Identifying Structure-Property Relationships through SMILES Syntax Analysis with Self-Attention Mechanism

Shuangjia Zheng[1], Xin Yan*[1], Yuedong Yang[2] and Jun Xu*[1,3]

[1]Research Center for Drug Discovery, School of Pharmaceutical Sciences, Sun Yat-sen University, 132 East Circle at University City, Guangzhou 510006, China

[2]National Supercomputer Center in Guangzhou, Sun Yat-sen University, Guangzhou 510006, China

[3]School of Computer Science & Technology, Wuyi University, 99 Yingbin Road, Jiangmen 529020, China

*To whom correspondence should be addressed.

Jun Xu：junxu@biochemomes.com or xujun9@mail.sysu.edu.cn

Abstract

Recognizing substructures and their relations embedded in a molecular structure representation is a key process for structure-activity or structure-property relationship (SAR/SPR) studies. A molecular structure can be either explicitly represented as a connection table (CT) or linear notation, such as SMILES, which is a language describing the connectivity of atoms in the molecular structure. Conventional SAR/SPR approaches rely on partitioning the CT into a set of predefined substructures as structural descriptors. In this work, we propose a new method to identifying SAR/SPR through linear notation (for example, SMILES) syntax analysis with self-attention mechanism, an interpretable deep learning architecture. The method has been evaluated by predicting chemical property, toxicology, and bioactivity from experimental data sets. Our results demonstrate that the method yields superior performance comparing with state-of-the-art models. Moreover, the method can produce chemically interpretable results, which can be used for a chemist to design, and synthesize the activity/property improved compounds.

Keywords

Virtual screening, Deep learning, Self-attention mechanism, Molecular descriptor, SAR/SPR

1. Introduction

Illuminating the relationship between molecular structures and chemical properties or bioactivity has always been a topic of significant interest in the chemical community.[1] However, this relationship is progressively difficult to clarify based on empirical measurements and heuristic rules with the explosive increase of experimental data.

Cheminformatics has been an area of active research by predicting the molecular properties or bioactivity from molecular structures with the aids of high performance computers and machine learning methods.[2] In the recent decades, with the emergence of deep learning methods,[3] machine learning has gathered increasing attention from the scientific community. Data-driven analysis has become a routine procedure in many chemical and pharmaceutical applications, including virtual screening,[4-5] chemical property prediction,[6-7] and *de novo* molecular design.[8-10] In many such applications, machine learning has shown strong potential to compete with or even outperform conventional approaches.

The Merck Molecular Activity Challenge sparked a trend of training deep learning networks with molecular fingerprints and other descriptors. The winning team used a multi-task model with a large number of precomputed molecular descriptors that improves upon a random forest baseline by a margin of 15%.[4] By using the same training strategy, Andreas and colleagues proposed the most accurate results of the Tox21 challenge for toxicity prediction.[11] Although many works demonstrated that massive multi-task networks trained with numerous molecular descriptors can provide significant boosts in the predictive power of conventional models for virtual screening and property prediction,[12-13] their inherent 'black-box' nature has persistently drawn considerable criticisms in the modeling community. Such models make the relationship between properties and structures more difficult to interpret.

Therefore, there are growing interests from both chemistry and machine learning field to directly learn molecular properties of compounds according to the atomic topology of a molecule, instead of predefined fingerprints or descriptors. Duvenaud and co-workers presented "neural fingerprints" (NFP) trying to extract data-driven features

instead of hand-crafted features from molecules.[14] The architecture was based on generalizing the fingerprints such that it can be learned via a back-propagation algorithm. Later, Kearnes and co-workers presented molecular graph convolutions, a deep learning system using a representation of small molecules as undirected graphs of atoms.[15] Following this idea, other researchers proposed several improved graph convolutional network (GCN) for dynamically extracting molecular feature vector to predict target properties.[16-18] Despite the considerable predictive performance, such congenital deficiencies of GCN like limited information propagation across the graph and unintuitive feature extraction indicates that the model still has space to improve.

Apart from graph representation, with the prevalence of generative model, researchers pay closer attention to molecular language representation, or molecular linear notation. Many unsupervised learning with diverse generative models were utilized for *de novo* molecular design.[9, 19-21] Most of them employed SMILES (the most popular molecular linear notation) as input to generate new molecules with specific properties. These studies proved that a molecular linear notation can be directly used in SAR/SPR studies. Compared to CT-based approaches, a structure linear notation input to sequence-based network is easier. However, to our best knowledge, there is no previous studies that directly input SMILES to deep learning model for biochemical properties prediction. We have two reasons to directly use SMILES in deep learning process:

(1) Substructures relations are embedded in SMILES, and represented in its syntaxes to be discovered *per se*.

(2) The SMILES notation fits more naturally as deep learning models require molecular sequences of various lengths as input.

We propose to apply self-attention mechanism with improved bi-directional long short-term memory (BiLSTM) model for high interpretability and convenient modeling as well as considerable predictive performance. The self-attention mechanism has been originally developed for a sentence analysis in computer science.[22] It has also been successfully applied in many fields, especially in nature language process (NLP) area such as machine translation,[23] sentiment classification and textual entailment.[22, 24]

The self-attention mechanism allows us to explicitly elucidate SAR/SPR for chemists. As a type of ‘end-to-end’ approach, we directly train the models with SMILES notations without any predefined assumptions to avoid biases. While the self-attention mechanism is learning from a large size of chemical structure data privileged substructures associated to the activity/property can be naturally picked up by weighting the critical parts (substructures) represented by SMILES syntaxes.

Thus, we demonstrate that this method can not only outperform most of state-of-the-art models on molecular property prediction, but identify important functional groups that are directly related to concerned activity/property, such as stability, toxicity or bioactivity.

2. Methods and materials

2.1 Molecular representation

Atomic connectivity in a molecule can be described by a SMILES notation, a text sequences.[25] This representation encodes the topological information of a molecule based on common chemical bonding rules. For example, the 6-carbon ringed molecule benzene can be encoded as ’c1ccccc1’. Each lowercase ‘c’ represents an aromatic carbon atom, and ’1’ the start and closing of a cycle/ring, hydrogen atoms can be deduced via simple rules. In a conventional network-based learning approach, each letter in a SMILES notation was sent to recurrent neural networks (RNNs) for training. This process cannot reflect the features of chiral centers, charges, cyclic connection descriptors. To preserve these critical chemical features, we train RNNs with normal tokens and combined tokens (the SMILES notations grouped by a pair of square brackets []).

