Supplementary Material for

"Predicting Drug-Target Interactions Using Restricted Boltzmann Machines"

Yuhao Wang[1]
Jianyang Zeng[2,*]

April 16, 2013

The following is supplementary material which provides additional information to substantiate the claims of the paper. Section S1 presents descriptive statistics of the MATADOR and STITCH-based datasets that were tested in the paper. Section S2 visualizes part of a DTI network constructed based on the prediction results. Section S3 describes details of the K-fold cross-validation procedure, and the results of a 5-fold cross-validation test performed in the paper. In Section S4, we describe additional cross-validation tests to further compare methods "integrating data with distinction" and "using only a single data type only" with training data of the same size. Section S5 presents details of a simple logic based approach which follows the basic premise that similar drugs and targets should have similar interactions.

S1 Descriptive Statistics of the MATADOR and STITCH-Based Datasets

Table S1 shows descriptive statistics of the MATADOR and STITCH-based datasets that were tested in the paper.

Statistics	**MATADOR-based data**	**STITCH-based data**
Number of drugs	784	598
Number of protein targets	2431	671
Number of drug-target interactions	13064	3296
Number of direct interactions	7862	2532
Number of indirect interactions	5202	764
Number of binding interactions	–	2589
Number of activation interactions	–	945
Number of inhibition interactions	–	1493
Average degree for a drug	16.7	5.5
Average degree for a target	5.4	4.9

Table S1: Descriptive statistics for both MATADOR and STITCH-based datasets.

[1] Department of Automation, Tsinghua University, Beijing, 100084, P. R. China

[2] Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, 100084, P. R. China

[*] Corresponding authors: Jianyang Zeng, zengjy321@tsinghua.edu.cn.

S2 Visualization of Drug-Target Interaction Networks

Fig. S1 visualizes part of a DTI network constructed based on the set of the 50 highest scoring interactions predicted by our algorithm using the MATADOR-based data.

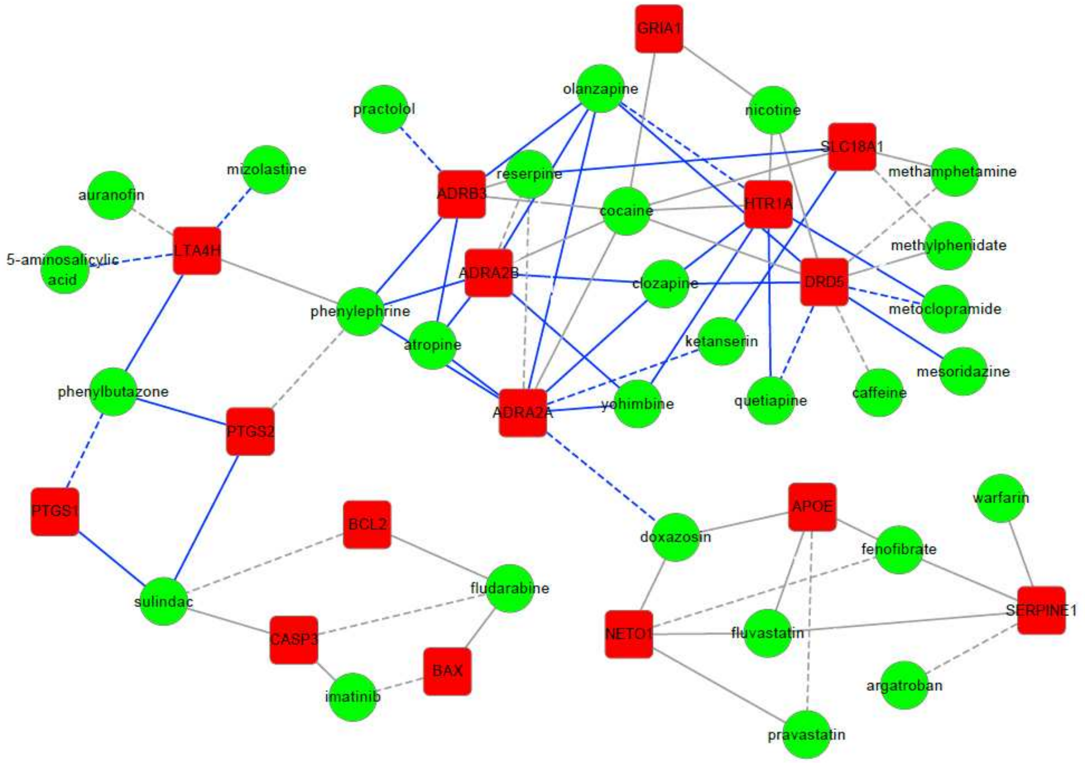

Figure S1: Part of the DTI network constructed based on the set of the 50 highest scoring interactions predicted using the MATADOR-based data. Solid links represent known interactions and dashed links represent predicted ones. Blue links represent direct interactions while grey ones represent indirect interactions. Green circles represent drugs while red squares represent target proteins. The network visualizations were prepared by Cytoscape [5].

