IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 30, 2020, accepted August 26, 2020, date of publication September 9, 2020, date of current version September 24, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3022850

Enhanced Graph Isomorphism Network for Molecular ADMET Properties Prediction

YUZHONG PENG[1,2], YANMEI LIN[2], XIAO-YUAN JING[3,5], (Member, IEEE), HAO ZHANG[3], YIRAN HUANG[4], AND GUANG SHENG LUO[1], (Member, IEEE)

[1]School of Computer Science, Fudan University, Shanghai 200433, China
[2]School of Computer and Information Engineering, Nanning Normal University, Nanning 530001, China
[3]School of Computer, Guangdong University of Petrochemical Technology, Maoming 525000, China
[4]Guangxi Key Laboratory of Multimedia Communications and Network Technology, School of Computer and Electronics and Information, Guangxi University, Nanning 530004, China
[5]School of Computer, Wuhan University, Wuhan 430072, China

Corresponding author: Yiran Huang (hyr@gxu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61862006 and Grant 61562008, in part by the Natural Science Foundation of Guangxi Zhuang Autonomous Region under Grant 2017GXNSFAA198228 and Grant 2020GXNSFAA159074, in part by the Natural Science Foundation of Guangdong Province under Grant 2019A1515011076, and in part by the BAGUI Scholar Program of Guangxi Zhuang Autonomous Region of China and the School Level Scientifific Research Project of Guangdong University of Petrochemical Technology (Talent Introduction) under Grant 2020rc021.

ABSTRACT The evaluation of absorption, distribution, metabolism, exclusion, and toxicity (ADMET) properties plays a key role in a variety of domains including industrial chemicals, agrochemicals, cosmetics, environmental science, food chemistry, and particularly drug development. Since molecules are often intrinsically described as molecular graphs, graph neural networks have recently been studied to improve the prediction of ADMET properties. Among many graph neural networks published in recent years, Graph Isomorphism Network (GIN) is a relatively recent and very promising one. In this paper, we propose an enhanced GIN, called MolGIN, via exploiting the bond features and differences influence of the atom neighbors to end-to-end predict ADMET properties. Based on GIN, MolGIN concatenates the bond feature together with node feature in the feature aggregator and applies a gate unit to adjust the atomic neighborhood weights to map the differences in the interaction strength between the central atom and its neighbors, such that more meaningful structural patterns of molecules can be explored toward better molecular modeling. Extensive experiments were conducted on seven public datasets to evaluate MolGIN against four baseline models with benchmark metrics. Experimental results of MolGIN were also compared with state-of-the-art results published in the last three years on each dataset. Experimental results in terms of RMSE and AUC show that MolGIN significantly boosts the prediction performance of GIN and markedly outperforms the baseline models, and achieves comparable or superior performance to state-of-the-art results.

INDEX TERMS ADMET prediction, graph convolutional network, graph isomorphism network, molecular property, quantitative structure-property relationship.

I. INTRODUCTION

Absorption, distribution, metabolism, exclusion, and toxicity (ADMET) properties are of great importance in drug development and risk assessment of compounds from industrial chemicals, food additives, pesticides, and environmental pollutants [1]. Good ADMET quality is indispensable for a good drug. However, a large number of candidate compounds have the unsuitable quality of ADMET in the process of screening drugs, which results in a long period, high cost, high candidate attrition rate, and many post-marketing withdrawals of the drug development. It was reported that the attrition rate of drug candidates has reached 90% [2] or even more [3], and the preclinical discovery cost of an FDA-approved drug and the overall estimated cost of developing a new drug have been up to 834 million U.S. dollars [4], [5] and 2.6 billion U.S. dollars in recent years [6], [7], respectively. Therefore, the early evaluation of ADMET properties is critical in drug development. The in-silico prediction tools, represented by quantitative structure-property relationship (QSPR) models,

The associate editor coordinating the review of this manuscript and approving it for publication was Mu-Yen Chen.

can be used to evaluate ADMET properties of candidate drug compounds and conduct preliminary virtual screening for drugs. QSPR models enable chemists and engineers to screen and prioritize compounds from vast libraries of molecular candidates before molecular synthesis or assay, eliminating compounds that are likely to possess unfavorable ADMET properties. This thereby greatly saves time and money. QSPR models, based on machine learning (ML) techniques, are powerful ways to boost drug screening and discovery in recent years.

Since QSPR can exploit the intrinsic relationships between molecular structure and the physicochemical and biochemical properties of molecules, QSPR models have been widely developed and applied to the ADMET properties prediction to provide faster, cost-effective and accurate prediction of unknown compounds ADMET properties [8], [9]. QSPR has become the most efficient and popular molecular properties prediction method due to its automatic, efficient, high-throughput, large-scale characteristics [10]. In QSPR models, multiple linear regression, neural network, random forest (RF), support vector machine, and fully connected-based deep neural network (FDNN, or DNN for short in some literature) are typically used to map the structure of compounds represented by hand-crafted features (i.e. fingerprints and descriptors) to ADMET properties [11]–[15]. In recent years, deep neural networks, based on their remarkable capability of learning concrete and sometimes implicit features, have won the huge success in a variety of fields [16]. In particular, deep neural network won the Merck molecular activity challenge in 2012 [17]. These successes have greatly encouraged many researchers to develop Deep Learning (DL)-based QSPR models for predicting the molecular property and drug ADME properties [11], [12], [14], [18]–[21]. DL-based QSPR became the most popular in-silico method, for DL can increase the performance of QSAR models [12], at least when sufficient data is available [17], [22].

In 2015, Duvenaud *et al.* [23] proposed Convolutional Networks on Graphs for Learning Molecular Fingerprints, which bring graph convolutions into the context of DL on molecules. Since then, Graph Neural Networks (GNNs) and their applications have become interesting emerging issues of molecular informatics, increasingly attracting the attention and research in cheminformatics and bioinformatics [17], [24]–[27]. GNNs directly utilize a graph-structured representation of the original molecule as the input data, in which atoms are represented as nodes and bonds are represented as edges of a graph [28]. The graph-structured data is processed by a series of convolutions or equivalent operations of convolutions to map the chemical information of the molecular graph into a distributed multi-dimensional feature vector. The GNN models often outperform the classical ML or DNN models with various fingerprints or descriptors [17], [24]. Unlike traditional feature engineering-based ML models, GNN models are generally less sensitive to the choice of atomic descriptors [26], [27]. During the past three years, many GNN variants have been developed for molecular properties prediction and achieved superior performance than other existing methods [24], [29], [30]. Furthermore, there are some interesting works using GNN for ADMET properties prediction. For example, Feinberg *et al.* [5] used a type of graph convolutional networks to model ADMET properties at Merck; Montanari *et al.* [31] demonstrated that a multitask graph convolutional approach for predicting physicochemical ADMET endpoints appears a highly competitive choice; Feinberg *et al.* [32] applied graph convolutions to an explicit molecular representation, and achieved unprecedented accuracy in the prediction of ADMET properties.

Among these variants of the GNN model, Graph Isomorphism Network (GIN) [33] has the maximum discriminative/representational power from different graph structures and quantifies generalization ability, quickly attracting extensive attention in GNN community. Since the original paper of Graph Isomorphism Network published by Xu *et al.* [33] in 2019, it has now been cited more than 160 times. Many works [33]–[39] have shown that GIN and GIN-based models significantly outperform other GNN models in many graph learning tasks, including molecular modeling and prediction. For example, Capela *et al.* [34] compared GIN with graph attention isomorphism network (GAIN) and gated graph recursive (isomorphism neural network, GGRNet) and their multitasking learning model on the molecular physicochemical properties prediction, which results show that GIN achieves the best performance and the multitasking learning model of each variant of GIN can achieve higher performance than its single-tasking learning model. Nguyen *et al.* [35] developed a model named GraphDTA for drug-target binding affinity prediction using GIN, and their experimental results show that this method not only predicts affinity better than non-deep learning models, but also outperforms the comparative deep learning models.

Despite the burst of successful excitements of GIN in the GNN community, as a general-purpose GNN model, GIN has not fully exploited the special information of the molecular graph, such as bond features, the differences of the interaction strength between the central atom and its neighbors. This special molecular information implies some meaningful molecular patterns. In this work, we aim to improve GIN toward better molecular modeling and ADMET properties prediction via exploiting the special molecular information. Our major contributions are summarized as follows:

1) We proposed a method called MolGIN to enhance the molecular modeling of GIN by exploiting bond features and the differences in the interaction strength between the central atom and its neighbors.
2) We designed a scheme to incorporate the molecular bond feature into the node information aggregation via concatenating the feature vectors of the atom neighbors and the bonds connecting with the atom.
3) We developed a control gate unit in the aggregator to adjust the neighborhood weights of the atom to map the differences in the interaction strength between the central atom and its neighbors.

4) We conducted extensive performance evaluations of our method on the classification and regression tasks of ADMET properties. The results show that incorporating the information of bond features and differences in the interaction strength between the central atom and its neighbors into the molecular modeling of GIN indeed markedly benefits ADMET properties prediction performance.

II. PRELIMINARIES

Before introducing MolGIN, we briefly introduce Graph Neural Network and Graph Isomorphism Network, which are the theoretical basic of MolGIN.

A. GRAPH NEURAL NETWORK

Graph neural network (GNN) is a special neural network architecture with the same essence as a convolutional neural network, but it is mainly used for processing and learning very irregular and unstructured graph data [40]. GNNs aim to learn the representation of each node in the graph, and hierarchically extract features of the nodes or graphs, and then use the final features for application modeling by a sub-model, such as multi-layer perceptron (MLP). GNNs leverage the graph structure and iteratively update the node representation from the node neighborhoods in a fashion of convolutional operation or equivalent to obtain the final feature representation of the nodes or the graph [40]. To explore the deeper and more extensive information of the node's receiving domain, multiple graphical convolution (or equivalent) layers are usually stacked together to update the node representation. In the simplest form, GNNs can update the node representation according to the following formula [41], [42]:

$$h_i^{l+1} = f(A) = \begin{cases} \frac{1}{|N_i|}\psi\left(\sum_{j\in N_i} h_j^l W^l\right), & l > 0 \\ A_i, & l = 0 \end{cases} \quad (1)$$

where h_i^{l+1} denotes the hidden state of the i-th node in the $(l+1)$-th layer; $h_i^0 = A_i$; ψ is a nonlinear transformation; N_i is the nodes connected to the i-th node on the graph; $|N_i|$ is the degree of the i-th node; W^l is a learnable parameter representing the weight matrix of the l-th layer; A_i is the attribute matrix of the i-th node.

B. GRAPH ISOMORPHISM NETWORK

Graph Isomorphism Network (GIN) [33] is one of the most potential GNN variations, and its discriminative/representational power is equal to the power of the Weisfeiler-Lehman (WL) graph isomorphism test [43], [44]. GIN replaces (1) with the following formula [33], [45]:

$$h_i^{l+1} = MLP^l\left((1+\varepsilon^l)h_i^l + \sum_{j\in N_i} h_j^l\right), \quad (2)$$

where $h_i^0 = A_i$, ε^l is a learnable parameter, *MLP* is a Multi-layer Perceptron. The (2) shows that: 1) GIN replaces the mean aggregator over nodes of the traditional GNN with a sum aggregator; 2) GIN adds an *MLP* after aggregating node features from the nodes' neighborhoods; and 3) in GIN, each neighbor contributes equally to the update of the central node. Furthermore, GIN concatenates the information of the nodes' representation across all layers of the model for the final representation according to the following formula:

$$h_G = CONCAT\left(\sum_{v\in G, k=0}^{L} h_v^k\right), \quad (3)$$

where v, G are the node and graph, respectively. $CONCAT(\cdot)$ denotes concatenate function.

It has been demonstrated theoretically and experimentally that GIN has more discriminative or representational power of graph structures than previous GNN models [33].

III. METHODOLOGY

A. MOLECULAR GRAPH AND ADMET PROPERTIES PREDICTION

Formally, a molecular graph $G = (V, E)$ is defined by a set of vertices V and a set of edges $E = \{(v, u)|v, u \in V\}$, corresponding the structure of the molecule, in which V refers to the set of atoms of the molecule and E refers to chemical bonds connecting the atoms. The neighbor of atom v is defined as $N_v = \{v \in V|(v, u) \in P\}$. P denotes a set of atoms pair connected by bonds.

