Drug-target interaction prediction with tree-ensembles on reconstructed networks

Konstantinos Pliakos[1,2] and Celine Vens[1,2]

[1] KU Leuven, Campus KULAK, Faculty of Medicine, Belgium
[2] ITEC, imec research group at KU Leuven
konstantinos.pliakos@kuleuven.be,celine.vens@kuleuven.be

Abstract. Computational prediction of drug-target interactions (DTI) is vital for drug discovery. Despite modern technological advances, drug development remains extremely expensive and time consuming. Therefore, *in silico* DTI predictions based on machine learning are needed. Here, we propose a new learning method which addresses DTI prediction as a multi-output prediction task by learning ensembles of multi-output bi-clustering trees (eBICT) on reconstructed networks. The proposed approach integrates background information from both drug and target protein spaces into the same global network framework. For evaluation purposes, we used several benchmark datasets that represent drug-protein networks. We showed that building tree-ensemble learning models with output space reconstruction leads to superior prediction results, while preserving the advantages of tree-ensembles, such as scalability, interpretability and inductive setting.

Keywords: drug-target networks · network reconstruction · interaction prediction · tree-ensembles · multi-output prediction

1 Introduction

Predicting drug-target interactions (DTI) is vital for the development of new drugs. Accurate and efficient identification of interactions between drugs and target proteins can accelerate the drug development process and reduce the required cost. It also assists scientists to foresee adverse effects of drugs [4, 7]. Apart from discovering new drugs, DTI prediction can also leverage drug repositioning [4, 2, 14, 5], which aims at revealing new uses for already approved drugs. However, despite the efforts made by the scientific community, experimentally identifying DTIs remains extremely time-consuming and expensive [9]. Therefore, effective machine learning models for DTI prediction are needed.

Particularly interesting is the machine learning task of multi-output (multi-target) prediction [13], where the model learns to predict multiple output variables at the same time. The interest in multi-output models is great for drug discovery as we have moved from the old paradigm of '*one target, one drug, one disease*' to the era of *polypharmacology*. It is known that drugs which interact with multiple target proteins are more effective [15]. Multi-output learning

can also contribute to investigating the off-target drug activity (i.e., unintended function of a drug), leading to new uses for existing drugs (drug repositioning) or contrarily, the identification of unwanted side-effects. Such adverse reactions of drug candidates are usually identified at a later stage of the drug development process, leading to extremely expensive *late stage failures*.

Nowadays, the interest of the scientific community is focused on the setting of *chemogenomics* [3]. The underlying idea behind this is that drug information is integrated with target information and thereby complement each other. However, typical approaches are mostly based on matrix factorization (MF) or graph learning, following the transductive setup (i.e. test instances are needed in the training phase). There are also other methods that train binary classifiers over the Cartesian product of drug and target-related feature sets. This leads to a huge data matrix and thus, these methods are computationally very expensive. Recently, another family of methods that draws increasing attention is GNN based approaches, such as [17, 1].

DTI networks are bi-partite ones that consist of two sets of nodes D and P, corresponding to drugs and targets, respectively. Each node is represented by a feature vector. Drug features may consist of chemical structure similarities, drug side effects, or drug-drug interactions. Target features may consist of protein sequence similarities, GO annotations, protein-protein interactions or protein functions. A link between two nodes of a DTI network represents an interaction between the corresponding drug and target. The set of existing or not existing links form an interaction matrix $\mathbf{Y} \in \Re^{|D|\times|P|}$. Every $y(i,j) \in \mathbf{Y}$ is 1 if an interaction between d_i and p_j exists and 0 otherwise. In Fig. 1, an illustration of a network in the aforementioned setting is displayed.

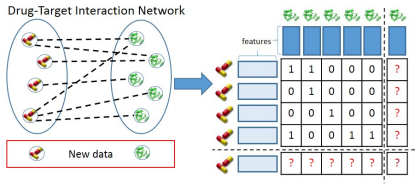

Fig. 1. Illustration of a (bi-partite) DTI interaction network.

2 Proposed Method

The proposed approach learns bi-clustering trees with output space reconstruction (BICTR), integrating tree-ensembles with semi-supervised approaches, such as MF. Here, we promote ensembles of bi-clustering trees (eBICT) [10] and NRLMF [6].

$$\begin{aligned}\min_{\mathbf{U},\mathbf{V}} \sum_{i=1}^{|D|}\sum_{j=1}^{|P|}(1+cY_{ij}-Y_{ij})\ln\left[1+\exp(u_i v_j^T)\right] - cY_{ij}u_i v_j^T \\ + \lambda_d||\mathbf{U}||_F^2 + \lambda_p||\mathbf{V}||_F^2 \\ + \alpha\mathrm{Tr}(\mathbf{U}^T\mathbf{L}^d\mathbf{U}) + \beta\mathrm{Tr}(\mathbf{V}^T\mathbf{L}^p\mathbf{V})\end{aligned} \tag{1}$$

