OXFORD

FP-GNN: a versatile deep learning architecture for enhanced molecular property prediction

Hanxuan Cai, Huimin Zhang, Duancheng Zhao, Jingxing Wu and Ling Wang [ID]

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

Abstract

Accurate prediction of molecular properties, such as physicochemical and bioactive properties, as well as ADME/T (absorption, distribution, metabolism, excretion and toxicity) properties, remains a fundamental challenge for molecular design, especially for drug design and discovery. In this study, we advanced a novel deep learning architecture, termed FP-GNN (fingerprints and graph neural networks), which combined and simultaneously learned information from molecular graphs and fingerprints for molecular property prediction. To evaluate the FP-GNN model, we conducted experiments on 13 public datasets, an unbiased LIT-PCBA dataset and 14 phenotypic screening datasets for breast cell lines. Extensive evaluation results showed that compared to advanced deep learning and conventional machine learning algorithms, the FP-GNN algorithm achieved state-of-the-art performance on these datasets. In addition, we analyzed the influence of different molecular fingerprints, and the effects of molecular graphs and molecular fingerprints on the performance of the FP-GNN model. Analysis of the anti-noise ability and interpretation ability also indicated that FP-GNN was competitive in real-world situations. Collectively, FP-GNN algorithm can assist chemists, biologists and pharmacists in predicting and discovering better molecules with desired functions or properties.

Keywords: artificial intelligence, drug design and discovery, machine learning, molecular representation, graph attention networks

Introduction

Accurate predicting molecular properties, including physico-chemical and bioactive properties, as well as ADME/T (absorption, distribution, metabolism, excretion and toxicity) properties, plays a key role in molecular design, especially in drug design and discovery. Quantitative structure–activity (property) relationship (QSAR/QSPR) modeling represents one of the most widely used and well-established computational approach for molecular properties prediction [1, 2]. QSAR/QSPR models are constructed using an empirical, linear or nonlinear function that estimates an activity/property based on a chemical structure, followed by the application of those models to predict and design novel molecules with desired functional properties [3, 4].

Typically, machine learning (ML)-based QSAR/QSPR models have relied heavily on appropriate molecular representations [5]. Currently, molecular representations can be divided into three main categories, including molecular descriptors, fingerprints and graphs. Molecular descriptors and fingerprints were derived from human expert domain knowledge for the comprehensive presentation of the constitutional, physicochemical, topological and structural features of molecules [6–8]. Molecular descriptors and fingerprints can be used as inputs for both conventional ML (e.g. Naive Bayes (NB) [9], Support Vector Machine (SVM) [10], Random Forest (RF) [11], eXtreme Gradient Boosting (XGBoost)

[12]) and deep learning (e.g. deep neural networks) algorithms for QSAR/QSPR modeling tasks. However, molecular descriptor-based models suffer from one major challenge in the era of big data: how to select the most important descriptors (called handcrafted descriptors) related to a property of interest from a large number of predefined and calculable molecular descriptors [13]. This step is not only significant for the performance accuracy of the model, but is also directly related to the interpretability of the model. Recently, the emergence of deep learning (DL) approaches enables the elimination of tiresome expert and domain-wise feature constructions by delegating this task to a neural network that can extract the most valuable traits of the raw input data that are required to model the problem at hand [14, 15]. In contrast, for graph-based molecular representations, the atoms and bonds of a molecule are regarded as nodes and edges, and the aggregated node features are used by DL architectures, such as the Graph Convolutional Network (GCN) [16], Graph Attention Network (GAT) [17], Attentive FP [18], Message Passing Neural Network (MPNN) [19] and Directed MPNN (D-MPNN) [20] for chemical learning tasks. Graph-based DL architectures have become popular and have been successfully employed in molecular property prediction tasks [21–26].

Although graph-based DL architectures are reported to yield state-of-the-art performance for molecular properties prediction

Hanxuan Cai is a graduate student at South China University of Technology. Her current research interests include machine learning and artificial intelligence-aided drug discovery (AIDD).
Huimin Zhang is a graduate student at South China University of Technology. Her research interests include AIDD.
Duancheng Zhao is a graduate student at South China University of Technology. His main research interests include deep learning for molecular generation.
Jingxing Wu is an undergraduate student at South China University of Technology. His research interests include machine learning and bioinformatics.
Ling Wang is an associate professor at South China University of Technology. He received a PhD from the School of Pharmaceutical Sciences at the Sun Yat-Sen University in 2014. His research focuses on computer-aided drug design, AIDD and medicinal chemistry.

tasks, whether graph-based DL models are better than conventional descriptors-based ML models for molecular properties learning tasks remains controversial. The majority of previous studies claimed that graph-based DL models were comparable or superior to conventional descriptors- or fingerprints-based ML models [20, 27, 28], while only few studies presented the opposite conclusion [29, 30]. For example, in 2021, Jiang *et al.* [30] demonstrated that the conventional descriptor-based models (especially RF and XGBoost methods) outperformed the graph-based DL models in terms of prediction accuracy and computational efficiency. Another recent study by Stepisnik and coworkers also reported a similar conclusion [31]. Currently, graph-based DL models still suffer from the potential limitation for insufficient modeling datasets, as it may be difficult for the automatic learning mechanism characteristics of graph neural networks (GNN) to learn robust graph representations from insufficient datasets [32]. In 2020, Rifaioglu *et al.* [28] discovered that graph- and fingerprint-based classifiers exhibited opposite trends when predicting the attributes of protein families. We hypothesize that the information captured by graphs or fingerprints is different, and may be complementary. Thus, the significant local chemical information contained in fingerprints may assist models to achieve superior results.

In this study, we introduce a new DL neural network architecture using graph neural networks with fingerprints information, called FP-GNN (Figure 1), for molecular properties prediction. FP-GNN first operates over a hybrid molecular representation that combines molecular graphs and molecular fingerprints. It not only learns to characterize the local atomic environment by propagating node information from nearby nodes to more distant nodes using the attention mechanism in a task-specific encoding, but also provides strong prior using fixed and complementary molecular fingerprints. We evaluated the FP-GNN model against 13 commonly used public benchmark datasets, LIT-PCBA datasets and 14 phenotypic screening datasets for breast cell lines. Compared to all baseline models, FP-GNN achieved comparable or superior performance on these datasets, illustrating its strong out-of-the-box and state-of-the-art performance when modeling a broad range of molecular properties. Anti-noise ability testing of FP-GNN also revealed its superiority over the Attentive FP, XGBoost and HRGCN+ models, while maintaining high predictive power. In addition, FP-GNN showed interpretability on both graph-based networks and fingerprint-based networks, which may help chemists capture important fragments in designing new molecules/drugs.

Materials and methods
Graph neural networks with attention mechanism

Molecules are natural graph-structured data and we therefore choose the spatial-GNN [33] to compute the information from molecular graphs. Before the data were inputted into the GNN model, we transformed each molecule into an undirected graph $G(V, E)$, where $V = \{x_1, x_2, \ldots, x_n\}$ is the node set representing atoms and E is the edge set representing chemical bonds. The spatial-GNN updates each node by aggregating the information of itself and neighbors according to

$$h'_i = Aggregate_1 \left(h_i, \sum_{j \in N(i)} h_j \right) \quad (1)$$

where h_i is the vector of node i and $N(i)$ means the neighbors of i.

Finally, the model aggregates the total graph to the output according to

$$H = Aggregate_2 \left(\sum_{i \in G} h'_i \right) \quad (2)$$

As shown in Figure 1A, we used the attention mechanism [17] to update the node message. The graph attention mechanism pays attention to the effect of neighbors and computes the attention from node j to node i according to

$$e_{ij} = LeakyRelu \left(a \cdot \left[W_1 h_i \| W_1 h_j \right] \right) \quad (3)$$

where $h_i \in R^l$, $W_1 \in R^{l' \times l}$, $a \in R^{1 \times 2l'}$ and $\|$ indicates the concatenation operation. The attentions of all neighbors were then normalized according to

$$\alpha_{ij} = softmax \left(e_{ij} \right) = \frac{\exp \left(e_{ij} \right)}{\sum_{k \in N(j)} \exp \left(e_{ik} \right)} \quad (4)$$

Attentions were used as weights to update node i as follows:

$$h'_i = elu \left(\sum_{j \in N(i)} \alpha_{ij} W_1 h_j \right) \quad (5)$$

where $h'_i \in R^{l'}$. We then computed attentions many times and calculated the mean as the final attention:

$$h'_i = elu \left(\frac{1}{K} \sum_{k=1}^{K} \sum_{j \in N(i)} \alpha_{ij}^k W_1^k h_j \right) \quad (6)$$

After updating all nodes, the output of the complete molecular graph was the mean of them:

$$H = \left(\frac{1}{N} \sum_{i=1}^{N} h'_i \right) \quad (7)$$

Initial molecule featurization

Similar to other graph-based methods [18, 20], we used the properties of molecules to initialize the nodes of the molecular graphs before the data were imported into the GNN model. The atomic features are given in Supplementary Table 1.

