ALADIN: A New Approach for Drug–Target Interaction Prediction

Krisztian Buza[1] and Ladislav Peska[2]

[1] Knowledge Discovery and Machine Learning, Rheinische
Friedrich-Wilhelms-Universität Bonn, Germany
buza@iai.uni-bonn.de
WWW home page: https://www.kdml.iai.uni-bonn.de/
[2] Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic
peska@ksi.mff.cuni.cz

Abstract. Due to its pharmaceutical applications, one of the most prominent machine learning challenges in bioinformatics is the prediction of drug–target interactions. State-of-the-art approaches are based on various techniques, such as matrix factorization, restricted Boltzmann machines, network-based inference and bipartite local models (BLM). In this paper, we extend BLM by the incorporation of a hubness-aware regression technique coupled with an enhanced representation of drugs and targets in a multi-modal similarity space. Additionally, we propose to build a projection-based ensemble. Our *Advanced Local Drug-Target Interaction Prediction* technique (ALADIN) is evaluated on publicly available real-world drug–target interaction datasets. The results show that our approach statistically significantly outperforms BLM-NII, a recent version of BLM, as well as NetLapRLS and WNN-GIP.

Keywords: drug–target interaction prediction, bipartite local models, ALADIN

1 Introduction

Prediction of drug–target interactions is one of the most prominent machine learning applications in the pharmaceutical industry, the importance of which is underlined by the fact that both time and expenditure related to drug development are enormous: on average, it costs ≈$1.8 billion and takes more than 10 years to bring a new drug to the market [17]. Drug–target interaction prediction (DTI) techniques promise to reduce the aforementioned cost and time, and to support drug repositioning [40], i.e., the use of an existing medicine to treat a disease that has not been treated with that drug yet.

Computational methods for DTI include approaches based on molecular docking simulations [9], [15] and ligand chemistry [21], [25]. Furthermore, text mining techniques have been proposed to identify biomedical entities and relations between them [7], [13], [28], [42]. However, a serious limitation of docking-based approaches is that they require information about the three-dimensional structure of candidate drugs and targets which is often not available, especially for G-protein coupled receptors (GPCRs) and ion channels. Additionally, the performance of ligand-based approaches is known to decrease if only few ligands are known. Therefore, machine learning techniques have been proposed for DTI [11], [19], [39]. Recent approaches are based on matrix factorization [5], [14], [41], support vector regression [34, 35], restricted Boltzmann machines [37], network-based inference [8], [10], decision lists [30] and bipartite local models (BLM) [4] with semi-supervised prediction [38], improved kernels [22] and the incorporation of neighbor-based interaction-profile inferring [23].

Real-world datasets in biology, chemistry and medicine [1], including drug–target interaction networks, have been shown to contain hubs, i.e., vertices that are connected to surprisingly many other vertices. For example, in the *Enzyme* dataset (described in Section 5.1), the vast majority of targets have less then 5 interactions, while some of the targets are very popular: each of 30 most popular targets interacts with 20 drugs at least. Despite such observations, none of the aforementioned variants of BLM took the presence of hubs into account. Furthermore, the presence of hubs has been observed in nearest neighbor graphs [29], which lead to the development of hubness-aware classifiers [33] and regression techniques [6]. Although hubness-aware techniques are among the most promising recent machine learning approaches, their potential to enhance drug–target interaction prediction methods has not been exploited yet.

In this paper, we extend BLM by the incorporation of a hubness-aware regression approach. Additionally, we propose an enhanced representation of drugs and targets in a multi-modal similarity space and build a projection-based ensemble. We call the resulting approach *Advanced Local Drug-Target Interaction Prediction*, or ALADIN for short. In order to assist reproducibility of our work, we perform experiments on publicly available real-world drug–target interaction datasets. The results show that our approach outperforms BLM-NII [23], a recent version of BLM, and two other drug–target prediction techniques.

The rest of this paper is organized as follows: in Section 2, we define the drug–target interaction prediction problem, this is followed by the review of BLM and hubness-aware regression in Section 3. We describe our approach, ALADIN, in Section 4 and present the results of experimental evaluation in Section 5. Finally, we conclude in Section 6.

2 Basic Notation and Problem Formulation

First, we define the Drug–Target Interaction Prediction problem. We are given a set $\mathcal{D} = \{d_1, \dots, d_n\}$ of n drugs, a set $\mathcal{T} = \{t_1, \dots, t_m\}$ of m pharmaceutical targets, an $n \times n$ drug similarity matrix $\mathcal{S}^D$, an $m \times m$ target similarity matrix $\mathcal{S}^T$ and an $n \times m$ interaction matrix $\mathcal{M}$. Each entry $s^D_{i,j}$ of $\mathcal{S}^D$ (and $s^T_{i,j}$ of $\mathcal{S}^T$, resp.) describe the similarity between drugs d_i and d_j (targets t_i and t_j). Each entry $m_{i,j}$ of $\mathcal{M}$ denotes if drug d_i and target t_j are known to interact:

$$m_{i,j} = \begin{cases} 1 & \text{if there is a known interaction between } d_i \text{ and } t_j \\ 0 & \text{otherwise.} \end{cases}$$

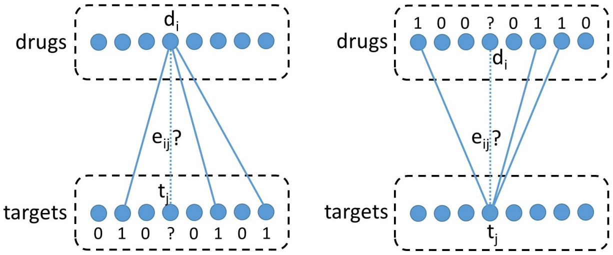


Fig. 1. Two independent predictions of Bipartite Local Models.

This formulation is in accordance with the usual setting in which only positive information is available: in case if $m_{i,j} = 0$, the corresponding drug d_i and target t_j *may* or *may not* interact, therefore, we call $u_{i,j} = (d_i, t_j)$ an *unknown* pair. The task is to predict the likelihood of interaction for each unknown pair.

