Meta Learning With Graph Attention Networks for Low-Data Drug Discovery

Qiujie Lv, Guanxing Chen, Ziduo Yang, Weihe Zhong, and Calvin Yu-Chian Chen

***Abstract*— Finding candidate molecules with favorable pharmacological activity, low toxicity, and proper pharmacokinetic properties is an important task in drug discovery. Deep neural networks have made impressive progress in accelerating and improving drug discovery. However, these techniques rely on a large amount of label data to form accurate predictions of molecular properties. At each stage of the drug discovery pipeline, usually, only a few biological data of candidate molecules and derivatives are available, indicating that the application of deep neural networks for low-data drug discovery is still a formidable challenge. Here, we propose a meta learning architecture with graph attention network, Meta-GAT, to predict molecular properties in low-data drug discovery. The GAT captures the local effects of atomic groups at the atom level through the triple attentional mechanism and implicitly captures the interactions between different atomic groups at the molecular level. GAT is used to perceive molecular chemical environment and connectivity, thereby effectively reducing sample complexity. Meta-GAT further develops a meta learning strategy based on bilevel optimization, which transfers meta knowledge from other attribute prediction tasks to low-data target tasks. In summary, our work demonstrates how meta learning can reduce the amount of data required to make meaningful predictions of molecules in low-data scenarios. Meta learning is likely to become the new learning paradigm in low-data drug discovery. The source code is publicly available at: https://github.com/lol88/Meta-GAT.**

***Index Terms*— Drug discovery, few examples, graph attention network, meta learning, molecular property.**

I. Introduction

DRUG discovery is a high-investment, long-period, and high-risk systems engineering [1]. When molecular biology studies have identified an effective target associated with a disease, the subsequent path of drug discovery becomes relatively clear [2]. With the help of various computer-aided virtual screening technologies and high-throughput omics technologies, researchers can integrate the relevant knowledge of computational chemistry, physics, and structural biology to effectively screen and design molecular compounds [3], [4], [5], [6], [7]. The key issue of drug discovery is the screening and optimization of candidate molecules, which must meet a series of criteria: the compound needs to have suitable potential for biological targets, and exhibit good physicochemical properties; absorption, distribution, metabolism, excretion, and toxicity (ADMET); water solubility; and mutagenicity [8], [9]. However, there are usually only a few validated leads and derivatives that can be used for lead optimization [10], [11]. Also, due to the possible toxicity, low activity, and low solubility, there are often only a few real biological data on candidate molecules and analog molecules. The accuracy of the physical chemical properties of candidate molecules directly affects the results of the drug development process. Therefore, researchers have paid more and more attention to accurately predict the physicochemical properties of candidate molecules with low data.

In the past few years, deep learning technology has been implemented to accelerate and improve the drug discovery process [12], [13], [14], [15], and some key advances have been made in molecular property prediction [16], [17], [18], [19], [20], side effect prediction [21], [22], [23], and virtual screening [24], [25]. In particular, the graph neural network (GNN), which can learn the information contained in the nodes and edges directly from the chemical graph structure, has aroused the strong interest of bioinformatics scientists [26], [27], [28], [29]. The performance of deep learning algorithm depends largely on the size of the training data, and a larger sample size usually produces a more accurate model. Given a large amount of labeled data, deep neural networks have enough ability to learn complex representations of inputs [30]. However, this is obviously in contradiction with insufficient data in the initial stage of drug discovery. Due to the scarcity of labeled data, achieving satisfactory results for low-data drug discovery remains a challenge. The paradigm of artificial intelligence for drug discovery has changed: from large-scale sample learning to small sample learning [31], [32], [33].

The human brain's understanding of objective things does not necessarily require large sample training, and it can be learned in many cases based on simple analogies [34], [35], [36]. DeepMind explores how the brain learns with few experience, that is, "meta learning" or "learning to learn" [37]. The understanding of meta learning mode is one of the important ways to achieve general intelligence.

Manuscript received 24 November 2021; revised 21 May 2022, 16 October 2022, and 26 December 2022; accepted 24 February 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62176272 and in part by the China Medical University Hospital under Grant DMR-112-085. *(Corresponding author: Calvin Yu-Chian Chen.)*

Qiujie Lv, Guanxing Chen, Ziduo Yang, and Weihe Zhong are with the Artificial Intelligence Medical Research Center, School of Intelligent Systems Engineering, Shenzhen Campus of Sun Yat-sen University, Shenzhen, Guangdong 518107, China.

Calvin Yu-Chian Chen is with the Artificial Intelligence Medical Research Center, School of Intelligent Systems Engineering, Shenzhen Campus of Sun Yat-sen University, Shenzhen, Guangdong 518107, China, also with the Department of Medical Research, China Medical University Hospital, Taichung 40447, Taiwan, and also with the Department of Bioinformatics and Medical Engineering, Asia University, Taichung 41354, Taiwan (e-mail: chenyuchian@mail.sysu.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2023.3250324.

Digital Object Identifier 10.1109/TNNLS.2023.3250324

Biswas et al. [38] developed UniRep for protein engineering to efficiently use resource-intensive high-fidelity assays without sacrificing throughput, and subsequent low-N supervision then identifies improvements to the activity of interest. Liu et al. [39] from the Chinese Academy of Sciences have established a complete and effective screening method for disease target markers based on few examples (or even one sample). Lin et al. [40] proposed a prototypical graph contrastive learning (PGCL) method for learning graph representation, which improved the results of molecular property prediction. Yu and Tran [25] proposed an XGBoost-based fitted Q iteration algorithm with fewer training data for finding the optimal structured treatment interruption (STI) strategies for HIV patients. They have made certain explorations and attempts in the field of drug virtual screening and combination drug prediction based on the few examples learning method [41], [42]. The abovementioned work is a useful attempt by meta learning for few samples learning problems, indicating that the meta learning method has the potential to be a useful tool in drug discovery and other bioinformatics research fields.

Meta learning uses meta knowledge to reduce requirement for sample complexity, thus solving the core problem of minimizing the risk of unreliable experience. However, the molecular structure is usually composed of the interaction between atoms and complex electronic configurations. Even small changes in the molecular structure may lead to completely opposite molecular properties. The model learns the complexity of molecular structure, which requires that the model should perfectly extract the local environmental influence of neighboring atoms on the central atom and the rich nonlocal information contained between pairs of atoms that are topologically far apart. Therefore, meta learning for low-data drug discovery is highly dependent on the structure of the network and needs to be redesigned for widely varying tasks.

Meta learning has made some representative attempts to predict molecular properties. Altae-Tran et al. [43] introduced an architecture of iteratively refined long short-term memory (IterRefLSTM) that uses IterRefLSTM to generate dually evolved embeddings for one-shot learning. Adler et al. [44] proposed cross-domain Hebbian ensemble few-shot learning (CHEF), which achieves representation fusion by an ensemble of Hebbian learners acting on different layers of a deep neural network. The Meta-molecular graph neural network (MGNN) leverages a pretrained GNN and introduces additional self-supervised tasks, such as bond reconstruction and atom-type prediction to be jointly optimized with the molecular property prediction tasks [45]. Meta-MGNN, CHEF, obtains meta knowledge through pretraining on a large-scale molecular corpus and additional self-supervised model parameters. IterRefLSTM trains the memory-augmented model, which restricts the model structure and can only be used in specific domain scenarios. How to represent molecular features effectively and how to capture common knowledge between different tasks are great challenges that exist in meta learning.

In this work, we propose a meta learning architecture based on graph attention network, Meta-GAT, to predict the biochemical properties of molecules in low-data drug discovery. The graph attention network captures the local effects of atomic groups at the atomic level through the triple attentional mechanism, so that the GAT can learn the influence of the atom group on the properties of the compound. At the molecular level, GAT treats the entire molecule as a supervirtual node that connects every atom in a molecule, implicitly capturing the interactions between different atomic groups. The gated recurrent unit (GRU) hierarchical model mainly focuses on abstracting or transferring limited molecular information into higher-level feature vectors or meta knowledge, improving the ability of the GAT to perceive chemical environment and connectivity in molecules, thereby efficiently reducing sample complexity. This is very important for low-data drug discovery. Meta-GAT benefits from meta knowledge and further develops a meta learning strategy based on bilevel optimization, which transfers meta knowledge from other attribute prediction tasks to low-data target tasks, allowing the model to quickly adapt to molecular attribute predictions with few examples. Meta-GAT achieved accurate prediction of few examples's molecular new properties on multiple public benchmark datasets. These advantages indicate that Meta-GAT is likely to become a viable option for low-data drug discovery. In addition, the Meta-GAT code and data are open source at https://github.com/lol88/Meta-GAT, so that the results can be easily replicated.

Our contributions can be summarized as follows.

1) We create a chemical tool to predict multiple physiological properties of new molecules that are invisible to the model. This tool could push the boundaries of molecular representation for low-data drug discovery.
2) The proposed Meta-GAT captures the local effects of atomic groups at the atomic level through the triplet attentional mechanism and can also model global effects of molecules at the molecular level.
3) We propose a meta learning strategy to selectively update parameters within each task through a bilevel optimization, which is particularly helpful to capture the generic knowledge shared across different tasks.
4) Meta-GAT demonstrates how meta learning can reduce the amount of data required to make meaningful predictions of molecules in low-data drug discovery.

II. Methods

In this section, we first briefly introduce the mathematical formalism of Meta-GAT and then introduce the meta learning strategy and graph attention network structure. Finally, the parameters and details of the model training are shown. Fig. 1 shows the overall architecture of Meta-GAT for low-data drug discovery.