ADMET properties prediction can be regarded as a binary classification or regression problem of molecular graphs in machine learning, which is formalized as follows: given a set of molecular graphs $\mathcal{G} = \{G_1, \cdots, G_n\}$ and their label set $Y = \{y_1, \cdots, y_m\}$, each molecular graph has an attribute vector of atoms, $a, a \in A$, and an attribute vector of bonds, $z, z \in Z$, respectively. Then the representation learning function $\theta(\cdot)$ learns a representation vector $h_{\mathcal{G}} = \theta(A, Z)$ that can help to predict the labels. Finally, the labeling function $\vartheta(\cdot)$ assigns the label of the entire molecular graph $y = \vartheta(h_{\mathcal{G}})$.

B. OUR ENHANCED GIN

The original work of GIN [33] focused on building a universal graph neural network with the same discriminative/representational power as the Weisfeiler-Lehman (WL) graph isomorphism test to represent the graph structure. Many researchers have used GIN to solve problems related to small molecules and have achieved good performance, however, GIN still has shortcomings in presenting small molecules. GIN neglects two important pieces of information and characteristics of small molecules: bond features and differences influence of neighbors, which imply local structures and some important patterns of molecular property. To tackle this problem, we enhanced the representational power of GIN on molecular graphs through two schemes: bond feature concatenation and neighborhood weight adjustment using a gate unit. This enhanced GIN called MolGIN that will be detailed in the following subsections.

1) EXPLOITING BOND FEATURE

We exploit the bond feature to mine local patterns of small molecules when aggregating node features from its neighborhoods. Let E_v be the set of bonds connected to atoms in N_v. z_v denotes the feature vector of E_v, which can be simply defined as below.

$$z_v = \varphi\left(\sum_{e_{uv}\in E_v} F_{uv}\right), \tag{4}$$

where φ is a real-valued function, simplified to a linear transform in this work. F_{uv} refers to the feature vector of the bond e_{uv} connecting with the atoms u and v. We concatenate z_v to the neighbors' feature vector of the central atom h_v on each layer of the GIN aggregator. Then (2) can be replaced by the formulation as below,

$$h_i^{l+1} = MLP^l\left((1+\varepsilon^l)h_i^l + \sum_{j\in N_i}\left(h_j^l \oplus z_j^l\right)\right), \tag{5}$$

where $\oplus$ denotes the concatenation. And (3) can be replaced by the formulation as below,

$$h_G = CONCAT\left(\sum_{v\in G, l=0}^{L}\left(h_v^l \oplus z_v^l\right)\right), \tag{6}$$

As a result, the proposed method can aggregate the information of the atom neighbors with the bonds and encode the local structure patterns into hidden vectors.

2) LEARNING NEIGHBORHOOD WEIGHTS

In GIN, all neighbors make an equal contribution to the update of the central node, which leads to neglecting the differences in influence strength between the central node and its different neighbors. To address this problem, inspired by the gate unit theme in LSTM [46], we introduce a control gate unit to adjust the contribution of neighbors in updating the central node feature. Concretely, the neighbors' feature vector h_v is multiplied by the control gate activation, and calculated by the Sigmoid function ψ in the range [0,1]. Hence, (5) can be redefined as follows:

$$h_i^{l+1} = MLP^l\left((1+\varepsilon^l)h_i^l + \sum_{j\in N_i}\left(\psi\left(h_j^l W^l + b^l\right)\odot h_j^l \oplus z_j^l\right)\right), \tag{7}$$

where $\odot$ is the element-wise multiplication, W^l and b^l are the weight matrix and bias of the l-th layer, respectively. In this way, $\psi(\cdot)$ is acted as a learnable controller of neighborhood weights, which can learn the weight matrix to tune different influence strengths between the central node and its neighbors during the training phase.

3) MOLECULAR GRAPH EMBEDDING ALGORITHM

Alg.1 outlines the general molecular graph embedding process of MolGIN based on the aforementioned theory. In Alg.1, $G(V, E)$ is the molecular graph to be embedded;

Algorithm 1 An Enhanced Graph Isomorphism Network for Molecular Modeling (MolGIN)

Require: the molecular graph $G(V, E)$; the input atomic vector $a_i, \forall i \in V$; the input bond vectors $z_i, \forall z \in E$; the learning depth of the aggregator L.

1: **for each** atom i in V **do**
2: $z_i \leftarrow \sum_{j\in N_i} z_{ij}$;
3: $h_i^0 \leftarrow a_i \oplus z_i$;
4: **end for**
5: **for** $(l = 1; l \le L; l++)$ **do**
6: **for each** atom i in V **do**
7: $jTempt \leftarrow \sum_{j\in N_i}(\psi(h_j^{l-1}W^{l-1} + b^{l-1}) \odot h_j^{l-1} \oplus z_j^{l-1})$;
8: $h_i^l \leftarrow MLP^{l-1}\left((1+\varepsilon^{l-1})h_i^{l-1} + jTempt\right)$;
9: **end for**
10: **end for**
11: $h_G \leftarrow CONCAT(\sum_{i\in V,\ l=0}^{L}(h_i^l \oplus z_i^l))$;
12: Output h_G.

N_i refers to the neighborhood node (atom) set; L refers to the learning depth (number of layers) of the aggregator, i.e. "search depths"; a_i denotes the feature vector of the i-th atom in G; z_{ij} denotes the feature vector of the bond connecting the i-th atom and the j-th atom; h_G denotes the molecular graph-level embedding representation. The loop shown at lines 1-4 is to initialize the representation (hidden state) of each node $h_i, i \in V$. The loop shown at lines 5-10 is to aggregates and update the representation (hidden state) of each node h_i iteratively. Line 11 is to sum the node's representation in each layer of the aggregator for each node and concatenate them to produce the molecular graph-level embedding representation h_G.

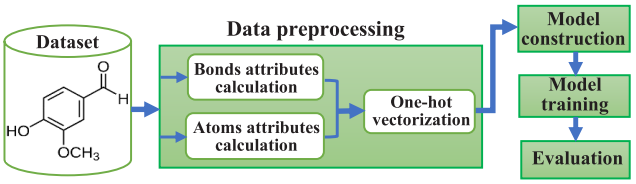

FIGURE 1. The pipeline of MolGIN-based ADMET properties prediction.

C. ADMET PROPERTIES PREDICTION USING MolGIN

Fig. 1 illustrates the pipeline of ADMET properties prediction using MolGIN. It can be summarized as four phases: data preprocessing, model construction, model training, and prediction. Since the prediction phase is the same as the standard machine learning, only the first three phases will be detailed in the subsequent sections.

1) DATA PREPROCESSING

In this work, dataset preprocessing mainly includes attributes calculation of each atom and bond, and initial featurization using one-hot encoding.

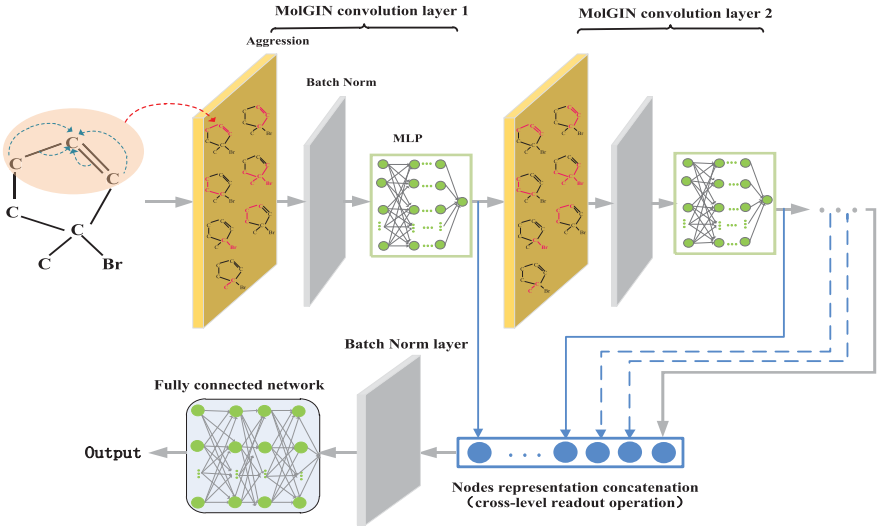

FIGURE 2. The semantic structure illustration of MolGIN. MolGIN mainly consists of one or more MolGIN convolutional layers, a node representation concatenation layer, some Batch Norm layers, and a fully connected neural network.

TABLE 1. Features characterizing the atoms and bonds.

Attribute name	#One-hot encoding dimension
Atomic attributes	
Atom type	45
Atom Degree	7
Formal charge	5
Implicit Valence	7
Number of Radical Electrons	8
Total Number of Hydrogens	5
Chirality type	4
Hybridization	5
InRingSize	6
In-aromaticity	1
IsAcceptor	1
IsDonor	1
Bond attributes	
Bond type	4
In-ring	1
IsConjugated	1
Bond stereo	6

Note: The atom type refers to the type of atom,including C, N, O, S, F, Si, P, Cl, Br, Mg, Na, Ca, Fe, As, Al, I, B, V, K, Tl, Yb, Sb, Sn, Ag, Pd, Co, Se, Ti, Zn, H, Li, Ge, Cu, Au, Ni, Cd, In, Mn, Zr, Cr, Pt, Hg, Pb, or "Unknown" in this work.

Firstly, given a molecular sample, the atoms and bonds were characterized by a set of attributes, as listed in Table 1. Then, the attributes of each atom and each bond of the molecular sample were respectively calculated by the open-source cheminformatics toolkit RDKit.

Next, the attributes of each atom and each bond were encoded as $a = \alpha_1 \oplus \alpha_2 \oplus \cdots \oplus \alpha_{lenA}$ and $z = \beta_1 \oplus \beta_2 \oplus \cdots \oplus \beta_{lenB}$, where a, $lenA$ and α_i denote the general atomic one-hot vector, the number of atomic attributes, and the i-th attribute feature of the atom, respectively; z, $lenB$ and β_i denote the general bond one-hot vector, the number of bond attributes, and the i-th attribute of the bond, respectively. The length of α_i (or β_i) is the number of possible values of α_i (or β_i). Note that $A = \{a_1, a_2, \cdots, a_{n_a}\}$ and $Z = \{z_1, z_2, \cdots, z_{n_b}\}$ were used as inputs of the aggregator component in our model, where n_a, n_b, $a_j (1 \le j \le n_a)$, and $z_k (1 \le k \le n_b)$ denote the total number of atoms, the total number of bonds, the general one-hot vector of the j-th atom, and the general one-hot vector of the k-th bond of the molecular sample, respectively.

2) MODEL CONSTRUCTION

We constructed a MolGIN network, as shown in Fig. 2, to learn a predictor and graph-level embedding that can be regarded as a task-optimized fingerprint. For simplicity, MolGIN construction mainly includes the following two steps. Firstly, we build a representation learning function $\theta(\cdot)$ for molecular graphs according to the method given in section III-B. In this way, MolGIN can learn the graph-level distributed representation vector h from the attributes of the atoms and bonds by iteratively aggregating messages and then performing a cross-level readout operation. We then add a predictor, i.e. the labeling function $\vartheta(\cdot)$, on the top of $\theta(\cdot)$ to produce the prediction result y for the given molecule. In this work, on the one hand, we use the linear regression layer formulated in (8) as a predictor for property regression.

$$y = \vartheta(h_{\mathcal{G}}) = hw^T + b, \tag{8}$$

where h is the input feature vector of the predictor, w^T and b are the weight vector and bias. On the other hand, we use the SoftMax layer formulated in (9) as a predictor for property classification.

$$Q_i = log\left(\frac{e^{h_i}}{\sum_{k=1}^{2} e^{h_k}}\right), \quad i = 1, 2 \tag{9}$$

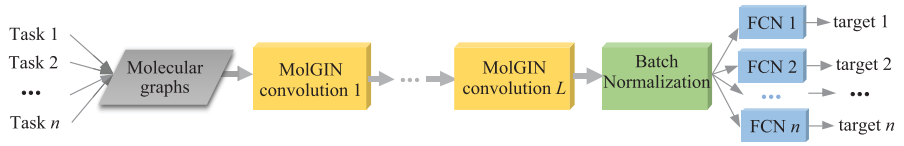

FIGURE 3. A semantic illustration of the multi-task MolGIN framework. FCN denotes the fully connected neural network. Each task shares the components for the molecular representation learning to learn more generalized features and has its own FCN for constructing the corresponding predictor of its target.