At the first glance, the above DTI problem seem to be similar to the problems considered in the recommender systems community. Note, however, that most recommender techniques consider only the interactions ("ratings") because even a few ratings are thought to be more informative than metadata, such as users' similarity based on their demographic information [27]. In contrast, drug–drug and target–target similarities play an essential role in DTI.

3 Background

In this section, we review the BLM approach and hubness-aware error correction for nearest neighbor regression.

3.1 Bipartite Local Model

BLM considers DTI as a link prediction problem in bipartite graphs [4]. The vertices in one of the vertex classes correspond to drugs, whereas the vertices in the other vertex class correspond to targets. There is an edge between drug d_i and target t_j if and only if $m_{i,j} = 1$.

The likelihood of unknown interactions is predicted as follows: we consider an unknown pair $u_{i,j} = (d_i, t_j)$ and calculate the likelihood of interaction as the aggregate of two independent predictions.

The first prediction (Fig. 1, left panel) is based on the relations between d_i and the targets. Each target t_k (except t_j) is labeled as "1" or "0" depending on $m_{i,k}$. Then a model is trained to distinguish "1"-labeled and "0"-labeled targets. Subsequently, this model is applied to predict the likelihood of interaction for the unknown pair $u_{i,j}$. This first prediction is denoted by $\hat{y}'_{i,j}$.

The second prediction, $\hat{y}''_{i,j}$, is obtained in a similar fashion, but instead of considering the interactions of drug d_i and labeling the targets, the interactions of target t_j are considered and drugs are labeled (Fig. 1, right panel). The models that make the first and second predictions are called *local models*.

In order to obtain the final prediction of the BLM, we average the predictions of the aforementioned local models:

$$\hat{y}_{i,j} = \frac{\hat{y}'_{i,j} + \hat{y}''_{i,j}}{2} \tag{1}$$

Note that instead of averaging, other aggregation functions, such as minimum or maximum are possible as well.

BLM is a generic framework in which various regressors or classifiers can be used as local models. Bleakley and Yamanishi [4] used support vector machines with a domain-specific kernel. In contrast, we propose to use a hubness-aware regression technique, EC*k*NN, which is described next.

3.2 EC*k*NN: *k*-Nearest Neighbor Regression with Error Correction

In the last decades, various regression schemes have been introduced, such as linear and polynomial regression, support vector regression, neural networks, etc. One of the most popular regression techniques is based on k-nearest neighbors: when predicting the numeric label on an instance x with k-nearest neighbor regression, the k-nearest neighbors of x (i.e., k instances that are most similar to x) are determined and the average of their labels is calculated as the predicted label of x. In our case, instances may either correspond to drugs or targets, depending on whether the first or the second BLM-prediction is calculated.

While being intuitive and simple to implement, k-nearest neighbor regression is well-understood from the point of view of theory as well, see e.g. [3], and the references therein for an overview of the most important theoretical results. The theoretical results are also justified by empirical studies: for example, in their recent paper, Stensbo-Smidt et al. found that nearest neighbor regression outperforms model-based prediction of star formation rates [31], while Hu et al. showed that a model based on k-nearest neighbor regression is able to estimate the capacity of lithium-ion batteries [18].

Despite all of the aforementioned advantages of k-nearest neighbor regression, one of its recently explored shortcomings is its suboptimal performance in the presence of bad hubs. Intuitively, bad hubs are instances that appear as nearest neighbors of many other instances, but have substantially different labels from those instances. The presence of bad hubs has been shown to be related to the intrinsic dimensionality of the data. This means, roughly speaking, that bad hubs are expected in complex data, such as drug–target interaction data. For a more detailed discussion, we refer to [6].

In order to alleviate the detrimental effect of bad hubs, in [6] we proposed an error correction technique which is reviewed next. We define the *corrected label* $y_c(x)$ of a training instance x as

$$y_c(x) = \begin{cases} \frac{1}{|\mathcal{R}_x|} \sum\limits_{x_i \in \mathcal{R}_x} y(x_i) & \text{if } |\mathcal{R}_x| \geq 1 \\ y(x), & \text{otherwise} \end{cases}, \tag{2}$$

where $y(x_i)$ denotes the original (i.e., uncorrected) label of instance x_i, and $\mathcal{R}_x$ is the set of "reverse neighbors", i.e. the set of training instances that have x as one of their k-nearest neighbors:

$$\mathcal{R}_x = \{\forall x_i | x \in \mathcal{N}(x_i)\} \tag{3}$$

where $\mathcal{N}(x_i)$ denotes the set of k-nearest neighbors of x_i.

In order to make predictions, k-nearest neighbor regression with error correction (ECkNN) uses the corrected labels. Given a "new" (unlabeled) instance x', its predicted label $\hat{y}(x')$ is calculated as follows:

$$\hat{y}(x') = \frac{1}{k} \sum_{x_i \in \mathcal{N}(x')} y_c(x_i). \tag{4}$$

4 Our approach

Next, we present ALADIN, our Advanced Local Drug-Target Interaction Prediction approach. Following subsections describe the components of ALADIN.

