Mohamed, S. K., Nováček, V., & Nounu, A. (2019). Discovering Protein Drug Targets Using Knowledge Graph Embeddings. *Bioinformatics*, [btz600]. https://doi.org/10.1093/bioinformatics/btz600

Peer reviewed version

Link to published version (if available):
10.1093/bioinformatics/btz600

Link to publication record in Explore Bristol Research
PDF-document

University of Bristol - Explore Bristol Research
General rights

Predicting Protein Drug Targets Using Knowledge Graph Embeddings

Sameh K. Mohamed [1,2*], Vít Nováček [1,2] and Aayah Nounu [3]

Abstract

Motivation: Computational approaches for predicting drug-target interactions (DTIs) can provide valuable insights into the drug mechanism of action. DTI predictions can help to quickly identify new promising (on-target) or unintended (off-target) effects of drugs. However, existing models face several challenges. Many can only process a limited number of drugs and/or have poor proteome coverage. The current approaches also often suffer from high false positive prediction rates.

Results: We propose a novel computational approach for predicting drug target proteins. The approach is based on formulating the problem as a link prediction in knowledge graphs (robust, machine-readable representations of networked knowledge). We use biomedical knowledge bases to create a knowledge graph of entities connected to both drugs and their potential targets. We propose a specific knowledge graph embedding model, TriModel, to learn vector representations (i.e. embeddings) for all drugs and targets in the created knowledge graph. These representations are consequently used to infer candidate drug target interactions based on their scores computed by the trained TriModel model. We have experimentally evaluated our method using computer simulations and compared it to five existing models. This has shown that our approach outperforms all previous ones in terms of both area under ROC and precision-recall curves in standard benchmark tests.

Availability: The data, predictions, and models are available at: drugtargets. insight-centre.org

Contact: sameh.kamal@insight-centre.org

1   Introduction

The development of drugs has a long history [1]. Until quite recently, pharmacological effects were often discovered using primitive trial and error procedures, such as applying plant extracts on living systems and observing the outcomes. Later, the

---

[1] Data Science Institute, College of Engineering and Informatics, NUI Galway, Galway, Ireland

[2] Insight Centre for Data analytics, NUI Galway, Galway, Ireland

[3] MRC Integrative Epidemiology Unit, Bristol Medical School, University of Bristol, Bristol, UK

*To whom correspondence should be addressed.

drug development process evolved to elucidating mechanisms of action of drug substances and their effects on phenotype. The ability to isolate pharmacologically active substances was a key step towards modern drug discovery [2, 3]. More recently, advances in molecular biology and biochemistry allowed for more complex analyses of drugs, their targets and their mechanisms of action. The study of drug targets has become very popular with the objective of explaining mechanisms of actions of current drugs and their possible unknown off-target activities. Knowing targets of potential clinical significance also plays a crucial role in the process of rational drug development. With such knowledge, one can design candidate compounds targeting specific proteins to achieve intended therapeutic effects.

However, a drug rarely binds only to the intended targets, and off-target effects are common [4]. This may lead to unwanted adverse effects [5], but also to successful drug re-purposing, i.e. use of approved drugs for new diseases [6]. To illustrate the impact off-target effects can have in new therapy development, let us consider *aspirin* that is currently being considered for use as a chemopreventive agent [7]. However, such a therapy would be hampered by known adverse side-effects caused by long-term use of the drug, such as bleeding of upper gastrointestinal tract [8]. After identifying the exact protein targets of *aspirin* that cause these adverse effects, the proteins can be targeted by newly developed and/or re-purposed drugs to avoid the unwanted side-effects of the proposed treatment.

Large-scale and reliable prediction of drug-target interactions (DTIs) can substantially facilitate development of such new treatments. Various DTI prediction methods have been proposed to date. Examples include chemical genetic [2] and proteomic methods [9] such as affinity chromatography and expression cloning approaches. These, however, can only process a limited number of possible drugs and targets due to the dependency on laboratory experiments and available physical resources. Computational prediction approaches have therefore received a lot of attention lately as they can lead to much faster assessments of possible drug-target interactions [10, 11].

The work of [10] was one of the first approaches to predict drug targets computationally. Their approach utilised a statistical model that infers drug targets based on a bipartite graph of both chemical and genomic information. The BLM-NII [11] model was developed to improve the previous approach by using neighbour-based interaction-profile inference for both drugs and targets. More recently, [12, 13] proposed a new way for predicting DTIs, where they have used a combination of drug similarity, target similarity and network-based inference. The COSINE [14] and NRLMF [15] models introduced the exclusive use of drug-drug and target-target similarity measures to infer possible drug targets. This has an advantage of being able to compute predictions even for drugs and targets with limited information about their interaction data. However, these methods only utilised a single measure to model components similarity. Other approaches such as the KronRLS-MKL [16] model used a linear combinations of multiple similarity measures to model the overall similarity between drugs and targets. Non-linear combinations were also explored in [11] and shown to provide better predictions.

Recently, [17] proposed a model called DNILMF that uses matrix factorisation to predict drug targets over drug information networks. This approach showed significant improvements over other methods on standard benchmarking datasets [17, 10].

All the previously discussed works were designed to operate on generic similarities of drug structure and protein sequence, therefore they can provide efficient predictions on new chemicals. More recently, approaches that incorporate prior knowledge about drugs and targets were proposed to enhance predictive accuracy on well-studied chemicals and targets. Such models may not be best suited to de novo drug discovery. However, they may provide valuable new insights in the context of drug repurposing and understanding the general mechanisms of drug action. The current state-of-the-art work in this context is arguably the DDR model [18], which uses a a multi-phase procedure to predict drug targets from relevant heterogeneous graphs. The gist of the approach is to combine various similarity indices and random walk features gained from the input graphs by means of non-linear fusion. Similarly, the NeoDTI model [19] predicts DTIs using supporting information about drugs and targets and a non-linear learning model over heterogeneous network data.