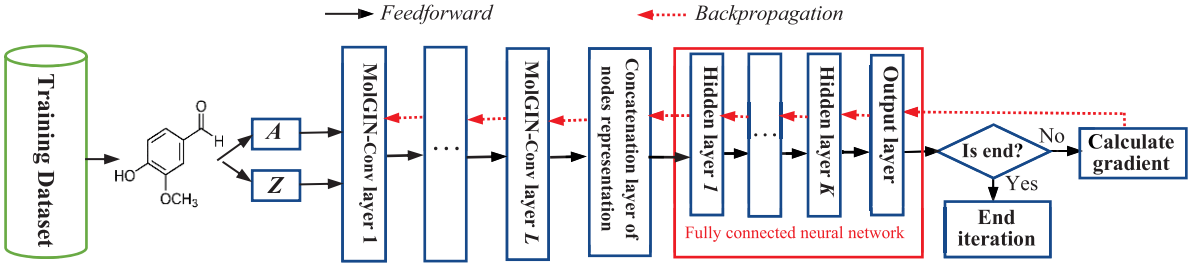

FIGURE 4. The training process of MolGIN. *A* and *Z* are the atom attribute set and bond attribute set of the given molecular sample in the dataset, respectively. *L* and *K* are the number of the MolGIN convolution layers and the number of the hidden layers of the fully connected neural network, respectively.

where Q_i is the probability that a chemical with feature vector h belongs to the i-th label (positive or negative in this paper). h_i refers to the i-th element value of the feature vector h.

Recently, some studies have demonstrated that multi-task learning models may offer improvements over the single-task learning model for modeling ADMET data sets [31], [32], [47], [48]. Following these studies, this work also builds a multi-task MolGIN model, as shown in Fig. 3, to demonstrate fair comparisons with the existing multi-task learning models on ADMET properties prediction.

3) MODEL TRAINING

Fig. 4 shows the workflow of MolGIN training. Given a training set $\{(X_i, y_i)|N = |X|, 1 \leq i \leq N\}$,where X_i and y_i are respectively the graph independent variables and ground-truth values of the input data for the i-th molecule in the training data. $X_i = A_i \cup Z_i$, where A_i and Z_i are the atoms attribute set and bonds attribute set of the i-th molecule, respectively. The MolGIN model is iteratively trained end-to-end until the preset criterion (e.g. maximum iterations, loss threshold) is reached. In each iteration, the model first calculates the sample output of the training set in the feedforward direction and tunes it by supervised backpropagation to minimize the loss function $L(\cdot)$. The loss function is defined by Mean Square Error formulated as (10) for a property regression task (e.g. logS), or defined by Cross Entropy formulated as (11) for a property classification task (e.g. Ames).

$$L(\tau) = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - (\hat{y}_i|X_i;\tau))^2}{\sigma^2}, \tag{10}$$

$$L(\tau) = \sum_{i=1}^{n} y_i \cdot log(P_i(\hat{y}_i|X_i;\tau)), \tag{11}$$

where τ denotes the parameter set in MolGIN, n denotes the mini-batch size, $\hat{y}_i$ denotes the predicted value of the i-th training sample.

For the multi-task learning MolGIN model (MT-MolGIN), we use the weighted sum of the losses in the backpropagation during the training process. The weighted sum of losses for MT-MolGIN is defined as

$$L_{sum}(\tau) = \frac{1}{M}\sum_{i=1}^{n} c_i L_i(\tau), \tag{12}$$

where, M and c_i are the number of tasks and the weights for the individual losses; $L_i(\tau)$ is the individual loss of each task-specific layer, which is defined by (10) for the property regression task or (11) for the property classification task.

IV. EXPERIMENTS AND RESULTS

A. BENCHMARK DATASETS

Seven public datasets are used in this work. The brief descriptions of these datasets are listed below:

- **Log$D_{7.4}$ dataset:** This is a benchmark dataset for lipophilicity prediction. It contains experimental results of octanol-water partition coefficients of 4200 compounds at physiological pH (Log$D_{7.4}$), which were curated from ChEMBL database [34].
- **LogS dataset:** In this work, the LogS dataset is a solubility dataset collected from [19], known as ESOL. It contains 1128 compounds with the logarithm of aqueous solubility value in mols per liter.
- **LD50 dataset:** This is a dataset about the oral 50% lethal dose in rats, usually used to evaluate the acute oral toxicity of compounds. It contains experimental values of 7413 chemicals that kill 50% of the treated rats immediately or in a short time, after single or multiple

doses administration within 24 hours. The LD50 values are relatively difficult to predict accurately due to the large difference between the maximum and minimum value [49].

- **PPB dataset:** Plasma protein binding (PPB) has a strong influence on the drug pharmacodynamic behavior and its oral bioavailability, which is necessary to predict and assay at the early stage of drug development. The PPB dataset in this work contains 1830 compounds collected from [50].
- **Pgp inhibitors dataset:** P-glycoprotein plays a vital role in pharmacokinetic processes such as absorption, distribution, metabolism, and excretion. The Pgp-inhibitor dataset in this work contains 1275 compounds that were collected from [51].
- **Ames dataset:** This Ames dataset contains the Ames testing experiment mutagenicity data of 7619 compounds collected from the previous work [52], and is widely used to develop good in-silico models for compounds mutagenicity prediction instead of Ames testing.
- **Tox21 dataset:** This Tox21 dataset is used in the Tox21 Data Challenge launched by the United States agencies (NIH, EPA and FDA) [53]. It is a benchmark toxicity prediction database that comprises 12 different cellular assays values of 8458 compounds corresponding to 12 different targets [19]. Twelves different cellular assays correspond to 12 classification tasks, although there are some missing values in each cellular assay [10], [19].

The statistical information of these datasets is shown in Table 2.

TABLE 2. Statistics of the ADMET properties datasets.

Category	Property	#Total	#Positives	#Negatives	Task type
Basic physicochemical property	Log$D_{7.4}$	4200	—	—	regression
	LogS	5220	—	—	regression
Absorption	Pgp-inhibitor	2297	1372	925	classification
Distribution	PPB	1830	—	—	regression
Toxicity	Ames	7619	4252	3376	classification
	Tox21	8014	—	—	Multi-classification
	LD50	7413	—	—	regression

Note: The 12 different targets on the Tox21 dataset have different numbers of positive and negative samples. The maximum and minimum numbers of positive samples are 1093 and 219 on SR-ARE and NR-PPAR-gamma subclasses, respectively.

B. EXPERIMENTAL SETUP

1) PERFORMANCE EVALUATION METHOD

To evaluate the model performance, following the previous works [19], [51], [54], we used the root mean squared error (RMSE) for regression tasks and area under the receiver-operating characteristic curve (AUC) for classification tasks. Note that RMSE and AUC are the most used benchmark metrics for evaluating molecular properties prediction models in recent year. Higher AUC is better for classification tasks, while lower RMSE is better for regression tasks.

To verify the advantage of the proposed method, we first evaluated MolGINs (i.e. single- and multi-task MolGIN models, denote by ST-MolGIN and MT-MolGIN, respectively.) against 4 baseline comparative models. Two of these baseline models were descriptor/fingerprint-based models and the other two are GNN models. These two descriptors/fingerprints-based models were fully connected Deep Neural Network (FDNN) and Random Forest (RF), which are the most widely used models by major pharmaceutical companies [5]. The two GNN baseline models were GIN detailed in Section II-B and GCN [31]. GCN is the most used GNN for ADMET properties prediction in recent years.

Besides, there are 39 state-of-the-art models published in the last three years for different ADMET datasets, as listed in Table 3. These models including single- and multi-task learning models can be roughly categorized into shallow machine learning models, deep learning models, and their hybrid models. To the best of our knowledge, no model has been demonstrated its performance on all datasets in this work, each model on only some of these ADMET datasets. Therefore, we also compared results of our proposed model with these 39 state-of-the-art results on different ADMET datasets to verify the advantage of our MolGIN.

To obtain unbiased and objective results, 5-fold cross-validation and external test were used to assess the performance comparison between MolGIN and all baseline models for each dataset. We randomly selected 85% samples for 5-fold cross-validation and then used the remaining samples as an external test set. Each comparison experiment was run 20 times independently, and its final experiment result was the average value of 20 runs.

Following the previous work [19], [51], this work used ECFP4 fingerprints as the input data for FDNN and RF models, while the input data for GCN, GIN, and MolGIN were the same as the molecular graphs described in Section III-C1. The initial feature data of molecular graphs and all ECFP4 fingerprints in this work were computed by the cheminformatics toolkit RDKit.

2) HYPERPARAMETER SETTING

For fairness and convenience, in this work, MolGIN and GIN used the same hyperparameter settings, following the hyperparameter settings of the previous work [33], as follows: the numbers of GNN layers (including the input layer) and MLP layers were 5 and 2, respectively; Batch normalization was applied on each hidden layer; an Adam optimizer with initial learning rate 0.01 and a learning rate decay with 0.5 every 50 epochs were applied; the number of hidden units $\in \{16, 32\}$ and the dropout ratio $\in \{0, 0.5\}$ and the batch size $\in \{32, 128\}$ were tuned for each dataset [33]. The same hyperparameters are used in the multi-task model MT-MolGIN.

The hyperparameter setting of GCN followed the hyperparameter settings of Montanari's single-task work [31].

TABLE 3. Summary of state-of-the-art models for different ADMET datasets.

