OXFORD

*Briefings in Bioinformatics*, 2023, **24(6)**, 1–13

https://doi.org/10.1093/bib/bbad398

**Problem Solving Protocol**

# FG-BERT: a generalized and self-supervised functional group-based molecular representation learning framework for properties prediction

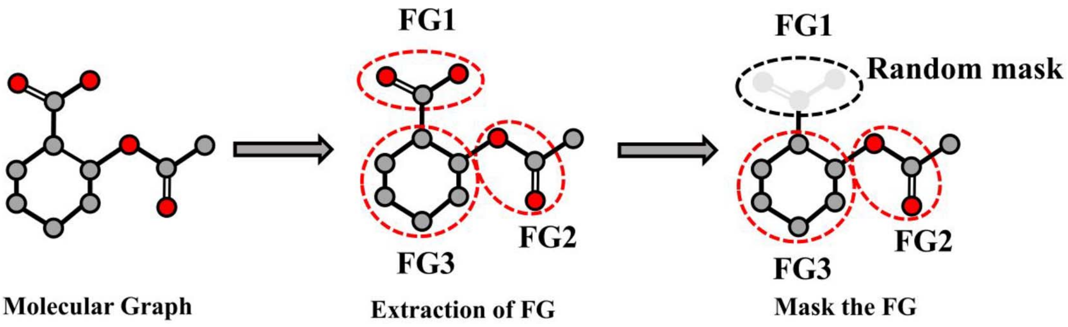
Biaoshun Li, Mujie Lin, Tiegen Chen and Ling Wang

Corresponding author: Ling Wang, Guangdong Provincial Key Laboratory of Fermentation and Enzyme Engineering, Joint International Research Laboratory of Synthetic Biology and Medicine, Ministry of Education, Guangdong Provincial Engineering and Technology Research Center of Biopharmaceuticals, School of Biology and Biological Engineering, South China University of Technology, Guangzhou 510006, China. Tel.: 020-39380602; E-mail: lingwang@scut.edu.cn

## Abstract

Artificial intelligence-based molecular property prediction plays a key role in molecular design such as bioactive molecules and functional materials. In this study, we propose a self-supervised pretraining deep learning (DL) framework, called functional group bidirectional encoder representations from transformers (FG-BERT), pertained based on ∼1.45 million unlabeled drug-like molecules, to learn meaningful representation of molecules from function groups. The pretrained FG-BERT framework can be fine-tuned to predict molecular properties. Compared to state-of-the-art (SOTA) machine learning and DL methods, we demonstrate the high performance of FG-BERT in evaluating molecular properties in tasks involving physical chemistry, biophysics and physiology across 44 benchmark datasets. In addition, FG-BERT utilizes attention mechanisms to focus on FG features that are critical to the target properties, thereby providing excellent interpretability for downstream training tasks. Collectively, FG-BERT does not require any artificially crafted features as input and has excellent interpretability, providing an out-of-the-box framework for developing SOTA models for a variety of molecule (especially for drug) discovery tasks.

## Graphical Abstract



***Keywords***: molecular property prediction; FG-BERT; molecular representations; deep learning; self-supervised learning

## INTRODUCTION

Accurate prediction of molecular properties is of great significance of functional molecule design and discovery, especially for drug molecule discovery, as it can be used in the early stage of drug discovery process to quickly identify active molecules with ideal properties and/or filter out unsuitable molecules [1]. Typically, molecular representation is the basis of molecular property prediction; therefore, how to obtain effective molecular characterization is an important problem to be solved in the field of molecular property prediction. Current molecular representations can be divided into five categories: molecular descriptors, fingerprints, graphs, molecular strings and molecular images. Based on these predefined molecular representations, practitioners can build quantitative structure–activity/property relationship (QSAR/QSPR) models with machine learning (ML) and deep learning (DL) for the prediction of molecular properties.

**Biaoshun Li** is a graduate student at South China University of Technology. His current research interests include machine learning, pretraining and artificial intelligence-aided drug discovery (AIDD).
**Mujie Lin** is an undergraduate student at South China University of Technology. Her research interests include machine learning and bioinformatics.
**Tiegen Chen** is a principal investigator in Medicinal Chemistry within Zhongshan Institute for Drug Discovery, Shanghai Institute of Materia Medica, Chinese Academy of Sciences. His current focus is organic synthesis; discovery and development of novel antiviral drugs.
**Ling Wang** is an associate professor at South China University of Technology. He received a PhD from the School of Pharmaceutical Sciences at the Sun Yat-Sen University in 2014. His research focus on computer-aided drug design (CADD), artificial intelligence-aided drug discovery (AIDD) and medicinal chemistry.
**Received:** June 30, 2023. **Revised:** September 25, 2023. **Accepted:** October 14, 2023
© The Author(s) 2023. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

Downloaded from https://academic.oup.com/bib/article/24/6/bbad398/7337693 by guest on 01 August 2024

In general, the accuracy of traditional ML-based QSAR/QSPR models depends heavily on how appropriate molecular representation is selected [2]. That is, traditional ML methods require chemists to manually develop a set of rules that encode the relevant structural information, pharmacophore characteristics and/or physicochemical properties of molecules into fixed-length vectors [3], such as commonly used molecular fingerprints and descriptors. However, the design and selection process is time-consuming and error-prone, making the scalability and versatility of molecular descriptors poor, which in turn leads to certain shortcomings in the field of molecular property prediction of descriptor-based ML models. DL-based models for predicting molecular properties differ from the traditional ML-based models in that they do not require the manual selection of the most important descriptors related to task attributes from a large number of predefined and computable molecular descriptors [4]. Currently, DL-based models can be usually constructed by using molecular graphs [5], SMILES sequences [6, 7] and molecular images [8] as input features. Among them, graph neural networks (GNNs) models have become a research hotspot in the field of molecular property prediction, because molecules are natural graph structures. Many supervised graph-based DL models [9–14] have been developed in recent years and showed considerable performance for molecular property tasks. However, the accuracy of such supervised GNN model depends on the size of the data volume and even performs worse than the descriptor/fingerprints-based traditional ML models on small datasets [15]. In addition, GNN is prone to the problem of oversmoothing, and the number of layers of the model is usually in the range of 2–4, which also limits the ability of the model to extract molecular features [6].

Very recently, pretraining models have been proposed to solve the above problems. They can learn useful molecular representations from a large amount of unlabeled data by setting certain pretraining strategies, such as contrast learning [16] and masked language learning [17], and then transfer the knowledge to downstream tasks for molecular property prediction. Some pretrained models have been developed and achieved competitive performance in the field of molecular property prediction [6, 18–21], compared with traditional supervised ML and DL models. Specifically, they pretrained the model by building their own pretraining tasks and then fine-tuning the model for molecular property prediction. For example, K-BERT can extract chemical information from SMILES like a chemist by utilizing three pretraining tasks based on atomic feature prediction, molecular feature prediction and comparative learning [6]. Mole-BERT employs an encoder of the VQ-VAE variant as a context-aware disambiguator to encode atoms into meaningful discrete values, expanding the atomic vocabulary and mitigating the significant quantitative divergence between atoms and rare atoms [21]. It predicts atomic discrete values by randomly masking them and pretraining the GNN to predict them. For graph-level pretraining, Mole-BERT [21] proposes Ternary Mask Comparison Learning to model heterogeneous semantic similarities between molecules, which is particularly effective for molecular retrieval and can match or exceed state-of-the-art (SOTA) methods in a fully data-driven manner. However, existing pretrained models do not pay attention to important functional group (FG) information in the molecular structure. As we all know, the structure of a molecule determines its various properties, and the structure of FGs within the molecule, as important components of the molecule, is often closely related to the properties of the molecule [22–25].

In this study, we developed a new masked chemical language pretraining framework (named functional group bidirectional encoder representations from transformers (FG-BERT), Figure 1) for learning chemical semantic and structural information from a large-scale unlabeled molecular corpus. FG-BERT has two important improvements: (1) it masks FGs in molecules to perform large-scale pretrained recovery predictions with high accuracy; (2) it utilizes a self-supervised pretraining learning framework to learn useful molecular representations from ∼1.45 million drug-like molecules with diverse biological activities. Compared with SOTA ML and DL methods, extensive experimental results demonstrate that FG-BERT is highly accurate in a variety of molecular properties prediction tasks on multiple benchmark datasets. In addition, FG-BERT can automatically learn to focus on FGs related to the target properties through attention mechanisms (AMs), providing valuable clues for further molecular analysis, design and optimization.