Molecular fingerprints

Molecular fingerprints are roughly divided into substructure key-based fingerprints, topological or path-based fingerprints and circular fingerprints [34]. Three complementary fingerprints (MACCS fingerprint [35], Pharmacophore ErG fingerprint [36] and PubChem fingerprint [37] were used in the FP-GNN model because they can complement and holographically express molecular characteristics [38]. The three fingerprints are described as follows:

MACCS fingerprint: a substructure key-based fingerprint using SMARTS pattern. MACCS contains most atomic properties, bond properties and atomic neighborhoods at diverse topological separations, which are meaningful for drug discovery. We choose the short variant of $1 + 166$ bits for this study.
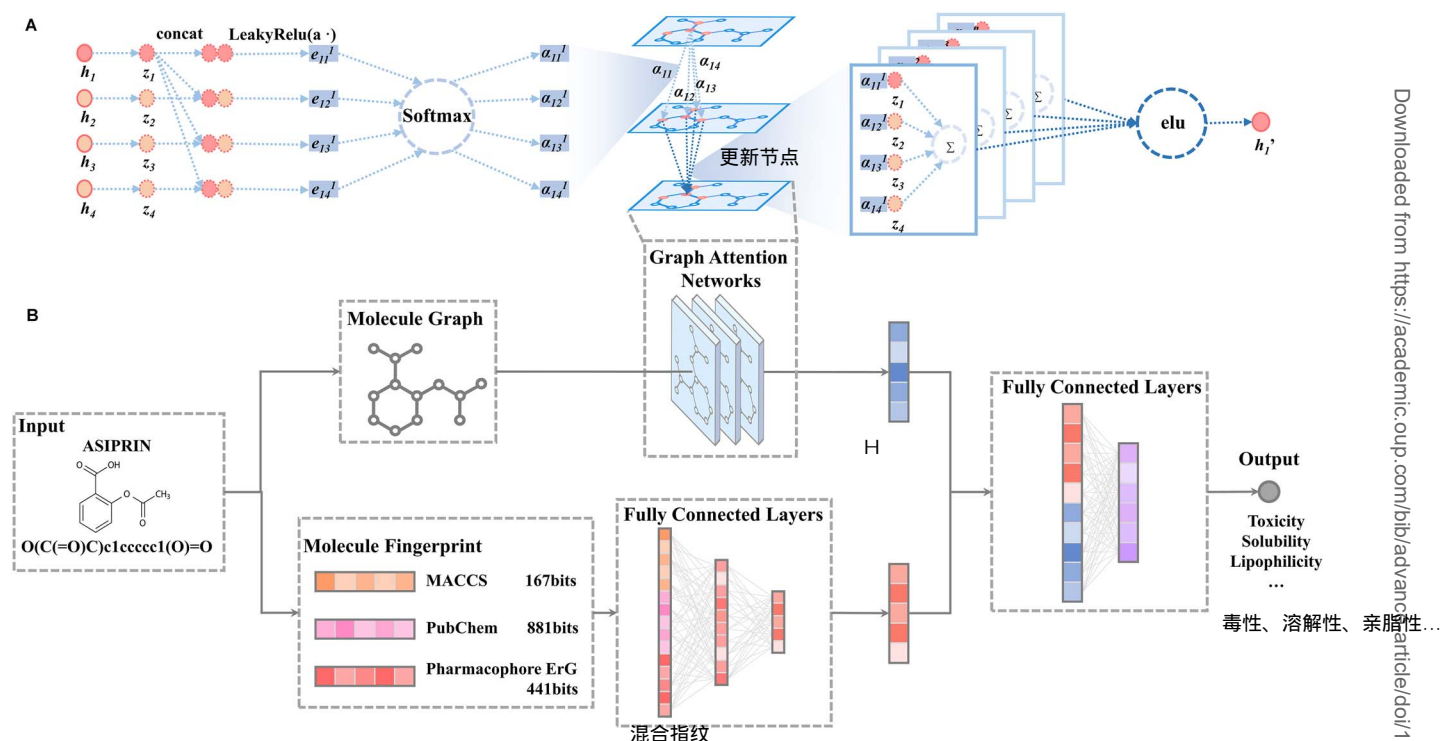
Figure 1. The Architecture of FP-GNN. (**A**) The graph attention network calculates attentions between each node and its neighbors, and then updates the node with those relative attentions. (**B**) The FP-GNN model combines the information from the molecular graphs and fingerprints to predict molecular properties.

PubChem fingerprint: a substructure key-based fingerprint of 881 bits with extensive coverage of chemical structures.

Pharmacophore ErG fingerprint: a 2D pharmacophore fingerprint using an extended reduce graph (ErG) method and applying pharmacophore-type node descriptions to encode molecular properties.

FP-GNN network architecture

As shown in Figure 1B, the FP-GNN architecture first combined the molecular graph and three complementary molecular fingerprints into the flexible and dynamic neural network. The simplified molecular input line entry system (SMILES) notation of the molecule was inputted to two paths of the FP-GNN architecture.

On one path, three complementary fingerprints (MACCS fingerprint, PubChem fingerprint and Pharmacophore ErG fingerprint), termed the mixed fingerprints, were concatenated according to

$$FP = \left(FP_{PubChem} \parallel FP_{MACCS} \parallel FP_{\text{Pharmacophore ErG}} \right) \quad (8)$$

The fingerprints vector was inputted into the artificial neural network (ANN) to obtain the following representation (Equation 9):

$$V' = W_2 \cdot FP + b \quad (9)$$

On the other path, the GNN model was used to capture the information of the molecular graph. The node representation was aggregated from itself and its neighbors by the attention mechanism. Finally, the average of all nodes was produced as the output to represent the molecular graph.

The outcomes received from the two paths were then fitted together and imported into fully connected layers to produce the ultimate output.

Hyperparameter optimization and training protocol

In this study, the Hyperopt Python package [39] was employed to conduct Bayesian optimization of hyperparameters. Six hyperparameters are chosen: the dropout rate of GNN, the number of multi-head attentions, the hidden size of attentions, the hidden size and dropout rate of the fingerprint networks (FPN), and the ratio of GNN in FP-GNN. FP-GNN uses dropout mechanism in both FPN and GNN to avoid overfitting during training. Meanwhile, FP-GNN can split datasets into three parts (training set, validation set and test set) and then utilize validation set to determine the final model for the better generalization ability.

FP-GNN was developed by the Pytorch framework. All FP-GNN models were trained on the SCUTGrid (SCUT supercomputing platform), which uses Matrox MGA G200e.

Benchmark datasets and performance evaluation metric

The performances of the FP-GNN models were extensively evaluated using three benchmark datasets. First, 13 commonly used public datasets (Supplementary Table 2) relevant to drug discovery were used to test the performance of FP-GNN, including three physicochemical datasets (ESOL [40], FreeSolv [41] and Lipophilicity [42]), 6 bioactivity and biophysics datasets (MUV [43], HIV [44], BACE [45], PDBbind-C, PDBbind-R and PDBbind-F [46]), and 4 physiology and toxicity datasets (BBBP [47], Tox21 [48], SIDER [49] and ClinTox [50, 51]). Second, LIT-PCBA [52], a recently developed unbiased and realistic dataset that consists of 15 targets and 7844 confirmed active and 407 381 confirmed inactive compounds

(Supplementary Table 4), was used to evaluate the performance of FP-GNN. Finally, 14 phenotypic screening datasets for breast cell lines (Table 2) were also employed to assess the predictive power of FP-GNN [53].

The regression tasks were evaluated by root-mean-square error (RMSE), while the classification tasks were evaluated by the area under the receiver operating characteristic curve (ROC-AUC) or the area under the precision recall curve (PRC-AUC).