2.2 Word embedding process

In one-hot encoding approach, each molecule is represented by a number of token vectors. All token vectors have the same number of components. Each component in a

vector is set to zero except the one at the token's index position. This data storage protocol requires great memory space and introduces inefficiency. Therefore, we employ word embedding algorithm [26-27] in SAR/SPR studies. To use this algorithm, each token vector is compressed to an information-enriched vector, and transformed from a space with one dimension per word to a continuous vector for unsupervised learning. This data representation can record the "semantic similarity" of every token. This process expedites the convergence of a training. In summary, each molecular structure is converted into a SMILES string, which is then encoded into a one-hot matrix, and then is transformed to a word embedding matrix at the embedding layer. Suppose we have a molecular, which has n tokens, represented in a sequence of molecular embeddings:

$$M = (t_1, t_2, \ldots t_n) \qquad (1)$$

where t_i is a vector standing for a d dimensional token embedding for the i-th token in a molecule. M is thus a molecule represented as a 2D matrix, which concatenates all the token embeddings together. M should have the shape n-by-d.

2.3 BiLSTM model & self-attention mechanism

Our proposed model consists of three parts. The first part is a bi-directional LSTM. The second part is the self-attention mechanism, which provides a set of summation weight vectors for the LSTM hidden states. The third part is a fully connected layer for property prediction. Figure 1 shows the proposed molecular attentive embedding model that was used in this work.

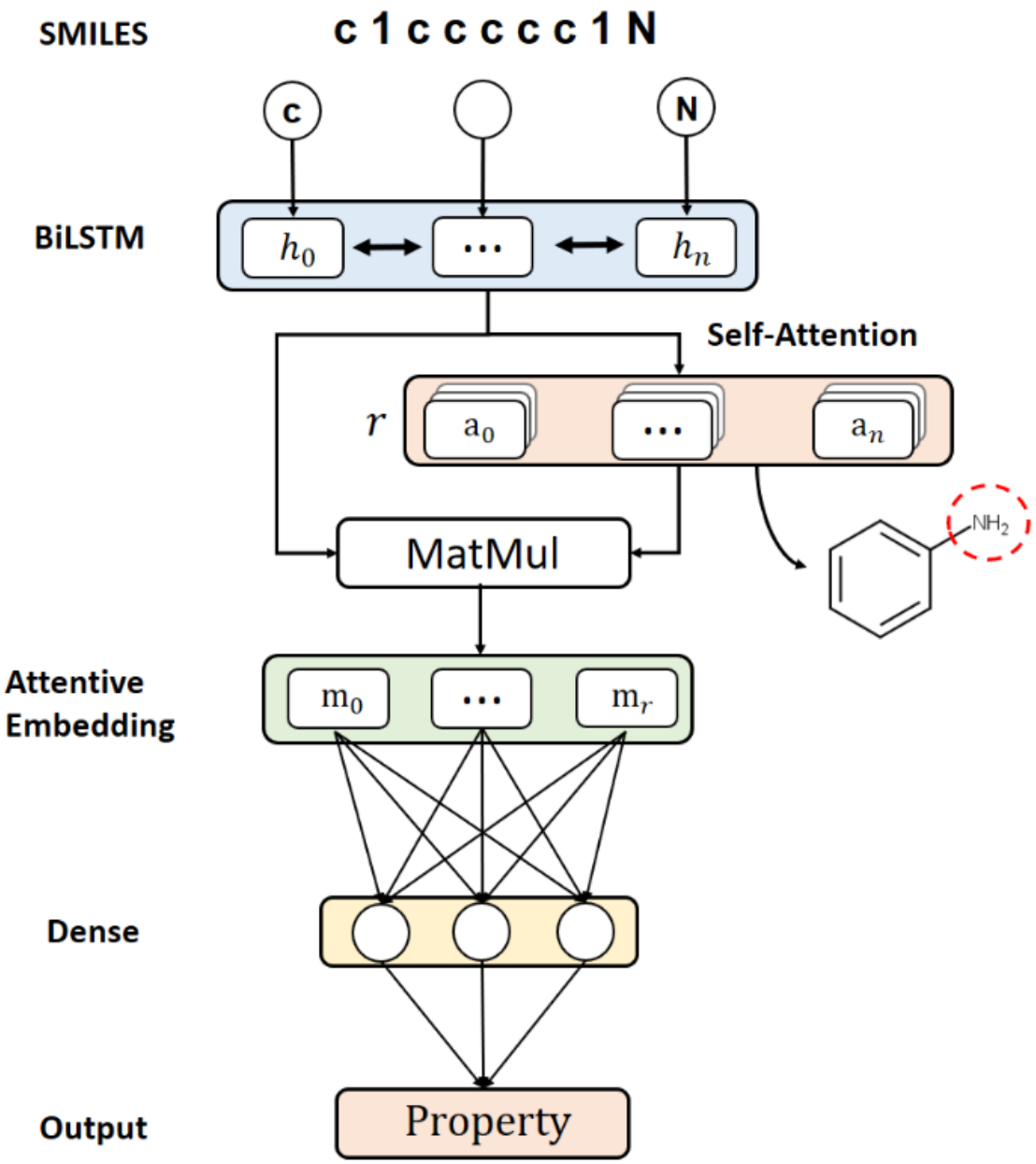

Figure 1. Model architecture. There are three fundamental components in our model: a BiLSTM structure, a self-attention metric and a fully connected layer.

In Figure 1, each SMILES string is converted to a two-dimensional embedding matrix, which has the shape of *n*-by-*d.* Token vectors in the molecular matrix *M* are independent to each other. To gain some dependency between adjacent tokens within a a molecule, a bi-directional LSTM is used to process a molecule:

$$\overrightarrow{h_i} = \overrightarrow{LSTM}(t_i, \overrightarrow{h_{i-1}}) \qquad (2)$$

$$\overleftarrow{h_i} = \overleftarrow{LSTM}(t_i, \overleftarrow{h_{i+1}}) \quad (3)$$

$\overrightarrow{h_i}$ is concatenated with $\overleftarrow{h_i}$, and a hidden state h_i is obtained to replace token embedding t_i, and thus h_t becomes a more information-enriched vector which gains some dependency between adjacent tokens in a molecule. For simplicity, we note all h_i in every time step *i* as *H*.

$$h_i = (\overrightarrow{h_i}, \overleftarrow{h_i}) \quad (4)$$

$$H = (h_0, h_1 \ldots h_n) \quad (5)$$

If the hidden unit number for each uni-directional LSTM is set as u, the shape of *H* would be *n*-by-2*u*.

The next goal is to know which part of the molecule is explicitly considered by the prediction model. In other words, we want to identify the relationship between tokens and concerned property/activity. We achieve this by introducing self-attention mechanism. The attention mechanism takes the whole LSTM hidden states *H* as input, and outputs a vector of weights *a*:

$$a = softmax(w_2 \tanh(W_1 H^T)) \quad (6)$$

where W_1 is a weight matrix with a shape of d_a-by-2u, and w_2 is a vector of parameters with size d_a, which is an adjustable hyper-parameter. As a result, the annotation vector *a* has a size of *n*, which is equal to the length of *H*. The *softmax* function ensures all the computed weights sum up to 1. The LSTM hidden states *H* are summed according to the weight provided by *a* to get a vector representation *m* of the input molecule. Intuitively, the attention coefficients directly determine which parts of the molecule are associated with the activity/property by highlight the related tokens' latent vectors h_ts in *m*.

