Supplementary Materials for DrugE-Rank

1 METHOD

1.1 Learning to rank in DrugE-Rank

Learning to rank (LTR) (Liu, 2009; LI, 2011) refers to a set of machine learning models for ranking objects. It has been widely used in document retrieval, web searching, collaborative filtering, expert finding, bioinformatics and many other applications. For example, with respect to users queries, LTR model has been employed to rank more relevant web pages higher. These methods can be divided into 3 types: (a) point-wise approaches, such as McRank (Li *et al.*, 2007) and Prank (Crammer *et al.*, 2001), where the ranking problem is approximately by a classical classification or ordinal regression problem; In this case, given a single query-document pair, the LTR model tries to predict its score. (b) pair-wise approaches, which transforms the ranking problem to a pairwise classification problem, such as LambdaMART (Burges, 2010) and RankingSVM (Herbrich *et al.*, 1999); Given a pairs of documents, the LTR model tries to tell which document is better with respect to a specific query. (c) list-wise approaches, which takes ranked lists as instances in both training and prediction, such as ListNet (Cao *et al.*, 2007) and ListMLE (Xia *et al.*, 2008). In this case, although the group structure of ranking is fully utilized in learning, the optimization of evaluation measure is difficult and approximations or bounds have to be used.

We develop DrugE-Rank to address the problem of predicting drug-target interactions for new drugs or new targets, which is especially challenging (Ding *et al.*, 2014). Identifying drug-target interactions can be considered as a multi-label classification problem, in the sense that drugs can be interacted (labeled) by multiple targets. The basic idea of DrugE-Rank is to solve the identification of drug-target interactions by learning to rank, where both feature-based and similarity-based methods are effectively integrated to improve the prediction performance. Taking the problem of predicting drug-target interactions for new targets as an example, we can view the targets as queries, drugs as documents and the problem then becomes to rank the most promising drugs with respect to the given targets. That is to say, to identify the interacting drugs of a given target is analogous to find relevant web pages with respect to a given query, which can be nicely solved by LTR.

LambdaMART (Burges, 2010) has been successfully applied to solve a number of real-world problems, such as Yahoo Learning to Rank competition (Chapelle and Chang, 2011) and BioASQ challenge (Liu *et al.*, 2015), and so we also utilize it for training LTR model in DrugE-Rank. In web searching, the power of LTR comes from the features used to characterize the relevance of web pages. Similarly, DrugE-Rank considers three types of important features, target features, drug features and pair features generated from component methods (see details in the Section 3.4 of the main text). These features are concatenated into a 189-dimension vector. Here we illustrate the workflow of DrugE-Rank using the prediction of interacting drugs of new targets. Given a new target t_{new}, each candidate drug then corresponds to a 189-dimension vector, and the feature vectors of candidate drugs are given to LambdaMart as input. Finally, a ranking list of drugs is returned as the prediction result. Note that when training the ranking model (function), we do not use all possible drug target pairs, and instead for each target, we use only top K drugs highly predicted to be interacting with the target by each component method. The drug-target interaction is highly sparse, in which false positives can occur easily. This manner of using only top K drugs is highly effective to avoid this type of false positives and eventually reduces the computational cost in training LTR. In fact, similar approaches are used to train LTR models in web searching since there are huge number of web pages.

We made use of RankLib[1] for the implementation of LambdaMART. To train the LambdaMART, the number of trees and the number of leaves for each tree were selected from $\{8,16,32,64,128\}$ and $\{4,8,16\}$, respectively.

