PAPER *Special Issue on Deep Learning Technologies: Architecture, Optimization, Techniques, and Applications*

MolHF:Molecular Heterogeneous Attributes Fusion for Drug-Target Affinity Prediction on Heterogeneity

Runze WANG[†], ***Student Member***, **Zehua ZHANG**[†a)], **Yueqin ZHANG**[†], **Zhongyuan JIANG**[††], **Shilin SUN**[†], ***and*** **Guixiang MA**[†††], ***Nonmembers***

SUMMARY Recent studies in protein structure prediction such as AlphaFold have enabled deep learning to achieve great attention on the Drug-Target Affinity (DTA) task. Most works are dedicated to embed single molecular property and homogeneous information, ignoring the diverse heterogeneous information gains that are contained in the molecules and interactions. Motivated by this, we propose an end-to-end deep learning framework to perform **Mol**ecular **H**eterogeneous features **F**usion (**MolHF**) for DTA prediction on heterogeneity. To address the challenges that biochemical attributes locates in different heterogeneous spaces, we design a Molecular Heterogeneous Information Learning module with multi-strategy learning. Especially, another Molecular Heterogeneous Attention Fusion module is present to obtain the gains of molecular heterogeneous features. With these, the diversity of molecular structure information for drugs can be extracted. Extensive experiments on two benchmark datasets show that our method outperforms the baselines in all four metrics. In addition, ablation studies validate the effect of attentive fusion and multi-group of drug heterogeneous features. Visual presentations demonstrate the impact of protein embedding level and the model ability of fitting data. In summary, the diverse gains brought by heterogeneous information contribute to drug-target affinity prediction.

key words: *DTA prediction on heterogeneity, molecular heterogeneous information learning, molecular heterogeneous attention fusion*

1. Introduction

Since the worldwide epidemic of SARS-CoV-2 in 2020, scholars have devoted to research SARS-CoV-2 gene sequences and related biological structures [1]. Recent studies show that there has been a number of variants of the virus [2]. Therefore, rapid screening drug compounds that interact with specific gene sequences or targets on proteins has become a key issue in vaccine development. Drug screening can be reliably performed by in vitro experiments [3], but the ligands screening in hug compounds database via wet-lab experiments is costly and time-consuming, which are unable to meet the rapid drug discovery demands [4].

Later, researchers try to use in silico approaches [5] to alleviate the high cost problems, such as Drug-Target Interaction (DTI) or Drug-Target Affinity (DTA) prediction. Molecular Docking [6] aims to explainably simulate the binding process of ligands and targets, but limited by extremely rare 3D structure data. Later, machine learning comes to the forefront since the capacity of taking complicated data as input. Perlman et al. [7] integrate similarity matrices related to drugs and proteins. Then the logistic repression algorithm is applied to predict drug-target interactions. DLGRMC [8] incorporates similarity information as Laplace Regular Terms and use matrix completion approach for unknown DTI prediction. SimBoost [9] introduces the gradient boosting machines to predict binding affinities for drug-target pairs. Besides, other studies similarly measure the biological entity similarities [10], the DTI matrices [11], and the feature vectors [12] as described above.

Since large amounts of biological data are available, deep learning contributes to drug discovery as it embeds raw biological data to non-linear hidden state [13]. AlphaFold [14] achieves notable results in predicting protein 3D structure, which has raised the boom in deep learning-assisted drug design. Furthermore, various deep models are proposed to embed molecular homogeneous information into feature vector space for target linkage prediction. Such homogeneous information, whether topological [15] or sequential structure [16], is taken as the model input and have shown encouraging performance. DeepDTA [17] is a typical quantitative prediction method, which explores the local sequential relationship between atoms and bonds. Karima et al. [18] attempt to capture correlation between drugs substructures and amino acid fragments to improve the DTA prediction performance. GANsDTA [19] uses Generative Adversarial Network (GAN) for drug sequential structure embedding in a semi-supervised view. Another group of methods mathematically regard molecular as topological graph, by representing atoms as nodes and the chemical bonds as edges. Gao et al. [20] define the graph convolution filters to produce dense vector representation for each atom. Tsubaki et al. [21] extract the topological features of drug molecules and learn their correlation to protein residues. GraphDTA [22] extends multi-type Graph Convolution Network (GCN) to capture local graph structure characteristics to improve drug-target affinity prediction performance. Aforementioned methods have made great performance on homogeneous DTA prediction task, but simply model single structure of each drug. Generally, the one-sided biochemical property is insufficient for learning comprehensive drug representation and the higher correlation to target.

Moreover, molecular heterogeneous attributes contain rich semantic and structural information. For example, local atoms associations can represent local sub-graph patterns (substructure) which are comprised of neighboring atoms

Manuscript received October 11, 2021.
Manuscript revised January 2, 2022.
[†]The authors are with Taiyuan University of Technology
[††]The author is with Xidian University
[†††]The author is with University Of Illinois at Chicago
a) E-mail: zhangzehua@tyut.edu.cn
DOI: 10.1587/transinf.E0.D.1

and bonds. Global molecular feature indicates its own specific structural property. Atoms-bonds arrangement shows the atoms and chemical bonds connections in sequential path, which is generalized by depth-first molecular graph search. Comprehensive learning of multi-type molecular heterogeneous information can provide biological information gains and enrich drug final features. However, the diverse structure information includes different biochemical attributes, which causes the challenge of heterogeneity fusion. Some existing works, DeepGS [23] concatenate learned drug graph structure and sequential structure feature directly, may ignore the selection of heterogeneous information and cause the inadequate feature fusion.

