Learning Deep Architectures for Interaction Prediction in Structure-based Virtual Screening

Adam Gonczarek, Jakub M. Tomczak, Szymon Zaręba, Joanna Kaczmar
Wrocław University of Science and Technology
adam.gonczarek@pwr.edu.pl

Piotr Dąbrowski, Michał J. Walczak
Indata SA

Abstract

We introduce a deep learning architecture for structure-based virtual screening that generates fixed-sized fingerprints of proteins and small molecules by applying learnable atom convolution and softmax operations to each compound separately. These fingerprints are further transformed non-linearly, their inner-product is calculated and used to predict the binding potential. Moreover, we show that widely used benchmark datasets may be insufficient for testing structure-based virtual screening methods that utilize machine learning. Therefore, we introduce a new benchmark dataset, which we constructed based on DUD-E and PDBBind databases.

1 Introduction

Virtual screening is one of the leading methods in computational drug discovery, which aims at identification of novel small molecules that are capable of binding a drug target, usually a protein. In short, there are two main approaches of virtual screening, ligand-based and structure-based. Ligand-based virtual screening relies on empirically established data, which provide information on active (binding compounds later called ligands) and inactive (not binding) molecules. This approach exploits chemical and spatial similarity among binders to identify new ligands of proteins. The second approach, structure-based virtual screening, requires structural information of a protein to dock a ligand candidate in the binding pockets of a target. Here, a large number of small molecules is screened against a structure of a target protein. Then, binding capacity between protein and compounds is assessed using scoring functions, and finally compounds are triaged according to their binding potential.

The main hurdles affecting virtual screening is complexity of chemical space comprising up to 10^{60} theoretical [1] and 10^7 of commercially available compounds [4][1], as well as high false positive rate of identified ligands and a lack of exhaustive training datasets.

Although the above mentioned hindrances are tackled by various approaches, *e.g.* Smina [5], with different success rate, it is the advent of deep learning that promises superior performance in high-throughput virtual screening [13]. Deep learning has already been successfully employed in ligand-based virtual screening [2, 7, 10] but only recently the very first attempts to the structure-based methods have emerged [9, 14].

[1] http://zinc15.docking.org/

Workshop on Machine Learning in Computational Biology. 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

In this study, we propose **a new deep architecture for predicting binding capacity of a protein-molecule pair**. In addition, we demonstrate the disadvantages of common benchmark datasets, which are used for training and testing screening methods. To fill this gap, we propose a **new benchmark dataset** that is more suitable for structure-based virtual screening.

2 Methodology

Model Our aim is to predict binding potential $y \in \{0, 1\}$ for given pair of a small molecule l and a target (represented by a pocket in protein structure to which a ligand may bind) p. We face three major problems in the stated task: **(i)** both target and small compound vary in size, **(ii)** each of them is represented by a list of atoms and therefore a method must be invariant to any permutation of the list, **(iii)** these are 3D structures, thus a method must be invariant to translations and rotations. To cope with these issues we propose to process the protein and the small molecule separately to obtain two fixed-size descriptions (*fingerprints*) that can be further transformed non-linearly *e.g.* by neural nets, and finally used for binding prediction. This approach aims at processing the protein-compound pair separately and then learning a relation of the interaction. A pipeline of the proposed method is presented in Figure 1(a).

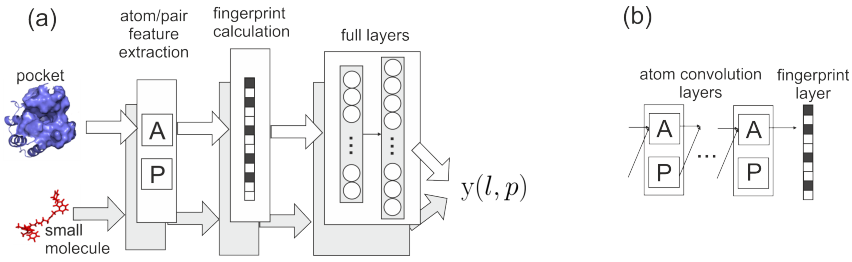

Figure 1: (a) Schema of the proposed approach. Letters A and P denote lists of atoms and connections, respectively. (b) Details about the neural fingerprint.