1.2 Component methods

Many computational methods have been proposed to tackle the problem of drug target interaction prediction (Ding *et al.*, 2014). The success of DrugE-Rank relies on the effective and efficient integration of the component methods. Considering this, the component method should be developed by using different principles and achieve reasonable and diverse results. In addition, the computational cost should be low. Finally, we select six well-known, cutting-edge similarity-based methods as component methods: k-Nearest Neighbor (k-NN), Bipartite Local Model with support vector classification (BLM-svc) (Bleakley and Yamanishi, 2009), Bipartite Local Model with support vector regression (BLM-svr) (Bleakley and Yamanishi, 2009), Laplacian regularized least squares (LapRLS) (Xia *et al.*, 2010), Network-based Laplacian regularized least squares (NetLapRLS) (Xia *et al.*, 2010), Weighted Nearest Neighbor based Gaussian Interaction Profile classifier (WNN-GIP) (van Laarhoven *et al.*, 2011; van Laarhoven and Marchiori, 2013). In DrugE-Rank, these component methods are used to learn pair features for drug-target pairs. Each method generates a feature for DrugE-Rank.

[1] http://sourceforge.net/p/lemur/wiki/RankLib/

Table 1. AUC for 10 × 5-fold cross-validation over Data-1 (p-values of paired t-test against DrugE-Rank with all features)

Method	drug-based interactions	target-based interactions
k-NN	$0.6974\ (5.07 \times 10^{-43})$	$0.7659\ (3.77 \times 10^{-18})$
BLM-svc	$0.7620\ (6.29 \times 10^{-37})$	$0.8381\ (1.58 \times 10^{-15})$
BLM-svr	$0.8109\ (2.35 \times 10^{-32})$	$0.8504\ (7.39 \times 10^{-09})$
LapRLS	$0.7826\ (7.28 \times 10^{-36})$	$0.8209\ (9.04 \times 10^{-11})$
NetLapRLS	$0.7693\ (3.19 \times 10^{-31})$	$0.8011\ (3.51 \times 10^{-11})$
WNN-GIP	$0.7512\ (8.54 \times 10^{-33})$	$0.8149\ (1.92 \times 10^{-13})$
RF	$0.7185\ (2.57 \times 10^{-37})$	$0.7811\ (5.79 \times 10^{-17})$
GBDT	$0.7013\ (1.15 \times 10^{-37})$	$0.7972\ (9.38 \times 10^{-14})$
DrugE-Rank DFV and TFV)	$0.7725\ (9.35 \times 10^{-23})$	$0.8446\ (5.50 \times 10^{-08})$
DrugE-Rank (PFV only)	$0.8597\ (9.17 \times 10^{-04})$	$0.9021\ (1.37 \times 10^{-03})$
DrugE-Rank (all features)	**0.8775**	**0.9142**

Table 2. AUC for independent testing data (FDA approved, new drugs and Experimental drugs)

Methods	Data-2 new drugs	Data-3 new targets	Data-5 new drugs
k-NN	0.5534	0.6493	0.5469
BLM-svc	0.6073	0.6932	0.5635
BLM-svr	0.6395	0.7392	0.5820
LapRLS	0.6016	0.7128	0.5527
NetLapRLS	0.5958	0.6912	0.5619
WNN-GIP	0.6119	0.6709	0.5677
RF	0.5762	0.6021	0.5514
GBDT	0.6074	0.6370	0.5697
DrugE-Rank (DFV and TFV)	0.6814	0.6877	0.5731
DrugE-Rank (PFV only)	0.6880 (0.0254)	0.7704 (0.0289)	0.6157 (0.0074)
DrugE-Rank (all features)	**0.7011** (0.0317)	**0.7937** (0.0274)	**0.6319** (0.0188)

2 EXPERIMENT RESULTS

2.1 AUC for cross-validation and independent testing data prediction

Table 1 shows the AUC performance of different methods by 10 five-fold cross validation over Data-1. Table 2 presents the AUC results of these models over Data-2, Data-3 and Data-5. These results are consistent with those in AUPR, which again demonstrates the performance advantage of DrugE-Rank.