Despite continuous advances of similarity based approaches like DDR, these models depended on time-consuming training and prediction procedures as they need to compute the similarity features for each drug and target pair during both training and prediction. Also, the models still have a high false positive rate, especially when using large drug target interaction datasets like DrugBank_FDA [18].

Here, we propose a method utilising prior knowledge about drugs and targets, similarly to the DDR and NeoDTI model. Our method overcomes the afore-mentioned limitations by approaching the problem as link prediction in knowledge graphs. Knowledge graphs are a data representation model that represents relational information as a graph, where the graph nodes represent entities and edges represent relations between them. Facts are modelled as (subject, predicate, object) (SPO) triples, e.g. (*Aspirin, Drug-Target, COX-1*), where a subject entity (drug) is connected to an object entity (target protein) through a predicate relation (Drug-Target). In recent years, knowledge graphs have been successfully used for knowledge representation and discovery in many different domains, including life sciences [20, 21, 22].

Our work utilises the fact that the current drug target knowledge bases like Drug-Bank [23] and KEGG [24] are largely structured as networks representing information about drugs in relationship with target proteins (or their genes), action pathways, and targeted diseases. Such data can naturally be interpreted as a knowledge graph. The task of finding new associations between drugs and their targets can then be formulated as a link prediction problem based on knowledge graph embeddings [25].

We have proposed a new knowledge graph embedding based approach, TriModel , for predicting drug target interactions in a multi-phase procedure. We first used the currently available knowledge bases to generate a knowledge graph of biological entities related to both drugs and targets. We then trained our model to learn efficient vector representations (i.e. embeddings) of drugs and target in the knowledge graph. These representations were then used to score possible drug target pairs using a scalable procedure that has a linear time and space complexity. We compared our method to other state-of-the-art models using experimental evaluation on standard benchmarks. Our results show that the TriModel model outperforms all other approaches in areas under ROC and precision recall curve, metrics that are well suited to assessing general predictive power of ranking models [26].

2 Materials

In this section we discuss the datasets that we used to train and evaluate our model. We present the standard benchmarking datasets: Yamanishi_08 [10] and DrugBank_FDA [27], and we present statistics for elements in both datasets. We also discuss some flaws in the Yamanishi_08 dataset, and we present a new KEGG based drug targets dataset that addresses these flaws.

2.1 Standard benchmarks

The Yamanishi_08 [10] and DrugBank_FDA [27] datasets represent the most frequently used gold standard datasets in the previous state-of-the-art models for predicting drug targets [18]. The DrugBank_FDA [27] dataset consists of a collection of DTIs of FDA approved drugs that are gathered from DrugBank Database [1]. The Yamanishi_08 dataset is a collection of known drug target interactions gathered from different sources like KEGG BRITE [28], BRENDA [29], SuperTarget [30], and DrugBank [27]. It consists of four groups of drug target interactions corresponding to four different target protein classes: (1) enzymes (E), (2) ion-channels (IC) (3) G-protein-coupled receptors (GPCR) and (4) nuclear receptors (NR). The data in these groups vary in terms of size and positive to negative ratios as shown in table 1, ranging from 90 known DTIs with 1:15 as in the NR group to 2926 DTIs with 1:100 in the E group. These properties of the datasets affect the effectiveness of both training and evaluating models that use them. For example, the NR DTIs group have the largest positive to negative ratio among all the groups in the Yamanishi_08 dataset and therefore they are the easiest for predictive models in terms of evaluation. Contrary to that, the state-of-the-art models show the worst evaluation results on the NR group compared to other groups. This happens due to the low number of available DTIs training instances, which affects the models' generalisation on the training data.

2.2 New KEGG based benchmarking dataset

The Yamanishi_8 benchmarking dataset was published in 2008, and it contained drug target interactions from various sources including the KEGG BRITE, BRENDA, and SuperTarget databases [10]. In recent years, these sources have witnessed multiple developments (modifications, deletions, additions of many brand new records to their data [31, 32]). These modification have directly affected the Yamanishi_08 dataset, where a subset of the identifiers of both its drugs and targets has been modified through these developments. This affects the ability to link these drugs and targets to their corresponding properties e.g. associated pathways, diseases, or other biological entities in the recent versions of biological knowledge bases. These modifications have also included various newly discovered drug target interactions that are not included in the Yamanishi_08 dataset. For example, the KEGG database alone contains 12112 drug target interactions, while the total number of drug target interactions in the Yamanishi_08 dataset is only 5127.

---

[1] https://www.drugbank.ca

Table 1: Statistics of elements in the benchmarking datasets used in this work. The DTIs column represent the number of known drug target interactions, the Corruptions column represent the number of all possible combinations of drugs and targets that are not in the known drug target interactions which is used as negative in model training and evaluation, and the P2N column represents the ratio of positive to negative instances.

