2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)

Gene Ontology aided Compound Protein Binding Affinity Prediction Using BERT Encoding

Lingling Zhao
Faculty of Computing
Harbin Institute of Technology
Harbin, China
zhaoll@hit.edu.cn

Peijin Xie
Faculty of Computing
Harbin Institute of Technology
Harbin, China
xiepeijin0729@163.com

Lingfeng Hao
Faculty of Computing
Harbin Institute of Technology
Harbin, China
fanganpai@vip.qq.com

Tiantian Li*
Editorial Office
Geriatric Hospital of Nanjing Medical University
Nanjing, China
litiantian315@126.com

Chunyu Wang*
Faculty of Computing
Harbin Institute of Technology
Harbin, China
chunyu@hit.edu.cn

2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 978-1-7281-6215-7/20/$31.00 ©2020 IEEE DOI: 10.1109/BIBM49941.2020.9312985

***Abstract*—The drug-target binding affinity(DTA) indicates the strength of the drug-target interaction; therefore, predicting DTA by computational approaches can considerably benefit drug discovery by narrowing down the searching space and pruning those drug-target pairs with low binding affinity scores. In the computational methods, feature representation of proteins is one of the most important parts due to its strong influence on the following regression task. This paper introduces the BERT-based language representation to embed the gene ontology annotations, combined with the raw sequence to characterize a protein by fusing its physical structure and human knowledge. We exploit CNN network stacked over full connected layers to learn the prediction of DTA scores in a supervised manner. This framework enhances the feature representation ability, leading to the improvement of the DTA prediction precision. The evaluation on the Davis and KIBA datasets compared to the state-of-the-art baselines demonstrates our feature representation's superiority.**

***Index Terms*—drug-target binding affinity, feature representation, Gene Ontology, semantic representation, BERT**

I. Introduction

Drug discovery and developing is a laborious, time consuming, expensive, and challenging process. The drug-target binding affinity (DTA) indicates the strength of the drug-target interaction; therefore, predicting DTA can considerably benefit drug discovery by narrowing down the searching space and pruning those drug-target pairs with low binding affinity scores. With the advances in computational methods and techniques, especially the application of machine learning in chemical and biological research fields, the computer-aided drug design method is a great opportunity to shorten the process and reduce costs. Many computational methods have been proposed to predict the interaction as well as the binding affinity between compound and protein in recent years, where the binding affinity of new compound-protein pairs is a crucial factor in candidate screening, which costs most resources.

With the quality and quantity increasing of available databases, machine learning and deep learning methods play a more significant role in drug discovery. Several computational models have been proposed for DTA prediction recently. For example, Gradient boosting machines are used in quantitative structure–activity relationship studies for regression and classification problems. The SimBoost [1] employs gradient boosting machines with a novel feature engineering to extract new features from drugs, targets and drug-target pairs in training datasets, and then these features are used as an input in the model to predict the binding affinity for unknown pairs. Regularized least-square (RLS) is another efficient model with kinds of applications. The KronRLS model [2] amends RLS with a Kronecker product of drug-drug and protein-protein to speed up the model training for DTA prediction and achieve promising performance. The Kronecker product part of this method is a similarity-based method in which any similarity measure could be used. Recently inspired by the successful applications in diverse research fields, deep learning approaches are also now intensively used in bioinformatics and cheminformatics. The first deep learning based DTA predict model is DeepDTA [4] which uses Simplified Molecular Input Line Entry System (SMILES), the one-dimensional representation of the drug compound chemical structure, as drug features, while the protein amino acid sequences are used to represent protein features. Another novel deep learning model for DTA is DeepAffinity which represents drugs with SMILES and proteins with structural property sequences. Because of the detailed structural information and higher resolution of sequences, DeepAffinity [3] benefits in the DTA regression tasks.

