Discovering Protein Drug Targets Using Knowledge Graph Embeddings

SUPPLEMENTARY MATERIAL

Sameh K. Mohamed [1 2 3] **Vít Nováček** [1 2 3] **Aayah Nounu** [4]

1. Supporting Knowledge Graphs

Knowledge graph embeddings are known to provide state-of-the-art results in link prediction on knowledge graphs (NICKEL ET AL.; LACROIX ET AL. 2016a; 2018). They operate by learning low rank representations of knowledge graphs' entities and relation, then they use these representation to assess the factuality of relation associations between entities. They encode knowledge of different types of association for each entity and relation in its embeddings, then they can efficiently learn their unknown link.

In the task of drug target prediction, a knowledge graph embedding model can be efficiently trained to predict drug target associations between drug and target entities in a a biological knowledge graph by learning efficient vector representations of both drugs and target in a knowledge graph context. Therefore, we train our knowledge graph embedding model, the TriModel model, on a knowledge graph that contains training drug target interactions along with other associations of drugs and targets as show in Table 1. For example, when training the TriModel model on the Yamanishi_08 and our KEGG_MED dataset (YAMANISHI ET AL. 2008), we include drug and target information from the KEGG (KANEHISA ET AL. 2017) and UniProt (CONSORTIUM 2017) knowledge bases. Information we include are for instance ATC codes of drugs, BRITE identifiers, classes, associated diseases, groups and associated pathways from the KEGG knowledge base. Similarly, we use the target's active sites, binding sites, conserved sites, domains, associated pathways, family, and protein-protein interactions and different gene ontology annotations from the UniProt (CONSORTIUM 2017) and InterPro (MITCHELL ET AL. 2019) databases.

For example, *Aspirin* has multiple ATC codes such as *B01AC06* and *C07FX04*. These codes are transformed into assertions in the knowledge graph, where each code is transformed into five assertions as follows:

1- (Aspirin, ATC-C1, B) 2- (Aspirin, ATC-C2, B01) 3- (Aspirin, ATC-C3, B01A) 4- (Aspirin, ATC-C4, B01AC 5- (Aspirin, ATC-C5, B01AC6	Aspirin ATC B01AC6

1- (Aspirin, ATC-C1, C) 2- (Aspirin, ATC-C2, C07) 3- (Aspirin, ATC-C3, C07F) 4- (Aspirin, ATC-C4, C07FX) 5- (Aspirin, ATC-C5, C07FX04)	Aspirin ATC C07FX04

Similarly, the enzyme classes of protein enzymes are also transformed into four assertions. For example the LATS1 protein has enzyme class number *EC:2.7.11.1*, this is transformed into four assertions as follows:

1- (LATS1, EC-C1, EC:2._._._) 2- (LATS1, EC-C2, EC:2.7._._) 3- (LATS1, EC-C3, EC:2.7.11._) 4- (LATS1, EC-C4, EC:2.7.11.1)	Aspirin EC EC:2.7.11.1

These generated assertions for both ATC and EC codes allow for grouping drugs and proteins with similar class levels by connecting them to the same nodes in the knowledge graph.

In the case of the DrugBank_FDA benchmark dataset (WISHART ET AL. 2008), the TriModel model uses information about drugs and targets from both the UniProt (CONSORTIUM 2017) and DrugBank (WISHART ET AL.; WISHART ET AL. 2006; 2008) knowledge bases. We use drugs' ATC codes, categories, associated pathways and the categories of these pathways. We also use the same protein information as for the Yamanishi_08 and KEGG_MED datasets. Table 1 contains a summary and statistics of the relation instances of each knowledge base.

[1]Data Science Institute [2]National University of Ireland Galway [3]Insight Centre for Data Analytics [4]MRC Integrative Epidemiology Unit, University of Bristol. Correspondence to: Sameh K. Mohamed <sameh.kamal@insight-centre.org>.

KEGG		**UniProt**		**DrugBank**	
Relation	Count	Relation	Count	Relation	Count
Drug ATC	902	Protein active site	281	Drug ATC	2541
Drug BRITE	3354	Protein binding site	153	Drug Category	21041
Drug Class	703	Protein conserved site	334	Drug Pathway	2013
Drug Disease	1696	Protein domain	2204	Pathway Category	4670
Drug Group	2714	Protein Pathway	3189		
Drug Pathway	1340	Protein Family	1584		
Protein BRITE	2572	PPI	7276		
Protein Disease	709	Protein GO: M.F.	6958		
Protein EC No.	753	Protein GO: C.C.	7588		
Protein Motif	4374	Protein GO: B.P	17639		
Protein Pathway	5551	Protein EC No.	840		

Table 1. Summary and statistics of a sample of relations extracted from supporting knowledge bases M.F, C.C and B.F refer to molecular function, cellular component and biological process respectively.

