Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 7181815, 7 pages
https://doi.org/10.1155/2021/7181815

Research Article

Mol-BERT: An Effective Molecular Representation with BERT for Molecular Property Prediction

Juncai Li[1] **and Xiaofei Jiang**[2]

[1]*Hunan Vocational College of Electronic and Technology, Changsha 410220, China*
[2]*College of Information Science and Engineering, Hunan University, Changsha 410082, China*

Correspondence should be addressed to Xiaofei Jiang; jiangxiaofei@hnu.edu.cn

Received 30 June 2021; Accepted 13 August 2021; Published 3 September 2021

Academic Editor: Yulin Wang

Molecular property prediction is an essential task in drug discovery. Most computational approaches with deep learning techniques either focus on designing novel molecular representation or combining with some advanced models together. However, researchers pay fewer attention to the potential benefits in massive unlabeled molecular data (e.g., ZINC). This task becomes increasingly challenging owing to the limitation of the scale of labeled data. Motivated by the recent advancements of pretrained models in natural language processing, the drug molecule can be naturally viewed as language to some extent. In this paper, we investigate how to develop the pretrained model BERT to extract useful molecular substructure information for molecular property prediction. We present a novel end-to-end deep learning framework, named Mol-BERT, that combines an effective molecular representation with pretrained BERT model tailored for molecular property prediction. Specifically, a large-scale prediction BERT model is pretrained to generate the embedding of molecular substructures, by using four million unlabeled drug SMILES (i.e., ZINC 15 and ChEMBL 27). Then, the pretrained BERT model can be fine-tuned on various molecular property prediction tasks. To examine the performance of our proposed Mol-BERT, we conduct several experiments on 4 widely used molecular datasets. In comparison to the traditional and state-of-the-art baselines, the results illustrate that our proposed Mol-BERT can outperform the current sequence-based methods and achieve at least 2% improvement on ROC-AUC score on Tox21, SIDER, and ClinTox dataset.

1. Introduction

Effectively identifying the molecular properties (e.g., bioactivity and toxicity) plays an essential part in drug discovery and material science, which can alleviate the costly and time-consuming process in comparison to the traditional experiment methods [1]. Such a process is usually known as molecular property prediction, and it is a fundamental task to explore the functionality of new drugs. A typical molecular property prediction system takes the drug features of descriptors as the input and outputs the predicted result of predefined chemical properties. The predicted value can benefit various subsequent tasks, including virtual screening [2–4] and drug repurposing [5–7]. However, accurately predicting molecular property with computational methods remains challenging.

Previous machine learning approaches focused on designing a variety of expert-engineered descriptors or molecular fingerprints manually based on experimental statistics to predict molecular property [8–10]. For example, extended-connectivity fingerprint (ECFP) [11], as the most representative fingerprint method, was designed to generate different types of circular fingerprints that extracted the molecular structures of atom neighborhoods by using a fixed hash function [12]. Then, these obtained fingerprint representations would be sent to traditional machine learning models to perform further predictions, and it can be applied to a wide range of different models, such as logistic regression, support vector classification, kernel ridge regression, random forest, influence relevance voting, and multitask networks [13]. However, this line of researches heavily depends on the design of hand-crafted features and domain

knowledge. Besides, the generated hash bit vectors are difficult to biologically understand the relationship between chemical properties and molecular structures.

Inspired by the remarkable achievements that deep learning has shown in a variety of domains, including computer vision [14] and natural language processing [15, 16], it also has gained lots of attention for molecular property prediction. The molecular representation methods being introduced can be mainly summarized into two parts: sequence-based and graph-based approaches. For sequence-based methods, simplified molecular input line entry specification, shortened as SMILES, is the most common molecular linear notation that encodes the molecular topology on the basis of chemical rules [17]. In this way, several methods are attempted to take SMILES representation as the input and use current successful models (e.g., recurrent neural networks) to obtain molecular representations [18], while this line of work suffered from insufficient labeled data for specific molecular tasks. More recently, researchers adopted the unsupervised and pretraining strategies in natural language processing (NLP) to learn contextual information from large unlabeled molecular datasets. For example, an unsupervised machine learning method named Mol2vec was developed to learn vector representations of molecular substructures [19]. And SMILES-BERT was proposed to pretrain the model through a masked SMILES recovery task by designing attention mechanism-based transformer layer [20]. These pretrained methods pay more attention to the contextual information of molecular sequences, but they hardly consider some molecular substructure (i.e., functional groups) that essentially contributes to the molecular property [21, 22].

