neighbors of v." $W_r \in \mathbb{R}^{d \times d}, b_r \in \mathbb{R}^d$ denote the model parameters depending only on the edge type, and $Z_{v,r} = \sum_{u \in N_r(v), e=(u,v,r)} m(e)$ denotes a normalization term. In CreaTDA, f^0 is initialized as a truncated normal sampler with mean 0, standard deviation 0.1, minimum cutoff value -0.2, and maximum cutoff value 0.2.

In other words, for each edge-type r, the embeddings of the $r-$neighbors of v are passed through a linear transformation and then weighed by the normalized edge weights $\frac{m(e)}{Z_{v,r}}$. After that, the results over all edge types are summed.

Definition 3 (Node embedding updating) Using a_v obtained from Eq. 1, the initial embeddings $f^0(v)$ are updated as follows:

$$f^1(v) = g(ReLU(W^1(f^0(v) \| a_v) + b^1)), \tag{2}$$

where "$\|$" denotes the concatenation operation, $ReLU(x) = \max\{0, x\}$, $g(\cdot)$ denotes the ℓ_2 normalization operation, and $W^1 \in \mathbb{R}^{d \times (2d)}$ and $b^1 \in \mathbb{R}^d$ denote global parameters shared by all nodes.

For each node v, its neighborhood information and initial embedding both contribute to its updated embedding, thus allowing the network topology information to be encoded.

2.2.2. *Reconstructing the credibility-encoding networks*

We seek to improve the credibility of the predicted TDAs, i.e., the reproducibility of the results indicating the TDAs, by encoding credibility information into the CreaTDA framework, such that credibility can be learned as part of the latent feature representations of nodes. While the credibility of an interaction/association is elusive to quantify, it can be reflected by the abundance of literature documenting this interaction/association, which can be approximated by the quantity of literature in which the two interacting/associated entities both appear.

We curated about three million papers in the PubMed database maintained by the United States National Library of Medicine (NLM).[23] The number of papers that a *drug-protein*, *protein-disease*, or *drug-disease* pair co-occurs in was computed by sub-string matching using the Trie hashing algorithm (see Supplementary Information for more details). These co-occurrence counts were then organized into co-occurrence matrices $C_r, r \in R_c = \{\textit{drug-protein}, \textit{protein-disease}, \textit{drug-disease}\}$, where $C_r[i, j]$ represents the number of co-occurring papers for entities i and j associated with edge-type r. We assumed that $C_r[i, j]$ is positively correlated with the credibility of the interaction/association between entities i and j. Hence, by incorporating C_r into CreaTDA, the notion of credibility can be introduced.

Here, we formally describe a method of integrating C_r into the CreaTDA framework. We first give mathematical definitions of the key terms used:

Definition 4 (Co-occurrence-dependent soft label) For an edge $e = (i, j, r)$ of edge-type $r \in R_c$ between entities i and j, its soft label is defined as:

$$l(e) = \sigma(C_r[i, j] + \alpha) \cdot m(e) \tag{3}$$

where α stands for a hyperparameter, $\sigma(x) = \frac{1}{1+e^{-x}}$, and $m(e)$ represents the edge embedding function defined in Eq. 1.

Definition 5 (Co-occurrence-dependent penalty weight) For an edge $e = (i, j, r)$ of edge-type $r \in R_c$ between entities i and j, the penalty weight of the reconstruction loss of e is defined as:

$$w(e) = \sigma(C_r[i, j] + \beta) \cdot m(e) + (1 - m(e)) \tag{4}$$

where β stands for a hyperparameter and $m(e)$, $\sigma(x)$ are the same as defined in Eq. 3.

In the implementation of CreaTDA, α and β are set to $\ln 3$ and 0, respectively, as they yielded the best performance according to the cross-validation results (Section 3.1).

