SMILES-BERT: Large Scale Unsupervised Pre-Training for Molecular Property Prediction

Sheng Wang
University of Texas at Arlington
Arlington, Texas
sheng.wang@mavs.uta.edu

Yuzhi Guo
University of Texas at Arlington
Arlington, Texas
yuzhi.guo@mavs.uta.edu

Yuhong Wang
National Center for Advancing
Translating Sciences, NIH
Rockville, Maryland
yuhong.wang@nih.gov

Hongmao Sun
National Center for Advancing
Translating Sciences, NIH
Rockville, Maryland
sunh7@mail.nih.gov

Junzhou Huang*
University of Texas at Arlington
Arlington, Texas
jzhuang@uta.edu

ABSTRACT

With the rapid progress of AI in both academia and industry, Deep Learning has been widely introduced into various areas in drug discovery to accelerate its pace and cut R&D costs. Among all the problems in drug discovery, molecular property prediction has been one of the most important problems. Unlike general Deep Learning applications, the scale of labeled data is limited in molecular property prediction. To better solve this problem, Deep Learning methods have started focusing on how to utilize tremendous unlabeled data to improve the prediction performance on small-scale labeled data. In this paper, we propose a semi-supervised model named SMILES-BERT, which consists of attention mechanism based Transformer Layer. A large-scale unlabeled data has been used to pre-train the model through a Masked SMILES Recovery task. Then the pre-trained model could easily be generalized into different molecular property prediction tasks via fine-tuning. In the experiments, the proposed SMILES-BERT outperforms the state-of-the-art methods on all three datasets, showing the effectiveness of our unsupervised pre-training and great generalization capability of the pre-trained model.

CCS CONCEPTS

• **Theory of computation → Semi-supervised learning**; **Structured prediction**; • **Applied computing → Molecular sequence analysis**; **Natural Language Modeling**; *Bioinformatics.*

KEYWORDS

Unsupervised Pre-training; Semi-supervised Learning; Molecular Property Prediction; Natural Language Modeling

ACM Reference Format:
Sheng Wang, Yuzhi Guo, Yuhong Wang, Hongmao Sun, and Junzhou Huang. 2019. SMILES-BERT: Large Scale Unsupervised Pre-Training for Molecular Property Prediction. In *10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (ACM-BCB '19), September 7–10, 2019, Niagara Falls, NY, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3307339.3342186

1 INTRODUCTION

The capability of accurate prediction of molecular properties is an essential key in the chemical and pharmaceutical industries. It benefits various academic areas and industrial applications such as improvement to rational chemical design, reducing R&D cost, decreasing the failure rate in potential drug screening trials, as well as speeding the process of new drug discovery [4]. The key problem of introducing Deep Learning into this area lies on embedding graph-like molecules onto a continuous vector space. Then the representations, as named molecular fingerprints, could be used for various applications such as molecular properties classification, regression, or generating new molecules. Instead of computing a basic property, traditional molecular fingerprints provide a description of a specific part of the molecular structure [27]. However, traditional molecular fingerprints require intensive manual feature engineering and strong domain knowledge. Besides, this kind of fingerprints is highly task-dependent, not general enough for other property prediction tasks [10].

The current success of deep learning in various areas and applications, e.g., image classification [12, 33], video understanding [1, 31, 34], medical imaging [15, 35, 42], and bioinformatics [39, 41], demonstrates that deep learning is a powerful tool in learning feature from data and good at task-related prediction. An increasing number of publications have introduced deep learning into molecular fingerprint learning [3, 39–41]. The models being introduced rely on two main deep learning structures: Recurrent Neural Networks (RNNs) [30] and Graph Convolutional Networks (GCNs) [17, 18]. For RNNs-based methods, molecules are represented as strings by Simplified Molecular-Input Line-Entry system (SMILES). In this way, the current successful models in natural language modeling could be utilized to extract high-quality

[*] Corresponding author.

ACM-BCB '19, September 7–10, 2019, Niagara Falls, NY, USA
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6666-3/19/09…$15.00
https://doi.org/10.1145/3307339.3342186

features from SMILES and make task-related predictions. GCNs-based methods consider the atoms in molecules as graph nodes and the chemical bonds as graph edges. These methods use graph convolutions to extract the feature then classify/regress the molecular properties. In general, it is not trivial to support RNNs-based methods for parallel training on multiple GPUs and multiple devices, and it needs different training tricks like gradient clipping and early stopping to assure the model convergence; GCNs-based methods usually have high computation complexity. It limits exploring more complicated methods for molecular properties prediction. Meanwhile, CNNs-based models [8, 31] for language translation and modeling have been developed and widely used. These methods could easily support parallel training. With the help of attention mechanism, the results even outperform a lot of RNN models.

The success of current deep learning methods highly relies on a large-scale labeled training samples. For many areas, the labeled sample number of image classification could easily reach several million or more. However, it is not the same situation with molecular property prediction. The cost of obtaining such scale of molecular properties with screening experiments is exceptionally high. It is similar to the case in natural language modeling that they have almost unlimited unlabeled data while a tiny portion has labels. The state-of-the-art framework to utilize the unlabeled data is the pre-training and fine-tuning framework [5]. It pre-trains the model in an unsupervised fashion then fine-tune the model on labeled data. Seq3seq Fingerprint model [41] first starts using this framework to involve large-scale unlabeled data in model training to improve the prediction performance. However, Seq3seq model is not very efficient since it uses an encoder-decoder structure, and the decoder is used as a scaffold and does not contribute to the final prediction.