The crucial part of the proposed approach is a *fingerprint*, *i.e.*, a description of a fixed size. One of the widely used fingerprints for virtual screening is *Extended Connectivity Fingerprint* (ECFP) [11]. ECFP is an automatic manner of determining fingerprints by consecutively applying a *hash function* on atom and its neighborhood followed by a *indexing operation*. The hash function allows to combine information about each atom and its neighboring substructures while the indexing operation is used to combine all the nodes' features into a single fingerprint of the whole compound. However, due to pre-determined form of hashing and indexing, ECFP is sensitive to small perturbations in molecule structure, and therefore the features obtained by this method are not very robust.

Very recently, the drawbacks of ECFP were alleviated by application of learnable operations similar to operations in convolutional neural nets [2]. Here the hashing is replaced with an adaptive convolutional-like operation and the indexing with a softmax operation. This could be formalized as follows. Let us denote an m^{th} atom in a compound described by F features by $\mathbf{a}_m$. Then hashing could be described in the following fashion:

$$\mathbf{a}_m := \sigma\left(\mathbf{W}\mathbf{a}_m + \sum_{i=1}^{I_m} \mathbf{H}_{I_m}\mathbf{a}_i + \mathbf{b}\right), \tag{1}$$

where $\mathbf{a}_i$ is i^{th} neighboring atom, I_m is the number of possible neighbors for the m^{th} atom[2], $\mathbf{W} \in \mathbb{R}^{R\times F}$ is a matrix of weights[3], $\mathbf{H}_1, \ldots, \mathbf{H}_5 \in \mathbb{R}^{R\times F}$ are matrices of weights for neighbors, $\mathbf{b} \in \mathbb{R}^R$ is a bias vector, and $\sigma(\cdot)$ is an element-wise non-linear function, *e.g.*, the sigmoid function or ReLU. We refer to this operation as *atom convolution* and it can be repeated K times which constitutes K layers, and each layer consists of own weights to learn (see Fig. 1(b)).

[2]Due to the physical properties of compounds there can be only up to 5 neighbors, so $I_m \in \{1, 2, \ldots, 5\}$.

[3]Notice that R is the number of new features and this could differ from F.

The indexing operation is then replaced with a softmax operation that consecutively applies the softmax function to each atom in the compound to yield the final *neural fingerprint* $\mathbf{n}$:

$$\mathbf{n} = \sum_m \mathrm{sofmax}(\mathbf{V}\mathbf{a}_m + \mathbf{c}) \tag{2}$$

where $\mathbf{V} \in \mathbb{R}^{S \times R}$ is a weight matrix, $\mathbf{c} \in \mathbb{R}^S$ is a bias vector and S is the size of the fingerprint.

Next, after obtaining the neural fingerprint for small compound ($\mathbf{n}_l$) and protein ($\mathbf{n}_p$), we apply a neural network (MLP) to obtain new representations: $\mathbf{w} = \mathrm{MLP}_l(\mathbf{n}_l)$ and $\mathbf{v} = \mathrm{MLP}_p(\mathbf{n}_p)$. Eventually, we calculate the bioactivity by transforming the inner product of $\mathbf{w}$ and $\mathbf{v}$ using the sigmoid function $\mathrm{sigm}(\cdot)$:

$$y(l, p) = \mathrm{sigm}(\mathbf{w}^\top \mathbf{v}). \tag{3}$$

Training Standard learning of neural networks utilizes the cross-entropy (CE) loss function. Typically databases contain only active ligand-protein pairs. Hence, we propose to add an additional term to the CE loss in order to avoid overfitting to the positive class:

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{n=1}^{N} \log y(l_n, p_n) + \mathbb{E}_{l,p \sim P(l,p)}[\log(1 - y(l, p))]. \tag{4}$$

Despite the fact that the expected value can be approximated using Monte Carlo methods, this formulation causes a problem since finding the joint probability of the small molecule-protein pair is a very complex task. We overcome this issue with the assumption that taking a random pair from the dataset would result in a negative (not binding) example. Obviously, such an approach introduces a bias, however, a chance of producing wrong label is negligible. The proposed loss function in (4) is closely related to the Noise Contrastive Estimation [3].

