Computers in Biology and Medicine 100 (2018) 253–258

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Interaction prediction in structure-based virtual screening using deep learning

Adam Gonczarek [a,c,*], Jakub M. Tomczak [a], Szymon Zaręba [a,c], Joanna Kaczmar [a], Piotr Dąbrowski [a,b], Michał J. Walczak [c]

[a] Department of Computer Science, Wrocław University of Science and Technology, Poland
[b] Indata SA, Wrocław, Poland
[c] Alphamoon, Wrocław, Poland

ARTICLE INFO

Keywords:
Virtual screening
Neural fingerprint
Graph convolution
Deep learning
PDBBind
DUD-E
MUV

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 molecule separately. These fingerprints are further non-linearly transformed, 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, MUV 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 general, there are two main approaches of virtual screening, ligand-based (LBVS, Ligand-Based Virtual Screening) and structure-based (SBVS, Structure-Based Virtual Screening). 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. Most importantly, LBVS is a method used when structures of targets coming from either X-ray crystallography or NMR spectroscopy are missing or the obtained respective structures lack accuracy to design binders via Structure-Based Drug Design (SBDD) methods.

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 problems of SBVS arise from exponentially growing number of protein structures [1] and the fact that proteins are dynamic ensembles often adopting different conformations [3]. One of the key challenges of SBVS is a faithful delineation of the relationships between ligands and a target protein. To do so, small molecules are represented by descriptors containing information on physico-chemical properties. They are later exploited to predict biological activity of the chemical compounds. This description and prediction is known in drug discovery as Quantitative Structure-Activity Relationship, QSAR, and constitutes the unmet need in virtual drug discovery [5,13,26].

Particular attention was drawn to QSAR in 2013, when pharmaceutical company, Merck, announced a competition for virtual screening techniques to predict actives for given targets based on Merck's in-house data [18,19].

The main hurdles affecting virtual screening is complexity of chemical space comprising up to 10^{60} theoretical [4] and 10^{7} of commercially available compounds [11],[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 [15] or homology-based methods [9], with different success rate, it is the advent of machine learning that promises superior performance in high-throughput virtual screening [2,29,30].

* Corresponding author. Department of Computer Science, Wrocław University of Science and Technology, Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland.
E-mail address: adam.gonczarek@pwr.edu.pl (A. Gonczarek).

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

http://dx.doi.org/10.1016/j.compbiomed.2017.09.007
Received 31 January 2017; Received in revised form 22 August 2017; Accepted 8 September 2017

An important group of methods applied to this task are kernel based approches. In Ref. [32] target is represented as fragment vector, where target information is based on physico-chemical properties of present binding sites and ligand is represented by binary vector indicating presence of common substructures. In Ref. [16] targets and ligands are arranged in an interaction network, where connections between targets and ligands indicate known interactions, thus binary interaction features for molecules are generated. In both cases interaction kernel is used to measure affinity between binding sites and ligands. More general approaches were also introduced. In Ref. [12] strategy for building target-ligand pair kernels is presented, then various kernels for ligands and targets are tested and compared. In Ref. [21] kernels are used to create target-ligand space and then feature selection on that space is performed.

Deep learning has already been successfully employed in ligand-based virtual screening [7,18,23] but only recently the very first attempts to the structure-based methods have emerged [22,27,31]. Particular hopes for deep learning are put on capturing relations between molecular descriptors of the small molecules and their activity to target desired protein, and hence to obtain QSAR models of high fidelity [8].

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

Our aim is to predict binding ability $y \in \{0,1\}$ 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 Fig. 1(a).

2.1. Feature extraction

In the proposed framework each molecule or protein pocket is represented by two lists of features specifying atoms and atom connections, respectively. Features used to describe a single atom involve type and partial charge. Atom type is encoded using one-hot vector. Note that the variety of atoms building the structure of small molecules is significantly greater than in case of proteins, therefore the length of encoded vectors differs for those two groups of molecules. Partial charges are transformed into vector of features using fuzzy Gaussian bins. Precisely, we take ten univariate Gaussian density functions uniformly distributed on interval $[-1,1]$. Then, for each partial charge we evaluate the density values, mapping each charge into ten-dimensional feature vector. Ultimately, we concatenate atom type and partial charge features into a single feature vector denoted by $\mathbf{a}_m^{(0)}$, where m is an atom index. Moreover, since the information about partial charges is typically unavailable for proteins, we first estimate the charges using PDB2PQR software [6].