2. Model Training and Evaluation Pipeline

The TriModel model uses a standard knowledge graph embeddings training procedure that consumes a set of graph assertions and initial embeddings (NICKEL ET AL. 2016b). The outcome of the training procedure is the update version of the embeddings, where these embeddings are used to provide scores for any given set of assertions. Fig. 2 shows a diagram of the training evaluation procedure of the TriModel model on the drug target interactions data. First, the model combines both the training drug target interactions and their corresponding supporting knowledge graph assertions to model the input knowledge graph *i.e.* triplets. These assertions are fed to the model along with an initial values for the embeddings for all the entities and relations in the graph. The learning process (training) involves updating the initial embeddings such that the updates embeddings values provide efficient scoring for all the input assertions.

3. Partial Knowledge Assessment

In this section, we explore the importance of different parts of the supporting knowledge graphs and their effect of the outcome accuracy on the TriModel model. We have evaluated the TriModel model on the Yamanishi_08 dataset while we remove parts of the supporting knowledge graphs for each group of drug target interactions. Our evaluation is performed in six configurations: (1) *none*: where the model do not use any supporting knowledge graph information during the training, (2) *C1*: the ATC codes of drugs are excluded (3) *C2*: the protein sequence related assertions such as motifs, domain, etc. (4) *C3*: the pathway related assertions for both drugs and proteins are excluded (5) *C4*: the disease related assertions are excluded. (6) *full*: the full supporting knowledge graph is used.

Table 2 shows a summary of the outcome results of the evaluation of the previously mentioned configurations. The results show that excluding parts of the knowledge graph have an effect on the outcome accuracy of the model. The results show that the best predictive accuracy is achieved on two configurations: *full* and *C1*. The *full* configuration denotes the use of the full knowledge graph as a supporting evidence, while the *C1* configuration represents the knowledge graph without the information related to the ATC codes. On the other hand, the worst results are achieved with the *none* configuration, where no supporting knowledge graph is used.

4. Experiments Setup Configurations

We learn the best parameters for the TriModel model using the grid search described in the experimental setup section. Table 3 shows the best parameters found using the grid search for each parameter, where the grid search is performed only on the supporting knowledge graphs of each of the investigated datasets with no drug target interactions. The grid search is then assessed as a classical link prediction task on knowledge graphs. We report the runtime of the TriModel model per each cross-validation iteration for each of the investigated benchmarking datasets in Fig. 1.

5. Novel Drug Target Interactions

We generated all possible unknown drug target combination for each of the investigated datasets, and we have used the TriModel model to predict scores for these combination. We have then ranked these combinations according to the

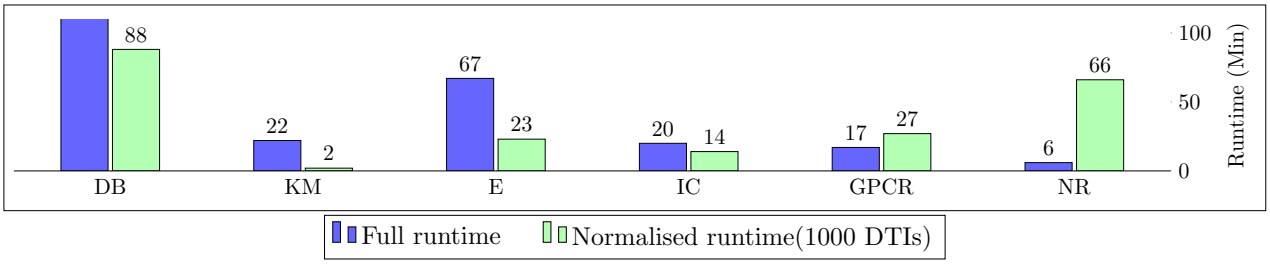

Figure 1. The runtime in minutes of the TriModel model for one cross-validation iteration (10-folds) on each of the investigated benchmarking datasets. The normalised runtime (t_n^d) of the dataset d is computed such that $t_n^d = t^d / N_d * 1000$, where n^d is the full runtime on the dataset d and N_d is the number of DTIs instances in the d dataset. DB full runtime = 867 minutes.

