MG-BERT: leveraging unsupervised atomic representation learning for molecular property prediction

Article *in* Briefings in Bioinformatics · May 2021

DOI: 10.1093/bib/bbab152

8 authors, including:

Chengkun Wu
National University of Defense Technology
59 PUBLICATIONS **564** CITATIONS
SEE PROFILE

Jiacai Yi
National University of Defense Technology
5 PUBLICATIONS **160** CITATIONS
SEE PROFILE

Kim Hsieh
University of Toronto
91 PUBLICATIONS **986** CITATIONS
SEE PROFILE

Tingjun Hou
Zhejiang University
455 PUBLICATIONS **18,437** CITATIONS
SEE PROFILE

Briefings in Bioinformatics, 00(00), 2021, 1–14

doi: 10.1093/bib/bbab152
Problem Solving Protocol

MG-BERT: leveraging unsupervised atomic representation learning for molecular property prediction

Xiao-Chen Zhang†, Cheng-Kun Wu†, Zhi-Jiang Yang, Zhen-Xing Wu, Jia-Cai Yi, Chang-Yu Hsieh, Ting-Jun Hou and Dong-Sheng Cao

Corresponding author: Dong-Sheng Cao, Xiangya School of Pharmaceutical Sciences, Central South University, Changsha, 410003, PR China. Tel.: +86-731-89824761; E-mail: oriental-cds@163.com; Ting-Jun Hou, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou, 310058, Zhejiang, PR China. Tel.: +86-571-88208412; E-mail: tingjunhou@zju.edu.cn
†The first two authors contribute equally to this paper.

Abstract

Motivation: Accurate and efficient prediction of molecular properties is one of the fundamental issues in drug design and discovery pipelines. Traditional feature engineering-based approaches require extensive expertise in the feature design and selection process. With the development of artificial intelligence (AI) technologies, data-driven methods exhibit unparalleled advantages over the feature engineering-based methods in various domains. Nevertheless, when applied to molecular property prediction, AI models usually suffer from the scarcity of labeled data and show poor generalization ability. Results: In this study, we proposed molecular graph BERT (MG-BERT), which integrates the local message passing mechanism of graph neural networks (GNNs) into the powerful BERT model to facilitate learning from molecular graphs. Furthermore, an effective self-supervised learning strategy named masked atoms prediction was proposed to pretrain the MG-BERT model on a large amount of unlabeled data to mine context information in molecules. We found the MG-BERT model can generate context-sensitive atomic representations after pretraining and transfer the learned knowledge to the prediction of a variety of molecular properties. The experimental results show that the pretrained MG-BERT model with a little extra fine-tuning can consistently outperform the state-of-the-art methods on all 11 ADMET datasets. Moreover, the MG-BERT model leverages attention mechanisms to focus on atomic features essential to the target property, providing

Xiao-Chen Zhang is currently a PhD student in State Key Laboratory of High-Performance Computing, School of Computer Science, National University of Defense Technology, China. His researches focus on the development of cheminformatics tools.

Cheng-kun Wu is currently an associate professor in State Key Laboratory of High-Performance Computing, School of Computer Science, National University of Defense Technology. His researches focus on Systems Biology, High-Performance Computing, Pattern Recognition, Machine Learning and Data Mining.

Zhi-Jiang Yang was born in Hunan, China. He is currently a graduate student at Xiangya School of Pharmaceutical Sciences, Central South University. His researches focus on leveraging artificial intelligence for drug discovery.

Zhen-Xing Wu was born in Hunan, China. He is currently a PhD student in the College of Pharmaceutical Sciences, Zhengjiang University, under the supervision of Prof. Hou. His interests mainly lie in the area of computer-aided drug design.

Jia-Cai Yi was born in Guangdong, China. He is currently a graduate student in State Key Laboratory of High-Performance Computing, School of Computer Science, National University of Defense Technology. His researches focus on leveraging artificial intelligence for drug discovery.

Chang-Yu Hsieh is currently a senior researcher at Tencent Quantum Laboratory since 2018. He received his PhD degree in Physics from the University of Ottawa in 2012 and worked as a postdoctoral researcher at the University of Toronto (2012–2013) and Massachusetts Institute of Technology (2013–2016), respectively. Before joining Tencent, he worked as a senior researcher at Singapore-MIT Alliance for Science and Technology (2017–2018). His research interests span across quantum information science, nonequilibrium statistical physics, theoretical chemistry and machine learning for chemistry.

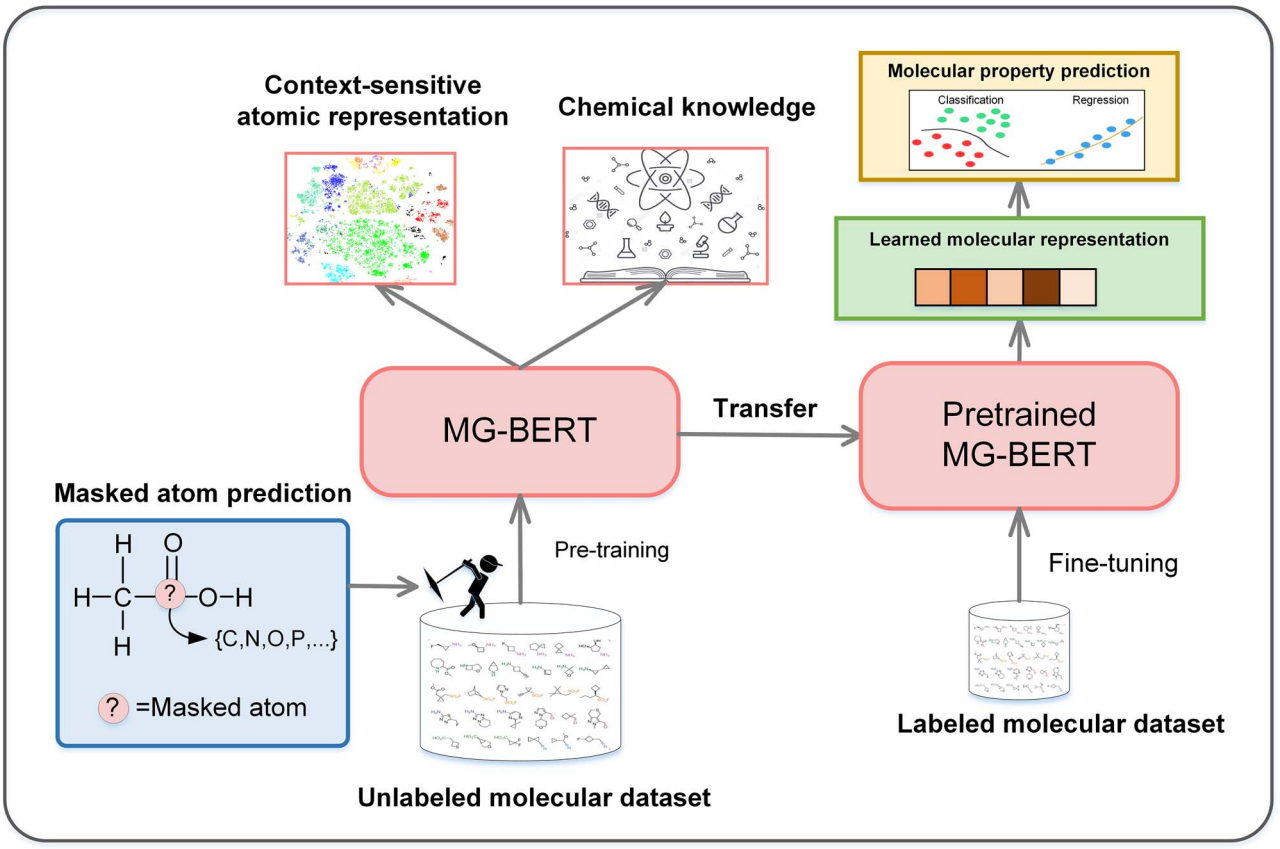
Ting-Jun Hou is currently a professor at the College of Pharmaceutical Sciences, Zhejiang University, China. His research interests include molecule simulation, drug design, machine learning, chemoinformatics and bioinformatics. Further information about Ting-Jun Hou can be found at the web site of his group: can be found at website of his group: http://cadd.zju.edu.cn.

Dong-Sheng Cao is currently a professor in the Xiangya School of Pharmaceutical Sciences, Central South University, China. His research interests include chemoinformatics, bioinformatics, drug design, chemo- and geoinformatics, web server and database, machine learning. Further information about Dong-Sheng Cao can be found at the website of his group: http://www.scbdd.com.

Submitted: 27 January 2021; **Received (in revised form):** 11 March 2021

excellent interpretability for the trained model. The MG-BERT model does not require any hand-crafted feature as input and is more reliable due to its excellent interpretability, providing a novel framework to develop state-of-the-art models for a wide range of drug discovery tasks.

Graphical Abstract



Key words: molecular property prediction; molecular graph BERT; atomic representation; deep learning; self-supervised learning

Introduction

Drug discovery is a risky, lengthy and resource-intensive process that usually takes around 10–15 years and billions of dollars [1]. To improve the efficiency of drug discovery, considerable efforts have been put into the development of computational tools and bioinformatics approaches [2, 3]. Among these methods, computational models for accurate prediction of molecular properties have a more significant and immediate impact on the drug discovery process since they can alleviate the excessive dependence on time-consuming and labor-intensive experiments and substantially reduce expenditure and time costs [4]. In this context, high-precision molecular property prediction models have become indispensable tools in many stages of the drug discovery process, covering hit identification, lead optimization, ADMET (Absorption, Distribution, Metabolism, Excretion and Toxicity) properties evaluation, etc. [5].