For extracting connection features a different approach is taken for small molecules and proteins. For the former we use the binding type (single, double, triple, aromatic), membership in the carbon ring and conjugation. For binding type we use one-hot encoding. In case of the proteins due to the use of pocket instead of the whole protein, we usually have a few separate groups of atoms, i.e., the detached fragments of a protein chain. Moreover, the binding ability is usually determined by smaller spatial substructures forming the entire pocket. Thus, instead of using chemical bonds, we define neighbors as the atoms within a certain radius around the atom of interest. For the purpose of our research we set the radius to be 5 Å. This value is chosen empirically, considering we do not allow isolated atoms, but also would like to keep the number of neighbors for each atom relatively small. Then the minimum spanning tree is found and the edges in the resulting tree are used as atom connections. The presented approach allows us to take into account the 3D structure of a pocket in the same manner as we take information about chemical bonds in small molecules. Finally, each connection is described by the distance between atoms transformed into feature vector using fuzzy Gaussian bins. We use analogous mapping as in case of partial charges, for each distance evaluating values of ten Gaussian densities uniformly distributed on interval [0,5] angstroms. We use the notation c_{ij} for either small molecules or protein pockets to denote the feature vector describing the connection between atoms i and j.

2.2. Fingerprint calculation

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) [24]. ECFP is an automatic manner of determining fingerprints by consecutively applying a *hash function* on atom and its neighborhood followed by an *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 [7]. Here the hashing is replaced with an adaptive convolutional-like operation and the indexing with a softmax operation. The convolution can be described in the following fashion:

$$\mathbf{a}_m^{(k)} = \sigma\left(\mathbf{W}^{(k)}\mathbf{a}_m^{(k-1)} + \sum_{i=1}^{N_m}\mathbf{W}_{N_m}^{(k)}\left[\mathbf{a}_i^{(k-1)};\mathbf{c}_{mi}\right] + \mathbf{b}^{(k)}\right), \tag{1}$$

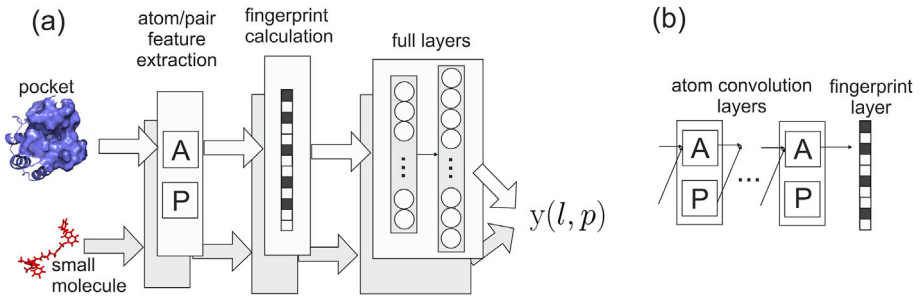

Fig. 1. (a) Schema of the proposed approach. Letters A and P denote lists of atoms and connections, respectively. (b) Details about the neural fingerprint.

where k specifies the number of convolution layer, $\mathbf{a}_i^{(k)}$ are the features of *ith* neighboring atom, $\mathbf{c}_{mi}$ are the features of a connection between atoms m and i, $[\cdot;\cdot]$ denotes a column vector resulting from concatenating two other column vectors, N_m is the number of neighbors for the *mth* atom,[2] $\mathbf{W}^{(k)}$ is a matrix of weights, $\mathbf{W}_1^{(k)},\ldots,\mathbf{W}_5^{(k)}$ are matrices of weights for different number of neighbors, $\mathbf{b}^{(k)}$ 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, each layer consists of own weights to learn and processes the output of the previous layer (see Fig. 1(b)). Notice that after a few convolution steps the information about small spatial substructure around a single atom m is gathered in the feature vector $\mathbf{a}_m^{(K)}$. We have also experimented with a slightly different type of learnable convolution for graphs, where atom neighbors are additionally weighted using the Laplacian matrix. For detailed description, we refer to [14].