Note that this vector representation usually focuses on a specific component of a molecule, like a special oxygen atom or a triple bond. However, there might be multiple components in a molecule that together results in a special function, like being toxic or being active to a specific target. (For example, 'c1ccccc1' and 'CCCC' are both hydrophobic groups.) We need multiple vector representations that focus on different functional group of the molecule. Thus, we need to perform multiple attentions.

Here, we extend the w_2 to a r-by-d_a matrix, note it as W_2, and the resulting annotation vector a becomes annotation matrix A. r is also an adjustable hyper-parameter. Formally,

$$A = softmax(W_2 \tanh(W_1 H^T)) \qquad (7)$$

We compute the r weighted sums by multiplying the annotation matrix A and LSTM hidden states H, the resulting matrix is the self-attentive molecular embedding:

$$M_a = AH \qquad (8)$$

where M_a is a self-attentive molecular embedding that contains the latent relationship between tokens and targeted chemical property. The size of M_a is r-by-$2u$. Finally, M_a is combined with a fully connected layer for property prediction.

2.4 Model Training and Evaluation

Both of the self-attention mechanism and BiLSTM models were implemented with Pytorch[28], an open-source library for deep learning. Classification tasks were evaluated by the area under the receiver operating characteristic curve (AUC) or the accuracy of model, and all regression tasks were evaluated by the mean squared errors (MSE). The AUC indicates classification (or ranked order) performed by measuring the area under the curve of true-positive rate versus the false-positive rate, AUC value of 1.0 means perfect separation whereas a value of 0.5 implies random separation. The MSE represents the error between the predicted value and the true value. The lower MSE values means better predictive performance. We used 5-fold cross-validation, where each model used 70% of the data for training, 10% for validation (parameters selection), and 20% as a test set. We repeated all the experiments three times with different choices of random seed. Self-attention mechanism BiLSTM (SA-BiLSTM) models were trained for 10–20 M steps using the Adam optimizer[29] with learning rate [0.001, 0.003, 0.005, 0.01], clip [-0.5, -0.3, 0, 0.3, 0.5] and batch size [64, 128, 256, 512]. We used grid search for parameter optimization.

To establish a baseline without the self-attention, we also trained conventional bi-directional LSTM model with the same data preprocessing. To compare with the SA-BiLSTM model, we used identical inputs and fold assignments. The hyper-parameters

were chosen by same grid search method as SA-BiLSTM model.

2.5 Visualization of the Attention

One of the most useful aspects of the self-attention mechanism is that the obtained attention weights allow us to identify SAR/SPR and interpret what the model has learned from the data. The elucidating the molecular embedding is quite straight forward due to the existence of annotation matrix A. Each row in the attentive embedding matrix M_a corresponds annotation vector a^i. Each element in this vector corresponds to the LSTM hidden state of a token on that position contributes to.

Thus, we can draw a heat map and know which part of molecular tokens are more important in one specific task. In addition, by summing up overall the annotation vectors and then normalizing the resulting weight vector, we can figure out which tokens the embedding takes into account, and which ones are skipped by the model. Then, we convert the SMILES representation to graph representation and maintain the tokens' weights that associated to each atom in molecules. The visualization procedure of attention weights is shown in Figure 2. The higher the attention coefficient, the darker the color.

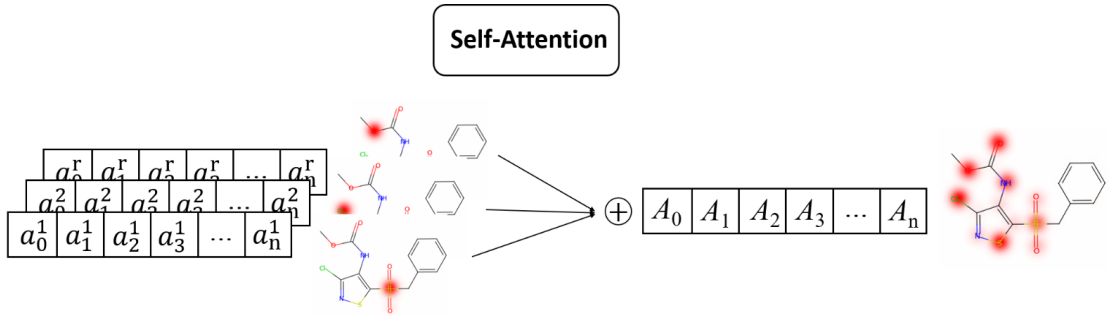

Figure 2. The visualization procedure of attention weights.

2.6 Benchmark datasets

In this work, we evaluate our model by applying it to three different learning tasks: chemical properties prediction, toxicity prediction and bioactivity prediction. For each task, we selected serval representative data sets for comparison. And we also selected

different task types (classification and regression) to better demonstrate the predictive performance. For chemical properties prediction, we chose three datasets: (1) aqueous solubility dataset of 1,144 molecules with their corresponding intrinsic solubilities in log10 (mol/L) as measured by Delaney and co-workers [30]; (2) a subset of the photovoltaic efficiency of 20,000 organic molecules previously used by Duvenaud and co-workers[14]; (3) COMDECOM stability dataset of 9,746 molecules measured in DMSO or H2O solutions for up to 105 days.[31] For toxicity prediction, we chose Tox21, a public database measuring toxicity of 7,831 compounds on 12 targets. This dataset has been used in the 2014 Tox21 Data Challenge.[32] For bioactivity prediction, we used (1) DUD-E database, which has 102 targets and comprising 22,886 active compounds and 50 chemically similar decoys for each active compound.[33] (2) a dataset of the half-maximal effective concentration (EC_{50}) *in vitro* of 10,000 molecules against a sulfide-resistant strain of *P. falciparum*, as measured by Gamo and co-workers[34] (3) HIV dataset introduced by the Drug Therapeutics Program (DTP) AIDS Antiviral Screen, which tested the ability to inhibit HIV replication for over 41,127 compounds. A summary of the three types of datasets is shown in Table 1.

Table 1. Details for dataset groups. The percentage of positive compounds are reported as mean of each task.