No	Name	Category	Datasets reported	Note
1	ADMETLab	Shallow ML models	$\text{Log}D_{7.4}$, LogS, PPB, Pgp-inhibitor, Ames, and Tox21	A platform for systematic ADMET evaluation containing a variety of shallow machine learning-based QSPR models with various descriptors/fingerprints published by Dong et al. in 2018 [51].
2	admetSAR 2.0	Shallow ML models and DL–>GNN	LD50, LogS, PPB, Pgp-inhibitor, and Ames	A comprehensive source and free platform for evaluating chemical ADMET properties using QSPR models containing a variety of shallow machine learning-based QSPR models and deep Graph convolutional neural network published by Yang et al. in 2019 [55].
3	SELU-MPNN	DL–>GNN	$\text{Log}D_{7.4}$, LogS, and Tox21	A message passing neural networks (MPNNs) using SELU activation function published by Withnall et al. in 2020 [56].
4	AMPNN	DL–>GNN	$\text{Log}D_{7.4}$, LogS, and Tox21	An attention MPNNs published by Withnall et al. in 2020 [56].
5	EMNN	DL–>GNN	$\text{Log}D_{7.4}$, LogS, and Tox21	An edge memory neural network based on MPNN architectures published by Withnall et al. in 2020 [56].
6	D-MPNN	DL–>GNN	$\text{Log}D_{7.4}$, LogS, and Tox21	A directed MPNN published by Yang et al. in 2019 [30].
7	D-MPNN-ensemble	Ensemble; DL–>GNN	$\text{Log}D_{7.4}$, LogS, and Tox21	An ensemble directed MPNN published by Yang et al. in 2019 [30].
8	MoleculeNET	Various shallow ML and DL models	$\text{Log}D_{7.4}$, LogS, and Tox21	MoleculeNET is large-scale benchmark for molecular machine learning that released as part of the DeepChem open source library in 2018, containing various Shallow and deep learning models, e.g. SVM, RF, multitask networks, GCN and Weave [19].
9	SAMPN	DL–>GNN	$\text{Log}D_{7.4}$ and LogS	A self-attention-based message passing neural network for predicting molecular lipophilicity and aqueous solubility published by Tang et al. in 2020 [57].
10	Multi-SAMPN	Multi-task; DL–>GNN	$\text{Log}D_{7.4}$ and LogS	A multi-task SAMPN published by Tang et al. in 2020 [57].
11	MPN	DNN–>GNN	$\text{Log}D_{7.4}$ and LogS	A message passing neural network developed by Tang et al. in 2020 [57].
12	Multi-MPN	Multi-task; DL–>GNN	$\text{Log}D_{7.4}$ and LogS	A multi-task message passing neural network developed by Tang et al. in 2020 [57].
13	GAIN	DL–>GNN	$\text{Log}D_{7.4}$ and LogS	A graph attention isomorphism network published by Capela et al. in 2020 [34].
14	GGRNET	DL–>GNN	$\text{Log}D_{7.4}$ and LogS	A gated graph recursive neural network developed by Capela et al. in 2020 [34].
15	MT-GAIN	Multi-task; DL–>GNN	$\text{Log}D_{7.4}$ and LogS	A multi-task graph attention isomorphism network published by Capela et al. in 2020 [34].
16	MT-GGRNET	Multi-task; DL–>GNN	$\text{Log}D_{7.4}$ and LogS	A multi-task gated graph recursive neural network published by Capela et al. in 2020 [34].
17	LightGBM-FP	Ensemble; Shallow ML	Ames and Tox21	A recent improvement of the gradient boosting algorithm, LightGBM, coupling with Morgan Fingerprints, which was published by Zhang et al. in 2019 [58].
18	LightGBM-MD	Ensemble; Shallow ML model	Ames and Tox21	A recent improvement of the gradient boosting algorithm, LightGBM, coupling with RDKit Molecular Descriptors, which was published by Zhang et al. in 2019 [58].
19	ProTox-II	Shallow ML models	Ames and Tox21	A freely available webserver for in silico toxicity prediction containing a variety of shallow machine learning-based QSPR models with various descriptors/fingerprints published by Banerjee et al. in 2018 [59].
20	Image-based CNN	DL–>CNN	Ames and Pgp-inhibitor	A molecular 2-D image-based CNN method for ADMET properties prediction, which was published by Shi et al. in 2019 [60].
21	novelBayes-1	Shallow ML model	Ames	A naive Bayes classification model for predicting Ames mutagenicity published by Zhang et al. in 2017 [61].
22	k-NN-emsemble	Ensemble; Shallow ML model	Ames	A K-NN ensembles classifier based on supervised projections for Ames mutagenicity prediction published by Garcia et al. in 2018 [62].
23	SVM	Shallow ML model	Ames	An SVM classifier for Ames mutagenicity prediction published by Garcia et al. in 2018 [62].
24	SVM-emsemble	Ensemble; Shallow ML model	Ames	An SVM ensembles classifier based on supervised projections for Ames mutagenicity prediction published by Garcia et al. in 2018 [62].
25	Autoencoder+MLP	DL–>hybrid	Tox21	A pre-trained auto-encoder coupled with MLP to build classifiers for the toxicity prediction published by Galushka et al. in 2018 [63].
26	Toxic Colors	DL–>CNN	Tox21	A supervised convolutional neural network based on molecular 2-D images for the toxicity prediction published by Fernandez et al. in 2018 [64].
27	DeepAOT	DL–>GNN	Tox21	An improved molecular graph encoding convolutional neural network architecture for acute oral toxicity prediction first presented by Xu et al in 2017 [65].
28	Multi-channel grid CNN	DL–>CNN	Tox21	A deep CNN based on multi-channel two-dimension grids of molecules for toxicity prediction published by Yuan et al. in 2019 [66].
29	BESTox	DL–>CNN	LD50	A convolutional neural network regression model Based on binary-encoded SMILES for acute oral toxicity prediction, which was published by Chen et al. in 2020 [67].
30	Top-MT-DNN-best	Multi-task; DL–>hybrid	LD50	The best model of multi-task deep neural networks using algebraic topology approach for quantitative toxicity prediction published by Wu et al. in 2018 [68].
31	Top-consensus	Ensemble; Multi-task; DL–>hybrid	LD50	An ensemble multi-task deep neural networks model using algebraic topology approach for quantitative toxicity prediction published by Wu et al. in 2018 [68].
32	Top-best	Multi-task; DL–>hybrid	LD50	The best model of deep neural networks using algebraic topology approach for quantitative toxicity prediction published by Wu et al. in 2018 [68].
33	GBDT-best	Shallow ML model	LD50	A gradient boosting decision tree model using algebraic topology approach for quantitative toxicity prediction published by Wu et al. in 2018 [68].
34	BTAMDL1	Multi-task; DL–>hybrid	LD50	A boosting tree-assisted multitask deep learning architecture integrating GBDT and multitask deep learning, but without additional features in GBDT, which published by Jiang et al. in 2020 [69].
35	BTAMDL2	Multi-task; DL–>hybrid	LD50	A boosting tree-assisted multitask deep learning architecture integrating GBDT and multitask deep learning with fingerprints in GBDT, which published by Jiang et al. in 2020 [69].
36	NSGA-II-SVM	Shallow ML model	PPB	A support vector machine model for ADME properties evaluation based on descriptors selection by non-dominated sorting genetic algorithm (NSGA-II) combining partial least square (PLS) regression published by Wang et al. in 2017 [50].
37	NSGA-II-RF	Shallow ML model	PPB	A random forest model for ADME properties evaluation based on descriptors selection by non-dominated sorting genetic algorithm (NSGA-II) combining partial least square regression published by Wang et al. in 2017 [50].
38	NSGA-II-GP	Shallow ML model	PPB	A Gaussian process model for ADME properties evaluation based on descriptors selection by non-dominated sorting genetic algorithm (NSGA-II) combining partial least square (PLS) regression published by Wang et al. in 2017 [50].
39	NSGA-II-consensu	Ensemble; Shallow ML model	PPB	A consensus model for ADME properties evaluation based on descriptors selection by non-dominated sorting genetic algorithm (NSGA-II) combining partial least square (PLS) regression published by Wang et al. in 2017 [50].

TABLE 4. Performance comparison between MolGIN and baseline models on regression datasets.

Model	Cross-validation (RMSE)				Test (RMSE)			
	LD50	Log$D_{7.4}$	ESOL	PPB	LD50	Log$D_{7.4}$	ESOL	PPB
RF	0.686±0.006	0.818±0.022	1.012±0.013	18.787±0.012	0.705±0.008	0.827±0.028	0.970±0.018	19.002±0.015
FDNN	0.633±0.015	0.793± 0.038	0.998±0.012	17.968±0.028	0.641±0.017	0.816± 0.017	0.869±0.016	18.067±0.032
GCN	0.628±0.013	0.675±0.037	1.009±0.12	16.315±0.025	0.621±0.015	0.654± 0.031	0.959±0.011	15.238±0.028
GIN	0.625± 0.012	0.666±0.005	0.664±0.006	15.651±0.017	0.599±0.014	0.648± 0.008	0.646±0.008	14.683± 0.018
ST-MolGIN	0.611±0.013	0.590± 0.0010	0.611±0.006	15.145±0.026	0.584±0.016	0.581±0.013	0.581±0.012	14.150±0.028
MT-MolGIN	0.592±0.0015	0.586±0.015	0.593± 0.009	15.139±0.031	0.575±0.018	0.569± 0.021	0.587±0.019	14.143±0.034

Note: The lower the score of RSME, the better the model performance.

TABLE 5. Performance comparison between MolGIN and baseline models on classification datasets.

Model	Cross-validation (AUC)			Test (AUC)		
	Pgp-inhibitor	Ames	Tox21	Pgp-inhibitor	Ames	Tox21
RF	0.899± 0.005	0.889±0.004	0.764±0.005	0.912± 0.005	0.893± 0.005	0.771± 0.011
FDNN	0.878± 0.010	0.817±0.008	0.778±0.012	0.895± 0.013	0.784± 0.012	0.789±0.018
GCN	0.886± 0.012	0.876±0.007	0.827±0.013	0.896± 0.015	0.849±0.015	0.829± 0.011
GIN	0.899± 0.009	0.891± 0.005	0.834± 0.012	0.903± 0.012	0.898± 0.08	0.831±0.013
ST-MolGIN	0.912± 0.012	0.915± 0.006	0.846± 0.012	0.930± 0.015	0.918± 0.012	0.851± 0.015
MT-MolGIN	0.905± 0.018	0.921± 0.009	0.848±0.012	0.933± 0.021	0.915± 0.016	0.858± 0.019

Note: The higher the score of AUC, the better the model performance.

The hyperparameter setting in FDNN followed the hyperparameter setting of Jan's single-task work [48]. Due to the space limitations, hyperparameters details for the GCN and FDNN are not listed here, see [31] and [48] for more details, respectively. The hyperparameters in RF are set to their default values in the Scikit-learn toolkit.

C. RESULTS AND DISCUSSION

We evaluated the effectiveness of our enhanced method for building the graph neural network to learn molecular representations from atoms and bonds data, comparing with RF, FDNN, GCN, GIN, and 39 state-of-the-arts models (see in Table 3) on 7 ADMET datasets. Among these models, there are various shallow machine learning and deep learning models with different architectures and input data. Some of them are multi-task models or ensemble models. Note that, in this work, all comparative state-of-the-art values are cited from the corresponding original references. The comparison results are detailed in the following subsections.

1) COMPARISON WITH BASELINE MODELS

In this section, we first compared our MolGIN with the baseline modes in terms of RMSE on four ADMET regression datasets. Table 4 shows the comparison results of these regression tasks.

The comparison results in Table 4 show that, no matter on 5-fold cross-validation or external test, MolGIN significantly achieved lower errors than not only the descriptors/fingerprints-based baseline models but also the end-to-end learning GNN baseline models on all 4 ADMET regression tasks. The ESOL and Log$D_{7.4}$ tasks were the two most profitable ADMET regression using MolGIN. For ESOL prediction, the single-task MolGIN improved 40.10%, 33.14%, 39.42%, and 10.06% comparing with RF, FDNN, GCN, and GIN on the test set, respectively. For Log$D_{7.4}$ prediction, the single-task MolGIN improved 29.75%, 28.80%, 11.16%, and 10.34% comparing with RF, FDNN, GCN, and GIN on the test set, respectively. Although LD50 prediction was the least beneficial among the four ADMET regression tasks using MolGIN, the single-task MolGIN improved 17.16%, 8.89%, 5.96%, and 2.50% comparing with RF, FDNN, GCN, and GIN, respectively. Besides, the experimental results in Table 4 show that the multi-task MolGIN (MT-MolGIN) achieved significant improvement on Log$D_{7.4}$, PPB, and LD50 datasets, but decreased a bit on LogS dataset. This largely may because there are correlations among the acute oral toxicity, octanol-water distribution coefficient, and plasma protein binding of chemicals.

Then, we compared MolGIN to the baseline models in terms of AUC on three ADMET datasets of classification. Table 5 shows the comparison results of these classification tasks, in which we can observe that, like on ADMET regression tasks, MolGIN significantly outperforms not only RF and FDNN models but also the GCN and GIN models on all classification tasks in this work.

Among these ADMET classification tasks, the mutagenicity predictions by MolGIN totally benefited the most comparing with the baseline models. The single-task MolGIN improved 2.57%, 17.09%, 8.13%, and 2.18% comparing with RF, FDNN, GCN, and GIN on the test set of the Ames dataset, respectively. Although Pgp-inhibitor prediction was the least beneficial among all 7 ADMET tasks using

MolGIN, the single-task MolGIN improved 1.97%, 4.14%, 3.79%, and 2.99% comparing with RF, FDNN, GCN, and GIN. For multi-task learning, MolGIN significantly benefited from multi-task learning on the Tox21 dataset, since there may be intrinsic relationships among the 12 subtasks in the Tox21 dataset as reported in [53]. In contrast, MT-MolGIN did not significantly improve on the Pgp-inhibitor dataset, even slightly declined on the Ames dataset than ST-MolGIN. This largely may because there is no strong correlation between the Pgp-inhibitor and Ames datasets. However, the multi-task models obtained a small improvement on the Pgp-inhibitor dataset, which may be because the multi-task data fusion properly added noise into the small Pgp-inhibitor dataset, thereby enhancing the robustness of the models.

Notably, the proposed MolGIN (single task) improved significantly when directly comparing with GCN on all ADMET datasets, by 2.65% to 39.42% in external tests and by 2.30% to 39.44% in 5-fold cross-validation, as shown in Table 6. This indicates that MolGIN's molecular modeling capabilities are much better than that of GCN, though GCN was the most commonly used GNN in molecular modeling for the last three years.

TABLE 6. Direct comparison with GCN and MolGIN (single-task).