Results and discussion
Performance of the FP-GNN network architecture on the public benchmark datasets

The 13 benchmark datasets related to drug discovery from Wu *et al.* [27] were used to evaluate the predictive power of the FP-GNN models. As shown in Supplementary Table 2, the benchmark datasets encompassed three categories: physicochemical, bioactivity and biophysics, and physiology and toxicity. The sizes of the datasets varied widely and included small datasets (e.g. PDBbind-C only contains 168 molecules) and large datasets (e.g. MUV datasets containing 17 learning tasks and consisting of 93 087 molecules). For multi-task datasets, the average performance metric of each model was calculated to represent the final performance. ROC-AUC was used as the evaluation metric for all classification tasks, except the MUV datasets. Since the ratio of actives to inactives in the MUV datasets was highly imbalanced, the PRC-AUC was used to evaluate the performance of the classification models based on the MUV datasets, which could better reflect the performance of classification models on imbalanced datasets than ROC-AUC. Regression models were evaluated using RMSE. To fairly compare the published performance of the advanced graph-based DL models (MoleculeNet [27], D-MPNN (Chemprop) [20], Attentive FP [18] and HRGCN+ [54]) and the advanced descriptors-based XGBoost [12] models on the public datasets, the same data-split code was adopted to randomly split each dataset into the training set, the validation set and the test set, with a ratio of 8:1:1. In addition, BACE, BBBP and HIV datasets were split at the same ratio based on molecular scaffolds. To reduce the occasionality in data splitting and to ensure the reliability of the results, we evaluated FP-GNN models based on 10 different random seeds and computed the average values of evaluation metrics as the final result. All results were counted after hyperparameter optimization.

The active compounds from the bioactivity and biophysics (Supplementary Table 2) datasets were measured based on their binding affinities for different biological targets. There is no doubt that accurately predicting the biological activities of small molecules for a given target can accelerate the discovery and development of new drug candidates. There was a total of eight learning tasks for this type of dataset (Supplementary Table 2), including four classification tasks based random- and scaffold-splitting methods for the HIV and BACE bioactivity datasets, one classification task based on the random-splitting method for the MUV bioactive dataset, and three regression tasks based random-splitting method for three biophysical datasets (PDBbind-C, PDBbind-R and PDBbind-F). As shown in Table 1, FP-GNN performed best on three of the eight learning tasks, including the two classification learning tasks based on the scaffold-splitting of BACE and HIV, and one regression task of PDBbind-C. Chemprop achieved the four best performance tasks, including one classification task based on the random-splitting of BACE and HIV, and two regression tasks of PDBbind-F and PDBbind-R. Graph-based weave models from MoleculeNet performed best

on the MUV dataset that contains 17 subtasks. Notably, FP-GNN achieved the second-best performance on the random-splitting of HIV, MUV, PDBbind-F and PDBbind-R. Although FP-GNN did not perform the best on some datasets, our model still performed comparatively well on those datasets.

Molecules from physiology and toxicity datasets record their effects on living bodies, such as the blood–brain barrier penetration dataset (BBBP), the side effect resource dataset (SIDER) and toxicities dataset (Tox21 and ClinTox). Thus, those datasets are closely related to the physiology and toxicity properties of drugs. Precisely predicting the physiological and toxicological properties of compounds can rule out improper molecules in the early stages of drug discovery, which is beneficial for reducing the cost reduction of new drug development. However, it remains challenging to predict physiological and toxicological properties accurately. As shown in Table 1, FP-GNN achieved the three best classification performance results on the BBBP (from random- and scaffold-splitting methods) and SIDER datasets, while Chemprop performed best on Tox21 and XGBoost performed best on ClinTox. FP-GNN also exhibited better performance than the Weave models of MoleculeNet on the ClinTox dataset.

The physicochemical properties of a given drug can reflect its pharmacokinetic phases in the body. The physicochemical properties of molecules play a key role in the development of candidate drugs. Therefore, the accurate prediction of the physicochemical properties of molecules facilitates drug discovery and development. FreeSolv, ESOL and Lipophilicity datasets were used to evaluate the predictive ability of the FP-GNN network architecture for physicochemical properties. Table 1 illustrates that FP-GNN performed best on the FreeSolv dataset, HRGCN+ performed best on the ESOL dataset and Attentive FP performed best on the Lipophilicity dataset. Although FP-GNN performed worse than Attentive FP on the Lipophilicity dataset, it outperformed the other graph-based DL methods (e.g. GCN, MPNN and Weave) in MoleculeNet.

The ultimate goal of building molecular property prediction models is to predict the properties of new molecules with novel scaffolds, to make them fall within the appropriate ranges of the desired properties. Consequently, the scaffold-based splitting method was used on the BACE, BBBP and HIV datasets to ensure that the scaffolds in the training set, validation set and test set were as distinct as possible. As shown in Table 1, the performance of scaffold-splitting classification models was lower than that of models based on random-splitting. Those data suggest that the scaffold-based splitting method was more challenging for learning tasks. Our FP-GNN models performed best on all three datasets using scaffold-based splitting and showed the same outstanding performance as the random-splitting method. All such results demonstrate that FP-GNN is stable in predicting molecules with new scaffolds.

Typically, DL models perform moderately on small datasets because insufficient samples could not provide adequate information. FP-GNN model aims to complement the information captured automatically from molecular graphs by using prior information from molecular fingerprints. As shown in Table 1 and Table S2, FP-GNN performed best on PDBbind-C and FreeSolv datasets with less than 1000 molecules, indicating that FP-GNN is also competitive on the datasets without enough samples.

Out of 16 learning tasks from 13 public benchmark datasets (Table 1), FP-GNN showed the best performance on seven tasks, while Chemprop exhibited the best performance on five tasks. MoleculeNet, Attentive FP, HRGCN+ and XGBoost performed best on one task each. Supplementary Table 3 summarizes how our

Table 1. Predictive performance results of FP-GNN on 13 commonly used public datasets

| | | | DL | | | | ML | |
Dataset	Split type	Metric	MoleculeNet (Graph) [27]	Chemprop (optimized) [20]	Attentive FP [54]	HRGCN+ [54]	XGBoost [54]	FP-GNN
BACE	Random	ROC–AUC		**0.898**	0.876	0.891	0.889	0.881
	Scaffold	ROC–AUC	0.806 (Weave)	0.857				**0.860**
HIV	Random	ROC–AUC		**0.827**	0.822	0.824	0.816	0.825
	Scaffold	ROC–AUC	0.763 (GC)	0.794				**0.824**
MUV	Random	PRC–AUC	**0.109 (Weave)**	0.053	0.038	0.082	0.068	0.090
Tox21	Random	ROC–AUC	0.829 (GC)	**0.854**	0.852	0.848	0.836	0.815
BBBP	Random	ROC–AUC		0.917	0.887	0.926	0.926	**0.935**
	Scaffold	ROC–AUC	0.690 (GC)	0.886				**0.916**
ClinTox	Random	ROC–AUC	0.832 (Weave)	0.897	0.904	0.899	**0.911**	0.840
SIDER	Random	ROC–AUC	0.638 (GC)	0.658	0.623	0.641	0.642	**0.661**
PDBbind-C	Random	RMSE		1.910				**1.876**
PDBbind-F	Random	RMSE		**1.286**				1.296
PDBbind-R	Random	RMSE		**1.338**				1.349
FreeSolv	Random	RMSE	1.150 (MPNN)	1.009	1.091	0.926	1.025	**0.905**
ESOL	Random	RMSE	0.580 (MPNN)	0.587	0.587	**0.563**	0.582	0.675
Lipophilicity	Random	RMSE	0.655 (GC)	0.563	**0.553**	0.603	0.574	0.625

Each dataset was split into training, validation and test sets using the corresponding data-split codes from published studies. The FP-GNN models used the same dataset and data split method to fairly compare the MoleculeNet, Chemprop, Attentive FP, HRGCN+ and XGBoost models. Bold font illustrates the models that outperformed all other models. The best graph-based models from MoleculeNet were used, the optimized results of the Chemprop models were from the original study [20], and the best performance results for the Attentive FP, HRGCN+ and XGBoost models were chosen from Wu *et al.* [54]. MPNN: message passing neural networks; GC: graph convolutional models and Weave: Weave models.

FP-GNN compared to each of the baseline models, including four commonly used advanced graph-based DL methods and the venerable descriptors-based ML method, XGBoost. Our FP-GNN model consistently matches or outperforms not only for each baseline individually (Supplementary Table 3), but also across all baselines (Table 1), indicating that coupling molecular graphs and fingerprints can improve the degree of generalization of graph-based DL algorithms to predict molecular properties better. The outstanding performance of FP-GNN on drug discovery-related datasets makes FP-GNN one of the most competitive DL methods in drug discovery practice.