# MATERIALS AND METHODS

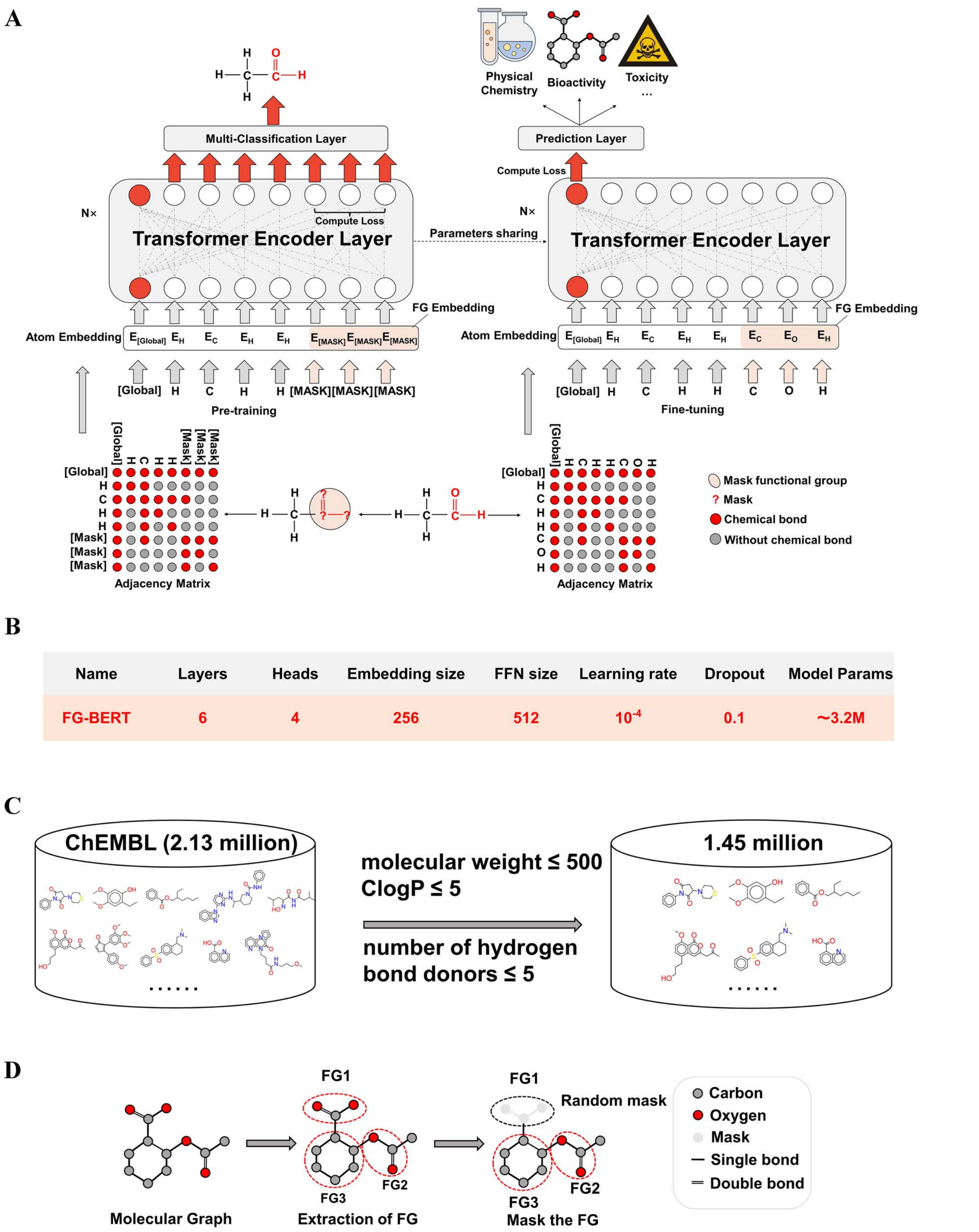
## Molecular corpus and benchmark datasets collection

The initial unlabeled molecular dataset (∼2.13 million molecules) was collected from ChEMBL (Version 30) [26], and then filtered through the three principles of drug-like properties (Figure 1C): molecular weight $\leq 500$, ClogP $\leq 5$ and number of hydrogen bond donors $\leq 5$. Eventually, a molecular corpus with ∼1.45 million molecules was obtained, and then was randomly split into training set and test set with a ratio of 9:1 for performing large-scale masking-recovery predictions during the pretraining process. For the fine-tuning process, we extensively evaluated the performance of the FG-BERT pretrained framework using three benchmark datasets. First, 15 commonly used public datasets (Table S1) relevant to drug discovery were collected and used to evaluate the performance of FG-BERT, including four physical chemistry datasets (ESOL, FreeSolv, Lipo and CEP) [27–30], two bioactivity and biophysical datasets (HIV, Malaria, MUV and BACE) [31–34], two quantum chemistry datasets (QM7 and QM8) [35] and four physiology and toxicity datasets (BBBP, Tox21, SIDER, ToxCast and ClinTox) [35–39]. Second, we also tested the performance of FG-BERT using 15 ADMET (Absorption, Distribution, Metabolism, Excretion and Toxicity) datasets [6] (Table S2). Finally, 14 breast cell line phenotype screening datasets [40] (Table S3) were used to evaluate the predictive power of FG-BERT.

## FG-BERT framework

FG-BERT is designed on the basis of the BERT model [41], which has two pretraining tasks, the masked language model (MLM) and the next sentence prediction task. In NLP, sentences are sequential, while molecular graphs differ from text in that the FGs and atoms in molecules are related through interconnected chemical bonds rather than a sequential order; therefore, FG-BERT does not require additional information about the binary values of FGs and atoms. In natural language processing, each word may be related to other words, so it is necessary to pay attention to all of them. However, in molecular graphs, atoms and FGs are primarily related to adjacent atoms or FGs connected by chemical bonds. Therefore, unlike BERT, we mainly focus on the localization connected by chemical bonds, so the information interaction between atoms and FGs is only through chemical bonds. Notably, FG-BERT can overcome the problem of oversmoothing due to the reconnection mechanism in BERT and has sufficient power to extract deep patterns in the molecular graph. As shown in Figure 1A, we use the adjacency matrix of molecules to control the information exchange of molecules. We add a GLOBAL node that exchanges information with all atomic and FG connections,

Downloaded from https://academic.oup.com/bib/article/24/6/bbad398/7337693 by guest on 01 August 2024



| Name | Layers | Heads | Embedding size | FFN size | Learning rate | Dropout | Model Params |
|---|---|---|---|---|---|---|---|
| FG-BERT | 6 | 4 | 256 | 512 | $10^{-4}$ | 0.1 | ~3.2M |

**Figure 1.** Schematic of the FG-BERT. (**A**) The FG-BERT framework and the corresponding pretraining and fine-tuning processes. (**B**) Hyperparameters of the FG-BERT pretraining model. (**C**) Molecular corpus screening process. (**D**) FGs masking process. FG, functional group.

Downloaded from https://academic.oup.com/bib/article/24/6/bbad398/7337693 by guest on 01 August 2024

the output of which can be considered as the final molecular representations for solving downstream classification or regression tasks [18]. Since the GLOBAL node is connected to all the nodes, it also takes into account the long distance dependency problem to some extent. Unlike BERT, our pretraining task directly masks language modeling i.e. masking FGs predictions, and then learns the chemical information in molecules and extracts the molecular representations to fine-tune the downstream tasks.

The FG-BERT framework consists of three parts: an embedding layer, a transformer layer and a pretraining/prediction head (Figure 1A). For each molecule fed into the model, we add a super node connected to all FGs and atoms and convert the SMILES-formatted molecule into molecular graph according to the chemical bond relationships.

At the embedding level, the atomic list $w = (a_1, a_2, a_3, \ldots, a_N)$ is embedded into the distribution space $x = (x_1, x_2, x_3, \ldots, x_N)$, $x_i \in R^{d_{\text{model}}}$ through an embedding matrix $D \in R^{V \times d_{\text{model}}}$, where $V$ is the size of the vocabulary and $d_{\text{model}}$ is the embedding size defined in Figure 1B. After the molecules pass through the embedding layer, an embedding vector is obtained, which is then transferred to the transformer layer. In the transformer layer, each node uses the AMs to aggregate information from neighboring nodes, and the message delivery process for a single node is described below

$$q_i = W_q x_i \tag{1}$$

$$k_i = W_k x_i \tag{2}$$

$$v_i = W_v x_i \tag{3}$$

$$s_{i,j} = \frac{\text{dot}\,(q_i, k_j)}{\sqrt{d_k}}, j \in N_i \tag{4}$$

$$a_{i,j} = \text{softmax}\,(s_{i,j}) = \frac{e^{s_{i,j}}}{\sum_{j \in N_i} e^{s_{i,j}}} \tag{5}$$

$$m_i = \sum_{j \in N_i} a_{i,j} v_j\ . \tag{6}$$

$x_i$ is the representation of input node $i$,$W_q$, $W_k$, $W_v$ is the learnable matrix shared by all nodes, $W_q, W_k, W_v \in R^{d_k \times d_{\text{model}}}$, $d_k = d_{\text{model}}/H$, $H$ is the number of heads of the model, as defined in Figure 1B, and $N_i$ denotes all the neighbors of node $i$.

We employ a multiheaded AM, where the above processes are executed $H$ times independently, and then the results are stitched together and passed into a linear transformation according to

$$M_i = W_0 \text{concat}\,(m_i^1, m_i^2 \ldots, m_i^K) \tag{7}$$

$W_0 \in R^{d_{\text{model}} \times d_{\text{model}}}$ is also the learnable matrix shared by all nodes.

To solve the oversmoothing problem in ordinary GNN, we use the same residual connectivity and layer normalization mechanism as transformer according to the following equation:

$$h_i = \text{layernorm}\,(x_i + M_i)\,. \tag{8}$$

To enhance the expressiveness of the model, we pass the output $h_i$ of the multi-headed attention sublayer to the feedforward sublayer, and we also use feed-forward network (FFN) sublayer connectivity and layer normalization mechanisms according to

$$\text{FFN}\,(h_i) = W_2\ \text{gelu}\,(\,W_1 h_i + b_1) + b_2 \tag{9}$$

$$o_i = \text{layernorm}\,(\,\text{FFN}\,(h_i) + h_i) \tag{10}$$

where $W_1 \in R^{d_{\text{hidden}} \times d_{\text{model}}}$, $W_2 \in R^{d_{\text{model}} \times d_{\text{hidden}}}$, $d_{\text{hidden}}$ equals FFN size defined in Figure 1B, and gelu denotes the activation function, called the Gaussian error linear unit.

The transformer layer in FG-BERT is executed multiple times, as determined by the parameters layers in Figure 1B. The pretraining head is different from the prediction head, but they both consist of two layers of FFN, using cross entropy as the loss function for the classification tasks and root mean square error (RMSE) as the loss function for the regression tasks. The activation function used for the pretrained model is GELU, and the activation function used for the downstream classification and regression tasks is LeakyReLU. Ultimately, the total number of parameters in the FG-BERT model is about 3.2 million.

## Input molecular representations of FG-BERT

To conveniently represent atoms and FGs in molecular graph, we construct an atom dictionary based on the frequencies of various atoms appearing in the pretrained molecular corpus. Statistics on the pretrained dataset used by FG-BERT show that there are 14 types of atoms that occur more than 1000 times, represented by their respective element symbols, while other atoms occur less than 1000 times, which we collectively refer to as [UNK] to represent them. In addition, to facilitate the subsequent downstream tasks, we added a super node to the molecular graph, denoted by [GLOBAL]. We use [MASK] to denote masked FGs (Figure 1D). Since a FG is usually composed of multiple atoms, when a FG is masked, there are usually multiple atoms that are [MASK]. We therefore construct a randomly selected list of FGs and use this list to identify the FGs that need to be masked, while the unselected FGs remain untouched. Thus, the dictionary includes the following tokens: [H], [C], [N], [O], [S], [F], [Cl], [Br], [P], [I], [Na], [B], [Se], [Si], [UNK], [MASK] and [GLOBAL]. In the present study, FGs are generated using RDKit software (http://www.rdkit.org/).

