2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)

CurrMG: A Curriculum Learning Approach for Graph Based Molecular Property Prediction

Yaowen Gu[1], Si Zheng[1], Jiao Li[1]*

[1]Institute of Medical Information Chinese Academy of Medical Sciences / Peking Union Medical College, Beijing , China
*Correspoding Authors: li.jiao@imicams.ac.cn

2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 978-1-6654-0126-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/BIBM52615.2021.9669478

Abstract—**Nowadays computational methods in bioinformatics and cheminformatics have been widely used in molecular property prediction, advancing activities such as drug discovery. Combining to expert manual annotation of molecular features, machine learning approaches have gained satisfying prediction accuracies in most molecular property prediction tasks. Recently, Graph neural networks (GNNs) have gained increasing popularity in cheminformatics, where a chemical molecule structure is represented as a graph, and have made monumental progress in molecular property prediction. However, GNNs models requires large amounts of training samples, and the diversified molecular structure information might under-utilized when the model is trained with traditional random sampling strategies, thus leading to redundancy and inefficiency. Similar to human learning procedures, training of molecule graph learning models can benefit from an easy-to-difficult curriculum. In this study, we proposed a curriculum learning approach for graph based molecular property prediction, called CurrMG. A data-aware integrated difficulty measurer was proposed to distinguish easy molecules from complex ones. Without any model redesign or external data, our training strategy improves model efficiency and accuracy in numerous molecular property prediction tasks and shows potential for low data drug discovery.**

Keywords—Molecular Property Prediction, Curriculum Learning, Molecule Graph Learning, Training Strategy

I. Introduction

Bioinformatics and cheminformatics technologies have promoted the field of drug discovery such as molecular property prediction [1, 2]. In the early years, machine learning approaches for molecular property prediction gained great interest [3, 4]. In this way, scientists were enthusiastic in designing molecular features by biological and chemical domain knowledge, converting molecule substructures and physicochemical properties to computer-readable formats like molecule fingerprint and descriptors[5]. Combining such molecular representations and a machine learning approach (e.g., logistic regression, random forest, support vector machine, extreme gradient boosting, etc.), the workflow has been widely used and gained moderate performance for molecular property prediction[6], absorption, distribution, metabolism, excretion, and toxicity (ADMET) prediction[3], drug-target interaction prediction (DTI)[7], etc. However, the information loss due to the incomplete molecular representations and the lack of capacities for machine learning models have restricted further development of molecular property prediction approaches.

In recent years, deep learning has achieved landmark improvement in computer vision (CV)[8], natural language processing (NLP)[9] and gradually been applied in the field of biology and chemistry. Researches have shown an excellent performance using deep learning models, such as Conventional Neural Networks (CNNs), Transformer[10], BERT[9], to solve protein structure prediction[11], biomarker discovery[12], drug discovery[13, 14] and drug repositioning[1, 15]. For molecular property prediction, as molecules can be transferred into a sequence format-SMILES, sequence-based models are used for molecule modeling, such as SMILES-BERT[16] and MG-BERT[17]. Moreover, due to consisting of atoms and bonds, molecular structures can also convert to graphs which are made up of nodes and edges, graph-based models called Graph Neural Networks (GNNs) can take molecule graphs as input to learn molecular representations for the downstream tasks. In molecular property prediction task, many GNNs (MPNN[18], AttentiveFP[19], and Pre-trained GIN[20], etc.) have been proposed and achieved excellent performance, which gradually become a paradigm for solving these prediction tasks. However, the recent improvements mainly focus on the redesign of GNNs model architecture, ignoring the re-examination of model training process and data utilization efficiency.

Curriculum learning was first proposed by Bengio in 2009 [21]. Inspired by the human learning process, curriculum learning aims to release the capacity of machine learning models by designing easy-to-difficult curriculums and training models on them. The frameworks of curriculum learning always include a difficulty measurer and a training scheduler [22]. The difficulty measurer is used for calculating the difficulty coefficient for each data in the train set, and the training scheduler is used to arrange them as a curriculum in training. Wide applications on CV and NLP tasks of curriculum learning have proved its ability to improve model generalization and performance [23-25]. Similar to image and sentence sequence, molecule graph can be distinguished by its structure complexity and molecule graph learning model can benefit on curriculum learning. However, the research of curriculum learning on molecular data is limited.

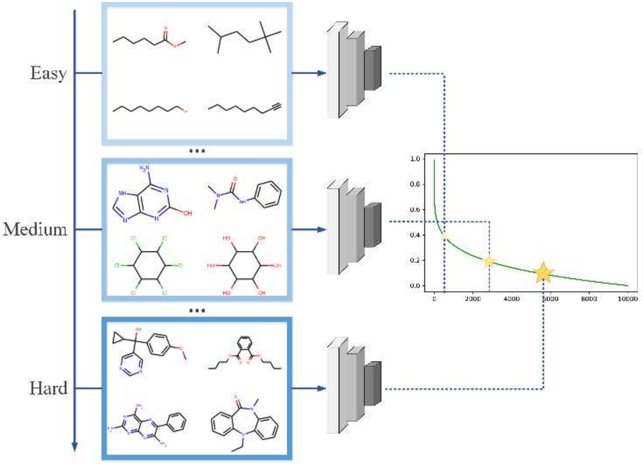

Fig. 1. A concept diagram of curriculum learning in molecule data

978-1-6654-0126-5/21/$31.00 ©2021 IEEE

To address these issues, we proposed a curriculum-based learning approach for molecular property prediction, called CurrMG. This data-aware optimization algorithm focuses on the training phase of molecule graph learning, rearranging training data through their difficulty coefficients calculated with domain knowledge. In this way, the molecule graph learning model is trained gradually from "data easy to learn" to "data difficult to learn" (shown as Fig I), contributing to training a better model. The experiments show a significant improvement of our approach on 6 molecular property benchmarks and 3 QSAR benchmarks compared with those without CurrMG. Especially, CurrMG also remains improvements combining with 5 different molecule graph learning models. In addition, the model convergence results indicate the potential application of CurrMG in low data resource drug discovery. In summary, our contributions can be concluded as follows:

- We explore and demonstrate the effectiveness of curriculum learning method in the field of molecular property prediction. This is the first time that curriculum learning is applied in this field.
- We propose a simple but solid curriculum-based learning framework consisting of difficulty measurer, CDF calculator, training scheduler, and GNN trainer, called CurrMG. Incorporating cheminformatics domain knowledge, CurrMG allows the GNN model to train from easy to difficult by three pre-design difficulty coefficients.
- Comprehensive analysis of various experiments indicates a universal improvement in several benchmarks and models. We also capture the further application potential for QSAR tasks and resource-economizing drug discovery.