Feature representation of proteins is one of most important parts due to its strong influence on the following regression task. Besides amino acid sequence, Gene Ontology (GO) annotation is another characterization method for proteins. The GO is a hierarchical vocabulary for annotating gene or gene product functions and their relationships with respect to molecular function (MF), cellular component (CC), and biological process (BP) [5]. Over the years, different families

978-1-7281-6215-7/20/$31.00 ©2020 IEEE

of GO semantic representation methods have been proposed utilizing different principles and types of knowledge to extract semantic evidences [6] including the ontology knowledge graphs and the related corpus. Language representation techniques based on deep learning has much inspired the study in biomedical fields to mine semantics from a large number of available textual resources or self-defined corpora based on the ontology knowledge graphs. Word embedding(WE) tools, such as Word2Vec [7], [8], GloVe [9], fastText [10], map words into vectors independent to the context the words involved and therefore are nonsensitive of the context. To capture the semantic difference of polysemous words, another type of prominent language representation approaches is context-dependent, such as ELMO [11] and Google's BERT [12]. These contextualized representation models proposed since 2018 are designed using whole sentences as context, therefore leading to a better polysemy and nuance handling. By applying the transfer learning technique, bioBERT [13], sci-BERT [14], and NCBI-BERT [15] are trained for specific domains based on BERT on textual database such as PubMed abstract(https://www.ncbi.nlm.nih.gov/pubmed/), showing promising performance compared to the state-of-the-art methods in various biomedical tasks.

In this paper, we introduce the BERT based language representation to embed the gene ontology annotations, combined with the sequence to characterize protein, accordingly the proposed framework can enhance the feature representation ability, leading to the improvement of the DTA prediction.

II. Methods

This section first introduces the datasets we used in subsection II-A and summarizes the entire BERT-encoding GO aided DTA prediction model (GOaidDTA in short) in subsection II-B. Then, we introduce the input representation for compounds and proteins repectively in subsection II-C and the encoding of GO annotation in subsection II-D. Finally, we give the neural network architecture in subsection II-E.

A. Datasets

For the protein feature representation, we utilized the GO and GO annotation data from QuickGO at https://www.ebi.ac.uk/QuickGO/. GO release version 2019-11-27 and GO annotation 2019-11-25 were used. This study evaluated the GOaidDTA using two common benchmark datasets: KIBA dataset (Tang, et al., 2014) and Davis dataset (Davis, et al.,2011). The Davis dataset consists of 442 proteins and 68 compounds forming 30,056 drug-target pairs while the KIBA dataset contains 229 proteins and 2111 compounds forming 118,254 drug-target pairs.Table 1 provides the statistics of these datasets.

It can be observed that these datasets differ significantly from the interaction distribution over the proteins or compounds, all of these datasets have skewed distribution.

B. Overview of the GOaidDTA

Figure 1 shows the overview of GOaidDTA. It takes the SMILES strings of compounds and the GO annotation as well as FASTA sequence of proteins as the input. With the help of pre-trained BERT model, it encodes term names composed of GO annotation into feature matrix as internal representation. Then it exploits Convolutional Neural Network(CNN) and fully connected (FC) networks to learn from known compound-protein pairs provided by Davis and KIBA. Finally, it output the binding affinity for new pairs. The design of GOaidDTA includes two main steps: 1. The pre-training of BERT model to encode each token in the term names of GO annotation for proteins; 2. The training and predicting of the binding affinity based on the designed deep neural network.

C. Input representation

Both the protein sequence and compound SMILES are encoded by one-hot coding, while GO terms annotating proteins are embedded by BERT model. Considering the SMILES strings and protein sequences are both in different lengths, we truncate them into fixed lengths for effective representation. As shown in [4], we also choose fixed lengths of 85 for SMILES and 1200 for protein sequences for Davis, but 100 for SMILES and 1000 for protein sequences respectively for KIBA. Then no sequences are longer than the maximum truncated length, but shorter ones are padding with zeros to the same length.

D. Encoding Gene ontology Annotation

To capture the GO annotation's semantic information for each protein, we adopted a pre-trained BERT model to map each GO term into a high-dimension feature vector. Regarding the WE tool input, Onto2vec and OPA2vec exploit term description as model input and convert the words in the description into a word vector sequence. Unlike the models mentioned above, GOaidDTA takes the term name as the WE model's input to represent the term itself. The reason is that the term names highly generalize the semantics of terms, thus being more concise compared to the description of terms. Also, considering that the semantics of words expressed by any WE technology inevitably has some deviations; as a result, more errors may take place due to the increasing number of words. However, it should be noted that there are disadvantages when taking terminology names as input. For context-sensitive WE methods such as BERT, terminology names provide less context information than terminology descriptions since terminology descriptions provide terminology interpretation.