Expressive molecular representations are essential for molecular property prediction. Traditional methods heavily rely on feature engineering, in which experts handcraft a set of rules to encode relevant structural information or physicochemical properties of molecules into fixed-length vectors [6]. Molecular fingerprints and molecular descriptors are two canonical categories of molecular features. Molecular fingerprints focus on recording information about molecular substructures [7]. For instance, extended connectivity fingerprints (ECFPs) specifies an initial feature to each nonhydrogen atom and iteratively combines the features of neighboring atoms until a specific diameter is reached [8, 9]. Fingerprints are usually not optimized for particular prediction tasks due to sparse encoding problems. Alternatively, descriptors consist of a collection of physiochemical properties and structural information selected by experts based on their professional insight and feature engineering practice [10, 11]. Molecular descriptors can reduce irrelevant features and improve performance to some extent. However, the design process is trivial, time-consuming and error-prone. Hence, molecular fingerprints and descriptors both suffer from low scalability and versatility.

Recently, deep learning (DL) methods have made significant breakthroughs in many fields, such as computer vision [12, 13], natural language processing (NLP) [14, 15], playing go game [16], etc. The fundamental principle behind DL is: designing a suitable deep neural network (DNN) and training it on a large amount of raw data to learn representations automatically, rather than relying on human-crafted features.

The successful applications of DL in various domains inspired its application in molecular property prediction. Many studies of molecular property prediction have tried to apply DNNs directly to low-level molecular representations, such as the sequential SMILES (Simplified molecular-input line-entry specification) strings or molecular graphs [17–22]. The SMILES string describes compositions and chemical structures of molecules by a line of ASCII strings. As a kind of text, some suitable text processing algorithms, such as CNN, LSTM and Transformer [9, 17, 23], can be directly applied to build the prediction models. However, these algorithms need to learn to parse out useful features of molecules from the complex syntax of SMILES, which greatly increased the difficulty of learning and generalizing. Notably, unsupervised methods based models like auto-encoders have been applied to SMILES to learning useful representation from a large amount of unlabeled data [24–26]. This type of model generally comprises two neural networks: an encoder and a decoder. The encoder network converts the input sequence (a variable-length SMILES sequence with discrete values) into a fixed-size continuous vector (latent representation). The decoder network takes the latent representation as input and aims to convert it back to the input sequence. These models can be trained to embed the discrete molecules into a continuous vector space through training on a large amount of unlabeled data. The latent representation can be adopted for the downstream prediction tasks. However, the SMILES recovery-based molecular representation may not quite be optimal for general prediction tasks and cannot be further optimized. Emerging GNNs can directly learn from graph data, which could be a great advantage in molecular property prediction. Specifically, GNNs first represent molecules as two-dimensional (2D) graphs according to the connection relations [27–29] or three-dimensional (3D) graphs according to the atom distance matrix [18, 26, 30]. Then atoms are embedded into vectors according to their atomic characteristics, such as atom type, valence electron number, bond number, etc. After that, the vector for each atom is iteratively updated by aggregating information of surrounding atoms. Finally, a graph-level vector is generated from vectors of all atoms through a specific readout mechanism and sent to the fully connected neural network for prediction. If needed, the information from atomic bonds can also be incorporated [22, 31]. However, limited by overfitting and oversmoothing problems, current GNNs are usually too shallow (generally 2–3 layers), which weakens their ability to extract deep-level patterns [32].

The common challenges faced by the DL models in molecular property prediction are the scarcity of labeled data. It is well-known that DL models usually require a large amount of labeled data to achieve high effectiveness and generalization [33]. For example, in image classification tasks, people usually collect millions of images to train their DL models [34]. Unfortunately, it is unrealistic to obtain so much molecular properties data, especially ADMET endpoint data, which often requires a large number of time-consuming, laborious and costly experiments [35]. This dilemma makes DL models usually overfitting, greatly hurting their generalization ability.

The scarcity of labeled data has motivated the development of self-supervised or semisupervised learning methods [15, 36] in other fields. In the NLP domain, the recently proposed BERT model can utilize a large number of unlabeled texts for pretraining and dramatically improve the performance of various downstream tasks. The success of the BERT model can be attributed to the masked tokens prediction in which the model learns to predict the masked or contaminated words according to other visible words in the same sentence. In this process, the model is driven to mine the context information in this sentence. This kind of context information can benefit the downstream task and greatly improve their prediction performance. Inspired by the BERT model, the SMILES-BERT model is proposed to directly apply the BERT model to SMILES strings [37]. Although the SMILES-BERT model suffers from a lack of interpretability due to the existence of auxiliary characters in SMILES strings. Additionally, the complex syntax of SMILES strings also increases the difficulty for model learning.

To address these issues, we proposed a novel molecular graph BERT (MG-BERT) model by integrating the local message passing mechanism of GNNs into the powerful BERT model. The proposed MG-BERT model can overcome the oversmoothing problem faced by common GNNs and provide enough capacity to extract deep-level features for the generation of molecular representations. We further proposed the masked atoms prediction pretraining as an effective strategy to mine the context information in molecules automatically. Experimental results illustrate that MG-BERT can generate context-sensitive atomic representations after pretraining and greatly boost the performance of molecular property prediction tasks on 11 practical tasks, in which MG-BERT can consistently outperform previous state-of-the-art models. Additionally, MG-BERT can learn to focus on atoms and substructure related to the target properties by attention mechanism, which provides valuable clues to analyze and optimize molecules.

Materials and methods

Dataset collection

The training process of the proposed MG-BERT model consists of two stages: pretraining and fine-tuning. In the pretraining stage, we took advantage of a large number of unlabeled molecules to mine context information in molecules. Herein, 1.7 million compounds were randomly selected from the ChEMBL [38] database as the pretraining data. To verify the pretraining model, we randomly keep 10% for pretraining evaluation. The number in the training set ends up to 1.53 million. In the fine-tuning stage, the pretrained model was further trained for specific molecular property prediction. Sixteen datasets (eight for regression and eight for classification) covering critical ADMET endpoints and various common molecular properties were collected from the ADMETlab [35] and MoleculeNet [39] to train and evaluate MG-BERT. Detailed information of these 16 datasets is listed in Table 1. All molecules in these datasets are stored in SMILES strings format. The datasets were split into the training, validation and test datasets by a ratio of 8:1:1. It is worth noting that SMILES strings have a wide span of length ranging from several characters to over 100 characters. Therefore, a stratified sampling by SMILES length was used to make dataset splitting more uniform.

Model architecture

The original BERT model consists of three components: an embedding layer, several Transformer encoder layers [14],

Table 1. The detailed information of the 11 datasets used in this study

Type	Dataset	Category	Number	Positive	Negative
Regression	Caco2	Absorption	979	—	—
Regression	logD	Physicochemical property	10 354	—	—
Regression	logS	Physicochemical property	5045	—	—
Regression	PPB	Distribution	1480	—	—
Regression	tox	Toxicity	7295	—	—
Regression	ESOL	Physicochemical property	1128	—	—
Regression	Freesolv	Physicochemical property	642	—	—
Regression	Lipo	Physicochemical property	4200	—	—
Classification	Ames	Toxicity	6719	3631	3088-
Classification	BBB	Distribution	1855	1437	418
Classification	FDAMDD	Toxicity	795	437	358
Classification	H_HT	Toxicity	2170	1434	736
Classification	Pgp_inh	Absorption	2125	1240	885
Classification	Pgp_sub	Absorption	1210	616	594
Classification	BACE	Biophysics	1513	691	822
Classification	BBBP	Physiology	2039	1560	479

and a task-related output layer. In the embedding layer, the input word token is embedded into continuous vector space through an embedding matrix. As the Transformer model cannot automatically learn positional information, a predefined positional encoding vector needs to be added to each embedding vector in the embedding layer. In the Transformer encoder layer, every word token exchanges information with each other through a global attention mechanism. The embedding and Transformer layers are shared during the pretraining and fine-tuning stages. The last layer is generally a fully connected neural network, which further processes the Transformer layer's output and performs specific classification or regression tasks. The last layer for the pretraining and fine-tuning stages are not shared and are called the pretraining head and the prediction head, respectively. We provide an introduction of the MG-BERT model in the supplementary part. More details about the BERT model are described in the original literature of BERT [15].

Unlike the original BERT model for unstructured NLP, MG-BERT makes a few modifications according to the characteristics of molecular graphs. In the embedding layer, word tokens are replaced by atom type tokens. As atoms in a molecule are no related sequentially, there is no need to assign positional information. In natural language sentences, one word may be related to any other word, so global attention is needed. However, in a molecule, the atom is primarily associated with its neighboring atoms linked by bonds. To effectively realize this kind of inductive bias [40], we modified the global attention in BERT to local attention based on chemical bonds, which only allow atoms to exchange information through chemical bonds. This kind of local message passing mechanism makes MG-BERT a new variant of GNN. Notably, MG-BERT can overcome the oversmoothing problem due to the res-connection mechanism in BERT and has enough capacity to extracting deep-level patterns in molecular graphs. As depicted in Figure 1, we use adjacent matrix of molecules to control information exchange in molecules.