II. METHOD

In this section, we introduce the framework of our approach, named CurrMG in detail, including the four core modules, and the whole workflow of our study.

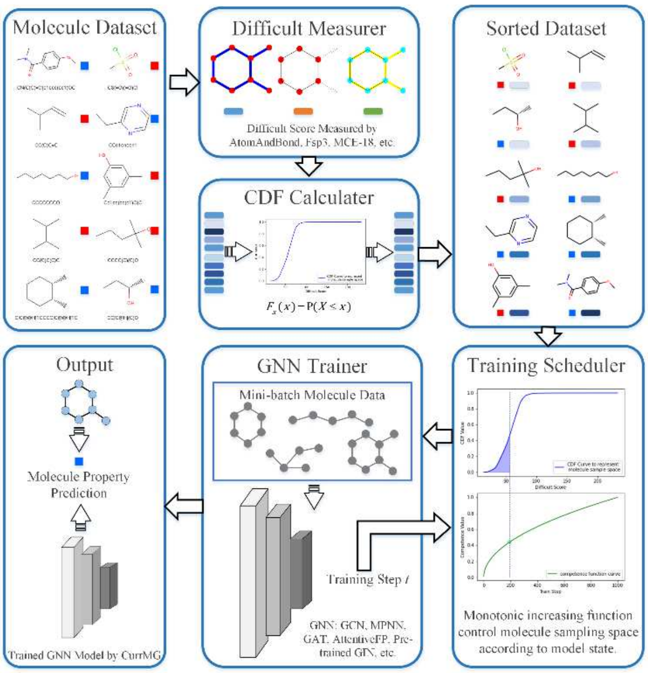
Molecule Dataset
Difficult Measurer
Difficult Score Measured by AtomAndBond, Fsp3, MCE-18, etc.
Sorted Dataset
CDF Calculater
$F_x(x) - P(X \le x)$
Output
Molecule Property Prediction
Trained GNN Model by CurrMG
GNN Trainer
Mini-batch Molecule Data
Training Step t
GNN: GCN, MPNN, GAT, AttentiveFP, Pre-trained GIN, etc.
Training Scheduler
CDF Value
Difficult Score
Competence Value
Train Step
Monotonic increasing function control molecule sampling space according to model state.

Fig. 2. The architecture of CurrMG

A. The Architecture of CurrMG

In this study, we proposed a curriculum-based learning approach for molecule property prediction, named CurrMG, which can accelerate convergence and improve the generalization of the model. Formally, CurrMG includes four core modules: difficulty measurer, CDF calculator, training scheduler, and a GNN trainer, which is shown in Fig 2.

B. Difficulty Measurer

In CurrMG, the difficulty measurer is a method of calculating the difficulty coefficient defined by cheminformatics domain knowledge. In our approach, we defined 3 base difficulty measurers calculated by molecular complexity, which we called $d_{\mathrm{AtomAndBond}}$, d_{Fsp3Ring} and $d_{\mathrm{MCE-18}}$.

$d_{\mathrm{AtomAndBond}}$ is a simple but effective difficulty measurer based on molecule structure. It only focuses on the sum of the number of molecular atoms and bonds so that it denotes the absolute molecular structure complexity which also corresponds to the message passing complexity in GNNs:

$$d_{\mathrm{AtomAndBond}}(x_i) = N_i^{Atom} + N_i^{Bond} \tag{1}$$

Where x_i is the i-th molecule, N_i^{Atom} and N_i^{Bond} represents the number of atoms and bonds in i-th molecule, respectively.

d_{Fsp3Ring} represents the content of cyclic carbon atoms which are sp^3 hybridized. d_{Fsp3Ring} is inspired by sp^3 index (Fsp^3) presented by Lovering[26]. Fsp^3 takes saturation as the molecular complexity, which allows a more complex structure without increasing molecular weight. In CurrMG, $d_{\mathrm{Fsp3Ring}}(x_i)$ is represented as:

$$d_{\mathrm{Fsp3Ring}}(x_i) = \frac{N_i^{sp^3\ cyclic\ carbon}}{N_i^{carbon}} \tag{2}$$

Where $N_i^{sp^3\ cyclic\ carbon}$ and N_i^{carbon} is the number of cyclic carbon atoms which are sp^3 hybridized and the total number of carbon atoms in i-th molecule, respectively.