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

$$\mathbf{n} = \sum_m \operatorname{softmax}\left(\mathbf{V}\mathbf{a}_m^{(K)} + \mathbf{d}\right) \tag{2}$$

where $\mathbf{V}$ is a learnable weight matrix, $\mathbf{d}$ is a bias vector. Notice that this results in fixed-size fingerprint regardless of the number of atoms. Moreover, this operation is invariant to the specific order of atoms in the input atom list.

Relationship modeling. 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 binding potential by transforming the inner product of $\mathbf{w}$ and $\mathbf{v}$ using the sigmoid function $\operatorname{sigm}(\cdot)$:

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

2.3. 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:

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

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 [10].

2.4. Implementation details

Our method was implemented and tested in Python 2.7 using TensorFlow[3] 0.12 as a computation framework with GPU support and RDKit[4] 2016.09.4 as a molecule processing library. The biggest challenge in using TensorFlow for implementing the presented method was handling the graph nature of molecule structure, which implies that the number of atom neighbors varies. This makes it hard to efficiently process the data as fixed sized tensors, what is a key assumption in the TensorFlow framework. To overcome this problem each molecule was transformed into a set of matrices, each containing atoms with a specific number of neighbors only.

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). The best result is bolded.

Dataset	Method	Mean AUC
DUD-E	Smina	0.700
	AtomNet [31]	0.855
	cmpds ECFP + LR	**0.904**
DUD	DeepVS [22]	0.800

3. Experiments

The efficacy of machine learning methods for virtual screening is typically evaluated with one of the renowned benchmarks, *e.g.*, DUD-E [20]. 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.

3.1. DUD-E experiment

3.1.1. Setup

We used DUD-E[5] benchmark, consisting of 102 proteins (targets), 22,886 active compounds (ligands or binders) and over 1 *M* decoys (non-binders). We randomly divided targets into training (72) and testing (30) parts, which is similarly to [31]. 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.

3.1.2. Results

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.

3.2. New benchmark: PDBBind + DUD-E and MUV

3.2.1. Setup

In our second experiment we employed PDBBind [17] for training and DUD-E and MUV for testing. PDBBind is a database of 11918 protein-ligand pairs with experimentally determined binding affinity of each complex. This dataset also contains binding pocket for each protein. Although it provides high quality set of positive examples it does not contain non-binders. DUD-E dataset contains 102 targets, each with assigned varying number (on average 224) of ligands per target (from a total set of 22,886) and 50 decoys for each active. The decoys were selected from the ZINC database using topological similarity fingerprint

[2] Due to the physical properties of atoms there can be only up to 5 neighbors, so $I_m \in \{1,2,\ldots,5\}$.
[3] https://www.tensorflow.org/.
[4] http://www.rdkit.org/.
[5] http://dude.docking.org/.

to minimize the resemblance between members of the ligands and decoys sets. Therefore, the binding affinity of decoys has not been experimentally verified. MUV dataset [25] contains only 17 targets, the number of actives and decoys per target is similar to DUD-E. To ensure manageable learning time, the number of small molecules was limited to a set of 1000 per protein chosen randomly from decoys and actives for both datasets.

One of our basic assumptions was that our method takes binding pocket as part of the input. This information is missing in DUD-E database. To overcome this issue the PDBBind and DUD-E datasets were cross referenced to find common proteins. 88 proteins occurred in both datasets and this common set was used for our experiment alongside with pocket information. This dataset will be called enriched DUD-E in this article. Similar cross referencing was done for the PDBBind and MUV datasets resulting in 7 proteins in the final enriched MUV dataset.