Dataset	Group	Drugs	Proteins	DTIs	Corruptions	P2N
	E	445	664	2926	$\approx$ 300K	1.00%
	IC	210	204	1476	$\approx$ 41K	3.57%
Yamanishi_08	GPCR	223	95	635	$\approx$ 21K	3.03%
	NR	54	26	90	1314	6.67%
	All	791	989	5127	$\approx$ 777K	0.66%
DrugBank_FDA	–	1482	1408	9881	$\approx$ 2.1M	0.48%
KEGG_MED	–	4284	945	12112	$\approx$ 4M	0.30%

To overcome these limitations, we propose a new drug target interaction benchmarking dataset that depends on recent versions of biological knowledge bases and includes a larger set of drug target interactions than the Yamanishi_08 dataset. We propose KEGG_MED, a dataset which is collected by extracting all the drug target interactions from the KEGG medicus database [2]. The KEGG_MED dataset contains 4284 drugs and 945 targets which are connected with 12112 drug target interactions. Table 1 shows a summary of statistics of the content on the dataset. Later in this paper, we report our results on this new suggested benchmark (in addition to the comparative validation on DrugBank_FDA) so that future approaches can be compared to our model.

2.3 Supporting knowledge graphs

Link prediction with knowledge graph embedding models require data to be modelled in a graph form, where the objective is to predict new links between graph entities. In the case of drug target discovery, we use supporting data from biomedical knowledge bases to generate informative graphs around drug target interactions. We generate a knowledge graph for each dataset to provide descriptive features for both drugs and targets. These knowledge graphs are extracted from different sources like KEGG [24], DrugBank [23], InterPro [33] and UniProt [34]. In our study we use a customised set of knowledge assertions about both drugs and targets. Appendix 1 and Table 1 in the supplementary material contain more information about the relation types present in each knowledge graph, and about their construction. For further information about the construction of such knowledge bases we refer to the work of [35] that provides a study of systematic integration of biological knowledge for learning drug-target interactions.

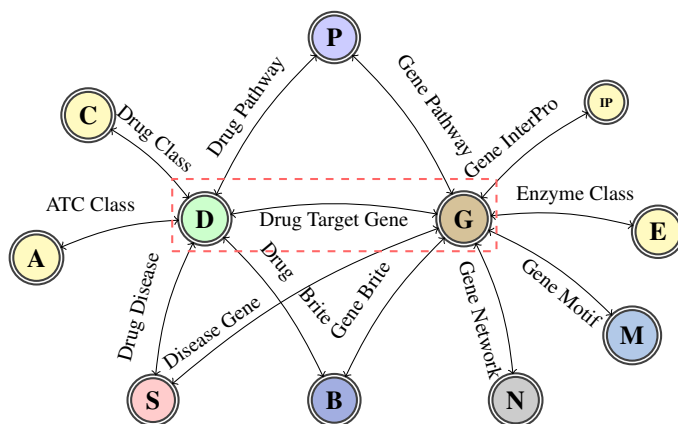[2]https://www.genome.jp/kegg/medicus.html

Figure 1: A graph schema for a knowledge graph about drugs, their target genes, pathways, diseases and gene networks extracted from KEGG and UniProt databases.

We generate a group-specific knowledge graph of information extracted from KEGG and UniProt for each DTI groups in the Yamanishi_8 dataset, while we use the DrugBank with UniProt knowledge bases to model information about DTIs of the DrugBank_FDA dataset. The information extracted in both cases is modelled as a graph of interconnected biological entities (schema shown in Fig. 1).

2.4 Methods

The knowledge graph embedding models we use follow a generative approach to learn low-rank embedding vectors for knowledge entities and relations. For learning the embeddings, multiple techniques can be used, such as tensor factorisation (c.f. the DistMult model [36]) or latent distance similarity (c.f. the TransE model [37]). The goal of all these techniques is to model possible interactions between graph embeddings and to provide scores for possible graph links. In the following, we provide details on the knowledge graph embedding procedure and the design of our proposed model, TriModel .

2.5 Knowledge graph embedding

Knowledge graph embedding (KGE) models learn a low rank vector representation of knowledge entities and relations that can be used to rank knowledge assertions according to their factuality. They are trained in a multi-phase procedure. First, a KGE model initialises all embedding vectors using random noise values. It then uses these embeddings to score the set of true and false training facts using a model-dependent scoring function. The output scores are then passed to the training loss function to compute training error. These errors are used by optimisers like AMSGrad [38] to generate gradients and update the initial embeddings, where the updated embeddings give higher scores for true facts and lower scores for false facts.

This procedure is performed iteratively for a set of iterations i.e. epochs in order to reach a state where the learnt embeddings provide best possible scoring for both true and false possible facts.

In the rest of this paper, we use $\mathbb{E}$ and $\mathbb{R}$ to denote the set all entities and relations in a knowledge graph respectively, where N_e and N_r represent the number of instances in $\mathbb{E}$ and $\mathbb{R}$ respectively. We also use Θ_E and Θ_R which denote the embeddings of entities and relations respectively, where $\Theta_E(i)$ is the embedding of entity i, $\Theta_R(j)$ is the embedding of relation j, and $f_m(s, r, o, \Theta)$ denotes the score of the fact that a subject entity s is connected to an object entity o with a relation r based on the embedding values Θ of the model m.

2.6 Embeddings representation

TriModel is a knowledge graph embedding model based on tensor factorisation that extends the DistMult [37] and ComplEx [39] models. It represents each entity and relation using three embedding vectors such that the embedding of entity i is $\Theta_E(i) = \{e_i^1, e_i^2, e_i^3\}$ where all embedding vectors have the same size K (a user-defined embeddings size). Similarly, the embedding of relation j is $\Theta_R(j) = \{w_j^1, w_j^2, w_j^3\}$. e^m and w^m denote the m part of the embeddings of the entity or the relation, and $m \in \{1, 2, 3\}$ represents the three embeddings parts. The embeddings in the Tri-Model model are initially with random values generated by the Glorot uniform random generator [40]. The embedding vectors are then updated during the training procedure to provide optimised scores for the knowledge graph facts.