$d_{\mathrm{MCE-18}}$ is a difficulty measurer that takes into account a variety of structural properties in molecules. MCE-18 is a robust molecule descriptor for the structure evolution analysis of medicinal chemistry initially proposed by Ivanenkov[27]. Due to MCE-18 has made the comprehensive consideration about multiple substructures closely related to molecular quality and complexity, we take it in as a base difficulty measurer in CurrMG framework. $d_{\mathrm{MCE-18}}$ is represented as:

$$d_{\mathrm{MCE-18}}(x_i) = \left[f_{AR}(x_i) + f_{NAR}(x_i) + f_{CHIRAL}(x_i) + f_{SPIRO}(x_i) + \frac{f_{sp^3}(x_i) + f_{Cyc}(x_i) + f_{Acyc}(x_i)}{1 + f_{sp^3}(x_i)} \right] \times Q^1 \tag{3}$$

Where $f_{AR}(x_i)$ denotes whether an aromatic or heteroaromatic ring is in x_i (0 or 1), $f_{NAR}(x_i)$ denotes whether an aliphatic or a heteroaliphatic ring is in x_i (0 or 1), $f_{CHIRAL}(x_i)$ denotes whether a chiral center is in x_i (0 or 1), $f_{SPIRO}(x_i)$ denotes whether a "spiro" point is in x_i, $f_{sp^3}(x_i)$ denotes Fsp^3 (from 0 to 1), $f_{Cyc}(x_i)$ denotes the content of cyclic carbon atoms which are sp^3 hybridized (from 0 to 1). $f_{Acyc}(x_i)$ denotes the content of acyclic carbon atoms which are sp^3 hybridized (from 0 to 1), while Q^1 is a normalized index.

Further, after defining 3 base difficulty measurers, we established a weighted fusion method to construct an integrated difficulty measurer *D*:

$$D(x_i) = \lambda_1 d_{\text{AtomAndBond}}(x_i) + \lambda_2 d_{\text{Fsp3Ring}}(x_i) + \lambda_3 d_{\text{MCE}-18}(x_i) \quad (4)$$

C. CDF Calculator

After calculating each molecular difficulty coefficient by difficulty measurer, we proposed a Cumulative Distribution Function (CDF) calculator to normalize the difficulty coefficient so that meeting the requirement of the training scheduler. In CurrMG framework, a discrete form molecular CDF value can be calculated by dividing the number of molecules whose difficulty coefficients are smaller than the current molecular into the total number of molecules in the training set. The CDF calculator $F(x)$ is represented as:

$$F(x_i) = P(D_X \leq D_i) = \sum_{d \leq D_X} P_{D_X}(d) \quad (5)$$

D. Training Scheduler

In curriculum learning approach, a training scheduler is necessary to guide the model on which mini-batch data to train according to their difficulty coefficient. In CurrMG, inspired by Platanios[23], we used a competence function as a monotonically increasing curve to control data sampling space. In the training phase, training scheduler starts at sampling data with a low difficulty coefficient and constructs a mini-batch dataset, gradually expanding data sampling space as the training progresses, which means that training scheduler transfers training data with a higher difficulty coefficient for the model. The training scheduler $C(t)$ is represented as:

$$C(t) = \min\left(1, \sqrt[\alpha]{t\frac{1-c_0^\alpha}{T} + c_0^\alpha}\right) \quad (6)$$

Where t and T are the number of current iterations and total iterations of the training phase, respectively. c_0 is the initial competence value. α is an adjustable hyperparameter, which controls the change of sample space.

Once we take t as the current model state, we can figure out the threshold for sampling data through competence function which takes t as the unique independent variable. In training scheduler, the competence value calculated by the model state can correspond to the molecular difficulty coefficient to complete the control of the current training data by the model state (e.g., when the competence value is 0.2, so training scheduler randomly samples training data whose difficulty coefficient are lower than 0.2 as a mini-batch data in this training iteration).

E. GNN Trainer

As the universality of CurrMG, we believe CurrMG is suitable for any deep learning model. In our study, we used several GNNs as model trainers. In general, GNN is trained on graph data, aiming to learn node representations by aggregating neighbor information. As for molecule graph learning, given a molecule graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ denotes the set of atoms and $\mathcal{E}$ denotes the set of bonds, the output of GNN model is:

$$h_G = \text{READOUT}(h_1, h_2, \ldots, h_i) \quad (7)$$

Where h_G is the graph-level representation variable of $\mathcal{G}$, and h_i is the node-level representations variable of $i \in \mathcal{V}$, which is updated layer by layer. READOUT is a function for aggregating atom representations as a molecule representation. After GNN layers output graph-level representation variables, a multilayer perceptron (MLP) is joined end-to-end as a predictor for the current task:

$$y_G = \text{MLP}(h_G) \quad (8)$$

F. Training Details

Loss function. The goal of the GNN trainer is to minimize the loss function. To improve numerical stability, in classification tasks, we used a BCEWithLogitsLoss as the loss function:

$$loss(x_i, y_i) = -w_i\left[y_i log\sigma(x_n) + (1 - y_i) log\sigma\big(1 - \sigma(x_i)\big)\right] \quad (9)$$

Where x_i and y_i are the output of GNN model and the true label of i-th molecule, respectively.

In regression tasks, we used a SmoothL1Loss as the loss function:

$$loss(x_i, y_i) = \frac{1}{N}\begin{cases} \frac{1}{2}(x_i - y_i)^2 & if|x_i - y_i| < 1 \\ |x_i - y_i| - \frac{1}{2} & otherwise \end{cases} \quad (10)$$

TABLE I. DETAILS OF MODEL HYPERPARAMETERS

Model	Hyperparameters
GCN	dropout=0.1, gnn_hidden_feats=128, num_gnn_layers=2, predictor_hidden_feats=64, learning rate=0.002, batch size=128, weight decay=0.001
GAT	alpha=0.5, dropout=0.05, gnn_hidden_feats=128, num_gnn_layers=2, num_heads=6, predictor_hidden_feats=128, learning rate=0.01, batch size=128, weight decay=0.0005
MPNN	edge_hidden_feats=64, node_out_feats=48, num_step_message_passing=2, learning rate=0.001, batch size=128, weight decay=0.0005
AttentiveFP	dropout=0.2, graph_feat_size=32, num_layers=2, num_timesteps=3, learning rate=0.01, batch size=128, weight decay=0.001
Pre-trained GIN	jk='concat', readout='sum', learning rate=0.001, batch size=128, weight decay=0.001

Model Hyperparameters. Since model hyperparameters are not the purpose of our study. Therefore, to focus on verifying the optimization effect of our approach on molecular graph model, we set a set of fixed parameters for each GNN model based on experience and used them on all benchmark dataset. The details of model hyperparameters are shown in Tabel I. As for CurrMG, we executed hyperparameters grid searches for 4 hyperparameters ($\lambda_1, \lambda_2, \lambda_3$ in {0, 0.33, 0.5, 1}, respectively. And α in {2, 3}) to find out suitable hyperparameters for each task.