The previous experiment shows that testing on a dataset with artificially created decoys can prove problematic and misleading owing to their resemblance throughout the set. Therefore, we decided to employ PDBBind for training and enriched DUD-E and MUV for validation and testing. Obviously all complexes containing targets that appeared in DUD-E or MUV were removed from PDBBind before the experiment. Next, the enhanced DUD-E dataset was divided into training and validation sets. As the number of targets in mentioned dataset is relatively low, we had to ensure good representation of proteins, for which the binding affinities are hard to predict, in both training and validation sets. This was done using the following procedure: (i) a baseline AUC of prediction was calculated, (ii) the targets were ordered according to this score and every fourth target was assigned to validation set resulting in 66 proteins in test set and 22 validation set. MUV dataset was used for testing purpose only.

We tested two different models based on the pipeline presented in Fig. 1(a). In the first model, we adopted Graph Convolutional Networks (GCN or graph convolutions) [14] and in the third model we utilized learnable neural fingerprints (or atom convolutions) presented in the previous section. We compared our approach to two widely used methods, *i.e.*, AutoDock Vina [28] and Smina [15].

3.2.2. Results

The results of total AUC are gathered in Table 2. The distribution of AUC for individual proteins are depicted as violin plots in Figs. 6 and 7. Additionally, we present the ROC curve for all proteins in Figs. 2 and 3 and for individual proteins in Figs. 4 and 5. First, we see that the choice of learnable fingerprint matters, since GCN performed worse than the reference methods. Second, we observe that the atom convolutional-based neural fingerprint outperformed both reference methods, achieving better mean AUC (see Table 2), and reaching more targets that exceeded high AUC thresholds (see Figs. 6 and 7).

In Fig. 8 we present four pockets that are representative for the analysis of the performance of our algorithm against the reference methods, where a) all the methods underperform, 3c4f, b) all the methods perform well, 3cqw, c) our methods under performs and SMINA and AutoDock Vina perform well, 2nnq, d) our method overperforms the remaining ones, 1b9v. In the first case, a), the binding cavity is rather deep with equal distribution of positive and negative charges and mildly hydrophobic. In the second case, b), the binding pocket is more hydrophobic with a relatively strong negatively charge gate around it. The

Table 2
Results obtained on the proposed benchmark. The presented approach with neural fingerprint (atom convolution) by Ours (NF) and the approach with graph convolutions are depicted as Ours (GCN). The best results are bolded.

Method	DUD-E	MUV
AutoDock Vina	0.633	0.503
Smina	0.642	0.503
Ours (GCN)	0.567	0.474
Ours (NF)	**0.704**	**0.575**

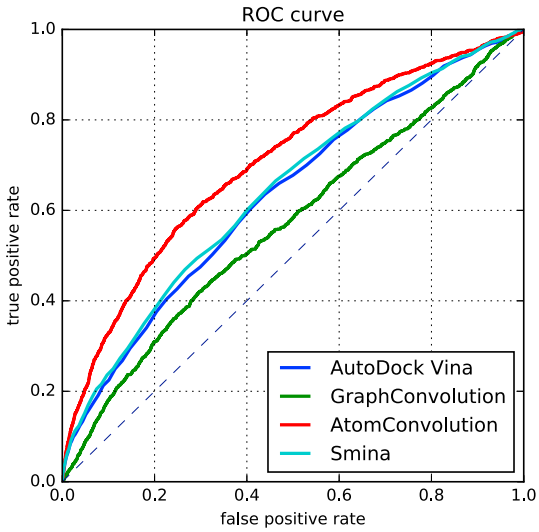

Fig. 2. Comparison of the considered methods using the ROC curve on DUD-E.

example c), exhibits only very shallow and electrostatically charged pockets and the example d), features deep and strongly negatively charged binding site with a strong positively charge gate outside of the pocket. The simplest explanation for a significant overperformance of the docking methods, Smina and AutoDock Vina, over our method in the case c) is that shallow and charged pockets are very sparsely represent in our training set (PDBBind). This due to experimental hindrances, in particular due to the fact that shallow and charged pockets bind ligands rather weakly and this prohibits co-crystallization of them. Moreover, NMR data are almost absent in PDBBind, and this results in substantially poorer training of our method. In contrast to this case, the example d) shows that our method overperforms the docking approaches. The type of pocket presented in d) is actually well represented in PDBBind data set, which leads to good performance of our method. It is however difficult to assess why all the methods perform poorly on the first case and well on the second one. We believe that there might be sequence-dependency but this is beyond the scope of this work.