The motivations of this paper are two-folded. First, we would like to build a powerful semi-supervised model utilizing the essential information in unlimited unlabeled data to improve the prediction performance with limited labeled data. Second, we would like our model to be efficient in training stage in two ways: 1) our model should naturally support parallel training to reduce pre-training time; 2) the model used for pre-training will all take part in the fine-tuning stage with no scaffolding part like the decoder of Seq3seq fingerprint [41]. Thus, in this paper, we propose a pre-training and fine-tuning two-stage framework named SMILES-BERT motivated by the recent natural language modeling work BERT [7]. The neural network structure is a fully convolutional net stacked of Transformer layers. In the pre-training task, SMILES-BERT is trained with unsupervised learning mechanism Masked SMILES Recovery on large scale unlabeled data. In the Masked SMILES Recovery task, the input SMILES will be randomly masked or corrupted, and the model is being trained to recover the original SMILES according to the information lying in the unmasked part of the input. After that, the model needs a slight fine-tuning with the labeled dataset to have good prediction performance. The proposed SMILES-BERT contains several benefits than the existing methods: 1) different from Seq2seq or Seq3seq model, SMILES-BERT does not require an encoder-decoder structure which is more efficient and the model could be more complicated given the same GPU memory; 2) SMILES-BERT is more natural to parallel training because of the fully convolutional structure; 3) The random masking method will having SMILES-BERT more general and able to avoid overfitting; 4) The attention mechanism is used in the Transformer layer which could potentially improve the prediction performance.

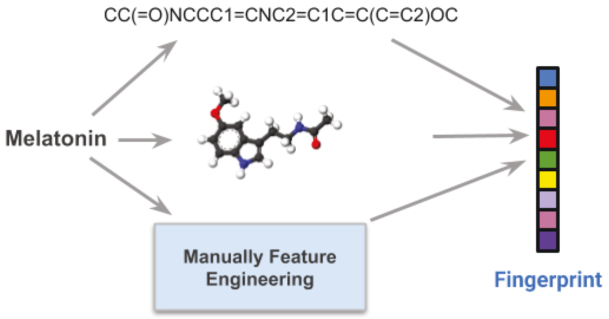

Figure 1: Mapping molecule to feature vector (Fingerprint) with different methods.

Our contributions of this paper could be summarize as:

- We propose a two-stage (pre-training and fine-tuning) model SMILES-BERT to utilize both unlabeled data and labeled data to have better molecular properties prediction performance.
- SMILES-BERT has better performance, outperforming a series of state-of-the art methods on three datasets.

The rest of the paper is organized as follows. Related work including both molecular property prediction and natural language modeling is summarized in Section 2. Section 3 gives a detailed introduction about the proposed SMILES-BERT including the two-stage training. We describe our experiment settings and results in Section 4. Following that is Section 5, which is the conclusion of this paper and potential future work.

2 RELATED WORK

Almost all the molecular property prediction methods or fingerprints could be concluded in Figure 1. The most important task is to embed the molecule into a continuous feature space for further task. Since molecules have different representation, these methods could be divided into three categories based on the input representation format being used: the manually feature engineering methods, the graph-based methods, and the sequence-based methods.

2.1 SMILES and canonical SMILES

To represent molecules with atoms and chemical bonds inside, the Simplified Molecular-Input Line-Entry system (SMILES) [37] is proposed to represent molecules in a simple way. SMILES is a line notation which represents the chemical structures in a graph-based definition, where the atoms, bonds and rings are encoded in a graph and represented in text sequences. One example of SMILES representation is shown in Figure 1: melatonin with structure $C_{13}H_{16}N_2O_2$, where corresponding SMILE representation is included as well as the 3D molecule structure. Simply speaking, the letters, e.g., C, N, generally represent the atoms, while some symbols like −, =, # represent the chemical bonds. SMILE system is not perfect given that the vanilla SMILE system is not a bijective mapping between SMILE sequence and a molecule. For example, a molecule could have multiple corresponding SMILE representations, e.g., *CCO*, *OCC* and *C*(*O*)*C*. To address this issue and provide a one-to-one mapping between SMILES and molecules, multiple canonicalization algorithms are invented to ensure the representation uniqueness of each molecular structure [21]. In this paper, all the SMILES are canonical.

2.2 Manually Designed Fingerprint

Traditionally, there is a class of molecular representation systems called molecular fingerprints. A fingerprint is basically a vector of a corresponding molecule as its continuous representation. Hence fingerprints can be thereafter fed into a machine learning system as an initial vector representation. A large number of previous studies have invented new fingerprint systems which can benefit future predictive tasks.

Many hash-based methods has been proposed to generate unique molecular feature representation [10, 11, 20]. One important class is called circular fingerprints. Circular fingerprints generate each layer's features by applying a fixed hash function to the concatenated features of the neighborhood in previous layer. One of the most famous ones is Extended-Connectivity FingerPrint (ECFP) [27]. However, due to the non-invertible nature of the hash function, the hash-bashed fingerprint methods usually do not encode enough task-related information and hence result in not good enough performance in properties prediction.