The information in C_r is then incorporated in the network reconstruction step to encode the credibility information of TDAs:

Definition 6 (Credibility-encoding network reconstruction) For the parameter set $\Theta = \{f^0, W_r, b_r, G_r, H_r, W^1, b^1\}$, the optimization objective of CreaTDA is:

$$\begin{aligned}\min_{\Theta} & \sum_{r \in R \setminus R_c} \sum_{\substack{u,v \in V \\ e=(u,v,r) \in E}} (m(e) - f^1(u)^T G_r H_r^T f^1(v))^2 \\ &+ \sum_{r \in R_c} \sum_{\substack{u,v \in V \\ e=(u,v,r) \in E}} w(e)(l(e) - f^1(u)^T G_r H_r^T f^1(v))^2,\end{aligned} \tag{5}$$

where $m(e)$ denotes the edge embedding function (Eq. 1), $w(e)$ denotes the co-occurrence-dependent penalty weight (Eq. 4), $l(e)$ denotes the co-occurrence-dependent soft label (Eq.

3), and $G_r, H_r \in \mathbb{R}^{d \times k}$ denote the edge-type specific projection matrices. In the implementation of CreaTDA, the ℓ_2-regularization terms on f^0, W_r, G_r, H_r, and W^1 are also summed. In addition, if $r \in \{$*drug-drug-structure-similarity*, *protein-protein-sequence-similarity*, *drug-drug-interaction*, *protein-protein-interaction*$\}$, where the corresponding adjacency matrix is symmetric, the constraint $G_r = H_r$ is imposed to enforce such a symmetry.

The network reconstruction step projects the node embeddings $f^1(\cdot)$ onto the edge-type-specific vector spaces such that the matrix products of the projected vectors best match the corresponding individual networks. Notably, the credibility information is *not* introduced for the negative interactions/associations in the HN, that is, when $m(e) = 0$, $l(e)$ and $w(e)$ are set to 0 and 1 (Eqs. 3 and 4), respectively, thus preventing the potential data leakage problem during the cross-validation process.

2.3. *Ablation studies*

To show that the integration of C_r into the CreaTDA framework is necessary for achieving better performance, we developed four models as the control in our ablation studies to nullify the credibility information encoded in the labels and/or weights: CreaTDA_og (no credibility encoded), CreaTDA_rl (random soft labels), CreaTDA_rw (random penalty weights), and CreaTDA_rlrw (both random soft labels and random penalty weights). More details about the mathematical definitions of these control models can be found in the Supplementary Information.

3. Results

3.1. *CreaTDA yields superior performance in predicting target-disease associations*

While the objective of CreaTDA is to reconstruct the HN, TDA prediction can be considered a binary classification task (i.e., whether an association exists or not). Though we used the modified labels for the optimization objective (Eq. 5), we still measured the prediction performance in terms of the area under the precision-recall curve (AUPR) and the area under the receiver operating characteristic curve (AUROC), using the *original* binary TDA labels as ground truth. We observed that the ratio between the numbers of "1"- and "0"-entries in the network is 0.232, suggesting data imbalance. As stated in previous works, AUPR generally presents a more informative metric than AUROC on the performance of models on those imbalanced datasets.[9,26]

Table 1. Cross-validation results, measured in terms of AUROC and AUPR, in the form of "mean ± standard deviation" over ten rounds of entry-wise cross-validation and cluster-wise cross-validation (Section 3.1), respectively. The results where CreaTDA outperformed all baseline methods are presented in boldface.

	GTN	RGCN	HGT	DTINet	CreaTDA
Entry-wise cross-validation					
AUROC	0.953 ± 0.002	0.974 ± 0.001	0.950 ± 0.002	0.859± 2e-5	**0.986 ± 2e-4**
AUPR	0.822 ± 0.017	0.915 ± 0.004	0.846 ± 0.006	0.658± 1e-5	**0.967 ± 5e-4**
Cluster-wise cross-validation					
AUROC	0.725 ± 0.003	0.738 ± 0.014	0.569 ± 0.012	0.815 ± 0.007	0.814 ± 0.007
AUPR	0.397 ± 0.004	0.332 ± 0.013	0.211 ± 0.006	0.503 ± 0.018	$\mathbf{0.516 \pm 0.016}$

We performed five-fold cross-validation, during which we conducted a random stratified splitting on the entries of the TDA matrix, which were divided into five folds, preserving the global positive-to-negative ratio as much as possible in each fold. For each of the five iterations, we sequentially chose one fold as test data and sampled 10% of the remaining four folds as validation data for hyperparameter tuning (the remaining 90% formed the training set). We refer to this cross-validation scheme as *entry-wise cross-validation.*

We computed the average AUROC and AUPR scores on the test sets of the five iterations as the performance statistics for one round of cross-validation. To account for the randomness effect, we performed ten rounds of five-fold cross-validation (with different random states) and recorded the means and standard deviations of the performance statistics (Table 1).