A. Problem Formulations

Consider several common drug discovery tasks T, such as predicting the toxicity and side effects of new molecules, x is the compound molecule to be measured, and the label y is the binary experimental label (positive/negative) of the molecular properties. Suppose that all some potential laws considered

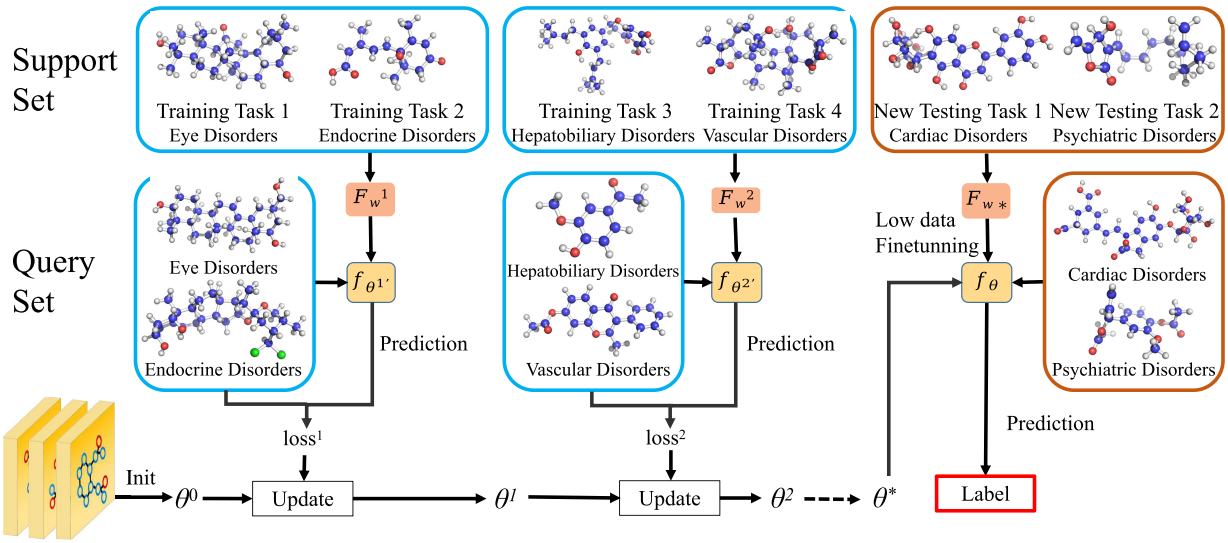

Fig. 1. Meta learning framework for few examples molecular property prediction. The blue box and the orange box represent the data flow in the training phase and the test phase, respectively.

by the model are called hypothesis space H. h is the optimal hypothesis from x to y. The expected risk $R(h)$ represents the prediction ability of the decision model for all samples. The empirical risk $R(h_I)$ represents the predictive ability of the model for samples in the training set by calculating the average value of the loss function, and I represents the number of samples in the training set. The empirical risk $R(h_I)$ is used to estimate expected risk $R(h)$. In real-world applications, only a few examples are available for a property prediction task of a new molecule, that is, $I \rightarrow$ few. According to the empirical risk minimization theory, if only a few training samples can be provided, which makes the empirical risk $R(h_I)$ far from the approximation of the expected risk $R(h)$, the obtained empirical risk minimizer is unreliable [46]. The learning challenge is to obtain a reliable empirical risk minimization from a few examples. This minimizer results in $R(h_I)$ approaching the optimal $R(h)$, as shown in the following equation:

$$\mathbb{E}[R(h_{I\rightarrow \text{few}}) - R(h)] = 0. \tag{1}$$

The empirical risk minimization is closely related to sample complexity. Sample complexity refers to the number of training samples required to minimize the loss of empirical risk $R(h_I)$. According to VapnikâĂŞ Chervonenkis (VC), when samples are insufficient, H needs less complexity, so that the few examples provided are sufficient for compensation. We use meta knowledge w to reduce the complexity of learning samples, thus solving the core problem of minimizing the risk of unreliable experience.

B. Meta Learning

Meta learning, also known as learning to learn, means learning a learning experience by systematically observing how the model performs in a wide range of learning tasks. This learning experience is called meta knowledge w. The goal of meta learning is to find the w shared across different tasks, so that the model can quickly generalize to new tasks that contain only a few examples with supervised information. The difference between meta learning and transfer learning is that transfer learning is usually fitting the distribution of one data, while meta learning is fitting the distribution of multiple similar tasks. Therefore, the training samples of meta learning are a series of tasks.

Model-agnostic meta-learning (MAML) [47] is used as a base meta learning algorithm for the Meta-GAT framework. Meta-GAT selectively updates parameters within each task through a bilevel optimization and transfers meta knowledge to new tasks with few label samples, as shown in Fig. 1. Bilevel optimization means that one optimization contains the another optimization as a constraint. In inner-level optimization, we hope to learn a general meta knowledge w from the support set of training tasks, so that the loss of different tasks can be as small as possible. The inner level optimization phase can be formalized, as shown in (3). In outer-level optimization, Meta-GAT calculates the gradient relative to the optimal parameter in the query set of each task and calculates the minimum total loss value of all training tasks to optimize the w parameter, thereby reducing the expected loss of the training task, as shown in (2). Algorithm 1 shows the specific algorithm details

$$w^* = \underset{w}{\text{argmin}} \sum_{i=1}^{M} \mathcal{L}_{f_\theta}^{\text{meta}}\left(\theta^{*(i)}(w), D_{\text{train}}^{q}\right) \tag{2}$$

$$\theta^{*(i)}(w) = \underset{\theta}{\text{argmin}}\, \mathcal{L}_{f_\theta}^{\text{task}}\left(\theta, w, D_{\text{train}}^{s(i)}\right) \tag{3}$$

where $\mathcal{L}_{\text{meta}}$ and $\mathcal{L}_{\text{task}}$ refer to the outer and inner objectives, respectively. i represents the ith training task.

Specifically, first, the train tasks T_{train} and test tasks T_{test} are extracted from a set of multitask T for drug discovery, where each task has a support set D^s and a query set D^q. Meta-GAT uses a large number of training tasks T_{train} to fitting the distribution of multiple similar tasks T. Second, Meta-GAT sequentially iterates a batch of training tasks, learns task-specific parameters, and tries to minimize the loss using

Algorithm 1 Pseudocode of Meta-GAT for Low-Data Drug Discovery

Require: A set of tasks for predicting molecular properties T;

Ensure: GAT parameters θ, step sizes α, β;

Randomly initialize θ;
while not done **do**
  Sample batch of tasks $T_{train} \sim T$;
  for all T_{train} **do**
    Sample m examples $D^s_{train} = \{D_1, D_2, \ldots, D_m\} \in D_{train}$;
    Evaluate $\nabla_\theta \mathcal{L}_{T_{train}}(f_\theta)$ by $y^s_{train} = GAT(D^s_{train}, \theta)$;
    Compute adapted parameters with gradient descent: $\theta'_i = \theta - \alpha \nabla_\theta \mathcal{L}_{T_{train}}(f_\theta)$;
    Sample n examples $D^q_{train} = \{D_1, D_2, \ldots, D_n\} \in D_{train} - D^s_{train}$;
    Evaluate $\mathcal{L}'_{T_{train}}$ by $y^q_{train} = GAT(D^q_{train}, \theta')$;
  end for
  Updat $\theta \leftarrow \theta - \beta \nabla_\theta \sum_{T_{train} \sim p(\mathcal{T})} \mathcal{L}'_{T_{train}}$
end while
Sample batch of tasks $T_{test} \sim T - T_{train}$
for all T_{test} **do**
  Sample k examples $D^s_{test} = \{D_1, D_2, \ldots, D_k\} \in D_{test}$
  // Similar to the training phase
  Evaluate and Compute adapted parameters with gradient descent
  Updat θ
  Sample j examples $D^q_{test} = \{D_1, D_2, \ldots, D_j\} \in D_{test} - D^s_{test}$
  $y^q_{test} = GAT(D^q_{test}, \theta)$
end for

gradient descent. The corresponding optimal parameters θ are obtained from each task's support set, as shown in (4). This parameters are not assigned to θ directly, but are cached

$$\theta'_i = \theta - \alpha \nabla_\theta \mathcal{L}_{T_{\text{train}}}(f_\theta) \tag{4}$$

$$\theta \leftarrow \theta - \beta \nabla_\theta \sum_{T_{\text{train}} \sim p(\mathcal{T})} \mathcal{L}'_{T_{\text{train}}}\left(f_{\theta'_i}\right). \tag{5}$$

Then, the outer-level optimization learns w, such that it produces models f_θ [see (5)]. Each task's query set is used to obtain a gradient value on each task-specific parameter θ. The vector sum of the gradient values, obtained from the above batch task query set, is used to update the parameters of the meta learner. The model continues iterating up to a preset number of times, and the best meta model is selected based on the query set

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \underset{T_{\text{test}} \in T}{\mathbb{E}} \mathcal{L}\left(\theta, w, D^s_{\text{test}}\right). \tag{6}$$

Finally, in the testing phase, Meta-GAT, which has learned meta knowledge w, learns the specificity of the new test task through a few inner optimizations on the support set of the new task, as shown in (6). Note that the model parameters θ exist separately or within meta knowledge w. We evaluate the performance of the model by the accuracy of θ on the query set of the new task. In the process of learning new tasks, the model benefits from meta knowledge to reduce the requirement of sample complexity, so as to realize the optimization strategy, which is faster to search the parameterized θ of the hypothesis $h \in H$ in the hypothesis space H.

Meta-GAT essentially searches for a hypothesis that is better for all tasks of predicting the properties of drug molecules. Therefore, when updating parameters, it combines the loss of all tasks on the query set to specify the gradient update. The parameter θ obtained in this way is already an approximate optimal hypothesis on the new task, and the optimal hypothesis can be reached with few inner iterations.

Our Meta-GAT uses meta knowledge w to guide the model to search for the parameter θ that approximates the optimal hypothesis h in the hypothesis space H, leading to the minimization of empirical risk. The meta knowledge w is obtained through limited analysis of new molecules and prior knowledge analysis of many similar molecules. The meta knowledge w changes the search strategy by providing a better parameter initialization or providing a search direction. Meta-GAT was rapidly migrated from a better hypothesis site to the new task through several inner optimizations on fewer new molecular instances, and then, the percentage of correctly assigned molecules with/without toxicity was increased.