To ensure the stability of our results, for each experiment, we set 5 random seeds for 8:1:1 (train set, validation set, and test set) data splitting and model training. The average performance was reported in the next section.

III. EXPERIMENT

In this section, we introduce comprehensive evaluation and analysis of our approach, including the datasets and models used in our study, experiment results on model performance and convergence, and ablation study.

A. Datasets

We evaluated CurrMG in 6 molecular property benchmark datasets and 3 Quantitative Structure-Activity Relationship (QSAR) benchmark datasets as additional evaluations. The benchmarks used in our study include FreeSolv, ESOL, Lipophilicity, Blood-Brain Barrier Penetration (BBBP), BACE, HIV, Monoamine, JAK2, and HERG, The molecular property benchmark datasets were downloaded from MoleculeNet[28] and the QSAR benchmark datasets were acquired from Cortés-Ciriano[29].

A brief introduction of the 9 benchmark datasets is shown in Table II.

TABLE II. DESCRIPTIONS OF BENCHMARK DATASETS

Category	**Dataset**	**Compounds**	**Task Type**	**Metric**	**Resource**
Physical Chemistry	FreeSolv	642	Regression	RMSE	[28]
Physical Chemistry	ESOL	1128	Regression	RMSE	[28]
Physical Chemistry	Lipophilicity	4200	Regression	RMSE	[28]
Physiology	BBBP	2053	Classification	AUC	[28]
Biophysics	BACE	1513	Classification	AUC	[28]
Biophysics	HIV	41127	Classification	AUC	[28]
QSAR	Monoamine	1379	Regression	RMSE	[29]
QSAR	JAK2	2655	Regression	RMSE	[29]
QSAR	HERG	5207	Regression	RMSE	[29]

FreeSolv is a database including experimental hydration free energy of small molecules; ESOL is a dataset for water solubility of molecules; Lipophilicity is a dataset about octanol/water distribution coefficient (logD at pH7.4) of compounds curated from ChEMBL database; BBBP is a dataset focusing on the barrier permeability of drugs which is a vital property for drugs targeting central nervous system; BACE is a drug-target binding affinity dataset (binary label) for a set of inhibitors of human β-secretase 1 (BACE-1); HIV, published by the Drug Therapeutics Program (DTP), contains an AIDS Antiviral Screen result for over 40,000 compounds. In MoleculeNet, the screening results are transformed to binary labels of inactive and active. Monoamine, JAK2, and HERG are bioactivity datasets from ChEMBL representing the drug-target binding affinity (pIC50 value) of Monoamine oxidase A, Tyrosine-protein kinase JAK2 and HERG, respectively.

B. Models

In our study, 5 GNN models (GCN, MPNN, GAT, AttentiveFP, and Pre-trained GIN) were tested to prove the model-independent of CurrMG. Each of them is partially different from the other:

GCN is a classical GNN model proposed by Kipf[30]. Comparing with CNN, GCN can execute the convolution process in a graph, which belongs to non-euclidean space. GCN aggregates and updates node representations by multiplying a Laplacian matrix in every layer.

MPNN is a general GNN architecture defining a message function and an update function to learn node representations originally used for quantum chemistry[18]. Different from GCN, MPNN also considers edge representations when calculating message function.

GAT is a GNN model proposed for the inductive task and assigning weights for different nodes[31]. GAT brings attention mechanism to GNN model, multiplying attention coefficients to neighbor nodes' representations when aggregating.

AttentiveFP is an interpretable graph attention network for molecular representation [19]. AttentiveFP uses the attention mechanism to learn atom-level local features and molecule-level global features orderly. Moreover, hidden variables from AttentiveFP also show high correlations with empirical descriptors related to certain tasks.

Pre-trained GIN is a graph isomorphism neural network pre-trained on large-scale molecule data using several self-supervised tasks[20]. Due to the well-designed tasks, the pre-trained GIN significantly improves performance in molecular property prediction downstream tasks. In our study, we use the GIN model pre-trained by context prediction task in our experiments.

C. Model Performance on Molecular Property Benchmarks

The performance of 5 GNN models on 6 molecular property benchmarks are shown as TABLE III. The overall performance results indicate that the trained model with CurrMG achieved better performance compared with those without CurrMG (except the GAT on FreeSolv). For every GNN model, the relative improvements when taking CurrMG as the training method are vary from **6.776**% (AttentiveFP) to **12.114**% (GAT) on regression tasks, and from **0.727**% (Pre-trained GIN) to **3.216**% (AttentiveFP) on classification tasks.

For every molecular property benchmark, the relative improvements when taking CurrMG as the training method are vary from **4.954**% (FreeSolv) to **13.941**% (ESOL) on regression tasks, and from **1.332**% (BBBP) to **2.209**% (HIV) on classification tasks.