Performance of the FP-GNN network architecture on an unbiased and realistic LIT-PCBA dataset

In 2020, Viet-Khoa Tran-Nguyen *et al.* [52] designed an unbiased and realistic dataset called LIT-PCBA, specifically dedicated to ML and virtual screening methods. LIT-PCBA consists of 7844 confirmed active and 407 381 confirmed inactive compounds toward 15 targets, which were collected from the PubChem BioAssay (PCBA) dataset [55]. For each target, unbiased training and validation sets were constructed using the asymmetric validation embedding method at the ratio of 3:1. The details of LIT-PCBA dataset are summarized in Supplementary Table 4. We therefore used this dataset to evaluate the predictive power of FP-GNN. Five fingerprint-based methods [56] (NB, SVM, RF, XGBoost and DNN) and two graph-based methods (GCN and GAT) were selected as the baseline models. All fingerprint-based models were constructed based on the Morgan fingerprint [57] and the mixed fingerprints (MACCS FP, PubChem FP and Pharmacophore ErG FP). According to the original paper and Jiang *et al.* [56], ROC-AUC was used to evaluate the performance of the classification models for the LITPCBA dataset.

As shown in Figure 2A, when compared to five Morgan fingerprint-based models and two graph-based models, FP-GNN exhibited the best performance on six targets (ADRB2, ALDH1, ESR1_ago, MAPK1, PPARG and TP53). Meanwhile, NB achieved the best performance on two targets (IDH1 and VDR); DNN achieved the best performance on two targets (FEN1 and OPRK1); GCN performed best on two targets (ESR1_ant and MTORC1); SVM, XGBoost and GAT achieved the best performance on one task each (PKM2, GBA and KAT2A, respectively). More importantly, FP-GNN performed the best on average, with the highest average AUC value of 0.739. Compared to the mixed fingerprints-based models, FP-GNN also showed a similar outstanding performance (Figure 2B). The details of the direct comparisons between the FP-GNN model and each of the baseline models are listed in Supplementary Table 5. It is clear that not only did our FP-GNN models outperform the fingerprint-based models but also exhibited comparable or superior performance to the two classical graph-based DL models (GCN and GAT). Even on the most challenging LIT-PCBA dataset, FP-GNN also exhibited strong competitiveness and can be used to accurately predict the biological activity of molecules for drug discovery campaigns.

Performance of FP-GNN compared to the advanced graph-based and fingerprint-based models on cell-based phenotypic screening datasets

Phenotypic-based screening (e.g. whole-cell activity), an original but indispensable drug screening method, has regained attention in recent years [58–62]. The phenotypic screening datasets (Table 2) for 13 breast cancer cell lines and 1 normal breast cell line were used to evaluate the performance of FP-GNN. Recently, He *et al.* [53] reported four graph-based DL models and one advanced fingerprint-based XGBoost model to predict the activities of molecules against those cell lines. Therefore, FP-GNN models were developed on the 14 cell-based phenotypic screening
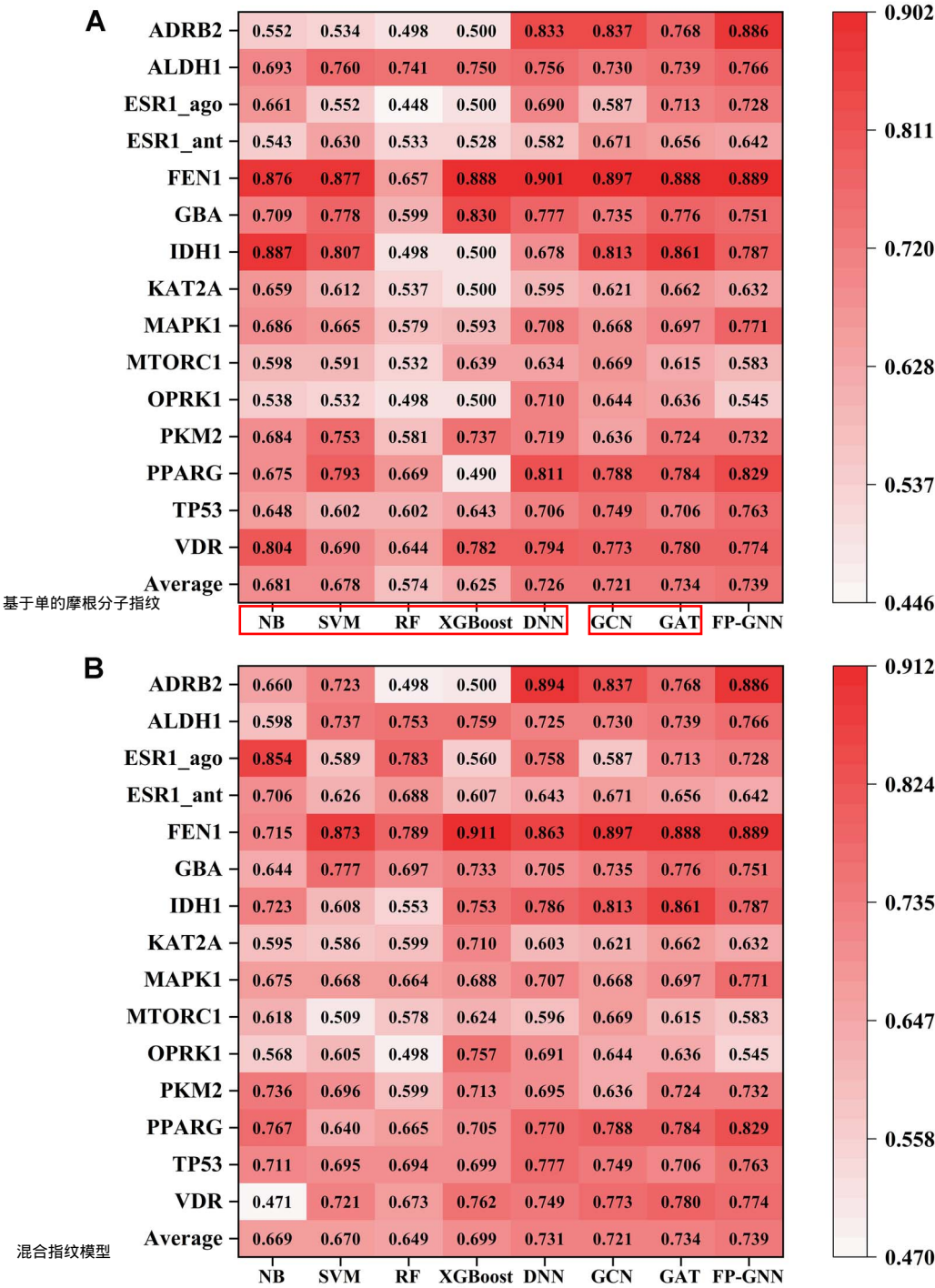
Figure 2. Performance of FP-GNN compared to the baseline models on the LIT-PCBA dataset. (**A**) The NB, SVM, RF, XGBoost and DNN models were established using the Morgan fingerprint as the molecular representation. (**B**) The NB, SVM, RF, XGBoost and DNN models were established using the mixed fingerprints (MACCS fingerprint, PubChem fingerprint and Pharmacophore ErG fingerprint) as the molecular representation. The performances of models based on the Morgan fingerprint were collected from Jiang *et al.* [56]. Three graph-based models (GCN, GAT and FP-GNN) as well as models based on the mixed fingerprint were constructed using the same benchmark.

datasets and compared to the performances published for the five models.

As shown in Table 2, FP-GNN performed best on 8 out of 14 cell lines (i.e. MDA-MB-453, SK-BR-3, T-47D, MCF-7, BT-474, BT-20, BT-549 and HBL-100), while Attentive FP achieved the best performance on three cell lines (HS-578 T, MDA-MB-231 and

Bcap37), XGBoost achieved the best performance on two cell lines (MDA-MB-361 and MDA-MB-468), and GCN performed best on MDA-MB-435. Notably, our FP-GNN models were the second-best in HS-578 T, MDA-MB-231 and MDA-MB-468 datasets. Importantly, FP-GNN achieved the overall best performance on these 14 cell lines, with the highest average AUC value of 0.849. FP-GNN

Table 2. Performance of FP-GNN on 14 breast cell line datasets compared to the graph-based DL models