In contrast to the state-of-the-art GO representation methods based on the word2vec, we exploited a NCBI-blueBERT model specifically-trained for the biomedical domain as the semantic embedding approach. Specifically, given a sentence $S = \{w_1, w_2, \ldots, w_N\}$ including N words, the BERT model first uses wordpiece tokenization, which means one word may break into several pieces. The aim of tokenization is to achieve a balance between vocabulary size and out-of-vocabulary words. In this way, BERT stores only 30K vocabulary words, yet very rarely encounters out-of-vocabulary words. Besides, two special tokens are added for each sentence, a start token [CLS] and an end token [SEP]. Finally, the sentence S is represents by a set of tokens $S = \{c_1, c_2, \ldots, c_N\}$. The input

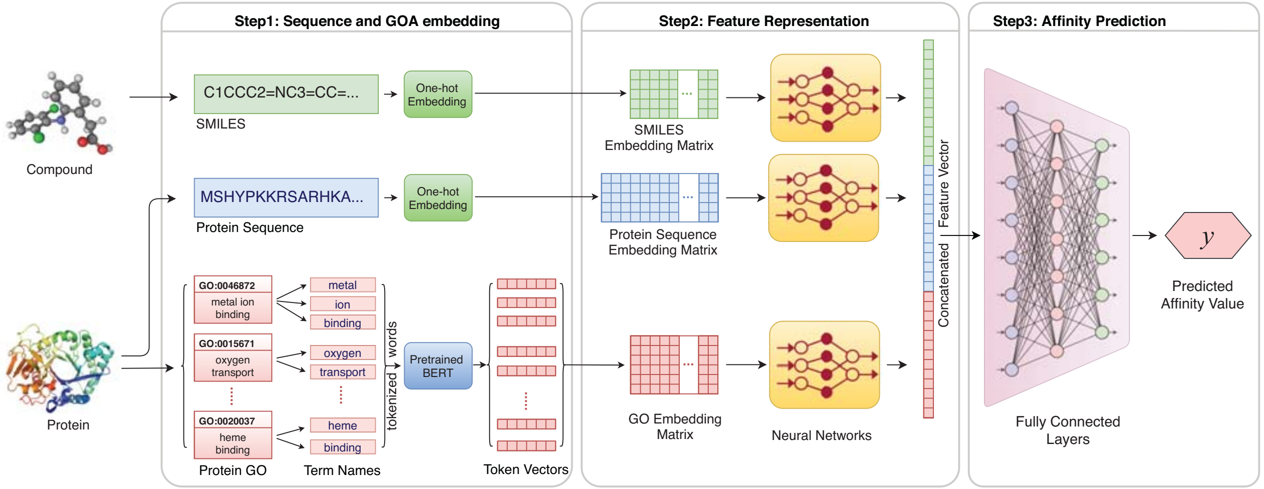

Fig. 1. Framework of GOaidDTA

TABLE I
THE DETAILED STATISTICS OF THE DATASETS, CONTAINING THE NUMBER OF PROTEINS, COMPOUNDS, INTERACTIONS AND AVERAGE NUMBER OF INTERACTIONS PER EACH PROTEIN AS WELL AS PER COMPOUND.

Dataset	No. Proteins	No. Compounds	No. Interaction	No. interactions per protein	No. interactions per compound	No. inactive interactions	No. non-inactive interactions
Davis	442	68	30056	68	442	2457	27554
KIBA	229	2111	118254	516.393	56.018	22729	93426

representation e_i for each token c_i by summing the corresponding token embedding, segment embedding, and position embedding. Input representations are then fed into a set of transformer blocks to obtain context-aware representations.

Through the word embedding, each term name is converted into a two-dimension vector $T_i = [v_1, v_2, \dots, v_{n_i}]$, where v_i denotes the vector converted from the $i-th$ token(a word or a word piece) and n_i is the number of tokens after tokenization. Then, all the terms annotating a protein form a semantic feature matrix $T = [T_1, T_2, \dots, T_M]$ with a width of 768, which is fixed by the BERT model. Further, to keep the same size for all proteins, the complimentary data obtained by filling zeros are exploited to obtain the input of the same size. Here, we selected 256 as the height of the semantic feature matrix. The final semantic feature matrix can be formulated as $T' = [T_1, T_2, \dots, T_M, [00\dots0], \dots, [00\dots0]]$.

E. The Neural Network Architecture

In this study, we developed a CNN-based neural network to capture both the spatial properties of the heterogeneous presentation of proteins, and this network consists of three blocks, a CNN block, a global-pooling layer, and a fully connected layer block. The whole architecture is described in Figure 2. The CNN block is an ensemble of CNN that combines different sizes of convolution kernel and then concatenates the spatial features together. The global pooling block was linked on top of the CNN block by integrating the global maximum pooling and the global average pooling to the given features. On top of the global pooling layer, protein and drug features from three separate CNN blocks were concatenated and fed into three fully connected layers to produce the predicted affinity score.