On the other hand, graph neural networks (GNNs) have been adopted to explore the graph-based representation for molecular property prediction [23–25]. Graph convolutions were the first work that applied the convolutional layers to encode molecular graph into neural fingerprints [26]. Similarly, much efforts are made to extend a variety of GNNs on property prediction tasks. For example, the weave featurization encoded chemical features to form molecule-level representations [27]. And some methods extended graph attention network [28] to learn the aggregation weights [25, 29]. Moreover, to better encode the interactions between atoms, a message passing neural network named MPNN was designed to utilize the attributed features of both atoms and edges [30]. More recently, DMPNN [31] and CMPNN [32] were further introduced to leverage the attributed information of nodes and edges during message passing. Although graph-based models have achieved great performance on molecular graph representation, they seldom make use of the vast available biological sequence data.

Recently, substantial pretrained models [33–37] trained on the large corpus or unlabeled data can learn universal representations, which are benefit for various downstream tasks, including protein sequence representation [38, 39], biomedical text mining [40, 41], and chemical reaction prediction [42]. Advances in pretrained models have shown their powerful ability for extracting information from unlabeled sequences, which raises a tantalizing question: can we develop a pretrained model to extract useful molecular substructure information from massive SMILES sequence datasets? To help solve this problem, we propose a novel neural framework, named Mol-BERT, tailored for molecular property prediction. The idea of Mol-BERT is natural and intuitive. Our framework consists of three types of modules. The feature extractor is first to extract atom-level and substructure features centered on the current atom, and the first module can be replaced with a wide range of different molecular representation methods. Then, the pretrained BERT module learns molecular substructure or fragment information from large pretraining corpus (i.e., unlabeled SMILES sequences). The final module is to predict the specific molecular property after fine-tuning the pretrained Mol-BERT via a multityped classifier. To illustrate the performance of our proposed method in various prediction tasks, Mol-BERT is fine-tuned and evaluated on 4 widely used molecular benchmark datasets. In comparison to state-of-the-art baselines (i.e., sequence- and graph-based methods), the experimental results prove the effectiveness of our proposed Mol-BERT.

This paper is organized as follows. Section 2 firstly introduces the preprocessed corpus for Mol-BERT pretraining and several molecular benchmark datasets used in this work. Then, Section 3 presents the molecular representation method, the pretraining, and fine-tuning of the Mol-BERT model, respectively. Moreover, Section 4 analyzes the prediction performance of our proposed method on several molecular datasets and compares it with state-of-the-art sequence-based and graph-based approaches. Finally, the conclusion of this work is summarized in Section 5.

2. Materials

The corpus of chemical compound (i.e., unlabeled SMILES) was obtained from the available ZINC and ChEMBL databases. As a free and available database for virtual screening, ZINC database contains over 230 million purchasable compounds in multiple formats, including ready-to-dock and 3D structures [43]. And ChEMBL database is a manually built database of bioactive molecules with drug-like properties, which collects 1,961,462 distinct compounds [44]. Specifically, we selected compound SMILES from ZINC version 15 and ChEMBL version 27 that can be processed by RDKit software [45], and the duplicates were removed in merged dataset. Moreover, we filtered them by following the same criteria of Mol2Vec [19]. Specifically, the two databases were firstly merged, and duplicates were removed. Then, only compounds SMILES that could be processed by RDKit were kept, and they were filtered according to the following cutoffs and criteria: molecular weight between 12 and 600; heavy-atom count between 3 and 50; clogP21between 5 and 7; and only H, B, C, N, O, F, P, S, Cl, and Br atoms allowed. Additionally, all counterions and solvents were removed, and canonical SMILES representations were generated by RDKit. Finally, this procedure yielded 4 million compounds. Detailed information on the pretraining corpus is provided in Supplementary (available here).