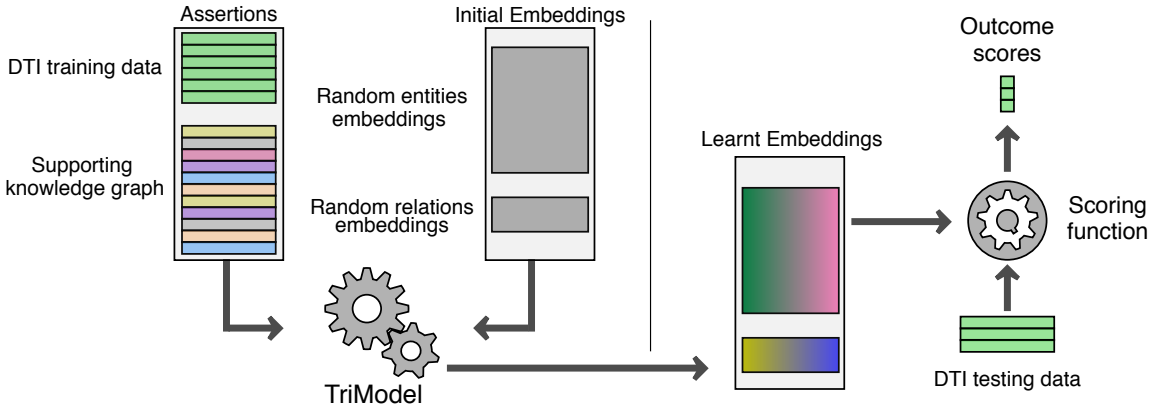

Figure 2. A diagram of the training pipeline of the TriModel model. Both drug target interactions and supporting knowledge graph assertions are combined and used as input to the model along with initial random embeddings for both entities and relations. The outcome of the training procedure is learnt embeddings which is used to score any drug target interaction data of drugs and proteins processed during the training processes.

Model	Config.	Metric	E			IC			GPCR			NR		
			Sd	St	Sp	Sd	St	Sp	Sd	St	Sp	Sd	St	Sp
TriModel	None	AUC-ROC	0.83	0.89	**0.99**	0.86	0.92	**0.99**	0.71	0.68	0.97	0.68	0.72	0.94
	C1		0.94	**0.96**	**0.99**	**0.93**	**0.98**	**0.99**	**0.92**	**0.89**	**0.99**	0.86	**0.85**	**0.99**
	C2		0.94	0.95	**0.99**	**0.93**	**0.98**	**0.99**	**0.92**	0.85	**0.99**	0.87	0.83	**0.99**
	C3		0.92	0.94	**0.99**	**0.93**	0.96	**0.99**	0.91	0.80	0.98	0.83	0.70	0.96
	C4		0.93	**0.96**	**0.99**	**0.93**	**0.98**	**0.99**	0.91	0.88	**0.99**	0.85	0.84	**0.99**
	Full		**0.95**	**0.96**	**0.99**	**0.93**	**0.98**	**0.99**	**0.92**	0.86	**0.99**	**0.89**	**0.85**	**0.99**
	None	AUC-PR	0.27	0.30	0.94	0.32	0.53	0.93	0.26	0.38	0.79	0.46	0.61	0.69
	C1		0.75	**0.83**	0.95	0.76	**0.89**	**0.95**	0.79	**0.75**	**0.81**	0.84	**0.78**	**0.84**
	C2		0.76	0.82	0.95	**0.79**	0.83	0.94	**0.81**	0.66	0.80	**0.87**	0.70	0.74
	C3		0.73	0.79	0.95	0.74	0.84	**0.95**	0.78	0.60	0.79	0.81	0.64	0.72
	C4		0.73	**0.83**	0.95	0.77	0.87	**0.95**	0.78	0.72	0.80	0.83	0.69	0.79
	Full		**0.78**	**0.83**	**0.96**	0.76	0.87	**0.95**	**0.81**	0.73	0.80	**0.87**	0.77	**0.84**

Table 2. Summary of evaluation results for experimenting the TriModel model on different benchmarking datasets with different excluded subsets of the supporting knowledge graph. The experiments have six different configurations: (1) None: where no supporting knowledge graphs are used (2) C1: excluded the ATC codes for drugs (3) C2: excluded protein sequence related assertions (4) C3: excluded pathway related assertions for both drugs and proteins (5) C4: excluded disease related assertions. (6) Full: using the full supporting knowledge graph.