## Pretraining strategy

The proposed pretraining strategy used in this study is very similar to BERT. First, we iterate over all molecules based on a predefined list of FGs (Figure 2), so that we can get a list of FGs corresponding to each molecule in the molecular corpus. Like BERT, we then randomly choose 15% of the FGs in a molecule for masking. Unlike BERT, the operation of FG substitution is abandoned because FG substitution of molecules may cause many nonconforming chemical rules to occur, so we have a 90% probability of being masked in the 15% of selected FGs. During pretraining, our loss is only calculated at the masked FGs, and the other 10% probability of remains unchanged.

## Training protocol, hyperparameter optimization and evaluation

### *Pretraining stage*

During the pretraining phase, each molecule in SMILES format is first converted into a two-dimensional (2D) undirected graph using RDKit software, and an additional super node is then added to each molecular graph. According to the FG-BERT pretraining strategy, FGs are masked randomly and finally the molecular graph is passed into the FG-BERT model to predict the masked FGs. For molecules with only a few FGs, we mask at least one FG. Meanwhile, we do not touch molecules that do not contain FG, although there are few such molecules in molecular corpus. Our model is trained on the batch gradient descent algorithm

Downloaded from https://academic.oup.com/bib/article/24/6/bbad398/7337693 by guest on 01 August 2024

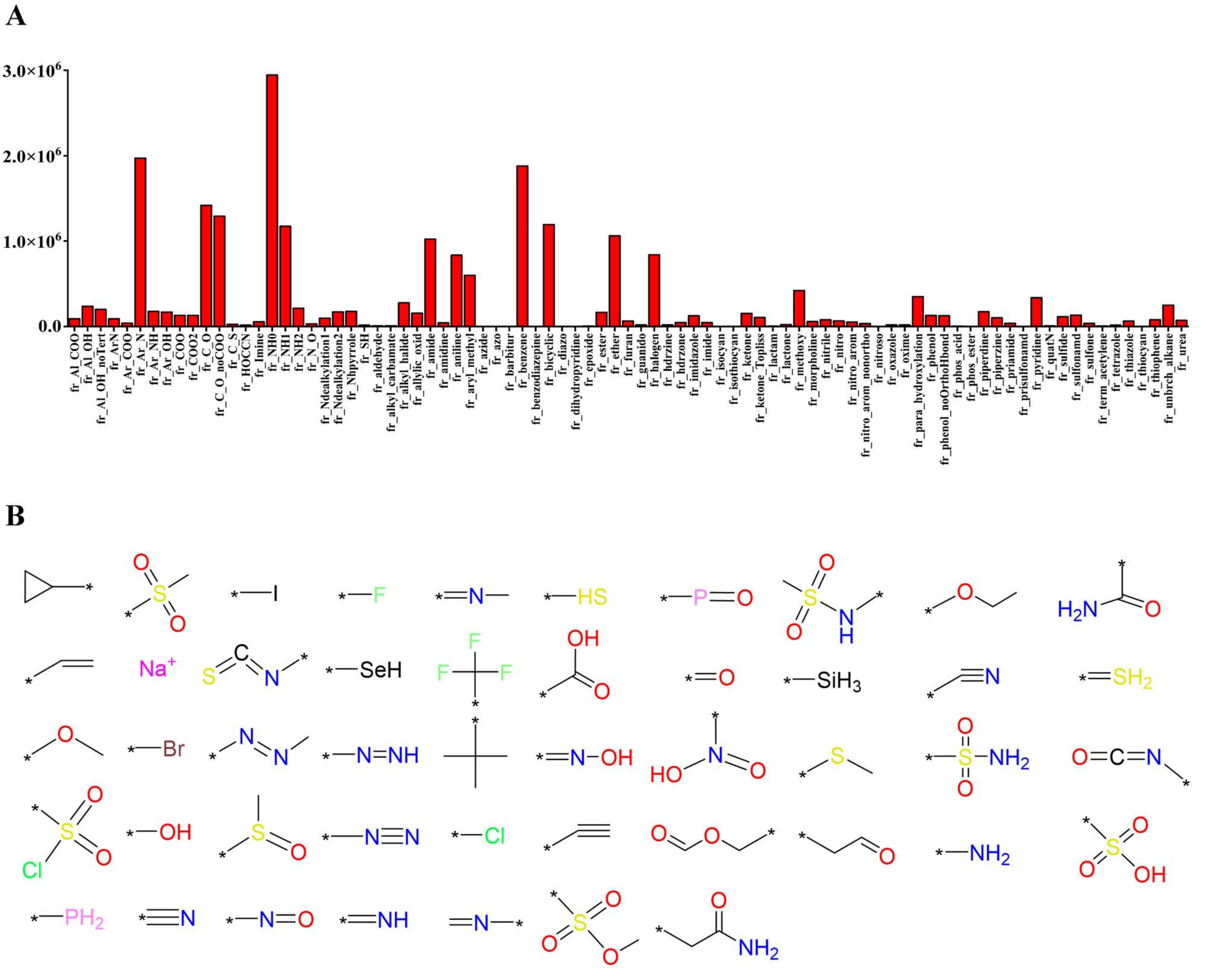


**Figure 2.** (**A**) Number of FGs statistics. (**B**) The FGs list. The detailed statistics on the number of FGs in the pretrained corpus and FGs list can be visualized at https://github.com/idrugLab/FG-BERT/tree/main/FGs. *FGs*, functional groups.

and the Adam optimizer [42]. The learning rate of the pretrained model is set to $10^{-4}$, and the batch size is set to 16. To evaluate the performance of FG-BERT pretraining, the pretraining masking strategy is employed to mask the molecules in the test set, and the recovery rate is then computed as an evaluation metric. The cross-entropy loss function is used to calculate the loss of FGs. FG usually consists of multiple atoms, so FG can only be fully restored to correctness i.e. if all atoms are correctly predicted, the loss is minimized and zero. If the masked FG is only partially recovered correctly, a loss will be incurred. The model minimizes the loss through multiple rounds of training and parameter updates so that the model performs prediction task more accurately. We therefore use the sum of cross-entropy losses of multiple atoms as a loss measure for FGs masking recovery pretraining.

### Fine-tuning stage

In the fine-tuning phase, we remove the pretrained model from its pretrained head and add a two-layer fully connected neural network on the Transformer encoder corresponding to the super node, called the prediction head. The dropout strategy [43] is used to avoid overfitting, and since the dropout has a large impact on the specific downstream task, different dropout values are therefore chosen according to different downstream tasks. According to the previous results of MG-BERT model [18], we also set such dropout values in the range of [0, 0.5] to make the optimal selection. In addition, the Adam optimizer is used as a fine-tuning optimizer for each task with a limited hyperparameter scan as follows: batch size: {8, 16, 32, 48, 64}, learning rate: [ln (3e-5), ln (15e-5)] and the number of head: is {4, 8}. The detailed hyperparameters for each downstream task are listed in Table S4.

To compare fairly with existing methods, we therefore chose the same data, data splitting method and ration, as well as evaluation metrics. For example, all three benchmark datasets are split into training, validation and testing sets with an 8:1:1 ratio. The Hyperopt-Python package [44] is employed to perform Bayesian optimization of hyperparameters that include the number of multihead concerns, the dropout rate, the batch size and the learning rate. The results of the validation set are used to determine the hyperparameters of the final model to improve the generalization ability, and the final performance of the FG-BERT model is represented by the results of the test set. In addition, FG-BERT uses early stopping to avoid overfitting during training, with the tolerance value set to 30 and the maximum epoch set to 200.

The RMSE and the area under the receiver operating characteristic curve (ROC-AUC) were selected as final evaluation metrics for downstream fine-tuned classification and regression tasks, respectively. In addition, to reduce random errors and to ensure the reliability of the results, we evaluated FG-BERT models based

Downloaded from https://academic.oup.com/bib/article/24/6/bbad398/7337693 by guest on 01 August 2024

on 10 different random seeds for each dataset and computed the average and standard deviate of evaluation metrics to represent the final result. FG-BERT DL framework was developed by the Tensorflow software, and all FG-BERT pretraining and fine-tuning are trained on SCUTGrid (SCUT Supercomputing Platform) GPU [NVIDIA Corporation GV100GL (Tesla V100 PCIe 32 GB)] and CPU [Intel(R) Xeon(R) Silver 4216 CPU@2.10 GHz] for training.

# RESULTS AND DISCUSSION

## The statistics and masking of FGs in FG-BERT model

It is well known that the properties of molecules are closely related to their structures, and FGs are important substructures in molecules. The presence of various FGs is a distinctive feature of molecular compounds [45]. We counted the number of FGs in the molecular corpus dataset (1 456 893 molecules) used for pretraining. As shown in Figure 2A, a large number of FGs are widely present in drug-like small molecules, suggesting that FGs play a crucial role in the properties of small molecule drugs.

Currently, our list of predefined FGs includes 47 common FGs (Figure 2B). The reason why this list is smaller than the 85 FGs shown in Figure 2A is that when counting the number of FGs, FGs attached to benzene rings and aliphatic chains are considered as separate FGs in the RDKit software. For example, a carboxyl group attached to a benzene ring and a methyl group are considered to be two different FGs. However, during the pretraining process of FG-BERT, we treat them as a unique FG, so our predefined list of FGs contains only 47 FGs instead of 85 FGs. In addition, considering the widespread presence of ring structures in small molecule drugs, these ring structures may be important substructures/fragments that affect the properties of drug molecules. Therefore, the ring structures are also considered as special FGs in this study, and subjected to perform the same masking and recovery pretraining operations as the other FGs (Figure 1D).