2.7 Training procedure

The TriModel is a knowledge graph embedding model that follows the multi-phase procedure discussed in section 2.5 to effectively learn a vector representation for entities and relation of a knowledge graph. First, the model initialises its embeddings with random noise. It then updates them by iterative learning on the training data. In each training iteration i.e. epoch, the model splits the training data into mini-batches and executes its learning pipeline over each batch. The learning pipeline of the model learns the embeddings of entities and relations by minimising a negative softmax log-loss that maximises the scores of true facts and minimises the scores of unknown facts (assumed false during training). This loss is defined as follows:

$$
\begin{aligned}
\mathcal{L}_{spo}^{\text{TriModel}} = & -\phi_{spo} + \log(\sum_{o'} \exp(\phi_{spo'})) \\
& -\phi_{spo} + \log(\sum_{s'} \exp(\phi_{s'po})) \\
& +\frac{\lambda}{3} \sum_{k=1}^{K} \sum_{m=1}^{3} (|e_s^m|^3 + |w_p^m|^3 + |e_o^m|^3)
\end{aligned}
\tag{1}
$$

where x' represents an entity $e : e \neq x, e \in \mathbb{E}$, e_i^m is the embedding part m of the entity embedding $\Theta_E(i)$, w_i^m is the embedding part m of the relation embedding $\Theta_R(i)$, ϕ_{spo} denotes the score of the triple (s, p, o), m denotes the embedding part index, λ denotes a configurable regularisation weight parameter and $|x|$ is the absolute of x. The term $\frac{\lambda}{3} \sum_{k=1}^{K} \sum_{m=1}^{3} (|e_s^m|^3 + |w_p^m|^3 + |e_o^m|^3)$ is the nuclear 3-norm, which is a
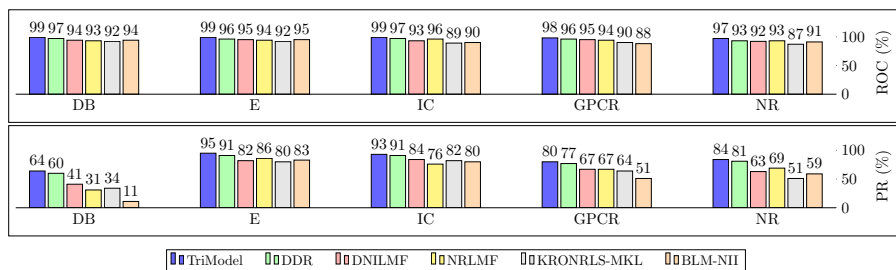
Figure 2: Bar chart for the values of the area under the roc curve (AUC-ROC) and area under the precision recall curve (AUC-PR) for the TriModel compared to other state-of-the-art models on standard benchmarking datasets. All values are rounded to two digits and multiplied by 100 to represent a percentage (%). DB represents the DrugBank_FDA dataset.

regularisation term [41] that enhances model generalisation over datasets with large entity vocabularies.

The scores of the TriModel model are computed using an embeddings interaction function (scoring function) that is defined as follows:

$$f_{\text{TriModel}}(s, r, o, \Theta) = \sum^K e_s^1 w_r^1 e_o^3 + e_s^2 w_r^2 e_o^2 + e_s^3 w_r^3 e_o^1. \tag{2}$$

It uses a set of three interactions: one symmetric interaction: $(e_s^2 w_p^2 e_o^2)$ and two asymmetric interactions: $(e_s^1 w_p^1 e_o^3)$ and $(e_s^3 w_p^3 e_o^1)$ for a convenient graphical explanation of the interaction, see Fig. 2 in the supplementary material. This approach models both symmetry and asymmetry in simple form similar to the DistMult [37] model where the DisMult model can be seen as a special case of the TriModel model if the first and third embeddings parts are equivalent ($e^1 = e^3$). We include more details about the training procedure in Appendix 2 in the supplementary material.

2.8 Results

In this section we describe the configuration of the data used in the experimentation, the evaluation protocol, the setup of our experiments and the results and findings of our experiments. We also compare the predictive accuracy of our model to selected existing approaches, including the state-of-the-art one.

2.9 Evaluation protocol

In order to facilitate comparison with the state-of-the-art models, we use a 10-fold cross validation (CV) to evaluate our model on the Yamanishi_08 and DrugBank_FDA datasets. First, we split the drug target interaction data into 10 splits i.e. folds. We then evaluate the model 10 times on each split, where the model is trained on the other 9 splits. This procedure is repeated 5 times and average results across these

runs are reported. This is to further minimise the impact of data variability on the result stability.

In each training configuration we use the known drug target interactions as positives, and all other possible combinations between the investigated dataset drugs and protein targets as negatives. This yields different positive to negative ratios since the datasets have different number of drugs, targets, and drug target interactions (see Table 1 for exact statistics of the ratios for each dataset).

We use the area under the ROC and precision recall curves (AUC-ROC and AUC-PR respectively) as an indication of the predictive accuracy of our model. We compute both metrics on the testing data (DTIs), where we divide the testing data into three groups: (1) S_p, containing testing drug target interactions where both the drug and the target are involved in known drug target interactions in the training data, (2) S_d, containing testing drug target interactions which contain drugs that have no known drug target interactions in the training data, (3) S_t, containing testing data of targets that has not involved in any known drug target interactions in the training data. The main reason for splitting the data this way was that one of the methods could not be compared with the others on the S_t, S_p data. The largest S_p group, however, generally exhibits least fluctuations across particular cross-validation runs, and therefore it is arguably most representative in terms of the comparative validation.