Another stream of traditional fingerprint methods are based on the biological experiments and the expertise knowledge and experience, e.g., [22, 28]. Biologists have figured out several important task-related sub-structures (fragments), e.g., *CC*(*OH*)*CC* for solubility prediction, and count those sub-structures as local features to produce molecular fingerprints. This kind of fingerprint methods usually work well for specific tasks, but could not generalize well for other tasks.

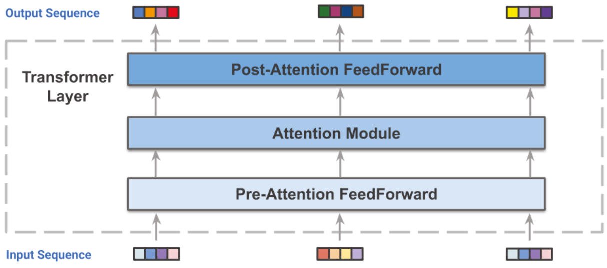

Figure 2: The structure of Transformer Layer.

2.3 Deep Fingerprints

The growth of deep learning has provided excellent flexibility and performance to learn molecular fingerprints from data samples, without explicit guides from experts [2, 8, 14, 29, 32, 39].

2.3.1 ***Graph-based fingerprint.***

Among all the graph-based molecular fingerprint, the state-of-the-art work is the neural fingerprint [3]. The neural fingerprint mimics the whole process of generating circular fingerprint but the hash function is replaced by a non-linear activated densely connected layer. The model of neural fingerprint is a deep neural network. To acquire enough labeled data, biologists need to perform a sufficiently large number of tests on chemical molecules, which is extremely expensive.

2.3.2 ***RNNs-based fingerprints.***

Recently, a few unsupervised fingerprint methods, e.g., seq2seq fingerprint [39], are proposed to alleviate the issue of insufficient labeled data. These models generally train deep neural networks to provide strong vector representations using a big pool of unlabeled data. The vector representation model is thereafter used for supervised training with any kind of classifiers. Since the deep models are trained with a sufficiently large data-set, the representation is expected to contain enough information to provide good inference performance. However, this type of methods are not trained with prediction tasks, meaning that the representation only adjusts to the recovery task of the original raw representation. It might not provide optimal inference performance for general prediction task. Seq3seq [41] is the first semi-supervised learning model for molecular property prediction. It has an Encoder-Decoder structure which could learn the fingerprints based on self-representation. Thus, it could utilize unlimited unlabeled data. However, the Encoder-Decoder framework limits its capability for property prediction. It is because the decoder of Seq3seq functions as a scaffold in pre-training stage and is barely useful in fine-tuning, but it has to consume the GPU memory in the pre-training stage. In this way, Seq3seq fingerprint is not computationally effective.

2.4 Transformer and BERT on Natural Language Modeling

Recently, there are several CNNs-based language models having excellent performance on various language modeling tasks [6, 7, 9, 24, 25, 31]. These methods use fully convolutional network structures instead of any RNNs blocks. With the help of self-attention mechanism [31], CNNs-based models could even outperform RNNs-based models. Among these methods, Transformer [31] is one of the most significant model building block. Furthermore, BERT [7] proposes to pre-train the Transformer encoders with two tasks: masked language learning, and continuous sentence classification. Both Transformer and BERT belongs to pre-train and fine-tuning framework, which could use the power of unlabeled data to initialize the parameters in the models, then promise good performance in following general language modeling tasks. This paper is inspired by Transformer and BERT, we keep the model used in BERT as our backbone with a few adaptations.

3 METHODOLOGY

In this section, the proposed SMILES-BERT is introduced step by step. First, we give the details of our backbone and its building block, i.e. Transformer Encoder. Then the Masked SMILES Recovery task used for pre-training our backbone on large scale unlabeled data will be introduced. Following that is the fine-tuning process for molecular properties prediction. The proposed model handles the molecules as sequences. Thus the inputs of SMILES-BERT are the tokenized molecules SMILES representations as shown in Figure3.