Cell lines	Classification	Compounds	Task metric	Attentive FP [53]	GAT [53]	GCN [53]	MPNN [53]	XGBoost [53]	FP-GNN
MDA-MB-453	HER-2+[a]	440	ROC–AUC	0.872	0.812	0.866	0.715	0.810	**0.886**
SK-BR-3	HER-2+	2026	ROC–AUC	0.805	0.840	0.839	0.760	0.848	**0.852**
MDA-MB-435	HER-2+	3030	ROC–AUC	0.824	0.830	**0.858**	0.749	0.853	0.820
T-47D	Luminal A[b]	3135	ROC–AUC	0.812	0.763	0.819	0.751	0.821	**0.846**
MCF-7	Luminal A	29 378	ROC–AUC	0.845	0.800	0.833	0.843	0.826	**0.866**
MDA-MB-361	Luminal B[c]	367	ROC–AUC	0.938	0.896	0.955	0.972	**0.976**	0.905
BT-474	Luminal B	811	ROC–AUC	0.787	0.657	0.866	0.847	0.827	**0.868**
BT-20	TNBC[d]	292	ROC–AUC	0.735	0.721	0.740	0.784	0.740	**0.887**
BT-549	TNBC	1182	ROC–AUC	0.630	0.710	0.669	0.634	0.651	**0.807**
HS-578 T	TNBC	469	ROC–AUC	**0.830**	0.758	0.636	0.665	0.753	0.770
MDA-MB-231	TNBC	11 202	ROC–AUC	**0.870**	0.770	0.859	0.850	0.865	0.866
MDA-MB-468	TNBC	1986	ROC–AUC	0.875	0.875	0.887	0.858	**0.896**	0.888
Bcap37	TNBC	275	ROC–AUC	**0.858**	0.767	0.693	0.807	0.744	0.779
HBL-100	Normal cell line	316	ROC–AUC	0.645	0.641	0.658	0.701	0.776	**0.850**
Average				0.809	0.774	0.798	0.781	0.813	**0.849**

[a]HER-2+: HER2-positive breast cancers. [b]Luminal A: Luminal A breast cancer is hormone-receptor positive (estrogen-receptor and/or progesterone-receptor positive), HER2-negative, with low levels of the protein Ki-67. [c]Luminal B: Luminal B breast cancer is hormone-receptor positive (estrogen-receptor and/or progesterone-receptor positive), HER2 positive or HER2 negative, with high levels of Ki-67. [d]TNBC: triple-negative breast cancer. Each dataset was split into training, validation and test sets using the corresponding data-split codes from He *et al*. [53]. The FP-GNN models used the same dataset and data split method to fairly compare the Attentive FP, GAT, GCN, MPNN and XGBoost models. Bold font illustrates the models that outperformed all other models. MPNN: message passing neural networks; GCN: graph convolutional networks; GAT: graph attention network.

exhibited excellent performance on the cell-based phenotypic screening datasets, suggesting that FP-GNN holds great potential for phenotype-based drug discovery.

The ablation experiment of FP-GNN

We investigated whether the information of local neighbors and global structures learned from the molecular graphs, and the chemical substructure information learned from the molecular fingerprints could complement each other and assist in optimizing our FP-GNN model.

To analyze the influence of the graph-based module and fingerprint-based module in the FP-GNN model, we counted the ratios of GNN in FP-GNN (Figure 3) based on the optimal set of hyperparameters for each of the 13 public datasets (Supplementary Table 6). As shown in Figure 3, more than half (54.3%) of the ratios of GNN in FP-GNN fell between 0.4 and 0.6, illustrating that the contributions of the two modules to the FP-GNN model were relatively balanced. In addition, pure GNN and pure FPN only accounted for approximately 4.3% of all models, demonstrating that coupling complementary molecular graph and fingerprint strategy to dynamic GNN can improve the performance of molecular property prediction.

The ablation experiment of FP-GNN was conducted on the unbiased and realistic LIT-PCBA dataset. The whole FP-GNN model for each target was split into FPN and GNN models with the original hyperparameters. FP-GNN also used the same hyperparameters except that the ratio of GNN in FP-GNN modules was set to 0.5. As shown in Figure 4, FP-GNN models outperformed FPN and GNN in 10 out of 15 targets. FP-GNN models performed medium, with slightly lower performance than GNN, but evidently higher than FPN models on the other five targets (ESR1_ago, FEN1, KAT2A, MTORC1 and OPRK1). These results illustrated that FP-GNN combines the advantages of FPN and GNN to capture the complementary information of molecular graphs and fingerprints to achieve better performance. A possible reason for this finding was that we used the default parameter 0.5 as the ratio of GNN in FP-GNN modules when building the FP-GNN model, and less information was captured from the unfavorable GNN or FPN module to influence the performance of FP-GNN
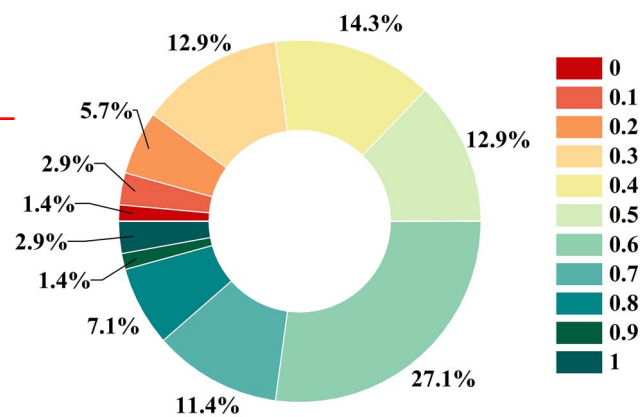

Figure 3. Distribution of the ratios of GNN in the FP-GNN models on the 13 public datasets. The details of optimal sets of hypermeters for the 13 public datasets are shown in Supplementary Table S6.

on the five targets. Collectively, combing molecular graphs and fingerprints can obtain local neighbor and complete structural information from the molecular graphs and substructures, as well as pharmacophore information from the molecular fingerprints, which will enable more accurate predictions of molecular properties.

The influence of different types of fingerprints

We explored the influence of different molecular fingerprints on the performance of our FP-GNN architecture. The Morgan fingerprint is the most common fingerprint used in QSAR/QSPR modeling [58, 63–67]. In addition to the mixture of three complementary fingerprints, we grafted the 1024-bits ECFP-4 fingerprint [68] (called Morgan fingerprint in the RDKit) into FP-GNN architecture and then tested it on public datasets.

As shown in Figure 5, FP-GNN models based on the mixed fingerprints performed better than the FP-GNN models based on the Morgan fingerprint, regardless of whether the analyses were performed on classification datasets (Figure 5A) or regression datasets (Figure 5B). In addition, we retrospect the performances
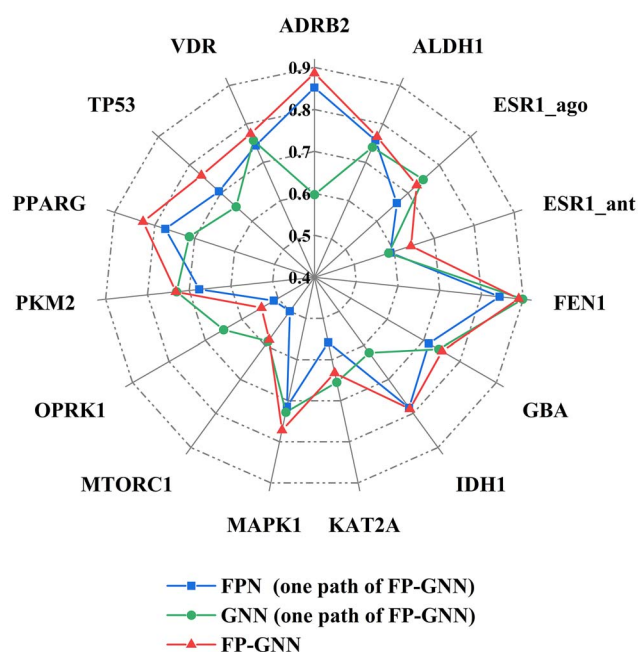
Figure 4. Results of the ablation study on the LIT-PCBA dataset.

of four classical ML methods (i.e. RF, SVM, NB and XGBoost) and one DNN DL method using the mixed fingerprints and the Morgan fingerprint on the LIT-PCBA dataset. As shown in Figure 2, by counting the number of best-prediction models based on the mixed fingerprints and the number of best models based on the Morgan fingerprint, it was seen that the former did not exhibit absolute superiority (42 versus 30, and the three results are equal). Furthermore, when comparing the performance of the Morgan fingerprint-based models and the mixed fingerprint-based models, there is a trend that simple NB and SVM methods can achieve more information from Morgan fingerprint, while advanced algorithms (RF, XGBoost and DNN) can capture more information from the mixed fingerprints. Meanwhile, the FP-GNN based on the mixed fingerprints exhibited better performance than FP-GNN based on the Morgan fingerprint (Figure 5). Thus, those data indicated that compared to the commonly used Morgan fingerprint, coupling the mixed fingerprints and molecular graph could achieve optimal complementarity to exhibit better performance.

The above-mentioned differences in complementarity may be related to the specific generation algorithms of fingerprints. The mixed fingerprints recorded most of the atomic and bond properties (MACCS fingerprint), extensive chemical structures and substructures (PubChem fingerprint) and pharmacophore feature (Pharmacophore ErG fingerprint) information, which may not be included in the features of molecular graphs. However, the Morgan fingerprint only records the local environmental information of atoms, which may be similar to the molecular graph features. Therefore, unlike the Morgan fingerprint, the mixed fingerprints can better complement the molecular graph features, and elicit better molecular representations.