To obtain the graph-level representation and facilitate the subsequent prediction tasks in the fine-tuning stage, we added a supernode connecting to all atoms for each molecule. On the one hand, this supernode can exchange information with other nodes, which can well solve long-distance dependence to some extent. On the other hand, this supernode output can be regarded as the ultimate molecular representation and used to solve the downstream classification or regression tasks.

Pretraining strategy

BERT leveraged two learning tasks to pretrain the model, including the masked language model (MLM) task and the next sentence prediction (NSP) task. The MLM task is a fill-in-the-blank task, where a model uses context words surrounding a mask token to predict what the masked word should be. The NSP task is to determine if two sentences are consecutive. As molecules lack ongoing relationships like sentences, we only used the masked atom prediction task to pretrain our model.

Our proposed pretraining strategy is very similar to BERT. Firstly, 15% of the atoms in a molecule will be randomly selected, and at least one atom will be selected for the molecules with only a few atoms. For each selected atom, there is an 80% probability of being replaced with [MASK] tokens, a 10% probability of being randomly replaced with other atoms, and a 10% probability of keeping unchanged. The original molecule is used as the ground truth to train the model and the loss is only calculated at the masked atoms.

Input representations

To represent and manipulate atoms in molecular graphs, we need to add all atom types to our dictionary. However, the number of atom types appearing in the molecules is very limited. After statistical analysis, 13 frequently encountered atom types were included in the dictionary, and the other rarely encountered atom types are uniformly denoted by [UNK]. To get the graph-level representation, we added a supernode to each molecular graph. This supernode is denoted by [GLOBAL]. Besides, the [MASK] token is needed for representing the masked atoms in the pretraining stage. Thus, our dictionary includes the following tokens: [H], [C], [N], [O], [F], [S], [Cl], [P], [Br], [B], [I], [Si], [Se], [UNK], [MASK], [GLOBAL].

Model training and evaluation

Pretraining stage

Each molecule was converted into a 2D undirected graph in the pretraining stage according to the constituent atoms and their connection relationship by RDKit [41]. Then a supernode connecting all the nodes was added to every molecular graph. After that, certain atoms were randomly selected for masking

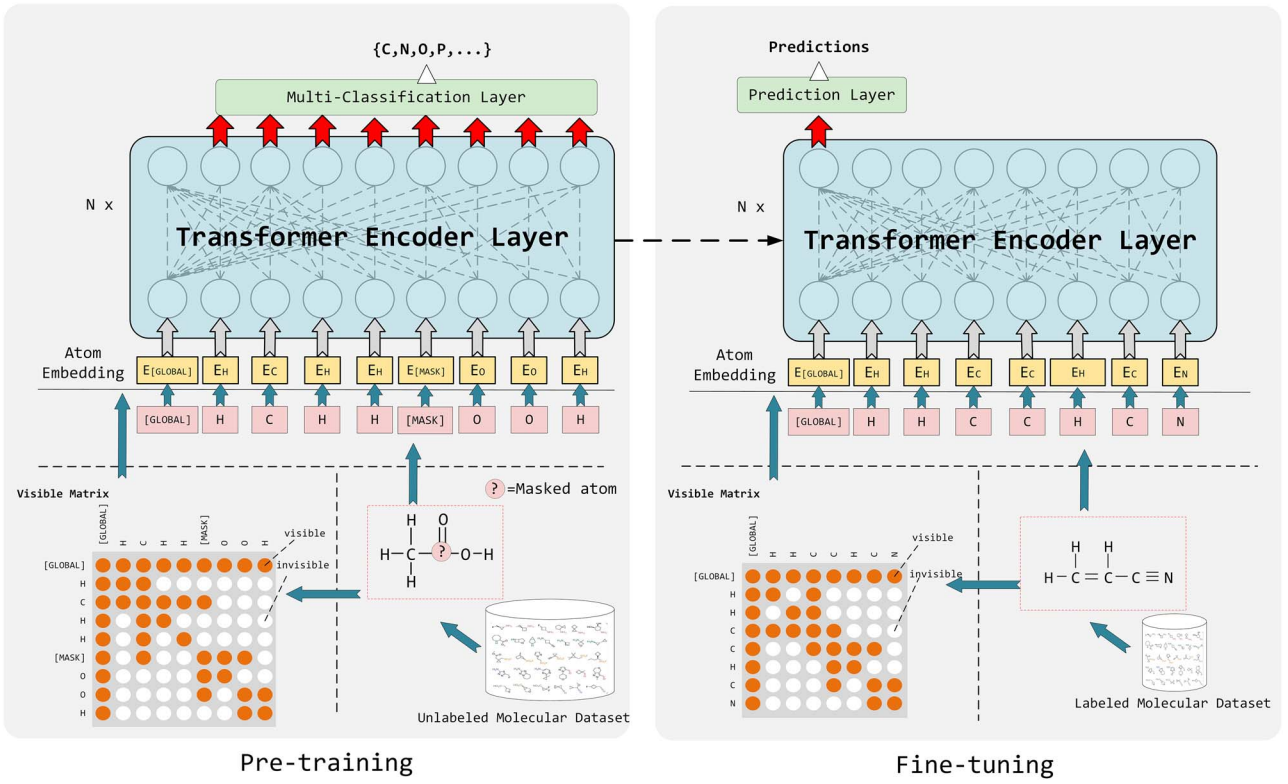


Figure 1. The overall pretraining and fine-tuning procedures of MG-BERT. The MG-BERT model uses bond-based local attention, in which every token can only exchange information with the tokens linked by chemical bonds. Apart from the output layers, the same architectures are used in both pretraining and fine-tuning. During the fine-tuning stage, the supernode denoted by [GLOBAL] was used to extract the global information to perform related prediction tasks. The visible matrix is used to modulate the attention matrix in the Transformer layer to control the information exchange.

according to the pretraining strategy. Finally, molecular graphs were sent to the MG-BERT model to predict the types of masked atoms. For some molecules with only a few atoms, we ensured as least one atom will be selected for masking. The model was trained via the standard batch gradient descent algorithm with an Adam optimizer [42]. The learning rate was set to 1e−4, and the batch size was set to 256. The model was pretrained for 10 epochs.

To evaluate the pretraining performance, the pretraining masking strategy was used to mask the molecules from the test set, and then the recovery rate was calculated as the evaluation metric.

Fine-tuning stage

After the pretraining, the pretraining head was removed. A two-layer task-related fully connected neural network was added to the output of the Transformer encoder layer corresponding to the supernode. The dropout strategy [43] was adopted to minimize overfitting. It should be noted that the dropout rate has a great impact on the final prediction and needs to be optimized according to specific tasks. According to our empirical results, the recommended range for the dropout rate is [0.0, 0.5]. Adam optimizer was used as the fine-tuning optimizer and a limited hyperparameter sweep was conducted for each task, with the batch sizes selected from {16,32,64} and the learning rates selected from {1e-5,5e-5,1e-4} [44].

The regression models were evaluated by the square determination coefficient (R2), and the classification models were evaluated by the area under the receiver operating characteristic (ROC-AUC). We used early stopping to avoid overfitting and set a maximum epoch to 100. To reduce random errors, every dataset was trained 10 times with random dataset splitting, and the calculated average and the standard deviation were reported as the final performance.

Results and discussion

Choice of MG-BERT model structure

To determine the better structure of the MG-BERT model for molecular property prediction tasks, we designed and compared three model structures. The specific parameters are listed in Table 2. The pretraining recovery accuracy and the averaged fine-tuning performance were used as the evaluation metrics. As listed in Table 2, the small MG-BERT model is inferior to the other two because of too few layers. Compared with the medium MG-BERT model, the large MG-BERT model performs better on the pretraining recovery task while performs slightly worse on molecular property prediction tasks. This phenomenon may be caused by the fact that the large MG-BERT model has an overfitting risk due to too many model parameters. The structure of the medium MG-BERT model was finally adopted since it can achieve the best performance on molecular property prediction.

