PLOS ONE

RESEARCH ARTICLE

SELF-BLM: Prediction of drug-target interactions via self-training SVM

Jongsoo Keum, Hojung Nam*

School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, 123 Cheomdangwgi-ro, Buk-gu, Gwangju, Republic of Korea

* hjnam@gist.ac.kr

Citation: Keum J, Nam H (2017) SELF-BLM: Prediction of drug-target interactions via self-training SVM. PLoS ONE 12(2): e0171839. doi:10.1371/journal.pone.0171839

Editor: Alexey Porollo, Cincinnati Children's Hospital Medical Center, UNITED STATES

Received: September 29, 2016

Accepted: January 26, 2017

Published: February 13, 2017

Data Availability Statement: The implemented software and supporting data are available at https://github.com/GIST-CSBL/SELF-BLM.

Funding: This work was supported by the Bio-Synergy Research Project (NRF-2014M3A9C4066449, http://nrf.re.kr/) of the Ministry of Science, ICT and Future Planning through the National Research Foundation, and supported by the National Research Foundation of Korea grant funded by the Korea government (MSIP) (NRF-2015R1C1A1A01051578, http://nrf.re.kr/). The funders had no role in study design,

Abstract

Predicting drug-target interactions is important for the development of novel drugs and the repositioning of drugs. To predict such interactions, there are a number of methods based on drug and target protein similarity. Although these methods, such as the bipartite local model (BLM), show promise, they often categorize unknown interactions as negative interaction. Therefore, these methods are not ideal for finding potential drug-target interactions that have not yet been validated as positive interactions. Thus, here we propose a method that integrates machine learning techniques, such as self-training support vector machine (SVM) and BLM, to develop a self-training bipartite local model (SELF-BLM) that facilitates the identification of potential interactions. The method first categorizes unlabeled interactions and negative interactions among unknown interactions using a clustering method. Then, using the BLM method and self-training SVM, the unlabeled interactions are self-trained and final local classification models are constructed. When applied to four classes of proteins that include enzymes, G-protein coupled receptors (GPCRs), ion channels, and nuclear receptors, SELF-BLM showed the best performance for predicting not only known interactions but also potential interactions in three protein classes compare to other related studies. The implemented software and supporting data are available at https://github.com/GIST-CSBL/SELF-BLM.

Introduction

In recent years, interest in identifying drug-target interactions has dramatically increased not only for drug development but also for understanding the mechanisms of action of various drugs. However, time and cost requirements associated with experimental verification of drug-target interactions cannot be disregarded. Many drug databases, such as DrugBank, KEGG BRITE, and SuperTarget, contain information about relatively few experimentally identified drug-target interactions [1–3]. Therefore, other approaches for identifying drug-target interactions are needed to reduce the time and cost of drug development. In this regard, *in silico* methods for predicting drug-target interactions can provide important information for drug development in a reasonable amount of time.

data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Various *in silico* screening methods have been developed to predict drug-target interactions. Among these methods, machine learning-based approaches such as bipartite local model (BLM) and MI-DRAGON which utilize support vector machine (SVM), random forest and artificial neural network (ANN) as part of their prediction model are widely used because of their sufficient performance and the ability to use large-scale drug-target data [4–9]. For these reasons, many machine learning based prediction tools and web-servers have been developed [10–13]. Especially, similarity-based machine learning methods which assume that similar drugs are likely to target similar proteins, have shown promising results [8, 9]. Although molecular docking methods also showed very good predictive performance, very few 3D structures of proteins are known, rendering docking methods unsuitable for large-scale screening [14, 15]. As such, a precise similarity-based method must be developed to predict interactions on a large-scale using the low-level features of compounds and proteins.

Previous similarity-based methods, such as the bipartite local model (BLM), Gaussian interaction profile (GIP), and kernelized Bayesian matrix factorization with twin kernel (KBMF2K), provide efficient ways to predict drug-target interactions and have shown very good performance [4, 16, 17]. BLM, which uses a supervised learning approach, has recently shown promising results using only similarities from each compound and each protein in the form of a kernel function. In the BLM method, the model for a protein of interest (POI) or compound of interest (COI) is learned from local information, which means that the model uses its own interactions of the COI or POI. This local-approach concept has been used in other methods, such as GIP, BLM-NII and others [17, 18].

Although such methods show very good performance, certain problems remain. Most previously developed methods categorize validated interactions between drugs and target proteins as positive, while unknown interactions are categorized as negative when constructing a predictive model. However, unknown interactions are not truly negative interactions, as they include potential interactions that have not yet been validated as positive interactions. To address this problem, Xia *et al.* developed a semi-supervised learning method (LapRLS) that regards known interactions as positive and unknown interactions as unlabeled data [19]. Chen *et al.* developed an algorithm using a network-based random walk with restart approach (RWRH) [20]. However, these methods demonstrate good performance in a limited set of conditions, where the drugs or targets use a drug-target network-based similarity score (NetLapRLS and NRWRH). Because these approaches are limited in predicting the interactions of novel compounds or proteins that do not have any known target or drug information (e.g., newly synthesized compounds or mutated protein sequences), other approaches are needed.

In this paper, we propose a drug-target interaction prediction method to predict potential interactions by using a modified BLM method. To classify unknown interactions into negative and unlabeled data, a clustering method was used before the training step [21]. Then, modified bipartite local models, termed self-training bipartite local models (SELF-BLMs), were constructed using a semi-supervised learning approach (self-training SVM) to improve a model's ability to find potential interactions [22]. Fig 1 shows the overall process of the method. Finally, to train the model, we used a previous dataset for humans involving enzymes, G-protein coupled receptors (GPCRs), ion channels, and nuclear receptors from previous studies [23]. We then constructed another drug-target interaction data set that contained recently updated interaction information for performance validation. As a result, the number of drug-target interactions increased by approximately 60% for each type of protein. Our model showed good performance based on the area under the ROC curve (AUC) and the area under the precision-recall curve (AUPR) values of the updated dataset. In addition, our proposed method found the highest number of potential drug-target interactions compared to other related methods in most cases.