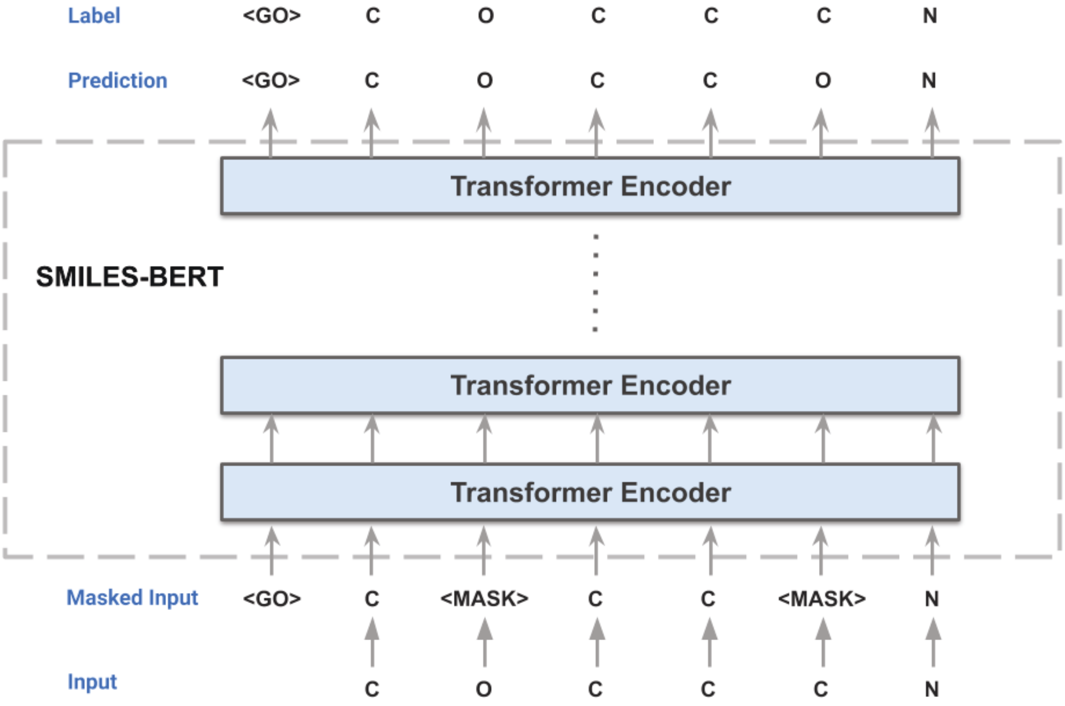

Figure 3: SMILES-BERT: pre-training stage.

3.1 Model Backbone and Transformer Layer

As shown in Figure 2, a transformer layer contains three components: a pre-attention feed forward neural network, a self-attention layer, and a post-attention feed forward neural network. The pre-attention feed forward is a fully-connected layer shared by all the input tokens. It maps the output features from former Transformer layer or the embedded features from the input into another nonlinear space. The post-attention works precisely in the same way, while the input is the output features after self-attention module.

RNNs-based methods utilize the sequential information naturally since the output from the former time step will be part of the input of the current time step. However, in Transformer Encoder, only using feed forward network could not bring temporal information from the sequence. The self-attention layer plays a crucial role to introduce the temporal relation into consideration for feature learning. For every time step, it could decide how to use information from other sequences by which is more related to itself.

The attention mechanism [31] used in Transformer encoder is named scaled dot-product attention. It maps the input data into three parts, a query matrix, a key matrix, and a value matrix. The query matrix works together with the key matrix to serve as the input of the Softmax. Then Softmax creates the attention weights, which will be applied to the value matrix to generate the output features with the attention on the whole sequence. The scaled doc-product attention is formulated as:

$$Z = Softmax\left(\frac{\left(XW^Q\right)\left(XW^K\right)^T}{\sqrt{d_k}}\right)XW^V,$$

Where $X \in \mathbb{R}^{N \times M}$ is the input feature matrix, W^Q, W^K, and $W^V \in \mathbb{R}^{M \times d_k}$ corresponds to the query weight matrix, the key weight matrix, and the value weight matrix. $\sqrt{d_k}$ is a scaling factor and Z is the output of the attention layer. It is the single head self-attention version. However, in the backbone, a more powerful version of the self-attention layer is used, the multi-head self-attention. Thus, different heads could pay attention to various aspects, making attention to the best power.

All the three components, the feed forward neural networks, and the self-attention layer are followed by a normalization layer to increase the generalization ability of the model. Besides, each of the components has a residual input to better utilize the original information.

The whole structure of the proposed model is shown in Figure 3. SMILES BERT contains a stack of Transformer Encoders with the self-attention mechanism.

3.2 Pre-training as Masked SMILES Recovery

The pre-training stage is shown in Figure 3. BERT uses a combination of two tasks to per-train the model, masked language learning and the consecutive sentences classification. Masked language learning is that given a partially masked sentence, using other visible tokens to predict the masked or corrupted ones. It is label-free so it could utilize all the unlabeled sentence in natural languages. The consecutive sentences classification is to classify if two sentences are consecutive, which is also label-free. However, different from natural language modeling, SMILES do not have a consecutive relationship. The masked language learning is still promising to pre-train the model with unlabeled SMILES and we name the task Masked SMILES Recovery.

We follow the way in BERT [7] to mask an input SMILES. First 15% tokens in a SMILES will be randomly selected for masking and the minimum token number per SMILES is one. For every selected token in a SMILES, it has 85% chance to be changed to <MASK> token. With 10% and 5% chances, it will be randomly changed to any other token in the dictionary or kept unchanged correspondingly. The original SMILES serve as ground truth for training the model but the loss is only computed based on the outputs of masked tokens and their ground truths. By randomly masking the input SMILES, the dataset used for pre-training model is enlarged. The randomness could increase the generalization ability of model and keep it from over-fitting.

The tokens are first embedded into the feature space. Besides the token embedding, positional embedding is also included to add sequential information used in self-attention layer to utilize the temporal information of the inputs.

The proposed SMILES-BERT differs from BERT in the following perspectives: 1) SMILES-BERT uses the single Masked SMILES Recovery on large scale unlabeled dataset. 2) We do not include the segmentation embedding used in BERT into our model since we do not involve the continuous sentences training.

3.3 Fine-tuning for Molecular Property Prediction

The fine-tuning stage is shown in Figure 4. After pre-training on the large scale unlabeled SMILES data, the model has a non-trivial initialization. During the pre-training, we pad every input SMILES with the leading token <GO>. In the fine-tuning stage, the model output corresponding to the <GO> token is used for molecular property prediction.