The anti-noise ability of FP-GNN

DL models place extensive demands on data quality and generally require a large quantity of correct data. Obtaining sufficient high-quality data is still the central challenge in computer-assisted drug discovery [32]. Actually, the available data used in drug discovery practices are usually scarce and of mediocre quality. When the model is used in real-world scenes, the noises in the data will affect the training process and reduce the practicality of the model. Therefore, we ran FP-GNN on the noisy data to test its anti-noise ability.

We divided the HIV dataset (41 127 compounds) at a ratio of 8:1:1 to generate the training set, validation set and test set. We ensured that the labels in the test set remained unchanged, and then changed the labels in the training set and validation set according to a predetermined ratio to generate noise artificially. The anti-noise ability of FP-GNN was compared with two DL methods (Attentive FP and HRGCN+) and one advanced method (XGBoost) from Wu *et al.* [54]. The same data, data split, evaluation metric and noise rates were also adopted from Wu *et al.* to ensure a fair comparison. Figure 6 indicates that FP-GNN achieved the state-of-the-art performance in the anti-noise tests. Based on the excellent anti-noise ability of our FP-GNN model, it is foreseeable that it can handle poor data situations in real drug discovery scenarios.

The interpretation of FP-GNN

The FP-GNN model based on the BBBP dataset containing the blood–brain barrier (BBB) permeability of molecules was used to analyze the interpretability of the model. Since the BBB can block most drugs and hormones, it is essential to accurately predict the BBB permeability of molecules for the development of drugs targeting central nervous system diseases. Facing the natural BBB that exists in the human body, hydrophobic molecules (low polarity and high ClogP) could bypass the BBB easily, while the converse is true for hydrophilic molecules.

The FP-GNN architecture can compute the attentions of adjacent atoms and then map the attentions to bonds connected to atoms (Figure 1). For a given molecule, the attention coefficient can be used to quantitatively characterize whether chemical fragments contribute more to the prediction of molecular properties. As shown in Figure 7, the portions of the molecule colored more darkly were more significantly in predicting whether the molecule can pass the BBB, while the role of the light-colored portions is less important. Considering an active molecule as an example (Figure 7A), most of the substructural groups of this compound are hydrophobic, laying the foundation for penetrating the BBB. The benzene ring (C7–C12, marked in red) of the molecule has the least polarity and maximum contribution to BBB penetration. We used ChemBioDraw (v.14.0.0.117) to further quantitatively analyze the ClogP values of these chemical fragments. Quantitative analyses of ClogP showed that the chemical portion marked in red had a lower polarity (ClogP = 2.142), while the gray mark portion had a higher polarity (ClogP = 1.389). In fact, our FP-GNN model paid great attention to the low-polarity benzene ring, which was also consistent with the prediction results as an active molecule. As shown in Figure 7B, for an inactive molecule, the dark portion (marked in red) represents an exposed substituent amino group that provides the majority of polarity to prevent the molecule from passing the BBB. The ClogP value of the chemical fragment in red is −0.905, while the ClogP of the fragment in gray is 0.934. The lower ClogP indicates that the red portion of the molecule was more hydrophilic and difficult to cross the BBB. The high attention marked in the red part from our FP-GNN model was consistent with the inactive prediction results. These cases not only demonstrate that our FP-GNN model was interpretable, but also hint the FP-GNN network architecture can learn the relationships
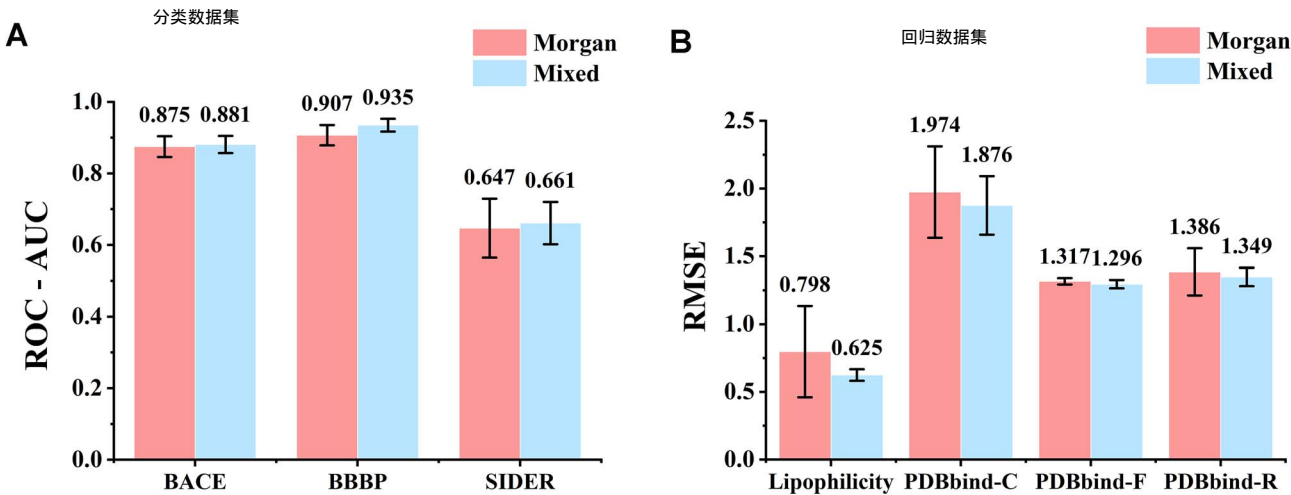
Figure 5. Comparisons of the performance of FP-GNN models based on the Morgan, as well as models based on the mixed of three complementary fingerprints. (**A**) The performance results for three classification datasets (BACE, BBBP and SIDER). (**B**) The performance results for four regression datasets (Lipophilicity, PDBbind-C, PDBbind-F and PDBbind-R). To ensure the reliability of the results, after optimizing the hyperparameters, the average metric value of the FP-GNN models based on 10 different random seeds was computed as the final result.
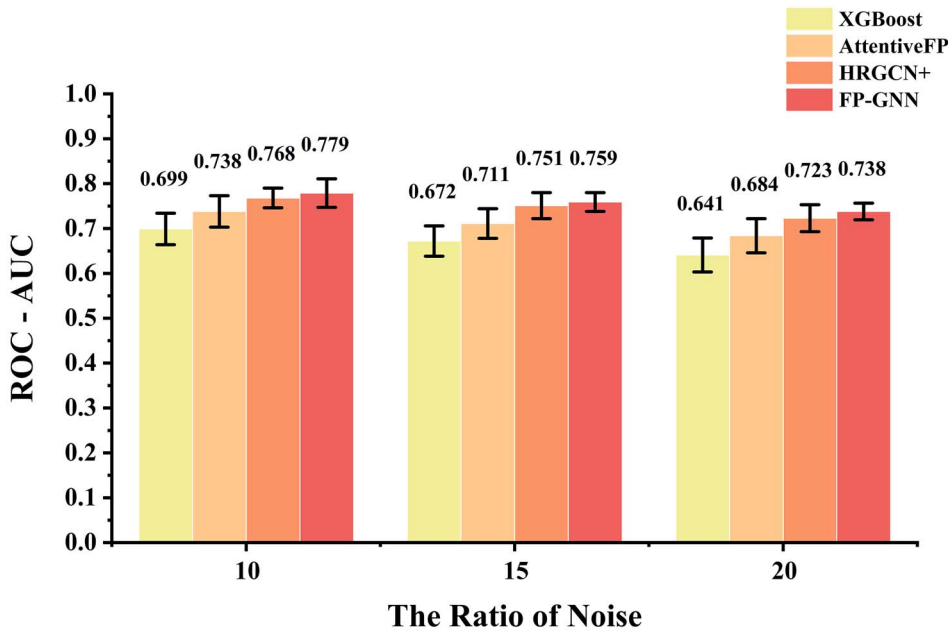

Figure 6. The anti-noise performances of Attentive FP, HRGCN+, XGBoost and FP-GNN models with different noise rates on the HIV dataset. The anti-noise results for Attentive FP, HRGCN+ and XGBoost models were collected from Wu *et al.* [54].