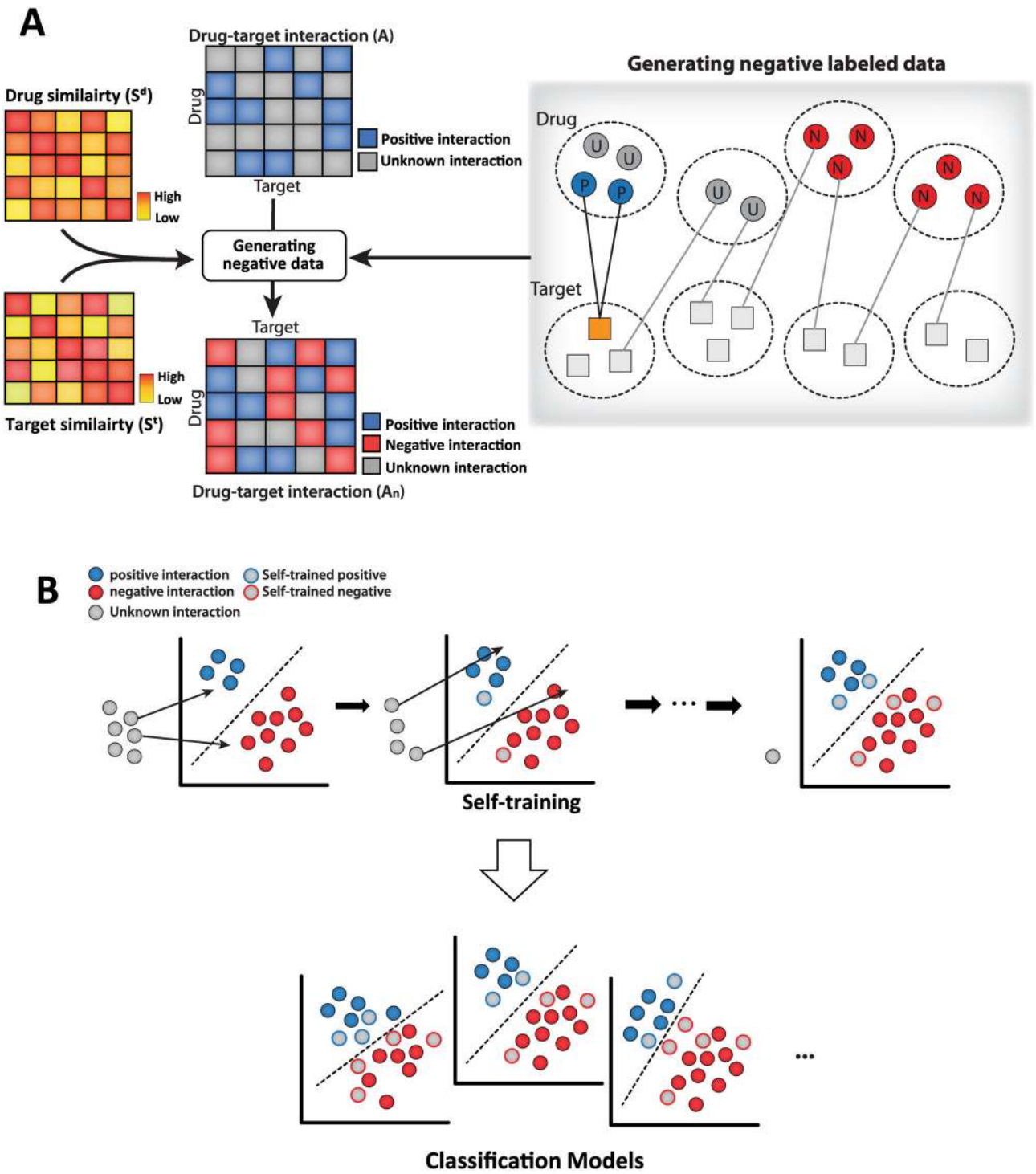

Fig 1. Overview of the proposed method. (A) From known information, drug-target interactions are classified into positive and unknown interactions (matrix A). Using similarity scores of drugs (matrix S^d) and targets (matrix S^t), we performed k-medoids clustering. If any of the drugs in a cluster do not interact with the cluster of the target protein, we considered the drugs in the cluster as having a negative interaction with the protein. Finally, drug-target interactions are classified into positive, negative and unknown interactions (matrix A_n). Yellow rectangle: target protein, blue circle: drugs having positive interactions with the target protein, red circle: drugs having negative interactions with the target protein, gray circle: drugs having unknown interactions with the target protein. **(B)** A self-training SVM repeatedly trains the unlabeled data (unknown) as positive or negative. Finally, local classification models that can find potential interactions are constructed.

doi:10.1371/journal.pone.0171839.g001

Materials and Methods

Drug-target interaction dataset for training

To train the model and cross-validate its performance, we used four types of drug-target datasets from humans, including enzymes, ion channels, GPCRs and nuclear receptors [23]. The data about the drugs, target proteins and drug-target interactions were derived from the KEGG BRITE, BRENDA, SuperTarget, and DrugBank databases [1–3, 24]. Table 1 shows details about the dataset information that was used.

Drug-target interaction dataset for validation

Because the previous dataset was constructed in 2007, many newly identified drug-target interactions have since been discovered. To validate the performance power of predicting potential drug-target interactions, we updated newly identified interactions among drugs and target proteins that belonged to the previous dataset using the DrugBank, KEGG BRITE, and DsigDB databases [1, 2, 25]. The drug-target interactions obtained from DrugBank and KEGG BRITE databases were credible [1, 2], but the DsigDB database provided manually curated data and text mining data [25]. Because text mining data are massive and not credible, we selectively took manually curated data from the DsigDB database [25]. For this update, the numbers of updated interactions for each interaction type were 4,449, 2,029, 1,268, and 168, respectively. The number of drug-target interactions increased by approximately 60% for each type of protein. Using the updated dataset, we compared the performance and potential identification capability of each method. Table 1 shows a summary of the previous and updated dataset.

Similarity metrics

The chemical similarities between drugs were calculated with the SIMCOMP method [26], which computes a global similarity score on the basis of common substructures between drugs using a graph alignment algorithm with the Eq (1)

$$S_d(d, d') = \frac{|d \cap d'|}{|d \cup d'|} \tag{1}$$

where d and d' are substructures of drugs

The structural information for the drugs was taken from the KEGG DRUG and KEGG COMPOUND sections of the KEGG LIGAND database [2].

The similarity between the proteins was calculated using a normalized version of the Smith-Waterman alignment score [27]. The normalized Smith-Waterman score between the proteins P_A and P_B was computed by the Eq (2)

$$S_p(P_A, P_B) = \frac{SW(P_A, P_B)}{\sqrt{SW(P_A, P_A)} \times \sqrt{SW(P_B, P_B)}} \tag{2}$$

where SW is the Smith-Waterman alignment score

Table 1. The number of drugs, target proteins, interactions and updated interactions of each type.

	Enzyme	Ion channels	GPCRs	Nuclear receptors
No. of drugs	445	210	223	54
No. of target proteins	664	204	95	26
No. of drug-target interactions (previous)	2,926	1,476	635	90
No. of drug-target interactions (updated)	4,449	2,029	1,268	168

doi:10.1371/journal.pone.0171839.t001

The amino acid sequences of the target proteins were derived from the KEGG GENES database [2].

Generating negative interactions

To categorize unknown interactions as negative or unlabeled interactions, first, for each target protein, if a compound interacted with the target protein, we considered the interaction to be positive. We then clustered drugs and proteins by means of k-medoids clustering [21]. If any of the compounds in a cluster do not interact with the cluster of the target protein, we considered the compounds in the cluster as having a negative interaction with the proteins (Fig 1A). The remaining unknown interactions were considered to be unlabeled interactions, which are potentially positive interactions. These unlabeled interactions may later be classified as negative or positive interactions using the semi-supervised learning method (Fig 1B).