TABLE I
CODED INFORMATION FOR ATOMIC AND BOND FEATURES

type	feature name	size	description
atom feature	atom symbol	16	[B, C, N, O, F, Si, P, S, Cl, As, Se, Br, Te, I, At, metal] (one-hot)
	degree	6	number of covalent bonds [0,1,2,3,4,5] (one-hot)
	formal charge	1	electrical charge (integer)
	radical electrons	1	number of radical electrons (integer)
	hybridization	6	[sp, sp2, sp3, sp3d, sp3d2, other] (one-hot)
	aromaticity	1	whether the atom is part of an aromatic system [0/1] (one-hot)
	hydrogens	5	number of connected hydrogens [0,1,2,3,4] (one-hot)
	chirality	1	whether the atom is chiral center [0/1] (one-hot)
	chirality type	2	[R, S] (one-hot)
bond feature	bond type	4	[single, double, triple, aromatic] (one-hot)
	conjugation	1	whether the bond is conjugated [0/1] (one-hot)
	ring	1	whether the bond is in ring [0/1] (one-hot)
	stereo	4	[StereoNone, StereoAny, StereoZ, StereoE] (one-hot)

C. Molecular Graph Representation

Molecules are coded into graphs with node features, edge features, and adjacency matrices for input into the graph network. We use a total of nine atomic features and four bond features to characterize the molecular graph's structure (see Table I). Atom features include hybridization, aromaticity, chirality, and so on, and bond features include type, conjugation, ring, and so on. Molecular structures usually involve atomic interactions and complex electronic structures, and bond features contain rich information about molecular scaffolds and conformational isomers. The encoded molecular graph can implicitly capture the local environment of the molecule and the key interactions between atoms and electrons

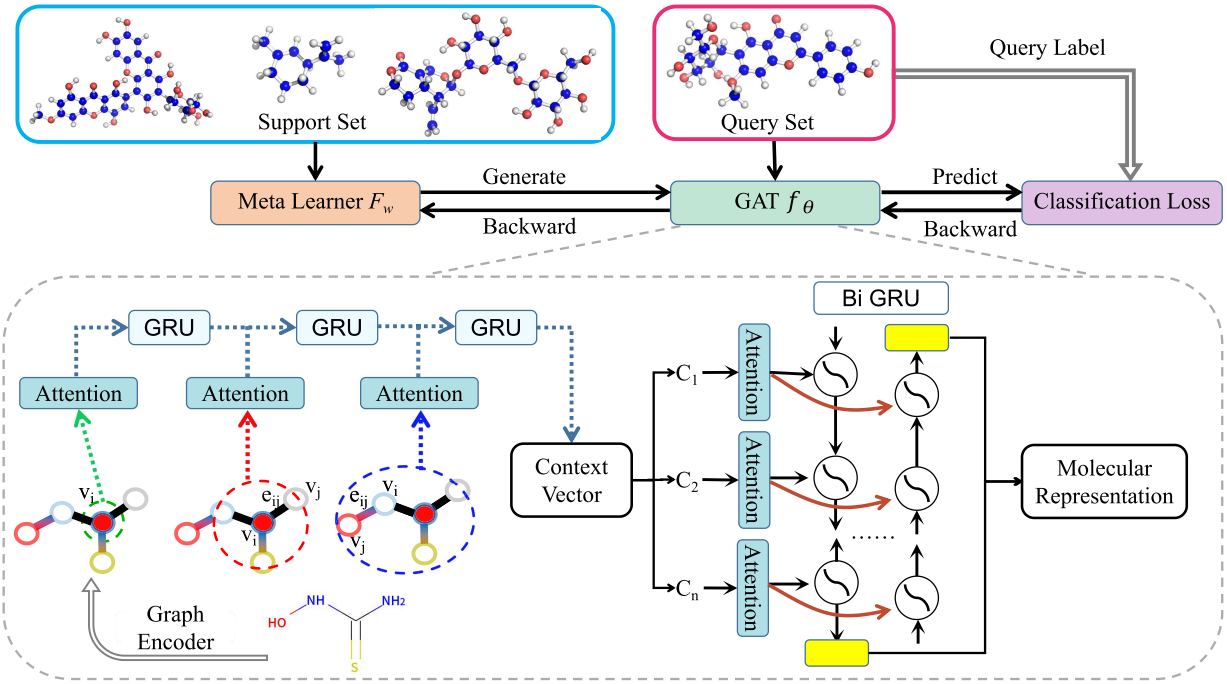

Fig. 2. Schematic of graph attention network architecture for meta learning.

and provide insight into the edge characteristics of molecular bonds.

D. Graph Attention Network

GNN has made substantial progress in the field of chemical informatics. It has the extraordinary capacity to learn the intricate relationships between structures and properties [16], [48], [49], [50]. The attention mechanism has proved its outstanding performance in predicting molecular properties.

Molecular structure involves the spatial position of atoms and the types of chemical bonds. Topologically adjacent nodes in molecules have a greater chance of interacting with each other. In some cases, they can also form functional groups that determine the chemical properties of the molecule. In addition, pairs of atoms that are topologically far apart may also have significant interactions, such as intramolecular hydrogen bonds. Our graph attention network extracts insights on molecular structure and features from both local and global perspectives, as shown in Fig. 2. GAT captures the local effects of atomic groups at the atomic level through the attentional mechanism and can also model global effects of molecules at the molecular level.

The molecule $G = (v, e)$ can be defined as a graph composed of a set of atoms (nodes) v and a set of bonds (edges) e. Similar to the previous study, we encode chemical information including nine atomic features and four bond features into the molecular graph as the input of graph attention network. For the local environment within the molecule, previous graph networks only aggregate the neighbor nodes' information, which may lead to insufficient edge (bond) information extraction. Our GAT gradually aggregates the triplet embedding of target node v_i, neighbor node v_j, and edge e_{ij} through the triple attention mechanism.

Specifically, GAT first performs linear transformation and nonlinear activation on the neighbor nodes' state vectors v_i,v_j and their edge hidden states e_{ij} to align these vectors to the same dimension, and concatenate them into triplet embedding vectors. Then, h_{ij} is normalized by the softmax function over all neighbor nodes to get attention weights a_{ij}. Finally, the node hidden state and edge hidden state elementwise multiplied by neighbor node representation, and the information of neighbors (including neighbor nodes and edges) is aggregated according to the attention weight to obtain the context state c_i of the atom i. The formula is shown below

$$h_{ij} = \text{LeakyReLU}\left(W \cdot \left[v_i, e_{ij}, v_j\right]\right) \tag{7}$$

$$a_{ij} = \text{softmax}\left(h_{ij}\right) = \frac{\exp\left(h_{ij}\right)}{\sum_{j \in N(i)} \exp\left(h_{ij}\right)} \tag{8}$$

$$c_i = \sum_{j \in N(i)} a_{ij} \cdot W \cdot \left[e_{ij}, v_j\right] \tag{9}$$

where $N(i)$ is the set of neighbor nodes for node i. W is the trainable weight matrix. Then, the GRU is used as a message transfer function to fuse messages with a farther radius to generate a new context state, as shown in Fig. 2 (bottom left). As the time step t increases, messages of nodes and edges in the range centered on node I and whose radius increases with t are collected successively to generate new states h_i^t, which is computed by

$$h_i^t = \text{GRU}\left(h_i^{t-1}, c_i^{t-1}\right). \tag{10}$$

In order to include more global information from the molecule, GAT aggregates the atomic level representation through the readout function, which treats the entire molecule as a supervirtual node that connects every atom in a molecule. We use the bidirectional GRU (BiGRU) with attention to

TABLE II
DETAILED DESCRIPTION OF THE BENCHMARK DATASETS

Category	Datasets	Data type	Metrics	Tasks	Molecules
Physiology	Tox21	SMILES	ROC-AUC	12	7831
	SIDER	SMILES	ROC-AUC	27	1427
Biophysics	MUV	SMILES	ROC-AUC	17	93087
Quantum mechanics	QM9	SMILES	MAE	12	133885

connect node features with historical information from two directions, so as to obtain a graph-level (molecular) representation M. The update gate in the GRU recurring network cell ensures that information is effectively transmitted to distant nodes, while the reset gate helps to filter out information that is not relevant to the learning task. Moreover, different attention weights can focus on the implicit interactions between distant atoms and extract more information related to learning tasks. The formula of readout function is shown below

$$M = \left\{\overleftarrow{\mathrm{GRU}}\left[\mathrm{att}\left(\overleftarrow{h^t}\right)\right]\overrightarrow{\mathrm{GRU}}, \left[\mathrm{att}\left(\overrightarrow{h^t}\right)\right]\right\} \tag{11}$$

where att uses the same attention mechanism as before. The final molecular representation M is used by the classifier for molecular property prediction.

GAT learns the contextual representation of each atom by aggregating the triple information from the atom feature, the neighboring atoms feature information, and the feature information of the connecting bond through the message passing mechanism and attention mechanism. Then, the context representations of atoms are gradually aggregated by BiGRU based on the attention mechanism to generate a global state vector for the entire molecule. The final vector representation is a high-quality descriptor of molecular structure information, which reduces the difficulty of learning the unsupervised information in molecular graph by the meta learning model.

E. Datasets

We report the experimental results of the Meta-GAT model on multiple public benchmark datasets. Table II shows the detailed information of the benchmark dataset, including categories, tasks, and the number of molecules. All datasets are available for download at the public project MoleculNet [51].