4. Conclusions

This study results in two contributions in the field of computational

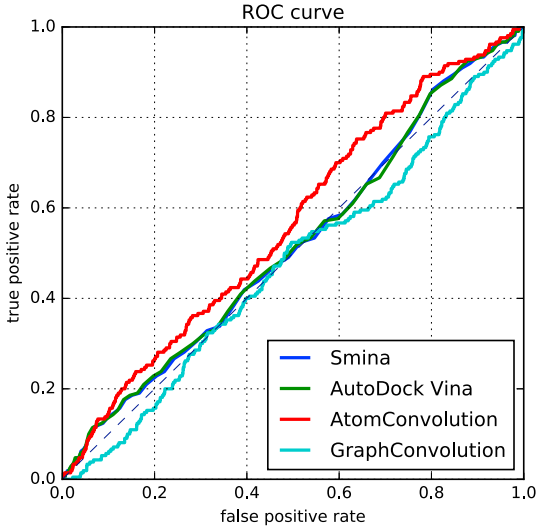

Fig. 3. Comparison of the considered methods using the ROC curve on MUV.

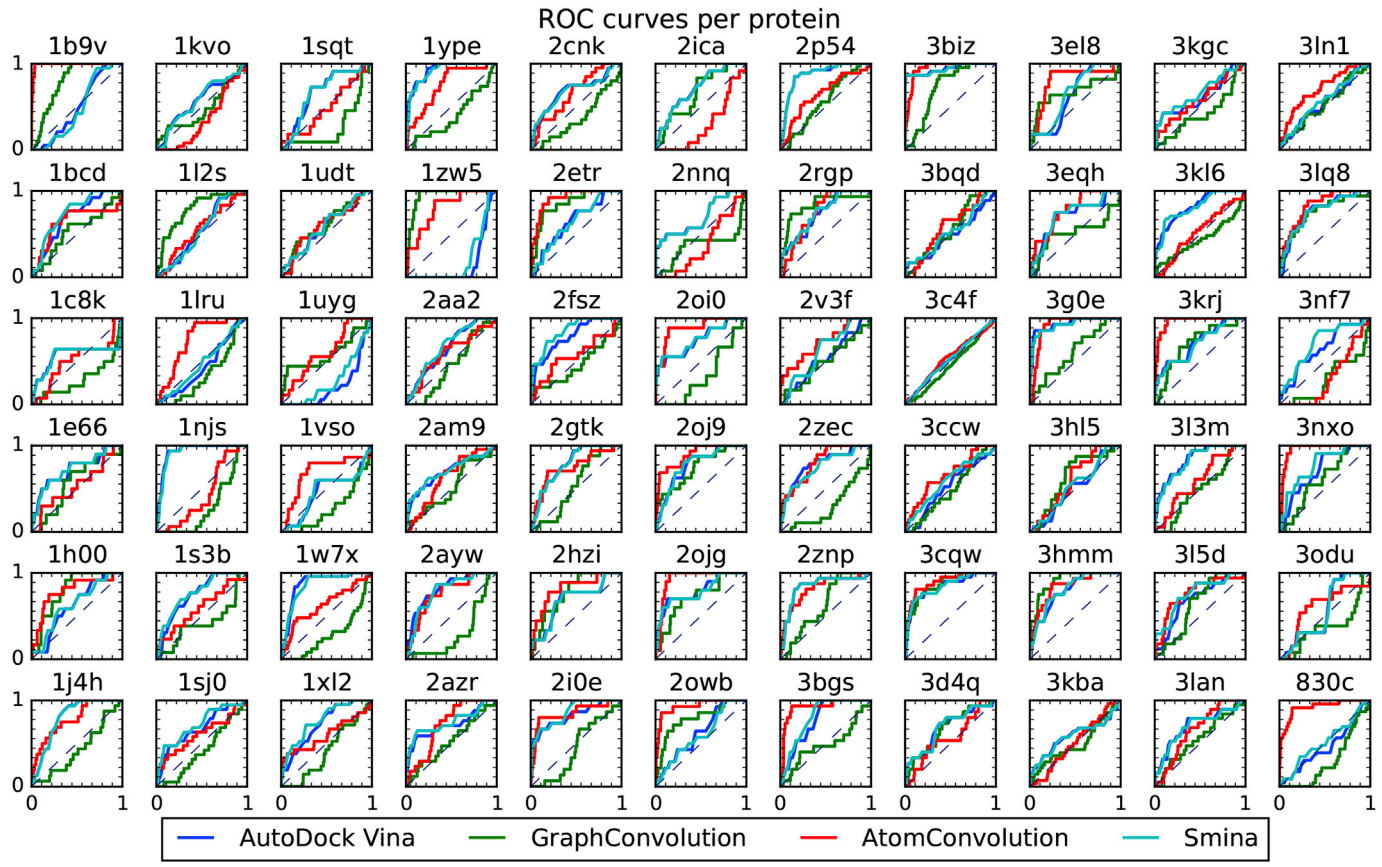

Fig. 4. Comparison of the considered methods using the ROC curve for individual proteins on DUD-E.

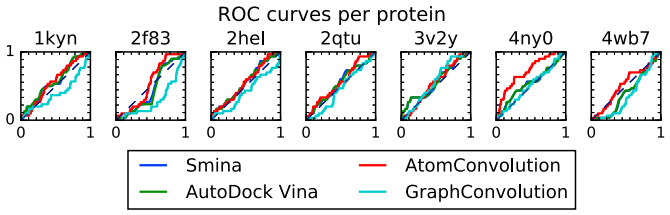