Table 3. The 10 most significant bits of the mixed fingerprints on the prediction of the FreeSolv dataset

| Rank	Importance	Mixed fingerprint bit	Fingerprint	Fingerprint bit	Meaning
1	0.276	190	Pharmacophore ErG	23	('Donor', 'Acceptor', 2)
2	0.226	189	Pharmacophore ErG	22	('Donor', 'Acceptor', 1)
3	0.196	93	MACCS	93	OC(N)C
4	0.184	1048	PubChem	440	C–C–O=O
5	0.177	1060	PubChem	452	C–O=O
6	0.170	111	MACCS	111	NCO
7	0.165	140	MACCS	140	OH
8	0.162	1015	PubChem	407	OCP
9	0.158	276	Pharmacophore ErG	109	('Donor', 'Aromatic', 4)
10	0.156	42	MACCS	42	C#N
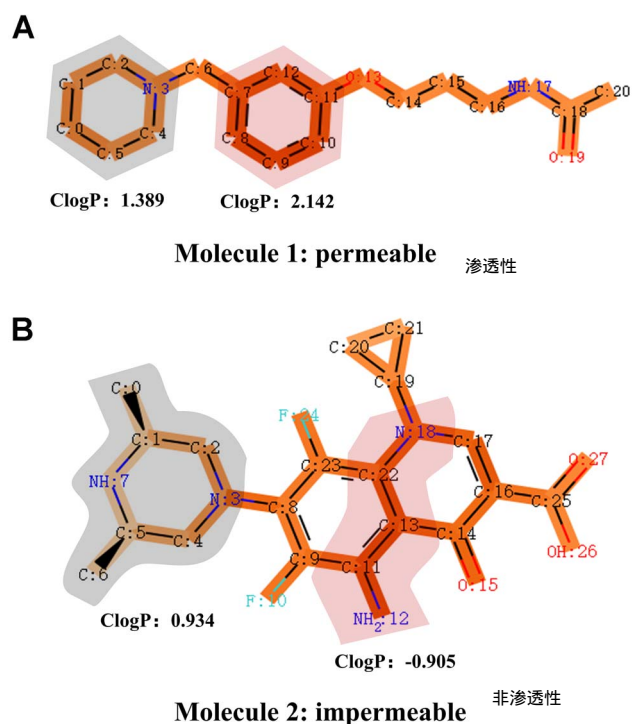
Figure 7. The importance of molecular structures during the prediction process. The darker the color, the more important are for the structures. Molecules were obtained from the BBBP (the blood–brain barrier penetration) dataset. (**A**) Molecule 1 is permeable, and the darker colored portion has a higher ClogP, which indicates a stronger lipophilicity. (**B**) Molecule 2 is impermeable, and the darker colored portion has lower ClogP, which means a weaker lipophilicity. The important portions that were captured by FP-GNN models were consistent with the prediction results.

between molecular substructures (chemical fragments) and their molecular properties.

Besides the GNN module, we analyzed the interpretation of the FPN module. We chose the FreeSolv (Free Solvation) dataset that contains the hydration free energy of small molecules in water. The mixed fingerprints (MACCS fingerprint, PubChem fingerprint and Pharmacophore ErG fingerprint) that we used in the FPN model had 1489 bits in total. We changed the values of each bit in order and then inputted the mixed fingerprints to the training model. The effects made by different changing bits indicated the importance of fingerprints in the model. The more a modified value deviated from the original prediction, the more critical the fingerprint bit was in predicting the free solvation of molecules. The ten most significant bits are shown in Table 3. As shown in Table 3, substructures represented by the 4th, 5th, 7th and 10th bits had strong polarity and high water solubility, which play essential roles in the free solvation of molecules. We calculated the Pearson correlation coefficient between the hydration-free energy of molecules and these 10 fingerprint bits. The Pearson values of the 3rd, 6th and 10th bits were above 0.7, indicating that they exhibited a strong correlation. Thus, it can be seen that our model captured the significant part of fingerprints, and the prediction results from the FP-GNN model can be explained. In the top 10 crucial bits, there are four, three and three bits coming from MACCS fingerprint, PubChem fingerprint and Pharmacophore ErG fingerprint, respectively. Such results illustrated that three fingerprints collectively played an important role in the FP-GNN model.

Conclusions

In this study, we advanced a new DL architecture called FP-GNN, which first couples the graph attention network based on a molecular graph and the artificial neural network based on the mixed molecular fingerprints to generate more comprehensive molecular representations. The performance of FP-GNN on 13 classical public datasets revealed that FP-GNN model performed outstandingly compared to four recently-published graph-based DL algorithms (MoleculeNet, Chemprop, Attentive FP and HRGCN+) and the venerable XGBoost algorithm. We also evaluated the predictive power of FP-GNN on an unbiased and realistic LIT-PCBA dataset and 14 phenotypic drug screening datasets related to breast cancer cell lines. The evaluation outcomes further indicated that our FP-GNN model was highly competitive. Analyses of the influence of molecular graphs and fingerprints on the FP-GNN model, as well as the results of ablation experiments found that (1) molecular graphs and mixed molecular fingerprints in the FP-GNN architecture contributed to improve model prediction performance; and (2) embedding different fingerprints in the FP-GNN architecture affected its predictive performance. In addition, the excellent anti-noise ability of FP-GNN indicated that our FP-GNN model could solve noisy (poor) data in the natural scenes of drug discovery. Importantly, the FP-GNN model has intuitive interpretability and can identify important chemical fragments in a molecule, which can assist in designing and optimizing new molecules with desired properties or functions.

Key Points

- We presented a deep learning (DL) architecture named fingerprints and graph neural networks (FP-GNN) to predict molecular properties.
- Extensive experimental results showed that FP-GNN was highly competitive compared to classic machine learning methods and state-of-the-art DL methods.
- The ablation experiments of FP-GNN indicated that information from molecular graphs and molecular fingerprints is complementary to improve the predictive power of molecular properties.
- The intuitive interpretability of the FP-GNN model can provide important chemical fragments to assist chemists and pharmacists in designing or optimizing new molecules with desired properties.

Supplementary data

Supplementary data are available online at https://academic.oup.com/bib.

Data and Code Availability

The full datasets and the source codes for FP-GNN are freely available on GitHub at https://github.com/idrugLab/FP-GNN.

Funding

the allocation time from the SCUTGrid at South China University of Technology.

References

1. Toropov AA, Toropova AP. QSPR/QSAR: state-of-art, weirdness, the future. *Molecules* 2020;**25**:1292.

2. Muratov EN, Bajorath J, Sheridan RP, *et al.* QSAR without borders. *Chem Soc Rev* 2020;**49**:3525–64.

3. Lewis RA. A general method for exploiting QSAR models in lead optimization. *J Med Chem* 2005;**48**:1638–48.

4. Cherkasov A, Muratov EN, Fourches D, *et al.* QSAR modeling: where have you been? Where are you going to? *J Med Chem* 2014;**57**:4977–5010.

5. Eklund M, Norinder U, Boyer S, *et al.* Choosing feature selection and learning algorithms in QSAR. *J Chem Inf Model* 2014;**54**: 837–43.

6. Moriwaki H, Tian YS, Kawashita N, *et al.* Mordred: a molecular descriptor calculator. *J Chem* 2018;**10**:1–14.

7. Cao DS, Xiao N, Xu QS, *et al.* Rcpi: R/Bioconductor package to generate various descriptors of proteins, compounds and their interactions. *Bioinformatics* 2015;**31**:279–81.

8. Yap CW. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J Comput Chem* 2011;**32**: 1466–74.

9. Clarke MRB, Duda RO, Hart PE. Pattern classification and scene analysis. *J R Stat Soc Ser A* 1974;**137**:442–3.

10. Cortes C, Vapnik V. Support-vector networks. *Mach Learn* 1995;**20**:273–97.

11. Breiman RF. Vaccines as tools for advancing more than public health: perspectives of a former director of the National Vaccine Program office. *Clin Infect Dis* 2001;**32**:283–8.

12. Chen T, Tong H, Benesty M. Xgboost: extreme gradient boosting. *R package version 04-2* 2015;**1**:1–4.

13. Dai H, Dai B, Song L. Discriminative embeddings of latent variable models for structured data. *PMLR* 2016;**48**:2702–11.

14. Coley CW, Barzilay R, Green WH, *et al.* Convolutional embedding of attributed molecular graphs for physical property prediction. *J Chem Inf Model* 2017;**57**:1757–72.

15. Duvenaud D, Maclaurin D, Aguilera-Iparraguirre J, *et al.* Convolutional networks on graphs for learning molecular fingerprints. *Adv Neural Inform Process Syst* 2015;**28**:2224–32.

16. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. *ICLR* 2017.

17. Veličković P, Cucurull G, Casanova A, *et al.* Graph attention networks. *ICLR* 2018; arXiv:1710.10903.

18. 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* 2020;**63**:8749–60.