We also compute aggregated weighted AU-ROC, AU-PR scores for comparing the different models regardless the data group. These scores are defined as follows:

$$M = \sum_g \omega_g \cdot M_g, \tag{3}$$

where $g \in \{S_p, S_d, S_t\}$, M represents the aggregated score (AUC-ROC or AUC-PR), M_g is the specific score value for the group g, and ω_g is the weight of the particular data group computed by dividing the number of instances in g by the total number of instances in $S_p \cup S_d \cup S_t$.

2.10   Experimental setup

We use the supporting knowledge graph to perform a grid search to learn the model's best hyperparameters. In all of our experiments we initialise our model embeddings using the Glorot uniform random generator [40] and we optimise the training loss using the AMSGrad optimiser [38], where the the learning rate (lr) $\in \{0.01, 0.02, 0.03\}$, embeddings size (K) $\in \{50, 100, 150, 200\}$ and batch size (b) $\in \{128, 256, 512, 1024, 4000\}$. The rest of the grid search hyper parameters are defined as follows: the regularisation weight (λ) $\in \{0.1, 0.3, 0.35, 0.01, 0.03, 0.035\}$, dropout ($d$) $\in \{0.0, 0.1, 0.2, 0.01, 0.02\}$. The number of training epochs is fixed to 1000. The outcome best parameter for this grid search is included in table 2 in the supplementary materials.

We use Tensorflow framework (GPU) along with Python 3.5 to perform our experiments. All experiments were executed on a Linux machine with processor Intel(R) Core(TM) i70.4790K CPU @ 4.00GHz, 32 GB RAM, and an nVidia Titan Xp GPU. We include the training runtime of the TriModel model for each cross-validation iteration for all the investigated benchmarks in Fig. 1 in the supplementary materials.

Table 2: A comparison with state-of-the-art models on standard datasets using multiple configurations (S_p, S_d, S_t). The state-of-the-art results were obtained from [18]. The count (%) represents the percentage of the configuration instances, and the DB and KM columns represent DrugBank_FDA and KEGG_MED respectively. All the experimental configurations on all the datasets are evaluated using a 10-fold cross validation which is repeated 5 times. Column M. represents metrics and column Ft. represent model's feature type. The *structure* feature type represents protein and drug structure based features and *Ext.* denotes extensive prior knowledge features. Underlined scores represent the best scores in their feature category while the overall best results are bold and highlighted in green.

M.	Model	Ft.	E S_d	E S_t	E S_p	IC S_d	IC S_t	IC S_p	GPCR S_d	GPCR S_t	GPCR S_p	NR S_d	NR S_t	NR S_p	DB S_d	DB S_t	DB S_p	KM S_d	KM S_t	KM S_p
	Config. (Count %)		4%	5%	91%	4%	1%	95%	5%	4%	91%	10%	4%	86%	4%	11%	85%	5%	3%	92%
AUC-ROC	BLM-NII	Structure	0.73	0.89	0.96	0.83	0.89	0.91	0.85	0.87	0.88	0.88	0.85	0.91	0.71	0.75	0.90	-	-	-
	COSINE	Structure	0.80	-	-	0.82	-	-	0.88	-	-	0.89	-	-	0.77	-	-	-	-	-
	KRONRLS-MKL	Structure	0.71	0.88	0.93	0.77	0.86	0.90	0.81	0.84	0.91	0.79	0.76	0.87	0.79	0.81	0.88	-	-	-
	NRLMF	Structure	0.75	0.90	0.95	0.80	0.93	0.98	0.87	0.92	0.95	0.88	0.83	0.93	0.89	0.80	0.93	-	-	-
	DNILMF	Structure	0.81	0.92	0.96	0.81	0.92	0.94	0.86	0.92	0.96	0.83	0.83	0.92	0.90	0.82	0.95	-	-	-
	DDR	Ext.	0.84	0.92	0.97	**0.94**	0.97	0.98	0.91	**0.93**	0.96	**0.90**	**0.88**	0.92	0.91	0.86	0.96	-	-	-
	TriModel	Ext.	**0.95**	**0.96**	**0.99**	0.93	**0.98**	**0.99**	**0.92**	0.86	**0.99**	0.89	0.85	**0.99**	**0.94**	**0.94**	**0.99**	**0.81**	**0.58**	**0.99**
AUC-PR	BLM-NII	Structure	0.22	0.73	0.86	0.37	0.61	0.83	0.35	0.37	0.53	0.35	0.41	0.62	0.03	0.05	0.12	-	-	-
	COSINE	Structure	0.35	-	-	0.36	-	-	0.40	-	-	0.56	-	-	0.30	-	-	-	-	-
	KRONRLS-MKL	Structure	0.07	0.07	0.87	0.23	0.23	0.86	0.31	0.37	0.67	0.49	0.46	0.51	0.22	0.18	0.35	-	-	-
	NRLMF	Structure	0.28	0.76	0.89	0.30	0.61	0.79	0.36	0.55	0.69	0.49	0.45	0.72	0.28	0.23	0.32	-	-	-
	DNILMF	Structure	0.30	0.76	0.85	0.30	0.61	0.87	0.31	0.56	0.70	0.41	0.52	0.66	0.24	0.21	0.42	-	-	-
	DDR	Ext.	0.73	0.82	0.92	0.69	0.80	0.92	0.63	0.61	0.79	0.71	0.64	0.83	0.44	0.39	0.61	-	-	-
	TriModel	Ext.	**0.78**	**0.83**	**0.96**	**0.76**	**0.87**	**0.95**	**0.81**	**0.73**	**0.80**	**0.87**	**0.77**	**0.84**	**0.59**	**0.62**	**0.64**	**0.18**	**0.18**	**0.94**

2.11 Comparison with state-of-the-art models

We evaluate our model on the Yamanishi_08 and DrugBank_FDA datasets, and we compare our results to the following state-of-the-art models: DDR [18], NRLMF [17], NRLMF [15], KRONRLS-MKL [16], COSINE [42], and BLM-NII [11]. The comparison is made using the metrics of area-under-the-ROC (AUC-ROC) and precision-recall (AUC-PR) curves.