2.2 Comparison with Shi and KBMF2K

We compared the performance of DrugE-Rank with two sate-of-the-art methods, Shi and KBMF2K. The default parameters in their original paper were used to run these methods. The 10 five-fold cross validation results on Data-1 in terms of AUPR and AUC are shown in Table 3 and 4, respectively. The performance of random prediction is also given as a baseline. Paired t-test was used to evaluate the statistical significance of result. From these results, we can see that DrugE-Rank significantly outperformed both Shi and KBMF2K. For example, DrugE-Rank achieved an AUPR of 0.4917, which is followed by Shi (0.3833) and KBMF2K (0.3698). Additionally, DrugE-Rank achieved an AUPR of 0.8775, which was followed by Shi (0.8067) and KBMF2K (0.7644). Furthermore, we compared their performance by using the independent test data of Data-2, Data-3 and Data-5, which is presented in Table 5 (AUPR) and 6 (AUC). These results also demonstrated the advantageous performance of DrugE-Rank. For example, DrugE-Rank achieved the highest AUC of 0.7937 on Data-3, which was followed by Shi (0.7104) and KBMF2K (0.6926).

Table 3. AUPR for 10 × 5-fold cross-validation over Data-1 comparing with state-of-the-art methods (p-values of paired t-test against DrugE-Rank with all features)

Method	drug-based interactions	target-based interactions
Random	0.0036 (7.79×10^{-53})	0.0182 (7.33×10^{-45})
KBMF2K	0.3698 (4.63×10^{-31})	0.5039 (5.08×10^{-10})
Shi	0.3833 (4.57×10^{-28})	0.5497 (1.21×10^{-03})
DrugE-Rank (all features)	**0.4917**	**0.5906**

Table 4. AUC for 10 × 5-fold cross-validation over Data-1 comparing with state-of-the-art methods (p-values of paired t-test against DrugE-Rank with all features)

Method	drug-based interactions	target-based interactions
Random	0.5007 (2.97×10^{-72})	0.5116 (9.23×10^{-53})
KBMF2K	0.7644 (7.54×10^{-43})	0.8057 (1.94×10^{-16})
Shi	0.8067 (8.01×10^{-29})	0.8332 (7.43×10^{-09})
DrugE-Rank (all features)	**0.8775**	**0.9142**

Table 5. AUPR for independent testing data (FDA approved, new drugs and Experimental drugs) comparing with state-of-the-art methods

Methods	Data-2 new drugs	Data-3 new targets	Data-5 new drugs
Random	0.0021 (0.0002)	0.0010 (8.72×10^{-05})	0.0009 (7.75×10^{-05})
KBMF2K	0.1220	0.1561	0.0310
Shi	0.1241	0.1614	0.0371
DrugE-Rank (all features)	**0.2031** (0.0127)	**0.2831** (0.0078)	**0.0997** (0.0205)

Table 6. AUC for independent testing data (FDA approved, new drugs and Experimental drugs) comparing with state-of-the-art methods

Methods	Data-2 new drugs	Data-3 new targets	Data-5 new drugs
Random	0.4929 (0.0291)	0.4902 (0.0189)	0.5007 (0.0067)
KBMF2K	0.5891	0.6926	0.5676
Shi	0.6174	0.7104	0.5660
DrugE-Rank (all features)	**0.7011** (0.0317)	**0.7937** (0.0274)	**0.6319** (0.0188)

2.3 Complementarity between component methods

The power of DrugE-Rank comes from the effective integration of different component methods, and high complementarity among different component methods will improve the performance of DrugE-Rank more. For the new target prediction on Data-3, Data-1 is randomly divided into five parts of an equal size, four among five being used to train the component methods. We check the complementarity of different component methods on the fifth part of Data-1, which is used to train the LTR model to make prediction on Data-3. Ideally, each component method can predict a different set of true drugs at top positions. We selected top 200 drugs predicted by each method, and plotted the ranks of true interacting drugs by each method for a pair of two component methods, such as BLM-svc vs. BLM-svr. Fig. 1 shows the results by eleven different pairs of component methods (the other four