3 Results

The efficacy of machine learning methods for virtual screening are typically evaluated with one of the renowned benchmarks, *e.g.*, DUD-E [8]. Generally, the evaluation dataset is divided into training and testing sets that contain different targets (together with their actives and decoys). Interestingly, it turns out that this testing protocol might be strongly biased due to similarity of artificial decoys for different targets. We addressed this problem with a newly developed benchmark based on two separate datasets.

DUD-E experiment We used DUD-E[4] benchmark, consisting of 102 proteins (targets), $22,886$ active compounds (ligands or binders) and over $1M$ decoys (non-binders). We randomly divided targets into training (72) and testing (30) parts, which is similarly to [14]. We applied the ECFP fingerprint with the size of 4096 to small compounds (cmpds) only and trained logistic regression (LR) to discriminate between actives and decoys. Notice that no information about targets was used. The method achieved 0.904 mean AUC, evaluated on the targets in the test set, and to the best of our knowledge it has outperformed other state-of-the-art methods for structure-based virtual screening trained in the similar manner (see Table 1). Thus, this suggests that datasets with many artificially generated decoys (like DUD-E) are prone to bias due to similarity of majority of the inactive compounds for one target to inactive compounds for other targets. Further, application of basic learning methods to small compounds only results in improved performance. Consequently, it is uncertain whether a method evaluated on this testing scheme learns the relationship between compounds and targets, or learns the discrimination between active and inactive molecules, where additional information about targets only contributes to noise.

Table 1: Results on DUD-E benchmark (70% of data for training and 30% of data for testing) and on DUD benchmark (leave-one-out cross-validation).

Dataset	Method	Mean AUC
DUD-E	Smina	0.700
	AtomNet [14]	0.855
	cmpds ECFP + LR	**0.904**
DUD	DeepVS [9]	0.800

[4] `http://dude.docking.org/`

PDBBind + DUD-E Next, we employed PDBBind [6] for training and DUD-E for testing. The original PDBBind database contains 3D structures of about $10k$ complexes, *i.e.*, structures of ligands docked in binding pockets of proteins. We have removed all the complexes containing targets from DUD-E, obtaining 8822 complexes for training. Notice that this dataset contains no artificially generated decoys, negative examples are generated by sampling random target-compounds pairs from the dataset (see Eq. 4). For testing we used 88 DUD-E targets represented by a binding pocket extracted from the original PDBBind. For each target we have randomly sampled 1000 compounds (actives and decoys) from DUD-E, resulting in $88,000$ testing examples. We tested two different models based on the pipeline presented in Fig. 1(a). In the first model, standard ECFP fingerprints were applied both to small compounds and pockets. In the second one we adopted learnable neural fingerprints. We compared our approach to two widely used methods, *i.e.*, AutoDock Vina [12] and Smina [5]. The results are presented in Table 2. First, we see the importance of applying learnable fingerprint, since ECFP performed worse than the reference methods. Second, we observe that the neural fingerprint-based approach outperformed both reference methods, achieving better mean AUC, and reaching more targets that exceeded high AUC thresholds.

Table 2: Results obtained on the proposed benchmark. The presented approach with ECFP is denoted by Ours(ECFP) and the one with neural fingerprint by Ours(NF). AUC $\geq \alpha$ denotes on how many targets (out of 88) a method performs better than α.