Fig. 2 presents overall results in terms of the AUC-ROC and AUC-PR scores for all compared models. The overall scores are combined across all testing configurations (S_p, S_d, S_t) for each dataset, where each specific score is computed as described in Eq. 3.

The results show that the TriModel model outperforms all other models in terms of AUC-ROC and AUC-PR on every benchmarking dataset. The TriModel model achieves a better AUC-PR score with a margin of 4%, 2%, 3%, 3%, 4% on E, IC, GPCR, NR, and DrugBank_FDA datasets respectively. It should be noted that we did not include the COSINE method in Fig. 2 as it is specifically designed to predict new drugs that do not have DTIs in the training phase. As such, the description of the method only reports accuracy on the new drug configuration (S_d), while the presented combined scores require values of all three evaluation configurations.

Table 2 shows a detailed comparison of the TriModel model and state-of-the-art models on all the standard benchmarking datasets for the the different evaluation settings S_p, S_d, and S_t. It also shows the relative number (in per cent) of drug-target statements available for each of the three validation settings.

The results in Table 2 show that the TriModel model outperforms other state-of-the-art models on 13 out of 15 different AUC-ROC experimentation configurations. In case of AU-PR, our model is better 14 out of 15 configurations. The results also show that the experimental configurations where our model is not the best represent a small portion of the total number of DTIs, while the TriModel model provides consistently better results for the largest S_p partition of the validation data.

Table 2 also show the results of the TriModel model on our proposed KEGG_MEDD dataset, where the model's AUC-PR scores are 0.18, 0.18, and 0.94 and its AUC-ROC scores are 0.81, 0.58, and 0.99 on the configurations S_d, S_t, and S_p respectively. No comparison with existing tools has been performed as their published versions cannot be directly applied to this data set.

2.12 Limitations

Despite the very promising results achieved by the prior knowledge-based models like DDR and TriModel , their predictive capabilities are best suited to finding new associations between well-studied drugs and targets (useful for instance in the drug repurposing context). If one needs predictions for de novo drug discovery, the models that utilise drug structure and target sequence similarities (e.g. BLM-NII, COSINE, KRONRLS-MKL, NRLMF or NRLMF) will likely deliver better results.

2.13 Web application for exploring the TriModel predictions

To let users explore our results, we have designed a web application [3]. The application allows for searching the predictions of the TriModel model. One can look for predictions using either drugs or targets as queries. Queries concerning multiple entities are possible simply by appending new terms to the search query. The results are presented as a table of the TriModel model scores of all the possible drug-target associations of the searched term.

The predictions provided by the web application are learnt by training the TriModel model on all the Yamanishi_08 dataset. The prediction scores are then computed for all possible drug-target combinations induced by the dataset. The scores of known drug interactions in the Yamanishi_08 dataset are set to 1, while the scores of all other drug target interactions are the normalised outcome of the TriModel predictions. The table of predictions in the application indicates the origin of each score, where a unique label "*Experimental Evidence*" is given to known DTIs and another label "*Model Prediction*" is assigned to the predicted scores.

2.14 Discussion

In the following we discuss possible reasons for the improved performance of our approach when compared to existing methods. We also review the limitations of the current DTI prediction benchmarks and discuss impact of data stratification on the predictive power of the models. Last but not least, we present tentative results in expert-based validation of predictions of our model that are not covered by the benchmark datasets. These results show high promise in terms of actual new discoveries predicted by our model.

2.15 Distinctive features of the presented approach

The relative success of the TriModel model can be attributed to two distinctive features not present in the state-of-the-art models. Firstly, we model input for the training as knowledge graphs. This allows for encoding multiple types of associations within the same graph and thus utilising more complex patterns. Other models that use graph-based data are limited in this respect as they only employ networks with single relation type. Secondly, the TriModel model uses a generative approach to learn efficient representations for both drugs and their targets. This approach enables scalable predictions of large volumes of drug-target interactions as it uses linear training time [25] and constant prediction time, which is not the case of the existing works. Furthermore, the TriModel model is able to predict other biological associations within the training data (e.g. drug and target pathways) with no extra computational effort. This shows substantial promise for further development of this technique.

---

[3]Hosted at: http://drugtargets.insight-centre.org.

2.16 Impact of data stratification on the predictive power

The Yamanishi_08 dataset is divided into four groups of DTIs according to the functionality of the target proteins. The groups are enzymes (E), ion-channels (IC) G-protein-coupled receptors (GPCR), and nuclear receptors (NR). The objective of this categorisation is to distinguish between models specifically tailored to predicting targets associated with a particular drug class [10]. [18] confirmed that organising the drug target interactions into groups according to the target's biological functionality enhances the predictive accuracy of models trained on such stratified data.

Based on our observations, we suggest a different explanation. The differences in performance appear to correlate with the relative numbers of negative examples in the grouped and full dataset configuration. Table 1 shows that the full Yamanishi_08 dataset configuration has a 0.66% positive to negative ratio, while the groups E, IC, GCPR, and NR have 1%, 3.57%, 3.03%, and 6.67% respectively. These differences can explain the variability of model performance quite well, since predicting positive instances is generally harder with more negatives present in the data [43]. In addition, dividing the DTI information gives rise to groups like the GPCR and NR groups. These contain only a small number of true DTIs (635 and 90 DTIs respectively), which further hampers the ability of models to generalise well (as we show in Section 2).