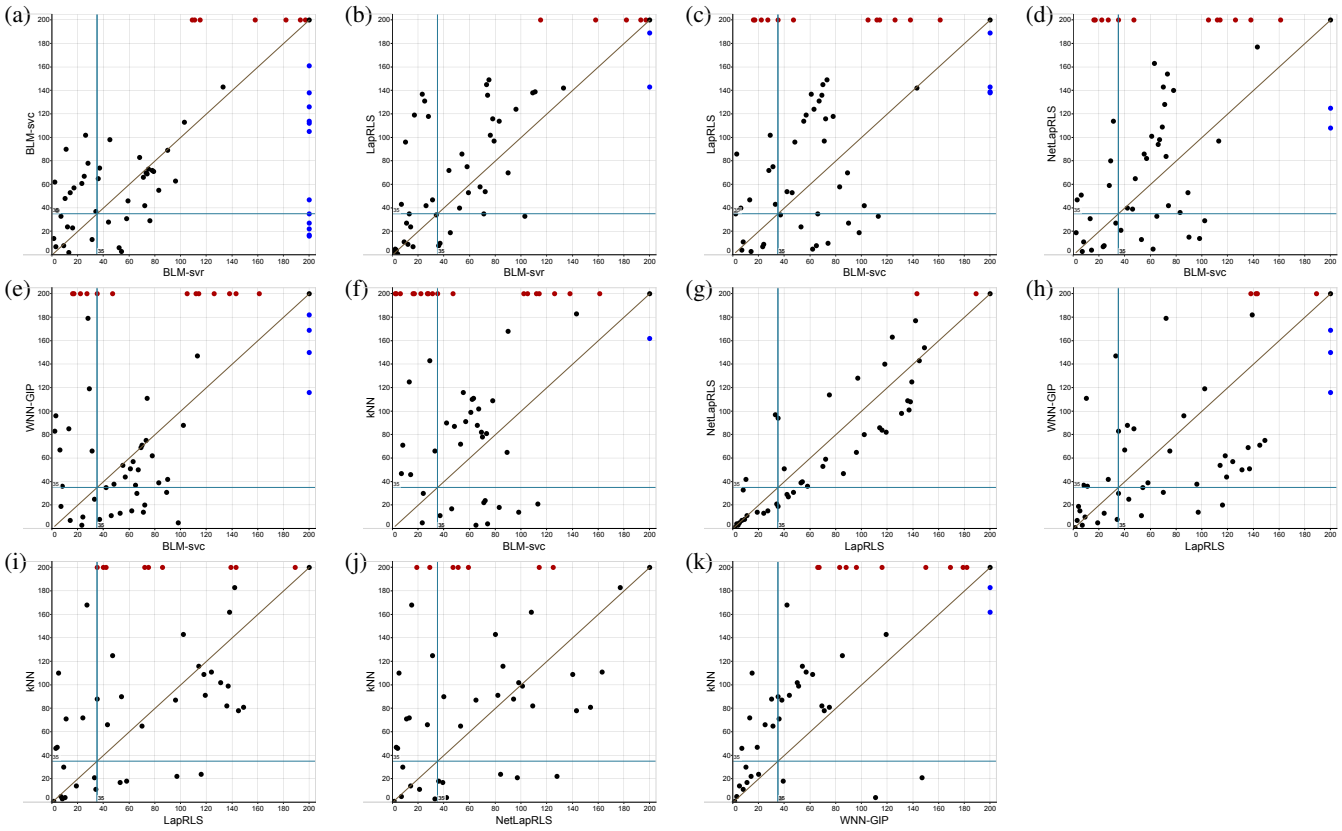

Fig. 1: Complementarity between component methods. The 200 drugs correctly predicted for a new given target by each method is compared with that of another. The comparison is between (a) BLM-svr vs. BLM-svc, (b) BLM-svr vs. LapRLS, (c) BLM-svc vs. LapRLS, (d) BLM-svc vs. NetLapRLS, (e) BLM-svc vs. WNN-GIP, (f) BLM-svc vs. k-NN, (g) LapRLS vs. NetLapRLS, (h) LapRLS vs. WNN-GIP, (i) LapRLS vs. k-NN, (j) NetLapRLS vs. k-NN and (k) WNN-GIP vs. k-NN.