Category	Dataset	Task	Task Type	Compounds	%Positive
Chemical property	Solubility	1	Regression	1,144	/
	Photovoltaics	1	Regression	29,978	/
	Stability	1	Classification	9,746	33.9
Toxicology	Tox21	12	Classification	11,764	4.7
Bioactivity	DUD-E	10	Classification	122,886	2
	HIV	1	Classification	41,128	3.5
	Drug efficacy	1	Regression	10,000	/

Eg. Properties, or output labels, are either 0/1 for classification tasks, or floating-point numbers for regression tasks

3. Experiments and discussion

3.1 Chemical properties prediction

3.1.1 Aqueous solubility & photovoltaic efficiency

Figure 3 compares single task (ST) language-based models results to published results on these datasets. The best-case MSE of the self-attentive BiLSTM (SA-BiLSTM) model on photovoltaics and solubility datasets were 1.18 ± 0.06 percent units and 0.38 ± 0.04 log M units. These results were much better than the original results reported by Duvenaud and co-workers (1.43 ± 0.09 and 0.52 ± 0.07) with the same fold assignments and were comparable to multiple task (MT) GCN model reported by Kearnes and co-workers (0.46 ± 0.08 and 1.10 ± 0.06).

Note that multiple task models used extra training data to improve the results, which were unfair to compare with single task model. Even so, SA-BiLSTM model improved MT GCN model by a margin of 0.8 log M units in solubility dataset.

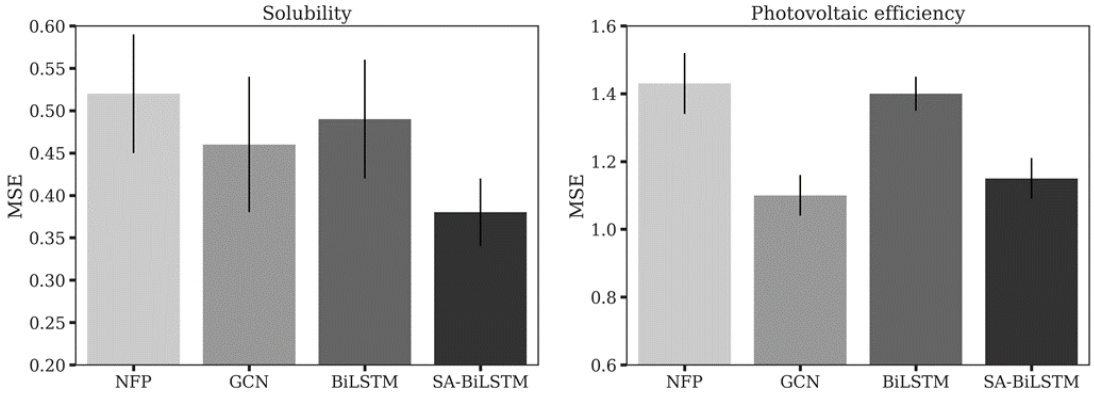

Figure 3. Comparison of language-based models to published graph models on two chemical property prediction datasets with same fold assignments.

3.1.2 Chemical stability

Bovens and co-workers constructed a random forest (RF) classification model with molecular fragments and reported the best-case accuracy of 72.9.[31] Later, Liu and co-workers applied naïve Bayesian model based upon atom center fragment features and achieved an accuracy of 76.5 %.[35] Their results with the same fold assignments of 9,764

compounds are listed in Table 2. Our SA-BiLSTM model improves RF model by a margin of 7.3%, and upon naïve Bayesian model by a margin of 3.7% for accuracy and a margin of 7.1% for recall.

Table 2. Comparison of language-based models to published conventional machine learning models on stable prediction dataset with same fold assignments.

Dataset	**Metric**	**RF**[a]	**Bayesian**[b]	**BiLSTM**	**SA-BiLSTM**
Stability	ACC (%)	72.9	76.5	79.1	**81.2**
	AUC (%)	--	83.5	84.2	**85.9**
	Recall (%)	--	73.0	77.2	**80.1**

[a]Random forest from ref 30. [b]Naive Bayesian from ref 34.

To further investigate the dependence of atom features on local chemical environments, we analyzed a few representative molecules in the stability dataset. We randomly selected 3 examples of negative and positive molecules from the test set, when the model had a high confidence (> 0.8) in predicting the label. As shown in Figure 4, we found that the model captured key factors strongly associated with the stability behind the molecule. For the unstable molecules, the model focused on heteroatoms, especially oxygens atom, sulfur atom and halogen atom (Figure 4(a)). It can also identify some important functional groups like chlorinated hydrocarbon (Figure 4b) and hydrazine in (Figure 4c). For the chemically stable molecules, the model paid attentions to the carbon atoms and related functional groups, like tert-butyl group (Figures 4d and 4f).

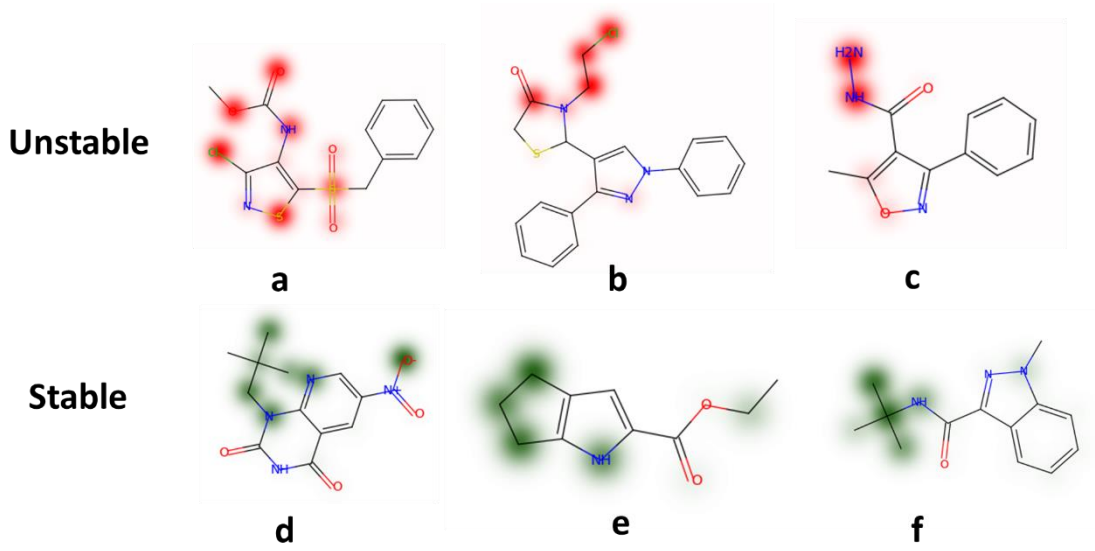

Figure 4. Heatmap of stable dataset molecules with the two-extreme score. *Red* unstable feature and *Green* Stable feature. The higher the attention coefficient, the darker the color.

3.2 Toxicity prediction

We used the modified Tox21 dataset reported in Wu's work.[36] In original paper, their graph convolutional network (GCN) model achieved the best performances in the test sets and reported the best-case average AUC of 0.829. In addition, Fernandez and co-workers trained 2D Convolution Networks (CCN) with molecular 2D images on the same dataset and reported the best-case average AUC of 0.708.[37] We also realized that there are many other better performance model, but we chose the CNN as reference because they utilized raw material of molecule (molecule image) to train the network, which was similar to our idea to some degree (molecular language). We should reiterate that we did not intend to emphasize the high performance of the model, but rather we demonstrate the utility of modern deep learning technique and molecular language representation can obtain both considerable results and good interpretability.