S3 Details of the K-Fold Cross-Validation Procedure

Our K-fold cross-validation test (K = 5 or 10) was performed on drug-target interactions (DTIs). Below we describe the details of our K-fold cross-validation test. Suppose in total we have N drug-target interactions (DTIs), and t types of DTI encoded in a visible unit. Let $\mathbf{x_i} = (x_i^1, \cdots, x_i^h, \cdots, x_i^t), i = 1, \cdots, N$, denote the state of the ith DTI, where $x_i^h = 1$ if the hth type of DTI is observed in visible data, and $x_i^h = 0$ otherwise. We randomly partitioned all DTIs, $\mathbf{x_1}, \cdots, \mathbf{x_N}$, into K non-overlapping subsets, each of which had approximately equal size. Each subset was in turn used as test set and the remaining K - 1 subsets were used as training data.

In the real application of the network based prediction of DTIs, we usually aim to predict a small number of unknown DTIs based on a large number of known DTIs. Thus, a cross-validation test with a small test data set and a large training data set should be sufficient enough to simulate the real scenario. Note that 10-fold cross-validation and leave-one-out cross-validation (LOOCV) tests have been widely used in previous work on DTI prediction [9, 1, 6, 4]. To check whether our algorithm can have a wider range of applications, we also performed a 5-fold cross-validation test. The results of this 5-fold cross-validation test are summarized in Table S2. Compared to our original 10-fold cross-validation test (Table 1 in the paper), we only found a small decrease in our algorithm's performance in the 5-fold cross-validation test.

Drug-target relationship	AUC	AUPR
Direct interaction	98.1	86.7
Indirect interaction	96.5	74.8

Table S2: The 5-fold cross-validation results on predicting direct and indirect interactions using our RBM model. Both known direct and indirect interactions from the MATADOR-based data were integrated with distinction in our RBM model.

S4 Additional Cross-Validation Tests

In our cross-validation tests, the size of training data are the same for the first two test methods, namely "integrating data with distinction" and "mixing data without distinction". Thus, in these two tests, AUC and AUPR are comparable. The third test method, i.e., "using direct (indirect) interaction only", used less training data than the first two test method. For example, when predicting direction interactions, the indirect interaction data was not used in the test. This may create bias when comparing two methods that use training data with different sizes. To make a fair comparison on methods "integrating data with distinction" and "using direct (indirect) interaction only", we have performed an additional test which used training data of the same size. In this test, when predicting direct interaction, we removed an indirect interaction if either the drug or target does not have any direct interaction with other drugs or targets in the dataset. By doing so, we maintained the same data size for both methods. We also performed a similar test on predicting indirect DTIs. Table S3 shows the descriptive statistics for the new data used in this additional test. As summarized in Table S4, our new comparison results confirmed that integrating data with distinction outperformed the method that uses a single interaction type only, when predicting direct and indirect DTIs.

Statistics	dataset for direct interaction prediction	dataset for indirect interaction prediction
Number of drugs	718	364
Number of protein targets	1568	1558
Number of drug-target interactions	10211	8228
Number of direct interactions	7862	3026
Number of indirect interactions	2349	5202
Average degree for a drug	14.2	22.6
Average degree for a target	6.5	5.3

Table S3: Descriptive statistics for the dataset with the same size that was used for comparing methods "integrating data with distinction" and "using direct (indirect) interactions only", when predicting direct and indirect DTIs.

In addition, we performed a similar comparison test for predicting different modes of action. Table S5 shows the descriptive statistics for the new data used for predicting different modes of action. As summarized in Table S6, the new comparison results also confirmed that integrating data with distinction outperformed the method that uses a single data type.

Drug-target relationship	**Test method**	**AUC**	**AUPR**
Direct interaction	Integrating data with distinction	98.3	**89.1**
	Using direct interactions only	98.0	78.9
Indirect interaction	Integrating data with distinction	96.9	**79.4**
	Using indirect interactions only	94.8	62.4

Table S4: Results on comparing methods “integrating data with distinction” and “using direct (indirect) interactions only” with training data of the same size, when predicting direct and indirect DTIs. The highest AUPR score is shown in bold.

Statistics	**dataset for binding interaction prediction**	**dataset for activation interaction prediction**	**dataset for inhibition interaction prediction**
Number of drugs	574	261	416
Number of protein targets	526	261	384
Number of drug-target interactions	2952	1454	2253
Number of direct interactions	2517	857	1673
Number of indirect interactions	435	597	580
Number of binding interactions	2589	899	1701
Number of activation interactions	713	945	617
Number of inhibition interactions	1326	614	1493
Average degree for a drug	5.1	5.6	5.4
Average degree for a target	5.6	5.6	5.9

Table S5: Descriptive statistics for the dataset with the same size used for comparing methods “integrating data with distinction” and “using a single interaction type only”, when predicting different modes of action.