Fig. 5. Comparison of the considered methods using the ROC curve for individual proteins on MUV.

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

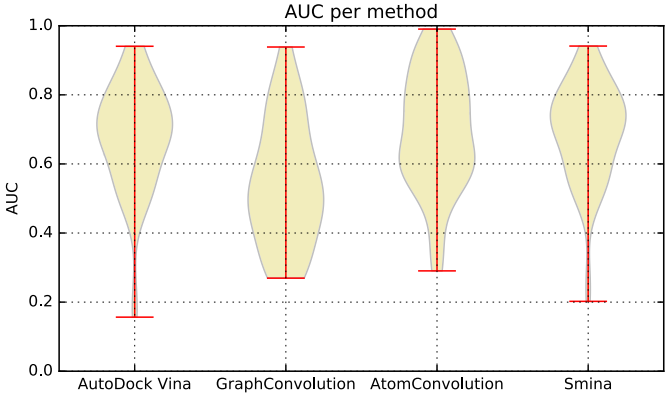

Fig. 6. Comparison of the considered methods using the AUC represented as a violin plot on DUD-E.

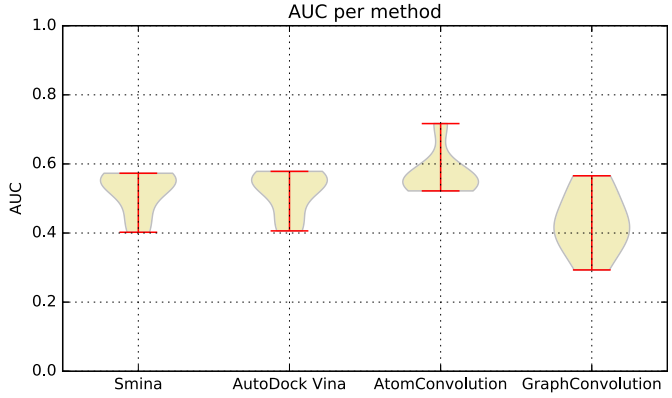

Fig. 7. Comparison of the considered methods using the AUC represented as a violin plot on MUV.