Pretraining is indeed effective

To verify the effectiveness of pretraining, we compared the performance of the pretrained and non-pretrained MG-BERT models on molecular property prediction under the same hyperparameter settings. According to the comparison results listed

Table 2. Parameters and performance between three structures of MG-BERT

Name	Layers	Heads	Embedding size	FFN size	Recovery accuracy	Performance
$MG\text{-}BERT_{SMALL}$	3	2	128	256	0.9527	0.8082
$MG\text{-}BERT_{MEDIUM}$	6	4	256	512	0.9831	0.8283
$MG\text{-}BERT_{LARGE}$	12	8	576	1152	0.9835	0.8253

Table 3. Performances comparison (R2; ROC-AUC) of the MG-BERT models with different settings (the results are the number of percentages)

Type	Dataset	MG-BERT (without pretraining)	MG-BERT (without hydrogens)	MG-BERT
Regression	Caco2	64.79 ± 3.57	70.78 ± 2.31	74.68 ± 3.89
Regression	logD	84.78 ± 1.09	82.88 ± 0.40	87.46 ± 0.80
Regression	logS	83.84 ± 0.73	85.39 ± 1.36	87.66 ± 0.42
Regression	PPB	62.07 ± 4.91	59.47 ± 5.12	65.94 ± 2.86
Regression	tox	58.33 ± 1.58	60.63 ± 1.50	63.68 ± 1.53
Regression	ESOL	79.01 ± 4.21	80.16 ± 3.84	84.74 ± 4.09
Regression	FreeSolv	76.82 ± 3.58	78.19 ± 4.14	84.63 ± 4.68
Regression	Lipo	67.57 ± 3.19	70.82 ± 3.07	76.50 ± 2.64
Classification	Ames	86.49 ± 1.00	87.45 ± 0.95	89.33 ± 0.83
Classification	BBB	91.06 ± 1.07	94.76 ± 1.64	95.41 ± 1.13
Classification	FDAMDD	80.76 ± 3.73	87.01 ± 1.76	88.23 ± 3.43
Classification	H_HT	67.54 ± 2.98	69.62 ± 3.54	72.87 ± 3.49
Classification	Pgp_inh	88.42 ± 0.64	90.18 ± 0.42	92.44 ± 1.14
Classification	Pgp_sub	83.16 ± 2.31	89.34 ± 0.79	91.57 ± 2.26
Classification	BACE	84.35 ± 3.04	86.59 ± 2.17	88.68 ± 2.51
Classification	BBBP	86.42 ± 1.93	88.93 ± 2.04	92.08 ± 1.89

in Table 3, the pretrained MG-BERT model can outperform the non-pretrained MG-BERT model by more than 2% on all datasets, clearly demonstrating the effectiveness of the pretraining strategy and the excellent generalization ability of the pretrained model. For some small datasets such as Caco2 and FDAMDD, the prediction performance is improved by more than 7%, suggesting that the pretraining strategy can improve the prediction performance more effectively for small datasets. These results indicate that the MG-BERT model can indeed learn useful knowledge and transfer the learned knowledge to the downstream tasks by providing a nontrivial neural network initialization.

Influence of hydrogen atoms on pretraining accuracy and prediction tasks

Hydrogen atoms are usually ignored in most reported molecular property prediction models. In this study, a controlled experiment was conducted to explore whether hydrogen atoms are necessary for our MG-BERT model. The hydrogen-free model based on the molecular graphs without all hydrogen atoms was developed under the same hyperparameter setting for the MG-BERT model.

As illustrated in Figure 2, the pretraining accuracy of the MG-BERT model with hydrogens can reach 98.31%, whereas that of the hydrogen-free model can only reach 92.25%. The fine-tuning results listed in Table 3 show that the performance of the MG-BERT model with hydrogens is much better than that of the hydrogen-free model. Especially on some regression tasks, the model with hydrogens can outperform the hydrogen-free model by more than 4%.

The logic behind this is that, MG-BERT only utilizes the composition and connection information of molecules. Under this setting, hydrogen atoms can help determine the numbers of the chemical bonds for atoms of other types. In the masked atom recovery task, the numbers of bonds are critical to determining the types of the masked atoms. Therefore, the hydrogen-free MG-BERT model shows a significant decrease in masked atoms recovery rate. Furthermore, the absence of hydrogen atoms will also affect the context-information mining process in the pretraining stage, thus weakening the generalization ability of the pretrained model. In addition, if hydrogen atoms are removed, some molecules can become indistinguishable. As shown in Figure 3, benzene and cyclohexane can be converted into the same graph if hydrogen atoms are removed. However, if hydrogen atoms are kept, they will be converted into two different graphs. In this way, the absence of hydrogen atoms has a great impact on the performance of the fine-tuned model.

Comparison with other machine learning methods

Based on different molecular representations, we selected some state-of-the-art models as the baselines to comprehensively evaluate our proposed MG-BERT model. The first is the XGBoost [45] model based on ECFP4 fingerprints (ECFP4-XGBoost). This combination is a classic paradigm for molecular property prediction tasks. The 2nd and 3rd are two of the most representative and widely used GNNs: graph attention network (GAT) [27] and graph convolutional network (GCN) [28]. The 4th is based on the continuous-and-data-driven descriptor (CDDD) [19], and it consists of a fixed RNN based encoder that has been pretrained on a large number of unlabeled SMILES strings and a fully connected neural network. We also included the SMILES-BERT model, which directly utilized the original BERT model for SMILES strings.

The prediction results are shown in Table 4 and Figure 4. Except for ROC-AUC and R2, we also utilized accuracy and root mean-squared-error (RMSE) for evaluation, which are shown in Table S1. The performance of the ECFP4-XGBoost model shows

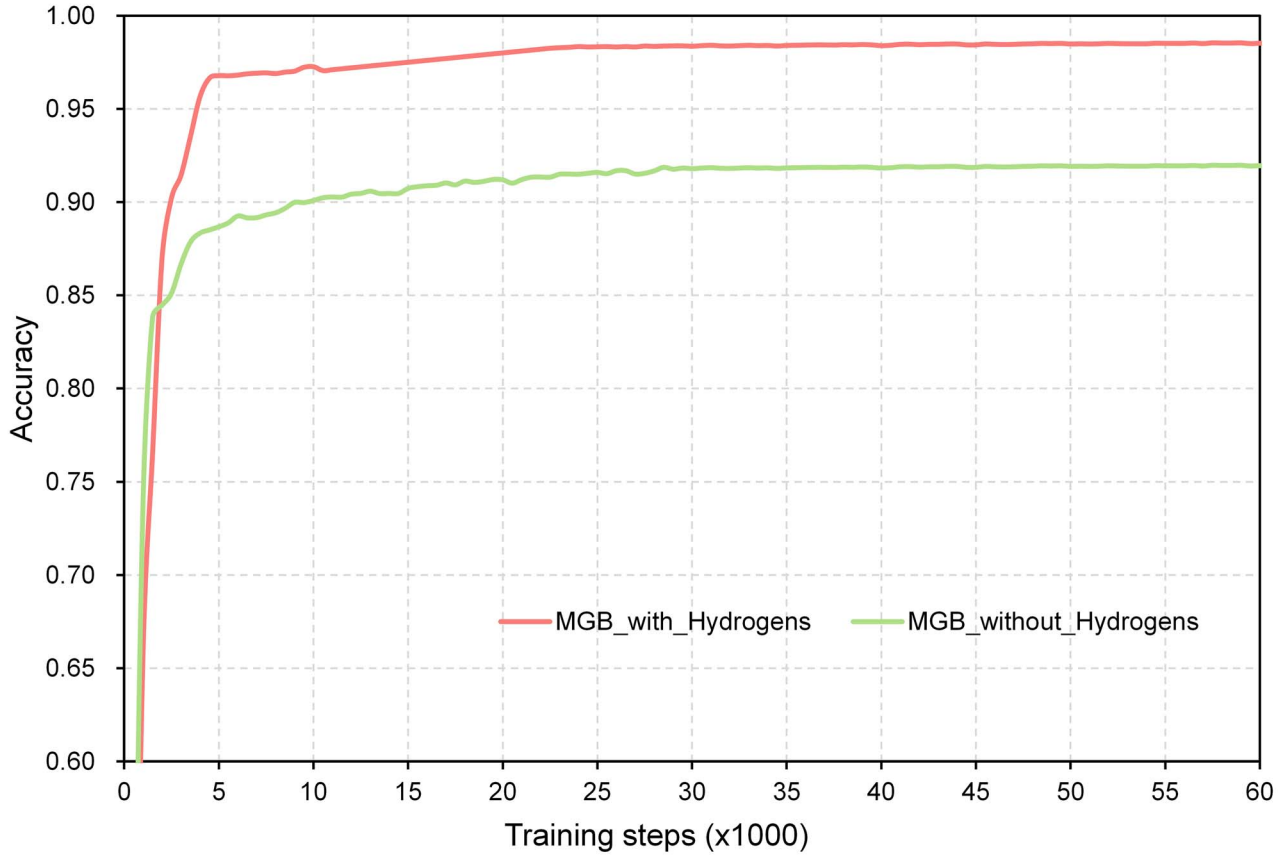


Figure 2. The pretraining accuracy of MG-BERT models versus the training steps.