In this paper, we selected 4 widely used benchmark datasets from MoleculeNet [13] to evaluate the performance of our proposed method. SMILES strings were used to encode the input chemical compound in all benchmark datasets. The benchmark datasets we used are introduced as follows:

(i) BBBP. The BBBP dataset provides 2,053 compounds on their permeability properties to predict the barrier permeability

(ii) Tox21. The Tox21 dataset measures 8,014 compounds with their corresponding toxicity data against 12 targets. The label of toxicity is recorded as binary task: if the label value is 1, then it means the compound has toxicity on specific target or 0 otherwise

(iii) SIDER. The SIDER dataset contains a total of 1,427 compounds and their adverse drug reactions (ADR) against 27 system-organ class. The ADR result is described as binary labels

(iv) ClinTox. The ClinTox dataset provides 2 classification tasks for 1,491 drug compounds with known chemical structures, including clinical trial toxicity and FDA approval status

In this paper, we followed the experimental setting of FP2VEC [46], and we split the datasets into the train, validation, and test set with a ratio of 8/1/1. Table 1 shows the detailed description of selected benchmark datasets. Please note that binary and multilabel correspond to the binary and multilabel classification tasks, respectively. And random splitting method randomly splits the samples into training, validation, and test subsets. Scaffold splitting method splits the samples on the basis of their 2D structural frameworks implemented by RDKit software.

Table 1: The detailed description of selected benchmark datasets.

Dataset	Category	Compound	Tasks	Task type	Split method
BBBP	Physiology	2,053	1	Binary	Scaffold
Tox21	Physiology	8,014	12	Multilabel	Scaffold
SIDER	Physiology	1,427	27	Multilabel	Scaffold
ClinTox	Physiology	1,491	2	Multilabel	Scaffold

3. Methods

In this section, we first describe the overview of our proposed Mol-BERT; then, we separately introduce three modules, which we refer to as the feature extractor, pretraining, and fine-tuning of Mol-BERT, respectively.

3.1. Overview. Figure 1 illustrates the overall process of Mol-BERT. As shown in Figure 1, Mol-BERT consists of three modules, including feature extractor, pretraining, and fine-tuning of Mol-BERT. The Mol-BERT framework learns to predict the molecular property as follows. Given the input drug data (i.e., canonical SMILES), the featurizer module adopts the effective molecular representation to transform them into a set of atom identifier (recall the detail in Feature Extractor). Then, the outputs are fed into a BERT module to obtain a contextual embedding of each molecular substructure through pretraining BERT on vast preprocessed corpus (recall the detail in Pretraining Mol-BERT). Finally, the fine-tuned Mol-BERT outputs a value indicating the probability of certain molecular property in classification task (recall the detail in Fine-Tuning Mol-BERT).

3.2. Feature Extractor. The molecular substructure is an important cue for molecular interactions [21, 22]. Therefore, the key idea behind Mol-BERT is that we strengthen to obtain a better representation of molecular substructures by pretraining BERT on the vast unlabeled SMILES sequences. Inspired by Mol2Vec [19] that considered molecular substructures or fragments derived from the Morgan algorithm as "words" and compound as "sentences," here we adopt a similar method to decompose the input SMILES sequences into biological words and sentences.

To achieve it, given an input compound SMILES string, we first obtain its standardize and canonical SMILES representation S generated by RDKit. Then, the Morgan algorithm [11] is used to generate all atom identifiers with radius 0 and 1, denoted by A_i^0 and $A_{i\,i}^1$, respectively, where the subscript i represents the index of each atom. As illustrated in the left part of Figure 1, A_i^0 (i.e., green node) represents the current node set traversed in an atom order while A_i^1 (i.e., Kelly node) represents the neighboring node set connecting directly to the current atom, so A_i^1an be viewed as a kind of substructure or fragment. And A_i are then hashed into a fixed-length vector. Take CC(N)C(=O)O as an example; it consists of six atoms, and we obtain its atom identifiers A_i^0 (i.e., A_1^0-A_6^0) and the corresponding substructures (i.e., A_1^1-A_6^1), and then, they are hashed into a fixed-length vector (e.g., A_1^1 corresponds to 3537119591). Finally, all vectors of the Morgan substructures are summed to obtain the molecular representation. Therefore, in this way, we can generate 119 atom identifiers at radius 0 and 13325 substructure identifiers at radius 1, respectively. The feature extractor module in Mol-BERT can be replaced with various molecular representation methods. For example, FP2Vec [46] can be used as the feature extractor to generate the 1024-bit Morgan (or circular) fingerprint with the predefined radius value.

