High throughput screening with machine learning

Oleksandr Gurbych
SoftServe Inc.,
2d Sadova Str., 79021 Lviv, Ukraine
ogurb@softserveinc.com

Maksym Druchok
SoftServe Inc.,
2d Sadova Str., 79021 Lviv, Ukraine
Institute for Condensed Matter Physics
1 Svientsitskii Str., 79011, Lviv, Ukraine
mdruc@softserveinc.com

Dzvenymyra Yarish
SoftServe Inc.,
2d Sadova Str., 79021 Lviv, Ukraine
dyari@softserveinc.com

Sofiya Garkot
Ukrainian Catholic University
17 Svientsitskii Str., 79011, Lviv, Ukraine
garkot@ucu.edu.ua

Abstract

This study assesses the efficiency of several popular machine learning approaches in the prediction of molecular binding affinity: CatBoost, Graph Attention Neural Network, and Bidirectional Encoder Representations from Transformers. The models were trained to predict binding affinities in terms of inhibition constants K_i for pairs of proteins and small organic molecules. First two approaches use thoroughly selected physico-chemical features, while the third one is based on textual molecular representations – it is one of the first attempts to apply Transformer-based predictors for the binding affinity. We also discuss the visualization of attention layers within the Transformer approach in order to highlight the molecular sites responsible for interactions. All approaches are free from atomic spatial coordinates thus avoiding bias from known structures and being able to generalize for compounds with unknown conformations. The achieved accuracy for all suggested approaches prove their potential in high throughput screening.

1 Introduction

Initial stages of drug discovery require localization of a disease cause, understanding of the molecular mechanism, then suggesting and testing drug leads. On a cellular level, a drug molecule binds to a target biomolecule and amends its function. Once the disease target has been identified a list of drug candidates is drafted and screened for the target-candidate binding affinities. Running *in silico* high throughput screening helps shrink the initial pool of *in vitro* tests by eliminating weakly scored candidates. We observe the increasing amount of physico- and bio-chemical data as well as difficulties with massive data arrays processing, which boosts the development and application of ML models to this field.

One of the screening techniques is the assessment of receptor-ligand binding affinities. The receptor-ligand binding can be illustrated with a classic example of competitive inhibition when malonate blocks the activity of succinate dehydrogenase. Succinate dehydrogenase is an enzyme that participates in a reaction chain that provides energy for cells. Its particular function is to conduct dehydrogenation on succinate molecules. However, if the active site of dehydrogenase is occupied by slightly different malonate (one CH_2 group is missing) no catalytic reaction occurs. Dehydrogenase is the receptor, whereas succinate and malonate are ligands competing for the active site of the receptor. The ligand with stronger attraction (binding affinity) wins the competition deciding whether the reaction chain is activated or blocked.

There is a number of studies applying ML techniques to the affinity prediction problem [1–8] and the methods vary from supporting vector machine, decision trees, random forest to deep neural networks. Of course, the list is not complete, it rather references representative papers in the field. It is difficult to compare them directly since they use different metrics, affinity characteristics, and datasets. In particular, the early researches started with only hundreds of affinity samples, while more recent studies advanced with growing availability of data.

In this contribution we test three machine learning algorithms – CatBoost [9], Graph Attention Neural Network, and Bidirectional Encoder Representations from Transformers [10, 11]. All these techniques are used to predict the binding affinity expressed as inhibitory constant K_i. We show that the considered approaches can be used for high throughput screening to save the time and resources on drug discovery stage by prioritizing more promising candidates for further development.

2   Materials and methods

The dataset used in this study is compiled out of four databases – BindingDB [12], PDBbind [13], Binding MOAD [14], BioLIP [15] A plain concatenation of datasets results in a number of duplicated protein-ligand pairs, whereas the inhibition constants K_i need to be inspected for nonphysical extreme concentrations. The inhibition constants are ligand concentrations required to produce half maximum inhibition; they are expressed in nanomoles. In concentration measurements the error increases proportionally to the concentration itself, therefore, it is convenient to work with decimal logarithm of inhibition constants $\log_{10} K_i$. Such a conversion helps balance the error contributions at different concentration ranges. After the whole data pre-processing the collected dataset consists of $\approx$350k unique receptor-ligand pairs with corresponding values of $\log_{10} K_i$.