Because we used a k-medoids clustering method, an appropriate and consistent number of clusters was needed to train various datasets. In this study, we allowed to find one or two new positive interactions for each known positive interaction. Therefore, we set the number of unlabeled interactions to be no more than double the number of positive interactions. For example, if a protein has two known positive interactions, we set the maximum number of unlabeled interactions for the protein as four. The reason why we set the stringent limit for the number of unlabeled interactions is that too many unlabeled interactions could generate a decreased number of negative interactions, thereby resulting in a loss of negative data information for model construction. Therefore, we defined the number of clusters of drugs and targets as the resulting number when the overall number was divided by an integer, and we calculated the ratio of unlabeled interactions to positive interactions for the following integers N (one to ten). Table 2 shows that the ratio was between one and two when the number of clusters was the number of drug and target proteins divided by two for each protein type. Therefore, we finally set k to be the number of drugs and target proteins divided by two. The detail steps of generating negative interactions are described in Algorithm 1

Bipartite local model

Bleakley *et al*. proposed a method called BLM to predict the interaction between a drug *i* and a target *j* [4]. BLM is described as follows. First, a local model for drug *i* is trained using an interaction profile of drug *i* and a similarity matrix of target proteins. Known interactions are regarded as positive, and unknown interactions are regarded as negative. Next, SVM constructs a classifier that distinguishes known interactions (positive) from unknown interactions (negative) using target similarity as a kernel. The model predicts the probability p_d (i,j) that a drug *i* and a query target *j* have an interaction by using the similarities between target *j* and the trained targets. Similarly, a local model for target *j* is trained using an interaction profile of target *j* and drug similarity. The model predicts the probability p_t (i,j) that a target *j*

Table 2. The number of drugs, target proteins, interactions and updated interactions of each type.

N	1	2	3	4	5	6	7	8	9	10
Enzymes	0	1.3	4.1	7.7	12.6	17.8	20.3	24	28.6	30.9
Ion channels	0	1.6	2.9	4	6.4	7.5	10.7	11.8	13.8	15.9
GPCRs	0	1.9	3.8	5.9	8	9.9	11.6	14.2	16.1	17.2
Nuclear receptors	0	1.5	3.7	5	7.6	9.4	9.6	11.9	11.8	12.7

doi:10.1371/journal.pone.0171839.t002

and a query drug *i* will have an interaction using the similarities between drug *i* and training drugs. Finally, we determine the predicted interaction value P(i,j) between drug *i* and target *j* with max(p_d (i,j), p_t (i,j)) or 0.5(p_d (i,j) + p_t (i,j)).

Algorithm 1: Generating negative interactions

```
Generating negative interactions (A, S^d, S^t);
 Input  : Drug-target interaction matrix A,
          Drug similarity matrix S^d,
          Target similarity matrix S^t
 Output: Negative labeled drug-target interaction matrix A_n
k_d := |D|) / 2;              // D: set of drugs, k_d: the number of drug cluster
k_t := |T| / 2;           // T: set of targets, k_t: the number of target cluster
C_d := k-medodids(k_d, S^d);                    //the set of drug clusters C_d
C_t := k-medodids(k_t, S^t);                  //the set of target clusters C_t
for i ← 1 to |D| do
  for j ← 1 to |T| do
    if A(i, j) = 1 then
      A_n(i, j) := 1;                                              //positive
    else
      SD_di := set of drugs in the cluster containing d_i;
      ST_tj := set of targets in the cluster containing t_j;
      if SC_t is not related SC_d then
        A_n(i, j) := -1;                                           //negative
      else
        A_n(i, j) := 0;                                           //unlabeled
      end
    end
  end
end
return A_n;
```

Self-training support vector machine

To classify the unlabeled data, a self-training SVM was used [22]. In a local prediction step, the SVM model was constructed as a BLM using only labeled data. The unlabeled data were then classified by this model. If the unlabeled data passed the threshold, the unlabeled data were classified as positive or negative. The next step was to iterate this process until no unlabeled data failed to pass the threshold. Finally, the model used all labeled data as a local classification model to predict whether a compound targets proteins of interest and whether a protein is targeted by a compound of interest. The detail steps of constructing SELF-BLM models M_t for prediction of drug are described in Algorithm 2. In similar manner, SELF-BLM models M_d for prediction of target proteins are constructed.

Algorithm 2: SELF-BLM

```
SELF-BLM (A, S^d, S^t);
 Input  : Drug-target interaction matrix A,
          Drug similarity matrix S^d,
          Target similarity matrix S^t
 Output: prediction model of target M_t
A_n := Generating negative interactions ((A, S^d, S^t));
I^t(i) := A_n(:, i);                          //A interaction vector of target t_i
set I^t_L(i);                               //interaction vector of labeled data
set I^t_U(i);                             //interaction vector of unlabeled data
set S^d_L;                                  //Similarity matrix of labeled data
M_t := train(S^d_L, I^t_L(i));                          //Train a local model for t_i
do
```

```
  set S_U^d XL;            //Similarity matrix of unlabeled data by labeled data
  P_U := test(M_t, S_U^d XL);                    //Predict unlabeled interactions
  if |P_U| > threshhold then
    change the unlabeled data to labeled data
    set I_L^t(i);
    set I_U^t(i);
    set S_L^d;
    M_t := train(S_L^d, I_L^t(i))
  end
while any unlabeled data is changed;
return M_t;
```

Results

We trained the model using a previous dataset constructed by Yamanishi *et al* and validated the model using a previous dataset and an updated dataset [23]. Because some unknown interactions in the previous dataset turned out to be positive in the updated dataset, we can measure the potential identification capability of models by comparing the performance results.

First, we compared the performance of SELF-BLM with that of BLM [4], BLM-RBF, which includes drug-target network-based similarity using an RBF kernel, such as GIP or BLM-NII, and semi-supervised learning approaches, such as LapRLS, and NetLapRLS, which include network-based similarity [17–19]. For BLM and BLM-RBF, we used the modified source code that was originally given by the authors. We used the LIBSVM (v.3.21) to use SVM implementation [28]. When implementing SVM, the similarity matrices were used as a kernel without any modification. For parameters of SVM, values of C and gamma were assigned as 1 and 1 over number of features, respectively. For LapRLS and NetLapRLS, we implemented the methods based on the original paper. In the papers reporting these methods, BLM takes the maximum value between a drug-predicted value calculated using drug similarity and target predicted value calculated using target similarity, whereas LapRLS and NetLapRLS take an average value between the drug-predicted value and the target-predicted value; hence, we followed such approaches when we implemented these methods in the present study. SELF-BLM also takes the maximum value between the drug-predicted value and the target-predicted value.

Because the all compared models are local models, the models are repeatedly constructed using associated interactions for a given drug or protein. If the methods are evaluated in k-fold cross-validation, positive interactions are frequently not included in the training step. For example, in case of *epinephrine* drug, the drug has three positive interactions with 95 target proteins in the GPCRs dataset. Because of the small number of positive interactions, positive labels are often not included in the training set when the data is segmented into k-sets. Thus, we evaluated the performance of the models using leave-one-out cross-validation (LOOCV). However, in order to confirm robustness of our model, we also evaluated the performance using 10-fold cross-validation (S1 Table).