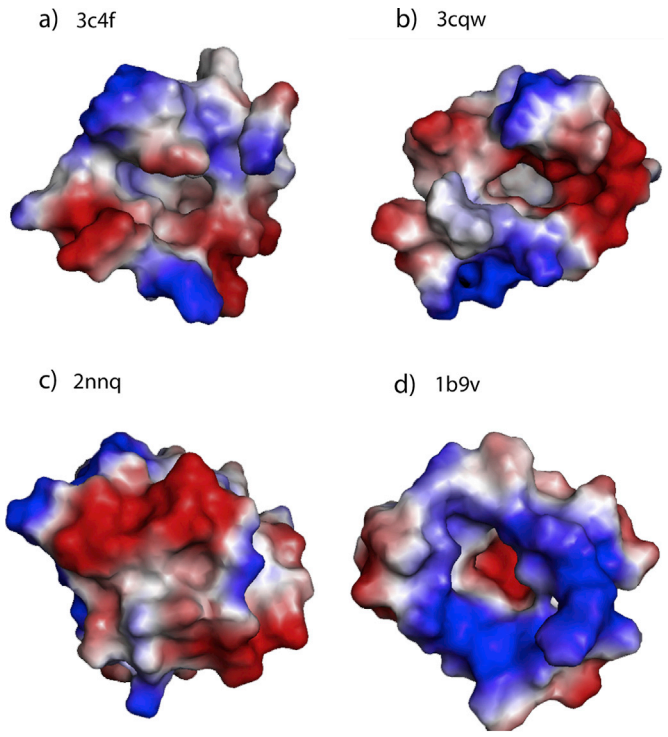

Fig. 8. Selected protein pockets from DUD-E.