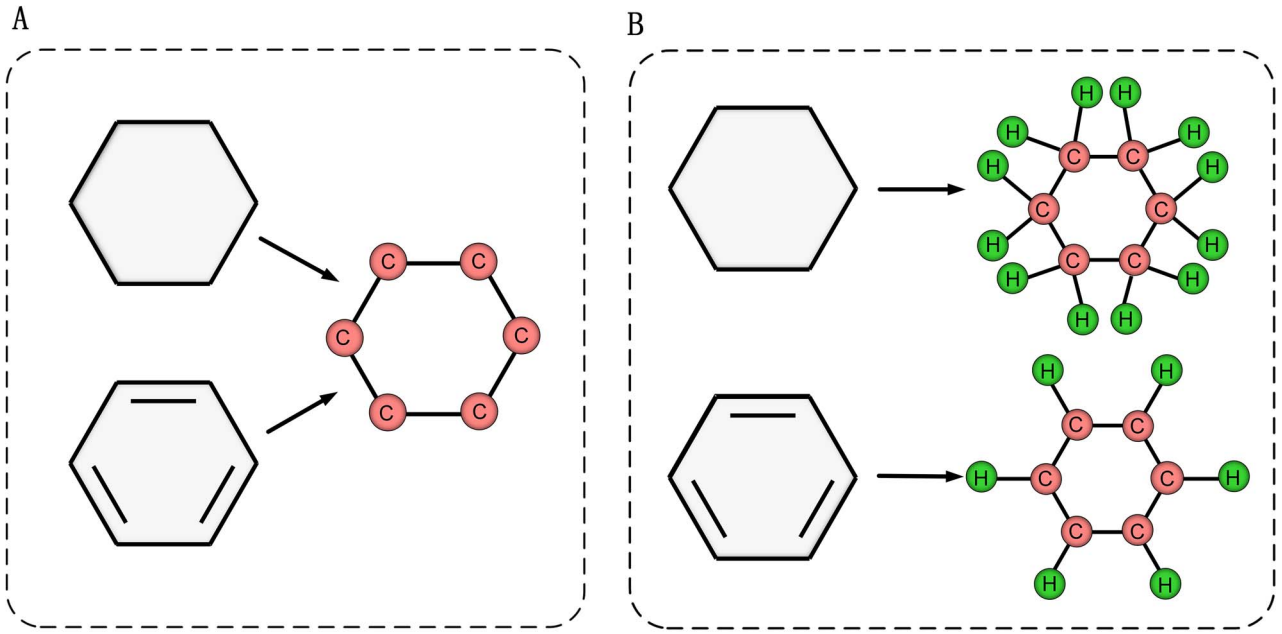


Figure 3. The influence of hydrogen atoms in converting molecules to graphs. (A) If hydrogens are removed, benzene and cyclohexane will be converted into the same graph. (B) If hydrogens are kept, benzene and cyclohexane will be converted into two different graphs.

a high variation on different datasets. This is quite possible that ECFP4 is a fixed-length molecular representation, leading to the information it represents may or may not work well for specified tasks. GNN models, including GAT and GCN, perform well when the labeled data are sufficient. However, when the labeled data are scarce, their performance becomes much worse, even worse

Table 4. The performance comparison (R2; ROC-AUC) of the proposed model and state-of-the-art models (the results are the number of percentages)

Type	Dataset	ECFP4-XGBoost	GAT	GCN	CDDD	SMILES-BERT	MG-BERT
Regression	Caco2	61.41 ± 4.61	69.16 ± 4.40	67.15 ± 3.67	73.42 ± 1.89	72.39 ± 2.85	74.68 ± 3.89
Regression	logD	70.84 ± 0.82	84.62 ± 2.29	86.22 ± 0.60	85.85 ± 2.03	86.31 ± 1.70	87.46 ± 0.80
Regression	logS	73.73 ± 2.32	84.06 ± 0.73	83.47 ± 1.18	84.01 ± 2.11	85.20 ± 1.31	87.66 ± 0.42
Regression	PPB	55.11 ± 4.35	59.96 ± 4.52	57.34 ± 4.74	54.12 ± 4.48	62.37 ± 3.89	65.94 ± 2.86
Regression	tox	59.02 ± 1.42	59.79 ± 1.79	56.71 ± 1.99	57.91 ± 4.19	60.49 ± 2.77	63.68 ± 1.53
Regression	ESOL	72.80 ± 2.93	76.63 ± 1.77	79.18 ± 2.88	80.21 ± 1.72	82.41 ± 3.74	84.74 ± 4.09
Regression	FreeSolv	73.70 ± 5.03	73.17 ± 4.86	77.81 ± 4.92	79.15 ± 3.12	81.47 ± 4.73	84.63 ± 4.68
Regression	Lipo	54.81 ± 1.64	67.56 ± 3.67	70.29 ± 3.16	69.24 ± 2.72	72.44 ± 2.57	76.50 ± 2.64
Classification	Ames	87.21 ± 0.91	86.38 ± 1.46	87.04 ± 1.55	86.82 ± 0.64	87.69 ± 0.98	89.33 ± 0.83
Classification	BBB	94.62 ± 0.82	93.03 ± 3.04	92.67 ± 1.44	94.44 ± 1.67	94.02 ± 1.79	95.41 ± 1.13
Classification	FDAMDD	88.14 ± 1.83	85.27 ± 1.39	85.12 ± 3.89	86.55 ± 2.54	87.94 ± 2.77	88.23 ± 3.43
Classification	H_HT	71.32 ± 2.41	67.48 ± 2.69	69.09 ± 2.54	70.81 ± 3.32	71.07 ± 3.81	72.87 ± 3.49
Classification	Pgp_inh	91.53 ± 1.52	90.88 ± 2.77	90.25 ± 1.99	91.54 ± 1.16	91.24 ± 1.42	92.44 ± 1.14
Classification	Pgp_sub	88.30 ± 2.60	90.73 ± 2.25	90.42 ± 2.79	89.88 ± 2.76	91.32 ± 2.35	91.57 ± 2.26
Classification	BACE	87.14 ± 2.14	85.58 ± 3.38	86.98 ± 1.39	86.58 ± 3.84	87.64 ± 2.94	88.68 ± 2.51
Classification	BBBP	89.16 ± 1.17	90.33 ± 3.02	90.74 ± 1.05	91.12 ± 1.52	91.32 ± 1.83	92.08 ± 1.89

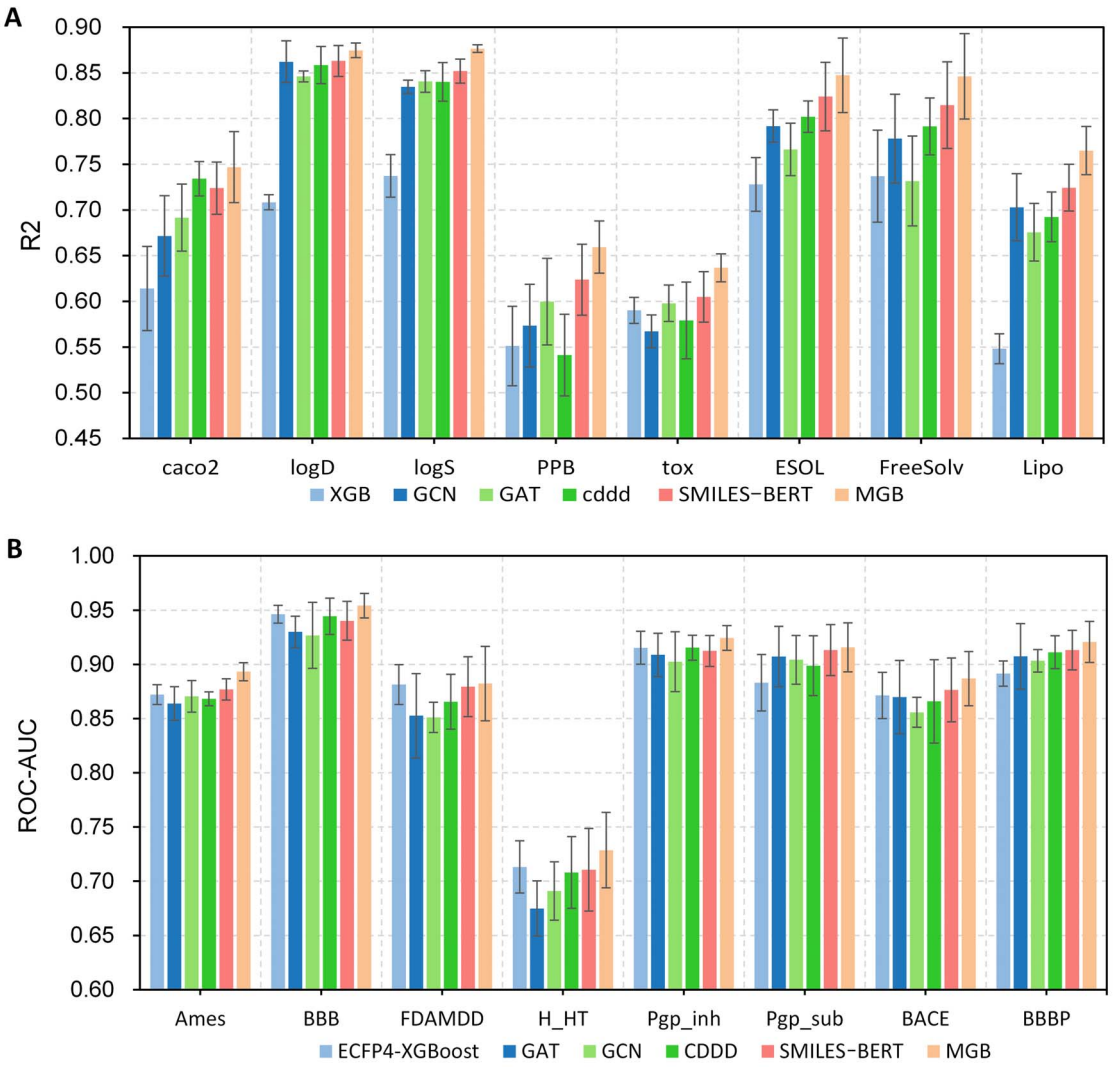


Figure 4. The performance comparison of MG-BERT and state-of-the-art models for (A) the regression tasks and (B) the classification tasks.