F. Model Implementation and Evaluation Protocols

Meta-GAT performs linear transformation and nonlinear activation from both atomic features and neighbor features to unify vector lengths [see (7)]. Then, the triplet embedding vectors of atoms are aligned using a fully connected layer, and the attention weights are calculated using softmax [see (8)]. The weighted sum of the atoms current state vectors is taken as the attention context vector of a single atom [see (9)], which is fed into the GRU along with the current state vector [see (10)]. This process is repeated twice to generate a new state vector for each atom. Finally, we assume that the molecular embedding is a virtual node embedding, so that the whole molecule can be embedded as if it was a single atom. Similar to the above process, we combine the context state vectors of each atom from both directions into a BiGRU [see (11)]. This process is also repeated twice to obtain a graph representation at the molecular level.

Meta-GAT is implemented using the pytorch framework and uses the Adam optimizer [52] with a 0.001 learning rate for gradient descent optimization. The learning rate for inner iterations is 0.1. Generates information around atoms with a radius of 2. The output unit of the full connection layer is 200. Both GRU and BiGRU also have 200 hidden units. Gradient descent is performed five times in each iteration of the training and testing phase, $\alpha, \beta = 5$. During training, 10 000 episodes are generated $K = N_{\mathrm{pos}} + N_{\mathrm{neg}}(N_{\mathrm{pos}}, N_{\mathrm{neg}} \in [1, 5, 10])$ in N-way K-shot. N_{pos} and N_{neg}, respectively, represent the number of positive and negative examples in the support set. We use CrossEntropyLoss as the loss function of the classification task. When training Meta-GAT on the task of molecular biochemical property prediction, multiple tasks are divided into two disjoint task sets, training task and test task. The training/testing division method for each dataset is the same as the comparison experiment. During the prediction phase, a batch of support sets with size $N_{\mathrm{pos}} + N_{\mathrm{neg}}$ and a batch of query sets with size $K = 128$ are randomly sampled from a task's dataset. For each test task, 20 independent runs were performed based on different random seeds, and the average value of area under the receiver operating characteristic curve (ROC-AUC) was calculated in the report of experimental section.

In addition, we also analyze the total training time, meta training time, meta testing time, number of multiply–accumulate operations (MACs), and model size to evaluate the computational complexity of the proposed method. Meta-GAT consists of two steps, the meta training phase and the meta testing phase. Total training time refers to the cost of stabilizing the performance of Meta-GAT on new tasks. Meta training time is the cost of one iteration in the meta training phase. Meta testing time refers to the cost of Meta-GAT learning the prediction task of molecular new property with few samples in the meta testing stage. Within an iteration, both the support set and the query set participate in the model forward calculation and perform one or more iterations of gradient descent. The cost of one iteration in meta training stage, namely, meta training time, is $2N * \alpha$ times of the model's forward calculation time, while meta testing time is $2 * \beta$ times longer than the model's forward computation time. GeForce RTX 2060 is used in this experiment, and N is 8, and α and β are 5. The average forward computation time of Meta-GAT on the Tox21 and Side Effect Resource (SIDER) datasets is 14.84 and 23.08 ms, respectively. Therefore, the meta training time is about 1187.2 and 1846.4 ms, the meta testing time is about 148.4 and 230.8 ms, and the total training time is about 7.3 and 6 h, respectively. The MACs of Meta-GAT are 3.17E9, and the model size is 4.8 M. The training time of meta-learning-based GAT is longer than that of GAT, but the size of obtained prediction model is the same as GAT.

We compare Meta-GAT with multiple baseline models, including random forest (RF) [53], Graph Conv [54], Siamese [55], MAML [47], attention LSTM (attnLSTM) [43], IterRefLSTM [43], Meta-MGNN [45], edge-labeling GNN (EGNN) [56], PreGNN [57], prototypical networks (PN) [58], CHEF [44], attentive fingerprint (Attentive FP) [16], communicative message passing neural network (CMPNN)

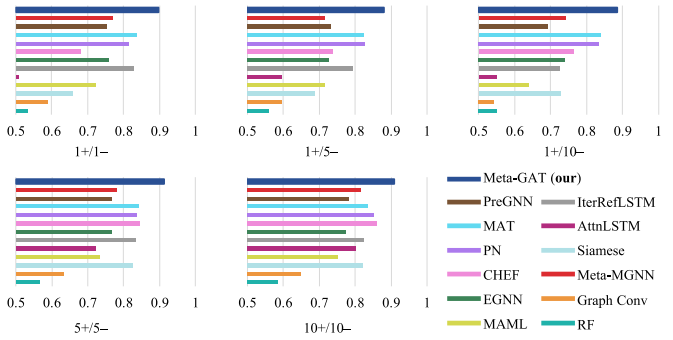

Fig. 3. ROC-AUC scores of Meta-GAT and previous models in the Tox21 few examples prediction task. 1+/5− represent the number of positive and negative examples is 1 and 5, respectively.

[59], Weave [60], continuous and data-driven descriptors (CDDD) [49], DeepTox [61], molecule attention transformer (MAT) [62], molecular prediction model fine-tuning (MolPMoFiT) [63], N-Gram [64], molecular to context vector (Mol2Context-vec) [17], and triplet message network (TrimNet) [29]. In the reproducibility settings, Siamese, MAML, AttnLSTM, IterRefLSTM, Meta-MGNN, EGNN, PN, and CHEF are based on meta learning methods, using the same training settings as Meta-GAT. RF and Graph Conv are single-task models. DeepTox, Weave, Attentive FP, and CMPNN are multitask models. For each assay prediction task, randomly select $N_{\text{pos}} + N_{\text{neg}}$ samples as the support set and 128 samples as the query set. Repeat this process 20 times, and calculate the final average value as the model result. MAT, PreGNN, CDDD, MolPMoFiT, and N-Gram, is a pretrained GNN model that uses self-supervised learning on a large-scale molecular corpus, resulting in better parameter initialization. Similarly, 128 samples were randomly collected for testing and repeated 20 times to avoid the randomness of model testing.

III. Results and Discussion

A. Tox21

The "Toxicology in the 21st Century" (Tox21), collected by the 2014 Tox21 Data Challenge, is a public database containing 12 assays measures the toxicity of biological target. We treat each assay as a single task. The first nine assays were used for training, including NR-AR, NR-AR-LBD, NR-AHR, NR-Aromatase, NR-ER, NR-ER-LBD, NR-PPAR-Gamma, SR-ARE, and SR-ATAD5. The last three assays were used for testing, including SR-HSE, SR-MMP, and SR-P53.

Meta-GAT is compared with other 12 models, and the experimental results are shown in Fig. 3. The numbers of positive and negative samples in the support set are both increased from 1 to 10, and the improvement of model performance is not obvious. However, the change in the ratio of positive and negative samples has great influence on the model performance. Interestingly, when there are only a few examples, the balanced ratio of positive and negative samples in the support sets may be more important than the increase in the number. To some extent, the ratio of positive and negative samples in the support set represents the distribution of task. A balanced ratio of positive and negative samples may better guide the model to search for the parameters of the optimal

TABLE III
Scores for Consistency Checks on the Tox21 Dataset Using Kappa and Paired Wilcoxon Tests

	SR-HSE	SR-MMP	SR-p53
p-value Graph Conv	6.23E-08	3.66E-10	3.80E-08
p-value Siamese	6.86E-06	6.99E-05	7.97E-06
p-value MAML	4.07E-04	1.29E-06	4.84E-07
p-value AttnLSTM	2.48E-05	5.70E-05	5.82E-06
p-value IterRefLSTM	5.93E-06	1.69E-11	6.03E-06
p-value EGNN	9.76E-06	8.87E-05	4.89E-05
p-value CHEF	1.56E-05	1.54E-05	5.82E-06
p-value PN	1.26E-04	5.94E-05	2.47E-05
p-value MAT	3.28E-04	2.96E-07	9.45E-05
p-value PreGNN	1.41E-06	4.55E-06	3.96E-05
p-value Meta-MGNN	7.90E-07	4.06E-06	9.89E-05
kappa	0.6718	0.8301	0.8562

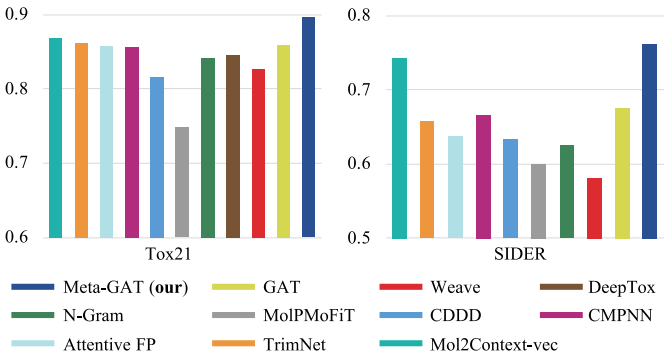

Fig. 4. Performance comparison of Meta-GAT with other representative molecular models.

hypothesis, making the Meta-GAT model easier to learn the meta knowledge of binary classification. Meta-GAT has shown impressive performance in toxicity assay tasks with few data.

We used Kappa and paired Wilcoxon Test to conduct consistency checks on the three test tasks of Tox21, and the results are shown in Table III. Kappa analysis is used to evaluate the consistency degree between predicted results of Meta-GAT and actual measured results. The paired Wilcoxon Test in nonparametric test is used to test whether the distribution of the predictions (independent samples) produced by the two models is equal. It is not limited by the data distribution, the test conditions are relatively loose, and it can be applied to the overall unknown samples. Wilcoxon *p-value* < 0.05 indicated that the distribution of Meta-GAT predicted results was different from that of other models. The results of Kappa analysis show that SR-HSE and real measurement results are highly consistent within the allowable error range, and SR-MMP and SR-p53 are extremely consistent. These statistical tests indicate that the prediction results of Meta-GAT can replace real measurements within the allowable error range.