SA-BiLSTM models achieved AUC scores between 0.72 and 0.91 for the 12 separate prediction targets, shown in Figure 5. The average AUC value is 0.842, which is significantly better than the CNN model and surpassing the GCN model results in

8/12 cases (p-values = 0.033).

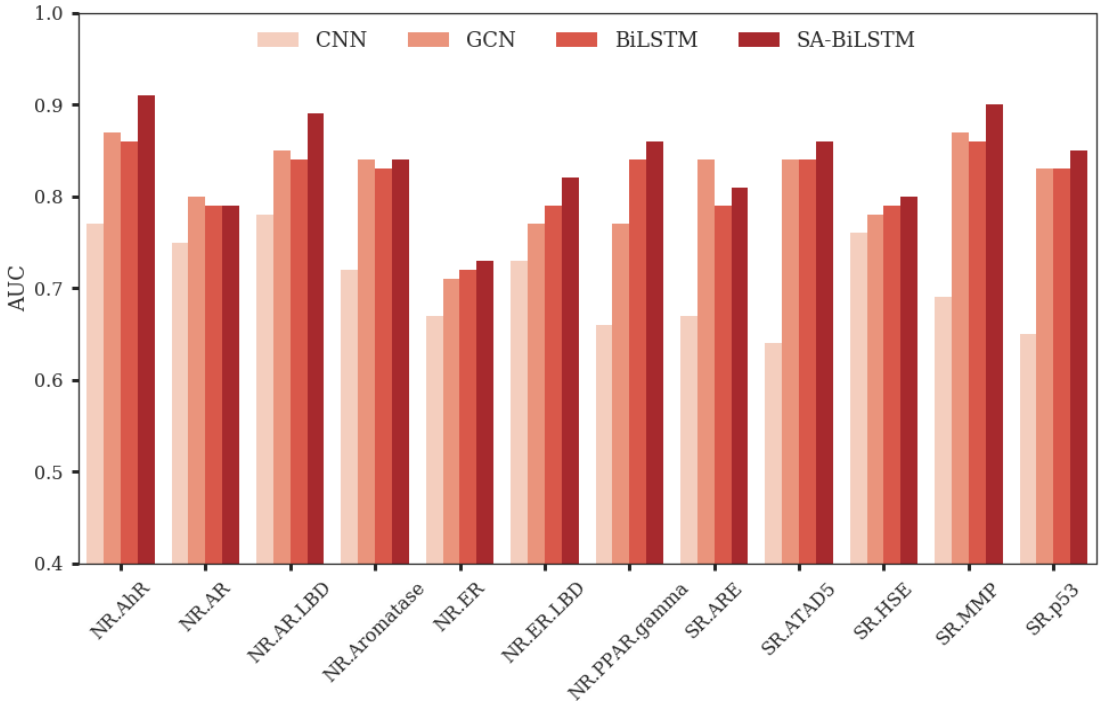

Figure 5. Language-based models' performance on the 12 targets in the Tox21 dataset in comparison to graph-based GCN and CNN models.

Selecting six representative toxic molecules from SR-MMP and NR-AhR datasets, we can color the molecules based on the results of the self-attention coefficients. As shown in Figure 6, SA-BiLSTM model can explicitly identify property related functional groups like phosphoric acid esters, long-chain quaternary ammonium salt, aliphatic halide, hydrazones and carboxylic acid, among which are well-known for reaction or toxicity.

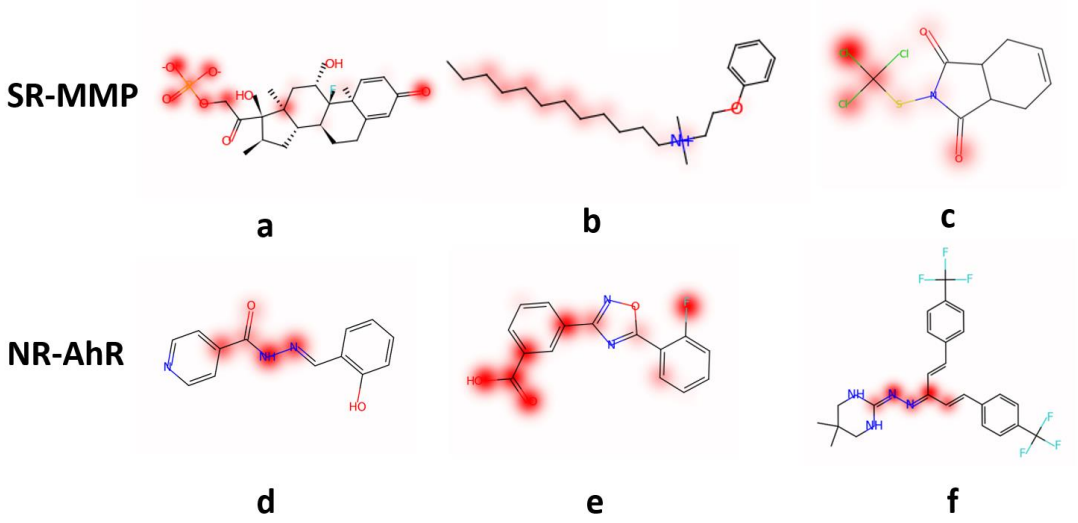

Figure 6. Heatmap of toxicity dataset molecules. *Red* predicted toxic or reactive features.

3.3 Bioactivity prediction

Duvenaud and Kearnes implemented their graph-based model in the drug efficacy datasets with the same condition as mentioned in **3.1.1 section**. They reported ST best-case MSE of 1.16 ± 0.03 and MT best-case MSE of 1.06 ± 0.06, respectively. Our results show that the SA-BiLSTM model have a significant improvement compare to neural fingerprints model and is comparable with the MT graph-based model, which is shown in Figure 7. For HIV dataset, we compared with the results reported in Wu's work,[36] where GCN and weave GCN achieved the best performances for the test set in the paper. Our SA-LSTM model improves upon weave GCN by a margin of 9.8% for the test set.

We do not report results for the DUD-E dataset group because our models performed extremely well on the DUD-E datasets (all subsets of DUD-E had median 5-fold-average AUCS > 0.99). The same situation has also been reported by Kearnes and co-workers[15]. The results can be attributed to the preprocessing of DUD-E datasets. However, we did not remove this dataset because it was clear enough to compare visualization of attention mechanism to protein-ligand interaction.