Prediction performance

We calculated the performance of the interaction prediction in terms of the area under the ROC curve (AUC) value and the area under the precision-recall curve (AUPR) value. The AUC value is a common evaluation approach for binary classification problems. However, the large bias between the negative and positive training data sets often weakens the power of AUC values. Meanwhile, because it is important to classify the positive labels with high accuracy, the AUPR value may be a more appropriate indicator than the AUC value.

Table 3. The AUC and AUPR values of the five methods for the four types of proteins in each validation set (previous and updated dataset).

		Enzymes		Ion channels		GPCRs		Nuclear receptors	
		Previous	Updated	Previous	Updated	Previous	Updated	Previous	Updated
SELF-BLM	AUC	**0.974**	0.859	**0.977**	**0.941**	**0.952**	0.914	0.890	**0.799**
BLM		0.968	0.846	0.972	0.923	0.94	0.893	0.869	0.767
BLM-RBF		0.974	0.880	0.975	0.903	0.930	0.880	**0.909**	0.792
LapRLS		0.954	**0.883**	0.960	0.915	0.894	0.872	0.816	0.778
NetLapRLS		0.960	0.869	0.958	0.928	0.926	**0.917**	0.867	0.789
SELF-BLM	AUPR	0.846	0.637	0.805	**0.762**	0.566	**0.614**	**0.625**	**0.573**
BLM		0.862	0.629	0.842	0.745	0.676	0.610	0.599	0.534
BLM-RBF		**0.891**	**0.652**	**0.922**	0.758	**0.709**	0.590	0.609	0.514
LapRLS		0.704	0.538	0.744	0.658	0.400	0.401	0.387	0.452
NetLapRLS		0.806	0.609	0.827	0.735	0.637	0.596	0.456	0.458

doi:10.1371/journal.pone.0171839.t003

Table 3 shows the AUC and AUPR values of the five methods for the four type of proteins in each data set (previous and updated datasets). As the results show, the AUPR values of BLM-RBF were high in most cases when we used the previous dataset for validation. However, with the updated dataset, the AUC and AUPR values of SELF-BLM were the highest for most protein types, except for enzymes (Fig 2, S1 Fig). In Table 3, it is noticeable that the AUC and AUPR values tend to be decreased in the updated data. The main reason for this result is that some negatively labeled interactions changed into positive interactions when the dataset was updated. Therefore, there are previously predicted a fair number of interactions as negative, the AUC and AUPR values decreased in the updated data. For instance, to predict the interaction between target HTR1E and drug Olanzapine according to type of GPCR, HTR1E was considered similar to HTR2A (0.23) and HTR2C (0.23), which bind to Olanzapine (positive); however, HTR1E is more similar to HTR1B (0.43), HTR1D (0.44), and HTR1F (0.55), which do not bind to Olanzapine (negative) in the training dataset. Thus, BLM does not receive a high indication that HTR1E will bind to Olanzapine. On the other hand, with the SELF-BLM methods, these negative targets were regarded as potential targets, and some targets were considered unlabeled as a result. Thus, SELF-BLM yields high marks using unlabeled data generated by clustering and the self-training SVM method (Fig 3). Moreover, in the case of the previous dataset, because the interaction between HTR1E and Olanzapine is regarded as negative, SELF-BLM seems incorrectly predicting the interaction. This is the main reason why SELF-BLM shows decreased performance in some cases using the previous dataset. However, in the updated dataset, the interaction is now regarded as positive, and the performance of SELF-BLM thus increased.

In addition, SELF-BLM could increase the prediction performance with the previous dataset by self-training unknown information. Because potential interactions are regarded as negative in the previous dataset, this approach makes it difficult for a model to be trained accurately. For example, in the case of predicting the positive interaction between target CHRM1 and drug Clozapine among GPCRs, the conditions are as follows. Target CHRM2 binds to Clozapine, and CHRM3, CHRM4, and CHRM5 do not bind to Clozapine (however, these targets actually do bind to Clozapine in the updated dataset). In similarity-based models, the CHRM1 model will choose a similar protein among targets. BLM does not indicate that CHRM1 will bind to Clozapine as CHRM1 is more similar to CHRM3 (0.45), CHRM4 (0.42), and CHRM5 (0.47) than to CHRM2 (0.42). In contrast, because SELF-BLM neither considers CHRM3, CHRM4, and CHRM5 as training data nor changes these targets to positive data beforehand, it predicts that CHRM1 will bind to Clozapine (S2 Fig). Therefore,