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] C. Abad-Zapatero, Notes of a protein crystallographer: on the high-resolution structure of the pdb growth rate, Acta Crystallogr. Sect. D. Biol. Crystallogr. 68 (2012) 613–617.
[2] Q. Ain, A. Aleksandrova, F. Roessler, P. Ballester, Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening, Wiley Interdiscip. Rev. Comput. Mol. Sci. 5 (2015) 405–424.
[3] R. Akbar, S. Jusoh, R. Amaro, V. Helms, Enri: a tool for selecting structure-based virtual screening target conformations, Chem. Biol. Drug Des. (2016).
[4] R. Bohacek, C. McMartin, W. Guida, The art and practice of structure-based drug design: a molecular modeling perspective, Med. Res. Rev. 16 (1996) 3–50.
[5] A. Cherkasov, E. Muratov, D. Fourches, A. Varnek, I. Baskin, M. Cronin, J. Dearden, P. Gramatica, Y. Martin, R. Todeschini, et al., Qsar modeling: where have you been? where are you going to? J. Med. Chem. 57 (2014) 4977–5010.
[6] T. Dolinsky, J. Nielsen, J. McCammon, N. Baker, Pdb2pqr: an automated pipeline for the setup of poisson–boltzmann electrostatics calculations, Nucleic acids Res. 32 (2004) W665–W667.
[7] D. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, R. Adams, Convolutional networks on graphs for learning molecular fingerprints, NIPS (2015) 2215–2223.
[8] E. Gawehn, J. Hiss, G. Schneider, Deep learning in drug discovery, Mol. Inf. 35 (2016) 3–14.
[9] M. Grant, Protein structure prediction in structure-based ligand design and virtual screening, Comb. Chem. high throughput Screen. 12 (2009) 940–960.
[10] M. Gutmann, A. Hyvärinen, Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics, J. Mach. Learn. Res. 13 (2012) 307–361.
[11] J. Irwin, T. Sterling, M. Mysinger, E. Bolstad, R. Coleman, Zinc: a free tool to discover chemistry for biology, J. Chem. Inf. Model. 52 (2012) 1757–1768.
[12] L. Jacob, J.P. Vert, Protein-ligand interaction prediction: an improved chemogenomics approach, Bioinformatics 24 (2008) 2149–2156.
[13] M. Khan, I. Sylte, Predictive qsar modeling for the successful predictions of the admet properties of candidate drug molecules, Curr. drug Discov. Technol. 4 (2007) 141–149.
[14] T. Kipf, M. Welling, Semi-supervised Classification with Graph Convolutional Networks, arXiv preprint arXiv:1609.02907, 2016.
[15] D. Koes, M. Baumgartner, C. Camacho, Lessons learned in empirical scoring with smina from the csar 2011 benchmarking exercise, J. Chem. Inf. Model 53 (2013) 1893–1904.
[16] T. van Laarhoven, S. Nabuurs, E. Marchiori, Gaussian interaction profile kernels for predicting drug–target interaction, Bioinformatics 27 (2011) 3036–3043.
[17] Z. Liu, Y. Li, L. Han, J. Li, J. Liu, Z. Zhao, W. Nie, Y. Liu, R. Wang, Pdb-wide collection of binding data: current status of the pdbbind database, Bioinformatics (2014) btu626.
[18] J. Ma, R. Sheridan, A. Liaw, G. Dahl, V. Svetnik, Deep neural nets as a method for quantitative structure–activity relationships, J. Chem. Inf. Model 55 (2015) 263–274.
[19] J. Markoff, Scientists see promise in deep-learning programs, N. Y. Times (2012).
[20] M. Mysinger, M. Carchia, J. Irwin, B. Shoichet, Directory of useful decoys, enhanced (dud-e): better ligands and decoys for better benchmarking, J. Med. Chem. 55 (2012) 6582–6594.
[21] S. Niijima, H. Yabuuchi, Y. Okuno, Cross-target view to feature selection: identification of molecular interaction features in ligand- target space, J. Chem. Inf. Model. 51 (2010) 15–24.
[22] J. Pereira, E. Caffarena, C. Santos, Boosting Docking-based Virtual Screening with Deep Learning, arXiv:1608.04844, 2016.
[23] B. Ramsundar, S. Kearnes, P. Riley, D. Webster, D. Konerding, V. Pande, Massively Multitask Networks for Drug Discovery, arXiv:1502.02072, 2015.
[24] D. Rogers, M. Hahn, Extended-connectivity fingerprints, J. Chem. Inf. Model 50 (2010) 742–754.
[25] S. Rohrer, K. Baumann, Maximum unbiased validation (muv) data sets for virtual screening based on pubchem bioactivity data, J. Chem. Inf. Model. 49 (2009) 169–184.
[26] S. Sahoo, C. Adhikari, M. Kuanar, B. K Mishra, A short review of the generation of molecular descriptors and their applications in quantitative structure property/activity relationships, Curr. Comput. Aided drug Des. 12 (2016) 181–205.
[27] K. Tian, M. Shao, Y. Wang, J. Guan, S. Zhou, Boosting compound-protein interaction prediction by deep learning, Methods 110 (2016) 64–72.
[28] O. Trott, A. Olson, Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading, J. Comput. Chem. 31 (2010) 455–461.
[29] T. Unterthiner, A. Mayr, G. Klambauer, M. Steijaert, J. Wegner, H. Ceulemans, S. Hochreiter, Deep learning as an opportunity in virtual screening, NIPS 27 (2014).
[30] A. Varnek, I. Baskin, Machine learning methods for property prediction in chemoinformatics: quo vadis? J. Chem. Inf. Model. 52 (2012) 1413–1437.
[31] I. Wallach, M. Dzamba, A. Heifets, AtomNet: a Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery, arXiv:1510.02855, 2015.
[32] C. Wang, J. Liu, F. Luo, Z. Deng, Q.N. Hu, Predicting Target-ligand Interactions Using Protein Ligand-binding Site and Ligand Substructures, in: BMC Systems Biology, BioMed Central Ltd., 2015, p. S2.