Method	Total AUC	Mean AUC ($\pm$ std.)	AUC ≥ 0.7	AUC ≥ 0.8	AUC ≥ 0.9
AutoDock Vina	0.644	0.691 ± 0.147	47	21	4
Smina	0.653	0.704 ± 0.138	**54**	23	4
Ours(ECFP)	0.600	0.551 ± 0.166	21	2	0
Ours(NF)	**0.714**	**0.705 $\pm$ 0.168**	47	**29**	**11**

4 Conclusions

This study results in two contributions in the field of computational drug discovery. First, we propose a new benchmark dataset built on top of the two established datasets, PDBBind and DUD-E. This benchmark provides suitable information for the development of structure-based virtual screening methods. At the same time, we demonstrate that currently available benchmarks represent mediocre training and testing sets due to insufficient coverage of chemical complexity. Second, we propose a novel deep learning-based approach able to identify ligands of target protein. The performed experiments showed that our approach outperforms two widely used methods, AutoDock Vina and Smina. Here developed method reaches AUC of 0.9 or greater for the 11 targets, while the reference methods exceed AUC of 0.9 for the 4 targets. We anticipate further evolution of the proposed approach by applying more sophisticated deep learning techniques, *e.g.* by developing more accurate learnable fingerprints.

Acknowledgments

The work conducted in this paper is partially co-financed by European Regional Development Fund within the framework of the Smart Growth Operational Programme 2014-2020, grant No. POIR.01.01.01-00-1083/15.

References

[1] RS Bohacek, C McMartin, and WC Guida. The art and practice of structure-based drug design: A molecular modeling perspective. *Medicinal research reviews*, 16(1):3–50, 1996.

[2] DK Duvenaud, D Maclaurin, J Iparraguirre, R Bombarell, T Hirzel, A Aspuru-Guzik, and RP Adams. Convolutional networks on graphs for learning molecular fingerprints. In *NIPS*, pages 2215–2223, 2015.

[3] MU Gutmann and A Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *J. Mach. Learn. Res.*, 13(Feb):307–361, 2012.

[4] JJ Irwin, T Sterling, MM Mysinger, ES Bolstad, and RG Coleman. Zinc: a free tool to discover chemistry for biology. *Journal of chemical information and modeling*, 52(7):1757–1768, 2012.

[5] DR Koes, MP Baumgartner, and CJ Camacho. Lessons learned in empirical scoring with smina from the csar 2011 benchmarking exercise. *J. Chem. Inf. Model.*, 53(8):1893–1904, 2013.

[6] Z Liu, Y Li, L Han, J Li, J Liu, Z Zhao, W Nie, Y Liu, and R Wang. Pdb-wide collection of binding data: current status of the pdbbind database. *Bioinformatics*, page btu626, 2014.

[7] J Ma, RP Sheridan, A Liaw, GE Dahl, and V Svetnik. Deep neural nets as a method for quantitative structure–activity relationships. *J. Chem. Inf. Model.*, 55(2):263–274, 2015.

[8] MM Mysinger, M Carchia, JJ Irwin, and BK Shoichet. Directory of useful decoys, enhanced (dud-e): better ligands and decoys for better benchmarking. *Journal of medicinal chemistry*, 55 (14):6582–6594, 2012.

[9] JC Pereira, ER Caffarena, and C Santos. Boosting docking-based virtual screening with deep learning. *arXiv:1608.04844*, 2016.

[10] B Ramsundar, S Kearnes, P Riley, D Webster, D Konerding, and V Pande. Massively multitask networks for drug discovery. *arXiv:1502.02072*, 2015.

[11] D Rogers and M Hahn. Extended-connectivity fingerprints. *J. Chem. Inf. Model.*, 50(5):742–754, 2010.

[12] O Trott and AJ Olson. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 31(2):455–461, 2010.

[13] T Unterthiner, A Mayr, G Klambauer, M Steijaert, JK Wegner, H Ceulemans, and S Hochreiter. Deep learning as an opportunity in virtual screening. *NIPS*, 27, 2014.

[14] I Wallach, M Dzamba, and A Heifets. AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery. *arXiv:1510.02855*, 2015.