In addition, Fig. 4 (left) also shows the performance comparison of Meta-GAT with other representative molecular models. We observe that Meta-GAT still achieves state-of-the-art (SOTA) performance compared with fully supervised models. Self-supervised models (CDDD, N-Gram, MolPMoFiT, and Mol2Context-vec) pretrain models from unlabeled large datasets and then fine-tune models on specific target datasets. Due to its powerful feature transfer capability, its model outperforms multitask models that only use the target dataset.

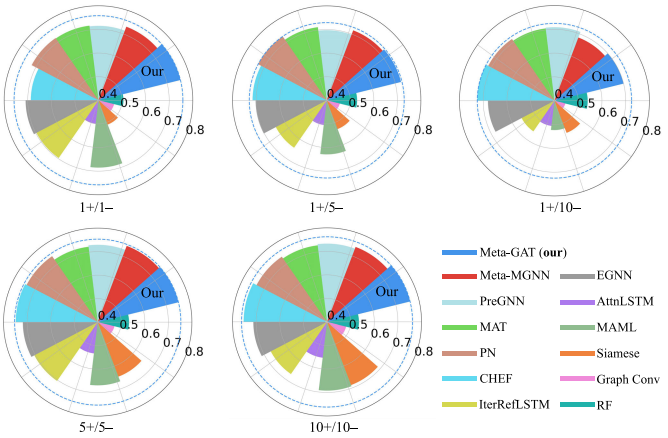

Fig. 5. ROC-AUC scores of Meta-GAT and previous models in the SIDER few examples prediction task. 1+/5− represent the number of positive and negative examples is 1 and 5, respectively.

The graph attention network introduces an attention mechanism by assigning different weights to different nodes, and the corresponding graph convolution operation aggregates the weighted sum of the atoms local information together, which can force the model to learn the most meaningful neighbors and local environments part. Compared with other common GCN architectures, graph attention networks (GAT, Attentive FP, and TrimNet) tend to achieve better performance. Overall, Meta-GAT shows a powerful improvement over the existing baseline model, indicating that meta learning method may be a better solution for low-data drug discovery.

B. SIDER

The SIDER is a public database containing 1427 marketed drugs and their adverse drug reactions [65]. According to the MedDRA classifications, drug side effects are grouped into 27 systemic organ classes. Among them, "renal and urinary disorders" (RUD), "pregnancy, puerperium and perinatal conditions" (PPPC), "ear and labyrinth disorders" (ELD), "cardiac disorders" (CD), "nervous system disorders" (NSD), and "injury, poisoning and procedural complications" (IPPC), six indications were used for testing, and the remaining 21 indications were used for training.

The performance comparison of Meta-GAT with other few examples methods is shown in Fig. 5. The meta learning model still shows a strong improvement in this set, demonstrating the potential of meta learning in few examples molecular property prediction tasks. As is shown before, a balanced ratio of positive and negative samples helps to further improve performance. The graph attention mechanism introduced in Meta-GAT can focus on task-related information from the neighborhood, which helps to achieve accurate iteration of few examples. It can be observed that the GNN based on meta learning (EGNN, Meta-MGNN, and Meta-GAT) obtains more advanced model performance than other meta learning methods (Siamese, MAML, AttnLSTM, and IterRefLSTM). Fig. 4 (right) shows that Meta-GAT achieves SOTA performance compared with fully supervised representative models. In addition, the distribution of Meta-GAT predicted results was

TABLE IV
SCORES FOR CONSISTENCY CHECKS ON THE SIDER DATASET USING KAPPA AND PAIRED WILCOXON TESTS

	RUD	PPPC	ELD	CD	NSD	IPPC
p-value Graph Conv	4.22E-07	4.59E-05	6.44E-05	7.29E-05	5.93E-06	9.76E-03
p-value Siamese	1.16E-05	4.46E-04	4.92E-03	9.49E-04	6.92E-05	9.59E-05
p-value MAML	1.26E-04	2.22E-02	3.96E-02	4.07E-04	9.20E-06	6.07E-03
p-value AttnLSTM	8.20E-05	3.74E-02	7.06E-03	1.62E-04	4.52E-06	8.58E-04
p-value IterRefLSTM	2.04E-05	2.85E-02	6.04E-04	2.61E-04	5.93E-06	2.70E-03
p-value EGNN	5.70E-05	9.44E-03	2.02E-03	1.59E-03	9.55E-06	9.41E-04
p-value CHEF	1.94E-04	3.86E-04	2.70E-03	6.56E-03	6.79E-05	6.49E-03
p-value PN	5.04E-04	3.83E-03	6.04E-04	3.76E-03	1.82E-06	1.48E-04
p-value MAT	8.58E-04	3.08E-03	3.86E-04	4.89E-05	9.89E-05	5.04E-04
p-value PreGNN	4.43E-03	8.58E-04	8.53E-04	1.47E-05	5.57E-06	3.50E-03
p-value Meta-MGNN	3.04E-03	2.80E-02	3.59E-04	3.59E-04	3.96E-05	4.63E-03
kappa	0.8437	0.6208	0.7031	0.8678	0.7151	0.7936

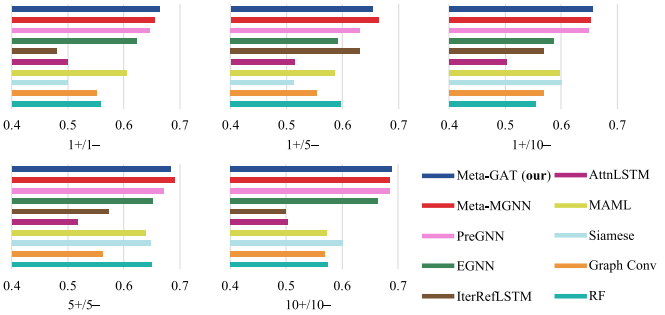

Fig. 6. ROC-AUC Scores of Meta-GAT and previous models in the MUV few examples prediction task. 1+/5− represent the number of positive and negative examples is 1 and 5, respectively.

different from that of other models (Wilcoxon p-$value < 0.05$) for the six indications in the SIDER dataset (see Table IV). PPPC, ELD, NSD, IPPC, and real measurement results are highly consistent within the allowable error range, while RUD and CD are extremely consistent.

C. MUV

The maximum unbiased validation (MUV) dataset contains 17 binary classification tasks for more than 90 000 molecules and is specifically designed to be challenging for standard virtual screening [51], [66]. The positives examples are selected to be structurally distinct from one another. MUV is a best-case scenario for baseline machine learning (since each data point is maximally informative) and a worst case test for the low-data methods, since structural similarity cannot be effectively exploited to predict behavior of new active molecules [43].

The first 12 assays were used for training. The five assays, MUV-832, MUV-846, MUV-852, MUV-858, and MUV-859, were used for model test. Fig. 6 reports the overall performance of all methods on the MUV dataset. Experimental results show that Meta-GAT outperforms other baseline models. In terms of average improvement, for one-shot learning, the average improvements are +0.72%. The value equals 0.39% for five-shot learning. Both Meta-MGNN and PreGNN provide considerable performance, with an average ROC-AUC of 0.6451 and 0.6554 on the test task set, respectively, which are slightly worse than that of Meta-GAT (ROC-AUC = 0.6626). Note that Meta-MGNN and PreGNN

TABLE V

SCORES FOR CONSISTENCY CHECKS ON THE MUV DATASET USING KAPPA AND PAIRED WILCOXON TESTS

	MUV-832	MUV-846	MUV-852	MUV-858	MUV-859
p-value Graph Conv	8.90E-05	1.92E-06	2.03E-07	1.76E-09	4.49E-06
p-value Siamese	1.76E-05	9.14E-09	1.26E-05	7.37E-06	3.70E-06
p-value MAML	3.77E-05	6.35E-07	6.38E-06	2.58E-05	2.69E-07
p-value AttnLSTM	8.45E-07	2.16E-06	2.64E-06	4.60E-05	9.84E-10
p-value IterRefLSTM	3.79E-05	2.32E-07	3.80E-08	6.92E-04	5.67E-08
p-value EGNN	1.65E-04	3.64E-06	4.64E-07	3.78E-06	2.75E-06
p-value PreGNN	1.85E-04	5.38E-06	7.24E-05	1.43E-09	7.43E-06
p-value Meta-MGNN	2.44E-07	4.66E-08	2.35E-09	4.88E-05	5.04E-06
kappa	0.6376	0.5251	0.4382	0.5934	0.4735

TABLE VI

COMPARISON OF PREDICTIVE PERFORMANCES (MAE) ON THE QM9 DATASET QUANTUM PROPERTIES. NOTE THAT FOR MAE, LOWER VALUE INDICATES BETTER PERFORMANCE

Method \ Task	LUMO	G	Cv
RF	0.01842	18.732	1.683
Graph Conv	0.00921	3.41	0.65
MAML	0.07353	0.6723	0.3152
AttnLSTM	0.0896	0.6372	0.2861
IterRefLSTM	0.09774	0.5426	0.2798
PreGNN	0.0144	0.9435	0.5283
Meta-MGNN	0.0763	0.5196	0.2196
Attentive FP	0.00415	0.893	0.252
D-MPNN	0.00852	0.962	0.537
N-Gram	0.005	0.428	0.334
Mol2Context-vec	**0.00321**	0.423	0.149
Meta-GAT (our)	0.05832	**0.05704**	**0.1046**

require a large-scale molecular corpus and additional self-supervised model parameters. Furthermore, we observe that IterRefLSTM and MAML baseline methods do not have stable performance on different tasks. In other words, they may perform well on Tox21 or SIDER, but perform poorly on the MUV task. In contrast, the performance of Meta-GAT on all three classification datasets is SOTA and stable. In addition, Wilcoxon *p-value* < 0.05 in Table V indicated that the distribution of Meta-GAT predicted results is different from that of other models on the five assays of the MUV dataset. Kappa analysis results in the last row of Table V show that the predicted results and real measurement results are moderately consistent within the allowable error range.