4.1 Similarity-based Representation

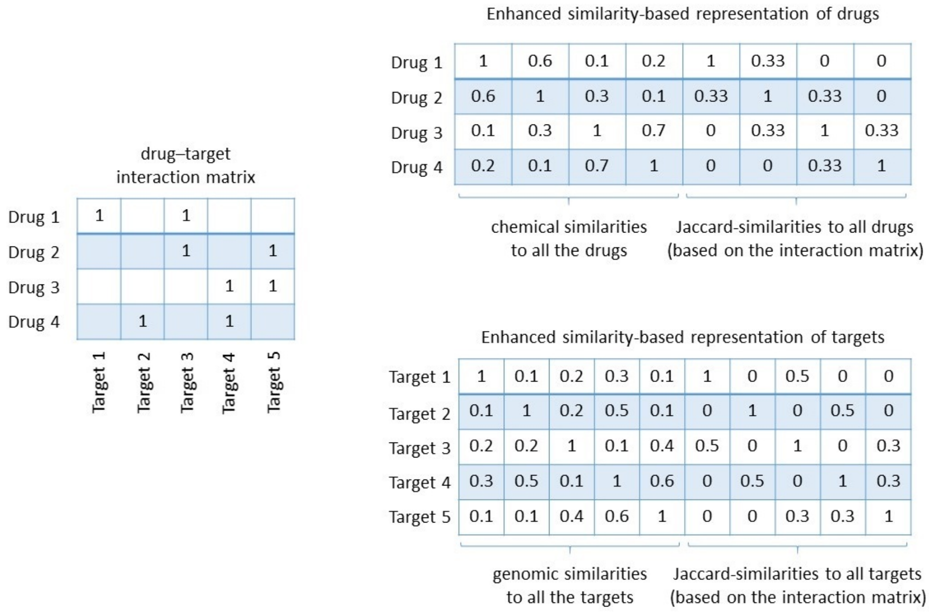
The given drug–drug similarities allow us to represent drugs in the similarity space: in particular, drug d_i is represented by the vector $(s^D_{i,1}, \dots s^D_{i,n})$. Given the target similarity matrices, targets may be represented in an analogous way, i.e., using their similarities to all the targets.

Additionally to the given drug–drug and target–target similarities, we propose to compute drug–drug and target–target similarities based on the known interactions (i.e., interactions in the training set). In particular, using the interaction matrix, we calculate the Jaccard-similarity between drugs as well as between targets. Thus the enhanced similarity-based representation of a drug (or target, respectively) consists of its chemical (genetic) similarity to all the drugs (targets) and its interaction-based similarity to all the drugs (targets). This is illustrated in Fig. 2.

4.2 Projection-based Ensemble

We propose to build a projection-based ensemble of BLMs as follows. Given the enhanced similarity-based representation of drugs and targets, we select a random subset of features and use only the selected features when training the local models (ECkNN) and making predictions. Denoting the size of the set of selected features by F_D and F_T (for drugs and targets, respectively), the above procedure first projects drugs into F_D-dimensional, and targets into an F_T-dimensional subspace. Subsequently, these lower dimensional representations are used with the prediction models.

The above process of random selection of features and making predictions using the resulting lower-dimensional representation is repeated N-times. This



drug–target interaction matrix

	Target 1	Target 2	Target 3	Target 4	Target 5
Drug 1	1		1		
Drug 2			1		1
Drug 3				1	1
Drug 4		1		1	

Enhanced similarity-based representation of drugs

	chemical similarities to all the drugs				Jaccard-similarities to all drugs (based on the interaction matrix)			
Drug 1	1	0.6	0.1	0.2	1	0.33	0	0
Drug 2	0.6	1	0.3	0.1	0.33	1	0.33	0
Drug 3	0.1	0.3	1	0.7	0	0.33	1	0.33
Drug 4	0.2	0.1	0.7	1	0	0	0.33	1

Enhanced similarity-based representation of targets

	genomic similarities to all the targets					Jaccard-similarities to all targets (based on the interaction matrix)				
Target 1	1	0.1	0.2	0.3	0.1	1	0	0.5	0	0
Target 2	0.1	1	0.2	0.5	0.1	0	1	0	0.5	0
Target 3	0.2	0.2	1	0.1	0.4	0.5	0	1	0	0.3
Target 4	0.3	0.5	0.1	1	0.6	0	0.5	0	1	0.3
Target 5	0.1	0.1	0.4	0.6	1	0	0	0.3	0.3	1

Fig. 2. Illustration of enhanced similarity-based representation of drugs and targets

results in an ensemble of N prediction models. As each member of the ensemble is constructed in the same way, their expected prediction accuracies will be similar, therefore, we propose to average the predictions of the members of the ensemble. Thus the final output of the ensemble is:

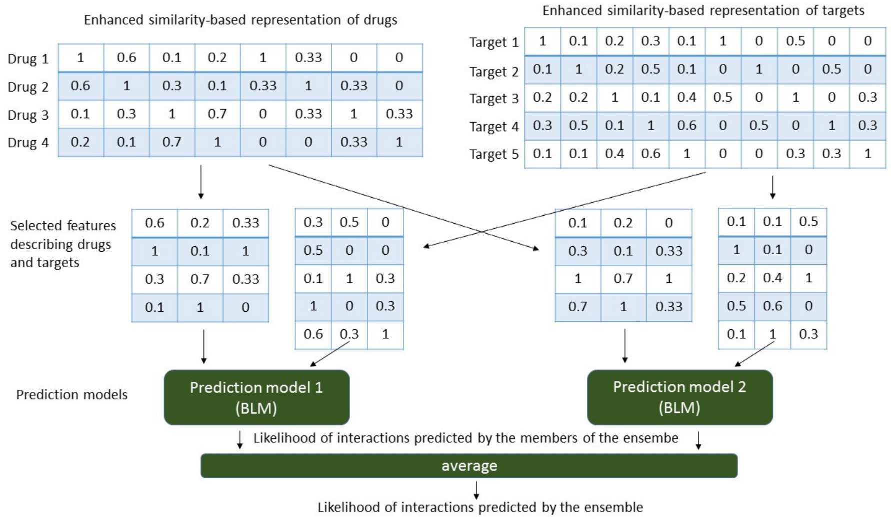
$$\hat{y}_{i,j} = \frac{1}{N} \sum_{l=1}^{N} \hat{y}_{i,j}^{(l)} \tag{5}$$

where $\hat{y}_{i,j}^{(l)}$ is the prediction of the l-th BLM for the unknown pair $u_{i,j}$.

The projection-based ensemble is illustrated in Fig. 3 for $N = 2$ base prediction models with $F_D = F_T = 3$ features selected from the enhanced similarity-based representation.

4.3 Prediction for New Drugs and Targets

One of the shortcomings of the BLM approach is that it does not handle the case of new drugs/targets. With new drug (target, resp.), we mean a drug d (target t) that does not have any known interaction in the training data. In such cases, BLM labels all targets (drugs) as “0”, consequently, no reasonable local model can be learned. In order to alleviate this problem, we use the weighted profile [39] approach to obtain predictions for new drugs/targets.