A simple trainable classifier/regressor is added to the output of the <GO> token. Then the small scale of the labeled dataset is used for fine-tuning the model to predict specific molecular property.

The proposed SMILES-BERT has several advantages. First, it could use large scale unlabeled dataset for model pre-training. It not only contains the dataset itself, by randomly masking the inputs, but the dataset could also be enlarged into theoretically infinite. Second, unlike encoder-decoder structures in [39], the whole model involving in pre-training will be used in fine-tuning. Thus, the model could be more complicated since it does not need scaffolding parameters (the decoder parameters in Seq2seq and Seq3seq fingerprint models).

3.4 Model Structure

In this paper, the proposed SMILES-BERT contains six Transformer Encoder layers. In each Transformer layer, the pre-attention and the post-attention fully-connected layers embed input features into a feature space with size 1024. For the self-attention block, SMILES-BERT uses a four-head multi-attention mechanism. Note that the layers and number of attention heads are less than the base BERT [7], which consists of twelve Transformer Encoder layers with 3072 fully-connected embedding size and twelve attention heads in attention block. It is because SMILES are relatively simpler than the natural language sequences. Besides, the vocabulary of SMILES is much less than the vocabulary of natural language. We have tried the base structure setting of BERT to molecular properties prediction and it does not provide a noticeable improvement. Then we keep the SMILES-BERT in the current setting since it is better for the model to have less computation and memory requirements in practice.

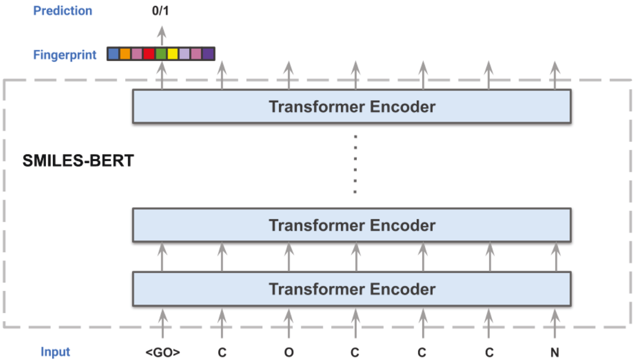

Figure 4: SMILES-BERT: fine-tuning stage.

4 EXPERIMENTAL RESULTS

In this section, we describe all our experiments related details. First, the implementation details are given. Then we include the detailed settings in both pre-training and fine-tuning stages. Following that is a brief introduction to the datasets we include in our experiments. At last, we list the state-of-the-art methods used in our comparison and demonstrate the power of the proposed SMILES-BERT with a thorough discussion of the experimental results.

4.1 Implementation Details

The proposed SMILES-BERT is implemented with the FairSeq [23], which is Facebook AI Research Sequence-to-Sequence Toolkit written in Python and PyTorch. Along with the proposed SMILES-BERT, we also implement a series of fingerprint models based on modern natural language sequence learning models including RNNs-based models [19, 38, 41] and CNNs-based models [6, 9, 24, 25, 31] models.

4.2 Experimental Settings

4.2.1 ***Pre-training.*** During the unsupervised pre-training stage, SMILES are tokenized into tokens as the feeding inputs to SMILES-BERT. As the Masked SMILES Recovery stage, the tokens are randomly selected to be masked with the masking strategy as described in Section 3.2. Note that the minimal number of masked token is set as one. Thus, each of the input SMILES contains at least one masked token. In this way, the pre-training dataset is enlarged with randomness. With training on such dataset, the generalization capability of proposed SMILES-BERT is enhanced.

The pre-trained dataset we use for SMILES is ZINC [13]. Zinc is a free database of commercially-available compounds for virtual screening. ZINC contains over 35 million purchasable compounds in ready-to-dock, 3D formats. ZINC is provided by the Irwin and Shoichet Laboratories in the Department of Pharmaceutical Chemistry at the University of California, San Francisco (UCSF). In SMILES-BERT, we only use the SMILES of the molecules with no additional label to pre-train the SMILES-BERT, strengthening the model prediction capability than only using the labeled dataset. To verify the pre-train model, we randomly keep 10000 samples for validation and another 10000 for evaluation. The number in the training set ends up to 18,671,355.

We use Adam optimizer [16] as the pre-training optimizer. To better initialize the proposed model, a warm-up strategy is introduced for the first 4000 training steps. During the warm-up, the learning rate increases from 10^{-9} to 10^{-4}. We notice that the warm-up stage is crucial in SMILES-BERT pre-training. Without it, the model tends not to converge even after a long time training. After the warm-up finishes, the learning rate starts from 10^{-4} with the inversed-square-root updating strategy. The Adam betas are $(0.9, 0.999)$ and the weight decay is 0.1. The batch size is set to 256 and the dropout is set to 0.1.

We pre-train SMILES-BERT for 10 epochs on ZINC dataset. We use the exact recovery rate to evaluate the pre-train model. The exact recovery rate on the ZINC validation dataset is 82.85%, meaning 82.85% masked SMILES could be exactly recovered by the information from the unmasked part.