D. QM9

Due to the huge computational cost of density functional theories approaches, there has been considerable interest in applying machine learning models to task of molecular quantum property prediction. QM9 is a comprehensive dataset that provides quantum mechanical properties, which include 12 calculated quantum properties for 134k stable small organic molecules composed of up to nine heavy atoms.

The three quantum properties of LUMO, G, and Cv were used as test tasks, and the other nine quantum properties were used for training tasks. QM9 is a regression dataset, and mean absolute error (MAE) is used to evaluate the performance of regression models. As shown in Table VI, Meta-GAT outcompetes other models on two out of three testing tasks in the QM9 datasets. Two pretrained-based models (N-Gram and Mol2Context-vec) provided more promising predictions than other meta learning models. Meta-GAT shows noticeable improvement in low-data drug discovery and is a promising meta learning method.

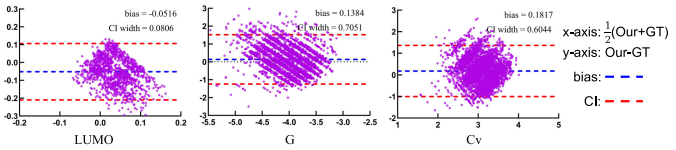

Fig. 7. High agreement between the predicted results of the Meta-GAT (Our) and GT assessed by the Bland–Altman analysis. The x-axis and y-axis represent the average values and the bias between the GT and the predicted values by the Meta-GAT, respectively. The blue dashed line indicates the mean bias. The red dashed lines indicate the 95% confidence intervals of the bias.

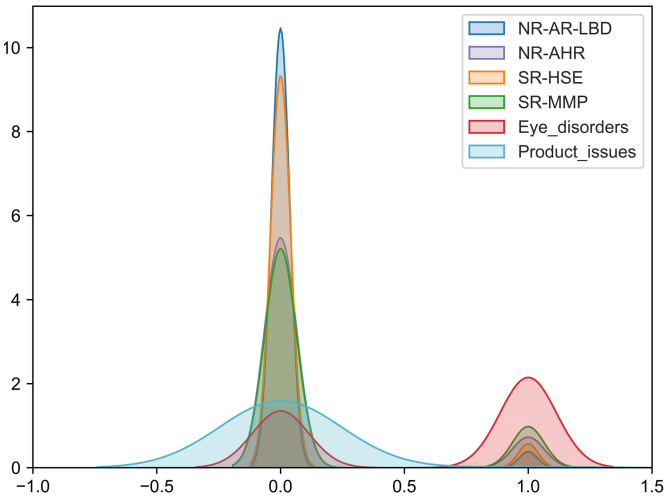

Fig. 8. Comparison of distribution differences between Tox21 and SIDER using kernel density estimation.

Moreover, Fig. 7 illustrates that the predicted results of the Meta-GAT are highly agreed with the ground truth (GT). Each subplot shows the results of Bland–Altman analysis for the three test task sets in QM9. The x-axis and y-axis represent the average values and the bias between the GT and the predicted values by the Meta-GAT, respectively. The blue dashed line indicates the mean bias. The red dashed lines indicate the 95% confidence intervals of the bias. The results show that the mean bias of the three test task sets is -0.0516, 0.1817, and 0.1384, respectively. The percentages of the scatter points falling within the 95% confidence interval are greater than 98%. Therefore, the results show the high agreement between the GT and the predicted results by the Meta-GAT. In other words, the prediction results of Meta-GAT can replace the GT measured by experiment within the allowable error range.

E. Transfer Learning to SIDER From Tox21

The experiments, thus far, have demonstrated that Meta-GAT is able to learn an efficient learning process to transfer meta knowledge from a range of training tasks, allowing the model to rapidly adapt to closely related molecular property predictions with few examples. Transferability and task-relatedness issues need to be carefully evaluated in real-world use cases for drug discovery to determine whether transfer

TABLE VII
ROC-AUC SCORES OF MODELS TRAINED ON TOX21 TESTED ON SIDER

	12 train tasks in Tox21	SIDER		
		1 test tasks	3 test task	27 test task
TrimNet	0.8615	0.4968	0.5080	0.4812
Attentive FP	0.8582	0.5690	0.5156	0.5093
CMPNN	0.8564	0.4968	0.4812	0.4985
CDDD	0.8156	0.5390	0.5454	0.5376
MolPMoFiT	0.7483	0.5804	**0.6025**	0.5183
N-Gram	0.8424	0.5220	0.5234	0.5186
Mol2Context-vec	0.8683	0.5376	0.5687	0.5340
Weave	0.8268	0.5127	0.4968	0.5028
GAT	0.8593	0.4922	0.5156	0.5063
MAML	0.7522	**0.6122**	0.5282	0.5021
AttnLSTM	0.8013	0.5165	0.4857	0.4921
IterRefLSTM	0.8239	0.5269	0.5038	0.5090
CHEF	0.8616	0.5668	0.5625	**0.5512**
PN	0.8523	0.5014	0.4872	0.5173
MAT	0.8362	0.5601	0.5317	0.5269
Meta-GAT (our)	**0.8962**	0.5264	0.5563	0.5264

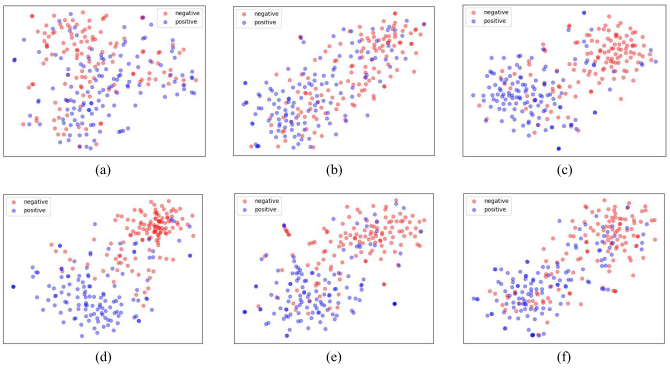

Fig. 9. Visualizations of molecular embeddings generated by Meta-GAT on each internal iteration during the test phase. (a)–(f) Number of iterations 0–5 using the support set.

learning can be used. We train the model on the Tox21 dataset and fine-tune on ten samples taken from the test task on the SIDER dataset, and then evaluate on the remaining samples. There is a large domain transfer in two datasets. The Tox21 measures the results of nuclear receptor assays, and the SIDER measures adverse effects from real patients. This problem becomes so challenging that even domain experts may not be able to accurately judge it.

We only use ten labeled data to transfer on one or more SIDER test tasks. The evaluation results in Table VII show that neither meta learning nor multitask models achieve generalization between unrelated tasks. The performance of these methods using knowledge transferred from Tox21 to SIDER is inferior to that of molecular models trained only on SIDER. Clearly, how to quantify the correlation between different tasks is important for transfer learning in drug discovery.

Kernel density estimation is used to estimate unknown density functions in probability theory. It does not attach any assumption to the data distribution and is a method to study the characteristics of the data distribution from the data sample itself. Fig. 8 shows the distribution differences of NR-AR-LBD, NR-AHR, SR-HSE, SR-MMP in Tox21, and eye disorders and product issues in SIDER using kernel density estimation. There was a strong correlation between the four Tox21 assays, so the training task NR-AR-LBD and NR-AHR could be transferred to the test tasks SR-HSE and SR-MMP. Due to the large distribution difference between Tox21 and SIDER, it may lead to negative transfer, overfitting problems in the case of data scarcity, thus failing to obtain meaningful molecular models. Identifying and addressing these possible limitations are research directions for our future work. Furthermore, kernel density estimation may be a method for assessing transferability in the field of drug discovery, which can measure the distributional differences between source and target tasks, thus revealing task relatedness. We hope our work can promote low-data drug discovery tasks.

F. Feature Visualization and Interpretation for Meta-GAT

The interpretability of the model is crucial, and reducing the gap between the visualization of the model and the chemical intuition of human understanding is conducive to the application of meta learning in drug discovery. When predicting new task of molecular properties, our model uses a few examples in the new task support set to perform several internal iterations, and then evaluates the performance of the query set in the new task. This raises an obvious question: can learning just a few molecules build a competitive classifier?

Taking the test dataset SR-HSE in the Tox21 toxicity prediction task as an example, we explored the performance of the molecular embedding representation generated by the Meta-GAT mode. Specifically, the high-dimensional feature generated by Meta-GAT is a 200-D embedding, similar to the fingerprint vector representation of a molecule. We reduce the high-dimensional vector to 2-D embedding space by t-distributed neighbor embedding (T-SNE) [67] to observe the distribution of molecular representations of different categories. Fig. 9 shows the distribution visualization of molecular embeddings in the SR-HSE dataset, and Fig. 9(a)–(f) represents the number of iterations (0–5) using the support set. The blue dots and red dots represent the numerators of positive and negative examples, respectively.

During model training, an initialization parameter that is approximately optimal for multiple toxicity prediction tasks has been searched. Before using the support set iteration in a new task of SR-HSE toxicity prediction [see Fig. 9(a)], the molecular representation had some degree of separation in 2-D mapping space. But, the model could not clearly classify positive and negative samples, and the blue dots and red dots were still mixed together. After one iteration using the support set [see Fig. 9(b)], the mixing degree of blue dots and red dots weakened and showed aggregation phenomenon to some extent. It shows that Meta-GAT iterates well in the right direction under the guidance of meta knowledge through feature analysis of limited data in the new task. Continue 1–2 iterates, it has been clearly observed that the model can better distinguish between blue and red dots in Fig. 9(c) and (d). The blue dots are mostly gathered in the bottom left corner of the space, and the red dots are mostly in the top-right corner. The model has reached the best performance on the new task. After 4–5 iterations using the same support set, the model may have overfitting. Therefore, we have to set an early stop to make the iterative process terminate early.