We compared the performance of CreaTDA to those of several baseline methods that have reached state-of-the-art performance on heterogeneous graph prediction tasks, including GTN,[32] RGCN,[28] HGT,[15] and DTINet[20] (see Supplementary Information for more details). We found that CreaTDA significantly outperformed all the baseline methods (Table 1), suggesting that CreaTDA can better learn the latent feature representations of the underlying network topology of the given HN.

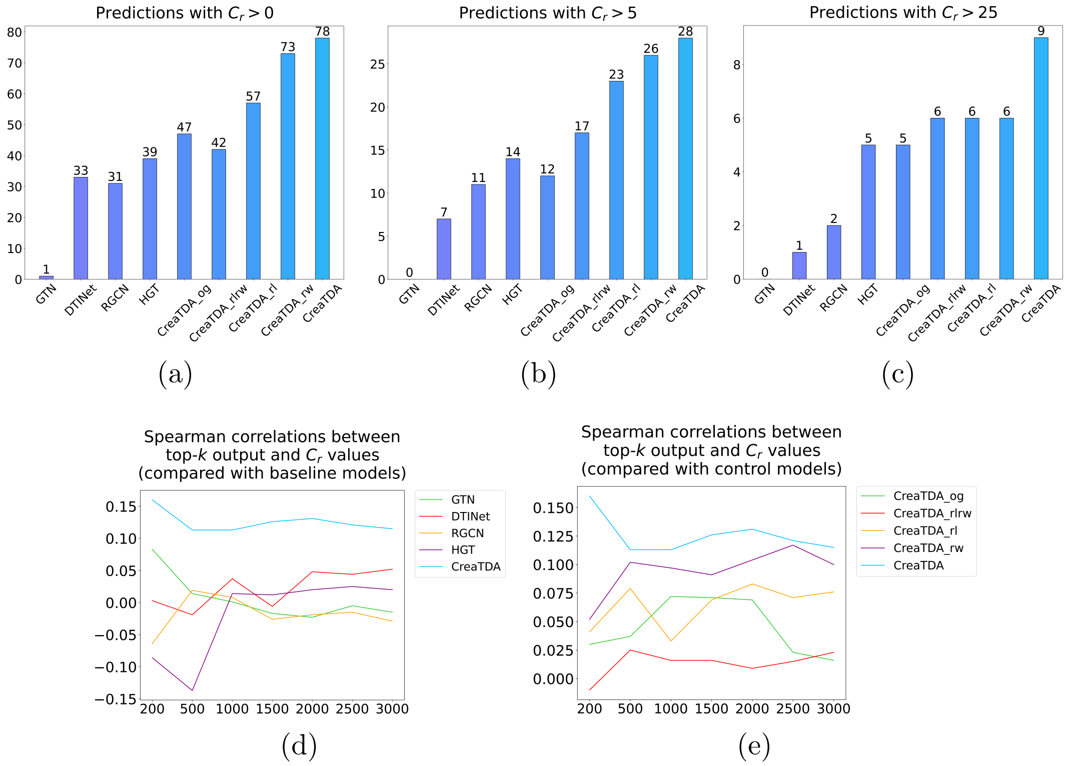
However, with CreaTDA yielding near-perfect performance, the prediction task may be trivial. Indeed, "similar" TDAs may appear in both training and test sets, thus constituting "easy" predictions that inflated the performance of the models. To more accurately gauge the performance and generalization capacity of the models, we conducted additional tests by reducing the similarity between training and test data. Specifically, we first performed agglomerative clustering on the disease entities according to the Jaccard similarities between their association profiles, i.e., the corresponding columns in the *protein-disease-association* adjacency matrix. We then developed a new cross-validation scheme by partitioning the resulting *clusters* of columns into training, validation, and test sets. The ratios between the sizes of the three datasets and the ratio between positive and negative samples in each dataset were roughly the same as those in the previous entry-wise cross-validation procedure. We refer to this new cross-validation scheme as *cluster-wise cross-validation.*

Table 1 shows that all models had a significant drop in performance when switching from entry-wise to cluster-wise cross-validation. However, CreaTDA still took the lead in performance (though DTINet yielded a comparable AUROC score with CreaTDA, the former achieved a poorer AUPR score), further verifying the superior predictive power of CreaTDA.