Drug-target relationship	**Test method**	**AUC**	**AUPR**
Binding interaction	Integrating data with distinction	94.7	**77.3**
	Using binding interactions only	94.1	74.4
Activation interaction	Integrating data with distinction	89.5	**62.6**
	Using activation interactions only	87.7	56.3
Inhibition interaction	Integrating data with distinction	90.7	**64.5**
	Using inhibition interactions only	89.5	60.2

Table S6: Results on comparing methods “integrating data with distinction” and “using a single interaction type only” with training data of the same size, when predicting different modes of action. The highest AUPR score is shown in bold.

S5 Details of the Simple Logic Based Approach

Previous network-based approaches for drug-target interaction prediction largely depended on the basic premise that similar drugs and targets should have similar interactions, and focused on integrating genomic and pharmacological data to represent the similarities of drugs, targets and their interactions and predict unknown interactions [9, 3, 2, 1, 7, 6, 4, 8]. Unfortunately, these previous approaches cannot be directly extended to represent the statistical structure of a multidimensional DTI network, and predict unknown types of DTIs. To capture the latent correlations among different types of DTIs on a multidimensional network, we have to resort to more effective prediction models. Our RBM-based approach extends the premise that similar drugs and targets should have similar interactions in that it not only considers the binary DTIs, but also captures the intrinsic correlations among different types of DTIs from the statistical structure of data.

As little work had been developed for predicting unknown types of DTIs on a multidimensional network, it was difficult for us to directly compare our work to other prediction approaches. Instead, we have compared our algorithm to a simple logic based approach on the MATADOR-based data. The simple logic based approach takes the same premise that similar drugs and targets should have similar interactions, which has been popularly used in previous DTI prediction approaches [9, 3, 2, 1, 7, 6, 4, 8]. In this simple logic based approach, we first defined a kernel, called *interaction type profile (ITP)* kernel, to measure the similarities of drugs and targets. The ITP kernel is similar to the Gaussian interaction profile kernel that has been used in [6], except that the interaction profiles are represented by different types of DTIs instead of binary DTIs. The basic idea underlying the simple logic based approach is that, the types of DTIs are predicted based on profiles of the drug-target pairs with the highest ITP kernel scores in training data. More details of this approach can be found in Algorithm 1.

Algorithm 1 Simple Logic Based Approach

```
Input: Training data $D$, kernels $K_d(\cdot,\cdot)$ and $K_t(\cdot,\cdot)$,
    drug-target pair $(d, t)$ in which types of DTI need to be predicted.
Find drug $d_{max}$ and target $t_{max}$ in training data $D$, such that drug-target pairs $(d, t_{max})$ and $(d_{max}, t)$
    maximize $K_d(\cdot,\cdot)$ and $K_t(\cdot,\cdot)$ respectively.
if both drug-target pairs $(d, t_{max})$ and $(d, t_{max})$ have direct (indirect) interaction then
    $\Pr[$ (d,t) has direct(indirect) interaction $] = 1$.
else
    if neither drug-target pairs $(d, t_{max})$ nor $(d, t_{max})$ has any interaction then
        $\Pr[$ (d,t) has direct or indirect interaction $] = 0$.
    else
        if only one pair of $(d, t_{max})$ or $(d, t_{max})$ has direct (indirect) interaction then
            $\Pr[$ (d,t) has direct(indirect) interaction $] = 1$.
        else
            if $(d, t_{max})$ and $(d, t_{max})$ have different interaction types then
                $\Pr[$ (d,t) has direct(indirect) interaction $] = \frac{1}{2}$.
            end if
        end if
    end if
end if
```

Suplementary References

[1] Kevin Bleakley and Yoshihiro Yamanishi. Supervised prediction of drug-target interactions using bipartite local models. *Bioinformatics*, 25(18):2397–2403, Sep 2009.

[2] Xing Chen, Ming-Xi Liu, and Gui-Ying Yan. Drug-target interaction prediction by random walk on the heterogeneous network. *Mol Biosyst*, 8(7):1970–1978, Jul 2012.

[3] Feixiong Cheng, Chuang Liu, Jing Jiang, Weiqiang Lu, Weihua Li, Guixia Liu, Weixing Zhou, Jin Huang, and Yun Tang. Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS Comput Biol*, 8(5):e1002503, 2012.

[4] Jian-Ping Mei, Chee-Keong Kwoh, Peng Yang, Xiao-Li Li, and Jie Zheng. Drug-target interaction prediction by learning from local information and neighbors. *Bioinformatics*, Nov 2012.

[5] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13(11):2498–2504, Nov 2003.

[6] Twan van Laarhoven, Sander B. Nabuurs, and Elena Marchiori. Gaussian interaction profile kernels for predicting drug-target interaction. *Bioinformatics*, 27(21):3036–3043, Nov 2011.

[7] Zheng Xia, Ling-Yun Wu, Xiaobo Zhou, and Stephen T C. Wong. Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. *BMC Syst Biol*, 4 Suppl 2:S6, 2010.

[8] Lei Xie, Li Xie, Sarah L. Kinnings, and Philip E. Bourne. Novel computational approaches to polypharmacology as a means to define responses to individual drugs. *Annu Rev Pharmacol Toxicol*, 52:361–379, Feb 2012.

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