The dataset consists of receptors (target proteins) represented in FASTA format [16, 17] and the ligands (inhibitors) – in SMILES format [18, 19]. Both formats are one-line sequence notations and do not use spatial coordinates; it was our essential requirement so that the trained models can make predictions on novel receptors and ligands with unknown conformations. Otherwise, if coordinates are known one can develop coordinate-based predictors, like K_{DEEP} approach [4] splitting space into 3D grid and assigning each elementary cube a set of corresponding descriptors.
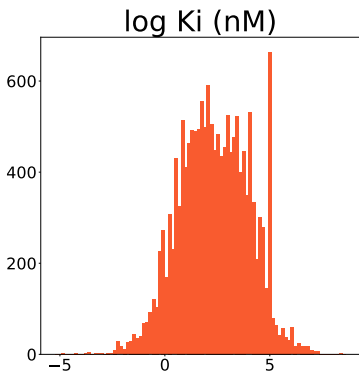

Figure 1: Histogram for binding affinities as decimal logarithm of inhibition constants $\log_{10} K_i$. The bare K_i constants are expressed in nanomoles.

In Fig. 1 we show the distribution of collected $\log_{10} K_i$ values. It has symmetric shape with a slight imbalance for the regions of strong and moderate inhibitors ($\log_{10} K_i < 4$, i.e. $K_i < 10\mu$M), while inactive compounds (with K_i's above 10μM) are less populated. One of possible ways to balance the dataset is to generate augmentations: most popular augmentations over SMILES strings are based on conformation variations or different atom enumerations (see for example Refs. [20, 21]). However, the ligand representations utilized in our study are insensitive to these techniques, thus we decided to keep the dataset in the original form rather than synthetically generating new inhibitors without knowledge about their inhibition constants.

The first machine learning technique of choice is **Cat-Boost Regressor** (CB), a gradient boosting algorithm on decision trees. We examined various physico-chemical features to represent receptors and ligands. As a result, we chose ECFP4 fingerprints for ligands, while receptors were characterized by a set of features based on different properties of amino acids. These properties include charge, van der Waals volume, solvent accessible area, buried area, flexibility, folding index, hydrophobicity, isoelectric point, polarity, refractivity, aliphatic index, numbers of H-bond donors and acceptors, and number of disulfides. All features for each receptor-ligand pair were combined into one vector and regressed towards corresponding $\log_{10} K_i$ value.

The second technique is **Graph Attention Neural Network** (GANN). We designed the architecture of this model to take two inputs: featurized graphs for receptor and ligand. The ligand graphs consist

of nodes (all heavy atoms) and edges (all intramolecular bonds). The ligand node and edge features encode atom and bond types, charges, number of electrons, hybridization, aromaticity, etc. The receptor graphs contain only nodes (amino acids along the protein chain), since all peptide bonds uniting amino acids are identical. The graphs for ligands were crafted with the help of standard featurizers, coming with DGL-LifeSci package (https://lifesci.dgl.ai/), whereas for the receptor graphs we developed a custom featurizer. The receptor node features partially resemble the features used in the CB approach and include charge, flexibility, number of H-bond donors and acceptors, hydrophobicity, solvent accessible area, molecular weight, polarity, van der Waals volume, etc. The graphs are fed into two separate "arms". The receptor arm is based on Graph Attention network [22] from DGL-LifeSci. In the SMILES arm we utilized the AttentiveFPGNN realization of Attentive FP network [23] from the DGL-LifeSci package as well. Outputs of these arms are then concatenated and regressed with a multilayer perceptron to predict a value of $\log_{10} K_i$.