3.3. Pretraining Mol-BERT. As a contextualized word representation model, BERT [33] adopted the masked technique to predict randomly masked words in a sequence, which can result in learning bidirectional representations. Therefore, Mol-BERT also uses a masked SMILES task (i.e., atom identifier) to predict random substructure in a SMILES string. Different from the traditional way of pretraining language models in NLP that BERT was trained on English Wikipedia and BooksCopus, in this paper, we pretrain Mol-BERT on our preprocessed corpus obtained from ZINC version 15 and ChEMBL version 27 datasets. Specifically, the input SMILES is transformed into a list of atom identifiers A_i

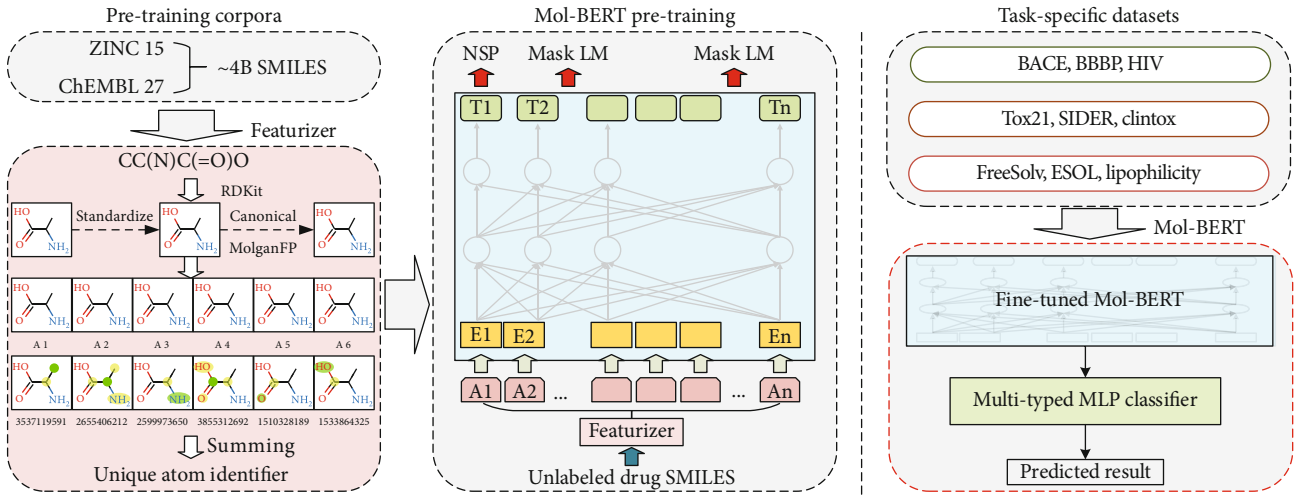

Figure 1: Overview of our proposed Mol-BERT for molecular property prediction.

via a previous module, rather than character-level for SMILES [20], and then, they are embedded as the input of BERT module for pretraining. We initialized our proposed Mol-BERT with weights from BERT [33] and follow the same way to randomly mask 15% tokens in a SMILES (i.e., atom identifier) as [MASK] token. The tokens are embedded into the feature vector. Here, we use token embedding and positional embedding since only the Masked Language Model (MLM) task is adopted in this paper. The proposed Mol-BERT is different from BERT in several ways as follows: (1) Mol-BERT adopted single masked SMILES task (i.e., MLM) on large-scale unlabeled datasets, while BERT uses two kinds of self-supervised tasks on English Wikipedia and BooksCopus, and (2) w exclude the segmentation embedding adopted in the BERT model since Mol-BERT does not require the continuous sentence training.