Similar to BERT and MG-BERT [18, 41], 15% of FGs in a molecule are randomly selected for masking, and the detailed masking process is shown in Figure 1D. Based on this, our FG-BERT focuses on the information of FGs in molecules, learns useful semantic and structural information in molecules by masking recovery of FGs for pretraining, and then extracts molecular representations to finally perform prediction of downstream tasks. After 20 epochs of pretraining, the FG-BERT model can recover FGs with an accuracy of 98.70%, demonstrating the effectiveness of masking FGs to construct chemical language pretraining model. After the pretraining, the obtained weights can be then used for downstream classification and regression tasks.

## Performance of the FG-BERT on the public benchmark datasets

We utilized 15 benchmark datasets related to drug discovery [20, 35], including eight classification tasks and seven regression tasks (Table S1), to evaluate the predictive ability of the FG-BERT pretrained model. Seventeen pretraining DL methods (Table 1) were collected from Xia *et al.* and used as baselines [21]. We strictly follow the data splitting strategy of Mole-BERT for fairly comparison. For example, we use scaffold splitting and use 10 different random seeds (0–9) for classification tasks and 3 different random seeds (1–3) for regression tasks to split the datasets into training, validation and test sets in an 8:1:1 ratio. The average and standard deviation of ROC-AUC or RMSE for each dataset were calculated as the final results of FG-BERT. Detailed performance results are summarized in Tables 1 and 2.

For classification tasks, FG-BERT performed best in five out of eight datasets (Table 1), including Tox21 (AUC = 0.784), ToxCast (AUC = 0.663), Sider (AUC = 0.640), ClinTox (AUC = 0.832) and Bace (AUC = 0.845). In addition, FG-BERT achieved overall best performance on these eight classification tasks, with the highest AUC value of 0.7492 (Table 2). For regression tasks, FG-BERT performed best on all four benchmark datasets (ESOL, Lipo, Malaria and CEP), with the lowest average RMSE value of 0.927. Moreover, our FG-BERT model showed an improvement of 7.2% in the overall average of regression tasks compared to other supervised learning models. Meanwhile, we counted how FG-BERT model compared to each baseline DL model on these commonly used public benchmark datasets. As shown in Table S5, our FG-BERT model consistently outperforms not only for each baseline individually, but also across all baselines (Table 1).

To further demonstrate the comprehensiveness and adaptability of the FG-BERT model, we added more advanced baseline models (Table S6) as a comparison against the regression datasets. As shown in Table S6, FG-BERT shows the best performance on FreeSolv and Lipo datasets and achieves the second-ranked performance on ESOL dataset. Additionally, two quantum chemistry datasets QM7 and QM8 were used to test the predictive performance of the FG-BERT model. FG-BERT performs the best in the QM8 dataset and achieves the second-ranked performance on the QM7 dataset (Table S6), implying the comprehensiveness and adaptability of the FG-BERT model. Taken together, these results indicate that FG-BERT outperforms SOAT DL models in predicting molecular properties. Therefore, our proposed FG-BERT model is highly competitive in predicting molecular properties in the field of drug discovery.

## Performance of the FG-BERT on the ADMET datasets

A total of 15 ADMET datasets (Table S2) related to drug discovery were used to further illustrate the superiority of FG-BERT in predicting molecular properties [6]. According to Wu *et al.* [6], we selected the following advanced models as baseline models for comparison, including two competing graph-based methods (HRGCN+ and Attentive FP) [9, 10], two fingerprint-based XGBoost models (XGBoost-MACCS and XGBoost-ECFP4) [46, 47] and one knowledge-based pretraining model K-BERT [6]. For a fair comparison, we used the same modeling data, splitting method and data-splitting ratio, reported from K-BERT. Furthermore, we averaged the results of the 10 different random seeds on the test set of each dataset as the final result of the model. Detailed performance results of FG-BERT on these ADMET datasets are shown in Figure 3A. As shown in Figure 3B, the FG-BERT model achieved the best molecular properties prediction performance compared with all baseline models, with the highest average AUC value of 0.813. The K-BERT pretraining model achieved the second-ranked performance, followed by HRGCN+, XGBoost-MACCS, XGBoost-ECFP4 and Attentive FP. Obviously, the two BERT-based pretrained models performed better than three non-pretrained models on these ADMET datasets, indicating that the pretrained models have certain advantages in the field of drug discovery. The reason may be that the BERT-based pretraining models can learn more accurate and useful molecular representations from a large number of unlabeled datasets. Meanwhile, FG-BERT outperforms K-BERT, mainly due to the fact that our model pays more attention to the important structural FGs related to the molecular properties, enabling the FG-BERT pretrained model to extract important chemical structure and semantic information from molecules. Collectively, the excellent predictive power of FG-BERT model

Downloaded from https://academic.oup.com/bib/article/24/6/bbad398/7337693 by guest on 01 August 2024

**Table 1:** Performance results (ROC-UAC) of FG-BERT on eight classification tasks from commonly used public datasets

| Methods | Tox21 | ToxCast | Sider | ClinTox | MUV | HIV | BBBP | Bace | Average |
|---|---|---|---|---|---|---|---|---|---|
| InfoGraph [57] | 73.3 ± 0.6 | 61.8 ± 0.4 | 58.7 ± 0.6 | 75.4 ± 4.3 | 74.4 ± 1.8 | 74.2 ± 0.9 | 68.7 ± 0.6 | 74.3 ± 2.6 | 70.10 |
| GPT-GNN [58] | 74.9 ± 0.3 | 62.5 ± 0.4 | 58.1 ± 0.3 | 58.3 ± 5.2 | 75.9 ± 2.3 | 65.2 ± 2.1 | 64.5 ± 1.4 | 77.9 ± 3.2 | 68.45 |
| EdgePred [59] | 76.0 ± 0.6 | 64.1 ± 0.6 | 60.4 ± 0.7 | 64.1 ± 3.7 | 75.1 ± 1.2 | 76.3 ± 1.0 | 67.3 ± 2.4 | 77.3 ± 3.5 | 70.08 |
| ContextPred [60] | 73.6 ± 0.3 | 62.6 ± 0.6 | 59.7 ± 1.8 | 74.0 ± 3.4 | 72.5 ± 1.5 | 75.6 ± 1.0 | 70.6 ± 1.5 | 78.8 ± 1.2 | 70.93 |
| GraphLoG [61] | 75.0 ± 0.6 | 63.4 ± 0.6 | 59.6 ± 1.9 | 75.7 ± 2.4 | 75.5 ± 1.6 | 76.1 ± 0.8 | 68.7 ± 1.6 | 78.6 ± 1.0 | 71.56 |
| G-Contextual [62] | 75.0 ± 0.6 | 62.8 ± 0.7 | 58.7 ± 1.0 | 60.6 ± 5.2 | 72.1 ± 0.7 | 76.3 ± 1.5 | 69.9 ± 2.1 | 79.3 ± 1.1 | 69.34 |
| G-Motif [62] | 73.6 ± 0.7 | 62.3 ± 0.6 | 61.0 ± 1.5 | 77.7 ± 2.7 | 73.0 ± 1.8 | 73.8 ± 1.2 | 66.9 ± 3.1 | 73.0 ± 3.3 | 70.16 |
| AD-GCL [63] | 74.9 ± 0.4 | 63.4 ± 0.7 | 61.5 ± 0.9 | 77.2 ± 2.7 | 76.3 ± 1.4 | 76.7 ± 1.2 | 70.7 ± 0.3 | 76.6 ± 1.5 | 72.16 |
| JOAO [64] | 74.8 ± 0.6 | 62.8 ± 0.7 | 60.4 ± 1.5 | 66.6 ± 3.1 | 76.6 ± 1.7 | 76.9 ± 0.7 | 66.4 ± 1.0 | 73.2 ± 1.6 | 69.71 |
| SimGRACE [65] | 74.4 ± 0.3 | 62.6 ± 0.7 | 60.2 ± 0.9 | 75.5 ± 2.0 | 75.4 ± 1.3 | 75.0 ± 0.6 | 71.2 ± 1.1 | 74.9 ± 2.0 | 71.15 |
| GraphCL [66] | 75.1 ± 0.7 | 63.0 ± 0.4 | 59.8 ± 1.3 | 77.5 ± 3.8 | 76.4 ± 0.4 | 75.1 ± 0.7 | 67.8 ± 2.4 | 74.6 ± 2.1 | 71.16 |
| GraphMAE [67] | 75.2 ± 0.9 | 63.6 ± 0.3 | 60.5 ± 1.2 | 76.5 ± 3.0 | 76.4 ± 2.0 | 76.8 ± 0.6 | 71.2 ± 1.0 | 78.2 ± 1.5 | 72.30 |
| 3D InfoMax [19] | 74.5 ± 0.7 | 63.5 ± 0.8 | 56.8 ± 2.1 | 62.7 ± 3.3 | 76.2 ± 1.4 | 76.1 ± 1.3 | 69.1 ± 1.2 | 78.6 ± 1.9 | 69.69 |
| GraphMVP [20] | 74.9 ± 0.8 | 63.1 ± 0.2 | 60.2 ± 1.1 | 79.1 ± 2.8 | 77.7 ± 0.6 | 76.0 ± 0.1 | 70.8 ± 0.5 | 79.3 ± 1.5 | 72.64 |
| MGSSL [68] | 75.2 ± 0.6 | 63.3 ± 0.5 | 61.6 ± 1.0 | 77.1 ± 4.5 | 77.6 ± 0.4 | 75.8 ± 0.4 | 68.8 ± 0.6 | 78.8 ± 0.9 | 72.28 |
| AttrMask [60] | 75.1 ± 0.9 | 63.3 ± 0.6 | 60.5 ± 0.9 | 73.5 ± 4.3 | 75.8 ± 1.0 | 75.3 ± 1.5 | 65.2 ± 1.4 | 77.8 ± 1.8 | 70.81 |
| Mole-BERT [21] | 76.8 ± 0.5 | 64.3 ± 0.2 | 62.8 ± 1.1 | 78.9 ± 3.0 | **78.6 ± 1.8** | **78.2 ± 0.8** | **71.9 ± 1.6** | 80.8 ± 1.4 | 74.04 |
| FG-BERT | **78.4 ± 0.8** | **66.3 ± 0.8** | **64.0 ± 0.7** | **83.2 ± 1.6** | 75.3 ± 2.4 | 77.4 ± 1.0 | 70.2 ± 0.9 | **84.5 ± 1.5** | **74.92** |

The data-split method from Mole-BERT was used to split each classification dataset into training, validation and test sets with an 8:1:1 ratio [21]. The FG-BERT used the same dataset and data split method for fairly comparison. The best result for each dataset is marked in bold. The standard deviation is located behind the mean. All comparison results are collected from Mole-BERT [21].