Dataset	Cross-validation			Test		
	GCN	MolGIN	Improvement	GCN	MolGIN	Improvement
LD50	0.628	0.611	2.71%	0.621	0.584	5.96%
Log$D_{7.4}$	0.675	0.590%	12.59	0.654	0.581	11.16%
LogS	1.009	0.611	39.44%	0.959	0.581	39.42%
PPB	16.315	15.145	7.17%	15.238	14.150	7.14%
Pgp-inhibitor	0.886	0.912	2.93%	0.896	0.930	3.79%
Ames	0.876	0.915	4.45%	0.849	0.918	8.13%
Tox21	0.827	0.846	2.30%	0.829	0.851	2.65%

Note: The evaluation metric for the LD50, Log$D_{7.4}$, LogS, and PPB datasets was RMSE, the smaller the better. The evaluation metric for the Pgp-inhibitor, Ames, and Tox21 datasets was AUC, the higher the better.

More importantly, MolGIN (single-task) does achieve better performance in direct comparison with GIN on all ADMET datasets, improving by 2.23% to 10.34% on external tests and by 1.44% to 11.41% on 5-fold cross-validation, as shown in Table 7. This shows that MolGIN has better molecular modeling capabilities than that of GIN, for MolGIN exploits the bond characteristics and the strength information of different atom-atom interactions based on GIN.

In conclusion, MolGIN significantly outperforms all baseline models on all ADMET datasets in this work, including the recent outstanding GNN models: GCN and GIN. This demonstrates the effectiveness and advantage of MolGIN.

In addition to RMSE and AUC, which are the most widely used evaluation metrics in molecular property prediction, we also used accuracy (ACC) for classification datasets and square correlation coefficients (R^2) for regression datasets to evaluate the performance of the proposed model and baseline models. Figs. 5 and 6 show the ACC and R^2 scores in external tests of the comparative models on the classification and regression datasets. Comparison results in Figs. 5 and 6 show that MolGIN outperformed all baseline models on 5/7 datasets, although it achieved lower R^2 score than RF on Log$D_{7.4}$ and LogS datasets. Moreover, MolGIN is the only one in the comparative models that satisfied $R^2 > 0.6$ on all four regression datasets. The evaluation of $R^2 > 0.6$ was suggested as a strict criterion for a practical regression-like QSAR/QSPR model [50], [70].

TABLE 7. Direct comparison with GIN and MolGIN (single-task).

Dataset	Cross-validation			Test		
	GIN	MolGIN	Improvement	GIN	MolGIN	Improvement
LD50	0.625	0.611	2.24%	0.599	0.584	2.50%
Log$D_{7.4}$	0.666	0.590	11.41%	0.648	0.581	10.34%
LogS	0.664	0.611	7.98%	0.646	0.581	10.06%
PPB	15.651	15.145	3.23%	14.683	14.150	3.63%
Pgp-inhibitor	0.899	0.912	1.45%	0.903	0.930	2.99%
Ames	0.891	0.915	2.69%	0.898	0.918	2.23%
Tox21	0.834	0.846	1.44%	0.831	0.851	2.41%

Note: The evaluation metric for the LD50, Log$D_{7.4}$, LogS, and PPB datasets was RMSE, the smaller the better. The evaluation metric for the Pgp-inhibitor, Ames, and Tox21 datasets was AUC, the higher the better.

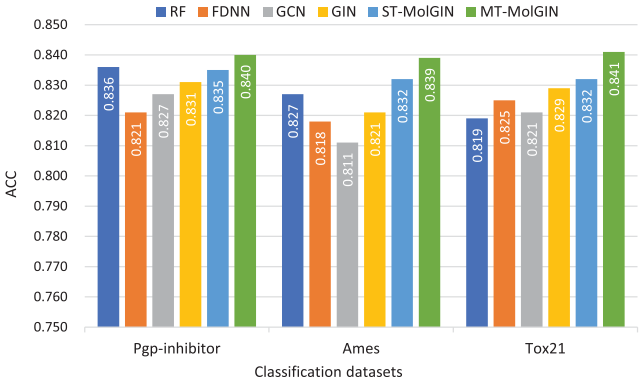

FIGURE 5. Comparison ACC results in external tests of different models on the classification datasets.

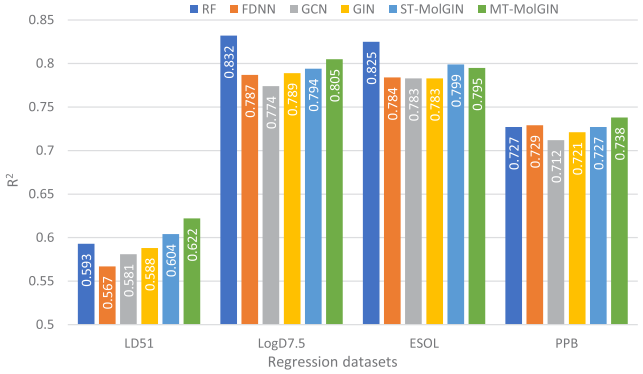

FIGURE 6. Comparison R^2 results in external tests of different models on the regression datasets.

2) COMPARISON WITH STATE-OF-THE-ART RESULTS

In this section, we compare MolGIN with state-of-the-art models including a variety of shallow machine learning models and deep-learning models on the test sets of the

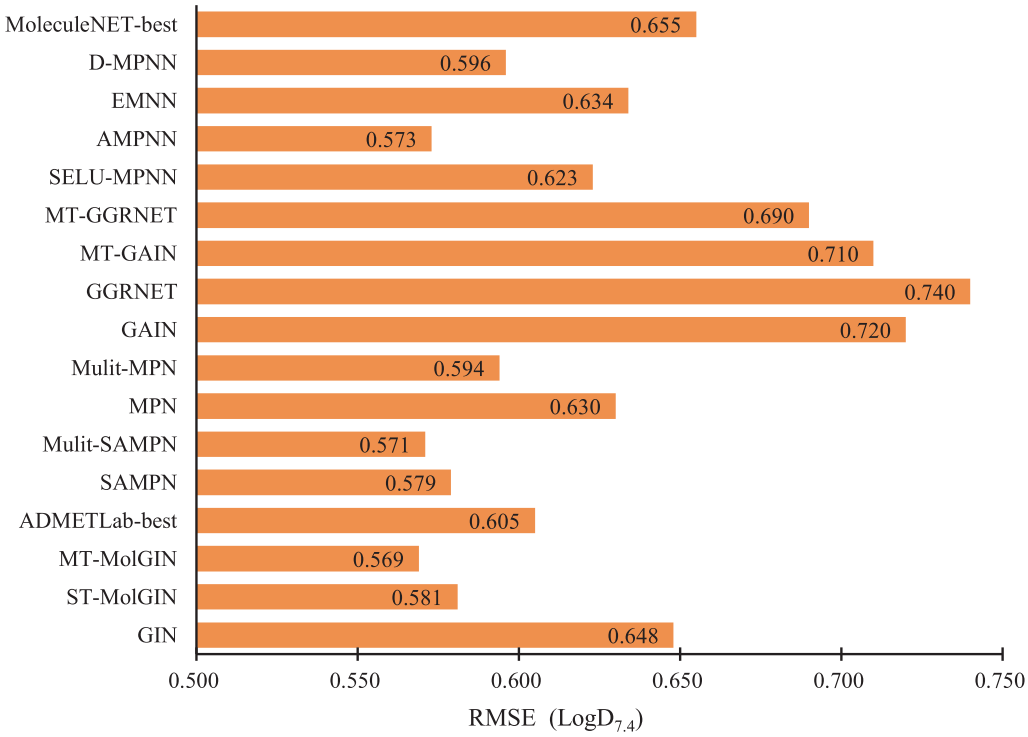

FIGURE 7. Comparison results of different methods on the Log$D_{7.4}$ dataset. The MoleculeNet-best and ADMETLab-best represent the best results of various models in MoleculeNet and ADMETLab platforms on the Log$D_{7.4}$ dataset, respectively. Our MT-MolGIN outperformed Multi-SAMPN (a multi-task SAMPN model, detailed in Table 3), leading state-of-the-art on the Log$D_{7.4}$ dataset.

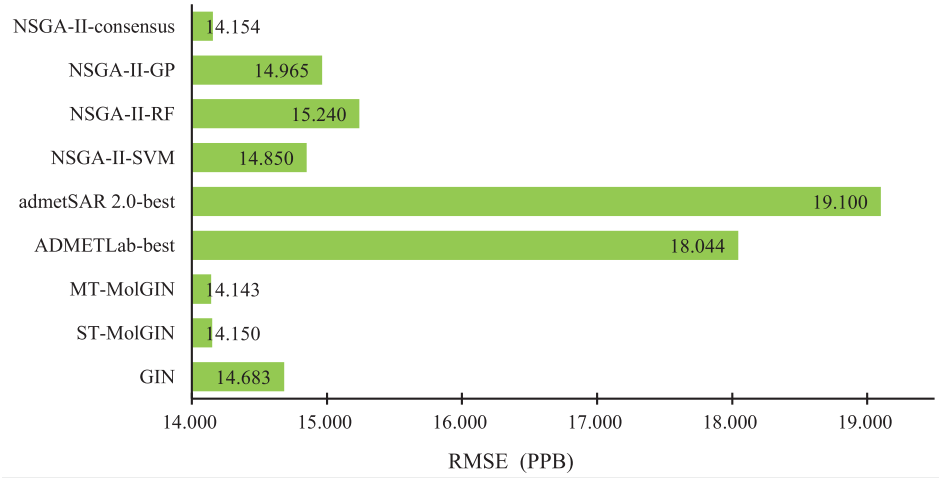

FIGURE 8. Comparison results of different methods on the PPB dataset. The admetSAR 2.0-best and ADMETLab-best represent the best result of various models in the admetSAR 2.0 and ADMETLab platforms on the PPB dataset, respectively. Our MT-MolGIN and ST-MolGIN won the first and second among all comparative models, respectively.

corresponding ADMET datasets. The comparison results on the Log$D_{7.4}$, PPB, LD50, ESOL, Ames, Pgp-inhibitor, and Tox21 datasets are shown in Figs.7-13, respectively.

From Figs.7-8 and 12-13, we observe that MolGIN outperformed all other models on the Log$D_{7.4}$, PPB, Pgp-inhibitor, and Tox21 datasets. MolGIN produced 4 new state-of-the-art results on 7 benchmark ADMET datasets, winning the most in all comparative models. Besides, we also observe from Figs.9 and 11 that MolGIN, respectively, ranked the second among all comparative models on the LD50 and Ames datasets, and ranked the fifth out of 21 comparative models on the ESOL dataset, as shown in Fig.10.

In summary, the overall performance of MolGIN is significantly better than that of other state-of-the-art models.

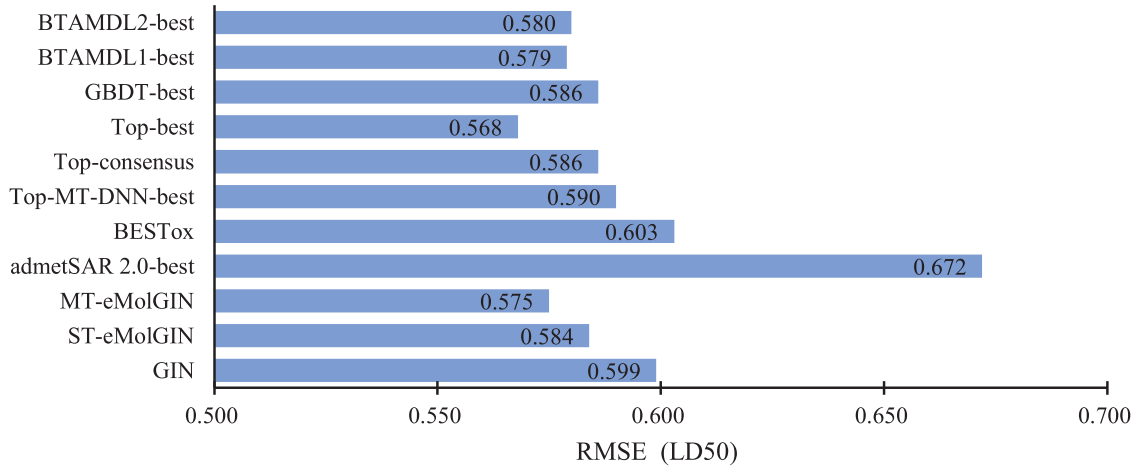

FIGURE 9. Comparison results of different methods on the LD50 dataset. The BTAMDL1-best, BTAMDL2-best, GBDT-best, Top-best, Top-MT-DNN-best, and admetSAR 2.0-best represent the best results of BTAMDL1, BTAMDL2, GBDT, Top, Top-MT-DNN, and admetSAR 2.0 models/platforms using various types of input data on the LD50 dataset, respectively. Our MT-MolGIN won the second among 11 comparative models.