Dataset	Best parameters				
	Batch size	Lambda	Dropout	Embedding size	Learning rate
Enzymes (E)	4000	0.03	0.2	150	0.01
Ion Channels (IC)	512	0.03	0.2	150	0.01
G-Protein Coupled Receptors (GPCR)	256	0.03	0.2	150	0.01
Nuclear Receptors (NR)	128	0.03	0.2	150	0.01
DrugBank_FDA (DB)	4000	0.3	0.2	200	0.01
KEGG_MEDICUS (KM)	4000	0.3	0.2	200	0.01

Table 3. The best parameters of the TriModel model on each of the investigated benchmarking datasets.

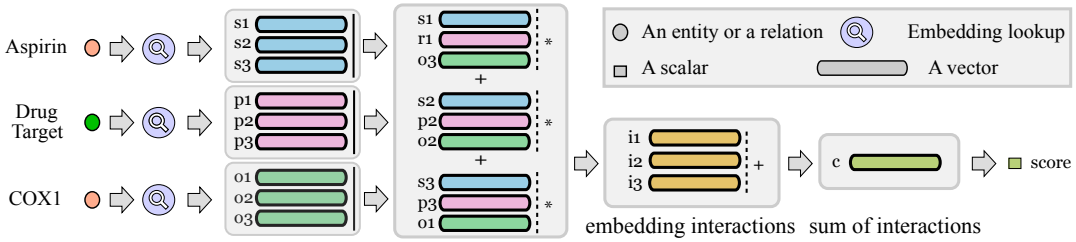

Figure 3. Flow diagram of the scoring function of TriModel. The subject, the relation, and the object are represented using three embedding vectors of size k. The score of a triple (s, p, o) is defined as $f(s, p, o) = \sum_k c_k$, where $c = i1 + i2 + i3$, $i1 = s1 \cdot r1 \cdot o3$, $i2 = s2 \cdot r2 \cdot o2$, and $i3 = s3 \cdot r3 \cdot o1$.

predicted scores and used the top 10 scored combinations of each datasets to asses the model's capabilities for predicting unknown drug target interactions. A domain expert annotated each of the top 10 combinations with three labels *YES*, *NO* and *UNKNOWN* which represent known valid drug target interactions, proven invalid drug target interactions and unknown interactions, respectively. Table 4 presents the top 10 predictions for each dataset along with the expert annotation and literature evidence (via PubMed ID).