pairs are shown in the main text). So each point in the figure is a true interacting drug, ranked higher than 200 in at least one of the two competitive methods. The red and blue points show those ranked lower than 200 by only one method, while the black points are those ranked lower than 200 by both methods. The points close to the diagonal line are the drugs for which two competing methods perform similarly, while the points far from the diagonal line are the drugs for which one method performs much better than the other method. Fig. 1 shows that no methods can beat other methods clearly. All these results confirm the complementarity of different component methods, which must bring diverse effects to allow DrugE-Rank to improve the predictive performance by their ensemble.

2.4 New prediction analysis

For drug 'Vecuronium' (DrugBank-ID: DB01338) in drug-based cross-validation experiments, target 'Muscarinic acetylcholine receptor M3' (UniProt-ID: P20309) was predicted by all 10 models. By checking DrugBank, we found that this target interacts with Pancuronium (DrugBank-ID: DB01337), which are the most structural similar drugs to Vecuronium. Both Vecuronium and Pancuronium are Androstanols, and can be used as non-depolarizing neuromuscular blocking agent. All these suggest that 'Muscarinic acetylcholine receptor M3' is a promising candidate target for Vecuronium.

For protein '5-hydroxytryptamine receptor 4' (UniProt-ID: Q13639) in target-based cross-validation experiments, drug Loxapine (DrugBank-ID: DB00408) was predicted by 7 out of 10 models. In fact, '5-hydroxytryptamine receptor 4' is in the same family (G-protein coupled receptor 1) with '5-hydroxytryptamine receptor 6' (UniProt-ID: P50406), '5-hydroxytryptamine receptor 7' (UniProt-ID: P34969), '5-hydroxytryptamine receptor 2A' (UniProt-ID: P28223), '5-hydroxytryptamine receptor 2C' (UniProt-ID: P28335), and '5-hydroxytryptamine receptor 1A' (UniProt-ID: P08908). All these five known proteins are the targets of Loxapine, which suggests that '5-hydroxytryptamine receptor 4' is a promising candidate target for Loxapine.

REFERENCES

Bleakley, K. and Yamanishi, Y. (2009). Supervised prediction of drug–target interactions using bipartite local models. *Bioinformatics*, **25**(18), 2397–2403.

Table 7. New interactions predicted by using target-based 10 × 5-fold cross validation results (models) on Data-1. Predicted (top 10) interactions were sorted by how many times they are predicted out of ten times (shown by #times), and the last column shows YES or NO, indicating if the interaction was found in the latest DrugBank database or not.

UniProt ID (Target Name)	DrugBank ID (Drug Name)	#times	Found?
Q8N1C3 (Gamma-aminobutyric acid receptor subunit gamma-1)	DB00898 (Ethanol)	8	YES
Q99928 (Gamma-aminobutyric acid receptor subunit gamma-3)	DB00898 (Ethanol)	8	YES
P78334 (Gamma-aminobutyric acid receptor subunit epsilon)	DB00898 (Ethanol)	8	YES
Q13639 (5-hydroxytryptamine receptor 4)	DB00408 (Loxapine)	7	NO
O94956 (Solute carrier organic anion transporter family member 2B1)	DB01045 (Rifampicin)	7	YES
Q9UM07 (Protein-arginine deiminase type-4)	DB00759 (Tetracycline)	7	YES
P19320 (Vascular cell adhesion protein 1)	DB00898 (Ethanol)	7	YES
P29475 (Nitric oxide synthase, brain)	DB01110 (Miconazole)	6	NO
Q12809 (Potassium voltage-gated channel subfamily H member 2)	DB00537 (Ciprofloxacin)	6	YES
P00374 (Dihydrofolate reductase)	DB00798 (Gentamicin)	6	YES

Table 8. New interactions found by using drug-based 10 × 5-fold cross validation on Data-1.