III. Results

A. Drug-Target Affinity Prediction

This study further evaluated our proposed method in a DTA prediction task using two benchmark data sets, the Davis [19] and KIBA [20] data set. To quantify the prediction performance in an unbiased manner, we carried out a 5-fold cross-validation experiment; meantime, the proposed method was compared with the state-of-the-art DTA prediction models using the Davis and KIBA datasets, DeepDTA and WideDTA, and two classic models SimBoost and KronRLS. The performance of all models was measured by five evaluation metrics; 1) Concordance Index [16] (CI), 2) Mean Squared Error (MSE), 3) r_m^2 index, 4) the Pearson correlation coefficient (R), 5) Area under the precision-recall curve (AUPR).

CI indicates the ranking performance of the models that output continuous values and can be calculated by the equation (1).

$$CI = \frac{1}{Z} \sum_{\delta_i > \delta_j} h\left(b_i - b_j\right) \tag{1}$$

where Z denotes the normalization constant and b_i, b_j represents the prediction value for the larger and small affinity δ_i and δ_J respectively. $h(\cdot)$ is the step function:

$$h(m) = \begin{cases} 1, & \text{if } m > 0 \\ 0.5, & \text{if } m = 0 \\ 0, & \text{if } m < 0 \end{cases} \tag{2}$$

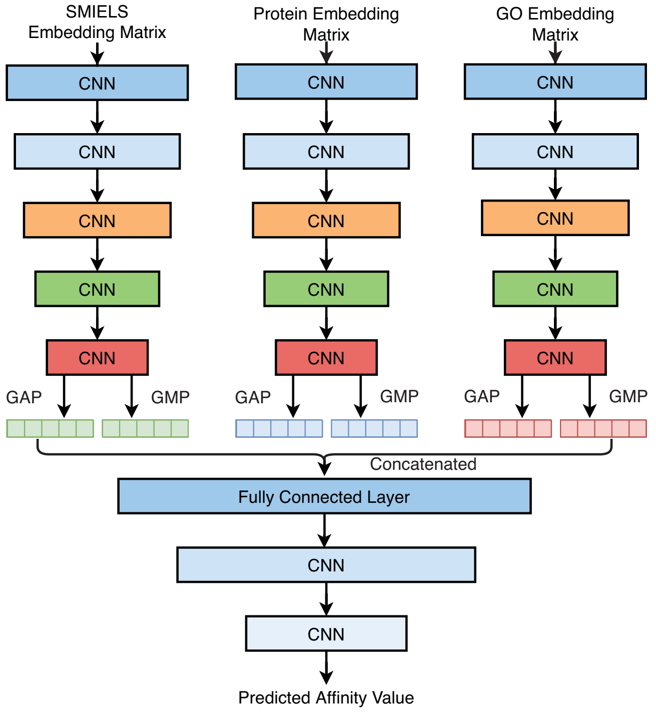

Fig. 2. The Neural Network Architecture

MSE is a well-defined metric to measure how close the fitted line in regression task. If $\hat{y}_k$ is the predicted value of the k-th drug-target pair, and y_k is the corresponding binding affinity, then the MSE estimated over n drug-target pairs is defined as:

$$\mathrm{MSE} = \frac{1}{n}\sum_{k=1}^{n}(y_k - \hat{y}_k)^2 \tag{3}$$

r_m^2 index can be used to evaluate the external predictive performance of DTA models where $r_m^2 > 0.5$ on the test set means the model is an acceptable model. The r_m^2 is described in (4), where r^2, r_0^2 are the squared correlation coefficients with and without intercept, respectively [17], [18].

$$r_m^2 = r^2 \times (1 - \sqrt{r^2 - r_0^2}) \tag{4}$$

The Pearson correlation (R) measures the linear correlation between the ground truth value and the predicted value. The range of value for R is between -1 and 1. The AUPR score is often utilized in binary prediction by averaging the precision across all recall values. We converted the binding affinity of datasets to binary datasets based on different threshold (7.0 for davis dataset and 12.1 for KIBA dataset).