19. Gilmer J, Schoenholz SS, Riley PF, *et al.* Neural message passing for quantum chemistry. *PMLR* 2017;**70**:1263–72.

20. Yang K, Swanson K, Jin W, *et al.* Analyzing learned molecular representations for property prediction. *J Chem Inf Model* 2019;**59**: 3370–88.

21. Kearnes S, McCloskey K, Berndl M, *et al.* Molecular graph convolutions: moving beyond fingerprints. *J Comput Aided Mol Des* 2016;**30**:595–608.

22. Withnall M, Lindelof E, Engkvist O, *et al.* Building attention and edge message passing neural networks for bioactivity and physical-chemical property prediction. *J Chem* 2020;**12**:1.

23. Rathi PC, Ludlow RF, Verdonk ML. Practical high-quality electrostatic potential surfaces for drug discovery using a graph-convolutional deep neural network. *J Med Chem* 2020;**63**:8778–90.

24. Pan X, Wang H, Li C, *et al.* MolGpka: a web server for small molecule pKa prediction using a graph-convolutional neural network. *J Chem Inf Model* 2021;**61**:3159–65.

25. Wang J, Cao D, Tang C, *et al.* DeepAtomicCharge: a new graph convolutional network-based architecture for accurate prediction of atomic charges. *Brief Bioinform* 2021;**22**:bbaa183.

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. Wu Z, Ramsundar B, Feinberg EN, *et al.* MoleculeNet: a benchmark for molecular machine learning. *Chem Sci* 2018;**9**:513–30.

28. Rifaioglu AS, Nalbat E, Atalay V, *et al.* DEEPScreen: high performance drug-target interaction prediction with convolutional neural networks using 2-D structural compound representations. *Chem Sci* 2020;**11**:2531–57.

29. Mayr A, Klambauer G, Unterthiner T, *et al.* Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chem Sci* 2018;**9**:5441–51.

30. Jiang D, Wu Z, Hsieh CY, *et al.* Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. *J Chem* 2021;**13**:12.

31. Stepisnik T, Skrlj B, Wicker J, *et al.* A comprehensive comparison of molecular feature representations for use in predictive modeling. *Comput Biol Med* 2021;**130**:104197.

32. Yang X, Wang Y, Byrne R, *et al.* Concepts of artificial intelligence for computer-assisted drug discovery. *Chem Rev* 2019;**119**: 10520–94.

33. Hamilton WL, Ying R, Leskovec J. Representation learning on graphs: methods and applications. *IEEE Data Eng Bull* 2017;**40**: 52–74.

34. Cereto-Massague A, Ojeda MJ, Valls C, *et al.* Molecular fingerprint similarity search in virtual screening. *Methods* 2015;**71**: 58–63.

35. Durant JL, Leland BA, Henry DR, *et al.* Reoptimization of MDL keys for use in drug discovery. *J Chem Inf Comput Sci* 2002;**42**: 1273–80.

36. Stiefl N, Watson IA, Baumann K, *et al.* ErG: 2D pharmacophore descriptions for scaffold hopping. *J Chem Inf Model* 2006;**46**: 208–20.

37. Bolton EE, Wang Y, Thiessen PA, *et al.* Chapter 12 - PubChem: integrated platform of small molecules and biological activities. *Annu Rep Comput Chem* 2008;**4**:217–41.

38. Shen WX, Zeng X, Zhu F, *et al.* Out-of-the-box deep learning prediction of pharmaceutical properties by broadly learned knowledge-based molecular representations. *Nat Mach Intell* 2021;**3**:334.

39. Hyperopt: distributed Hyperparameter optimization. https://github.com/hyperopt/hyperopt (24 May 2019, accessed).

40. Delaney JS. ESOL: estimating aqueous solubility directly from molecular structure. *J Chem Inf Comput Sci* 2004;**44**:1000–5.

41. Mobley DL, Guthrie JP. FreeSolv: a database of experimental and calculated hydration free energies, with input files. *J Comput Aided Mol Des* 2014;**28**:711–20.

42. Mendez D, Gaulton A, Bento AP, *et al.* ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res* 2019;**47**:D930–40.

43. Rohrer SG, Baumann K. Maximum unbiased validation (MUV) data sets for virtual screening based on PubChem bioactivity data. *J Chem Inf Model* 2009;**49**:169–84.

44. AIDS antiviral screen data. In: NIH/NCI (ed). 2017.

45. Subramanian G, Ramsundar B, Pande V, *et al.* Computational modeling of beta-secretase 1 (BACE-1) inhibitors using ligand based approaches. *J Chem Inf Model* 2016;**56**:1936–49.

46. Wang R, Fang X, Lu Y, *et al*. The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J Med Chem* 2004;**47**: 2977–80.

47. Martins IF, Teixeira AL, Pinheiro L, *et al*. A Bayesian approach to in silico blood-brain barrier penetration modeling. *J Chem Inf Model* 2012;**52**:1686–97.

48. Tox21 data challenge. *NIH* 2017.

49. Kuhn M, Letunic I, Jensen LJ, *et al*. The SIDER database of drugs and side effects. *Nucleic Acids Res* 2016;**44**:D1075–9.

50. Gayvert KM, Madhukar NS, Elemento O. A data-driven approach to predicting successes and failures of clinical trials. *Cell Chem Biol* 2016;**23**:1294–301.

51. Artemov GN, Bondarenko SM, Shirokova VV, *et al*. Spatial organization of chromosomes in malaria mosquitoes. *Tsitologiia* 2016;**58**:315–9.

52. Tran-Nguyen VK, Jacquemard C, Rognan D. LIT-PCBA: an unbiased data set for machine learning and virtual screening. *J Chem Inf Model* 2020;**60**:4263–73.

53. He S, Zhao D, Ling Y, *et al*. Machine learning enables accurate and rapid prediction of active molecules against breast cancer cells. *Front Pharmacol* 2021;**12**:796534.

54. Wu Z, Jiang D, Hsieh CY, *et al*. Hyperbolic relational graph convolution networks plus: a simple but highly efficient QSAR-modeling method. *Brief Bioinform* 2021;**22**:bbab112.

55. Wang Y, Bryant SH, Cheng T, *et al*. PubChem BioAssay: 2017 update. *Nucleic Acids Res* 2017;**45**:D955–63.

56. Jiang Z, Xu J, Yan A, *et al*. A comprehensive comparative assessment of 3D molecular similarity tools in ligand-based virtual screening. *Brief Bioinform* 2021;**22**:bbab231.

57. Morgan HL. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *J Chem Doc* 2002;**5**:107–13.

58. Luo Y, Zeng R, Guo Q, *et al*. Identifying a novel anticancer agent with microtubule-stabilizing effects through computational cell-based bioactivity prediction models and bioassays. *Org Biomol Chem* 2019;**17**:1519–30.

59. Guo Q, Luo Y, Zhai S, *et al*. Discovery, biological evaluation, structure-activity relationships and mechanism of action of pyrazolo[3,4-b]pyridin-6-one derivatives as a new class of anticancer agents. *Org Biomol Chem* 2019;**17**:6201–14.

60. Moffat JG, Vincent F, Lee JA, *et al*. Opportunities and challenges in phenotypic drug discovery: an industry perspective. *Nat Rev Drug Discov* 2017;**16**:531–43.

61. Malandraki-Miller S, Riley PR. Use of artificial intelligence to enhance phenotypic drug discovery. *Drug Discov Today* 2021;**26**: 887–901.

62. Berg EL. The future of phenotypic drug discovery. *Cell Chem Biol* 2021;**28**:424–30.

63. Guo Q, Zhang H, Deng Y, *et al*. Ligand- and structural-based discovery of potential small molecules that target the colchicine site of tubulin for cancer treatment. *Eur J Med Chem* 2020;**196**:112328.

64. Wang L, Chen L, Yu M, *et al*. Discovering new mTOR inhibitors for cancer treatment through virtual screening methods and in vitro assays. *Sci Rep* 2016;**6**:18987.

65. Wang L, Li Y, Xu M, *et al*. Chemical fragment-based CDK4/6 inhibitors prediction and web server. *RSC Adv* 2016;**6**:16972–81.

66. Wang L, Le X, Li L, *et al*. Discovering new agents active against methicillin-resistant Staphylococcus aureus with ligand-based approaches. *J Chem Inf Model* 2014;**54**:3186–97.

67. Wang L, Chen L, Liu Z, *et al*. Predicting mTOR inhibitors with a classifier using recursive partitioning and naive Bayesian approaches. *PLoS One* 2014;**9**:e95221.

68. Rogers D, Hahn M. Extended-connectivity fingerprints. *J Chem Inf Model* 2010;**50**:742–54.