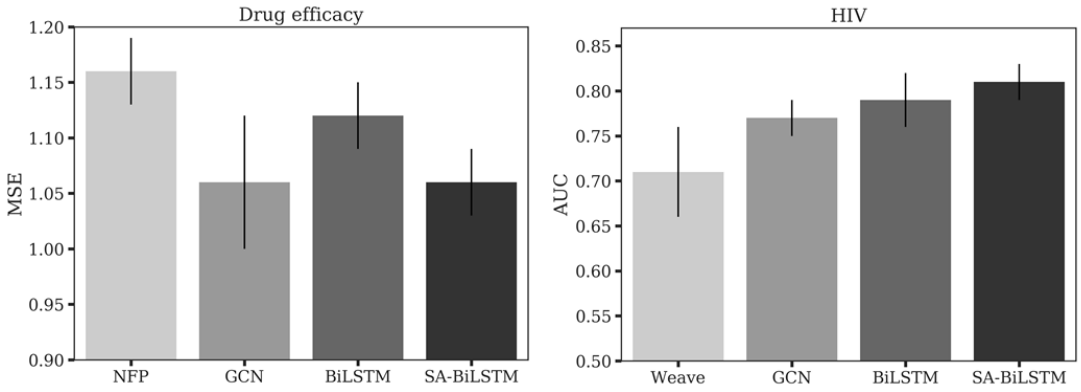

Figure 7. Comparison of language-based models to published models on drug efficacy and HIV datasets with same fold assignments.

The two properties (stability and toxicity) visualized above are directly related to molecular substructures. Thus, it would be easy for our model to identify specific functional groups relevant to the target properties. However, identifying key molecular substructures related to the bioactivity of a molecule can be more challenging.

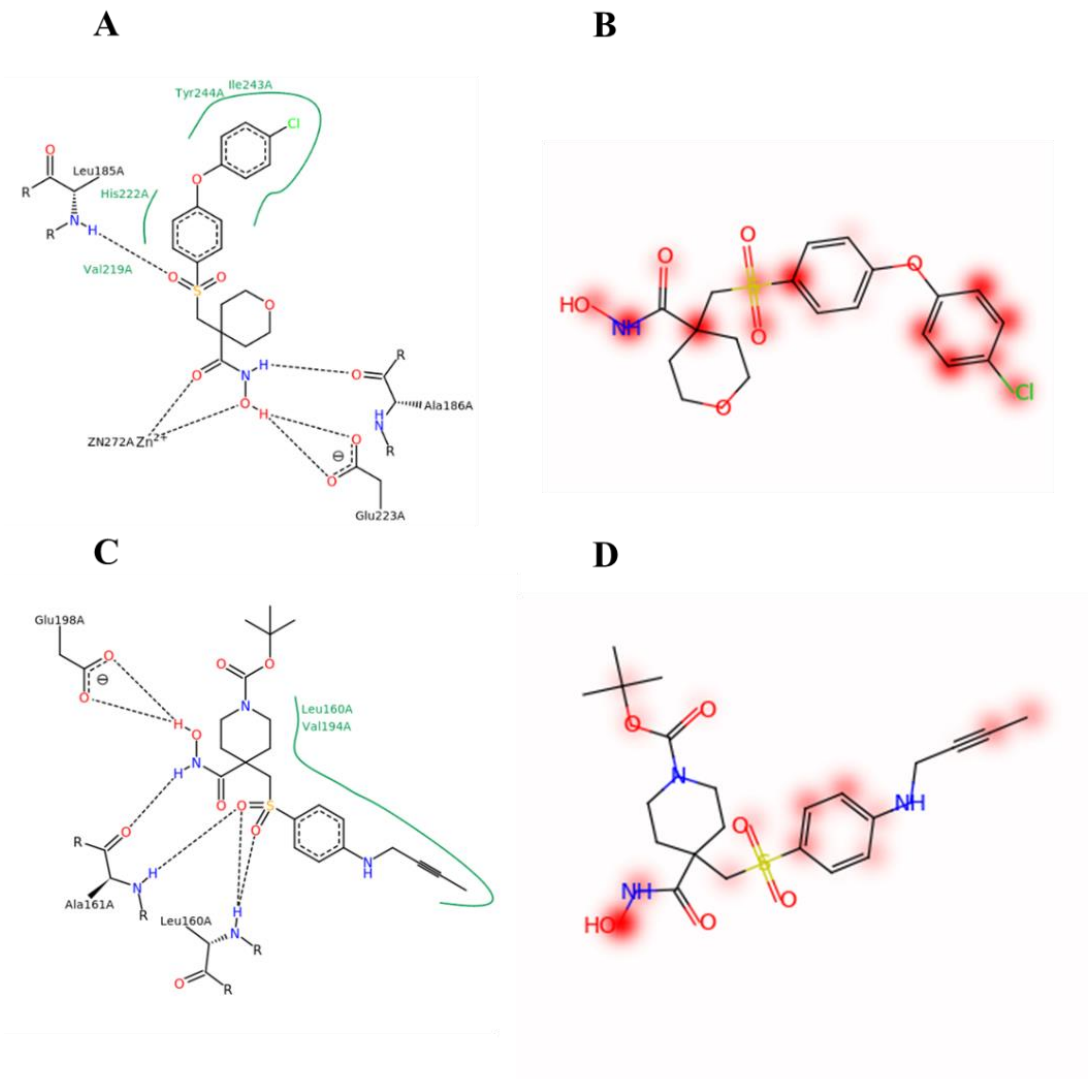

Figure 8. Examples for visualization of bioactive compounds with self-attention weights. Two authentic ligand-target interactions of RS1 (A) and WAY-344 (C) are shown on left. The predicted interaction sites of RS1 (B) and WAY-334 (D), colored in red, are shown as comparisons on right.

To exemplify this, we used a well-trained model to predict the bioactive compound of MMP-13. For typical active molecules selected from the test set, our model predicted with high confidences (> 0.8) comparing with the highlight parts with the authentic interaction sites.

Figure 8A shows the protein-ligand interaction of RS1 and MMP-13 (PDB ID: 830C)[38], and Figure 8B shows the attention weight of RS1. Figure 8C shows the protein-ligand interaction of WAY-344 and MMP-13 (PDB ID: 2PJT)[39], and Figure 8D

shows the attention weight of WAY-344. In both cases, molecular components with weights higher than 0.8 overlap substantially with the interaction sites between a compound and a protein. In the case of 830C, there are four main interaction sites with different kind of interactions, including hydrogen bonds, metal and hydrophobic interactions.

All of the key components of RS1 that related to the interactions were recognized and highlighted by the attention weight, even though there are some redundant parts like sulfur atom. In the case of 2PJT, the highlighted part of WAY-344 almost correspond to the part of interaction sites as well. This result suggests that the self-attention mechanism in the proposed model is superior in finding the relations of substructures and bioactivities in an interpretable format.

3.4 Exploratory experiments

Effect of multiple attentive vectors

Another intuitively important parameter is the number of attention rows in the molecular embedding, which we assigned *r* in the **2.3 section**. Having multiple rows in the molecular embedding is expected to provide more abundant information about the encoded content.