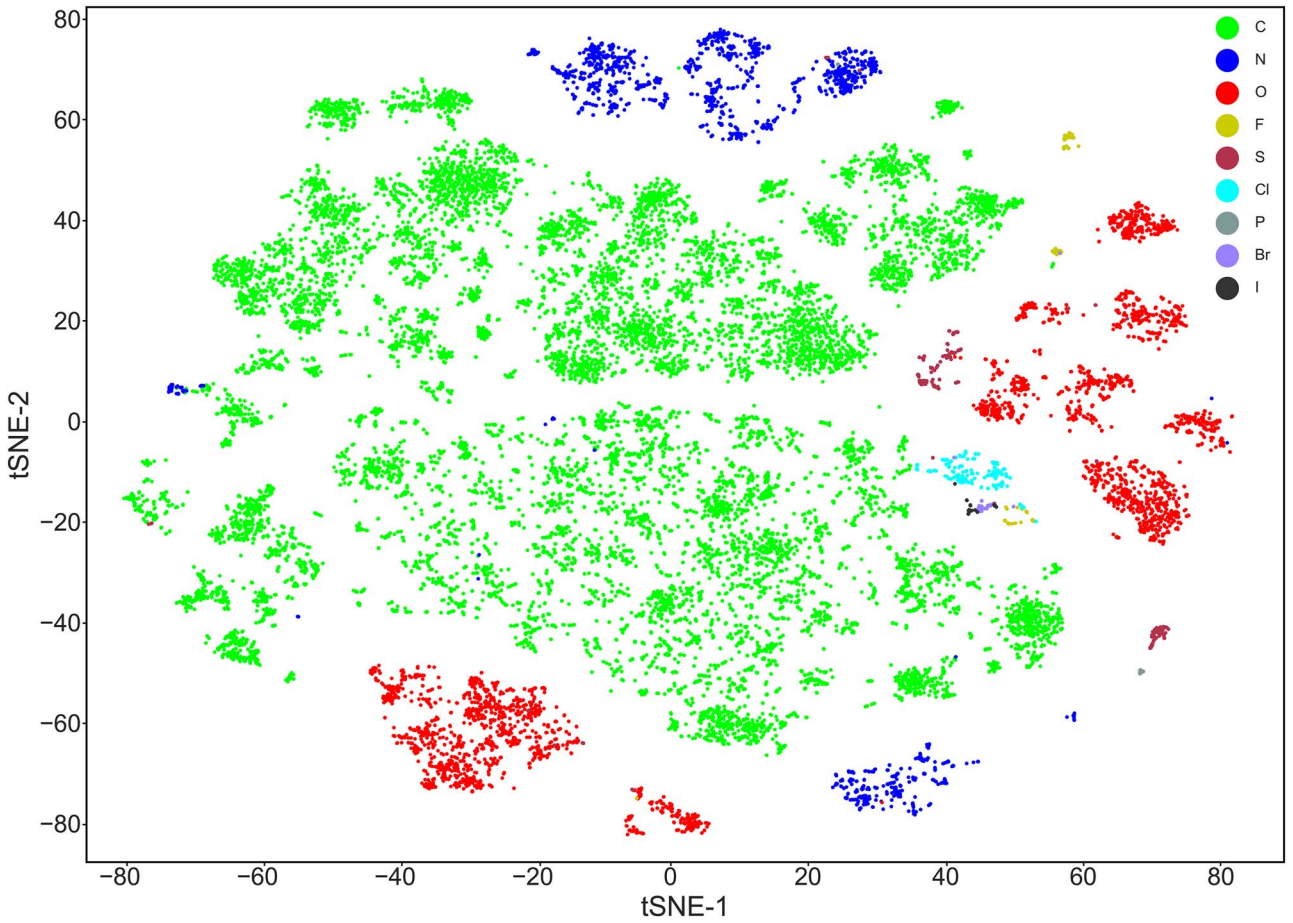


Figure 5. The t-SNE plot of the atomic embedding vectors colored by atomic types. The hydrogen atoms are ignored before t-SNE for better visualization.

Table 5. Symbols and defined categories of atoms

Symbols	Categories
C–N	Carbon atoms with four single bonds, in which one bond is connected to a nitrogen atom.
C–O	Carbon atoms with four single bonds, in which no bond is connected to the nitrogen atom and one bond is connected to the oxygen atom.
C–C	Carbon atoms with four single bonds, in which no bond is connected to nitrogen atom or oxygen atom.
C=C	Carbon atoms with a double bond that is connected to another carbon atom.
C=O	Carbon atoms with a double bond that connected to an oxygen atom.
Aromatic C	Carbon atoms in a benzene ring.
O=	Oxygen atoms with a double bond.
–OH.	Oxygen atoms with a single bond connected to hydrogen.
–O–	Oxygen atoms with two single bonds, in which no bond is connected to hydrogen atoms.
N=	Nitrogen atoms with a double bond.
$–NH_2$	Nitrogen atoms have three single bonds, in which two bonds are connected to hydrogen atoms.
–NH.	Nitrogen atoms have three single bonds, in which one bond is connected to hydrogen atoms.
N	Nitrogen atoms have three single bonds, in which no bond is connected to hydrogen atoms.
Others	Atoms not defined above

than the model based on the molecular fingerprints. The CDDD model shows certain competitiveness. However, the molecular representations of the CDDD model were obtained through the SMILES encoding and decoding tasks, which cannot be further optimized for specific tasks. In contrast, The SMILES-BERT model and our MG-BERT model can also learn rich context-sensitive information in the pretraining stage and can be further optimized for specific tasks. SMILES-BERT models are slightly behind our MG-BERT model. This may be caused by the fact that SMILES strings are much more complex to learn from than molecular graphs, which means the SMILES-BERT models have to parse out the molecular information hidden in the complex syntax of SMILES strings. Although the MG-BERT model can directly learn from the molecular graphs, which are naturally representation of molecules. The proposed MG-BERT model can consistently outperform the other methods. The overall improvement is 28.1%

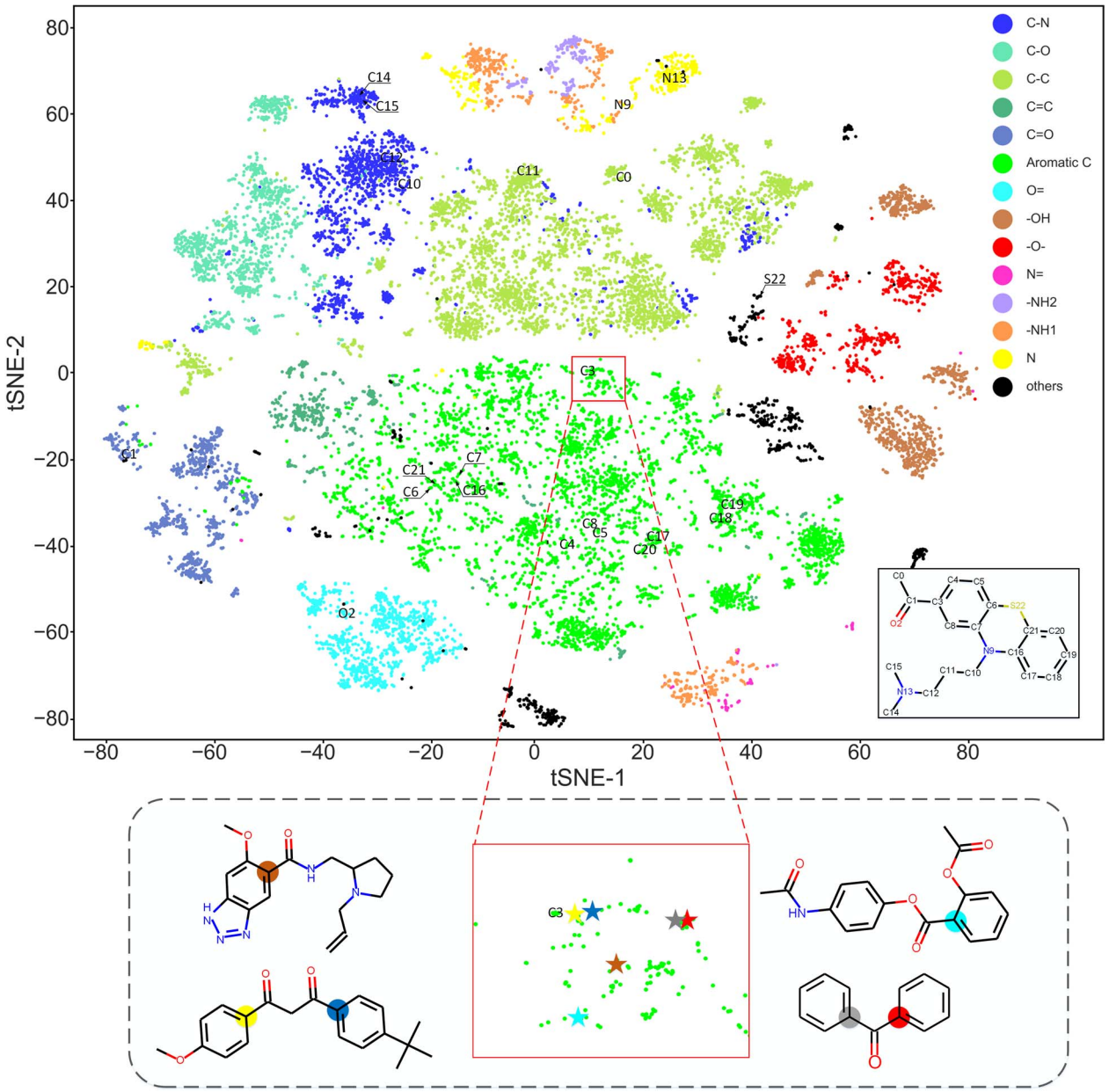


Figure 6. The t-SNE plot of the atomic embedding vectors colored by atom categories. In the main graph, the positions of the atoms belonging to the displayed molecule are marked. In the enlarged graph, the atoms and their corresponding positions are marked by the same color. In the enlargement, all colored atoms are in a benzene ring and linked to a carbonyl group.