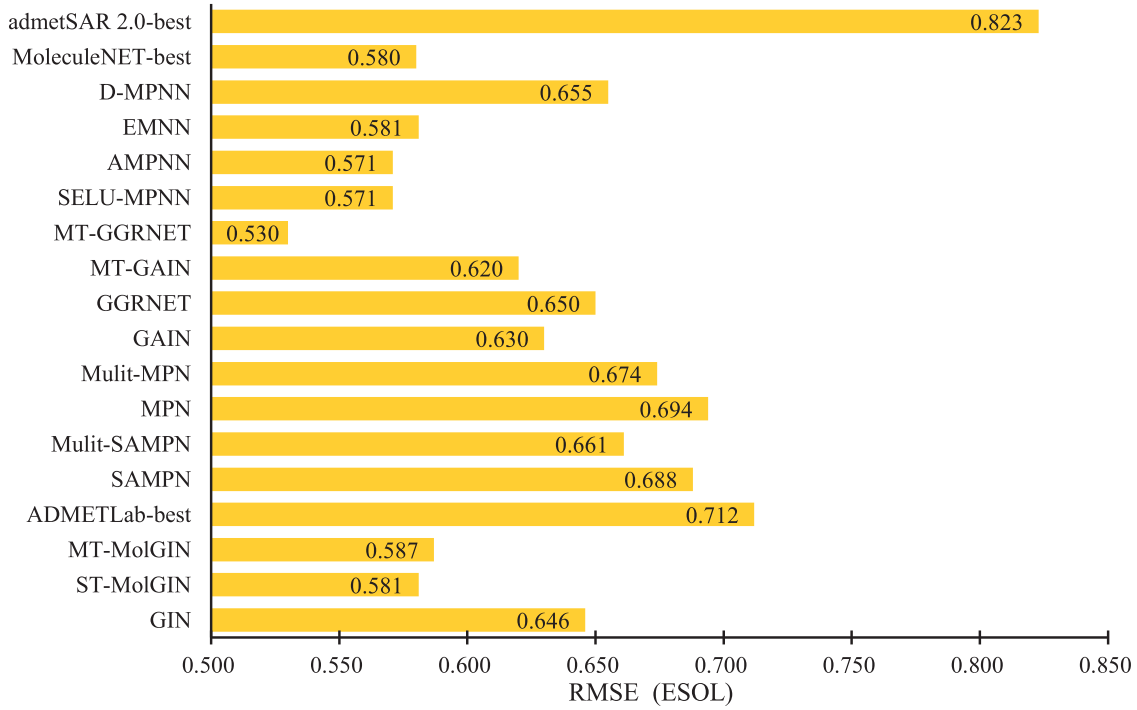

FIGURE 10. Comparison results of different methods on the ESOL dataset. The admetSAR 2.0-best, MoleculeNet-best, and ADMETLab-best represent the best results of various models in the admetSAR 2.0, MoleculeNet, and ADMETLab platforms on the ESOL dataset, respectively. Our ST-MolGIN won the fifth among 18 comparative models.

MolGIN thus is very powerful alternative framework for molecular ADMET properties prediction.

3) ABLATION STUDY

To explore the effect of individual components in MolGIN, we conducted an ablation study on the ADMET classification datasets. Specifically, we separately removed the bond feature concatenation and neighborhood weight adjustment from the MolGIN architecture respectively, while keeping the other components unchanged. The model that removed the bond feature concatenation from MolGIN named MolGIN-bond, while the model that removed the neighborhood weight adjustment named MolGIN-interact. The ablation study results are showed in Table 8.

From Table 8, we can observe that the MolGIN performance is significantly affected by the bond feature concatenation and neighborhood weight adjustment. When the

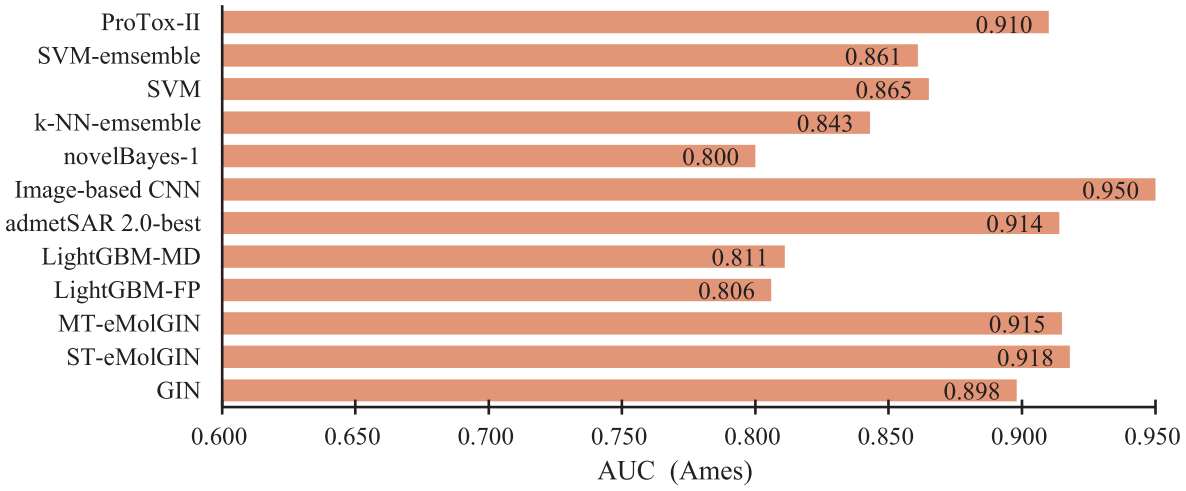

FIGURE 11. Comparison results of different methods on the Ames dataset. The admetSAR 2.0-best represents the best result of various models in the admetSAR 2.0 platform on the Ames dataset. Our ST-MolGIN and MT-MolGIN won the second and third among 12 comparative models, respectively.

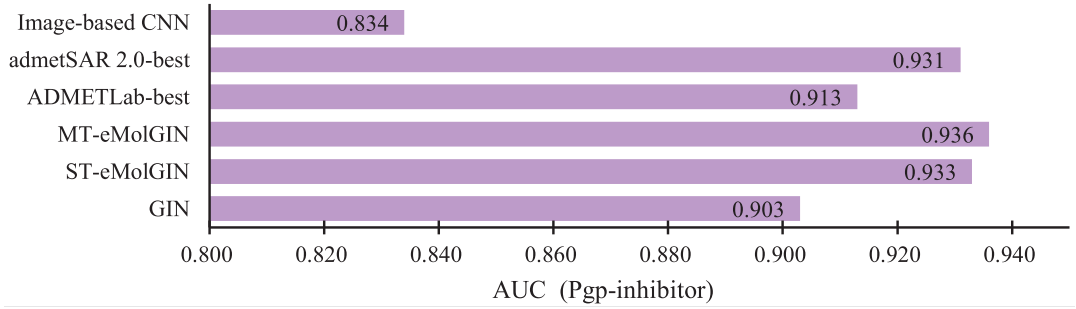

FIGURE 12. Comparison results of different methods on the Pgp-inhibitor dataset. The admetSAR 2.0-best and ADMETLab-best represent the best results of various models in the admetSAR 2.0 and ADMETLab platforms on the Pgp-inhibitor dataset, respectively. Our MT-MolGIN and ST-MolGIN won the first and second among all comparative models, respectively.

TABLE 8. Ablation study results (AUC) on ADMET classification datasets.

	GIN	MolGIN	MolGIN-interact	MolGIN-bond
Pgp-inhibitor	0.903± 0.012	0.930± 0.015	0.918± 0.025	0.909± 0.031
Ames	0.898± 0.008	0.918± 0.012	0.909± 0.018	0.903± 0.026
Tox21	0.831±0.013	0.851± 0.015	0.842± 0.015	0.835± 0.027

Note: All the results in this Table are obtained by the corresponding single-task models.

bond feature concatenation was removed, in terms of AUC on the Pgp-inhibitor, Ames, and Tox21 datasets, MolGIN respectively reduced by 1.61%, 0.98%, and 1.06% but still respectively improved by 1.66%, 1.22%, and 1.32% compared with GIN. When the neighborhood weight adjustment was removed, in terms of AUC on the Pgp-inhibitor, Ames, and Tox21 datasets, MolGIN respectively reduced by 2.57%, 1.63%, and 1.88% but still respectively improved by 0.66%, 0.56%, and 0.48% compared with GIN.

4) DISCUSSION

Our experimental comparison results on 7 ADMET datasets showed that GIN significantly outperforms GCN on not only all 5-fold cross-validations but also all external tests. It is largely because GIN has the more expressive power to capture different graph structures than GCN. This finding is consistent with the finding of [33] that GIN has the most expressive among the class of GNNs including GCN not only empirically but also theoretically. Xu *et al.* [33] has also theoretically proved that the GIN architecture is as powerful as the Weisfeiler-Lehman graph isomorphism test [44].

More importantly, the experimental comparison results showed that in terms of RMSE and AUC, our enhanced GIN, i.e. MolGIN, significantly outperformed all baseline models including GIN on all datasets, and achieved the best performance on 4 datasets and the second-best performance on 2 datasets when comparing to 39 state-of-the-art models. MolGIN overall won the best among all comparative models. These results demonstrate that MolGIN has powerful molecular modeling and ADMET properties prediction capabilities based on the GIN model. This is largely because MolGIN not only inherits the strong discriminative/representational power of GIN, but also exploits the molecular characteristic information in molecular modeling.

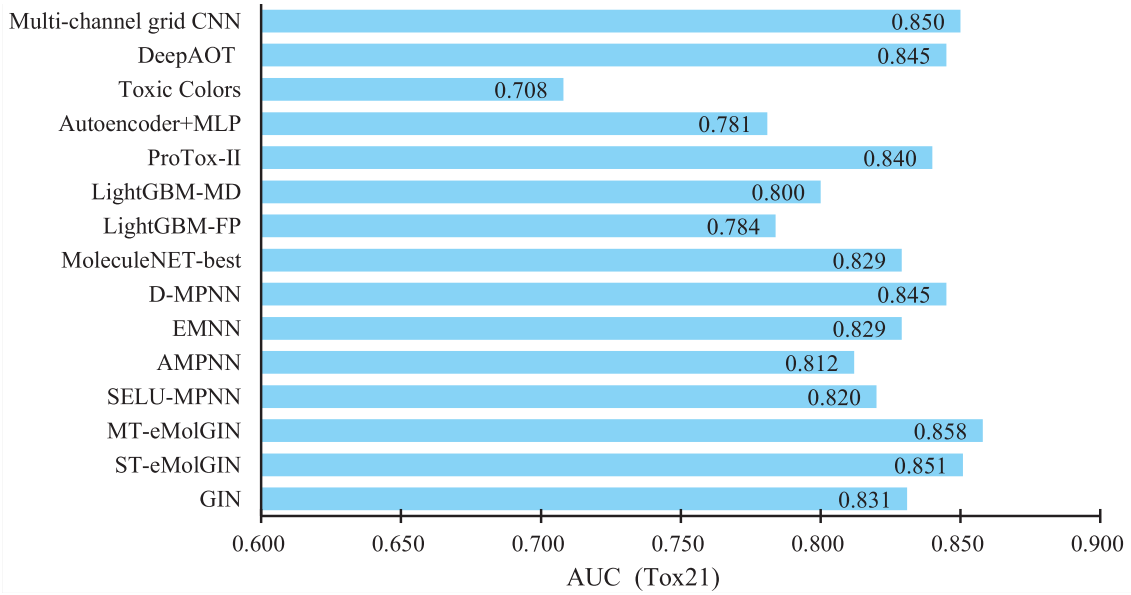

FIGURE 13. Comparison results of different methods on the Tox21 dataset. The MoleculeNET-best represents the best result of various models in the MoleculeNET platform on the Tox21 dataset. Our MT-MolGIN and ST-MolGIN won the first and second among 15 comparative models, respectively.