3.4. Fine-Tuning Mol-BERT. After pretraining on the vast of unlabeled SMILES compounds, with minimal modification of hyperparameters, Mol-BERT can be applied to molecular property prediction on various downstream tasks. We mostly follow the same architecture, optimization, and hyperparameter choices used in [8]. For classification task (i.e., BBBP and Tox21), we feed the final BERT vector into a linear classification layer to predict the molecular property. A simple classifier is adopted to output the binary value. Then, the labeled sample is used for fine-tuning the model. Mol-BERT feeds the learned drug embeddings into a multi-typed MLP classifier to generate predictions. Output scores include both continuous scores, such as the solubility value and as binary outputs indicating whether a molecule is toxic or nontoxic. The multityped classifier detects whether the task is regression or classification and switches to the correct loss function and evaluation metrics. In the case of regression, we use the mean square error (MSE) as the loss function and root mean square error (RMSE) as performance metrics. In the classification case, we use binary cross entropy as the loss function and area under the receiver operating characteristics (AUC-ROC) as performance metrics. Given a set of SMILES compounds and the ground-truth labels in the training dataset, we used the crossentropy and the mean square error as loss function for classification and regression tasks, respectively.

Table 2: The fine-tuning hyperparameters.

Parameter	Value/range
Learning rate	1*e*-5~1*e*-3
Batch size	8
Epoch	100
Optimizer	Adam
Embedding dimension	300
Size of dictionary	13,325
Number of attention head	6
Layers of fully connected neural network	6

4. Results and Discussion

In this section, we first introduce the experimental settings. Then, we demonstrate the performance of our proposed Mol-BERT in comparison to state-of-the-art methods to predict the molecular property on 4 wildly used benchmark datasets.

4.1. Baseline Methods. We compare Mol-BERT with many state-of-the-art sequence-based and graph-based baselines which can be categorized as follows:

(i) ECFP: extended-connectivity fingerprints, referred to as ECFP [11], are a type of widely used circular or Morgan fingerprints for encoding the substructures in a molecule

(ii) GraphCov: graph convolutions are proposed by [26] to apply the convolutional networks for learning molecular fingerprints. Here, we term it as GraphCov

Table 3: The metric scores of the test set against BBBP, Tox21, SIDER, and ClinTox datasets.

Model/dataset	BBBP	Tox21	SIDER	ClinTox
ECFP	0.702 ± 0.006	0.810 ± 0.013	0.673 ± 0.025	0.783 ± 0.023
GraphCov	0.877 ± 0.036	0.772 ± 0.041	0.593 ± 0.035	0.845 ± 0.051
Weave	0.837 ± 0.065	0.741 ± 0.044	0.543 ± 0.034	0.823 ± 0.023
MPNN	0.913 ± 0.041	0.808 ± 0.024	0.595 ± 0.030	0.879 ± 0.054
FP2VEC	0.874 ± 0.023	0.730 ± 0.006	0.582 ± 0.008	0.643 ± 0.032
SMILES-BERT	0.814 ± 0.093	0.732 ± 0.025	0.601 ± 0.010	0.872 ± 0.017
Mol-BERT	0.875 ± 0.048	0.839 ± 0.075	0.695 ± 0.071	0.923 ± 0.025

(iii) Weave: similar to GraphCov, the weave featurization [27] encodes meaningful features of atom, bond, and graph distances between matching pairs to form molecule-level representations

(iv) MPNN: a novel message passing method is proposed to be operated on undirected graph [30]

(v) FP2VEC: based on Morgan or circular fingerprint, it introduces and encodes a molecule as trainable vectors [46]

(vi) SMILES-BERT: [20] proposes a semisupervised BERT model that takes the SMILES representation as input

We report the results of these baselines in FP2Vec [46], including ECFP, GraphCov, Weave, and FP2VEC. And we reimplemented MPNN and SMILES-BERT, respectively. As for MPNN [30], it is a graph-based model considering the edge features during message passing. And SMILES-BERT [20] is a sequence-based model based on transformer layer and attention mechanisms entirely to encode compound SMILES. These models are relied on the public code and kept the same settings of models the same as reported in the original papers.

4.2. Evaluation Metrics. We applied the area under the receiver operating characteristic curve (AUC-ROC) metric for classification task. Following [46], we train the prediction model with a train set and optimize the model based on the AUC-ROC metric of validation set for classification task. And the prediction results are measured using those optimized models on the test set. For all experiments in this paper, we repeated the same procedures on each task for 5 times and reported the mean and standard deviation of AUC scores. Besides, we evaluated all models on the scaffold splitting method as reported by [46].