Table 1: Parameters and Performances Contrast between Two Structures of SMILES-BERT

	layers	att-heads	ffn-dim	accuracy
SMILES-BERT	6	4	1024	**0.9154**
SMILES-BERT (large)	12	12	3072	0.9147

4.2.2 ***Fine-tuning.*** The supervised fine-tuning stage is based on the pre-trained model. As the pre-training stage, we use Adam optimizer for fine-tuning. The learning rate is not sensitive. We have tried several learning rates such as 10^{-5}, 10^{-6}, 10^{-7} and all the learning rate could get very good prediction results. Besides, we also test several different learning rate updating strategies such as no-updating, inversed-square-root updating. It turns out the updating strategy is not important for the training results. Thus, we simply choose not to update the learning rate in the fine-tuning stage.

In all our experiments, we fine-tune the model with each of the labeled datasets for 50 epochs and we choose the best model on validation data for the final evaluation.

4.3 Datasets Description

To evaluate our methods we use three datasets, LogP dataset, PM2 dataset and PCBA-686978 dataset in our experiments. The three datasets vary in not only properties but also the size of datasets. We would like to see if the pre-trained model could adapt well to fine-tuning with different molecular properties and different dataset sizes. The intrinsic logic of the experimental settings is from small-scale dataset (LogP) to large-scale datasets (PM2 and PCBA), from nonpublic datasets (LogP and PM2) to public dataset (PCBA).

4.3.1 ***LogP.*** LogP dataset is obtained from the National Center for Advancing Translational Sciences (NCATS) at National Institutes of Health (NIH). LogP dataset contains a total of 10,850 samples. Each sample contains a pair of a SMILES string and a water-octanol partition coefficient (LogP) value. The value is continuous and we use the threshold of 1.88 suggested by an NCATS expert to convert the dataset as a classification task. Samples with LogP value larger than 1.88 will be classified as the positive samples, while the opposites are considered the negative ones.

4.3.2 ***PM2.*** PM2 dataset is also obtained from NCATS at NIH. PM2 has 323,242 data samples with PM2 labels. Similarly, the continuous PM2 labels are set as positive if it is larger than 0.024896; otherwise as negative.

4.3.3 ***PCBA-686978.*** PCBA [26] is a group of public available dataset containing 128 datasets from PubChem [36]. We select one of the largest datasets, the dataset with ID 686978 among the 128 datasets to evaluate our method. PCBA-93 contains 302,175 samples.

For each of the three datasets, we randomly select 80% for training, 10% as the validation set and the rest 10% for evaluation.

4.4 Experimental Results

4.4.1 ***SMILES-BERT Structure Study.*** To compare what kind of structure of SMILES-BERT could have better performance on molecular properties prediction tasks, we compare two structures. We have not explored more structures for the following two reasons. 1) Any of the two structures has better prediction performance (accuracy) than state-of-the-art method but they do not have a noticeable performance difference. 2) SMILES-BERT training could take a long time. For a single GPU, it could take more than a week to train the model for 10 epochs. The detailed parameters of the two structures are listed in Table 1 as well as the performance on LogP dataset.

In Table 1, the att-heads stands for the attention heads in self-attention layers and the ffn-dim stands for the dimension for the shared fully-connected layer in each Transformer Layer. As shown in Table 1, the SMILES-BERT(large) is much more complicated than SMILES-BERT in all settings, while the performance is slightly worse. The performance difference could be caused by noise or randomness. Thus, we choose SMILES-BERT as our structure since it takes much less training cost and could have very good performance.

4.4.2 ***Comparison Methods.*** To prove the capability of molecular properties prediction performance of the proposed SMILES-BERT, we choose four state-of-the-art methods [3, 10, 39, 41] for comparison. These methods include one state-of-the-art manually designed fingerprint Circular Fingerprint [10], one graph-based neural network Neural Fingerprint [3], one unsupervised RNNs-based deep learning model Seq2seq Fingerprint [39], and one semi-supervised RNNs-based model [41]. We note the four methods as CircularFP, NeuralFP, Seq2seqFP, Seq3seqFP in all the following tables and figures.

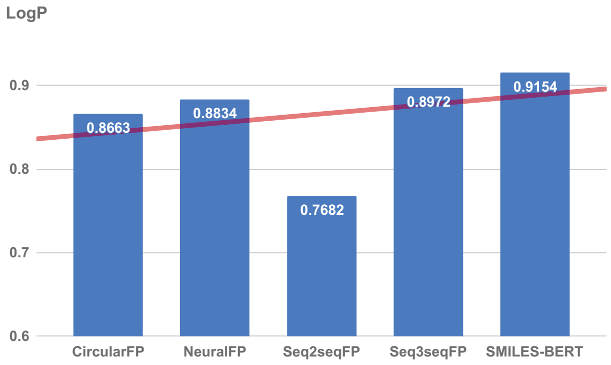

Figure 5: Prediction Results (Accuracy) on LogP Dataset.

4.4.3 ***Results of LogP.*** The prediction results for LogP data are shown in Figure 5. In this experiment, we use classification accuracy as the prediction metric to evaluate our model. As an unsupervised fingerprint, Seq2seq has reasonably lower performance than other methods. As a graph-based neural network, NeuralFP is slightly better than the manually designed CircularFP. Seq3seqFP and the proposed SMILES-BERT are both semi-supervised methods, which utilize large-scale unlabeled data. These semi-supervised methods have better performance than others. The proposed SMILES-BERT improves accuracy by around 2%. Since both of the SMILES-BERT and Seq3seq are pre-trained on Zinc, it shows that SMILES-BERT could better utilize the unsupervised information with the Masked SMILES Recovery task.