TABLE III. COMPARISON OF PERFORMANCE ON MOLECULAR PROPERTY BENCHMARKS

Model	**Trainer**	**Dataset**											
		FreeSolv		***ESOL***		***Lipophlicity***		***BBBP***		***BACE***		***HIV***	
		RMSE	***Δ%***[b]	***RMSE***	***Δ%***	***RMSE***	***Δ%***	***ROC-AUC***	***Δ%***	***ROC-AUC***	***Δ%***	***ROC-AUC***	***Δ%***
GCN	Random	0.881	2.001	0.789	13.825	0.699	9.462	0.912	0.883	0.902	0.536	0.752	2.629
	CurrMG	**0.864**		**0.675**		**0.632**		**0.920**		**0.907**		**0.772**	
MPNN	Random	1.375	12.316	0.826	19.339	0.689	4.472	0.852	3.135	0.824	2.387	0.705	2.467
	CurrMG	**1.205**		**0.667**		**0.658**		**0.880**		**0.845**		**0.722**	
GAT	Random	**1.258**	-1.566	0.975	13.120	1.162	24.789	0.870	1.080	0.845	1.634	0.640	0.664
	CurrMG	1.277		**0.847**		**0.874**		**0.880**		**0.859**		**0.645**	
AttentiveFP	Random	1.078	7.065	0.694	9.481	0.760	3.781	0.884	0.808	0.834	4.838	0.646	4.004
	CurrMG	**1.002**		**0.628**		**0.731**		**0.891**		**0.876**		**0.673**	
Pre-trained GIN[a]	Random							0.921	0.755	0.891	0.146	0.746	1.281
	CurrMG							**0.928**		**0.892**		**0.756**	

[a] We only report classification results for Pre-trained GIN since the original implementation[20] do not admit regression task.

[b] For regression task, Δ% denotes 100×(Metric Random - Metric CurrMG)/ Metric Random, where for classification tasks Δ% denotes 100×(Metric CurrMG - Metric Random)/ Metric CurrMG

It is also important to note that through the experiments on 6 molecular property benchmarks, we can observe a stable improvement of our approach-CurrMG compared with the conventional training method which only randomly samples data. The results also proved the sufficient stability and robustness of CurrMG, which is applicable to different GNN models and molecular datasets.

D. Model Convergence on Molecular Property Benchmark

Due to the efficient data utilization efficiency of curriculum learning that it can improve model prediction capacity by just controlling the training data sequences. As for our approach, we also attempted to explore how CurrMG can help with model convergence including the convergence rate and degree. We printed the train loss and validation score curve of GCN trained on molecular property benchmarks to conclude this topic.

The 6 train loss curves shown in Fig 3 indicate that CurrMG caused a rapid model convergence during the early train phase and the final train loss is commonly lower than those without CurrMG. An important explanation is, as an adequate training using CurrMG, the data sampling space is gradually expanded so that the model is trained from a local convergence state to a global convergence state smoothly.

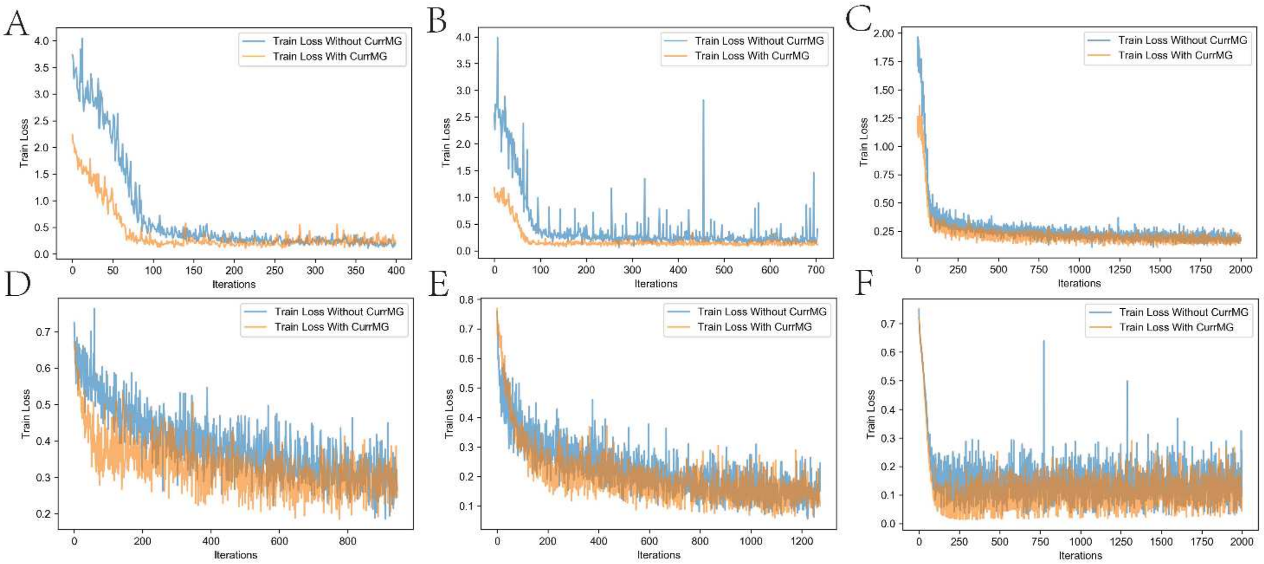

Fig. 3. The train loss curve of GCN on molecular property benchmarks. Fig 3A is GCN train loss curve on FreeSolv; Fig 3B is GCN train loss curve on ESOL; Fig 3C is GCN train loss curve on Lipophilicity; Fig 3D is GCN train loss curve on BACE; Fig 3E is GCN train loss curve on BBBP; Fig 3F is GCN train loss curve on HIV.

The 6 validation score curves shown in Fig 4 provided two findings: (I) Model trained by CurrMG achieved a comparable final validation score compared with those without CurrMG; (II) although it seems that CurrMG caused the model performance improvement on validation set more slowly, combining to the characteristics of CurrMG that the control of data sampling space, the true result is that CurrMG reaches a higher data utilization efficiency. When the model trained by CurrMG reached convergence at about 60% of total iterations, the training dataset is not been "used up", which means CurrMG made a trade-off to release the model capacity with fewer data but longer training iterations. For example, as for HIV benchmark, the model used all training data, and convergence reached less than 10,000 iterations when training without CurrMG. But when trained with CurrMG, the model convergence reached about 15,000 iterations but the used training data are less (77.5% of all training data calculated by competence function). These delightful findings proved the solid improvement in model convergence with CurrMG, while also revealed a feasible prospect for efficient data utilization and low data drug discovery.