Especially, in direct comparison with GIN, MolGIN achieved significantly better performance than GIN across all datasets, improving performance from 1.44% to 11.41% on 5-fold cross-validations and from 2.23% to 10.34% on external tests. This reveals the advantages of our MolGIN exploiting the molecular bond features and differences influence between the central atom and its neighbors base on GIN. Furthermore, when we removed the bond feature concatenation or neighborhood weight adjustment from the MolGIN architecture, the AUC of MolGIN dropped but still improved compared with GIN, as shown in Table 8. Results above strongly supports that the molecular characteristics, i.e. bond features and the differences of the atom-atom interactions, can effectively help the GNN to learn more local structural features of molecules, thereby boosting molecular ADMET properties prediction. The results in Table 8 show that MolGIN removing bond feature concatenation has a greater effect on performance than removing the neighborhood weight adjustment. It reveals that the bond feature is meaningful than the feature of differences influence between the central atom and its neighbors in the MolGIN architecture for ADMET properties prediction. More interestingly, MolGIN simultaneous removing the bond feature and the feature of differences influence strength of atom-atom interactions had a greater impact on model performance than the sum of MolGIN removing only bond information and MolGIN removing only the feature of differences influence between the central atom and its neighbors. It reveals that the feature combination of molecular bond and atom-atom interactions' strength differences is more meaningful than the simple sum of the bond feature and the feature of atom-atom interactions' strength differences for MolGIN.

The evaluation of the models performance in terms of R^2 shown in Fig. 6 demonstrates that MolGIN satisfies the strict evaluation of $R^2 > 0.6$ on all four regression datasets, this strict evaluation is for a practical regression-like QSAR/QSPR model according to the suggestions from Wang *et al.* [50], Tropsha *et al.* [70]. This means that our proposed model can be a practical predictive model and could be helpful for fast estimating ADMET properties of new chemical entities even to the high-throughput screening.

It is noteworthy that the finding that exploiting bond features can improve the molecular properties prediction is consistent with the EMNN model in Withnall's study [56]. However, our MolGIN significantly outperforms EMNN on 2/3 datasets, as shown in Figs.7 and 13, and achieved comparable performance to EMNN on 1/3 datasets in this work, as shown in Fig.10. This may be because MolGIN uses better network architecture, the special information of atom-atom interactions, and different way of exploiting bond features comparing to EMNN.

Furthermore, the finding in this work will facilitate further research on the boosting molecular modeling and its application in bioinformatics and cheminformatics.

V. CONCLUSION

GIN has the maximum discriminative/representational power from different graph structures and quantifies generalization ability than other previous GNN architectures [33], thereby being quickly attracted extensive attention in GNN research community and used in graph data problems. However, GIN has not fully exploited the characteristic information of the molecular graph, failing to fully exploit its molecular modeling capability.

To tackle the aforementioned problem, based on GIN, this work mainly exploited molecular characteristics, i.e. bond features and differences in the interaction strength between the central atom and its neighbors, to enhance the molecular modeling capability and to boost ADMET properties prediction. We hence proposed an enhanced GIN, called MolGIN, for ADMET properties prediction. MolGIN was implemented by concatenating feature vectors of the atom neighbors and the bonds connecting with the atom and adding a control gate unit to adjust the neighborhood weights in the information aggregator.

We carried out extensive performance evaluation by comparing our method with not only GIN and other 3 baseline models but also a collection of state-of-the-art results on 7 different ADMET datasets. Experimental results in terms of RMSE and AUC show that MolGIN significantly outperformed GIN and other baseline models in effectiveness measure, and achieved comparable or superior performance to state-of-the-art models on the corresponding tasks. MolGIN provides a potential way for large-scale chemical ADMET properties prediction. The proposed method is also promising for other molecular modeling problems in bioinformatics and cheminformatics.

However, similar to other deep learning methods, the MolGIN model still lacks interpretability. It is difficult to clearly recognize which substructures of the molecule play important roles in the prediction task. We will study to improve the model interpretability in the future.

REFERENCES

[1] K. R. Przybylak, J. C. Madden, E. Covey-Crump, L. Gibson, C. Barber, M. Patel, and M. T. D. Cronin, "Characterisation of data resources for in silico modelling: Benchmark datasets for ADME properties," *Expert Opinion Drug Metabolism Toxicol.*, vol. 14, no. 2, pp. 169–181, Feb. 2018.

[2] C.-Y. Jia, J.-Y. Li, G.-F. Hao, and G.-F. Yang, "A drug-likeness toolbox facilitates ADMET study in drug discovery," *Drug Discovery Today*, vol. 25, no. 1, pp. 248–258, Jan. 2020.

[3] S. M. Paul, D. S. Mytelka, C. T. Dunwiddie, C. C. Persinger, B. H. Munos, S. R. Lindborg, and A. L. Schacht, "How to improve R&D productivity: The pharmaceutical industry's grand challenge," *Nature Rev. Drug Discovery*, vol. 9, no. 3, pp. 203–214, 2010.

[4] S. Morgan, P. Grootendorst, J. Lexchin, C. Cunningham, and D. Greyson, "The cost of drug development: A systematic review," *Health Policy*, vol. 100, no. 1, pp. 4–17, Apr. 2011.

[5] E. N. Feinberg, R. Sheridan, E. Joshi, V. S. Pande, and A. C. Cheng, "Step change improvement in ADMET prediction with PotentialNet deep featurization," 2019, *arXiv:1903.11789*. [Online]. Available: http://arxiv.org/abs/1903.11789

[6] *2015 Biopharmaceutical Research Industry Profile*, Pharmaceutical Res. Manuf. Amer., Washington, DC, USA, 2015.

[7] N. Fleming, "How artificial intelligence is changing drug discovery," *Nature*, vol. 557, no. 7707, pp. S55–S57, May 2018.

[8] A. Sazonovas, P. Japertas, and R. Didziapetris, "Estimation of reliability of predictions and model applicability domain evaluation in the analysis of acute toxicity (LD50)," *SAR QSAR Environ. Res.*, vol. 21, nos. 1–2, pp. 127–148, Jan. 2010.

[9] X. Li and D. Fourches, "Inductive transfer learning for molecular activity prediction: Next-gen QSAR models with MolPMoFiT," *J. Cheminformatics*, vol. 12, no. 1, pp. 1–5, Dec. 2020.

[10] Y. Peng, Z. Zhang, Q. Jiang, J. Guan, and S. Zhou, "TOP: Towards better toxicity prediction by deep molecular representation learning," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Nov. 2019, pp. 318–325.

[11] A. Lusci, G. Pollastri, and P. Baldi, "Deep architectures and deep learning in chemoinformatics: The prediction of aqueous solubility for drug-like molecules," *J. Chem. Inf. Model.*, vol. 53, no. 7, pp. 1563–1575, Jul. 2013.

[12] J. Ma, R. P. Sheridan, A. Liaw, G. E. Dahl, and V. Svetnik, "Deep neural nets as a method for quantitative structure–activity relationships," *J. Chem. Inf. Model.*, vol. 55, no. 2, pp. 263–274, 2015.

[13] T. Lei, Y. Li, Y. Song, D. Li, H. Sun, and T. Hou, "ADMET evaluation in drug discovery: 15. Accurate prediction of rat oral acute toxicity using relevance vector machine and consensus modeling," *J. Cheminformatics*, vol. 8, no. 1, p. 6, Dec. 2016.

[14] C. A. S. Bergström and P. Larsson, "Computational prediction of drug solubility in water-based systems: Qualitative and quantitative approaches used in the current drug discovery and development setting," *Int. J. Pharmaceutics*, vol. 540, nos. 1–2, pp. 185–193, Apr. 2018.

[15] L. L. Ferreira and A. D. Andricopulo, "ADMET modeling approaches in drug discovery," *Drug Discovery Today*, vol. 24, no. 5, pp. 1157–1165, 2019.

[16] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[17] K. Liu, X. Sun, L. Jia, J. Ma, H. Xing, J. Wu, H. Gao, Y. Sun, F. Boulnois, and J. Fan, "Chemi-net: A molecular graph convolutional network for accurate drug property prediction," *Int. J. Mol. Sci.*, vol. 20, no. 14, p. 3389, Jul. 2019.

[18] G. B. Goh, N. O. Hodas, and A. Vishnu, "Deep learning for computational chemistry," *J. Comput. Chem.*, vol. 38, no. 16, pp. 1291–1307, Jun. 2017.

[19] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, and V. Pande, "MoleculeNet: A benchmark for molecular machine learning," *Chem. Sci.*, vol. 9, no. 2, pp. 513–530, 2018.

[20] A. S. Rifaioglu, H. Atas, M. J. Martin, R. Cetin-Atalay, V. Atalay, and T. Doğan, "Recent applications of deep learning and machine intelligence on in silico drug discovery: Methods, tools and databases," *Briefings Bioinf.*, vol. 20, no. 5, pp. 1878–1912, Sep. 2019.

[21] H. Sun, P. Shah, K. Nguyen, K. R. Yu, E. Kerns, M. Kabir, Y. Wang, and X. Xu, "Predictive models of aqueous solubility of organic compounds built on a large dataset of high integrity," *Bioorganic Med. Chem.*, vol. 27, no. 14, pp. 3110–3114, Jul. 2019.

[22] F. Ghasemi, A. Mehridehnavi, A. Pérez-Garrido, and H. Pérez-Sánchez, "Neural network and deep-learning algorithms used in QSAR studies: Merits and drawbacks," *Drug Discovery Today*, vol. 23, no. 10, pp. 1784–1790, Oct. 2018.

[23] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams, "Convolutional networks on graphs for learning molecular fingerprints," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2224–2232.

[24] S. Kearnes, K. McCloskey, M. Berndl, V. Pande, and P. Riley, "Molecular graph convolutions: Moving beyond fingerprints," *J. Comput.-Aided Mol. Des.*, vol. 30, no. 8, pp. 595–608, Aug. 2016.

[25] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, "SchNet—A deep learning architecture for molecules and materials," *J. Chem. Phys.*, vol. 148, no. 24, Jun. 2018, Art. no. 241722.

[26] M. Sun, S. Zhao, C. Gilvary, O. Elemento, J. Zhou, and F. Wang, "Graph convolutional networks for computational drug development and discovery," *Briefings Bioinf.*, vol. 21, no. 3, pp. 919–935, 2020.

[27] C. Chen, W. Ye, Y. Zuo, C. Zheng, and S. P. Ong, "Graph networks as a universal machine learning framework for molecules and crystals," *Chem. Mater.*, vol. 31, no. 9, pp. 3564–3572, May 2019.

[28] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, vol. 70, 2017, pp. 1263–1272.

[29] C. W. Coley, R. Barzilay, W. H. Green, T. S. Jaakkola, and K. F. Jensen, "Convolutional embedding of attributed molecular graphs for physical property prediction," *J. Chem. Inf. Model.*, vol. 57, no. 8, pp. 1757–1772, Aug. 2017.

[30] K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea, A. Palmer, V. Settels, T. Jaakkola, K. Jensen, and R. Barzilay, "Analyzing learned molecular representations for property prediction," *J. Chem. Inf. Model.*, vol. 59, no. 8, pp. 3370–3388, Aug. 2019.

[31] F. Montanari, L. Kuhnke, A. Ter Laak, and D.-A. Clevert, "Modeling physico-chemical ADMET endpoints with multitask graph convolutional networks," *Molecules*, vol. 25, no. 1, p. 44, Dec. 2019.

[32] E. N. Feinberg, E. Joshi, V. S. Pande, and A. C. Cheng, "Improvement in ADMET prediction with multitask deep featurization," *J. Med. Chem.*, vol. 63, no. 16, pp. 8835–8848, Aug. 2020.