We also found that all control models yielded performance inferior to CreaTDA on the cluster-wise cross-validation (Supplementary Table 1), suggesting that the encoded credibility information in both the designed labels and weights can effectively advance CreaTDA to accurately capture the latent feature representations of the underlying network topology.

3.2. *CreaTDA improves the credibility of TDA predictions*

To evaluate the credibility of the novel predictions of CreaTDA, we investigated their corresponding C_r values, which approximate the abundance of literature documenting the entailed TDAs (Section 2.2.2). Here, the "novel" predictions were obtained through the following process: (i) training CreaTDA on the whole HN using the hyperparameters that yielded the best performance in the cluster-wise cross-validation scheme (Section 3.1); (ii) selecting those "significant" predictions whose output values in the reconstructed TDA matrix were greater



(a) (b) (c)

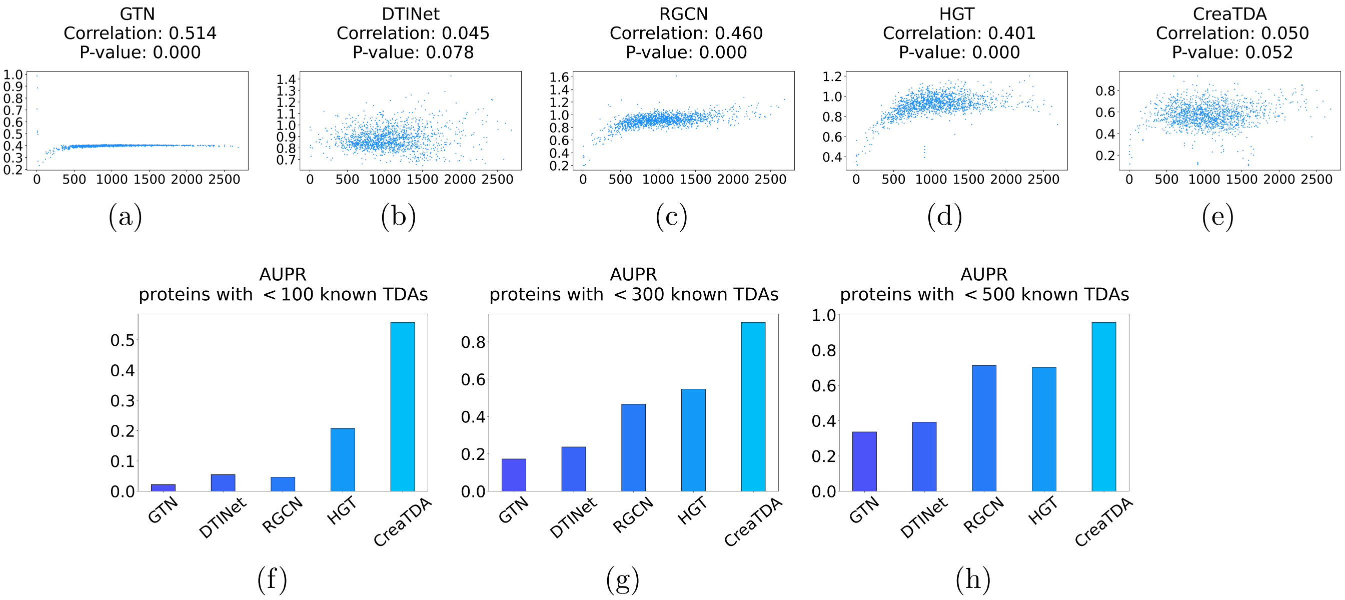
(d) (e)

Fig. 2. Examining the credibility of model predictions. **(a)**, **(b)**, and **(c)** document the numbers of predictions among the top-200 novel predictions with $C_r > 0, 5, 25$, respectively. **(d)** and **(e)** plot the Spearman correlations between the output values of the top-k ($k = 200, 500, 1000, 1500, 2000, 2500, 3000$) predictions and their corresponding C_r values, with **(d)** comparing CreaTDA with the baseline models and **(e)** comparing CreaTDA with the four control models developed in our ablation study. The P-values of the correlations, calculated using the *sklearn* package,[22] can be found in Supplementary Table 2.