In contrast to the two above approaches purely dealing with physico-chemical properties, **Bidirectional Encoder Representations from Transformers** (BERT) treats receptor and ligand sequences as text strings and constructs representations by jointly conditioning on both left and right context in all layers, which is crucial to fully capture the intricate inter-dependencies in the chemical structures. Our model is built of two pre-trained BERT heads (for FASTA and SMILES) and a multilayer perceptron on top of them. BERT architecture is the same for two inputs: 6 layers, 12 self-attention heads, hidden size 768. FASTA BERT is pre-trained with the masked language modelling objective on the subset of UniProtKB protein sequences database, Swiss-Prot consisting of 506k entries [24]. Pre-trained SMILES BERT weights were taken from Ref. [25], optimized on 155k sequences from publicly available PubChem dataset [26]. We concatenate the averaged final hidden states from FASTA BERT and SMILES BERT, and use it as an input to the perceptron, which then outputs $\log_{10} K_i$ value. To our best knowledge, it is one of the first attempts to utilize Transformer-like models as sole predictors of the binding affinity. Worth mentioning here a recent paper by Morris *et al.* [27] adopting Transformer approach for the affinity prediction, however, their setup is limited to single receptor task, thus embeddings are learned for ligand SMILES only.

3 Results

The dataset was randomly split into train, validation, and test subsets with the 80:10:10 ratios. No specific pre-selection by receptor or ligand families was done. The train and validation parts were used to train the ML models and monitor their accuracy during the train routine, whereas with the test subset we evaluated the final models. Such a unified data split was intended to evaluate all three approaches on the same footing. The best achieved Mean Absolute Error (MAE) scores (in $\log_{10} K_i$ units) on the test subset are CB — 0.60, GANN — 0.64, BERT — 0.51. For example, the MAE of ± 0.6 for $\log_{10} K_i$ corresponds to a factor 4 (multiplier or divider) to the bare K_i value, which is only a bit weaker accuracy than experimentally acceptable factor of ≤ 2. Worth noting that our values of MAE are lower than $0.71 \div 1.03$ values obtained in Ref. [7] on PDBbind and Astex datatsets [13, 28].
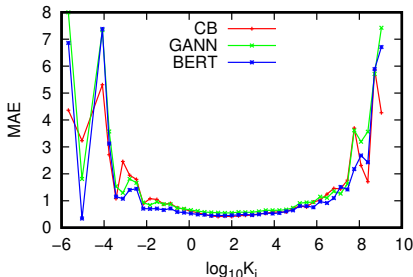

Figure 2: Mean absolute errors as a function of experimental inhibition constants.

We plotted bin-averaged MAE as a function of actual $\log_{10} K_i$ for the three methods of choice (see Fig. 2). One can see high MAE peaks at low and high $\log_{10} K_i$, with a broad well inbetween. We attribute these peaks to a low number of data points – the ML models lack the examples to properly learn extreme cases (see Fig. 1).

In a simpler case of a binary classification "inhibitor vs. non-inhibitor" (either $\log_{10} K_i$ is less or greater than 4) even high MAE values for low $\log_{10} K_i$ (left band of peaks in Fig. 2) still provide a decent chance that compound will be correctly classified as inhibitor. These are the high MAE values for high $\log_{10} K_i$ (peaks on the right) that pose a risk of misclassified type, since the boundary of $\log_{10} K_i = 4$ is closer to this weakly reproduced region. In particular, the inhibitor vs. non-inhibitor classification yields accuracy of 94% for CB and BERT, while GANN provides 93%. These values
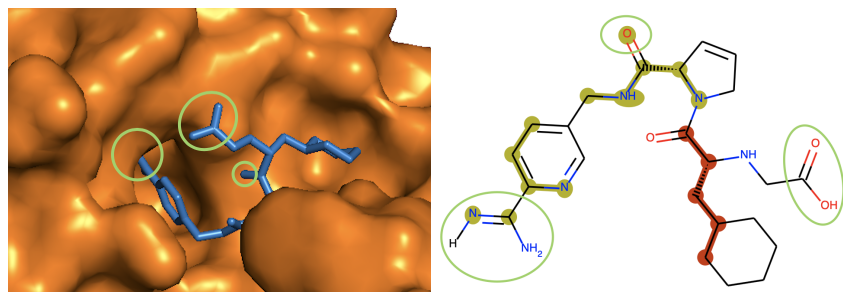
Figure 3: Left – fragment of human thrombin (orange surface) and a ligand (blue sticks) occupying its active site with binding atoms (green ovals) defined experimentally. Right – 2D projection of the ligand with binding atoms highlighted by attention. Green and red spots mark atoms with attention scores $<5\%$ and $\geq 5\%$. The plots are made with PyMol (https://pymol.org) and RDKit (https://www.rdkit.org/) packages.