Enhanced similarity-based representation of drugs
Enhanced similarity-based representation of targets
Selected features describing drugs and targets
Prediction models
Prediction model 1 (BLM)
Prediction model 2 (BLM)
Likelihood of interactions predicted by the members of the ensembe
average
Likelihood of interactions predicted by the ensemble

Drug 1	1	0.6	0.1	0.2	1	0.33	0	0
Drug 2	0.6	1	0.3	0.1	0.33	1	0.33	0
Drug 3	0.1	0.3	1	0.7	0	0.33	1	0.33
Drug 4	0.2	0.1	0.7	1	0	0	0.33	1

Target 1	1	0.1	0.2	0.3	0.1	1	0	0.5	0	0
Target 2	0.1	1	0.2	0.5	0.1	0	1	0	0.5	0
Target 3	0.2	0.2	1	0.1	0.4	0.5	0	1	0	0.3
Target 4	0.3	0.5	0.1	1	0.6	0	0.5	0	1	0.3
Target 5	0.1	0.1	0.4	0.6	1	0	0	0.3	0.3	1

0.6	0.2	0.33
1	0.1	1
0.3	0.7	0.33
0.1	1	0

0.3	0.5	0
0.5	0	0
0.1	1	0.3
1	0	0.3
0.6	0.3	1

0.1	0.2	0
0.3	0.1	0.33
1	0.7	1
0.7	1	0.33

0.1	0.1	0.5
1	0.1	0
0.2	0.4	1
0.5	0.6	0
0.1	1	0.3

Fig. 3. Projection-based ensemble of BLMs using the enhanced similarity-based representation of drugs and targets.

Given a new drug d_i, and a target t_j, we predict the likelihood of the interaction between d_i and t_j as follows:

$$\hat{y}'_{i,j} = \frac{\sum\limits_{d_k \in \mathcal{D} \setminus \{d_i\}} m_{k,j} \mathcal{S}^D_{i,k}}{\sum\limits_{d_k \in \mathcal{D} \setminus \{d_i\}} \mathcal{S}^D_{i,k}}. \tag{6}$$

The intuition behind Eq. (6) is that similar drugs are likely to behave similarly in terms of their interaction with a given target. Therefore, drugs are weighed according to their similarity to the new drug d_i and we calculate the weighted average of the known interactions of other drugs with the same target.

The case of new targets is analogous. Given a new target t_j and a drug d_i, the weighted profile approach can be used to calculate the prediction for the likelihood of the interaction between d_i and t_j as follows:

$$\hat{y}''_{i,j} = \frac{\sum\limits_{t_k \in \mathcal{T} \setminus \{t_j\}} m_{i,k} \mathcal{S}^T_{j,k}}{\sum\limits_{t_k \in \mathcal{T} \setminus \{t_j\}} \mathcal{S}^T_{j,k}}. \tag{7}$$

Although the weighted profile approach is more general than BLM, in the sense that it can be used for new drugs/targets as well, the predictions of the weighted profile approach are less accurate than the predictions of BLM. Therefore, we use the weighted profile approach instead of BLM *only* in case of *new* drugs and targets. We summarize the proposed approach in Alg. 1.

Algorithm 1 Advanced Local Drug-Target Interaction Prediction (ALADIN)

```
Require: Drug–Target interaction matrix $I$, Drug–drug similarity matrix $S^D$, Target–
    target similarity matrix $S^T$, number of nearest neighbors $k$, ensemble size $N$, num-
    ber of selected features $F_D$, $F_T$
Ensure: Likelihood of drug–target interactions
$D \leftarrow$ enhanced similarity-based representations of drugs
$T \leftarrow$ enhanced similarity-based representations of targets
for $l = 1 \dots N$ do
    $D' \leftarrow$ random subset of $D$ with $F_D$ features
    $T' \leftarrow$ random subset of $T$ with $F_T$ features
    Predict interaction scores with BLM using EC$k$NN as local model and
      $D'$ and $T'$ as the representation of drugs and targets.
      (Use the weighted profile approach instead of BLM in case of new
      drugs/targets.)
end for
Average the predictions made in each execution of the loop
```

5 Experimental Evaluation

In order to assist reproducibility of our work, we evaluated our approach on publicly available real-word drug–target interaction data. Next we describe the data and the experimental protocol in detail. This is followed by the discussion of our experimental results.[3]

5.1 Experimental Settings

Datasets We performed experiments on five drug–target interaction datasets (Tab. 1), namely Enzyme, Ion Channel, G-protein coupled receptors (GPCR), Nuclear Receptors (NR), and Kinase.[4] These datasets have been used in various studies previously, see e.g. [4], [12], [14], [24], [38], [39].

The first four datasets contain binary interaction matrices between drugs and targets, each entry of which indicates whether the interaction between the corresponding drug and target is known. In contrast, Kinase contains continuous values of binding affinity for all drug–target pairs of the dataset. In order to produce a binary interaction matrix, we used the same cutoff threshold as Pahikkala et al. [24].

Additionally, each dataset contains a drug–drug similarity matrix and a target–target similarity matrix. In case of the Enzyme, Ion Channel, GPCR and NR datasets, chemical structure similarities between drugs were computed using the SIMCOMP algorithm [16], while the Kinase dataset contains 2D Tanimoto coefficients. Similarities between targets were determined by the Smith-Waterman algorithm, see [12] and [39] for details.

[3] See `http://www.biointelligence.hu/dti` for further results.

[4] The datasets are available at
`https://zenodo.org/record/556337#.WPiAzIVOIdV`

Table 1. Number of drugs, targets and interactions in the datasets used in our study.

Dataset	# Drugs	# Targets	# Interactions
Enzyme	445	664	2926
Ion Channels	210	204	1476
G-protein coupled receptors (GPCR)	223	95	635
Nuclear Receptors (NR)	54	26	90
Kinase	68	442	1527

Evaluation Protocol Although leave-one-out cross-validation is popular in the DTI literature [4], [22] [23], in their recent study, Pahikkala et al. [24] argue that it may lead to overoptimistic results. Thus, we performed experiments according to the interaction-based 5×5-fold cross-validation protocol (in each round of the cross-validation, the test set contains one fifth of all the drug–target pairs).