4.3. Implementation Details. To optimize all trainable parameters, we adopt Adam optimizer for pretraining and fine-tuning. The dynamic learning rate technique is adopted to adjust the learning rate during training and fine-tuning according to various downstream tasks. We use PyTorch to implement Mol-BERT. And we use 3 NVIDIA GTX 1080Ti GPUs to pretrain Mol-BERT. All fine-tuning tasks are run on a single NVIDIA GTX 2080Ti GPU. Table 2 shows all the hyperparameters of the fine-tuning model.

4.4. Comparison Results. To examine the competitiveness of the proposed model, we compared Mol-BERT with state-of-the-art models used for molecular property prediction on classification task. Table 3 reports the mean and standard deviation of ROC-AUC score on BBBP, SIDER, Tox21, and ClinTox datasets. From this table, we can observe that the proposed Mol-BERT significantly outperforms the baselines across three datasets, including Tox21, SIDER, and ClinTox. More specifically, our proposed Mol-BERT achieved at least 2.9% on Tox21, 2.2% on SIDER, and 4.4% on ClinTox higher ROC-AUC metric than baselines. For example, on the Tox21 dataset, Mol-BERT achieved a ROC-AUC score of 0.839 with 2.9% absolute gain compared to ECFP (the second best method). This is because Mol-BERT leverages the molecular representation pretrained on large- scale unlabeled SMILES sequences, while ECFP heavily relied on feature engineering. Compared with graph-based methods that explore the molecular graph features, the proposed Mol-BERT outperformed them on three datasets while it achieved comparable performance with MPNN on the BBBP dataset. This is due to the fact that the contextual information learned from large unlabeled datasets can benefit a lot to the model performance. Moreover, in comparison to the sequence-based pretrained model (i.e., SMILES-BERT), our proposed Mol-BERT achieved stable performance across all datasets. This is a very encouraging result. The reason could be that our method adopted the molecular representation to consider the structural feature of molecular substructures, which benefits to the performance. Overall, it is essentially a nontrivial achievement in terms of molecular property prediction.

5. Conclusions

In this paper, we proposed an effective molecular representation method with the pretrained BERT model, named Mol-BERT, to resolve the molecular property prediction. Our proposed Mol-BERT leverages the molecular representation of substructures pretrained on large-scale unlabeled SMILES dataset, which is able to learn both structural and the contextual information of drug. We implement the proposed method and conduct experimental comparisons on four

widely used benchmarks. The experimental results show that Mol-BERT outperforms the classic and state-of-the-art graph-based models on molecular property prediction.

While our proposed method achieves good performance on classification tasks, there are still some limitations expected to be overcome. First, our method achieves relatively poorer performance on regression task, mainly owing to the small number of samples in the dataset (e.g., FreeSolv). We would like to investigate metalearning strategies for data augmentation, which results in great success in natural language processing. Second, molecular property prediction is the primary step in drug discovery; we will continue to improve our method to further investigate the following prediction task (e.g., protein-protein interaction, drug-disease associations) in the future.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request (https://github.com/cxfjiang/MolBERT).

Conflicts of Interest

The authors declare no competing financial interest.

Supplementary Materials

The pretraining corpus are available at "https://drive.google.com/drive/folders/1ST0WD1-hX9XtiPWwCceZbgZlBV0fKPbe." *(Supplementary Materials)*

References

[1] S. Ekins, A. C. Puhl, K. M. Zorn et al., "Exploiting machine learning for end-to-end drug discovery and development," *Nature Materials*, vol. 18, no. 5, pp. 435–441, 2019.

[2] X. Lin, Z. Quan, Z. J. Wang, H. Huang, and X. Zeng, "A novel molecular representation with BiGRU neural networks for learning atom," *Briefings in Bioinformatics*, vol. 21, no. 6, pp. 2099–2111, 2020.

[3] X. Lin, Z. Quan, Z. J. Wang, T. Ma, and X. Zeng, "KGNN: knowledge graph neural network for drug-drug interaction prediction," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pp. 2739–2745, Yokohama, Japan, 2020.

[4] B. K. Shoichet, "Virtual screening of chemical libraries," *Nature*, vol. 432, no. 7019, pp. 862–865, 2004.