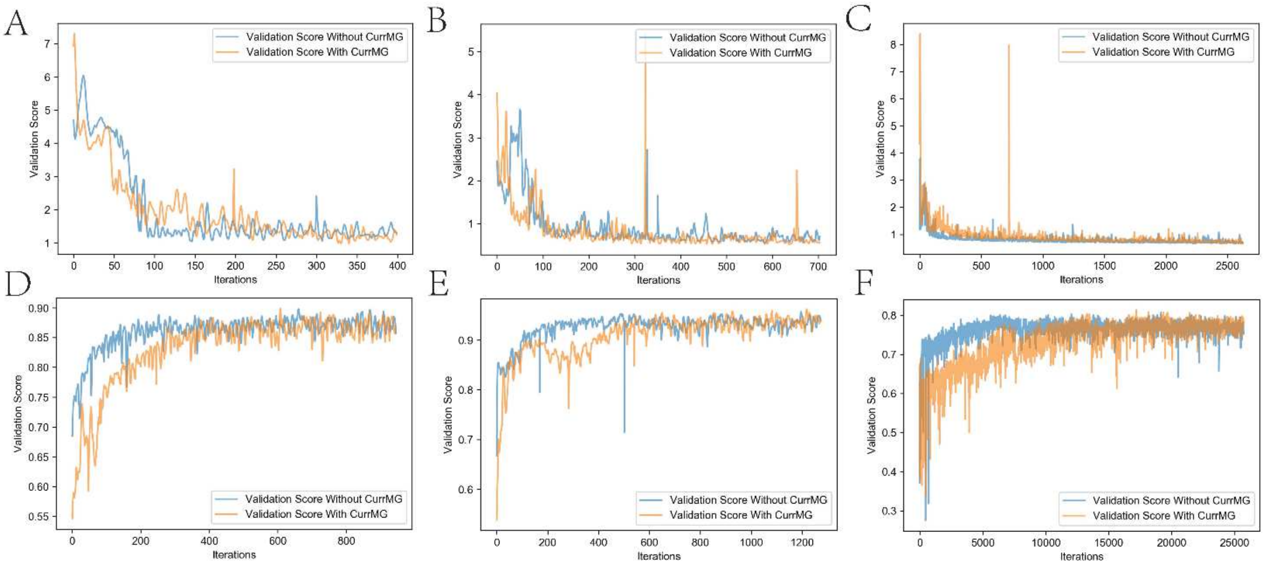

Fig. 4. The validation score curve of GCN on molecular property benchmarks.. Fig 4A is GCN validation score curve on FreeSolv; Fig 4B is GCN validation score curve on ESOL; Fig 4C is GCN validation score curve on Lipophilicity; Fig 4D is GCN validation score curve on BACE; Fig 4E is GCN validation score curve on BBBP; Fig 4F is GCN validation score curve on HIV

E. *Ablation Study*

In this section, we delved into our approach on two interesting topics: (I) Are there some part of CurrMG redundant and can be abandoned? (II) What is the best combination of hyperparameters of CurrMG in a certain benchmark? We designed a module replacement experiment and a hyperparameter combination search experiment to answer these questions.

Module Replacement Experiment. We replaced the core modules of our curriculum learning approach CurrMG-difficulty measurer, CDF calculator, and training scheduler, to conclude that each module is indispensable. For difficulty measurer, we used a random number generator as the replacement of difficulty measurer to calculate the difficulty coefficient. For the CDF calculator, we used another normalization calculator-Min Max Scaler as the replacement. For training scheduler, we simply set $c_0 = 1$ so that training scheduler can sample data with the whole sampling space in the training phase.

We used GCN as the GNN trainer and finished Module Replacement Experiment on 6 molecular property benchmark datasets. The model performance result is shown in Table IV. CurrMG reached the optimal performance on FreeSolv, ESOL, Lipophilicity, BACE, and HIV, and the Min-Max Scaler ablation method reached the optimal performance on BBBP. From our perspective, the reason for the suboptimal performance on BBBP of CurrMG is that the Min-Max Scaler is also an effective method that is widely used in normalization. In our approach, Min-Max Scaler is regarded as an alternative solution in the normalization method which has achieved suboptimal performances on HIV and Lipophilicity.

TABLE IV. COMPARISON OF PERFORMANCE ON MOLECULAR PROPERTY BENCHMARKS

Dataset	**Metric**	**Ablation Method**				
		Baseline	***Random Difficulty Measurer***	***C_0=1***	***Min Max Scaler***	***CurrMG***
FreeSolv	RMSE	0.881	1.045	0.928	0.923	**0.864**
ESOL	RMSE	0.784	0.706	0.695	0.723	**0.687**
Lipophilicity	RMSE	0.699	0.675	0.685	0.662	**0.651**
BBBP	ROC-AUC	0.912	0.916	0.917	**0.927**	0.920
BACE	ROC-AUC	0.902	0.904	0.900	0.903	**0.907**

HIV	ROC-AUC	0.752	0.755	0.747	0.759	**0.772**

Hyperparameter Combination Search Experiment. There are 4 hyperparameters existing in CurrMG-λ_1, λ_2, λ_3 and α, which represent the use of $d_{\mathtt{AtomAndBond}}$, the use of $d_{\mathtt{Fsp3Ring}}$, the use of $d_{\mathtt{MCE-18}}$ and the steepness of the monotonically increasing curve in training scheduler, respectively. Due to the difference in difficulty coefficient distribution in different benchmarks leading to a slight influence on CurrMG, it's necessary to design a hyperparameter search to figure out the optimal hyperparameter combination in a current task. Represented by GCN, we executed hyperparameters grid searches (see *model hypterparameter* in METHOD section) and recorded the top3 hyperparameter combinations (shown as TABLE V). Results indicate that at least 3 hyperparameter combinations of CurrMG have achieved improvement compared to those without CurrMG in almost all molecular property prediction tasks. Moreover, the easiest difficulty measurer $d_{\mathtt{AtomAndBond}}$ got the best performance which participated in the 5 best hyperparameter combinations of 6 tasks. Meanwhile, integrating different difficulty measurer is beneficial to gain steady improvements which occupy 11 of 18 top3 performance records. As for α, related to the smoothness of sample difficulty change during training, $\alpha = 2$ is the optimum value for most tasks.