In this paper, we run the program on the computer of Intel(R) Xeon(R) CPU E5-2643 CPU, NVIDIA GeForce GTX 1080 Ti, and 128G DDR4 RAM. We implemented our method using python 3.6, TensorFlow, and Keras. The hyper-parameters of our neural network are reported in Tabel II.

Four baselines are selected to compare with our model, including SimBoost, KronRLS, DeepDTA and WideDTA. SimBoost is a gradient boosting machine based method that depends on feature engineering to represent drug-target interactions; KronRLS algorithm utilizes only 2D based compound similarity-based representations of the drugs and Smith-Waterman similarity representation of the targets; both the DeepDTA and its evolutionary version WideDTA employ same CNN architecture to capture the interaction features from text-information of proteins and ligands, their difference mainly lies in the inclusion of new hand-crafted features such as Maximum Common Substructures(LMCS) and protein domain and/or motif(PDM) information in the WideDTA.

TABLE II
THE HYPER-PARAMETERS OF OUR NEURAL NETWORK

Parameters	Range
Number of filters	[32, 64, 96, 128, 256]
Filter length (compounds)	3
Filter length (protein sequence)	7
Filter length (GO)	5
hidden neurons	1024,1024,512
batch size	256
optimizer	Adam
learning rate	0.001

TABLE III
COMPARISON OF OUR APPROACH AND BASE COMPARABLE APPROACHES ON THE KIBA DATASET

Model	CI	MSE	r_m^2	R	AUPR
SimBoost	0.836	0.222	0.629	-	0.760
KronRLS	0.782	0.411	0.342	-	0.635
DeepDTA	0.863	0.194	0.673	0.848	0.788
WideDTA	0.875	0.179	-	-	-
GOaidDTA	0.876	0.179	0.706	0.868	0.782

TABLE IV
COMPARISON OF OUR APPROACH AND BASE COMPARABLE APPROACHES ON DAVIS DATASET

Model	CI	MSE	r_m^2	R	AUPR
SimBoost	0.872	0.282	0.644	-	0.709
KronRLS	0.871	0.379	0.407	-	0.661
DeepDTA	0.878	0.261	0.630	0.846	0.714
WideDTA	0.886	0.262	-	-	-
GOaidDTA	0.891	0.229	0.654	0.850	0.717

To evaluate the proposed approach, we applied it to Davis and KIBA datasets. The comparation to the state-of-the-art baselines demonstrates the superiority of our feature representation.On KIBA dataset,it is apparent from the Table III that GOaidDTA becomes the best-performing one on the measures of CI scores.And the number of r_m^2 grows from 0.673 to 0.706 compared with the first-rate model DeepDTA. Also Table IV shows the achieved results on the Davis dataset. GOaidDTA gets a better result in all measures compared to the 4 baselines. Remarkably on CI index, our approach gets to 0.891, while the rest never reached 0.89. As for the evaluation metric MSE scores, GOaidDTA also outperforms DeepDTA from 0.261 to 0.229,which is a great leap compared with the improvement from DeepDTA(0.262) to WideDTA(0.261). Consequently,our model that uses GO annotation is an effective approach for

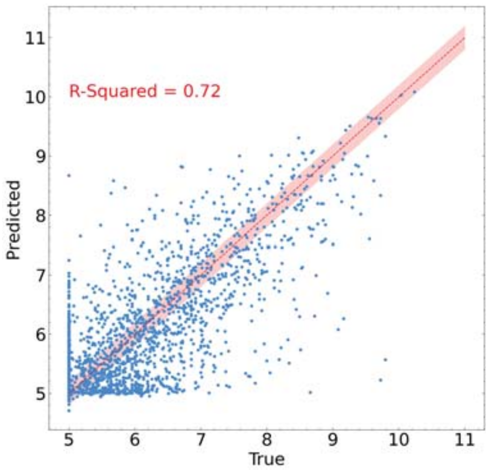

Fig. 3. Predictions from our method against measured (real) binding affinity values for Davis dataset

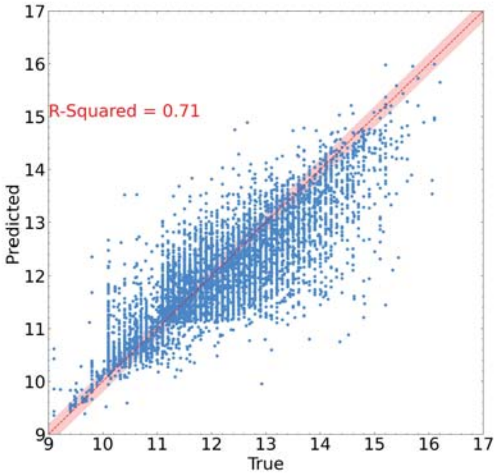

Fig. 4. Predictions from our method against measured (real) binding affinity values for KIBA dataset

drug target binding affinity prediction.Simultaneously the improvement of the other 3 evaluation metrics shows the advantages of accuracy, acceptability and predictability compared with other models.