than $\mu + 2\sigma$, where μ and σ stand for the mean and the standard deviation of the predicted values of elements in each row, respectively; and (iii) choosing the "novel" predictions, which were assigned with the label "0" in the original TDA matrix (i.e., $m(e) = 0$), from the above "significant" predictions. Since these novel predictions had edge weights equal to 0, their corresponding C_r values were not encoded (Eqs. 3 and 4), hence precluding data leakage.

We first examined the C_r values of the novel predictions with the top-200 output values. We found that compared with all baseline and control models, among their corresponding top-200 novel predictions, CreaTDA predicted more novel TDAs with C_r values greater than 0, 5, and 25, respectively (Fig. 2a-2c). Such results showed that CreaTDA could produce novel predictions with more evidence support from PubMed, even though their credibility information was not encoded in CreaTDA during the prediction process.

We next examined the Spearman correlation between the output and the corresponding C_r values of the top-k predictions. We found that CreaTDA yielded a stronger correlation than all baseline (Fig. 2d) and control models (Fig. 2e). We also conducted a hypothesis test (two-sided t-test), in which the null hypothesis meant that the output and C_r values were uncorrelated. We found that CreaTDA yielded overall lower P-values than all baseline and control models (Supplementary Table 2). Here, a stronger correlation (with a lower P-value) indicated that the model predicted TDAs with higher credibility (i.e., larger C_r values). Such results illustrated that the novel TDAs predicted by CreaTDA were more likely to be valid.



(a) (b) (c) (d) (e)

(f) (g) (h)

Fig. 3. Examining the robustness against the effect of exposure bias for different models. **(a)-(e)** plot the row-wise maximum values over the 0-labeled entries of the reconstructed TDA matrix (y-axis) against the row-wise sums of the original TDA matrix (x-axis) for the baseline models and CreaTDA. The Spearman correlations between these two vectors and their P-values, calculated using the *sklearn* package, are also reported. **(f)-(h)** present the AUPR scores on the sparse sub-networks of the whole TDA network containing proteins associated with few known TDAs.

3.3. *CreaTDA is robust to the effect of exposure bias*

In this section, we showed that CreaTDA was robust to the effect of exposure bias, a common phenomenon in recommendation systems where the unobserved interactions are often misrepresented as negative preferences.[6] This phenomenon also arises in our TDA prediction task, where those TDAs with 0-labels in the input data are not necessarily "negative" associations. Due to exposure bias, the models generally produce fewer meaningful TDA predictions for those proteins/diseases with few known TDAs and often have difficulty learning their latent feature representations. To investigate the robustness of the models against the effect of exposure bias, we computed the Spearman correlation between the row-wise maximum values over the 0-labeled entries of the reconstructed TDA matrix and the row-wise sums of the original TDA matrix (i.e., the number of diseases associated with the corresponding protein) for the baseline models and CreaTDA. We found that CreaTDA yielded a significantly lower correlation than GTN, RGCN, and HGT, only slightly exceeding the correlation yielded by DTINet (Fig. 3a-3e). Here, a strong correlation indicates two possible drawbacks: (i) the predicted values of TDAs depend heavily on the amount of known information, i.e., the number of diseases known to be associated with the involved protein; and (ii) the top predictions of the model are likely to leave out biologically significant TDAs for those proteins with less available information. Therefore, the above results indicated that with a significantly weaker correlation, CreaTDA suffered less from these two drawbacks.

We then examined the prediction performance (AUPR scores) on the sparse sub-networks of the original TDA network for different models trained on the whole HN. More specifically, we selected those rows of the original TDA matrix with a sum less than $100, 300$, and 500, respectively, to simulate three sparse sub-networks. We found that CreaTDA consistently achieved higher AUPR scores than the baseline methods on these sparse sub-networks (Fig. 3f-3h). Here, a higher AUPR score indicated that for proteins with few known TDAs, CreaTDA could generate more accurate predictions and better learn their latent feature representations. These results suggested that CreaTDA is robust to the effect of exposure bias and thus can provide a helpful tool to predict novel TDAs, especially for those proteins with less information.