are comparable with accuracy of 95% reported in a study by Li *et al.* [6] utilizing Bayesian additive regression trees.

We also visualized the attention scores ascribed by BERT model to ligands. In general, the attention scores prioritize elements most relevant to a target task, thus comparing the scores for different ligand fragments we expected to highlight the ones responsible for binding. However, we noticed that attention focus is often spread across ligand molecules instead of particular sites. One of such examples is visualized in Fig. 3 with a complex of human thrombin and a ligand taken from the Protein Data Bank (http://www.rcsb.org/structure/2ANM). This ligand is known to be a strong inhibitor to the human thrombin. The left snapshot shows an active site of thrombin molecule (depicted as an orange surface) occupied by the ligand (blue sticks), the green ovals show the ligand atoms interacting with the receptor (according to the Protein Data Bank). The right plot shows a planar structure of the ligand: again, the green ovals denote the interacting atoms, while red and yellow spots indicate molecular fragments assumed by attention to contribute most to the prediction of affinity of the ligand to thrombin. One can see that the attention is not only focused on the atoms in ovals. It might showcase that binding is actually dependent on a full topology of a ligand rather than some particular sites.

4 Conclusions

This study is aimed to compare three popular machine learning methods in prediction of molecular binding affinity. The following methods were tested: CatBoost, Graph Attention Neural Network, and Bidirectional Encoder Representations from Transformers. For development of the first two we used physico-chemical features in representations of molecules, whereas the third one operated with FASTA and SMILES notations as text sequences.

The above approaches were tested on the same dataset with fixed split onto same train/validation/test subsets in order to maintain equal evaluation conditions. The achieved accuracy for the three considered approaches shows a fair level of prediction, which is only slightly weaker than the acceptable experimental accuracy. Besides the single-number metric Mean Absolute Error we also inspect how the deviations behave along the data range of affinities and map the regions of stronger and weaker confidence. All the considered techniques prove to be lightweight and efficient for high throughput screening.

We also examine and visualize attention scores for ligands within the Transformer approach. We expected them to indicate molecular sites responsible for interactions with proteins. However, the attention often highlights broader parts of molecules, which might be an indication that binding mechanisms involve larger structural fragments.

Though the class of proteins is wide *per se*, we believe that the considered approaches can be extended to a broader range of biomolecules, making them more universal.

References

[1] R.D. King, J.D. Hirst, and M.J.E. Sternberg. Comparison of artificial intelligence methods for modeling pharmaceutical QSARs. *Applied Artificial Intelligence*, 9(2):213–233, 1995.

[2] R.N. Jorissen and M.K. Gilson. Virtual screening of molecular databases using a support vector machine. *Journal of Chemical Information and Modeling*, 45(3):549–561, 2005. PMID: 15921445.

[3] K. Yugandhar and M.M. Gromiha. Feature selection and classification of protein–protein complexes based on their binding affinities using machine learning approaches. *Proteins: Structure, Function, and Bioinformatics*, 82(9):2088–2096, 2014.

[4] J. Jiménez, M. Škalič, G. Martínez-Rosell, and G. De Fabritiis. K_{DEEP}: Protein–ligand absolute binding affinity prediction via 3d-convolutional neural networks. *Journal of Chemical Information and Modeling*, 58(2):287–296, 2018.