4.4.4 ***Results of PM2.*** PM2 is a much larger dataset than LogP. It contains 300 times data than LogP. It favors the supervised learning method because they could get better performance from more data samples. As shown in Table 2, unsupervised Seq2seqFP could not generate label-related fingerprint to have good prediction. The results of CircularFP and NeuralFP are similar. That CircularFP is slightly better than NeuralFP could be caused by that the graph-based neural network tends hard to train and tune in practice. Seq3seqFP slightly improves the performance compared to supervised method. The proposed SMILES-BERT achieves the better accuracy and it could get more than 5% improvement than Seq3seqFP. The results in Table 2 show that with the help of unsupervised pre-training, the proposed SMILES-BERT could have better representation and prediction capability after fine-tuning on the large dataset.

Table 2: Prediction Results (Accuracy on PM2 Dataset.)

Method	Accuracy
Circular Fingerprint [10]	0.6858
Neural Fingerprint [3]	0.6802
Seq2seq Fingerprint [39]	0.6112
Seq3seq Fingerprint [41]	0.7038
SMILES-BERT	**0.7589**

4.4.5 ***Results of PCBA-686978.*** We introduce a public dataset PCBA-686978 to compare the molecular property prediction performance on all the state-of-the-art methods. Figure 6 shows the results of five models. The trend is the same as the LogP and PM2 datasets. The proposed SMILES-BERT has 87.84% accuracy, which is 8% higher than the unsupervised Seq2seqFP.

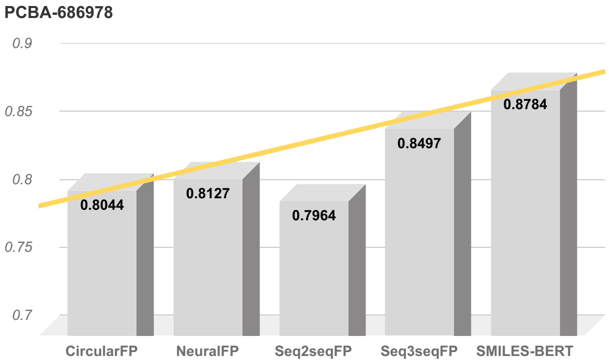

Figure 6: Prediction Results (Accuracy) on PCBA-686978 Dataset.

All the experiments on the three datasets demonstrate the power of the proposed SMILES-BERT. With the help of the large-scale unsupervised pre-training via the Masked SMILES Recovery task, SMILES-BERT could easily be fine-tuning towards the labeled dataset. It could have outstanding molecular property prediction performance, independently from whether the scale of the labeled dataset is small or large.

5 CONCLUSION AND FUTURE WORK

In the paper, to better use the numerous unlabeled molecular data and overcome some problems in current models, we have proposed a novel semi-supervised learning method SMILES-BERT for molecular properties prediction. The backbone of SMILES-BERT is BERT, a combination of Transformer Layer and attention mechanism. The semi-supervised method utilizes the power of unlabeled data through a large scale pre-training through a Masked SMILES Recovery task. The labeled dataset could be easily fine-tuned on the pre-trained model and could have very good prediction performance. In our experiments on three datasets, i.e., LogP, PM2 and PCBA, the proposed SMILES-BERT over the performance of various of state-of-the-art methods and future potential to deal with most kind of label datasets with a good generalization capability.

In this work, we utilize the Masked SMILES Recovery task in the pre-training stage corresponding to the masked language learning task in BERT [7]. However, BERT has another task to classify if two concatenated sentences are originally continuous. This task is to pre-train the classification with the input <GO> token. In SMILES-BERT, the classification capability of the model has not been involved in the pre-training stage. Thus, we could have the setting to include Quantitative Estimate of Druglikeness (QED) prediction as another task into the pre-training stage to warm up the classification capability of SMILES-BERT. It could potentially to increase the classification in the fine-tuning stage. We plan to design and include the QED prediction pre-training task in our future work.

ACKNOWLEDGMENTS

This work was partially supported by US National Science Foundation IIS-1718853 and the NSF CAREER grant IIS-1553687. The authors would like to thank NVIDIA for GPU donation and the NIH and UCSF for sharing the drug discovery datasets. The statements contained herein are solely of the authors and do not represent or imply concurrence or endorsement by NCI.

REFERENCES

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675* (2016).

[2] Han Altae-Tran, Bharath Ramsundar, Aneesh S Pappu, and Vijay Pande. 2016. Low Data Drug Discovery with One-shot Learning. *arXiv preprint arXiv:1611.03199* (2016).

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[4] Travers Ching, Daniel S Himmelstein, Brett K Beaulieu-Jones, Alexandr A Kalinin, Brian T Do, Gregory P Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz, Michael M Hoffman, et al. 2018. Opportunities and obstacles for deep learning in biology and medicine. *bioRxiv* (2018), 142760.

[5] Andrew M Dai and Quoc V Le. 2015. Semi-supervised sequence learning. In *Advances in Neural Information Processing Systems*. 3079–3087.

[6] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. 2017. Language modeling with gated convolutional networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 933–941.

[7] 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* (2018).

[8] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. 2015. Convolutional networks on graphs for learning molecular fingerprints. In *Advances in neural information processing systems*. 2224–2232.

[9] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 1243–1252.