3.4. *CreaTDA is able to predict novel TDAs with literature support*

To show that CreaTDA can help scientists find reliable TDAs, we validated the top-200 novel predictions of CreaTDA by searching for literature support and presented several representative cases (see the complete list of the top-200 predictions in Supplementary Table 3).

3.4.1. *CreaTDA reveals potential targets with literature support*

Respiratory syncytial virus (RSV) is a major cause of severe lower respiratory tract illness in children, including bronchiolitis. CreaTDA predicted an association between bronchiolitis and the epidermal growth factor receptor (EGFR). Previous studies showed that EGFR interacts with the RSV 2-20 F protein in a strain-specific manner and is thus a potential target for RSV diseases,[7] which exactly supported our prediction result. We also extended to a general category of virus diseases as an example. CreaTDA predicted an association between virus diseases and vascular endothelial growth factor-A (VEGF-A, also known as VEGF), a principal pro-angiogenic factor. This association can also be supported by previous research,[1] which illustrated that viruses, e.g., the human papillomavirus[19] and herpes simplex virus-1[31] exploit cell signaling mechanisms to upregulate VEGF expression and thus benefit their pathogenesis. In addition, recent research on COVID-19 has shown that anti-VEGF medication may be a potential treatment for those critically ill patients.[25] These validation results showed that CreaTDA could successfully identify novel targets critically involved in specific diseases.

3.4.2. *CreaTDA provides new perspectives for understanding diseases*

CreaTDA predicted an association between the fragile X syndrome (FXS) and the glucocorticoid receptor gene NR3C1. This prediction can be supported by previous research, which showed that the G allele in the BclI polymorphism of NR3C1 has a protective effect among female individuals against FXS and is associated with altered patterns of the anxiety/fear network of the brain.[2] Hence, our prediction about NR3C1 may help understand the diverse clinical outcomes associated with FXS and thus inspire effective therapies for individuals with specific polymorphisms.

3.4.3. *CreaTDA discovers new biomarkers for disease studies*

CreaTDA detected an association between sleep apnea syndromes and the intercellular adhesion molecule 1 (ICAM-1). ICAM-1 has been known as a marker widely used in studies on obstructive sleep apnea syndrome (OSAS) to investigate inflammation.[3] In a previous study, scientists found that OSAS patients displayed a significant decrease in ICAM-1 level after nasal continuous positive airway pressure (nCPAP) therapy, suggesting that OSAS-induced hypoxia activates ICAM-1.[21] CreaTDA also predicted an association between retinopathy of prematurity (ROP) and myeloperoxidase (MPO). This finding was consistent with a previous result that MPO is one of the nine proteins with the potential to increase the ROP risk.[14] All these findings verified that CreaTDA could provide an effective tool to identify novel biomarkers useful in clinical studies.

4. Conclusion

In this paper, we presented CreaTDA, an end-to-end deep learning-based framework to predict novel TDAs. CreaTDA first learns the node embeddings that encode features of the network topology and then reconstructs the modified biological networks with the encoded credibility information of TDAs. We showed that compared with state-of-the-art baseline methods, CreaTDA achieved superior performance on both the standard TDA prediction task and a more challenging task with a low similarity between training and test data. Moreover, comprehensive tests demonstrated that CreaTDA could predict novel TDAs with improved credibility and more literature support. In addition, we discovered that CreaTDA was robust to the effect of exposure bias and maintained decent performance for those entities with less information. All these results suggest CreaTDA can provide a powerful and helpful tool to advance the drug discovery process.

Acknowledgements

This work was supported in part by the National Key Research and Development Program of China (2021YFF1201300), the National Natural Science Foundation of China (61872216, T2125007 to JZ, 31900862 to DZ), the Turing AI Institute of Nanjing, the Tsinghua-Toyota Joint Research Fund, the Senior and Junior Technological Innovation Team (20210509055RQ), and the Jilin Provincial Key Laboratory of Big Data Intelligent Computing (20180622002JC). The authors thank Dr. Fangping Wan for helpful discussions.

Availablility

The source code, data, and supplementary information of this study can be accessed at `https://github.com/Dr-Patient/CreaTDA`.