[5] H. Öztürk, A. Özgür, and E. Ozkirimli. DeepDTA: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17):i821–i829, 09 2018.

[6] L. Li, C.C. Koh, D. Reker, J.B. Brown, H. Wang, N.K. Lee, H. Liow, H. Dai, H.-M. Fan, L. Chen, and D.-Q. Wei. Predicting protein-ligand interactions based on bow-pharmacological space and bayesian additive regression trees. *Scientific Reports*, 9:7703, 2019.

[7] Y. Li, M.A. Rezaei, C. Li, and X. Li. DeepAtom: A framework for protein-ligand binding affinity prediction. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 303–310, 2019.

[8] Y. Kwon, W.-H. Shin, J. Ko, and J. Lee. AK-score: Accurate protein-ligand binding affinity prediction using the ensemble of 3d-convolutional neural network. *ChemRxiv*, 12015045, 2020.

[9] A.V. Dorogush, V. Ershov, and A. Gulin. CatBoost: gradient boosting with categorical features support. *arXiv*, 1810.11363, 2018.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv*, 1810.04805, 2018.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

[12] X. Chen, M. Liu, and M.K. Gilson. BindingDB: A web-accessible molecular recognition database. *Combinatorial Chemistry and High Throughput Screening*, 4(8):719–725, 2001.

[13] R. Wang, X. Fang, Y. Lu, and S. Wang. The PDBbind database: Collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *Journal of Medicinal Chemistry*, 47(12):2977–2980, 2004.

[14] L. Hu, M.L. Benson, R.D. Smith, M.G. Lerner, and H.A. Carlson. Binding MOAD (Mother Of All Databases). *Proteins: Structure, Function, and Bioinformatics*, 60(3):333–340, 2005.

[15] J. Yang, A. Roy, and Y. Zhang. BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions. *Nucleic Acids Research*, 41(D1):D1096–D1103, 10 2012.

[16] D.J. Lipman and W.R. Pearson. Rapid and sensitive protein similarity searches. *Science*, 227(4693):1435–1441, 1985.

[17] W.R. Pearson and D.J. Lipman. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences*, 85(8):2444–2448, 1988.

[18] D. Weininger. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988.

[19] D. Weininger, A. Weininger, and J.L. Weininger. SMILES. 2. algorithm for generation of unique SMILES notation. *Journal of Chemical Information and Computer Sciences*, 29(2):97–101, 1989.

[20] J. Hemmerich, E. Asilar, and G.F. Ecker. COVER: conformational oversampling as data augmentation for molecules. *Journal of Cheminformatics*, 12(18), 2000.

[21] E.J. Bjerrum. SMILES enumeration as data augmentation for neural network modeling of molecules. *arXiv*, 1703.07076, 2017.

[22] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio. Graph attention networks. *arXiv*, 1710.10903, 2017.

[23] Z. Xiong, D. Wang, X. Liu, F. Zhong, X. Wan, X. Li, Z. Li, X. Luo, K. Chen, H. Jiang, and M. Zheng. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of Medicinal Chemistry*, 63(16):8749–8760, 2020.

[24] A. Bairoch and R. Apweiler. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic acids research*, 28(1):45–52, 2000.

[25] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T.L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A.M. Rush. HuggingFace's transformers: State-of-the-art natural language processing. *arXiv*, 1910.03771, 2019.

[26] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. Shoemaker, P.A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang, and E. Bolton. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Research*, 47:D1102–D1109, 2019.

[27] P. Morris, R.St. Clair, W.E. Hahn, and E. Barenholtz. Predicting binding from screening assays with transformer network embeddings. *Journal of Chemical Information and Modeling*, 60(9):4191–4199, 2020.

[28] M.J. Hartshorn, M.L. Verdonk, G. Chessari, S.C. Brewerton, W.T.M. Mooij, P.N. Mortenson, and C.W. Murray. Diverse, high-quality test set for the validation of protein-ligand docking performance. *Journal of Medicinal Chemistry*, 50(4):726–741, 2007.