(7.02% on classification tasks and 21.28% on regression tasks). Notably, for the PPB dataset, the improvement of the MG-BERT model is more than 6%. The improvements of our model relative to baselines are statistically significant (95% confidence interval, CI) according to the paired t-test ($P \leq 0.001$). These results convincingly highlight MG-BERT's potential to become a good choice for molecular property prediction tasks in drug design.

Analysis of the atomic representations from the pretrained MG-BERT model by t-SNE

To analyze what the MG-BERT model learned in the pretraining stage, we visualized the atomic representations generated by the pretrained model and tried to find some interesting patterns. Specifically, 1000 molecules (including approximately 22 000 atoms) were randomly selected from the fine-tuning dataset and fed into the pretrained model without masking, and the output of the Transformer encoder layer was gathered. In this way, a 256-dimensional vector was obtained for each atom, and about 22 000 vectors were obtained in total. The classical dimensionality reduction method t-SNE [46, 47] was used to visualize these high-dimensional vectors. As shown in Figure 5, atoms of different types can be easily distinguished, demonstrating that the generated representations contain information of atomic types.

Further observation shows that the atoms of the same type can be divided into several different groups. It seems

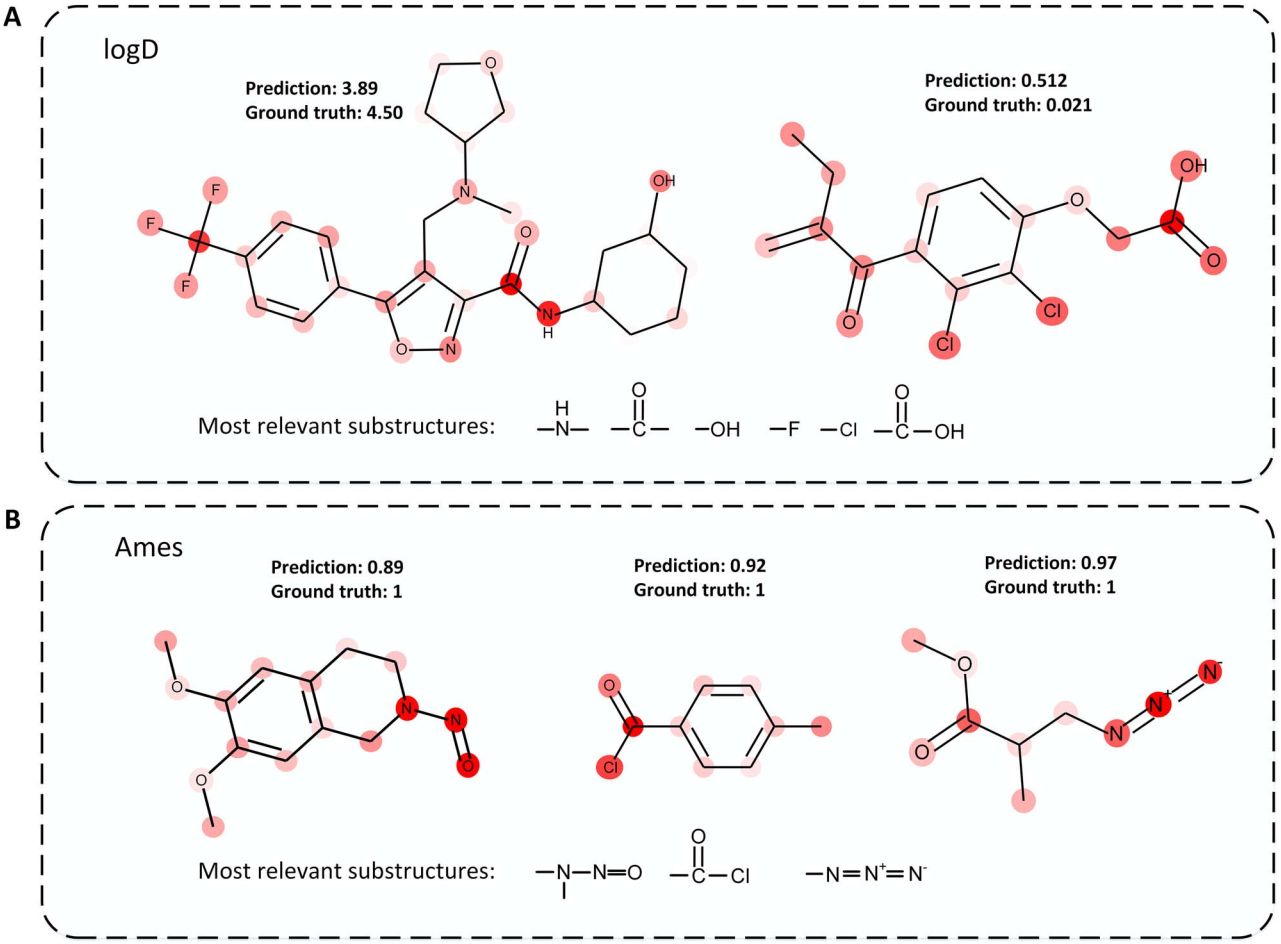


Figure 7. The attention weight visualization of the supernodes for some molecules in the (A) logD prediction task and (B) Ames prediction task. A darker the color denotes a bigger attention weight. The attention weights for hydrogens are transferred to their neighbors for convenience.

that the generated atomic representations contain richer information than atomic types. This finding inspires us to define some atomic categories according to atoms' surrounding environments and visualize the atoms by their categories. The categories were mainly defined by the possible varieties of the 1st-order neighborhood (Table 5). As shown in Figure 6, atoms are clustered together according to their categories. These results indicate that the generated atomic representation for each atom contains information on its 1st-order neighborhood. Notably, the carbon atoms in the benzene ring are distinguishable from those in other environments, indicating that the pretrained model can capture high-order neighborhood information. To make further analysis, we randomly selected a complex molecule and marked the positions of its atoms in Figure 6. It can be found that atoms in the benzene ring tend to be close to each other if their branching environments are similar. We further made a local enlargement to figure out whether the environments of atoms in a small region are consistent. In the enlarged graph, we randomly marked some atoms and displayed their corresponding molecules. It can be found that all these atoms are in a benzene ring and linked to a carbonyl group. These results can prove that our pretrained model can surely capture high-order neighborhood information to some extent.

The above results fully prove that the generated atomic representations can successfully capture the 1st-order or even high-order neighborhood information. In this way, the learned atomic representations can be regarded as the representations of molecular substructures, which can be very beneficial for downstream tasks.

Analysis of MG-BERT's attention

Understanding the relationship between molecular structures and molecular properties is quite beneficial for analyzing and optimizing molecules, and MG-BERT provides a natural way to reveal these relationships. MG-BERT leverages attention mechanisms to aggregate information from all atomic representations to form molecular representations. In this way, the attention weights represent the contribution of each atomic representation in the final molecular representation and can be regarded as a relevant measurement to the target property. To note, each atomic representation also aggregates information from its neighbors, so the attention weights are not only just for individual atoms but also shared by its neighbors to some extent.

To find out whether our model can reasonably allocate the attention weights, we randomly selected some molecules and visualized their attention weights according to specific tasks.

The results for several molecules in the logD and Ames prediction tasks are shown in Figure 7. The property of logD is related to molecular lipophilicity. According to Figure 7A, we can conclude that more attention is distributed to polar groups, which play an important role in determining molecular lipophilicity. The Ames task is to determine whether a molecule belongs to mutagens or not. Figure 7B shows that the attention predominately is distributed to the acylchloride, nitrosamide and azide groups, which have been demonstrated to be mutagenic structural alerts [48]. These results demonstrate that MG-BERT can reasonably allocate attention weight according to specific tasks, which is of great significance for medicinal chemists to explore the relationship between substructures and molecular properties.

Conclusion

In this study, we presented a novel semisupervised learning approach called MG-BERT to alleviate the data scarcity problem in molecular property prediction. The proposed MG-BERT model modifies the original BERT model according to the characteristic of molecular graphs. MG-BERT takes advantage of large amounts of unlabeled molecular data through the masked atom recovery task to mine the context information in molecular graphs for effective atomic and molecular representation learning. After the pretraining, MG-BERT could easily be fine-tuned on small labeled datasets and achieved very competitive prediction performance. In experiments, the MG-BERT model consistently outperformed the-state-of-art models on 11 representative ADMET tasks, fully demonstrating the effectiveness of our proposed method. Furthermore, we visualized the atom representations from the pretrained model and found that the generated atomic representations can fully capture the 1st-order neighborhood information and even high-order neighborhood information to some extent. Through this, we effectively explained why pretraining is beneficial for downstream tasks. Although DL models have strong learning and predictive capabilities, their interpretability is generally so poor that they are called black-box models. MG-BERT provides a natural way to measure the relevance between atoms or substructures with the target property by attention mechanisms. These features have established MG-BERT as an effective and interpretable computational tool in solving challenges of molecular property prediction and molecular optimization.