Dataset	#	Drug Name	Drug Id	Target Name	Target ID	Score	Valid	Evidence
E	1	Halothane	D00542	CYP2E1	hsa:1571	8.820	YES	PubMed:19442086
	2	Aminocaproic acid	D00160	PROC	hsa:5624	8.601	Unknown	-
	3	Imatinib mesylate	D01441	MAPK1	hsa:5594	8.355	YES	PubMed: 22089930
	4	Methoxsalen	D00139	CYP1A1	hsa:1543	8.323	YES	PubMed: 7702611
	5	Isoflurophate	D00043	ELANE	hsa:1991	8.311	Unknown	-
	6	Imatinib mesylate	D01441	MAPK3	hsa:5595	8.295	YES	PubMed: 15100154
	7	Metyrapone	D00410	CYP1A1	hsa:1543	8.275	YES	PubMed: 9512490
	8	Salicylic acid	D00097	PTGS2	hsa:5743	8.184	No	-
	9	Nifedipine	D00437	CYP2C9	hsa:1559	8.140	YES	PubMed: 9929518
	10	Aminoglutethimide	D00574	CYP21A2	hsa:1589	8.132	YES	PubMed: 8201961
IC	1	Nicotine	D03365	CHRNA4	hsa:1137	6.486	YES	PubMed:17590520
	2	Zonisamide	D00538	SCN5A	hsa:6331	6.468	YES	PubMed:20025128
	3	Benzocaine	D00552	SCN5A	hsa:6331	6.380	YES	PubMed:19661462
	4	Nimodipine	D00438	CACNA1S	hsa:779	6.297	YES	PubMed:16675661
	5	Metoclopramide	D00726	CHRNA5	hsa:1138	6.285	Unknown	-
	6	Isoflurane	D00545	GLRA2	hsa:2742	6.262	Unknown	-
	7	Diazoxide	D00294	ABCC9	hsa:10060	6.198	YES	PubMed: 21428460
	8	Prilocaine	D00553	SCN10A	hsa:6336	5.992	YES	PubMed:17139284
	9	Verapamil hydrochloride	D00619	CACNA1F	hsa:778	5.961	YES	PubMed:19125880
	10	Nimodipine	D00438	CACNA2D1	hsa:781	5.940	Unknown	PubMed: 29176626
GPCR	1	Isoetharine	D04625	ADRB2	hsa:154	7.148	YES	PubMed:21948594
	2	Octreotide acetate	D02250	SSTR1	hsa:6751	6.752	YES	PubMed:16438887
	3	Clonidine hydrochloride	D00604	ADRA1B	hsa:147	6.650	YES	PubMed: 17584443
	4	Metoprolol	D02358	ADRB2	hsa:154	6.499	YES	PubMed:19637941
	5	Epinephrine	D00095	ADRA1D	hsa:146	6.489	YES	PubMed:20954794
	6	Theophylline	D00371	ADORA2A	hsa:135	6.407	YES	PubMed:16357952
	7	Denopamine	D02614	ADRB2	hsa:154	6.388	NO	PubMed: 22505670
	8	Risperidone	D00426	DRD2	hsa:1813	6.386	YES	PubMed:17059881]
	9	Bosentan	D01227	AGTR1	hsa:185	6.347	Unknown	-
	10	Epinephrine	D00095	ADRA1B	hsa:147	6.306	YES	PubMed:20954794
NR	1	Medroxyprogesterone acetate	D00951	ESR1	hsa:2099	6.314	YES	PubMed:17094978
	2	Mometasone furoate	D00690	NR3C1	hsa:2908	6.066	YES	PubMed:8439518
	3	Ethinyl estradiol	D00554	ESR2	hsa:2100	6.038	NO	PubMed: 15878629
	4	Dydrogesterone	D01217	ESR1	hsa:2099	5.968	YES	PubMed: 22878119
	5	Norethindrone	D00182	ESR1	hsa:2099	5.893	YES	PubMed: 27245768
	6	Etretinate	D00316	RORB	hsa:6096	5.848	Unknown	-
	7	Mifepristone	D00585	ESR1	hsa:2099	5.841	YES	PubMed: 15001543
	8	Tretinoin	D00094	RORA	hsa:6095	5.679	YES	CheMBL
	9	Tazarotene	D01132	RORC	hsa:6097	5.463	Unknown	-
	10	Testosterone	D00075	ESR1	hsa:2099	5.453	YES	PubMed:12676605
DB	1	Methysergide	DB00247	HTR1D	P28221	6.421	YES	PubMed: 7984267
	2	Phenoxymethylpenicillin	DB00417	SLC15A2	Q16348	6.295	Unknown	-
	3	L-Valine	DB00161	BCAT2	O15382	6.263	YES	PubMed: 6933702
	4	Corticorelin ovine triflutate	DB09067	GHRHR	Q02643	6.237	Unknown	-
	5	Acarbose	DB00284	GANC	Q8TET4	6.232	YES	KEGG
	6	Halothane	DB01159	GRIA1	P42261	6.226	YES	PubMed: 14739810
	7	Hydroxocobalamin	DB00200	GIF	P27352	6.226	Unknown	-
	8	Quazepam	DB01589	GABRB2	P47870	6.215	YES	PubMed:6738302
	9	Nintedanib	DB09079	KIT	P10721	6.202	Unknown	-
	10	Miglitol	DB00491	SI	P14410	6.201	YES	CheMBL

Table 4. Validation of the top 10 scored combination for each of the investigated datasets. The DrugBank, CheMBL and KEGG DTIs are used as evidence the different interactions, where the PubMed ID is listed when possible.

References

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

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

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

Mitchell, Alex L, Attwood, Teresa K, Babbitt, Patricia C, Blum, Matthias, Bork, Peer, Bridge, Alan, Brown, Shoshana D, Chang, Hsin-Yu, El-Gebali, Sara, Fraser, Matthew I, Gough, Julian, Haft, David R, Huang, Hongzhan, Letunic, Ivica, Lopez, Rodrigo, Luciani, Aurlien, Madeira, Fabio, Marchler-Bauer, Aron, Mi, Huaiyu, Natale, Darren A, Necci, Marco, Nuka, Gift, and et. al. Interpro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Research*, 47(D1):D351–D360, 2019.

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

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

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

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

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