2.17 Validating the discovery potential of TriModel

Good performance of a model in benchmark tests is no doubt important. For various reasons like overfitting or training data imbalances, however, good benchmark results may not necessarily mean that the model can effectively support new discoveries.

Laboratory validation can ultimately confirm the model predictions as actual discoveries, but this is costly and time-consuming to be done at large scale. One can, however, perform alternative validations of the predictions using data that was not used for training the model. Such complementary validation can provide stronger foundations for claiming a model has high generalisation power.

We have performed a complementary validation of the TriModel 's predictions by manual analysis of top-10 drug-target associations per each of the examined benchmarking datasets. To decide whether or not the associations are true positives, we reviewed available literature. We only validated the predictions that were not part of the training data. The validation outcome shows that the TriModel model achieves 7 out of 10, 7 out of 10, 8 out of 10, 7 out of 10 and 6 out of 10 true predictions on the E, IC, GPCR, NR, DB datasets respectively. A detailed version of the validated predictions is included in Table 3 in the supplementary materials.

One can easily see that our model puts actual drug-target introductions (some of which were only recently discovered) high up in the result list. This is very promising for further development of the model and its deployment in clinical application scenarios.

2.18 Conclusions and future work

In this work, we have approached the problem of predicting new drug targets as a link prediction task in biomedical knowledge graphs. We have presented the Tri-Model model, a knowledge graph embedding model that can efficiently predict new drug target interactions. We have generated knowledge graphs of biological entities related to drugs and targets using available biological knowledge bases like KEGG, UniProt, and DrugBank. We have then used these knowledge graphs to train the TriModel model to learn efficient vector representation for both drugs and targets. In experiments using a standard benchmark data, we have demonstrated that the TriModel model outperforms state-of-the-art models in terms of both the area under ROC and precision recall curves.

Our study has also led to several secondary findings and contributions. We have shown that dividing datasets of drug target interactions into groups based on target properties does not positively affect the predictive accuracy of computation models. It can result in groups with very few drug target interactions, which negatively affects the accuracy of learnt models. Last but not least, we have developed a new KEGG based drug target interactions dataset that tackles the issues in the Yamanishi_08 dataset, and provides a richer set of up-to-date drug target interactions.

In future, we intend to explore how incorporation of more context data relevant to the target prediction problem can further improve the accuracy of our model. Last but not least, we will validate selected predictions of our model in laboratory experiments to demonstrate the clinical relevance of our results.

3 Funding

References

[1] Jürgen Drews. Drug discovery: A historical perspective. *Science*, 287(5460):1960–1964, 2000.

[2] Georg C Terstappen, Christina Schlüpen, Roberto Raggiaschi, and Giovanni Gaviraghi. Target deconvolution strategies in drug discovery. *Nature Reviews Drug Discovery*, 6(11):891, 2007.

[3] Walter Sneader. *Drug discovery: a history.* John Wiley & Sons, 2005.

[4] Lei Xie, Li Xie, Sarah L Kinnings, and Philip E Bourne. Novel computational approaches to polypharmacology as a means to define responses to individual drugs. *Annual review of pharmacology and toxicology*, 52:361–379, 2012.

[5] Joanne Bowes, Andrew J Brown, Jacques Hamon, Wolfgang Jarolimek, Arun Sridhar, Gareth Waldron, and Steven Whitebread. Reducing safety-related drug attrition: the use of in vitro pharmacological profiling. *Nature reviews Drug discovery*, 11(12):909, 2012.

[6] Anne Corbett, James Pickett, Alistair Burns, Jonathan Corcoran, Stephen B Dunnett, Paul Edison, Jim J Hagan, Clive Holmes, Emma Jones, Cornelius Katona, et al. Drug repositioning for alzheimer's disease. *Nature Reviews Drug Discovery*, 11(11):833, 2012.

[7] Peter M Rothwell, Michelle Wilson, Carl-Eric Elwin, Bo Norrving, Ale Algra, Charles P Warlow, and Tom W Meade. Long-term effect of aspirin on colorectal cancer incidence and mortality: 20-year follow-up of five randomised trials. *The Lancet*, 376(9754):1741–1750, 2010.

[8] Linxin Li, Olivia C Geraghty, Ziyah Mehta, Peter M Rothwell, and Oxford Vascular Study. Age-specific risks, severity, time course, and outcome of bleeding on long-term antiplatelet treatment after vascular events: a population-based cohort study. *The Lancet*, 390(10093):490–499, 2017.

[9] Lekha Sleno and Andrew Emili. Proteomic methods for drug target discovery. *Current opinion in chemical biology*, 12(1):46–54, 2008.

[10] Yoshihiro Yamanishi, Michihiro Araki, Alex Gutteridge, Wataru Honda, and Minoru Kanehisa. Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, 24(13):i232–i240, 2008.

[11] Jian-Ping Mei, Chee-Keong Kwoh, Peng Yang, Xiao-Li Li, and Jie Zheng. Drug–target interaction prediction by learning from local information and neighbors. *Bioinformatics*, 29(2):238–245, 2012.

[12] Feixiong Cheng, Yadi Zhou, Weihua Li, Guixia Liu, and Yun Tang. Prediction of chemical-protein interactions network with weighted network-based inference method. In *PloS one*, 2012.

[13] Feixiong Cheng, Chuang Liu, Jing Jiang, Weiqiang Lu, Weihua Li, Guixia Liu, Wei-Xing Zhou, Jin Huang, and Yun Tang. Prediction of drug-target interactions and drug repositioning via network-based inference. In *PLoS Computational Biology*, 2012.