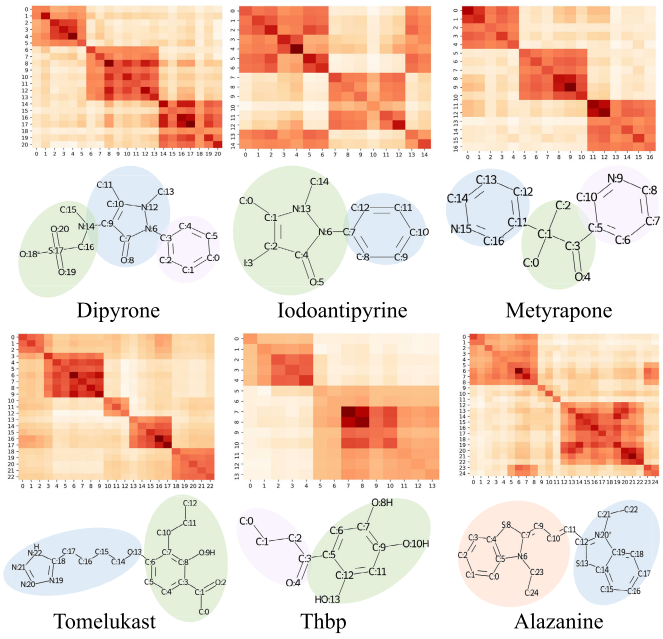

Fig. 10. Heatmap of atomic similarity matrices for six molecules.

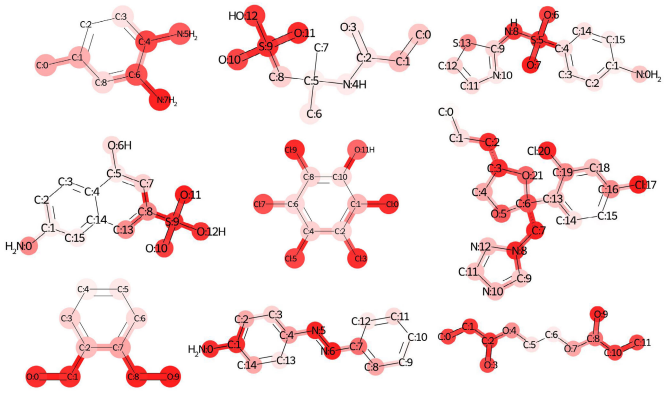

Fig. 11. Attention weights learned from the Meta-GAT are used to highlight each atom in nine molecules in the toxicity prediction task on Tox21 datasets.

In addition, we conducted two visualization experiments on the atom similarity matrix and attention weights to rationalize Meta-GAT. We obtained the similarity coefficient between atom pairs by calculating the Pearson correlation coefficient for those feature vectors and plotted the heatmap of the atomic similarity matrices for the six molecules, as shown in Fig. 10. Taking the molecule structure of Dipyrone as an example, the atoms in Dipyrone are clearly separated into three clusters, as follows: a benzene (atoms 0–5), an aminomethanesulfonic acid (atoms 6–13), and a pyrazolidone (atoms 14–20). The first impression of the visual pattern in the heat map for the compound iodoantipyrine may show some degree of chaos, which is caused by the disorder of the atom numbers in SMILES. Combining atoms 0–6, atom N13, and atom C14 of iodoantipyrine, the atoms in iodoantipyrine are clearly divided into two clusters. The visual pattern of these heat maps strongly agrees with our chemical intuition regarding these molecular structure.

Fig. 11 shows that the attention weights learned from the Meta-GAT model are used to highlight each atom in nine molecules on Tox21 datasets. Meta-GAT model may pay more attention to atomic groups that may cause toxicity, such as sulfonic acid or aniline. The sulfonic acid has potential hazards, including eye burns, skin burns, digestive tract burns if swallowed, and respiratory tract burns if inhaled. Aniline leakage may cause combustion, and explosion hazards, and it is very toxic to the blood and nerves, can be poisoned by the skin or the respiratory tract absorption. These observations suggest that Meta-GAT has indeed successfully extracted relevant information by learning from a specific task, and the attention weight at the atom level indeed has chemical implications. For more intricate problems, attention weight may also be taken as hints for discovering new knowledge.

IV. Conclusion

Drug discovery is the process of discovering new molecules properties and identifying the useful molecules as new drugs after optimization. In the initial stage of optimization of candidate molecules, due to low solubility or possible toxicity, new molecules or analog molecules do not have many records of real physicochemical properties and biological activities. Therefore, the key problem of AI-assisted drug discovery is few examples learning. Here, we propose a meta learning method based on graph attention network, Meta-GAT, which uses graph attention network to extract the interaction of atom pairs and the edge features of bonds in molecules. Also, the meta learning algorithm trains a well-initialized parameter through multiple prediction tasks and, on this basis, performs one or more steps of gradient adjustment to achieve the purpose of quickly adapting to a new task with only few data. Meta-GAT achieves SOTA performance on multiple public benchmark datasets, indicating that it can adapt to new tasks faster than other models. This algorithm is expected to fundamentally solve the problem of few samples in drug discovery. We have proved that Meta-GAT can provide a powerful impetus for low-data drug discovery. The development of meta learning is an important direction of AI-assisted drug discovery. It is believed that the new learning paradigm can be applied in the field of drug discovery in the future.

Acknowledgment

The authors would like to thank the anonymous reviewers for their valuable suggestions.

References

[1] H. Dowden and J. Munro, "Trends in clinical success rates and therapeutic focus," *Nature Rev. Drug Discovery*, vol. 18, no. 7, pp. 495–497, 2019.

[2] L. Wang et al., "Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field," *J. Amer. Chem. Soc.*, vol. 137, no. 7, pp. 2695–2703, Feb. 2015.

[3] G. Sliwoski, S. Kothiwale, J. Meiler, and E. W. Lowe, "Computational methods in drug discovery," *Pharmacological Rev.*, vol. 66, no. 1, pp. 334–395, 2014.

[4] Z. Yang, W. Zhong, L. Zhao, and C. Y.-C. Chen, "ML-DTI: Mutual learning mechanism for interpretable drug–target interaction prediction," *J. Phys. Chem. Lett.*, vol. 12, no. 17, pp. 4247–4261, 2021.

[5] J.-Q. Chen, H.-Y. Chen, W.-J. Dai, Q.-J. Lv, and C. Y.-C. Chen, "Artificial intelligence approach to find lead compounds for treating tumors," *J. Phys. Chem. Lett.*, vol. 10, no. 15, pp. 4382–4400, Aug. 2019.

[6] J.-Y. Li, H.-Y. Chen, W.-J. Dai, Q.-J. Lv, and C. Y.-C. Chen, "Artificial intelligence approach to investigate the longevity drug," *J. Phys. Chem. Lett.*, vol. 10, no. 17, pp. 4947–4961, Sep. 2019.

[7] C. Y. Lee and Y.-P.-P. Chen, "New insights into drug repurposing for COVID-19 using deep learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 11, pp. 4770–4780, Nov. 2021.

[8] M. J. Waring et al., "An analysis of the attrition of drug candidates from four major pharmaceutical companies," *Nature Rev. Drug Discovery*, vol. 14, no. 7, pp. 475–486, Jul. 2015.

[9] 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.

[10] 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.

[11] R. S. Simões, V. G. Maltarollo, P. R. Oliveira, and K. M. Honorio, "Transfer and multi-task learning in QSAR modeling: Advances and challenges," *Frontiers Pharmacol.*, vol. 9, p. 74, Feb. 2018.

[12] C. Li et al., "Geometry-based molecular generation with deep constrained variational autoencoder," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, 2022, doi: 10.1109/TNNLS.2022.3147790.

[13] C. Ji, Y. Zheng, R. Wang, Y. Cai, and H. Wu, "Graph polish: A novel graph generation paradigm for molecular optimization," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Sep. 14, 2021, doi: 10.1109/TNNLS.2021.3106392.

[14] P. Schneider et al., "Rethinking drug design in the artificial intelligence era," *Nature Rev. Drug Discovery*, vol. 19, no. 5, pp. 353–364, May 2020.

[15] X. Jing and J. Xu, "Fast and effective protein model refinement using deep graph neural networks," *Nature Comput. Sci.*, vol. 1, no. 7, pp. 462–469, Jul. 2021.

[16] Z. Xiong et al., "Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism," *J. Medicinal Chem.*, vol. 63, no. 16, pp. 8749–8760, Aug. 2019.

[17] Q. Lv, G. Chen, L. Zhao, W. Zhong, and C. Yu-Chian Chen, "Mol2Context-vec: Learning molecular representation from context awareness for drug discovery," *Briefings Bioinf.*, vol. 22, no. 6, Nov. 2021, Art. no. bbab317.

[18] L. A. Bugnon, C. Yones, D. H. Milone, and G. Stegmayer, "Deep neural architectures for highly imbalanced data in bioinformatics," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 8, pp. 2857–2867, Aug. 2020.

[19] J. Song et al., "Local–global memory neural network for medication prediction," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 4, pp. 1723–1736, Apr. 2021.

[20] R. Huang, X. Tan, and Q. Xu, "Learning to learn variational quantum algorithm," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Feb. 28, 2022, doi: 10.1109/TNNLS.2022.3151127.

[21] Y. Yamanishi, E. Pauwels, and M. Kotera, "Drug side-effect prediction based on the integration of chemical and biological spaces," *J. Chem. Inf. Model.*, vol. 52, no. 12, pp. 3284–3292, Dec. 2012.

[22] Á. Duffy et al., "Tissue-specific genetic features inform prediction of drug side effects in clinical trials," *Sci. Adv.*, vol. 6, no. 37, Sep. 2020, Art. no. eabb6242.

[23] G. Yu, Y. Xing, J. Wang, C. Domeniconi, and X. Zhang, "Multiview multi-instance multilabel active learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 9, pp. 4311–4321, Sep. 2022.

[24] A. Morro et al., "A stochastic spiking neural network for virtual screening," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 4, pp. 1371–1375, Apr. 2018.