TABLE V. COMPARISON OF PERFORMANCE ON MOLECULAR PROPERTY BENCHMARKS

Dataset	**1st**		**2nd**		**3rd**	
	Parameters	***Metrics***[b]	***Parameters***	***Metrics***	***Parameters***	***Metrics***
FreeSolv	λ_1=1,λ_2=0,λ_3=0,α=2	0.864(+)	λ_1=0,λ_2=0,λ_3=1,α=2	0.867(+)	λ_1=0,λ_2=0,λ_3=1,α=3	0.916(-)
ESOL	λ_1=0.33,λ_2=0.33,λ_3=0.33,α=2	0.675(+)	λ_1=0.5,λ_2=0.5,λ_3=0,α=2	0.682(+)	λ_1=0.5,λ_2=0,λ_3=0.5,α=3	0.687(+)
Lipophilicity	λ_1=0.5,λ_2=0.5,λ_3=0,α=2	0.632(+)	λ_1=0.5,λ_2=0,λ_3=0.5,α=2	0.651(+)	λ_1=0,λ_2=0,λ_3=1,α=3	0.653(+)
BACE	λ_1=1,λ_2=0,λ_3=0,α=3	0.907(+)	λ_1=0,λ_2=0,λ_3=1,α=2	0.905(+)	λ_1=0.5,λ_2=0.5,λ_3=0,α=2	0.904(+)
BBBP	λ_1=0.5,λ_2=0.5,λ_3=0,α=3	0.920(+)	λ_1=0,λ_2=0.5,λ_3=0.5,α=3	0.916(+)	λ_1=0.5,λ_2=0,λ_3=0.5,α=2	0.916(+)
HIV	λ_1=0,λ_2=0,λ_3=1,α=2	0.772(+)	λ_1=0.5,λ_2=0,λ_3=0.5,α=2	0.764(+)	λ_1=0.5,λ_2=0,λ_3=0.5,α=3	0.760(+)

[c.] The metrics of FreeSolv, ESOL, and Lipophilicity are RMSE. In BACE, BBBP, and HIV, the metrics are ROC-AUC. (+) indicates the current result exceeds baseline result, while (-) is converse.

F. Results on QSAR Benchmark

For further comparison, we finished experiments on 3 QSAR benchmark datasets as additional estimations. The performance of 4 GNN models on 3 QSAR benchmarks are shown as TABLE VI. Similar to the model performance comparisons on molecular property benchmarks, the results on 3 QSAR benchmarks also show an outperforming performance of CurrMG. For every GNN model, the relative improvements when taking CurrMG as the training method are **4.297**% (GCN), **6.729**% (MPNN), **22.058**% (GAT), and **8.938**% (AttentiveFP). For every molecular property benchmark, the relative improvements when taking CurrMG as the training method are **14.165**% (Monoamine), **10.604**% (JAK2), and **6.747**% (HERG). The results show CurrMG can not only apply in molecular property prediction but is also able to expand in other molecular graph-level prediction tasks.

TABLE VI. COMPARISON OF PERFORMANCE ON QSAR BENCHMARKS

Model	**Trainer**	**Dataset**					
		Monoamine		***JAK2***		***HERG***	
		RMSE	***Δ%***	***RMSE***	***Δ%***	***RMSE***	***Δ%***
GCN	Random	0.727	2.733	0.811	4.885	0.720	5.274
	CurrMG	**0.708**		**0.771**		**0.682**	
MPNN	Random	0.908	11.891	0.890	4.653	0.714	3.642
	CurrMG	**0.800**		**0.849**		**0.688**	
GAT	Random	1.349	35.035	1.104	20.080	0.990	11.060
	CurrMG	**0.876**		**0.882**		**0.880**	
AttentiveFP	Random	0.928	7.002	0.995	12.800	0.841	7.013
	CurrMG	**0.863**		**0.868**		**0.782**	

IV. Discussion

In this paper, our proposed approach-CurrMG significantly improved the performance of the molecular graph learning model on molecular property prediction tasks. Experiments on 5 molecular graph learning models and 9 benchmark datasets proved the effectiveness of CurrMG as a model-independent data-aware curriculum-based learning method. Compared with traditional training methods without curriculum learning, CurrMG sorts molecular samples through an integrated difficulty measurer and uses a model-aware training scheduler to control the data sampling space during the training phase, so that the model can be trained from simple samples and gradually transferred to more difficult samples. A comprehensive analysis has proved that this curriculum learning method is helpful to the convergence and generalization of the model. Moreover, results on model convergence also show a feasible prospect for low data drug discovery based on CurrMG. However, the hyperparameters of CurrMG- calculation method in difficulty measurer and the power term in training scheduler are task-related. It is still necessary to simplify the number of hyperparameters as much as possible to reduce the computation time. What's more, as the model-independence of CurrMG, it may suitable for any deep learning models which are not been explored in our study. In the future, we will focus on proposing more effective task-specific difficulty measurers through label distance, and task-aware descriptors. While the complete evaluations of curriculum learning on molecular property prediction (e.g. influence on the computing complexity and suitable models, etc.) are also in the top priorities.

V. CONCLUSIONS

In this work, we have proposed a curriculum-based learning approach for molecular property prediction, called CurrMG. In particular, CurrMG achieves a universal improvement of model convergence and performance on 5 typical molecule graph learning models and 9 benchmark datasets, which also indicate the model-independence and robustness of CurrMG. In addition, benefiting from its easy-to-difficult training curriculum, we observe a moderate potential of data economization for CurrMG when achieving the same performance. These results have established CurrMG as a powerful and robust training optimization algorithm in solving the challenge of molecular property prediction.