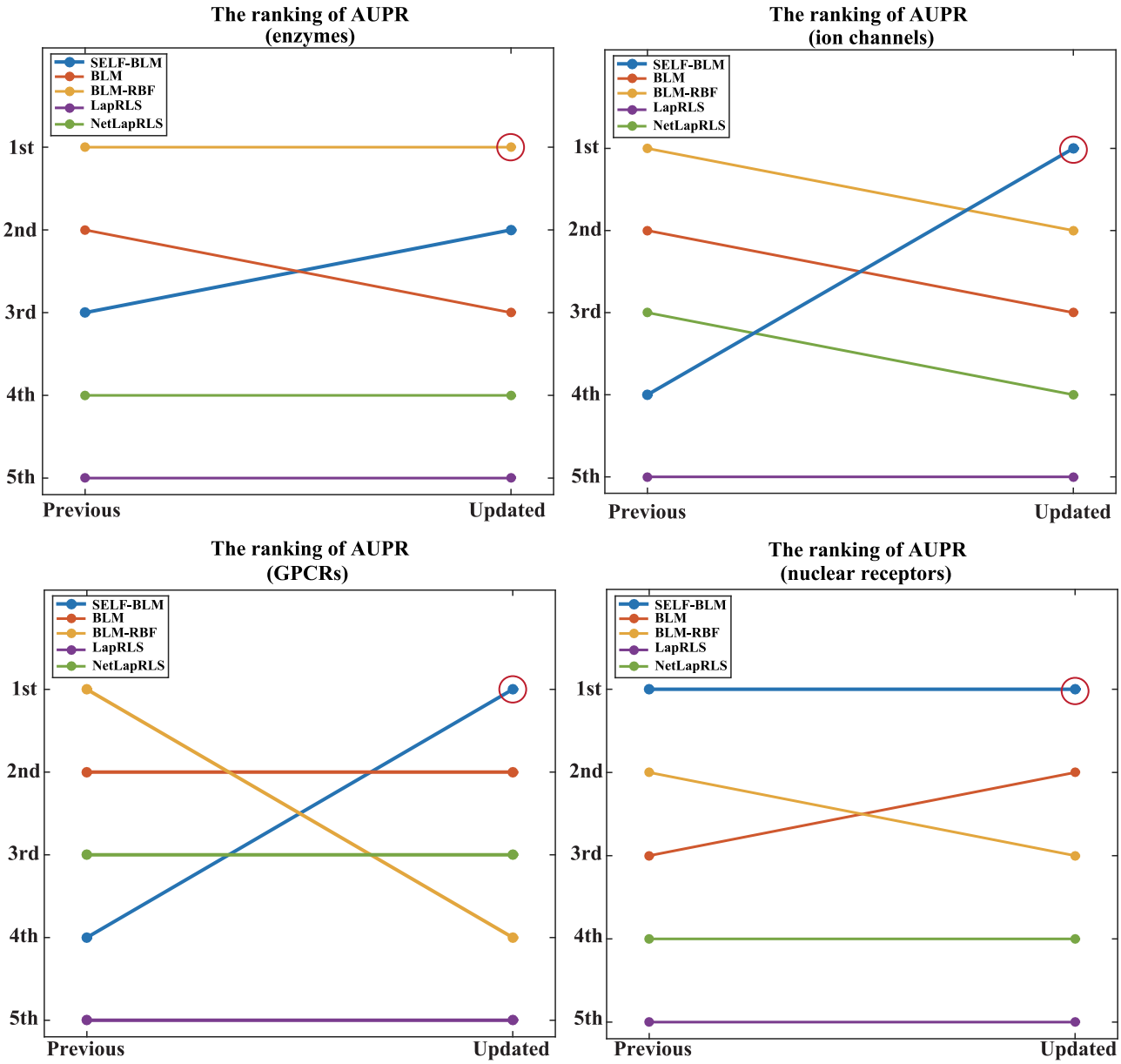

Fig 2. Rankings of AUPR trends by the different methods according to the updated dataset. In each panel, y-axis shows the rank representation of the AUPR value. A) the ranking in type of enzymes, B) the ranking in type of ion channels, C) the ranking in type of GPCRs, D) the ranking in type of nuclear receptor.

doi:10.1371/journal.pone.0171839.g002

SELF-BLM can yield high performance not only for the updated dataset but also for the previous validation dataset. Furthermore, additional experiment was conducted using up-to-dated drug-target information to show that the results are consistent in other dataset (see S1 File).

Prediction performance for new interactions

Next, we evaluated the performance of models regarding potential interaction identification. We compared the number of potential interactions at each percentage of positive interactions from the top 1% to 100% of the ranked score. For example, the targets of GPCRs have 635

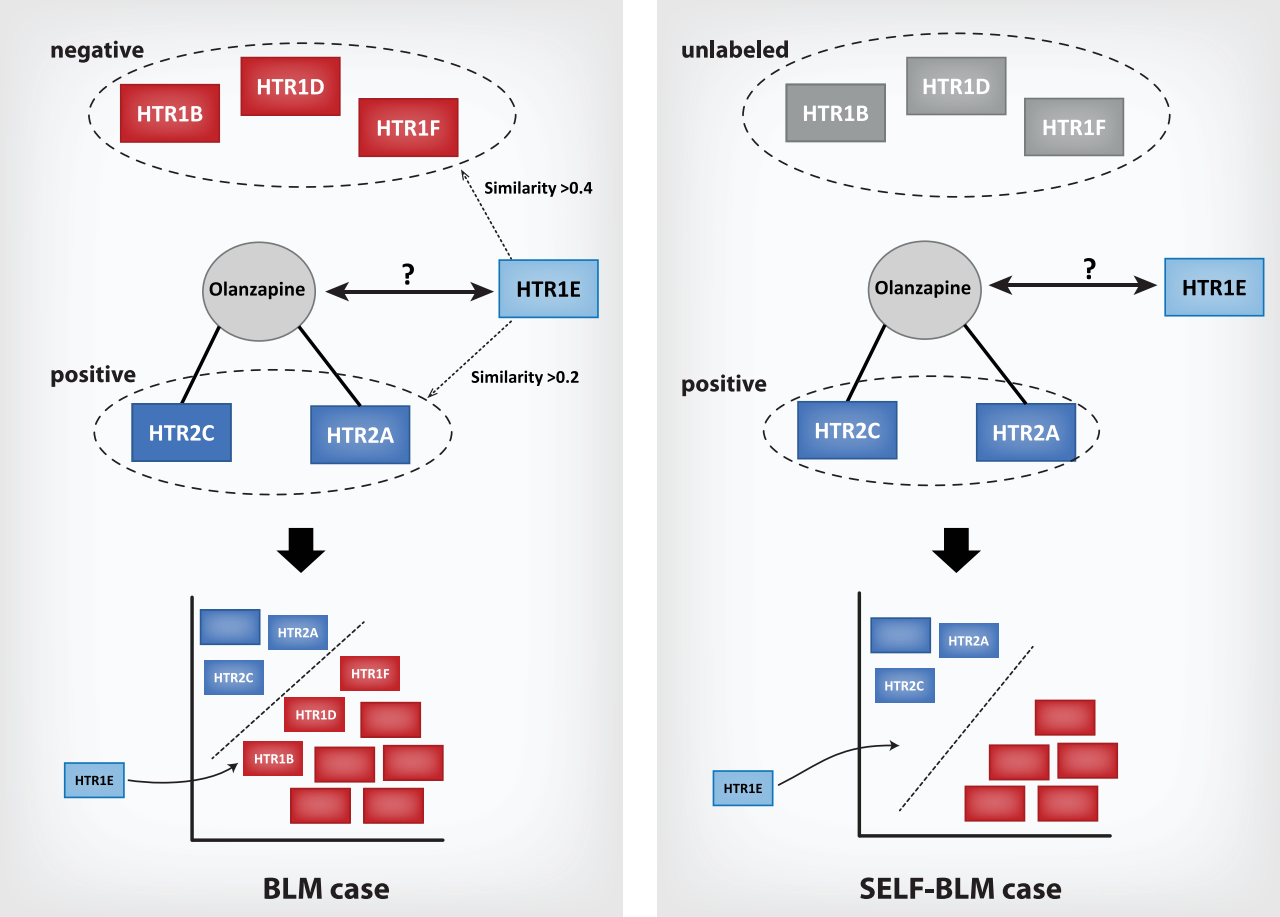

Fig 3. An example of SELF-BLM predicting the targets of a drug. In the previous dataset, it was known that proteins (HTR2A and HTR2C) bind to a drug (Olanzapine), but it was not known that other proteins (HTR1B, HTR1D, and HTR1F) also bind to the drug. Thus, in BLM, HTR2A and HTR2C are labeled as positive, and HTR1B, HTR1D and HTR1F are labeled as negative. Because the protein (HTR1E) is more similar to negatively labeled proteins than to positively labeled proteins, the protein is predicted to be negative. However, in SELF-BLM, these proteins (HTR1B, HTR1D, and HTR1F) are unlabeled. Therefore, the protein (HTR1E) is predicted as positive. There was no information suggesting that the protein (HTR1E) binds to the drug (Olanzapine) in the previous data, but it was later revealed that the protein indeed binds to the drug.

doi:10.1371/journal.pone.0171839.g003

known interactions, so we set the positive as the top six (1%) to 635 (100%) from a total of 21,185 interactions, and the number of potential interactions were compared within the percentage range. As shown in Fig 4, SELF-BLM finds the most number of potential interactions than other methods for all of the protein types, except for nuclear receptors (see S2 File for details).