[14] Ayeshah A Rosdah, Jessica K. Holien, Lea MD Delbridge, Gregory J Dusting, and Shiang Y Lim. Mitochondrial fission–a drug target for cytoprotection or cytodestruction? *Pharmacology research & perspectives*, 4(3):e00235, 2016.

[15] Hui Liu, Jianjiang Sun, Jihong Guan, Jie Zheng, and Shuigeng Zhou. Improving compound–protein interaction prediction by building up highly credible negative samples. *Bioinformatics*, 31(12):i221–i229, 2015.

[16] André CA Nascimento, Ricardo BC Prudêncio, and Ivan G Costa. A multiple kernel learning algorithm for drug-target interaction prediction. *BMC bioinformatics*, 17(1):46, 2016.

[17] Ming Hao, Stephen H Bryant, and Yanli Wang. Predicting drug-target interactions by dual-network integrated logistic matrix factorization. *Scientific reports*, 7:40376, 2017.

[18] Rawan S Olayan, Haitham Ashoor, and Vladimir B Bajic. Ddr: efficient computational method to predict drug–target interactions using graph mining and machine learning approaches. *Bioinformatics*, 34(7):1164–1173, 2017.

[19] Fangping Wan, Lixiang Hong, An Xiao, Tao Jiang, and Jianyang Zeng. Neodti: neural integration of neighbor information from a heterogeneous network for discovering new drug-target interactions. *Bioinformatics*, 35 1:104–111, 2019.

[20] Jens Lehmann and Robert Isele et. al. DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*, 6(2):167–195, 2014.

[21] Michel Dumontier, Alison Callahan, Jose Cruz-Toledo, Peter Ansell, Vincent Emonet, François Belleau, and Arnaud Droit. Bio2rdf release 3: A larger, more connected network of linked data for the life sciences. In *Proceedings of the ISWC 2014*, pages 401–404, 2014.

[22] Emir Muñoz, Vít Nováček, and Pierre-Yves Vandenbussche. Facilitating prediction of adverse drug reactions by using knowledge graphs and multi-label learning models. *Briefings in Bioinformatics*, 20(1):190–202, 8 2017.

[23] David S. Wishart, Craig Knox, An Chi Guo, Savita Shrivastava, Murtaza Hassanali, Paul Stothard, Zhan Chang, and Jennifer Woolsey. Drugbank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research*, 34:D668–D672, 2006.

[24] Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. Kegg: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45(D1):D353–D361, 2017.

[25] Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1):11–33, 2016.

[26] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240. ACM, 2006.

[27] David S. Wishart, Craig Knox, An Chi Guo, Dean Cheng, Savita Shrivastava, Dan Tzur, Bijaya Gautam, and Murtaza Hassanali. Drugbank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Research*, 36:D901–D906, 2008.

[28] Minoru Kanehisa and Susumu et. al. Goto. From genomics to chemical genomics: new developments in kegg. *Nucleic acids research*, 34(suppl_1):D354–D357, 2006.

[29] Ida Schomburg, Antje Chang, Christian Ebeling, Marion Gremse, Christian Heldt, Gregor Huhn, and Dietmar Schomburg. Brenda, the enzyme database: updates and major new developments. *Nucleic acids research*, 32(suppl_1):D431–D433, 2004.

[30] Stefan Günther, Michael Kuhn, Mathias Dunkel, Monica Campillos, Christian Senger, Evangelia Petsalaki, Jessica Ahmed, Eduardo Garcia Urdiales, Andreas Gewiess, Lars Juhl Jensen, et al. Supertarget and matador: resources for exploring drug-target relationships. *Nucleic acids research*, 36(suppl_1):D919–D922, 2007.

[31] Sandra Placzek, Ida Schomburg, Antje Chang, Lisa Jeske, Marcus Ulbrich, Jana Tillack, and Dietmar Schomburg. Brenda in 2017: new perspectives and new tools in brenda. In *Nucleic Acids Research*, 2017.

[32] Nikolai Hecker and Jessica Ahmed et. al. Supertarget goes quantitative: update on drug–target interactions. In *Nucleic Acids Research*, 2012.

[33] Alex L Mitchell and Teresa K et. al. Attwood. Interpro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Research*, 47(D1):D351–D360, 2019.

[34] The UniProt Consortium. Uniprot: the universal protein knowledgebase. In *Nucleic Acids Research*, 2017.

[35] Daniel Scott Himmelstein, Antoine Lizée, Christine S Hessler, Leo Brueggeman, Sabrina L Chen, Dexter Hadley, Ari J Green, Pouya Khankhanian, and Sergio E Baranzini. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. In *eLife*, 2017.

[36] Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *NIPS*, pages 2787–2795, 2013.

[37] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. In *ICLR*, volume abs/1412.6575, 2015.

[38] Sashank Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *ICLR*. OpenReview.net, 2018.

[39] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In *ICML*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 2071–2080. JMLR.org, 2016.

[40] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, volume 9 of *JMLR Proceedings*, pages 249–256. JMLR.org, 2010.

[41] Timothée Lacroix, Nicolas Usunier, and Guillaume Obozinski. Canonical tensor decomposition for knowledge base completion. In *ICML*, volume 80 of *JMLR Workshop and Conference Proceedings*, pages 2869–2878. JMLR.org, 2018.

[42] Hansaim Lim, Paul Gray, Lei Xie, and Aleksandar Poleksic. Improved genome-scale multi-target virtual screening via a novel collaborative filtering approach to cold-start problem. *Scientific reports*, 6:38860, 2016.

[43] Tie-Yan Liu et al. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, 3(3):225–331, 2009.