UniProt ID (Target Name)	DrugBank ID (Drug Name)	#times	Found?
Q16850 (Lanosterol 14-alpha demethylase)	DB01045 (Rifampicin)	10	YES
O94956 (Solute carrier organic anion transporter family member 2B1)	DB01045 (Rifampicin)	10	YES
P48051 (G protein-activated inward rectifier potassium channel 2)	DB00898 (Ethanol)	10	YES
P20309 (Muscarinic acetylcholine receptor M3)	DB01339 (Vecuronium)	10	NO
P21728 (D(1A) dopamine receptor)	DB00933 (Mesoridazine)	10	NO
Q12809 (Potassium voltage-gated channel subfamily H member 2)	DB00537 (Ciprofloxacin)	10	YES
Q02641 (Voltage-dependent L-type calcium channel subunit beta-1)	DB00898 (Ethanol)	9	YES
P46098 (5-hydroxytryptamine receptor 3A)	DB00898 (Ethanol)	8	YES
P00374 (Dihydrofolate reductase)	DB00798 (Gentamicin	7	YES
P20309 ((Muscarinic acetylcholine receptor M3)	DB00209 (Trospium)	7	NO
P20309 (Muscarinic acetylcholine receptor M3)	DB00771 (Clidinium)	7	NO
Q16445 (Gamma-aminobutyric acid receptor subunit alpha-6)	DB00898 (Ethanol)	7	YES
P21728 (D(1A) dopamine receptor)	DB01242 (Clomipramine)	7	NO

Burges, C. (2010). From ranknet to lambdarank to lambdamart: An overview. *Learning*, **11**, 23–581.

Cao, Z., Qin, T., Liu, T.-Y., Tsai, M.-F., and Li, H. (2007). Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*, pages 129–136. ACM.

Chapelle, O. and Chang, Y. (2011). Yahoo! learning to rank challenge overview. In *Proceedings of the Yahoo! Learning to Rank Challenge, held at ICML 2010, Haifa, Israel, June 25, 2010*, pages 1–24.

Crammer, K., Singer, Y., *et al.* (2001). Pranking with ranking. In *Nips*, volume 1, pages 641–647.

Ding, H., Takigawa, I., Mamitsuka, H., and Zhu, S. (2014). Similarity-based machine learning methods for predicting drug–target interactions: a brief review. *Briefings in Bioinformatics*, **15**(5), 734–747.

Herbrich, R., Graepel, T., and Obermayer, K. (1999). Large margin rank boundaries for ordinal regression. *Advances in neural information processing systems*, pages 115–132.

LI, H. (2011). A short introduction to learning to rank. *IEICE TRANSACTIONS on Information and Systems*, **94**(10), 1854–1862.

Li, P., Wu, Q., and Burges, C. J. (2007). Mcrank: Learning to rank using multiple classification and gradient boosting. In *Advances in neural information processing systems*, pages 897–904.

Liu, K., Peng, S., Wu, J., Zhai, C., Mamitsuka, H., and Zhu, S. (2015). Meshlabeler: improving the accuracy of large-scale mesh indexing by integrating diverse evidence. *Bioinformatics*, **31**(12), 339–347.

Liu, T.-Y. (2009). Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, **3**(3), 225–331.

van Laarhoven, T. and Marchiori, E. (2013). Predicting drug-target interactions for new drug compounds using a weighted nearest neighbor profile. *PloS one*, **8**(6), e66952.

van Laarhoven, T., Nabuurs, S. B., and Marchiori, E. (2011). Gaussian interaction profile kernels for predicting drug–target interaction. *Bioinformatics*, **27**(21), 3036–3043.

Xia, F., Liu, T.-Y., Wang, J., Zhang, W., and Li, H. (2008). Listwise approach to learning to rank: theory and algorithm. In *Proceedings of the 25th international conference on Machine learning*, pages 1192–1199. ACM.

Xia, Z. *et al.* (2010). Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. *BMC System Biology*, **4**(Suppl 2), S6.