[33] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?" in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019, pp. 1–17.
[34] F. Capela, V. Nouchi, R. Van Deursen, I. V. Tetko, and G. Godin, "Multitask learning on graph neural networks applied to molecular property predictions," 2019, *arXiv:1910.13124*. [Online]. Available: http://arxiv.org/abs/1910.13124
[35] T. Nguyen, H. Le, and S. Venkatesh, "GraphDTA: Prediction of drug-target binding affinity using graph convolutional networks," *bioRxiv*, Jan. 2019, Art. no. 684662, doi: 10.1101/684662.
[36] R. L. Murphy, B. Srinivasan, V. Rao, and B. Ribeiro, "Relational pooling for graph representations," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019, pp. 4663–4673.
[37] F. Errica, M. Podda, D. Bacciu, and A. Micheli, "A fair comparison of graph neural networks for graph classification," in *Proc. 8th Int. Conf. Learn. Represent. (ICLR)*, Addis Ababa, Ethiopia, 2020, pp. 1–14.
[38] B.-H. Kim and J. C. Ye, "Understanding graph isomorphism network for rs-fMRI functional connectivity analysis," 2020, *arXiv:2001.03690*. [Online]. Available: https://arxiv.org/abs/2001.03690
[39] T. Nguyen, T. Nguyen, and D.-H. Le, "Graph convolutional networks for drug response prediction," *bioRxiv*, Jan. 2020, doi: 10.1101/2020.04.07.030908.
[40] S. Zhang, H. Tong, J. Xu, and R. Maciejewski, "Graph convolutional networks: A comprehensive review," *Comput. Social Netw.*, vol. 6, no. 1, pp. 1–23, Dec. 2019.
[41] S. Sukhbaatar, A. Szlam, and R. Fergus, "Learning multiagent communication with backpropagation," in *Proc. 29th Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2016, pp. 2252–2260.
[42] T. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. 5th Int. Conf. Learn. Represent. (ICLR)*, Toulon, France, 2017, pp. 1–14.
[43] B. Weisfeiler and A. Lehman, "A reduction of a graph to a canonical form and an algebra arising during this reduction," *Nauchno-Technicheskaya Informatsia*, vol. 2, no. 9, pp. 12–16, 1968.
[44] L. Babai and L. Kucera, "Canonical labelling of graphs in linear average time," in *Proc. 20th Annu. Symp. Found. Comput. Sci.*, Oct. 1979, pp. 39–46.
[45] V. Prakash Dwivedi, C. K. Joshi, T. Laurent, Y. Bengio, and X. Bresson, "Benchmarking graph neural networks," 2020, *arXiv:2003.00982*. [Online]. Available: http://arxiv.org/abs/2003.00982
[46] F. A. Gers, N. N. Schraudolph, and J. Schmidhuber, "Learning precise timing with LSTM recurrent networks," *J. Mach. Learn. Res.*, vol. 3, no. 1, pp. 115–143, 2003.
[47] B. Ramsundar, S. Kearnes, P. Riley, D. Webster, D. Konerding, and V. Pande, "Massively multitask networks for drug discovery," 2015, *arXiv:1502.02072*. [Online]. Available: http://arxiv.org/abs/1502.02072
[48] J. Wenzel, H. Matter, and F. Schmidt, "Predictive multitask deep neural network models for ADME-tox properties: Learning from large data sets," *J. Chem. Inf. Model.*, vol. 59, no. 3, pp. 1253–1268, Mar. 2019.
[49] H. Zhu, T. M. Martin, L. Ye, A. Sedykh, D. M. Young, and A. Tropsha, "Quantitative structure-activity relationship modeling of rat acute toxicity by oral exposure," *Chem. Res. toxicology*, vol. 22, no. 12, pp. 1913–1921, 2009.
[50] N.-N. Wang, Z.-K. Deng, C. Huang, J. Dong, M.-F. Zhu, Z.-J. Yao, A. F. Chen, A.-P. Lu, Q. Mi, and D.-S. Cao, "ADME properties evaluation in drug discovery: Prediction of plasma protein binding using NSGA-II combining PLS and consensus modeling," *Chemometric Intell. Lab. Syst.*, vol. 170, pp. 84–95, Nov. 2017.
[51] J. Dong, N.-N. Wang, Z.-J. Yao, L. Zhang, Y. Cheng, D. Ouyang, A.-P. Lu, and D.-S. Cao, "ADMETlab: A platform for systematic ADMET evaluation based on a comprehensively collected ADMET database," *J. Cheminform.*, vol. 10, no. 1, p. 29, Dec. 2018.
[52] C. Xu, F. Cheng, L. Chen, Z. Du, W. Li, G. Liu, P. W. Lee, and Y. Tang, "In silico prediction of chemical AMES mutagenicity," *J. Chem. Inf. Model.*, vol. 52, no. 11, pp. 2840–2847, Nov. 2012.
[53] A. Mayr, G. Klambauer, T. Unterthiner, and S. Hochreiter, "DeepTox: Toxicity prediction using deep learning," *Frontiers Environ. Sci.*, vol. 3, p. 80, Feb. 2016.
[54] Y. Zhou, S. Cahya, S. A. Combs, C. A. Nicolaou, J. Wang, P. V. Desai, and J. Shen, "Exploring tunable hyperparameters for deep neural networks with industrial ADME data sets," *J. Chem. Inf. Model.*, vol. 59, no. 3, pp. 1005–1016, Mar. 2019.
[55] H. Yang, C. Lou, L. Sun, J. Li, Y. Cai, Z. Wang, W. Li, G. Liu, and Y. Tang, "AdmetSAR 2.0: Web-service for prediction and optimization of chemical ADMET properties," *Bioinform.*, vol. 35, no. 6, pp. 1067–1069, Mar. 2019.
[56] M. Withnall, E. Lindelöf, O. Engkvist, and H. Chen, "Building attention and edge message passing neural networks for bioactivity and physical–chemical property prediction," *J. Cheminform.*, vol. 12, no. 1, p. 1, Dec. 2020.
[57] B. Tang, S. T. Kramer, M. Fang, Y. Qiu, Z. Wu, and D. Xu, "A self-attention based message passing neural network for predicting molecular lipophilicity and aqueous solubility," *J. Cheminform.*, vol. 12, no. 1, Dec. 2020.
[58] J. Zhang, D. Mucs, U. Norinder, and F. Svensson, "LightGBM: An effective and scalable algorithm for prediction of chemical toxicity—Application to the Tox21 and mutagenicity data sets," *J. Chem. Inf. Model.*, vol. 59, no. 10, pp. 4150–4158, 2019.
[59] P. Banerjee, A. O. Eckert, A. K. Schrey, and R. Preissner, "ProTox-II: A Webserver for the prediction of toxicity of chemicals," *Nucleic Acids Res.*, vol. 46, no. W1, pp. W257–W263, Jul. 2018.
[60] T. Shi, Y. Yang, S. Huang, L. Chen, Z. Kuang, Y. Heng, and H. Mei, "Molecular image-based convolutional neural network for the prediction of ADMET properties," *Chemometric Intell. Lab. Syst.*, vol. 194, Nov. 2019, Art. no. 103853.
[61] H. Zhang, Y.-L. Kang, Y.-Y. Zhu, K.-X. Zhao, J.-Y. Liang, L. Ding, T.-G. Zhang, and J. Zhang, "Novel Naïve Bayes classification models for predicting the chemical AMES mutagenicity," *Toxicology Vitro*, vol. 41, pp. 56–63, Jun. 2017.
[62] G. Cerruela García, N. García-Pedrajas, I. Luque Ruiz, and M. Á. Gómez-Nieto, "An ensemble approach for in silico prediction of AMES mutagenicity," *J. Math. Chem.*, vol. 56, no. 7, pp. 2085–2098, Aug. 2018.
[63] M. Galushka, F. Browne, M. Mulvenna, R. Bond, and G. Lightbody, "Toxicity prediction using pre-trained autoencoder," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2018, pp. 299–304.
[64] M. Fernandez, F. Ban, G. Woo, M. Hsing, T. Yamazaki, E. LeBlanc, P. S. Rennie, W. J. Welch, and A. Cherkasov, "Toxic colors: The use of deep learning for predicting toxicity of compounds merely from their graphic images," *J. Chem. Inf. Model.*, vol. 58, no. 8, pp. 1533–1543, Aug. 2018.
[65] Y. Xu, J. Pei, and L. Lai, "Deep learning based regression and multiclass models for acute oral toxicity prediction with automatic chemical feature extraction," *J. Chem. Inf. Model.*, vol. 57, no. 11, pp. 2672–2685, Nov. 2017.
[66] Q. Yuan, Z. Wei, X. Guan, M. Jiang, S. Wang, S. Zhang, and Z. Li, "Toxicity prediction method based on multi-channel convolutional neural network," *Molecules*, vol. 24, no. 18, p. 3383, Sep. 2019.
[67] J. Chen, H.-H. Cheong, and S. W. I. Siu, "Bestox: A convolutional neural network regression model based on binary-encoded smiles for acute oral toxicity prediction of chemical compounds," in *Proc. Int. Conf. Algorithms Comput. Biol.* Cham, Switzerland: Springer, 2020, pp. 155–166.
[68] K. Wu and G.-W. Wei, "Quantitative toxicity prediction using topology based multitask deep neural networks," *J. Chem. Inf. Model.*, vol. 58, no. 2, pp. 520–531, Feb. 2018.
[69] J. Jiang, R. Wang, M. Wang, K. Gao, D. D. Nguyen, and G.-W. Wei, "Boosting tree-assisted multitask deep learning for small scientific datasets," *J. Chem. Inf. Model.*, vol. 60, no. 3, pp. 1235–1244, Mar. 2020.
[70] A. Tropsha, P. Gramatica, and V. Gombar, "The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models," *QSAR Combinat. Sci.*, vol. 22, no. 1, pp. 69–77, Apr. 2003.

YUZHONG PENG was born in Guangxi, China, in 1980. He received the B.S. degree in computer science and technology and the M.S. degree in computer application technology from Guangxi Teachers Education University, Nanning, China, in 2003 and 2009, respectively. He is currently pursuing the Ph.D. degree in computer software and theory with Fudan University, Shanghai, China.

From 2009 to 2018, he was a Teacher with Guangxi Teachers Education University. He is also a Professor with Nanning Normal University. He is the author of a book and more than 30 articles. He holds three patents. His research interests include machine learning, data modeling/mining, and biological/medical informatics.

Prof. Peng is a CCF member and an ACM member. He was a recipient of the IEEE BIBM Best Student Paper Award, in 2019.

YANMEI LIN was born in Guangxi, China, in 1988. She received the M.S. degree in molecular biology from Guangxi University, Nanning, China, in 2016. Since 2016, she has been a Teacher with Nanning Normal University. Her research interest includes biological/medical informatics.

XIAO-YUAN JING (Member, IEEE) received the Ph.D. degree in pattern recognition and intelligent system from the Nanjing University of Science and Technology, in 1998.

He is currently a Professor and the Dean of the School of Computer, Guangdong University of Petrochemical Technology. He is also a Professor with the School of Computer, Wuhan University, China. He has published more than 100 papers in various international journals and conference proceedings, such as the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, *Pattern Recognition*, the IEEE TRANSACTIONS ON IMAGE PROCESSING, *Transfusion Clinique et Biologique*, the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, the IEEE TRANSACTIONS ON RELIABILITY, CVPR, AAAI, IJCAI, WWW, ICSE, and ASE. His research interests include pattern recognition, machine learning, artificial intelligence, information security, and software engineering.

HAO ZHANG received the Ph.D. degree in computer science from Fudan University, in 2019. He is currently a Teacher with the Guangdong University of Petrochemical Technology. His research interests include data mining and causal inference.

YIRAN HUANG received the M.Eng. degree in computer science from Guangxi University, China, in 2008, and the Ph.D. degree in computer science from the South China University of Technology, in 2017. He is currently a Lecturer with Guangxi University. His main research interests include bioinformatics and parallel and distributed computing.

GUANG SHENG LUO (Member, IEEE) was born in Huangshi, Hubei, China, in 1982. He received the B.S. degree in computer science and technology from the Huazhong University of Science and Technology, Wuhan, China, in 2005, and the M.S. degree in software engineering from Fudan University, Shanghai, China, in 2015. He is currently pursuing the Ph.D. degree with the Shanghai Key Laboratory of Intelligent Information, Fudan University. He has more than ten years of software development experience. His research interests include machine learning and computational mathematics.