**Table 2:** Performance results (RMSE) of FG-BERT on four regression tasks from commonly used public datasets

| Methods | ESOL | Lipo | Malaria | CEP | Average |
|---|---|---|---|---|---|
| ContextPred [60] | 1.196 ± 0.037 | 0.702 ± 0.020 | 1.101 ± 0.015 | 1.243 ± 0.025 | 1.061 |
| JOAO [64] | 1.120 ± 0.019 | 0.708 ± 0.007 | 1.145 ± 0.010 | 1.293 ± 0.003 | 1.067 |
| GraphMVP [20] | 1.064 ± 0.045 | 0.691 ± 0.013 | 1.106 ± 0.013 | 1.228 ± 0.001 | 1.022 |
| AttrMask [60] | 1.112 ± 0.048 | 0.730 ± 0.004 | 1.119 ± 0.014 | 1.256 ± 0.000 | 1.054 |
| Mole-BERT [21] | 1.015 ± 0.030 | 0.676 ± 0.017 | 1.074 ± 0.009 | 1.232 ± 0.009 | 0.999 |
| FG-BERT | **0.944 ± 0.025** | **0.655 ± 0.009** | **1.057 ± 0.006** | **1.051 ± 0.029** | **0.927** |

The data-split method from Mole-BERT was used to split each regression dataset into training, validation and test sets with an 8:1:1 ratio [21]. The FG-BERT used the same dataset and data split method for fairly comparison. The best result for each dataset is marked in bold. The standard deviation is located behind the mean. All comparison results are collected from Mole-BERT [21].

makes it one of the most competitive methods to predict the ADMET properties of molecules.

## Performance of FG-BERT compared to the advanced graph-based and fingerprint-based models on cell-based phenotypic screening datasets

Phenotype-based screening, such as whole-cell viability, is an original but indispensable drug screening method that has received renewed attention in recent years [48–52]. Therefore, we evaluated the predictive performance of FG-BERT in phenotypic screening datasets for 13 breast cancer cell lines and 1 normal breast cell line (Table S3) [40]. For fair comparison with other models, we also used a consistent dataset, splitting method, as well as the same data splitting ratio (8:1:1 between training, validation and test sets), and then we calculated the average of AUC values obtained from 10 different random seeds as the final result of FG-BERT. Figure 3C and D shows the detailed performance results of FG-BERT on the 14 cell lines screening datasets. It can be found that FG-BERT performs better than these baseline models across the board. For example, FG-BERT performed best on 9 of the 14 cell lines (i.e. MDA-MB-453, SK-BR-3, T-47D, MCF-7, BT-474, BT-20, BT-549, MDA-MB-231 and HBL-100), while Attentive FP achieved the best performance on 2 cell lines (HS-578T and Bcap37), XGBoost performed best on 2 cell lines (MDA-MB-361 and MDA-MB-468) and GCN performed best on MDA-MB-435. Importantly, FG-BERT achieved the best overall performance on these 14 cell lines, with the highest average AUC value of 0.856 (Figure 3D). The statistics results show that the overall accuracy of the FG-BERT model is improved by 4.3% compared to the second-ranked XGBoost model (AUC = 0.813). These results fully demonstrate the excellent performance of FG-BERT on the cell-based phenotypic screening datasets, indicating that FG-BERT has great potential in cell-based phenotypic screening drug discovery.

## Ablation studies

### *Whether it is useful to mask the FGs*

To demonstrate the effectiveness of masking FGs for pretraining, we compared FG-BERT with MG-BERT [18], which was pretrained with random masked atoms. Since its original paper uses the random split method, we also use the random split method to split the same datasets to fairly evaluate the model performance. As shown in Figure 4A and B, the predictive performance of FG-BERT outperformed MG-BERT in all five datasets, with an overall relative improvement of 11.7%, of which the average performance of the classification and regression tasks were 1.9% and 18.2%, respectively. There is no doubt that BERT model is effective in pretraining by masking FGs.

### *Whether pretraining is indeed effective*

To demonstrate that pretraining can indeed improve the accuracy of molecular property prediction tasks, we tested the performance of pretrained and unpretrained FG-BERT under the

Downloaded from https://academic.oup.com/bib/article/24/6/bbad398/7337693 by guest on 01 August 2024

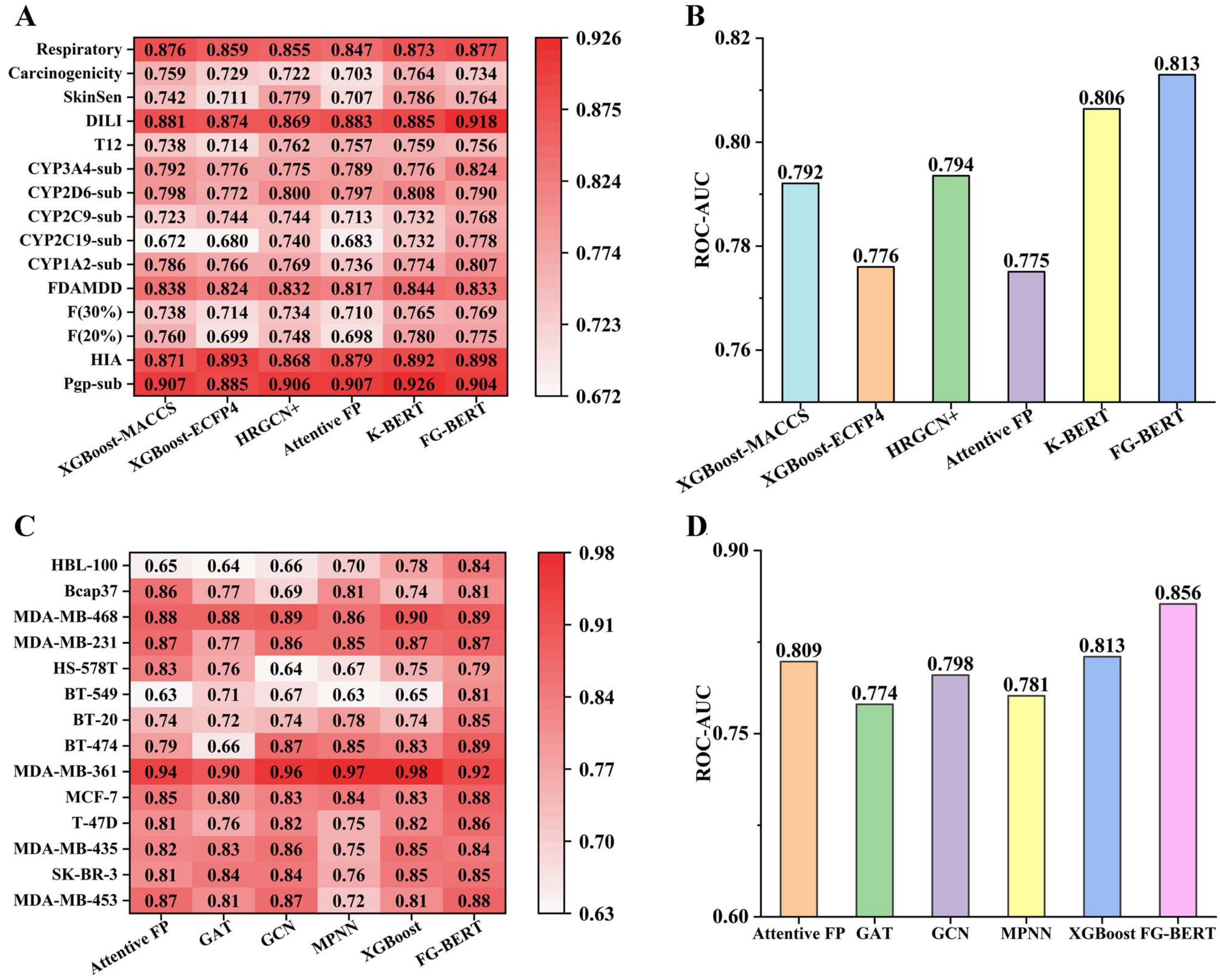


**Figure 3.** Performance of FG-BERT compared to the baseline models on 15 ADMET datasets. (**A**) Detailed AUC values of FG-BERT and baseline models for the test sets. (**B**) The average AUC values of FG-BERT and baseline models. The performances of all five baseline models were collected from Wu *et al.* [6]. Performance of FG-BERT compared to the baseline models on 14 breast cell-based phentypical sreening datasets. (**C**) Detailed AUC values of FG-BERT and baseline models for the test sets. (**D**) The average AUC values of FG-BERT and baseline models. The performances of all five baseline models were collected from He *et al.* [40].

same set of hyperparameters. Under the nonpretraining condition, the model parameters are fine-tuned using the initialized weights for downstream tasks. Detailed comparison results are listed in Figure 4C and D. It shows that under the same set of hyperparameters, the pretrained FG-BERT model outperformed the unpretrained FG-BERT model, with an average performance improvement of nearly 7.9% for the classification tasks and nearly 9.6% for the regression tasks. These results demonstrate that pretraining actually enables FG-BERT to capture rich structural and semantic information from large-scale unlabeled molecules, extract effective molecular representations and easily migrate them to specific downstream tasks through simple neural networks to enhance the predictive power of the model.

### *Whether it is useful to filter the pretrained dataset*

We further tested the performance of FG-BERT model constructed based on an unfiltered molecular dataset collected from ChEMBL to obtain weights for six classification tasks. Figure 5 shows that the predictive performance of the FG-BERT model built on the filtered dataset is better than that of the FG-BERT model based on unfiltered dataset. This may be due to the fact that the filtered corpus based on the three principles of drug-like properties contains higher quality of molecules, which makes it easier for the FG-BERT model to recognize the drug-like molecules, and thereby extracting better molecular representation, which ultimately makes the model more accurate in downstream prediction tasks.