ACKNOWLEDGMENT

This work is supported by The National Key Research and Development Program of China (Grant No. 2017YFC0907503 and No.2016YFC0901901) and the National Natural Science Foundation of China (Grant No. 81601573).

REFERENCES

[1] J. Li, S. Zheng, B. Chen, A. J. Butte, S. J. Swamidass, and Z. Lu, "A survey of current trends in computational drug repositioning," *Briefings in bioinformatics,* vol. 17, no. 1, pp. 2-12, Jan 2016.

[2] F. Zhong *et al.*, "Artificial intelligence in drug design," *Science China. Life sciences,* vol. 61, no. 10, pp. 1191-1204, Oct 2018.

[3] J. Dong *et al.*, "ADMETlab: a platform for systematic ADMET evaluation based on a comprehensively collected ADMET database," *J Cheminform,* vol. 10, no. 1, p. 29, Jun 26 2018.

[4] A. S. Rifaioglu, H. Atas, M. J. Martin, R. Cetin-Atalay, V. Atalay, and T. Doğan, "Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases," *Briefings in bioinformatics,* vol. 20, no. 5, pp. 1878-1912, Sep 27 2019.

[5] I. Muegge and P. Mukherjee, "An overview of molecular fingerprint similarity search in virtual screening," *Expert opinion on drug discovery,* vol. 11, no. 2, pp. 137-48, 2016.

[6] A. Varnek and I. Baskin, "Machine learning methods for property prediction in chemoinformatics: Quo Vadis?," *J Chem Inf Model,* vol. 52, no. 6, pp. 1413-37, Jun 25 2012.

[7] C. Chen *et al.*, "DNN-DTIs: Improved drug-target interactions prediction using XGBoost feature selection and deep neural network," *Comput Biol Med,* vol. 136, p. 104676, Jul 29 2021.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems,* vol. 25, pp. 1097-1105, 2012.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805,* 2018.

[10] M. Jaderberg, K. Simonyan, and A. Zisserman, "Spatial transformer networks," *Advances in neural information processing systems,* vol. 28, pp. 2017-2025, 2015.

[11] K. Tunyasuvunakool *et al.*, "Highly accurate protein structure prediction for the human proteome," *Nature,* Jul 22 2021.

[12] O. J. Skrede *et al.*, "Deep learning for prediction of colorectal cancer outcome: a discovery and validation study," *Lancet (London, England),* vol. 395, no. 10221, pp. 350-360, Feb 1 2020.

[13] F. Hu, D. Wang, Y. Hu, J. Jiang, and P. Yin, "Generating Novel Compounds Targeting SARS-CoV-2 Main Protease Based on Imbalanced Dataset," in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2020, pp. 432-436.

[14] Y. Long *et al.*, "Predicting Drugs for COVID-19/SARS-CoV-2 via Heterogeneous Graph Attention Networks," in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2020, pp. 455-459.

[15] H. Liu *et al.*, "Drug Repositioning for SARS-CoV-2 Based on Graph Neural Network," in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2020, pp. 319-322.

[16] S. Wang, Y. Guo, Y. Wang, H. Sun, and J. Huang, "SMILES-BERT: large scale unsupervised pre-training for molecular property prediction," in *Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics*, 2019, pp. 429-436.

[17] X. C. Zhang *et al.*, "MG-BERT: leveraging unsupervised atomic representation learning for molecular property prediction," *Briefings in bioinformatics,* May 5 2021.

[18] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *International conference on machine learning*, 2017: PMLR, pp. 1263-1272.

[19] Z. Xiong *et al.*, "Pushing the Boundaries of Molecular Representation for Drug Discovery with the Graph Attention Mechanism," *Journal of medicinal chemistry,* vol. 63, no. 16, pp. 8749-8760, Aug 27 2020.

[20] W. Hu *et al.*, "Strategies for pre-training graph neural networks," *arXiv preprint arXiv:1905.12265,* 2019.

[21] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th annual international conference on machine learning*, 2009, pp. 41-48.

[22] X. Wang, Y. Chen, and W. Zhu, "A Survey on Curriculum Learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* pp. 1-1, 2021.

[23] E. A. Platanios, O. Stretcu, G. Neubig, B. Poczos, and T. Mitchell, "Competence-based Curriculum Learning for Neural Machine Translation," Minneapolis, Minnesota, 2019: Association for Computational Linguistics, pp. 1162-1172.

[24] B. Xu, L. Zhang, Z. Mao, Q. Wang, H. Xie, and Y. Zhang, "Curriculum learning for natural language understanding," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 6095-6104.

[25] C. Gong, D. Tao, S. J. Maybank, W. Liu, G. Kang, and J. Yang, "Multi-modal curriculum learning for semi-supervised image classification," *IEEE Transactions on Image Processing,* vol. 25, no. 7, pp. 3249-3260, 2016.

[26] F. Lovering, J. Bikker, and C. Humblet, "Escape from flatland: increasing saturation as an approach to improving clinical success," *Journal of medicinal chemistry,* vol. 52, no. 21, pp. 6752-6756, 2009.

[27] Y. A. Ivanenkov, B. A. Zagribelnyy, and V. A. Aladinskiy, "Are We Opening the Door to a New Era of Medicinal Chemistry or Being Collapsed to a Chemical Singularity?," *Journal of medicinal chemistry,* vol. 62, no. 22, pp. 10026-10043, Nov 27 2019.

[28] Z. Wu *et al.*, "MoleculeNet: a benchmark for molecular machine learning," *Chemical science,* vol. 9, no. 2, pp. 513-530, Jan 14 2018.

[29] I. Cortés-Ciriano and A. Bender, "Deep Confidence: A Computationally Efficient Framework for Calculating Reliable Prediction Errors for Deep Neural Networks," *Journal of Chemical Information and Modeling,* vol. 59, no. 3, pp. 1269-1281, 2019/03/25 2019.

[30] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907,* 2016.

[31] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903,* 2017.