Figure 3 and 4 show the regression line for the predicted by our model against ground truth binding affinity values on Davis and KIBA test datasets. If the prediction binding affinity is close to the truth value, the points would form a diagonal line on the graph. We observe that especially for KIBA dataset, the density is high around the diagonal line.

IV. Conclusion

We have presented a deep learning architecture to predict drug-target binding affinity, which we refer to as GOaidDTA. GOaidDTA combines GO annotation, amino acid sequence of protein and SMIELS of compound. We use three separate 5-layer convolution blocks to learn representations for drug, protein sequences and GO, respectively. Finally, the affinity prediction task is completed by a 4-layer fully connected network. Experimental evaluation shows that our model achieves a more accurate prediction. In future work, we will explore the interpretability of our model by adopting the attention mechanism on the amino acid sequence of protein and GO annotation.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (NSFC, Grant no. 61872114 and 61305013). The funders have no role in study design, data collection, data analysis, data interpretation, or writing of the manuscript. We thank all the reviewers for their valuable comments and suggestions.

References

[1] He T, Heidemeyer M, Ban F, et al. SimBoost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines[J]. Journal of cheminformatics, 2017, 9(1): 1-14.

[2] Pahikkala T.et al. . (2014) Toward more realistic drug–target interaction predictions. Brief. Bioinformatics, 16, 325–327.

[3] Karimi M, Wu D, Wang Z, et al. DeepAffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks[J]. Bioinformatics, 2019, 35(18): 3329-3338.

[4] Öztürk H, Özgür A, Ozkirimli E. DeepDTA: deep drug–target binding affinity prediction[J]. Bioinformatics, 2018, 34(17): i821-i829.

[5] M. Ashburner, C. Ball, J. Blake, D. Botstein, H. Butler, J. Cherry, A. Davis, K. Dolinski and S. Dwight et al., Nat. Genet, 2000, 25, 25–29

[6] Zhao C, Wang Z. GOGO: An improved algorithm to measure the semantic similarity between gene ontology terms[J]. Scientific reports, 2018, 8(1): 1-10.

[7] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)

[8] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems. 3111–3119.

[9] Pennington J, Socher R, Manning C D. Glove: Global vectors for word representation[C]//Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014: 1532-1543.

[10] Bojanowski P, Grave E, Joulin A, et al. Enriching word vectors with subword information[J]. Transactions of the Association for Computational Linguistics, 2017, 5: 135-146.

[11] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pages 2227–2237.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 .

[13] Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining[J]. Bioinformatics, 2020, 36(4): 1234-1240.

[14] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3615–3620, Hong Kong, China, November 2019. Association for Computational Linguistics

[15] Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. arXiv preprint arXiv:1906.05474.

[16] Gonen, Mithat, and Glenn Heller. “Concordance Probability and Discriminatory Power in Proportional Hazards Regression.” Biometrika, vol. 92, no. 4, 2005, pp. 965–970.

[17] Roy, Partha Pratim, et al. “On Two Novel Parameters for Validation of Predictive QSAR Models.” Molecules, vol. 14, no. 5, 2009, pp. 1660–1701.

[18] Roy, Kunal, et al. “Some Case Studies on Application of ‘r(m)2’ Metrics for Judging Quality of Quantitative Structure-Activity Relationship Predictions: Emphasis on Scaling of Response Data.” Journal of Computational Chemistry, vol. 34, no. 12, 2013, pp. 1071–1082.

[19] Davis, M. I., Hunt, J. P., Herrgard, S., Ciceri, P., Wodicka, L. M., Pallares, G., et al. (2011). Comprehensive analysis of kinase inhibitor selectivity. Nat. Biotechnol. 29, 1046.

[20] Tang, J., Szwajda, A., Shakyawar, S., Xu, T., Hintsanen, P., Wennerberg, K., et al. (2014). Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. J. Chem. Inf. Modeling 54, 735–743.