### *The influence of different FG masking rate*

We also conducted ablation experiments on the FGs masking rates of the pretrained model. Four masking rates of 0.15, 0.2, 0.3 and 0.5 were set for pretraining, and public commonly datasets were used for fine-tuning. Notably, HIV dataset is not adopted because it is too large and would require a lot of computation time. We used random scaffold to split datasets for downstream tasks, and we computed the average values of evaluation metrics based on three random seeds as the final results. As shown in Figure 4E and F, FG-BERT model achieves the average optimal performance at the FGs masking rate of 0.15, which is different from the best results obtained by KPGT [53] at a masking rate of 0.5 and by ATMOL [54] at a masking rate of 0.25. A recent study in the field of CV has shown that the high masking rate of

Downloaded from https://academic.oup.com/bib/article/24/6/bbad398/7337693 by guest on 01 August 2024

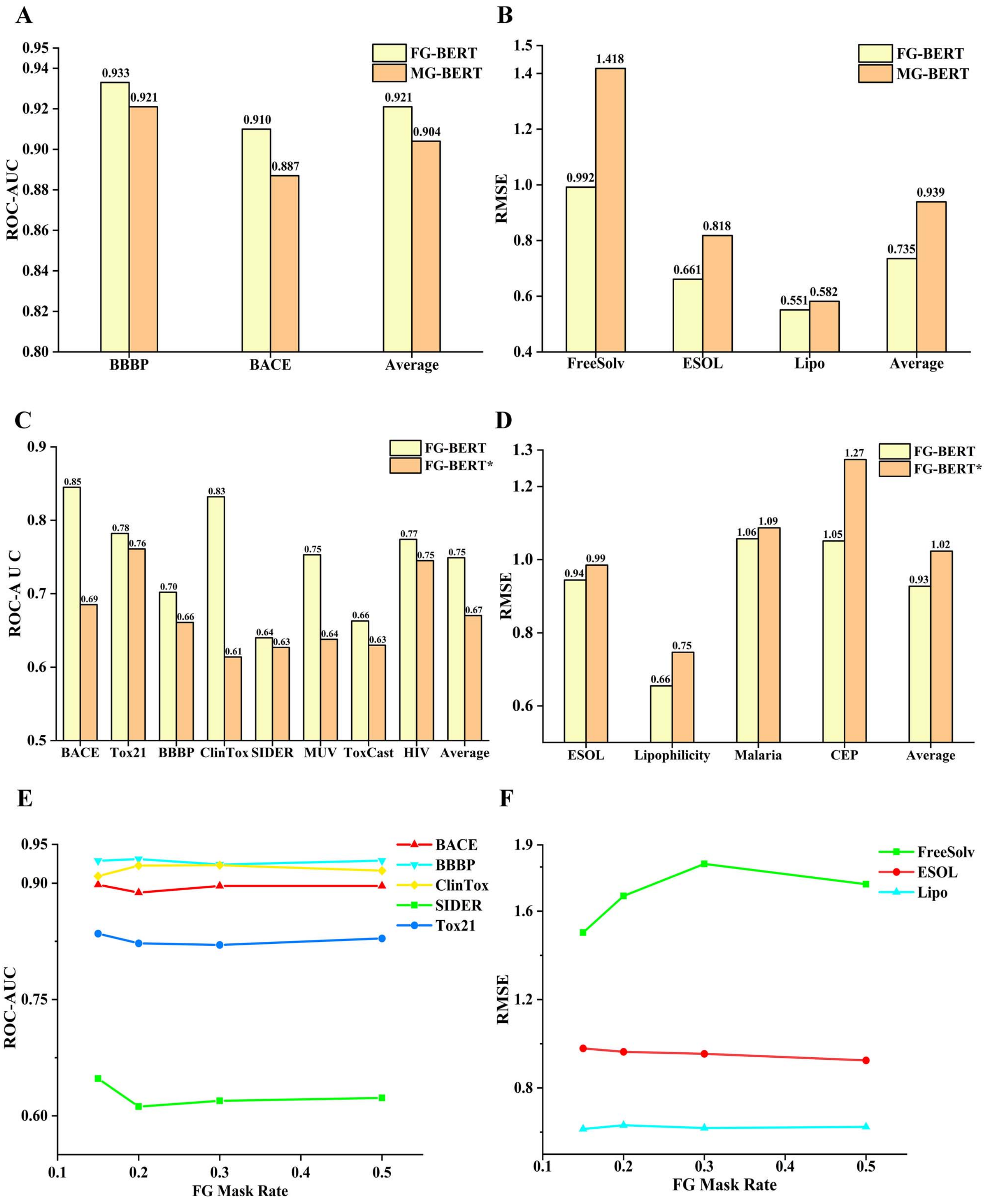


**Figure 4.** Ablation experiment of FG-BERT. Comparison performance of FG-BERT and MG-BERT on classification tasks (**A**) and regression tasks (**B**). Data are from MG-BERT [18]. Comparison performance of pretrained FG-BERT and without pretrained FG-BERT on classification tasks (**C**) and regression tasks (**D**). FG-BERT* indicates the without pretrained FG-BERT. Effects of FG masking rates on downstream classification tasks (**E**) and regression tasks (**F**).

the MLM has improved the model performance, which conflicts with our conclusion. The reason maybe that the high masking rate may make it difficult for the model to learn the semantic and structural information in chemical molecules, which in turn makes the quality of molecular representations extracted by the model insufficient and affects the performance of the model.

Downloaded from https://academic.oup.com/bib/article/24/6/bbad398/7337693 by guest on 01 August 2024

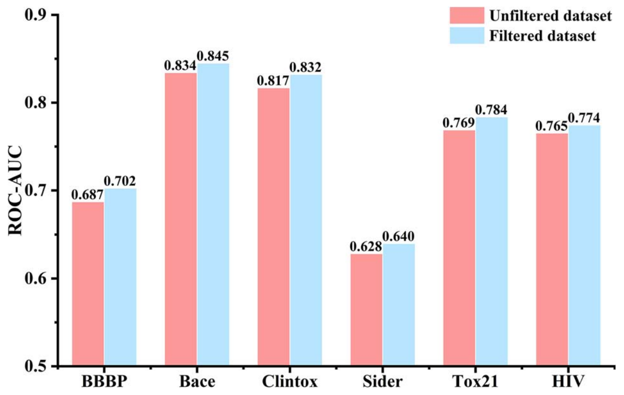


**Figure 5.** Comparison of FG-BERT model results with or without filtering pretrained corpus.

## The interpretation of FG-BERT

A comprehensive understanding of the relationship between the molecular structure and its properties is essential for analysis and for further lead compound optimization, which necessitates further exploration of the interpretability of the FG-BERT model. Currently, FG-BERT can reveal such relationship by aggregating information from all atomic and FG representations through the AM to generate a representation of the entire molecule, while attention weights can be generated and used to indicate the importance of atoms and/or FGs in the molecular representation, so it can be considered as a measure of the target property correlation measure.

For one thing, the FG-BERT model based on the blood–brain barrier permeability (BBBP) dataset was utilized to analyze the interpretability of the model. Since the blood–brain barrier (BBB) prevents the entry of most drugs and hormones, accurate prediction of the BBBP of molecules is essential for the development of drugs for central nervous system diseases treatment. Typically, hydrophobic molecules are more likely to cross the BBB due to their low polarity and high ClogP, while the converse is true for hydrophilic molecules. As shown in Figure 6, we easily visualized the attention weights of the molecules of interest, where darker red color indicates that they are given higher weights, and vice versa. Taking a permeable molecule as an example (Figure 6A), the benzene ring and cyclohexane (high hydrophobic FGs with the lowest polarity) in the molecule contribute the most to the BBB. We further quantified the ClogP values of these FGs by using ChemBioDraw (v. 14.0.0.117). It is easy to see that the left part of the molecule (Figure 6A) has a ClogP value of 0.547, while the right part has a ClogP value of 5.291, which is consistent with the prediction of the FG-BERT model. Meanwhile, for an impermeable molecule, FG-BERT tends to focus on the amino and hydroxyl groups on the left part of the molecule (Figure 6B), which provides most of the polarity to prevent the molecule from crossing the BBB. Furthermore, the ClogP value of the left FGs of the molecule in red is −0.575, indicating that the red portion of the molecule is more hydrophilic and may face difficulties in crossing the BBB. The high attention marked in the red part from our FG-BERT model was consistent with the inactive prediction results. Clearly, the high concern of the FGs marked in red from our FG-BERT model is consistent with the predicted results for the impermeable molecule.

$\beta$-secretase 1 (BACE-1) is an enzyme in human body involved in the cleavage of amyloid precursor proteins (APPs) [55]. The cleavage of APP will produce $\beta$-amyloid, one of the main components deposited in Alzheimer's disease (AD), so BACE-1 is considered an important target for the treatment of AD [34]. Herein, the optimal model based on the BACE dataset was chosen to further probe the interpretability of FG-BERT model. As shown in Figure 6C and D, two molecules (BACE_350 and BACE_1015) were selected from the test set for case studies. Notably, these two molecules have the same scaffold, but BACE_350 ($pIC_{50} = 8.22$) is active, while BACE_1015 (pIC50 = 6.35) is inactive [56]. Based on the visualization of the atomic attention weights of the two molecules (Figure 6C and D), it is clear that the model captures the important FGs in the molecules. We speculate that FG-BERT focuses on the difference between BACE_350 and BACE_1015, because BACE_350 contains a six-membered ring and an alkyne substituent, while the corresponding portion of BACE_1015 contains a five-membered ring without alkyne substituent. Furthermore, the binding modes of BACE_350 and BACE_1015 to BACE-1 were investigated using Glide SP docking, and the 2D protein-ligand interactions of the two molecules are shown in Figure 6C and D, respectively. Glide scores indicated that BACE_350 (docking score = −6.786 kcal/mol) exhibited better inhibitory activity against BACE-1 than BACE_1015 (docking score = −3.823 kcal/mol), which is consistent with the results of enzyme inhibition experimental. The highlighted amino group in the molecule can form two key hydrogen bonds with ASP32 and ASP228 according to the docking results, indicating that our model can automatically learn key FGs information from molecules and apply it to molecular property prediction tasks. In addition, Figure 6E presents the detailed 3D binding modes of the two molecules to BACE-1. Notably, the alkyne FG of BACE_350 is oriented toward the S3 hydrophobic pocket of BACE-1, resulting in a strong hydrophobic interactions, which is also in accordance with the highlighted FG of BACE_350. However, these hydrophobic interactions were not observed due to the lack of the alkyne FG in BACE_1015, which may account for the poor inhibitory activity of BACE_1015 against BACE-1. Such results shows that our FG-BERT model can identify key interaction patterns associated with biological activity.