We first reconstruct the output space, exploiting neighborhood information, revealing underlying manifolds in the topology of the DTI network (i.e. $\mathbf{Y}$) and alleviating class-imbalance. The input of our approach is the drug feature space $\mathbf{X_d}$, the target feature space $\mathbf{X_p}$, and the interaction matrix $\mathbf{Y}$. We reconstruct the DTI network by learning matrices $\mathbf{U}$ and $\mathbf{V}$ based on Eq. 1 [6]. The new interaction matrix is denoted as $\hat{\mathbf{Y}}$ and every $\hat{y_{ij}} \in \hat{\mathbf{Y}}$ is computed as: $\hat{y}_{ij} = \frac{\exp(\mathbf{u}_i \mathbf{v}_j^T)}{1+\exp(\mathbf{u}_i \mathbf{v}_j^T)}$.

Next, we learn eBICT on the reconstructed output space. In more detail, the input for every tree in our ensemble is the drug feature space $\mathbf{X_d}$, the target feature space $\mathbf{X_p}$, and the reconstructed interaction matrix $\hat{\mathbf{Y}}$. The root node of every tree in our setting contains the whole interaction network and a partitioning of this network is conducted in every node. The tree growing process is based on both vertical and horizontal splits of the reconstructed interaction matrix $\hat{\mathbf{Y}}$. The variance reduction is computed as $Var = \sum_j^{|P|} Var(\hat{\mathbf{Y}}_j)$ when the split test is on $\phi_d \in \mathbf{X_d}$ and $Var = \sum_i^{|D|} Var(\hat{\mathbf{Y}}_i^T)$ when the split test is on a $\phi_p \in \mathbf{X_p}$.

The NRLMF-based reconstruction step of the proposed strategy boosts the predictive performance of the eBICT while preserving all the advantages of tree-ensembles, such as scalability, computational efficiency, and interpretability. Our approach, despite being integrated with MF, continues to follow the inductive setup. In more detail, the output space reconstruction process takes place only in the training process. After the training model is complete, new instances that may arrive (e.g., new candidate drugs) just traverse the grown bi-clustering trees and predictions are assigned to them based on the leaves in which they end up.

3 Data and Results

The benchmark datasets [16] that were used are displayed in Table 1. Our evaluation study begins with comparing the proposed approach BICTR against eBICT without output space reconstruction. Next, we compare BICTR to three state of the art DTI prediction methods, BLMNII [8], STC [12], and NRLMF [6]. The methods are compared in three prediction settings; $T_d \times L_p$, predicting interactions between new drug candidates and known targets, $L_d \times T_p$, predicting

Table 1. The DTI networks used in the experimental evaluation are presented.

DPN	$\|drugs\| \times \|proteins\|$	$\|Features\|$	$\|interactions\|$
NR	54×26	$54 - 26$	90/1404 (6.4%)
GR	223×95	$223 - 95$	635/21185 (3%)
IC	210×204	$210 - 204$	1476/42840 (3.4%)
E	445×664	$445 - 664$	2926/295480 (1%)

interactions between known drugs and new targets, and $T_d \times T_p$, predicting interactions between new drug candidates and new targets. In $T_d \times L_p$ and $L_d \times T_p$ we used 10-fold cross validation (CV) on nodes while in $T_d \times T_p$, we used 5-fold CV over blocks of drugs and targets. The number of trees in tree-ensembles was set to 100 and the weight of positive interactions c in Eq.1 to 5. All other parameters of NRLMF, BLMNII, and STC were optimized in 5-fold CV inner tuning (nested CV) following grid search. Additional experimental results are shown in [11].

The area under the receiver operating characteristic curve (AUROC) results are presented in Table 2. Best results are shown in bold faces and * indicates that the results between BICTR and its competitor were found statistically significantly different ($p < 0.05$) based on a Wilcoxon Signed-Ranks Test run on the CV-folds. BICTR outperforms eBICT in all three prediction settings. Thus, the original hypothesis that network reconstruction can boost the predictive performance of multi-output learning models is verified. BICTR also outperforms all other competitors, affirming its effectiveness.

Table 2. AUROC results for the compared methods.

AUROC					
	$T_d \times L_p$				
Data	$BICTR$	$eBICT$	$NRLMF$	$BLMNII$	STC
NR	**0.875**	0.787*	0.851*	0.807*	0.794*
GR	**0.894**	0.857*	0.867*	0.842*	0.847*
IC	**0.811**	0.780*	0.792	0.737*	0.783*
E	**0.891**	0.827*	0.777*	0.815*	0.794*
Avg	**0.868**	0.813	0.822	0.800	0.805
	$L_d \times T_p$				
Data	$BICTR$	$eBICT$	$NRLMF$	$BLMNII$	STC
NR	**0.905**	0.614*	0.747*	0.667*	0.525*
GR	**0.951**	0.846*	0.861*	0.776*	0.800*
IC	**0.968**	0.931*	0.949*	0.887*	0.909*
E	**0.973**	0.924*	0.940*	0.904*	0.906*
Avg	**0.949**	0.829	0.874	0.809	0.785
	$T_d \times T_p$				
Data	$BICTR$	$eBICT$	$NRLMF$	$BLMNII$	STC
NR	0.676	0.634*	**0.683**	0.554*	0.469*
GR	**0.811**	0.792*	0.800*	0.475*	0.630*
IC	**0.733**	0.719*	0.731	0.466*	0.649*
E	**0.812**	0.785*	0.749*	0.490*	0.682*
Avg	**0.758**	0.733	0.741	0.496	0.608