Furthermore, we calculated the potential AUPRs of the four methods for the four types of proteins. In the potential precision-recall curve, positive labels were the potential interactions that were identified in the updated dataset, and negative labels were unknown interactions in the updated dataset. Therefore, we confirmed how the methods found the potential interactions simply by drawing a plot of the potential precision-recall curve (S3 Fig). The curves show that SELF-BLM finds many potential interactions with high accuracy. Thus, the AUPR of SELF-BLM was the greatest among the methods for all of the protein types, except for the nuclear receptor type. Fig 5 shows the potential AUPRs of the five methods for the four types of proteins.

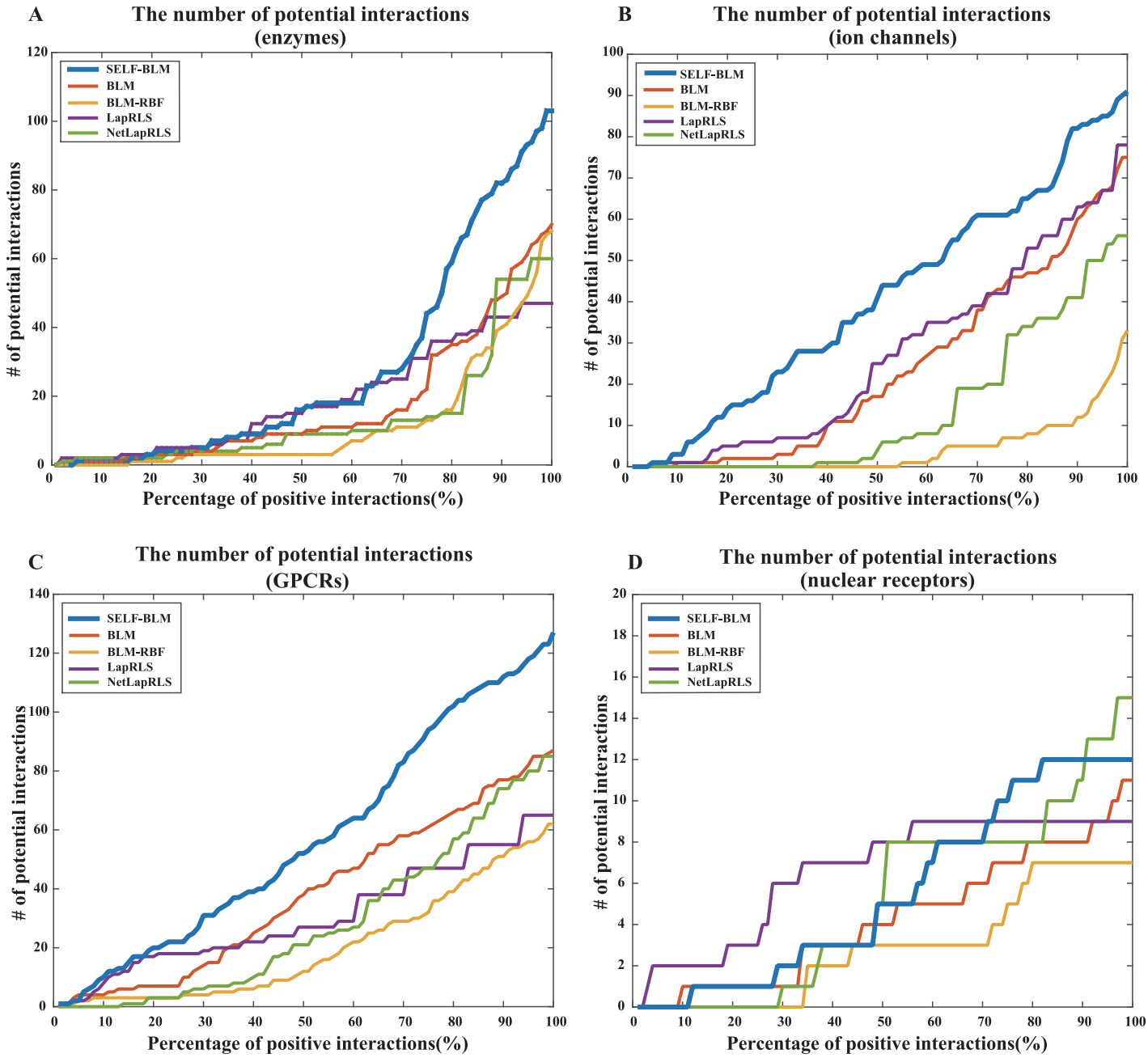

Fig 4. The number of potential interactions found by each method. X-axis represents the accumulated percentage of positively predicted interactions in each method, y-axis represents the number of correctly predicted potential interactions. A) The number of potential interactions according to type of enzyme. B) The number of potential interactions according to type of ion channel. C) The number of potential interactions according to type of GPCR. D) The number of potential interactions according to type of nuclear receptor.

doi:10.1371/journal.pone.0171839.g004

In our results, BLM-RBF found few potential interactions and had low values of potential AUPR; also, the performance of BLM-RBF showed a greater drop than the other methods in most cases. Because BLM-RBF uses network-based similarities as an important factor for identifying a drug-target interaction, if a COI or POI had few interactions with the training set, the interaction similarities made it difficult to predict potential interactions of a COI or