## CONCLUSIONS

Given the scarcity of labeled data in the field of molecular property prediction, in this study we propose a new self-supervised learning framework called FG-BERT. The FG-BERT model enables efficient recovery pretraining by masking FGs in molecular graphs, and comprehensively mine chemical structure and semantic information from ∼1.45 million of unlabeled molecules to learn useful molecular representations. The FG-BERT pretrained model can be easily used for downstream molecular property prediction tasks though fine-tuning strategy. Based on three large-scale benchmark datasets involved in various molecular properties, including physical chemistry, biophysics and physiology, the systematic evaluation results indicate that the FG-BERT pretrained model is highly competitive, even compared with the most advanced traditional supervised ML and DL models, as well as self-supervised pretraining models. In addition, the high interpretability of the FG-BERT model allows users to fully understand the relationship between molecular structure and its properties, which can provide valuable fragments/FGs information to help scientists more accurately design and identify molecules with desired functions and/or therapeutic effects. All in all, we anticipate that FG-BERT as an out-of-the-box, effective and interpretable computational tool can be utilized for various drug discovery-related tasks. FG-BERT currently primarily focuses on FG information in molecules and does not fully consider

Downloaded from https://academic.oup.com/bib/article/24/6/bbad398/7337693 by guest on 01 August 2024

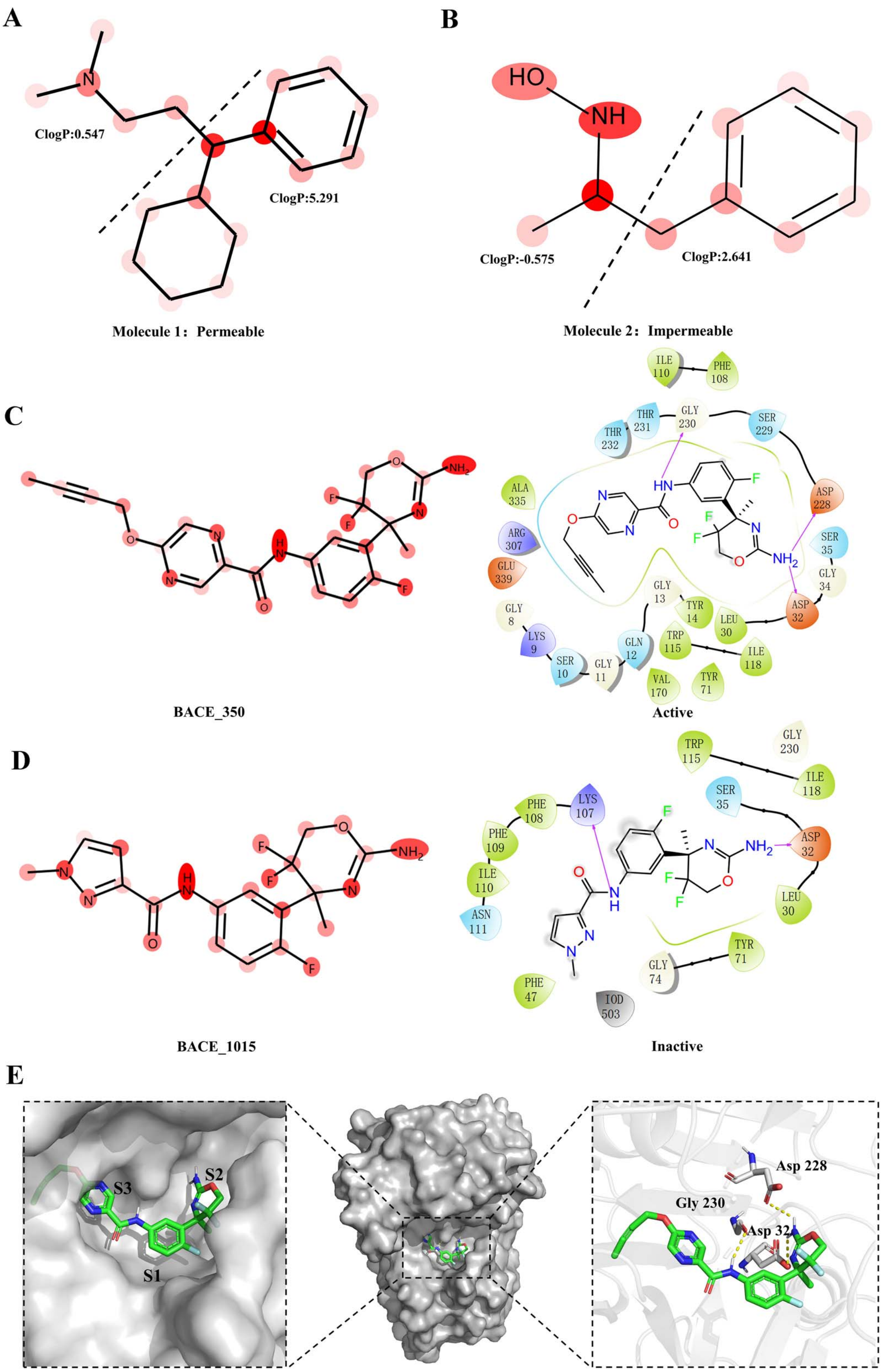


**Figure 6.** The importance of molecular structures during the prediction process. The darker the color, the more important are for the structures. Molecule 1 and molecule 2 were obtained from the BBBP (the BBB penetration) dataset. BACE_350 and BACE_1015 were obtained from the BACE dataset. (**A**) Molecule 1 is permeable, and the darker colored portion has a higher ClogP, which indicates a stronger lipophilicity. (**B**) Molecule 2 is impermeable, and the darker colored portion has lower ClogP, which means a weaker lipophilicity. The important portions that were captured by FG-BERT models were consistent with the prediction results. (**C**) and (**D**) Demonstration of the colored active and inactive molecules against BACE-1, and the 2D protein–ligand interaction binding modes generated by Glide SP docking. (**E**) Representation of the binding pocket (left), the solid surface of protein and ligand binding sites (middle) and the predicted 3D binding modes of BACE_350 and BACE_1015 to BACE-1 (right). All figures were generated using PyMOL software (https://pymol.org/2/).

Downloaded from https://academic.oup.com/bib/article/24/6/bbad398/7337693 by guest on 01 August 2024

molecular scaffold information. By integrating molecular scaffold information with FG information, it is possible to enhance the ability of model to extract molecular features, which in turn improves its molecular property prediction performance. We will explore this further in future work.

**Key Points**

- We presented a DL pretraining model named FG-BERT to predict molecular properties.
- It utilizes a self-supervised pretraining learning framework to learn useful molecular representation from ∼1.45 million drug-like molecules with diverse biological activities.
- Extensive experimental results have shown that FG-BERT is highly competitive with classical ML methods and SOTA pretraining and graph-based DL methods.
- The ablation experiments of FG-BERT demonstrate that the model can improve the downstream task prediction performance of the model after recovering pretraining by masking FGs.
- The intuitive interpretability of the FG-BERT model can provide important chemical fragments to assist chemists and pharmacists in designing or optimizing new molecules with desired properties.

## SUPPLEMENTARY DATA

Supplementary data are available online at http://bib.oxfordjournals.org/.

## FUNDING

This work was supported in part by the Natural Science Foundation of Guangdong Province (2023B1515020042), and the National Natural Science Foundation of China (81973241).

## DATA AVAILABILITY

The datasets used in this study and the source code for FG-BERT are publicly available at https://github.com/idrugLab/FG-BERT.