[25] Y. Yu and H. Tran, "An XGBoost-based fitted Q iteration for finding the optimal STI strategies for HIV patients," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jun. 2, 2022, doi: 10.1109/TNNLS.2022.3176204.

[26] K. V. Chuang, L. M. Gunsalus, and M. J. Keiser, "Learning molecular representations for medicinal chemistry: Miniperspective," *J. Medicinal Chem.*, vol. 63, no. 16, pp. 8705–8722, Aug. 2020.

[27] 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, May 2020.

[28] D. Duvenaud et al., "Convolutional networks on graphs for learning molecular fingerprints," in *Proc. Adv. Neural Inf. Process. Syst., Annu. Conf. Neural Inf. Process. Syst.* Montreal, QC, Canada: Curran Associates, Inc., Dec. 2015, pp. 2224–2232.

[29] P. Li et al., "TrimNet: Learning molecular representation from triplet messages for biomedicine," *Briefings Bioinf.*, vol. 22, no. 4, Jul. 2021, Art. no. bbaa266.

[30] Q.-J. Lv et al., "A multi-task group bi-LSTM networks application on electrocardiogram classification," *IEEE J. Transl. Eng. Health Med.*, vol. 8, pp. 1–11, 2020.

[31] C. Cai et al., "Transfer learning for drug discovery," *J. Medicinal Chem.*, vol. 63, no. 16, pp. 8683–8694, 2020.

[32] S. Guo, L. Xu, C. Feng, H. Xiong, Z. Gao, and H. Zhang, "Multi-level semantic adaptation for few-shot segmentation on cardiac image sequences," *Med. Image Anal.*, vol. 73, Oct. 2021, Art. no. 102170.

[33] M. Huisman, J. N. Van Rijn, and A. Plaat, "A survey of deep meta-learning," *Artif. Intell. Rev.*, vol. 54, pp. 1–59, Aug. 2021.

[34] A. Banino et al., "Vector-based navigation using grid-like representations in artificial agents," *Nature*, vol. 557, no. 7705, pp. 429–433, May 2018.

[35] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, "Meta-learning in neural networks: A survey," 2020, *arXiv:2004.05439*.

[36] J. Vanschoren, "Meta-learning: A survey," 2018, *arXiv:1810.03548*.

[37] J. X. Wang et al., "Prefrontal cortex as a meta-reinforcement learning system," *Nature Neurosci.*, vol. 21, no. 6, pp. 860–868, May 2018.

[38] S. Biswas, G. Khimulya, E. C. Alley, K. M. Esvelt, and G. M. Church, "Low-N protein engineering with data-efficient deep learning," *Nature Methods*, vol. 18, no. 4, pp. 389–396, Apr. 2021.

[39] R. Liu, X. Yu, X. Liu, D. Xu, K. Aihara, and L. Chen, "Identifying critical transitions of complex diseases based on a single sample," *Bioinformatics*, vol. 30, no. 11, pp. 1579–1586, Jun. 2014.

[40] S. Lin et al., "Prototypical graph contrastive learning," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jul. 27, 2022, doi: 10.1109/TNNLS.2022.3191086.

[41] Y. Sun et al., "Combining genomic and network characteristics for extended capability in predicting synergistic drugs for cancer," *Nature Commun.*, vol. 6, no. 1, pp. 1–10, Sep. 2015.

[42] Q. Liu, H. Zhou, L. Liu, X. Chen, R. Zhu, and Z. Cao, "Multi-target QSAR modelling in the analysis and design of HIV-HCV co-inhibitors: An in-silico study," *BMC Bioinf.*, vol. 12, no. 1, pp. 1–20, Dec. 2011.

[43] H. Altae-Tran, B. Ramsundar, A. S. Pappu, and V. Pande, "Low data drug discovery with one-shot learning," *ACS Central Sci.*, vol. 3, no. 4, pp. 283–293, 2017.

[44] T. Adler et al., "Cross-domain few-shot learning by representation fusion," 2020, *arXiv:2010.06498*.

[45] Z. Guo et al., "Few-shot graph learning for molecular property prediction," in *Proc. Web Conf.*, J. Leskovec, M. Grobelnik, M. Najork, J. Tang, and L. Zia, Eds., Ljubljana, Slovenia, Apr. 2021, pp. 2559–2567.

[46] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM Comput. Surv.*, vol. 53, no. 3, pp. 1–34, 2020.

[47] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, Sydney, NSW, Australia, Aug. 2017, pp. 1126–1135.

[48] D. Jiang et al., "Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models," *J. Cheminformatics*, vol. 13, no. 1, pp. 1–23, Feb. 2021.

[49] R. Winter, F. Montanari, F. Noé, and D.-A. Clevert, "Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations," *Chem. Sci.*, vol. 10, no. 6, pp. 1692–1701, Jul. 2019.

[50] J. Cui, B. Yang, B. Sun, X. Hu, and J. Liu, "Scalable and parallel deep Bayesian optimization on attributed graphs," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 1, pp. 103–116, Jan. 2020.

[51] Z. Wu et al., "MoleculeNet: A benchmark for molecular machine learning," *Chem. Sci.*, vol. 9, no. 2, pp. 513–530, 2018.

[52] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[53] F. Fabris, A. Doherty, D. Palmer, J. P. De Magalhães, and A. A. Freitas, "A new approach for interpreting random forest models and its application to the biology of ageing," *Bioinformatics*, vol. 34, no. 14, pp. 2449–2456, Jul. 2018.

[54] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. 5th Int. Conf. Learn. Represent. (ICLR)*, Toulon, France, Apr. 2017, pp. 1–14.

[55] G. Koch et al., "Siamese neural networks for one-shot image recognition," in *Proc. ICML Deep Learn. Workshop*, vol. 2, Lille, France, 2015, pp. 1–30.
[56] J. Kim, T. Kim, S. Kim, and C. D. Yoo, "Edge-labeling graph neural network for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Long Beach, CA, USA: Computer Vision Foundation, Jun. 2019, pp. 11–20.
[57] W. Hu et al., "Strategies for pre-training graph neural networks," in *Proc. 8th Int. Conf. Learn. Represent. (ICLR)*, Addis Ababa, Ethiopia, Apr. 2020, pp. 1–22.
[58] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
[59] Y. Song, S. Zheng, Z. Niu, Z.-H. Fu, Y. Lu, and Y. Yang, "Communicative representation learning on attributed molecular graphs," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, C. Bessiere, Ed., Jul. 2020, pp. 2831–2838, doi: 10.24963/IJCAI.2020/392.
[60] 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.
[61] A. Mayr, G. Klambauer, T. Unterthiner, and S. Hochreiter, "DeepTox: Toxicity prediction using deep learning," *Frontiers Environ. Sci.*, vol. 3, p. 80, Feb. 2016.
[62] L. Maziarka, T. Danel, S. Mucha, K. Rataj, J. Tabor, and S. Jastrzebski, "Molecule attention transformer," 2020, *arXiv:2002.08264*.
[63] 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. 1–15, Dec. 2020.
[64] S. Liu, M. F. Demirel, and Y. Liang, "N-gram graph: Simple unsupervised representation for graphs, with applications to molecules," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 8464–8476.
[65] M. Kuhn, I. Letunic, L. J. Jensen, and P. Bork, "The SIDER database of drugs and side effects," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D1075–D1079, Jan. 2016.
[66] S. G. Rohrer and K. Baumann, "Maximum unbiased validation (MUV) data sets for virtual screening based on PubChem bioactivity data," *J. Chem. Inf. Model.*, vol. 49, no. 2, pp. 169–184, Feb. 2009.
[67] L. Van Der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

Qiujie Lv is currently pursuing the Ph.D. degree with the Artificial Intelligence Medical Research Center, School of Intelligent Systems Engineering, Shenzhen Campus of Sun Yat-sen University, Shenzhen, Guangdong, China.

His research interests include graph neural network, drug discovery, artificial intelligence, and bioinformatics.

Guanxing Chen is currently pursuing the Ph.D. degree with the Artificial Intelligence Medical Research Center, School of Intelligent Systems Engineering, Shenzhen Campus of Sun Yat-sen University, Shenzhen, Guangdong, China.

His research interests include explainable artificial intelligence, drug discovery, deep learning, biosynthesis, and vaccine design.

Ziduo Yang is currently pursuing the Ph.D. degree with the Artificial Intelligence Medical Research Center, School of Intelligent Systems Engineering, Shenzhen Campus of Sun Yat-sen University, Shenzhen, Guangdong, China.

His main research interests include explainable graph neural network, computer vision, reinforcement learning, and chemoinformatics.

Weihe Zhong is currently pursuing the Ph.D. degree with the Artificial Intelligence Medical Research Center, School of Intelligent Systems Engineering, Shenzhen Campus of Sun Yat-sen University, Shenzhen, Guangdong, China.

His main research interests include graph neural network, chemoinformatics, and drug discovery.

Calvin Yu-Chian Chen is currently the Director of the Artificial Intelligent Medical Center and a Professor with the School of Intelligent Systems Engineering, Shenzhen Campus of Sun Yat-sen University, Shenzhen, Guangdong, China. He also serves as an Advisor at China Medical University Hospital, Taichung, China, and Asia University, Taichung, and a Guest Professor at the Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, and the University of Pittsburgh, Pittsburgh, PA, USA. He has published more than 300 SCI articles and with H-index more than 47. In 2020–2023, he is the highly cited candidate (in the field of computer science and technology). In 2021–2023, he was also selected as the world's top 100 000 scientists. In 2018–2023, he was also selected as the world's top 2% scientists. He had built several artificial intelligence medical systems for hospital, including various pathological image processing, MRI image processing, and big data modeling. He also built the world's largest traditional Chinese medicine database (http://TCMBank.cn/). His laboratory general research interests include developing structured machine learning techniques for computer vision tasks to investigate how to exploit the human commonsense and incorporate them to develop the advanced artificial intelligence system.