Evaluation Metrics We evaluated the predictions both in terms of Area Under ROC Curve (AUC) and Area Under Precision-Recall Curve (AUPR). AUC and AUPR values were calculated in each round of the cross-validation. We report averaged values. Additionally, we performed paired t-test at significance level $p = 0.01$ in order to judge if the observed differences are statistically significant.

Baselines We compared our approach, ALADIN, with other drug–target interaction prediction techniques, such as BLM-NII, NetLapRLS and WNN-GIP. BLM-NII is a recent version of BLM that extends BLM with "neighbor-based interaction-profile inferring" [23]. NetLapRLS stands for "net Laplacian regularized least squares" [38], while WNN-GIP is a combination of weighted nearest neighbor and Gaussian interaction profile kernels [36].

Parameter Settings We set the number of base prediction models (N) to 25 for ALADIN.[5] Other hyperparameters of ALADIN, whenever not indicated otherwise, were learned via grid-search in internal 5-fold cross-validation on the training data. In particular: the number of nearest neighbors for the local model, ECkNN, and the number of selected features, were chosen from $\{3, 5, 7\}$ and $\{10, 20, 50\}$ respectively.

Hyperparameters of the baselines were learned similarly. In particular: for BLM-NII, the `max` function was used to generate final predictions and the weight

[5] In our initial experiments, we observed that increasing the number of base models results in asymptotically increasing performance. For example, we obtained AUPR of 0.835, 0.867 and 0.871 with 5, 25 and 100 base models on the Ion Channel dataset. We made similar observations on the other datasets both in terms of AUC and AUPR. Therefore, using $N = 25$ base models seems to be a fair compromise between runtime and prediction quality.

α for the combination of structural and collaborative similarities was chosen from $\{0.0, 0.1, \ldots, 1.0\}$. In WNN-GIP, the decay hyperparameter T was chosen from $\{0.1, 0.2, \ldots, 1.0\}$ and the weight α for combination of structural and collaborative similarities was chosen from $\{0.0, 0.1, \ldots, 1.0\}$. The hyperparameters[6] of NetLapRLS, were chosen from $\{10^{-6}, 10^{-5}, \ldots, 10^{2}\}$.

Implementation We implemented our approach, ALADIN, in Python.[7] We used the EC*k*NN implementation from the publicly available PyHubs library[8] and methods from the NumPy machine learning library for the calculation of AUC and AUPR. We used implementations of NetLapRLS, BLM-NII and WNN-GIP from the publicly available PyDTI software library.[9]

5.2 Experimental Results

Our results are shown in Fig. 4 and Fig. 5. The symbols +/− denote if the differences between the best-performing approach and other methods are statistically significant (+) or not (−).

As one can see, our approach, ALADIN outperformed its competitors, NetLapRLS, BLM-NII and WNN-GIP, on the Enzyme, Ion Channel, GPRC and Kinase datasets both in terms of AUC and AUPR. In the vast majority of the cases, the difference is statistically significant. In case of the NR dataset, the difference between ALADIN, BLM-NII and WNN-GIP is not significant. Note, however, that NR is an exceptionally small dataset, therefore, the results obtained on NR are likely to be less stable compared to other datasets.

Additionally, we examined the contribution of hubness-aware error correction: in particular, we run ALADIN with simple *k*NN regression instead of EC*k*NN. We found that *ALADIN with ECkNN* systematically outperformed *ALADIN with kNN* on all the examined datasets. The difference was statistically significant in most of the cases. In terms of AUC, we observed the largest difference on the Kinase dataset (0.93 versus 0.90), whereas in terms of AUPR, the largest difference was observed on the Enzyme dataset (0.83 versus 0.73). These results indicate that error correction is essential for accurate predictions.[10]

Furthermore, we examined how ALADIN's performance depend on k, the number of nearest neighbors in EC*k*NN. As one can see in Fig. 6, high performance is maintained for various k values and $k = 3$ seems to result in good results both in terms of AUC and AUPR.

[6] $\beta = \beta_{drug} = \beta_{target}$ and $\gamma = \gamma_{drug} = \gamma_{target}$

[7] See `https://github.com/lpeska/ALADIN` for our codes.

[8] `https://sourceforge.net/projects/pyhubs/`

[9] `https://github.com/stephenliu0423/PyDTI`

[10] These results are in accordance with our further observations: considering the input data of the local models, the skewness of the distribution of bad k-nearest neighbor occurrences (with $k = 3$), which is often used to quantify the presence of bad hubs [33], is remarkably high, between 1.61 and 11.13.