References

1. Alkharsah KR. Vegf upregulation in viral infections and its possible therapeutic implications. *International journal of molecular sciences*, 19(6):1642, 2018.
2. Bruno JL, et al. Glucocorticoid regulation and neuroanatomy in fragile x syndrome. *Journal of Psychiatric Research*, 134:81, 2021.
3. Carpagnano GE, et al. Systemic and airway inflammation in sleep apnea and obesity: the role of icam-1 and il-8. *Translational Research*, 155(1):35, 2010.
4. Chen H, et al. Modeling relational drug-target-disease interactions via tensor factorization with multiple web sources. In *The World Wide Web Conference*, 218–227. 2019.
5. —. Learning data-driven drug-target-disease interaction via neural tensor network. In *International Joint Conference on Artificial Intelligence (IJCAI)*. 2020.
6. Chen J, et al. Bias and debias in recommender system: A survey and future directions. *arXiv preprint arXiv:201003240*, 2020.
7. Currier MG, et al. Egfr interacts with the fusion protein of respiratory syncytial virus strain 2-20 and mediates infection and mucin expression. *PLoS pathogens*, 12(5):e1005622, 2016.
8. Davis AP, et al. The comparative toxicogenomics database: update 2017. *Nucleic acids research*, 45(D1):D972, 2017.
9. Davis J, et al. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, 233–240. 2006.

10. Defferrard M, et al. Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems*, 29, 2016.
11. Failli M, et al. Prioritizing target-disease associations with novel safety and efficacy scoring methods. *Scientific reports*, 9(1):1, 2019.
12. Ferrero E, et al. In silico prediction of novel therapeutic targets using gene–disease association data. *Journal of translational medicine*, 15(1):1, 2017.
13. Hamilton W, et al. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.
14. Holm M, et al. Systemic inflammation-associated proteins and retinopathy of prematurity in infants born before the 28th week of gestation. *Investigative ophthalmology & visual science*, 58(14):6419, 2017.
15. Hu Z, et al. Heterogeneous graph transformer. In *Proceedings of The Web Conference 2020*, 2704–2710. 2020.
16. Keshava Prasad T, et al. Human protein reference database—2009 update. *Nucleic acids research*, 37(suppl_1):D767, 2009.
17. Knox C, et al. Drugbank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic acids research*, 39(suppl_1):D1035, 2010.
18. Kuhn M, et al. A side effect resource to capture phenotypic effects of drugs. *Molecular systems biology*, 6(1):343, 2010.
19. Li G, et al. Overexpression of human papillomavirus (hpv) type 16 oncoproteins promotes angiogenesis via enhancing hif-1α and vegf expression in non-small cell lung cancer cells. *Cancer letters*, 311(2):160, 2011.
20. Luo Y, et al. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nature communications*, 8(1):1, 2017.
21. Ohga E, et al. Effects of obstructive sleep apnea on circulating icam-1, il-8, and mcp-1. *Journal of applied physiology*, 94(1):179, 2003.
22. Pedregosa F, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825, 2011.
23. Roberts RJ. Pubmed central: The genbank of the published literature, 2001.
24. Rogers D, et al. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742, 2010.
25. Sahebnasagh A, et al. Anti-vegf agents: As appealing targets in the setting of covid-19 treatment in critically ill patients. *International Immunopharmacology*, 101:108257, 2021.
26. Saito T, et al. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3):e0118432, 2015.
27. Schlander M, et al. How much does it cost to research and develop a new drug? a systematic review and assessment. *PharmacoEconomics*, 39(11):1243, 2021.
28. Schlichtkrull M, et al. Modeling relational data with graph convolutional networks. In *European semantic web conference*, 593–607. Springer, 2018.
29. Smith TF, et al. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195, 1981.
30. Wan F, et al. Neodti: neural integration of neighbor information from a heterogeneous network for discovering new drug–target interactions. *Bioinformatics*, 35(1):104, 2019.
31. Wuest TR, et al. Vegf-a expression by hsv-1–infected cells drives corneal lymphangiogenesis. *Journal of Experimental Medicine*, 207(1):101, 2010.
32. Yun S, et al. Graph transformer networks. *Advances in neural information processing systems*, 32, 2019.