It makes sense to evaluate how significant the improvement can be brought by *r*. Taking the models we used for Photovoltaic Efficiency and Drug Efficacy dataset as an example, we change *r* from 1 to 30 for each task, and train the resulting 10 models in photovoltaic efficiency and drug efficacy datasets independently (Figure 9).

Figure 9 shows that there is significant difference between having only one attention row for the molecular embedding and multiple rows. The models are also quite stable with respect to *r*, since in the two figures a wide range of values between 5 to 30 are all generating comparable curves. Apart from this, we found that the number of *r* significantly affected the visualization.

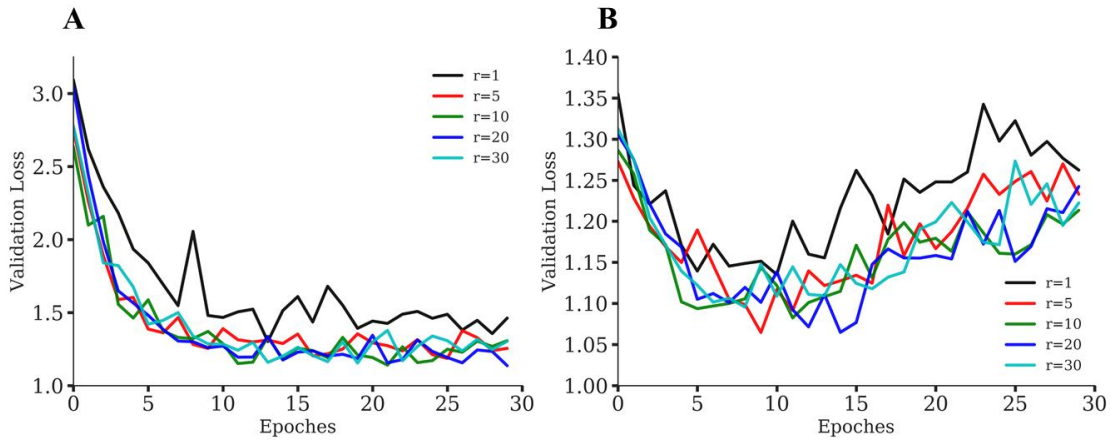

Figure 9. Effect of the number of rows (r) in model training on photovoltaic efficiency (left) and drug efficacy dataset (right).

Taking two molecules in drug efficacy dataset as examples, the model focused on different part of molecule with different number of rows (Figure 10). Redundant attention rows can lead to meaningless visualization. We observed that for most of cases, a few attention rows (5-10) were enough.

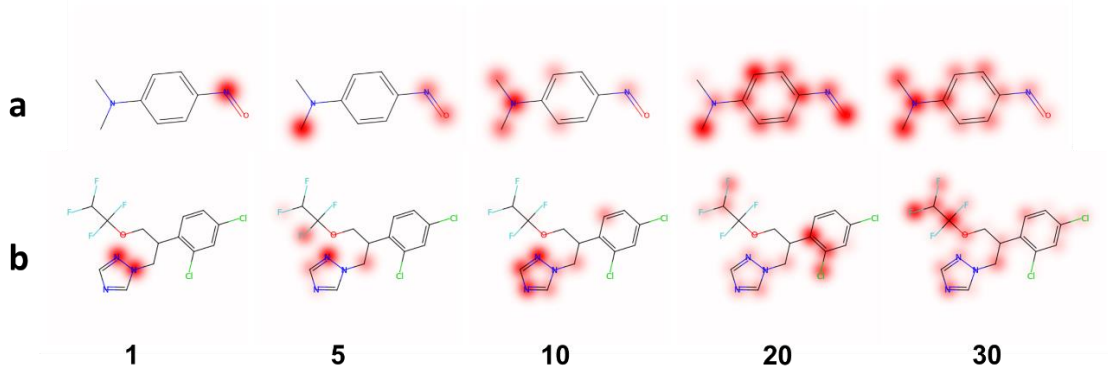

Figure 10. Effect of the number of rows (*r*) in visualization.

4 Conclusions

In this paper, we have proposed a deep learning method with a molecular self-attention mechanism. The results are interpretable for SAR/SPR studies when using SMILES as input. The model is successful because the substructures and their relations were already encoded in SMILES syntaxes. What we do is to consider SMILES as a chemical natural language, and use deep learning methods to figure out the relations among the

substructures and activities/properties from the chemical databases.

Our self-attention model outperforms the previous graph-based models and conventional machine learning models in most of cases. Furthermore, we demonstrated that the self-attention mechanism can explicitly identify the relationship between SMILES tokens and targeted chemical property. We believe that this study will provide new insights into the structure-activity relationship, and may find more applications in other fields in the future.

Acknowledgement

This work has funded in part of the national science & technology major project of the ministry of science and technology of China (2018ZX09735010), GD Frontier & Key Techn. Innovation Program (2015B010109004), GD-NSF (2016A030310228), Natural Science Foundation of China (U1611261) and the program for Guangdong Introducing Innovative and Enterpreneurial Teams (2016ZT06D211).

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

SZ contributed concept and implementation. SZ wrote the manuscript. All authors contributed to the interpretation of results. All authors reviewed and edited the manuscript.

Reference

1. Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, II; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; Consonni, V.; Kuz'min, V. E.; Cramer, R.; Benigni, R.; Yang, C.; Rathman, J.; Terfloth, L.; Gasteiger, J.; Richard, A.; Tropsha, A., QSAR modeling: where have you been? Where are you going to? *J Med Chem* **2014,** *57* (12), 4977-5010.
2. Lo, Y. C.; Rensi, S. E.; Torng, W.; Altman, R. B., Machine learning in chemoinformatics and drug discovery. *Drug Discov Today* **2018,** *23* (8), 1538-1546.