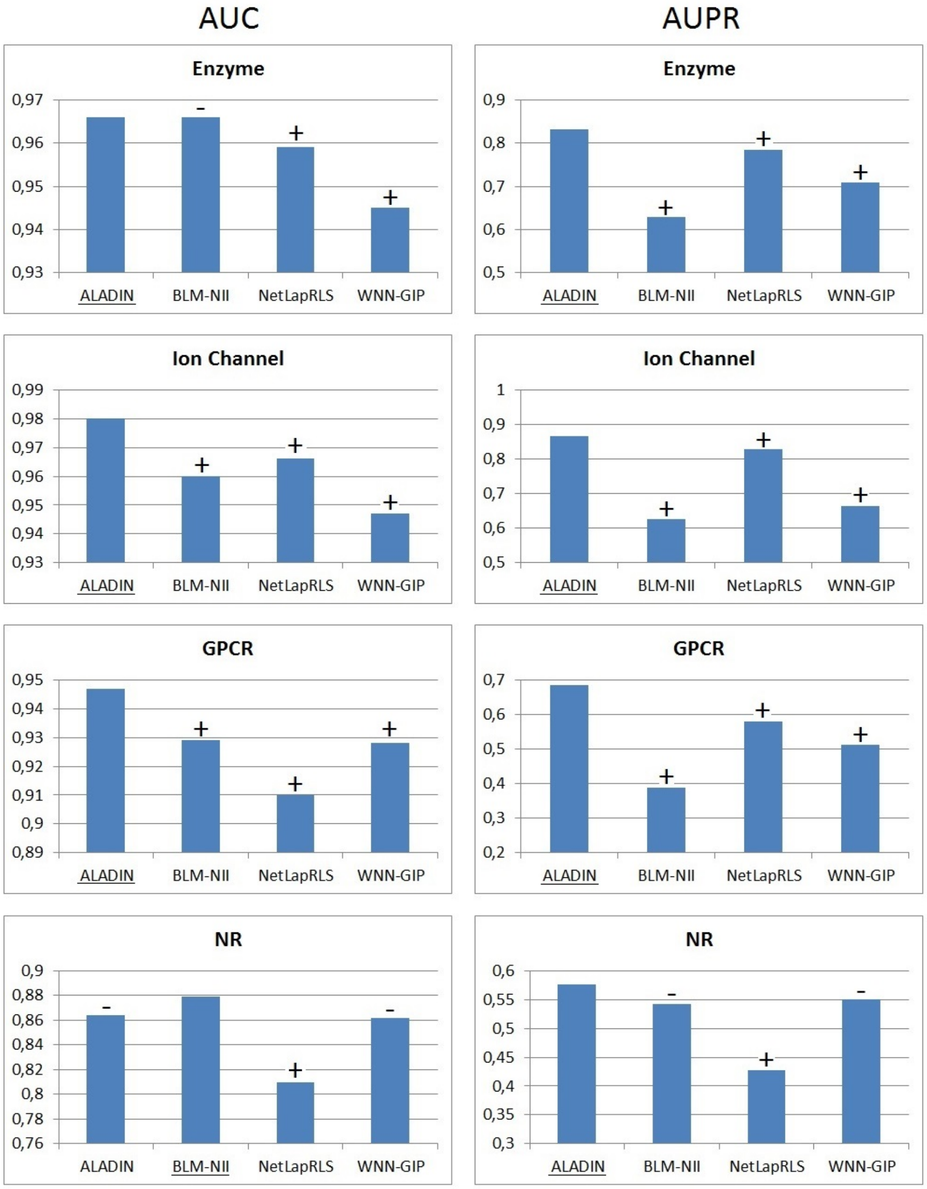


Fig. 4. Experimental results: the performance of ALADIN and its competitors in terms of AUC (left) and AUPR (right) on the Enzyme, Ion Channel, GPCR and NR datasets. The best-performing method is underlined. The symbols +/− denote if the differences between the best-performing approach and other methods are statistically significant (+) or not (−).

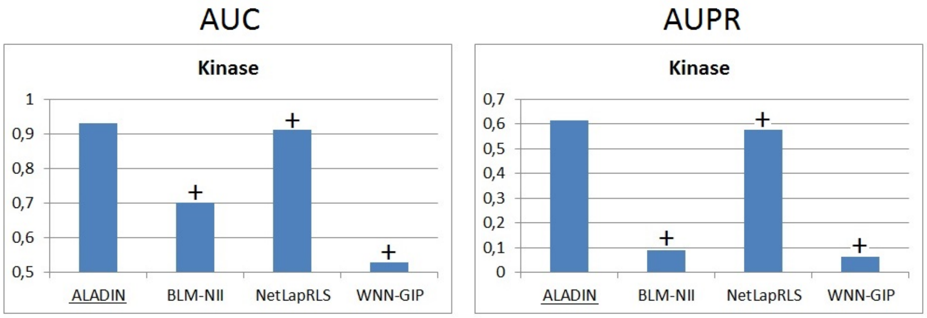


Fig. 5. Experimental results: the performance of ALADIN and its competitors in terms of AUC (left) and AUPR (right) on the Kinase dataset. The best-performing method is underlined. The symbols +/− denote if the differences between the best-performing approach and other methods are statistically significant (+) or not (−).

5.3 Application for the Prediction of New Interactions

Next, we illustrate that, besides achieving high accuracy in terms of AUC and AUPR, the predictions of ALADIN may be relevant for pharmaceutical applications as well. We begin this discussion by noting that the drug–target interactions contained in the Enzyme, Ion Channel, GPCR and NR datasets were extracted from the Kyoto Encyclopedia of Genes and Genomes[11] (KEGG) *several years ago* and, in order to allow for comparison of prediction techniques, they have been kept unchanged. However, in the mean time, additional drug–target interactions have been validated chemically and the results have been uploaded to databases, such as KEGG, DrugBank[12] or Matador[13].

Therefore, in order to demonstrate that our approach is able to predict new interactions, we trained ALADIN and its competitors, BLM-NII, NetLapRLS and WNN-GIP using all the interactions of the original datasets, and ranked the non-interacting drug–target pairs of the original datasets according to their predicted interaction scores. For simplicity, we use the term *predicted new interactions* for the top-ranked 20 drug–target pairs. We say that a predicted new interaction is *validated* if it is included in the current version of KEGG, DrugBank or Matador.

In terms of the number of validated interactions, ALADIN had the best overall performance. For example, on the Ion Channel and NR datasets, ALADIN was able to predict 12 and 8 validated interactions, whereas none of its competitors was able to predict more than 6 validated interactions on these datasets.