Abbreviation

BERT, bidirectional encoder representations from Transformers; AI, artificial intelligence; ASCII, American Standard Code for Information Interchange; GNN, graph neural network; ADMET, absorption, distribution, metabolism, excretion and toxicity; ECFP, extended connectivity fingerprints; DL, deep learning; DNN, deep neural network; SMILES, simplified molecular input line entry specification; CNN, convolutional neural network; LSTM, long short-term memory.

Key Points

- MG-BERT integrates the local message passing mechanism of GNNs into the powerful BERT. As a new variant of GNNs, MG-BERT can overcome the oversmoothing problem and has enough capacity to extracting deep-level patterns in molecular graphs.
- MG-BERT can take advantage of a large number of unlabeled molecules through the masked atom recovery task to mine the context information in molecular graphs and transfer the learned knowledge to benefit molecular property prediction.
- MG-BERT can outperform the state-of-the-art models on molecular property prediction without any handcrafted features and provides interpretability by reasonably allocating attention weights to atoms or substructures according to the relevance with the target property.

Availability

All datasets and codes used in this study are available at GitHub: https://github.com/zhang-xuan1314/Molecular-graph-BERT.

Supplementary data

Supplementary data are available online at *Briefings in Bioinformatics*.

Authors' contributions

XCZ and DSC developed the algorithms; XCZ wrote the codes and drafted the manuscript; ZJY and JCY helped prepared the datasets and figures. DSC, CKW, CYH, TJH and ZXW helped check and improve the manuscript. All authors read and approved the final manuscript.

Funding

Changsha Municipal Natural Science Foundation [kq2014144]; Changsha Science and Technology Bureau project [kq2001034]; National Key Research & Development project by the Ministry of Science and Technology of China (2018YFB1003203); State Key Laboratory of High-Performance Computing (No. 201901-11); National Science Foundation of China (U1811462).

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. Zhou S-F, Zhong W-Z. Drug design and discovery: principles and applications. *Molecules* 2017;**22**(2):279.
2. Marshall GRJ. Computer-aided drug design. *Annu Rev Pharmacol* 1987;**27**:193–213.
3. Veselovsky A, Ivanov A. Strategy of computer-aided drug design. *Current Drug Targets-Infectious Disorders* 2003;**3**:33–40.
4. Song CM, Lim SJ, Tong JC. Recent advances in computer-aided drug design. *Brief Bioinform* 2009;**10**:579–91.
5. Inza I, Calvo B, Armañanzas R, *et al*. Machine learning: an indispensable tool in bioinformatics. *Methods Mol Biol* 2010;**593**:25–48.

6. Phillips J, Gibson W, 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.
7. Livingstone DJ. The characterization of chemical structures using molecular properties. *A Survey, J Chem Inf Comput Sci* 2000;**40**(2):195–209.
8. Rogers D, Hahn M. Extended-connectivity fingerprints. *J Chem Inf Model* 2010;**50**:742–54.
9. Chen J-H, Tseng YJ. Different molecular enumeration influences in deep learning: an example using aqueous solubility. *Brief Bioinform* 2020:bbaa092.
10. Consonni V, Todeschini R. Molecular descriptors. In: T. Puzyn, J. Leszczynski, and M.T.D. Cronin (Eds.), *Recent advances in QSAR studies, Methods and applications*. New York: Springer, 2010, 20–102.
11. Todeschini R, Consonni V. *Handbook of Molecular Descriptors*. Weinheim: Wiley-VCH, 2002.
12. Szegedy C, Vanhoucke V, Ioffe S, *et al*. Rethinking the inception architecture for computer vision. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. Piscataway, NJ: IEEE, 2016, 2818–26.
13. He K, Zhang X, Ren S, *et al*. Identity mappings in deep residual networks. In: *European Conference on Computer Vision*. Cham, Switzerland: Springer, 2016, 630–45.
14. Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. *arXiv preprint*, arXiv:1706.03762, 2017.
15. Devlin J, Chang M-W, Lee K, *et al*. Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv* preprint, arXiv:1810.04805, 2018.
16. Silver D, Schrittwieser J, Simonyan K, *et al*. Mastering the game of go without human knowledge. *Nature* 2017;**550**:354–9.
17. Bjerrum EJ. SMILES enumeration as data augmentation for neural network modeling of molecules. *arXiv* preprint, arXiv:1703.07076, 2017.
18. Gilmer J, Schoenholz SS, Riley PF, *et al*. Neural message passing for quantum chemistry. In: *Proceedings of the 34th International Conference on Machine Learning,* Sydney, NSW, Australia, 2017. p. 1263–1272. Proceedings of Machine Learning Research, Cambridge, MA , USA.
19. Winter R, Montanari F, Noé F, *et al*. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chem Sci* 2019;**10**:1692–701.
20. Feinberg EN, Sur D, Wu Z, *et al*. Potential net for molecular property prediction. *ACS Central Science* 2018;**4**:1520–30.
21. Gomes J, Ramsundar B, Feinberg EN, *et al*. Atomic convolutional networks for predicting protein-ligand binding affinity. *arXiv* preprint, arXiv:1703.10603, 2017.
22. Kearnes S, McCloskey K, Berndl M, *et al*. Molecular graph convolutions: moving beyond fingerprints. *J Comput Aided Mol Des* 2016;**30**:595–608.
23. Karpov P, Godin G, Tetko IV. Transformer-CNN: Swiss knife for QSAR modeling and interpretation. *J Cheminform* 2020;**12**(1):17.
24. Xu Z, Wang S, Zhu F, *et al*. Seq2seq fingerprint: An unsupervised deep molecular embedding for drug discovery. In: *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. New York, USA: Association for Computing Machinery, 2017, pp. 285–94.
25. Kadurin A, Nikolenko S, Khrabrov K, *et al*. druGAN: an advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico. *Mol Pharm* 2017;**14**:3098–104.
26. Feinberg EN, Joshi E, Pande VS, *et al*. Improvement in ADMET prediction with multitask deep featurization. *J Med Chem* 2020;**63**:8835–48.
27. Veličković P, Cucurull G, Casanova A, *et al*. Graph attention networks. International Conference on Learning Representations, Vancouver, BC, Canada, 2018. OpenReview.net. International Conference on Representation Learning, La Jolla, CA, USA.
28. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. In: *International Conference on Learning Representations*, Toulon, France, 2017. OpenReview.net. International Conference on Representation Learning, La Jolla, CA, USA.
29. 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**(16): 8749–60.
30. Gao P, Zhang J, Sun Y, *et al*. Accurate predictions of aqueous solubility of drug molecules via the multilevel graph convolutional network (MGCN) and SchNet architectures. *Journal of Machine Learning Research* 2020;**22**:23766–72.
31. Shang C, Liu Q, Chen K-S, *et al*. Edge attention-based multi-relational graph convolutional networks. *arXiv* e-prints, arXiv:1802.04944, 2018.
32. Li G, Müller M, Qian G, *et al*. Deepgcns: making gcns go as deep as cnns. *arXiv* preprint, arXiv:1910.06849, 2019.
33. Zhang Q, Yang LT, Chen Z, *et al*. A survey on deep learning for big data. *Inform Fusion* 2018;**42**:146–57.
34. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. *Commun ACM* 2017;**60**:84–90.
35. Dong J, Wang N-N, Yao Z-J, *et al*. ADMETlab: a platform for systematic ADMET evaluation based on a comprehensively collected ADMET database. *J Chem* 2018;**10**:29.
36. Chen T, Kornblith S, Norouzi M, *et al*. A simple framework for contrastive learning of visual representations. In: *Proceedings of the 37th International Conference on Machine Learning, Virtual Event*. Cambridge, MA , USA: Proceedings of Machine Learning Research, 2020, p. 1597-1607.
37. 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*. New York, USA: Association for Computing Machinery, 2019, p. 429–36.
38. Gaulton A, Bellis LJ, Bento AP, *et al*. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* 2012;**40**:D1100–7.
39. Wu Z, Ramsundar B, Feinberg EN, *et al*. MoleculeNet: a benchmark for molecular machine learning. *Chem Sci* 2018;**9**:513–30.
40. Battaglia PW, Hamrick JB, Bapst V, *et al*. Relational inductive biases, deep learning, and graph networks. *arXiv* preprint, arXiv:1806.01261, 2018.
41. Landrum G. RDKit: Open-Source Cheminformatics Software. http://www.rdkit.org (accessed Aug 20, 2020).
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. *The Journal of Machine Learning Research* 2014;**15**:1929–58.

44. Liu Y, Ott M, Goyal N, *et al*. Roberta: A robustly optimized bert pretraining approach. In: *International Conference on Learning Representations*, Virtual Conference, 2020. OpenReview.net. International Conference on Representation Learning, La Jolla, CA, USA.
45. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, USA: Association for Computing Machinery, 2016, p. 785–94.
46. Wattenberg M, Viégas F, Johnson I. How to use t-SNE effectively. *Distill* 2016;**1**(10):e2.
47. Van der Maaten L, Hinton G. Visualizing data using t-SNE. *Journal of Machine Learning Research* 2008;**9**(11):2579–605.
48. Plošnik A, Vračko M, Sollner Dolenc M. Mutagenic and carcinogenic structural alerts and their mechanisms of action. *Arh Hig Rada Toksikol* 2016;**67**:169–82.