[5] S. Pushpakom, F. Iorio, P. A. Eyers et al., "Drug repurposing: progress, challenges and recommendations," *Nature Reviews Drug Discovery*, vol. 18, no. 1, pp. 41–58, 2019.

[6] Z. Quan, Y. Guo, X. Lin, Z. J. Wang, and X. Zeng, "GraphCPI: graph neural representation learning for compound-protein interaction," in *2019 IEEE International Conference on Bioinformatics and Biomedicine*, pp. 717–722, San Diego, CA, USA, 2019.

[7] Y. Zhou, Y. Hou, J. Shen, Y. Huang, W. Martin, and F. Cheng, "Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2," *Cell Discovery*, vol. 6, no. 1, pp. 1–18, 2020.

[8] D. S. Cao, Q. S. Xu, Q. N. Hu, and Y. Z. Liang, "ChemoPy: freely available python package for computational biology and chemoinformatics," *Bioinformatics*, vol. 29, no. 8, pp. 1092–1094, 2013.

[9] A. Mauri, V. Consonni, M. Pavan, and R. Todeschini, "Dragon software: an easy approach to molecular descriptor calculations," *Match*, vol. 56, no. 2, pp. 237–248, 2006.

[10] H. Moriwaki, Y. S. Tian, N. Kawashita, and T. Takagi, "Mordred: a molecular descriptor calculator," *Journal of Cheminformatics*, vol. 10, no. 1, p. 4, 2018.

[11] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," *Journal of Chemical Information and Modeling*, vol. 50, no. 5, pp. 742–754, 2010.

[12] R. C. Glen, A. Bender, C. H. Arnby, L. Carlsson, S. Boyer, and J. Smith, "Circular fingerprints: flexible molecular descriptors with applications from physical chemistry to ADME," *IDrugs*, vol. 9, no. 3, p. 199, 2006.

[13] Z. Wu, B. Ramsundar, E. N. Feinberg et al., "MoleculeNet: a benchmark for molecular machine learning," *Chemical Science*, vol. 9, no. 2, pp. 513–530, 2018.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, Las Vegas, United States, 2016.

[15] C. Xia, C. Zhang, X. Yan, Y. Chang, and P. S. Yu, "Zero-shot user intent detection via capsule neural networks," 2018, https://arxiv.org/abs/1809.00385.

[16] J. Yin, C. Gan, K. Zhao, X. Lin, Z. Quan, and Z. J. Wang, "A novel model for imbalanced data classification," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 4, pp. 6680–6687, 2020.

[17] D. Weininger, A. Weininger, and J. L. Weininger, "SMILES. 2. Algorithm for generation of unique smiles notation," *Journal of Chemical Information and Computer Sciences*, vol. 29, no. 2, pp. 97–101, 1989.

[18] Z. Xu, S. Wang, F. Zhu, and J. Huang, "Seq2seq fingerprint: an unsupervised deep molecular embedding for drug discovery," in *Proceedings of the 8th ACM international conference on bioinformatics, computational biology, and health informatics*, pp. 285–294, New York, NY, USA, 2017.

[19] S. Jaeger, S. Fulle, and S. Turk, "Mol2vec: unsupervised machine learning approach with chemical intuition," *Journal of Chemical Information and Modeling*, vol. 58, no. 1, pp. 27–35, 2018.

[20] S. Wang, Y. Guo, Y. Wang, H. Sun, and J. Huang, "SMILES-BERT: large scale unsupervised pre-training for molecular property prediction," in *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pp. 429–436, New York, NY, USA, 2019.

[21] K. Huang, C. Xiao, T. Hoang, L. Glass, and J. Sun, "Caster: predicting drug interactions with chemical substructure representation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 1, pp. 702–709, 2020.

[22] R. B. Silverman and M. W. Holladay, *The Organic Chemistry of Drug Design and Drug Action*, Academic Press, 2014.

[23] K. Schütt, P. J. Kindermans, H. E. S. Felix, S. Chmiela, A. Tkatchenko, and K. R. Müller, "SchNet: a continuous-filter convolutional neural network for modeling quantum interactions," *Advances in neural information processing systems*, pp. 991–1001, 2017, https://arxiv.org/abs/1706.08566.