## REFERENCES

1. Song CM, Lim SJ, Tong JC. Recent advances in computer-aided drug design. *Brief Bioinform* 2009;**10**:579–91.
2. Eklund M, Norinder U, Boyer S, Carlsson L. Choosing feature selection and learning algorithms in QSAR. *J Chem Inf Model* 2014;**54**:837–43.
3. Phillips JC, Gibson WB, Yam J, *et al.* Survey of the QSAR and in vitro approaches for developing non-animal methods to supersede the in vivo LD50 test. *Food Chem Toxicol* 1990;**28**:375–94.
4. Dai H, Dai B, Song L. Discriminative embeddings of latent variable models for structured data. *Int Conf Mach Learn* 2016; 2702–11.
5. Li Y, Hsieh C-Y, Lu R, *et al.* An adaptive graph learning method for automated molecular interactions and properties predictions. *Nat Mach Intell* 2022;**4**:645–51.
6. Wu Z, Jiang D, Wang J, *et al.* Knowledge-based BERT: a method to extract molecular features like computational chemists. *Brief Bioinform* 2022;**23**:bbac131.
7. Wang S, Guo Y, Wang Y, *et al.* 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*, Niagara Falls, New York, September 7–10, 2019; 429–36.
8. Zeng X, Xiang H, Yu L, *et al.* Accurate prediction of molecular properties and drug targets using a self-supervised image representation learning framework. *Nat Mach Intell* 2022; 1–13.
9. Xiong Z, Wang D, Liu X, *et al.* Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *J Med Chem* 2019;**63**:8749–60.
10. Wu Z, Jiang D, Hsieh C-Y, *et al.* Hyperbolic relational graph convolution networks plus: a simple but highly efficient QSAR-modeling method. *Brief Bioinform* 2021;**22**:bbab112.
11. Cai H, Zhang H, Zhao D, *et al.* FP-GNN: a versatile deep learning architecture for enhanced molecular property prediction. *Brief Bioinform* 2022;**23**:bbac408.
12. Wu J, Xiao Y, Lin M, *et al.* DeepCancerMap: a versatile deep learning platform for target-and cell-based anticancer drug discovery. *Eur J Med Chem* 2023;**255**:115401.
13. Ai D, Wu J, Cai H, *et al.* A multi-task FP-GNN framework enables accurate prediction of selective PARP inhibitors. *Front Pharmacol* 2022;**13**:971369.
14. Zhu W, Zhang Y, Zhao D, *et al.* HiGNN: a hierarchical informative graph neural network for molecular property prediction equipped with feature-wise attention. *J Chem Inf Model* 2023;**63**: 43–55.
15. Jiang D, Wu Z, Hsieh C-Y, *et al.* Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. *J Chem* 2021;**13**:1–23.
16. Liu X, Zhang F, Hou Z, *et al.* Self-supervised learning: generative or contrastive. *IEEE Trans Knowl Data Eng* 2021;**35**:857–76.
17. Vaswani A, Shazeer N, Parmar N, *et al.* Attention is all you need[J]. *Advances in Neural Information Processing Systems* 2017;30.
18. Zhang X-C, Wu C-K, Yang Z-J, *et al.* MG-BERT: leveraging unsupervised atomic representation learning for molecular property prediction. *Brief Bioinform* 2021;**22**:bbab152.
19. Stärk H, Beaini D, Corso G, *et al.* 3d infomax improves gnns for molecular property prediction. *Int Conf Mach Learn* 2022; 20479–502.
20. Liu S, Wang H, Liu W, *et al.* Pre-training molecular graph representation with 3d geometry. ICLR, ArXiv Prepr. ArXiv211007728. 2021.
21. Xia J, Zhao C, Hu B, *et al.* Mole-BERT: rethinking pre-training graph neural networks for molecules. *Elev Int Conf Learn* 2023.
22. Ertl P, Altmann E, McKenna JM. The most common functional groups in bioactive molecules and how their popularity has evolved over time. *J Med Chem* 2020;**63**:8408–18.
23. Wadhwa K, Hennissen J, Shetty S, *et al.* Influence of substitution of various functional groups on inhibition efficiency of TEMPO analogues on styrene polymerization. *J Polym Res* 2017; **24**:1–8.
24. Assad H, Kumar A. Understanding functional group effect on corrosion inhibition efficiency of selected organic compounds. *J Mol Liq* 2021;**344**:117755.
25. Iqbal J, Vogt M, Bajorath J. Learning functional group chemistry from molecular images leads to accurate prediction of activity cliffs. *Artif Intell Life Sci* 2021;**1**:100022.
26. Gaulton A, Bellis LJ, Bento AP, *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* 2012;**40**:D1100–7.

Downloaded from https://academic.oup.com/bib/article/24/6/bbad398/7337693 by guest on 01 August 2024

27. Delaney JS. ESOL: estimating aqueous solubility directly from molecular structure. *J Chem Inf Comput Sci* 2004;**44**:1000–5.
28. Mobley DL, Guthrie JP. FreeSolv: a database of experimental and calculated hydration free energies, with input files. *J Comput Aided Mol Des* 2014;**28**:711–20.
29. Mendez D, Gaulton A, Bento AP, *et al.* ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res* 2019;**47**:D930–40.
30. Hachmann J, Olivares-Amaya R, Atahan-Evrenk S, *et al.* The Harvard clean energy project: large-scale computational screening and design of organic photovoltaics on the world community grid. *J Phys Chem Lett* 2011;**2**:2241–51.
31. AIDS antiviral screen data. In: NIH/NCI (ed). 2017.
32. Gamo F-J, Sanz LM, Vidal J, *et al.* Thousands of chemical starting points for antimalarial lead identification. *Nature* 2010;**465**: 305–10.
33. Rohrer SG, Baumann K. Maximum unbiased validation (MUV) data sets for virtual screening based on PubChem bioactivity data. *J Chem Inf Model* 2009;**49**:169–84.
34. Subramanian G, Ramsundar B, Pande V, *et al.* Computational modeling of $\beta$-secretase 1 (BACE-1) inhibitors using ligand based approaches. *J Chem Inf Model* 2016;**56**:1936–49.
35. Martins IF, Teixeira AL, Pinheiro L, *et al.* A Bayesian approach to in silico blood-brain barrier penetration modeling. *J Chem Inf Model* 2012;**52**:1686–97.
36. Tox21 data challenge. *NIH* 2017.
37. Kuhn M, Letunic I, Jensen LJ, *et al.* The SIDER database of drugs and side effects. *Nucleic Acids Res* 2016;**44**:D1075–9.
38. Wu Z, Ramsundar B, Feinberg EN, *et al.* MoleculeNet: a benchmark for molecular machine learning. *Chem Sci* 2018;**9**: 513–30.
39. Gayvert KM, Madhukar NS, Elemento O. A data-driven approach to predicting successes and failures of clinical trials. *Cell Chem Biol* 2016;**23**:1294–301.
40. He S, Zhao D, Ling Y, *et al.* Machine learning enables accurate and rapid prediction of active molecules against breast cancer cells. *Front Pharmacol* 2021;**12**:3766.
41. Devlin J, Chang M-W, Lee K, *et al.* BERT: pre-training of deep bidirectional transformers for language understanding. ArXiv Prepr. ArXiv1810.04805, 2018.
42. Kingma DP, Ba J. Adam: a method for stochastic optimization. In: *International Conference on Learning Representations*, San Diego, CA, USA, 2015. OpenReview.net. International Conference on Representation Learning, La Jolla, CA, USA.
43. Srivastava N, Hinton G, Krizhevsky A, *et al.* Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res* 2014;**15**:1929–58.
44. Bergstra J, Yamins D, Cox DD. Hyperopt: distributed asynchronous hyper-parameter optimization. *Astrophys Source Code Libr* 2022.
45. Ji Z, Shi R, Lu J, *et al.* ReLMole: molecular representation learning based on two-level graph similarities. *J Chem Inf Model* 2022;**62**: 5361–72.
46. Durant JL, Leland BA, Henry DR, *et al.* Reoptimization of MDL keys for use in drug discovery. *J Chem Inf Comput Sci* 2002;**42**: 1273–80.
47. Rogers D, Hahn M. Extended-connectivity fingerprints. *J Chem Inf Model* 2010;**50**:742–54.
48. Luo Y, Zeng R, Guo Q, *et al.* Identifying a novel anticancer agent with microtubule-stabilizing effects through computational cell-based bioactivity prediction models and bioassays. *Org Biomol Chem* 2019;**17**:1519–30.
49. Guo Q, Luo Y, Zhai S, *et al.* Discovery, biological evaluation, structure–activity relationships and mechanism of action of pyrazolo [3, 4-b] pyridin-6-one derivatives as a new class of anticancer agents. *Org Biomol Chem* 2019;**17**:6201–14.
50. Moffat JG, Vincent F, Lee JA, *et al.* Opportunities and challenges in phenotypic drug discovery: an industry perspective. *Nat Rev Drug Discov* 2017;**16**:531–43.
51. Malandraki-Miller S, Riley PR. Use of artificial intelligence to enhance phenotypic drug discovery. *Drug Discov Today* 2021;**26**: 887–901.
52. Berg EL. The future of phenotypic drug discovery. *Cell Chem Biol* 2021;**28**:424–30.
53. Li H, Zhao D, Zeng J. KPGT: knowledge-guided pre-training of graph transformer for molecular property prediction. In: *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Washington, D.C., USA. 2022; 857–67. ACM.
54. Liu H, Huang Y, Liu X, *et al.* Attention-wise masked graph contrastive learning for predicting molecular property. *Brief Bioinform* 2022;**23**:bbac303.
55. Hunt CE, Turner AJ. Cell biology, regulation and inhibition of $\beta$-secretase (BACE-1)[J]. *FEBS J* 2009;**276**(7):1845–59.
56. Malamas MS, Erdei J, Gunawan I, *et al.* Aminoimidazoles as potent and selective human $\beta$-secretase (BACE1) inhibitors. *J Med Chem* 2009;**52**:6314–23.
57. Sun F-Y, Hoffmann J, Verma V, *et al.* Infograph: unsupervised and semi-supervised graph-level representation learning via mutual information maximization. ArXiv Prepr ArXiv190801000. 2019.
58. Hu Z, Dong Y, Wang K, *et al.* GPT-GNN: generative pre-training of graph neural networks. In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Virtual Event, CA, USA. 2020; 1857–67. ACM.
59. Hamilton W, Ying Z, Leskovec J. Inductive representation learning on large graphs. *Adv Neural Inf Process Syst* 2017;**30**.
60. Hu W, Liu B, Gomes J, *et al.* Strategies for pre-training graph neural networks. ArXiv Prepr. ArXiv190512265 2019.
61. Xu M, Wang H, Ni B, *et al.* Self-supervised graph-level representation learning with local and global structure. *Int Conf Mach Learn* 2021;11548–58.
62. Rong Y, Bian Y, Xu T, *et al.* Self-supervised graph transformer on large-scale molecular data. *Adv Neural Inf Process Syst* 2020;**33**: 12559–71.
63. Suresh S, Li P, Hao C, *et al.* Adversarial graph augmentation to improve graph contrastive learning. *Adv Neural Inf Process Syst* 2021;**34**:15920–33.
64. You Y, Chen T, Shen Y, *et al.* Graph contrastive learning automated. *Int Conf Mach Learn* 2021;12121–32.
65. Xia J, Wu L, Chen J, *et al.* SimGRACE: a simple framework for graph contrastive learning without data augmentation. *Proc ACM Web Confs* 2022;**2022**:1070–9.
66. You Y, Chen T, Sui Y, *et al.* Graph contrastive learning with augmentations. *Adv Neural Inf Process Syst* 2020;**33**:5812–23.
67. Hou Z, Liu X, Cen Y, *et al.* Graphmae: self-supervised masked graph autoencoders. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2022; 594–604.
68. Zhang Z, Liu Q, Wang H, *et al.* Motif-based graph self-supervised learning for molecular property prediction. *Adv Neural Inf Process Syst* 2021;**34**:15870–82.

Downloaded from https://academic.oup.com/bib/article/24/6/bbad398/7337693 by guest on 01 August 2024