Most notably, numerous validated interactions were only predicted by our approach, for example, on the Enzyme dataset, the interactions between *Ibuprofen* (D00126) and *arachidonate 15-lipoxygenase* (hsa:246) and its second type

[11] `http://www.kegg.jp/`
[12] `https://www.drugbank.ca/`
[13] `http://matador.embl.de/`

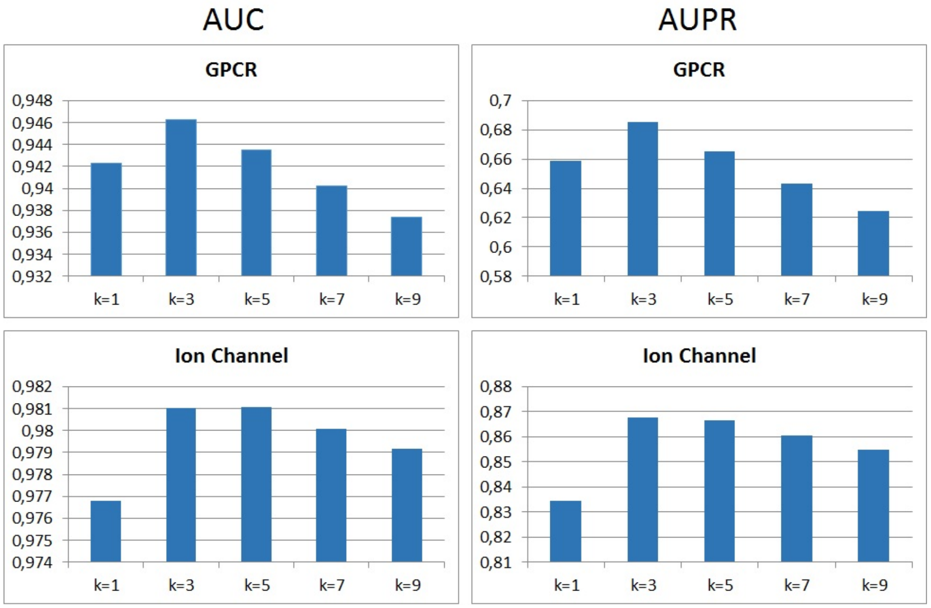


Fig. 6. ALADIN's performance in case of various k values in ECkNN.

(hsa:247); as well as the interaction between *Phentermine* (D05458) and *monoamine oxidase A* (hsa:4128); and the interaction between *Dyphylline* (D00691) and *phosphodiesterase 7A* (hsa:5150). On the GPCR dataset, only ALADIN was able to predict the validated interaction between *Theophylline sodium acetate* (D01712) and *adenosine A2b receptor* (hsa:136), as well as the interaction between *Loxapine* (D02340) and *dopamine receptor D1* (hsa:1812).

6 Conclusions and Outlook

In this paper, we considered the drug–target interaction prediction problem which has important applications in understanding the mechanisms of how drugs effect, drug repositioning and prediction of adverse effects. We proposed an extension of BLM, one of the most prominent DTI models. In particular, we proposed the ALADIN approach which represents drugs and targets in a multimodal similarity space, uses ECkNN, a hubness-aware regression approach as local model in BLM and builds a projection-based ensemble.

We performed experiments on widely-used publicly-available datasets, the results of which show that our approach is superior to BLM-NII, NetLapRLS and WNN-GIP. We also demonstrated that our approach is able to predict chemically validated new drug–target interactions.

While DTI is an essential task, we point out that ALADIN may be adapted for the prediction of interactions between other biomedical entities, such as protein–RNA interactions [32] or protein–protein interactions [2].

Furthermore, we believe that our approach may motivate new recommender systems techniques as well. Although it was shown that only a few ratings per user may be more relevant than content-based metadata [27], we argue that the continuous flow of new users causes ongoing cold start problem [20], [26] in many cases, such as small e-commerce enterprises. This indicates that hybrid prediction models incorporating both relevance feedback and metadata may be desirable. Methods like ALADIN can be applied in such domains, e.g., as a part of an alternating hybrid approach, where users with sufficient feedback receive purely collaborative recommendations.

Acknowledgment

Ladislav Peska was supported by the Charles University grant P46.

References

1. Barabási, A.L., Gulbahce, N., Loscalzo, J.: Network medicine: a network-based approach to human disease. Nature Reviews Genetics 12(1), 56–68 (2011)
2. Besemann, C., Denton, A., Yekkirala, A.: Differential association rule mining for the study of protein-protein interaction networks. In: 4th International Conference on Data Mining in Bioinformatics. pp. 72–80. Springer-Verlag (2004)
3. Biau, G., Cérou, F., Guyader, A.: On the rate of convergence of the bagged nearest neighbor estimate. The Journal of Machine Learning Research 11, 687–712 (2010)
4. Bleakley, K., Yamanishi, Y.: Supervised prediction of drug–target interactions using bipartite local models. Bioinformatics 25(18), 2397–2403 (2009)
5. Bolgar, B., Antal, P.: Bayesian matrix factorization with non-random missing data using informative gaussian process priors and soft evidences. Journal of Machine Learning Research 52, 25–36 (2016)
6. Buza, K., Nanopoulos, A., Nagy, G.: Nearest neighbor regression in the presence of bad hubs. Knowledge-Based Systems 86, 250–260 (2015)
7. Cellier, P., Charnois, T., Plantevit, M.: Sequential patterns to discover and characterise biological relations. In: International Conference on Intelligent Text Processing and Computational Linguistics. pp. 537–548. Springer (2010)
8. Chen, X., Liu, M.X., Yan, G.Y.: Drug–target interaction prediction by random walk on the heterogeneous network. Molecular BioSystems 8(7), 1970–1978 (2012)
9. Cheng, A.C., Coleman, R.G., Smyth, K.T., Cao, Q., Soulard, P., Caffrey, D.R., Salzberg, A.C., Huang, E.S.: Structure-based maximal affinity model predicts small-molecule druggability. Nature biotechnology 25(1), 71–75 (2007)
10. Cheng, F., Liu, C., Jiang, J., Lu, W., Li, W., Liu, G., Zhou, W., Huang, J., Tang, Y.: Prediction of drug-target interactions and drug repositioning via network-based inference. PLoS Comput Biol 8(5), e1002503 (2012)
11. Davis, J., Santos Costa, V., Ray, S., Page, D.: An integrated approach to feature invencion and model construction for drug activity prediction. In: Proceedings of the 24th International Conference on Machine Learning. pp. 217–224 (2007)
12. Davis, M.I., Hunt, J.P., Herrgard, S., Ciceri, P., Wodicka, L.M., Pallares, G., Hocker, M., Treiber, D.K., Zarrinkar, P.P.: Comprehensive analysis of kinase inhibitor selectivity. Nature Biotechnology 29(11), 1046–1051 (2011)