[10] Robert C Glen, Andreas Bender, Catrin H Arnby, Lars Carlsson, Scott Boyer, and James Smith. 2006. Circular fingerprints: flexible molecular descriptors with applications from physical chemistry to ADME. *IDrugs* 9, 3 (2006), 199.

[11] Ye Hu, Eugen Lounkine, and Jürgen Bajorath. 2009. Improving the Search Performance of Extended Connectivity Fingerprints through Activity-Oriented Feature Filtering and Application of a Bit-Density-Dependent Similarity Function. *ChemMedChem* 4, 4 (2009), 540–548.

[12] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4700–4708.

[13] John J Irwin, Teague Sterling, Michael M Mysinger, Erin S Bolstad, and Ryan G Coleman. 2012. ZINC: a free tool to discover chemistry for biology. *Journal of chemical information and modeling* 52, 7 (2012), 1757–1768.

[14] Steven Kearnes, Kevin McCloskey, Marc Berndl, Vijay Pande, and Patrick Riley. 2016. Molecular graph convolutions: moving beyond fingerprints. *Journal of computer-aided molecular design* 30, 8 (2016), 595–608.

[15] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. 2018. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* 172, 5 (2018), 1122–1131.

[16] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[17] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

[18] Ruoyu Li, Sheng Wang, Feiyun Zhu, and Junzhou Huang. 2018. Adaptive Graph Convolutional Neural Networks. *arXiv preprint arXiv:1801.03226* (2018).

[19] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025* (2015).

[20] HL Morgan. 1965. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *J. Chemical Documentation* 5 (1965), 107–113.

[21] Greeshma Neglur, Robert L Grossman, and Bing Liu. 2005. Assigning unique keys to chemical compounds for data integration: Some interesting counter examples. In *International Workshop on Data Integration in the Life Sciences*. Springer, 145–157.

[22] Noel M O'Boyle, Casey M Campbell, and Geoffrey R Hutchison. 2011. Computational design and selection of optimal organic photovoltaic materials. *The Journal of Physical Chemistry C* 115, 32 (2011), 16200–16210.

[23] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038* (2019).

[24] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365* (2018).

[25] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/languageunsupervised/language understanding paper. pdf* (2018).

[26] Bharath Ramsundar, Steven Kearnes, Patrick Riley, Dale Webster, David Konerding, and Vijay Pande. 2015. Massively multitask networks for drug discovery. *arXiv preprint arXiv:1502.02072* (2015).

[27] David Rogers and Mathew Hahn. 2010. Extended-connectivity fingerprints. *Journal of chemical information and modeling* 50, 5 (2010), 742–754.

[28] Chetan Rupakheti, Aaron Virshup, Weitao Yang, and David N Beratan. 2015. Strategy to discover diverse optimal molecules in the small molecule universe. *Journal of chemical information and modeling* 55, 3 (2015), 529–537.

[29] Govindan Subramanian, Bharath Ramsundar, Vijay Pande, and Rajiah Aldrin Denny. 2016. Computational Modeling of β-secretase 1 (BACE-1) Inhibitors using Ligand Based Approaches. *Journal of Chemical Information and Modeling* 56, 10 (2016), 1936–1949.

[30] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*. 3104–3112.

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[32] Izhar Wallach, Michael Dzamba, and Abraham Heifets. 2015. AtomNet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery. *arXiv preprint arXiv:1510.02855* (2015).

[33] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. 2017. Residual attention network for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3156–3164.

[34] Sheng Wang, Zheng Xu, Chaochao Yan, and Junzhou Huang. 2019. Graph Convolutional Nets for Tool Presence Detection in Surgical Videos. In *International Conference on Information Processing in Medical Imaging*. Springer, 467–478.

[35] Sheng Wang, Jiawen Yao, Zheng Xu, and Junzhou Huang. 2016. Subtype Cell Detection with an Accelerated Deep Convolution Neural Network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 640–648.

[36] Yanli Wang, Jewen Xiao, Tugba O Suzek, Jian Zhang, Jiyao Wang, Zhigang Zhou, Lianyi Han, Karen Karapetyan, Svetlana Dracheva, Benjamin A Shoemaker, et al. 2011. PubChem's BioAssay database. *Nucleic acids research* 40, D1 (2011), D400–D412.

[37] David Weininger. 1970. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. In *Proc. Edinburgh Math. SOC*, Vol. 17. 1–14.

[38] Sam Wiseman and Alexander M Rush. 2016. Sequence-to-sequence learning as beam-search optimization. *arXiv preprint arXiv:1606.02960* (2016).

[39] Zheng Xu, Sheng Wang, Feiyun Zhu, and Junzhou Huang. 2017. Seq2seq Fingerprint: An Unsupervised Deep Molecular Embedding for Drug Discovery. In *BCB*.

[40] Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, et al. 2019. Are Learned Molecular Representations Ready For Prime Time? *arXiv preprint arXiv:1904.01561* (2019).

[41] Xiaoyu Zhang, Sheng Wang, Feiyun Zhu, Zheng Xu, Yuhong Wang, and Junzhou Huang. 2018. Seq3seq fingerprint: towards end-to-end semi-supervised deep drug discovery. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. ACM, 404–413.

[42] Xinliang Zhu, Jiawen Yao, Feiyun Zhu, and Junzhou Huang. 2017. WSISA: Making Survival Prediction from Whole Slide Pathology Images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.