3. LeCun, Y.; Bengio, Y.; Hinton, G., Deep learning. *Nature* **2015,** *521* (7553), 436-44.
4. Ma, J.; Sheridan, R. P.; Liaw, A.; Dahl, G. E.; Svetnik, V., Deep neural nets as a method for quantitative structure-activity relationships. *J Chem Inf Model* **2015,** *55* (2), 263-74.
5. Jimenez, J.; Skalic, M.; Martinez-Rosell, G.; De Fabritiis, G., KDEEP: Protein-Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks. *J Chem Inf Model* **2018,** *58* (2), 287-296.
6. Lusci, A.; Pollastri, G.; Baldi, P., Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules. *J Chem Inf Model* **2013,** *53* (7), 1563-75.
7. Mobley, D. L.; Wymer, K. L.; Lim, N. M.; Guthrie, J. P., Blind prediction of solvation free energies from the SAMPL4 challenge. *J Comput Aided Mol Des* **2014,** *28* (3), 135-50.
8. Olivecrona, M.; Blaschke, T.; Engkvist, O.; Chen, H., Molecular de-novo design through deep reinforcement learning. *J Cheminform* **2017,** *9* (1), 48.
9. Segler, M. H. S.; Kogej, T.; Tyrchan, C.; Waller, M. P., Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent Sci* **2018,** *4* (1), 120-131.
10. Putin, E.; Asadulaev, A.; Vanhaelen, Q.; Ivanenkov, Y.; Aladinskaya, A. V.; Aliper, A.; Zhavoronkov, A., Adversarial Threshold Neural Computer for Molecular de Novo Design. *Mol Pharm* **2018,** *15* (10), 4386-4397.
11. Mayr, A. K., G.; Unterthiner, T.; Hochreiter, S, DeepTox: Toxicity Prediction using Deep Learning. *Front Environ Sci* **2016,** *3, 3.*
12. Bharath Ramsundar, S. K., Patrick Riley, Dale Webster, David Konerding, Vijay Pande. Massively Multitask Networks for Drug Discovery. *arXiv priprint arXiv:1502.02072v1*, 2015.
13. Ramsundar, B.; Liu, B.; Wu, Z.; Verras, A.; Tudor, M.; Sheridan, R. P.; Pande, V., Is Multitask Deep Learning Practical for Pharma? *J Chem Inf Model* **2017,** *57* (8), 2068-2076.
14. David K Duvenaud, D. M., Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. *arXiv priprint arXiv:1509.09292*, 2015.
15. Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P., Molecular graph convolutions: moving beyond fingerprints. *J Comput Aided Mol Des* **2016,** *30* (8), 595-608.
16. Coley, C. W.; Barzilay, R.; Green, W. H.; Jaakkola, T. S.; Jensen, K. F., Convolutional Embedding of Attributed Molecular Graphs for Physical Property Prediction. *J Chem Inf Model* **2017,** *57* (8), 1757-1772.
17. Seongok Ryu, J. L., Woo Youn Kim. Deeply learning molecular structure-property relationships using graph attention neural network. *arXiv priprint arXiv:1805.10988*, 2018.
18. Chao Shang, Q. L., Ko-Shin Chen, Jiangwen Sun, Jin Lu, Jinfeng Yi, Jinbo Bi. Edge Attention-based Multi-Relational Graph Convolutional Networks. *arXiv*

priprint arXiv:1802.04944, 2018.
19. Gomez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernandez-Lobato, J. M.; Sanchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A., Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS central science* **2018,** *4* (2), 268-276.
20. Putin, E.; Asadulaev, A.; Ivanenkov, Y.; Aladinskiy, V.; Sanchez-Lengeling, B.; Aspuru-Guzik, A.; Zhavoronkov, A., Reinforced Adversarial Neural Computer for de Novo Molecular Design. *J Chem Inf Model* **2018,** *58* (6), 1194-1204.
21. Putin, E.; Asadulaev, A.; Vanhaelen, Q.; Ivanenkov, Y.; Aladinskaya, A. V.; Aliper, A.; Zhavoronkov, A., Adversarial Threshold Neural Computer for Molecular de Novo Design. *Mol Pharm* **2018**.
22. Zhouhan Lin, M. F., Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, Yoshua Bengio. A STRUCTURED SELF-ATTENTIVE SENTENCE EMBEDDING. *arXiv priprint arXiv:1703.03130*, 2017.
23. Dzmitry Bahdanau, K. C., Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv priprint arXiv:1409.0473*, 2014.
24. Vaswani, A. S., Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N.; Kaiser, Lukasz; Polosukhin, Illia. Attention Is All You Need. *arXiv priprint arXiv:1706.03762*, 2017.
25. SMILES. http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html. Accessed 15 Sep 2018.
26. Jaeger, S.; Fulle, S.; Turk, S., Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition. *Journal of chemical information and modeling* **2018,** *58* (1), 27-35.
27. Yang, K. K.; Wu, Z.; Bedbrook, C. N.; Arnold, F. H., Learned protein embeddings for machine learning. *Bioinformatics* **2018**.
28. Pytorch. Version: 0.4.0. https://pytorch.org/.
29. Diederik P. Kingma, J. B. Adam: A Method for Stochastic Optimization. *arXiv priprint arXiv:1412.6980*, 2014.
30. Delaney, J. S., ESOL: estimating aqueous solubility directly from molecular structure. *J Chem Inf Comput Sci* **2004,** *44* (3), 1000-5.
31. Zitha-Bovens, E.; Maas, P.; Wife, D.; Tijhuis, J.; Hu, Q. N.; Kleinoder, T.; Gasteiger, J., COMDECOM: predicting the lifetime of screening compounds in DMSO solution. *J Biomol Screen* **2009,** *14* (5), 557-65.
32. Tox21 Challenge. https://tripod.nih.gov/tox21/challenge/. Accessed 2018-08-05.
33. Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K., Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J Med Chem* **2012,** *55* (14), 6582-94.
34. Gamo, F. J.; Sanz, L. M.; Vidal, J.; de Cozar, C.; Alvarez, E.; Lavandera, J. L.; Vanderwall, D. E.; Green, D. V.; Kumar, V.; Hasan, S.; Brown, J. R.; Peishoff, C. E.; Cardon, L. R.; Garcia-Bustos, J. F., Thousands of chemical starting points for antimalarial lead identification. *Nature* **2010,** *465* (7296), 305-10.
35. Liu, Z.; Zheng, M.; Yan, X.; Gu, Q.; Gasteiger, J.; Tijhuis, J.; Maas, P.; Li, J.; Xu,

J., ChemStable: a web server for rule-embedded naive Bayesian learning approach to predict compound stability. *J Comput Aided Mol Des* **2014,** *28* (9), 941-50.

36. Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V., MoleculeNet: a benchmark for molecular machine learning. *Chem Sci* **2018,** *9* (2), 513-530.
37. Fernandez, M.; Ban, F.; Woo, G.; Hsing, M.; Yamazaki, T.; LeBlanc, E.; Rennie, P. S.; Welch, W. J.; Cherkasov, A., Toxic Colors: The Use of Deep Learning for Predicting Toxicity of Compounds Merely from Their Graphic Images. *J Chem Inf Model* **2018,** *58* (8), 1533-1543.
38. Lovejoy, B.; Welch, A. R.; Carr, S.; Luong, C.; Broka, C.; Hendricks, R. T.; Campbell, J. A.; Walker, K. A.; Martin, R.; Van Wart, H.; Browner, M. F., Crystal structures of MMP-1 and -13 reveal the structural basis for selectivity of collagenase inhibitors. *Nat Struct Biol* **1999,** *6* (3), 217-21.
39. Huang, A.; Joseph-McCarthy, D.; Lovering, F.; Sun, L.; Wang, W.; Xu, W.; Zhu, Y.; Cui, J.; Zhang, Y.; Levin, J. I., Structure-based design of TACE selective inhibitors: manipulations in the S1'-S3' pocket. *Bioorg Med Chem* **2007,** *15* (18), 6170-81.