13. Fayruzov, T., De Cock, M., Cornelis, C., Hoste, V.: Linguistic feature analysis for protein interaction extraction. BMC Bioinformatics 10(1), 374 (2009)
14. Gönen, M.: Predicting drug–target interactions from chemical and genomic kernels using bayesian matrix factorization. Bioinformatics 28(18), 2304–2310 (2012)
15. Halperin, I., Ma, B., Wolfson, H., Nussinov, R.: Principles of docking: An overview of search algorithms and a guide to scoring functions. Proteins: Structure, Function, and Bioinformatics 47(4), 409–443 (2002)
16. Hattori, M., Okuno, Y., Goto, S., Kanehisa, M.: Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. Journal of the American Chemical Society 125(39), 11853–11865 (2003)
17. Hopkins, A.L.: Drug discovery: predicting promiscuity. Nature 462(7270), 167–168 (2009)
18. Hu, C., Jain, G., Zhang, P., Schmidt, C., Gomadam, P., Gorka, T.: Data-driven method based on particle swarm optimization and k-nearest neighbor regression for estimating capacity of lithium-ion battery. Applied Energy 129, 49–55 (2014)
19. Jamali, A.A., Ferdousi, R., Razzaghi, S., Li, J., Safdari, R., Ebrahimie, E.: Drug-miner: comparative analysis of machine learning algorithms for prediction of potential druggable proteins. Drug discovery today 21(5), 718–724 (2016)
20. Kaminskas, M., Bridge, D., Foping, F., Roche, D.: Product-seeded and basket-seeded recommendations for small-scale retailers. Journal on Data Semantics pp. 1–12 (2016)
21. Keiser, M.J., Roth, B.L., Armbruster, B.N., Ernsberger, P., Irwin, J.J., Shoichet, B.K.: Relating protein pharmacology by ligand chemistry. Nature biotechnology 25(2), 197–206 (2007)
22. van Laarhoven, T., Nabuurs, S.B., Marchiori, E.: Gaussian interaction profile kernels for predicting drug–target interaction. Bioinformatics 27(21), 3036–3043 (2011)
23. 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 (2013)
24. Pahikkala, T., Airola, A., Pietilä, S., Shakyawar, S., Szwajda, A., Tang, J., Aittokallio, T.: Toward more realistic drug-target interaction predictions. Briefings in Bioinformatics 16(2), 325–337 (2015)
25. Pérot, S., Regad, L., Reynès, C., Spérandio, O., Miteva, M.A., Villoutreix, B.O., Camproux, A.C.: Insights into an original pocket-ligand pair classification: a promising tool for ligand profile prediction. PloS one 8(6), e63730 (2013)
26. Peska, L., Vojtas, P.: Recommending for disloyal customers with low consumption rate. In: International Conference on Current Trends in Theory and Practice of Informatics. pp. 455–465. Springer (2014)
27. Pilászy, I., Tikk, D.: Recommending new movies: even a few ratings are more valuable than metadata. In: 3rd ACM Conf. on Recommender Systems. pp. 93–100 (2009)
28. Plantevit, M., Charnois, T., Klema, J., Rigotti, C., Crémilleux, B.: Combining sequence and itemset mining to discover named entities in biomedical texts: a new type of pattern. International Journal of Data Mining, Modelling and Management 1(2), 119–148 (2009)
29. Radovanović, M., Nanopoulos, A., Ivanović, M.: Hubs in space: Popular nearest neighbors in high-dimensional data. The Journal of Machine Learning Research 11, 2487–2531 (2010)

30. Sönströd, C., Johansson, U., Norinder, U., Boström, H.: Comprehensible models for predicting molecular interaction with heart-regulating genes. In: 7th IEEE International Conference on Machine Learning and Applications. pp. 559–564 (2008)
31. Stensbo-Smidt, K., Igel, C., Zirm, A., Pedersen, K.S.: Nearest neighbour regression outperforms model-based prediction of specific star formation rate. In: IEEE International Conference on Big Data. pp. 141–144 (2013)
32. Stražar, M., Žitnik, M., Zupan, B., Ule, J., Curk, T.: Orthogonal matrix factorization enables integrative analysis of multiple RNA binding proteins. Bioinformatics 32(10), 1527–1535 (2016)
33. Tomašev, N., Buza, K., Marussy, K., Kis, P.B.: Hubness-aware classification, instance selection and feature construction: Survey and extensions to time-series. In: Feature selection for data and pattern recognition, pp. 231–262. Springer (2015)
34. Ullrich, K., Kamp, M., Gärtner, T., Vogt, M., Wrobel, S.: Ligand-based virtual screening with co-regularised support vector regression. In: 16th IEEE International Conference on Data Mining Workshops. pp. 261–268 (2016)
35. Ullrich, K., Mack, J., Welke, P.: Ligand affinity prediction with multi-pattern kernels. In: International Conference on Discovery Science. pp. 474–489. Springer (2016)
36. van Laarhoven, Twan and Marchiori, Elena: Predicting drug-target interactions for new drug compounds using a weighted nearest neighbor profile. PloS one 8(6), e66952 (2013)
37. Wang, Y., Zeng, J.: Predicting drug-target interactions using restricted Boltzmann machines. Bioinformatics 29(13), i126–i134 (2013)
38. Xia, Z., Wu, L.Y., Zhou, X., Wong, S.T.: Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. BMC Systems Biology 4(Suppl 2), S6 (2010)
39. 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 (2008)
40. Zhang, P., Agarwal, P., Obradovic, Z.: Computational drug repositioning by ranking and integrating multiple data sources. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 579–594. Springer (2013)
41. Zheng, X., Ding, H., Mamitsuka, H., Zhu, S.: Collaborative matrix factorization with multiple similarities for predicting drug-target interactions. In: 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1025–1033 (2013)
42. Zhu, S., Okuno, Y., Tsujimoto, G., Mamitsuka, H.: A probabilistic model for mining implicit chemical compound–gene relations from literature. Bioinformatics 21(Suppl 2), ii245–ii251 (2005)