[24] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko, "Quantum-chemical insights from deep tensor neural networks," *Nature Communications*, vol. 8, no. 1, pp. 1–8, 2017.

[25] Z. Xiong, D. Wang, X. Liu et al., "Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism," *Journal of Medicinal Chemistry*, vol. 63, no. 16, pp. 8749–8760, 2020.

[26] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre et al., "Convolutional networks on graphs for learning molecular fingerprints," *Advances in Neural Information Processing Systems*, pp. 2224–2232, 2015, https://arxiv.org/abs/1509.09292.

[27] S. Kearnes, K. McCloskey, M. Berndl, V. Pande, and P. Riley, "Molecular graph convolutions: moving beyond fingerprints," *Journal of Computer-Aided Molecular Design*, vol. 30, no. 8, pp. 595–608, 2016.

[28] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," 2017, https://arxiv.org/abs/1710.10903.

[29] S. Ryu, J. Lim, S. H. Hong, and W. Y. Kim, "Deeply learning molecular structure-property relationships using attention- and gate-augmented graph convolutional network," 2018, https://arxiv.org/abs/1805.10988.

[30] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, *Neural Message Passing for Quantum Chemistry. in International Conference on Machine Learning*, PMLR, 2017.

[31] K. Yang, K. Swanson, W. Jin et al., "*Are learned molecular representations ready for prime time?, [Ph.D. thesis]*," Massachusetts Institute of Technology, 2019.

[32] Y. Song, S. Zheng, Z. Niu, Z. H. Fu, Y. Lu, and Y. Yang, "Communicative representation learning on attributed molecular graphs," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pp. 2831–2838, Yokohama, Japan, 2020.

[33] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, vol. 1, pp. 4171–4186, Minneapolis, United States, 2019.

[34] W. Hu, B. Liu, J. Gomes et al., "Strategies for pre-training graph neural networks," 2019, https://arxiv.org/abs/1905.12265.

[35] K. Li, Y. Zhong, X. Lin, and Z. Quan, "Predicting the disease risk of protein mutation sequences with pre-training model," *Frontiers in Genetics*, vol. 11, p. 1535, 2020.

[36] B. Song, Z. Li, X. Lin, J. Wang, T. Wang, and X. Fu, "Pretraining model for biological sequence data," *Briefings in Functional Genomics*, vol. 20, no. 3, pp. 181–195, 2021.

[37] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, NIPS (Conference and Workshop on Neural Information Processing Systems), 2017.

[38] S. Min, S. Park, S. Kim, H. S. Choi, and S. Yoon, "Pre-training of deep bidirectional protein sequence representations with structural information," 2019, https://arxiv.org/abs/1912.05625.

[39] R. Rao, N. Bhattacharya, N. Thomas et al., "Evaluating protein transfer learning with tape," in *Advances in Neural Information Processing Systems*, NIPS (Conference and Workshop on Neural Information Processing Systems), 2019.

[40] K. Huang, J. Altosaar, and R. Ranganath, "ClinicalBERT: modeling clinical notes and predicting hospital readmission," 2019, https://arxiv.org/abs/1904.05342.

[41] J. Lee, W. Yoon, S. Kim et al., "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[42] P. Schwaller, T. Laino, T. Gaudin et al., "Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction," *ACS Central Science*, vol. 5, no. 9, pp. 1572–1583, 2019.

[43] J. J. Irwin, T. Sterling, M. M. Mysinger, E. S. Bolstad, and R. G. Coleman, "ZINC: a free tool to discover chemistry for biology," *Journal of Chemical Information and Modeling*, vol. 52, no. 7, pp. 1757–1768, 2012.

[44] D. Mendez, A. Gaulton, A. P. Bento et al., "ChEMBL: towards direct deposition of bioassay data," *Nucleic Acids Research*, vol. 47, no. D1, pp. D930–D940, 2019.

[45] J. Woosung and K. Dongsup, *RDKit: Open-Source Cheminformatics*, 2006, https://www.rdkit.org.

[46] W. Jeon and D. Kim, "FP2VEC: a new molecular featurizer for learning molecular properties," *Bioinformatics*, vol. 35, no. 23, pp. 4979–4985, 2019.