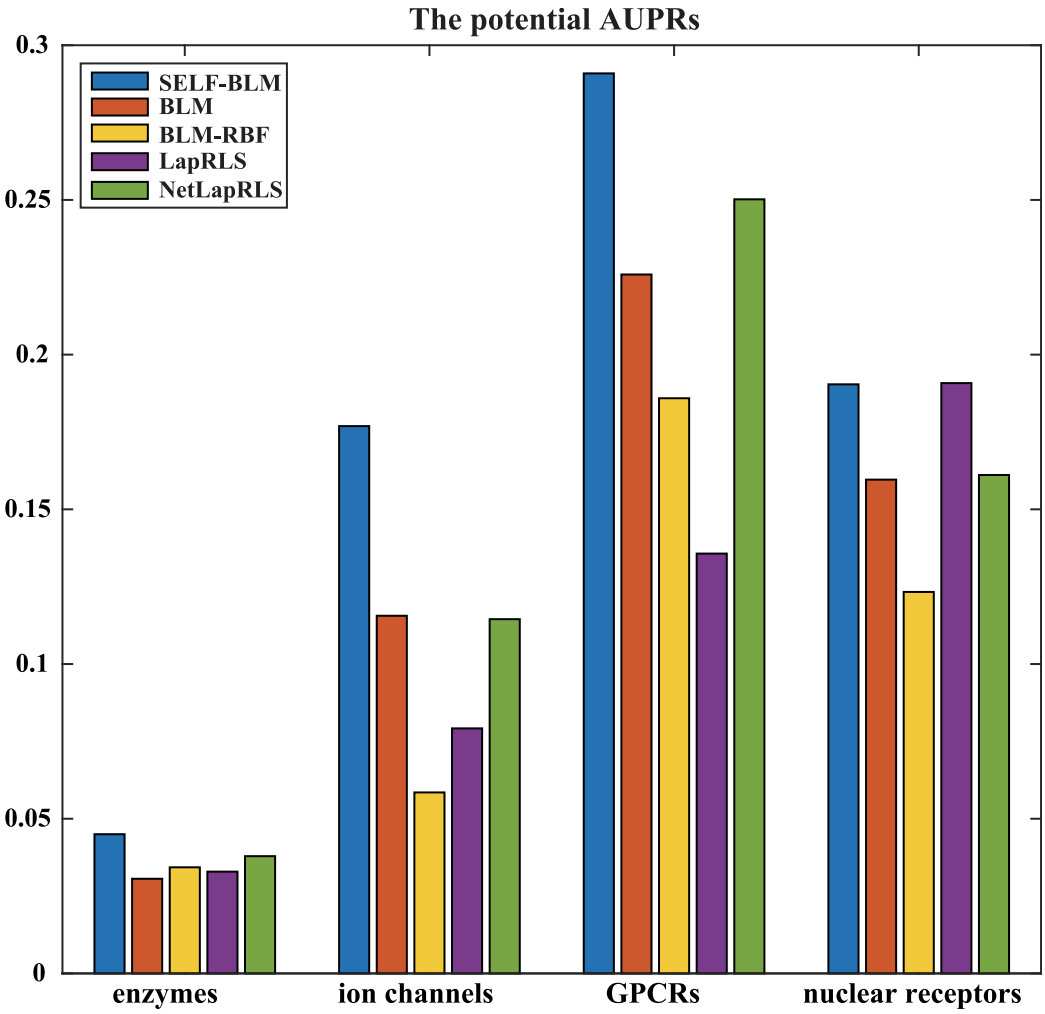

Fig 5. The potential AUPRs of the five methods for the four types of proteins.

doi:10.1371/journal.pone.0171839.g005

POI. This result shows that network-based similarity helps to find the interaction of a COI or POI that has a large amount of interaction information, but it is unsuitable for finding interactions of compounds or proteins for which little information about interactions is available. Although LapRLS and NetLapRLS are semi-supervised learning methods, we can confirm that these methods do not show good performance or a strong ability to identify potential interactions.

Conclusion

In this study, we proposed a modified BLM, termed SELF-BLM, to accurately predict potential drug-target interactions. SELF-BLM uses k-medoids clustering and a self-training SVM algorithm to identify potential interactions among unknown interactions. To validate the performance of the method, we used benchmark datasets and updated recently verified interactions as potential interactions to the dataset using the DrugBank, KEGG, and DsigDB databases. Finally, we demonstrated the capability of SELF-BLM to predict potential interactions between drugs and target proteins. Notably, in most cases, SELF-BLM showed best validation

performance with respect to AUC and AUPR for the updated dataset and found more potential interactions with high confidence prediction score compared to other methods.

In our study, we used a benchmark dataset for training to compare SELF-BLM with other methods and to validate its capability to identify interactions. However, as the research proceeded, various other similarity methods were developed. Like other similarity based-methods, SELF-BLM majorly depends on drug similarity and target similarity. Therefore, the performance of the model may be improved by using more-effective similarity methods such as kernel fusion method for various data fusion and/or efficient novel similarity features [29–31]. We emphasize that our SELF-BLM could show the best performance in the field of novel drugs or novel targets identification researches because our method does not require any known drug-target interaction information that is hardly known in novel molecules. Furthermore, in addition to drug-target protein interaction, it is important to deal with data imbalance problems or unlabeled data in many other areas so that, our method as well as the methods used in these areas can help to deal the problems [32, 33].

Supporting information

S1 Fig. Ranking of AUC trend according to the updated dataset among the methods. In each panel, y-axis shows the rank representation of the AUC value. A) the ranking in type of enzymes, B) the ranking in type of ion channels, C) the ranking in type of GPCRs, D) the ranking in type of nuclear receptors.
(EPS)

S2 Fig. An example of SELF-BLM predicting the targets of a drug. In the previous dataset, it was known that a protein (CHRM2) bind to a drug (Clozapine), but it was not known that other proteins (CHRM3, CHRM4, and CHRM5) also bind to the drug. Thus, in BLM, CHRM2 is labeled as positive, and CHRM3, CHRM4, and CHRM5 are labeled as negative. Because the protein (CHRM1) is more similar to negatively labeled proteins than to positively labeled proteins, a predicted score of the protein is not high. However, in SELF-BLM, these proteins (CHRM3, CHRM4, and CHRM5) are unlabeled, therefore, the protein (CHRM1) is predicted as positive. In this process, SELF-BLM finds positive interactions confidently.
(EPS)

S3 Fig. The potential precision-recall curve of the five methods for the four types of proteins.
(EPS)

S1 Table. The AUC and AUPR values of the five methods for the four types of proteins in each validation set (previous and updated dataset) using 10-fold cross-validation.
(DOCX)

S1 File. Additional experiments with up-to-dated drug-target interaction dataset.
(PDF)

S2 File. The number of potential interactions which are found by each method.
(XLSX)

Acknowledgments

Authors thank the lab members for their valuable feedback and comments of the study.

Author Contributions

Conceptualization: HN JK.

Data curation: JK.

Formal analysis: JK.

Funding acquisition: HN.

Investigation: JK.

Methodology: JK.

Project administration: HN.

Resources: HN.

Software: JK.

Supervision: HN.

Validation: HN JK.

Visualization: HN JK.

Writing – original draft: HN JK.

Writing – review & editing: HN JK.