4 Conclusion

Here, we have presented a new drug-target interaction prediction approach based on multi-output prediction with output space reconstruction. We showed that multi-output learning models can manifest superior DTI predictive performance when built on reconstructed networks.

References

1. Alet, F., Weng, E., Lozano-Pérez, T., Kaelbling, L.P.: Neural relational inference with fast modular meta-learning. In: Advances in Neural Information Processing Systems, pp. 11827–11838 (2019)
2. Ashburn, T.T., Thor, K.B.: Drug repositioning: identifying and developing new uses for existing drugs. Nature Reviews Drug Discovery **3**(8), 673–683 (aug 2004)
3. Ezzat, A., Wu, M., Li, X.L., Kwoh, C.K.: Computational prediction of drug-target interactions using chemogenomic approaches: an empirical survey. Briefings in Bioinformatics (jan 2018)
4. Keiser, M.J., Setola, V., Irwin, J.J., Laggner, C., Abbas, A.I., Hufeisen, S.J., Jensen, N.H., Kuijer, M.B., Matos, R.C., Tran, T.B., Whaley, R., Glennon, R.A., Hert, J., Thomas, K.L.H., Edwards, D.D., Shoichet, B.K., Roth, B.L.: Predicting new molecular targets for known drugs. Nature **462**(7270), 175–181 (nov 2009)
5. Li, J., Zheng, S., Chen, B., Butte, A.J., Swamidass, S.J., Lu, Z.: A survey of current trends in computational drug repositioning. Briefings in Bioinformatics **17**(1), 2–12 (jan 2016)
6. Liu, Y., Wu, M., Miao, C., Zhao, P., Li, X.L.: Neighborhood Regularized Logistic Matrix Factorization for Drug-Target Interaction Prediction. PLOS Computational Biology **12**(2), e1004760 (feb 2016)
7. Lounkine, E., Keiser, M.J., Whitebread, S., Mikhailov, D., Hamon, J., Jenkins, J.L., Lavan, P., Weber, E., Doak, A.K., Côté, S., Shoichet, B.K., Urban, L.: Large-scale prediction and testing of drug activity on side-effect targets. Nature **486**(7403), 361–367 (jun 2012)
8. Mei, J.P., Kwoh, C.K., Yang, P., Li, X.L., Zheng, J.: Drug-target interaction prediction by learning from local information and neighbors. Bioinformatics **29**(2), 238–245 (jan 2013)
9. Paul, S.M., Mytelka, D.S., Dunwiddie, C.T., Persinger, C.C., Munos, B.H., Lindborg, S.R., Schacht, A.L.: How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nature Reviews Drug Discovery **9**(3), 203–214 (mar 2010)
10. Pliakos, K., Vens, C.: Network inference with ensembles of bi-clustering trees. BMC Bioinformatics **20**(1), 525 (2019)
11. Pliakos, K., Vens, C.: Drug-target interaction prediction with tree-ensemble learning and output space reconstruction. BMC Bioinformatics **21**(49) (2020)
12. Shi, J.Y., Yiu, S.M., Li, Y., Leung, H.C., Chin, F.Y.: Predicting drug-target interaction for new drugs using enhanced similarity measures and super-target clustering. Methods **83**, 98–104 (jul 2015)
13. Waegeman, W., Dembczyński, K., Hüllermeier, E.: Multi-target prediction: a unifying view on problems and methods. Data Mining and Knowledge Discovery pp. 1–32 (nov 2018)

14. Wu, Z., Cheng, F., Li, J., Li, W., Liu, G., Tang, Y.: SDTNBI: an integrated network and chemoinformatics tool for systematic prediction of drug-target interactions and drug repositioning. Briefings in Bioinformatics **18**(2), bbw012 (mar 2016)
15. Xie, L., Xie, L., Kinnings, S.L., Bourne, P.E.: Novel Computational Approaches to Polypharmacology as a Means to Define Responses to Individual Drugs. Annual Review of Pharmacology and Toxicology **52**(1), 361–379 (feb 2012)
16. Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W., Kanehisa, M.: Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. Bioinformatics **24**(13), i232–i240 (jul 2008)
17. Zitnik, M., Agrawal, M., Leskovec, J.: Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics **34**(13), i457–i466 (2018)