References

1. Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, et al. DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Res. 2014; 42(Database issue):D1091–7. doi: 10.1093/nar/gkt1068 PMID: 24203711
2. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, et al. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res. 2006; 34(Database issue):D354–7. doi: 10.1093/nar/gkj102 PMID: 16381885
3. Gunther S, Kuhn M, Dunkel M, Campillos M, Senger C, Petsalaki E, et al. SuperTarget and Matador: resources for exploring drug-target relationships. Nucleic Acids Res. 2008; 36(Database issue):D919–22. doi: 10.1093/nar/gkm862 PMID: 17942422
4. Bleakley K, Yamanishi Y. Supervised prediction of drug-target interactions using bipartite local models. Bioinformatics. 2009; 25(18):2397–403. doi: 10.1093/bioinformatics/btp433 PMID: 19605421
5. Prado-Prado F, Garcia-Mera X, Escobar M, Alonso N, Caamano O, Yanez M, et al. 3D MI-DRAGON: new model for the reconstruction of US FDA drug- target network and theoretical-experimental studies of inhibitors of rasagiline derivatives for AChE. Curr Top Med Chem. 2012; 12(16):1843–65. doi: 10.2174/156802612803989228 PMID: 23030618
6. Prado-Prado F, Garcia-Mera X, Escobar M, Sobarzo-Sanchez E, Yanez M, Riera-Fernandez P, et al. 2D MI-DRAGON: a new predictor for protein-ligands interactions and theoretic-experimental studies of US FDA drug-target network, oxoisoaporphine inhibitors for MAO-A and human parasite proteins. Eur J Med Chem. 2011; 46(12):5838–51. doi: 10.1016/j.ejmech.2011.09.045 PMID: 22005185
7. Romero-Duran FJ, Alonso N, Yanez M, Caamano O, Garcia-Mera X, Gonzalez-Diaz H. Brain-inspired cheminformatics of drug-target brain interactome, synthesis, and assay of TVP1022 derivatives. Neuropharmacology. 2016; 103:270–8. doi: 10.1016/j.neuropharm.2015.12.019 PMID: 26721628
8. Ding H, Takigawa I, Mamitsuka H, Zhu S. Similarity-based machine learning methods for predicting drug-target interactions: a brief review. Brief Bioinform. 2014; 15(5):734–47. doi: 10.1093/bib/bbt056 PMID: 23933754
9. Chen X, Yan CC, Zhang X, Zhang X, Dai F, Yin J, et al. Drug-target interaction prediction: databases, web servers and computational models. Brief Bioinform. 2016; 17(4):696–712. doi: 10.1093/bib/bbv066 PMID: 26283676
10. Gonzalez-Diaz H, Prado-Prado F, Garcia-Mera X, Alonso N, Abeijon P, Caamano O, et al. MIND-BEST: Web server for drugs and target discovery; design, synthesis, and assay of MAO-B inhibitors

and theoretical-experimental study of G3PDH protein from Trichomonas gallinae. J Proteome Res. 2011; 10(4):1698–718. doi: 10.1021/pr101009e PMID: 21184613

11. Gonzalez-Diaz H, Prado-Prado F, Sobarzo-Sanchez E, Haddad M, Maurel Chevalley S, Valentin A, et al. NL MIND-BEST: a web server for ligands and proteins discovery–theoretic-experimental study of proteins of Giardia lamblia and new compounds active against Plasmodium falciparum. J Theor Biol. 2011; 276(1):229–49. doi: 10.1016/j.jtbi.2011.01.010 PMID: 21277861
12. Gfeller D, Grosdidier A, Wirth M, Daina A, Michielin O, Zoete V. SwissTargetPrediction: a web server for target prediction of bioactive small molecules. Nucleic Acids Res. 2014; 42(Web Server issue): W32–8. doi: 10.1093/nar/gku293 PMID: 24792161
13. Nickel J, Gohlke BO, Erehman J, Banerjee P, Rong WW, Goede A, et al. SuperPred: update on drug classification and target prediction. Nucleic Acids Res. 2014; 42(Web Server issue):W26–31. doi: 10.1093/nar/gku477 PMID: 24878925
14. Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, et al. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. Journal of computational chemistry. 1998; 19(14):1639–1662. doi: 10.1002/(SICI)1096-987X(19981115)19:14%3C1639::AID-JCC10%3E3.0.CO;2-B
15. Cheng AC, Coleman RG, Smyth KT, Cao Q, Soulard P, Caffrey DR, et al. Structure-based maximal affinity model predicts small-molecule druggability. Nat Biotechnol. 2007; 25(1):71–5. doi: 10.1038/nbt1273 PMID: 17211405
16. Gonen M. Predicting drug-target interactions from chemical and genomic kernels using Bayesian matrix factorization. Bioinformatics. 2012; 28(18):2304–10. doi: 10.1093/bioinformatics/bts360 PMID: 22730431
17. van Laarhoven T, Nabuurs SB, Marchiori E. Gaussian interaction profile kernels for predicting drug-target interaction. Bioinformatics. 2011; 27(21):3036–43. doi: 10.1093/bioinformatics/btr500 PMID: 21893517
18. Mei JP, Kwoh CK, Yang P, Li XL, Zheng J. Drug-target interaction prediction by learning from local information and neighbors. Bioinformatics. 2013; 29(2):238–45. doi: 10.1093/bioinformatics/bts670 PMID: 23162055
19. Xia Z, Wu LY, Zhou X, Wong ST. Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. BMC Syst Biol. 2010; 4 Suppl 2:S6. doi: 10.1186/1752-0509-4-S2-S6 PMID: 20840733
20. Chen X, Liu MX, Yan GY. Drug-target interaction prediction by random walk on the heterogeneous network. Mol Biosyst. 2012; 8(7):1970–8. doi: 10.1039/c2mb00002d PMID: 22538619
21. Kaufman L, Rousseeuw PJ. Finding groups in data: an introduction to cluster analysis. vol. 344. John Wiley & Sons; 2009.
22. Li Y, Guan C, Li H, Chin Z. A self-training semi-supervised SVM algorithm and its application in an EEG-based brain computer interface speller system. Pattern Recognition Letters. 2008; 29(9):1285–1294. doi: 10.1016/j.patrec.2008.01.030
23. 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. 2008; 24(13):i232–40. doi: 10.1093/bioinformatics/btn162 PMID: 18586719
24. Schomburg I, Chang A, Ebeling C, Gremse M, Heldt C, Huhn G, et al. BRENDA, the enzyme database: updates and major new developments. Nucleic Acids Res. 2004; 32(Database issue):D431–3. doi: 10.1093/nar/gkh081 PMID: 14681450
25. Yoo M, Shin J, Kim J, Ryall KA, Lee K, Lee S, et al. DSigDB: drug signatures database for gene set analysis. Bioinformatics. 2015; 31(18):3069–71. doi: 10.1093/bioinformatics/btv313 PMID: 25990557
26. 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. J Am Chem Soc. 2003; 125(39):11853–65. doi: 10.1021/ja036030u PMID: 14505407
27. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol. 1981; 147 (1):195–7. doi: 10.1016/0022-2836(81)90087-5 PMID: 7265238
28. Chang CC, Lin CJ. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST). 2011; 2(3):27.
29. Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P. Drug target identification using side-effect similarity. Science. 2008; 321(5886):263–6. doi: 10.1126/science.1158140 PMID: 18621671
30. Perlman L, Gottlieb A, Atias N, Ruppin E, Sharan R. Combining drug and gene similarity measures for drug-target elucidation. J Comput Biol. 2011; 18(2):133–45. doi: 10.1089/cmb.2010.0213 PMID: 21314453

31. Wang YC, Zhang CH, Deng NY, Wang Y. Kernel-based data fusion improves the drug-protein interaction prediction. Comput Biol Chem. 2011; 35(6):353–62. doi: 10.1016/j.compbiolchem.2011.10.003 PMID: 22099632

32. Wang YC, Chen SL, Deng NY, Wang Y. Computational probing protein-protein interactions targeting small molecules. Bioinformatics. 2016; 32(2):226–34. PMID: 26415726

33. Zhao XM, Wang Y, Chen L, Aihara K. Gene function prediction using labeled and unlabeled data. BMC bioinformatics. 2008; 9(1):1. doi: 10.1186/1471-2105-9-57