In this paper, we define the heterogeneous DTA prediction task ($\mathcal{H}$-DTA) and present a **Mol**ecular **H**eterogeneous features **F**usion model **MolHF** to obtain the diverse heterogeneous gains. As aforementioned, different types of heterogeneous attributes locate in different spaces. So we design a Molecular Heterogeneous Information Learning module (MHIL) to unify the drug characteristics from multiple spaces. Moreover, a multi-strategy learning for heterogeneous structure of local atoms association, molecular topology and atoms-bonds arrangement is introduced. Considering the different levels of heterogeneous gain brought by meaningful biochemical feature, we devise a Molecular Heterogeneous Attention Fusion (MHAF) module to weight the various heterogeneous attributes before the fusion of drug features. Based on the two modules, comprehensive drug representation can be achieved for better prediction on $\mathcal{H}$-DTA task. The main works are summarized as follows :

- We denote and formulate $\mathcal{H}$-DTA issue for Drug-Target Affinity (DTA) prediction on heterogeneous information level, and propose a heterogeneous fusion method MolHF for $\mathcal{H}$-DTA task, focusing on molecular heterogeneous information.
- We devise a MHIL module to leverage diverse heterogeneous attributes for molecular features fusion. a MHAF module is present to learn the contribution of each biochemical feature for the final drug representation.
- Empirical studies show that attentive fusion of diverse heterogeneous information outperforms the homogeneous molecular properties projecting. Visual means for protein features illustrate that the embedding level of proteins exerts impact on molecular heterogeneous feature selection.

In the urgent period of vaccine or drug development, MolHF can speed the screening process. The predicted results of MolHF are continuous values symbolizing the tightness of drug-target linkage, where the weak interaction pairs will be discarded to reduce the candidate drug space.

2. Formulating $\mathcal{H}$-DTA prediction

In this section, we will define the $\mathcal{H}$-DTA prediction as a regression task to discover the binding strength of a pair of drug and target, as shown in Fig 1. We denote drug set as $\mathcal{D}$ and target set as $\mathcal{T}$, for each drug $d_i \in \mathcal{D}$ and each target $t_j \in \mathcal{T}$, which are associated with the mapping functions $\phi(d_i) : \mathcal{D} \rightarrow \mathcal{A}$ and $\psi(t_j) : \mathcal{T} \rightarrow \mathcal{P}$, respectively. $\mathcal{A}$ denotes the set of drug heterogeneous information where $|\mathcal{A}| \geq 2$. Similarly, $\mathcal{P}$ denotes the set of protein structure information. The mathematical description of $\mathcal{H}$-DTA prediction task can be expressed as to learn heterogeneous function cluster:

$$\mathcal{F}_{heter} = \{f_1(\cdot), f_2(\cdot), \cdots, f_{|\mathcal{D}|}(\cdot)\} \tag{1}$$

where $f_i(\cdot)$ denotes the embedding function for the i-th type of heterogeneous information. Given the triple $\{d_i, t_j, y_{(i,j)}\}$, $y_{(i,j)}$ is the binding affinity between d_i and t_j, the optimization of the whole model is defined as follows:

$$\text{argmin}\{f_{att}[\mathcal{F}_{heter}(\mathcal{A})] \otimes f_{\mathcal{T}}(\mathcal{P}) - y_{(i,j)}\} \tag{2}$$

$f_{\mathcal{T}}(\cdot)$ is protein learning function and $f_{att}(\cdot)$ is denotes the heterogeneous attention fusion function.

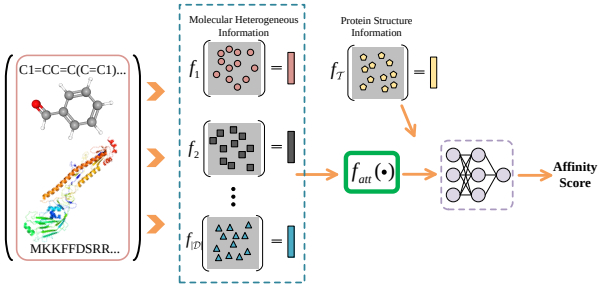


Fig. 1: The formulation of $\mathcal{H}$-DTA prediction task.

3. Proposed Method

MolHF is proposed to obtain the gains of molecular heterogeneous attributes for $\mathcal{H}$-DTA task. The framework can be seen in Fig 2. Given the drug-target pairs as input, MHIL module and protein encoder are devised to firstly transform the structure-data into common feature space. Then, the various types of molecular heterogeneous attributes are fused based on learned attention weights in MHAF. Finally, the final representations of drugs and proteins are fed into Fully Connection Network (FCN) for binding affinity prediction.

3.1 MHIL Module

Since different types of heterogeneous information locate in multiple attribute spaces, MHIL module is needed to unify the learned heterogeneous features into the same feature space. A multi-strategy feature learning will be adopted for representing local atoms association, molecular topology and atoms-bond arrangement of drug. Local atoms association feature reflects the local sub-graph structure that show the specific biochemical property. For this, we use chemical informatics tool RDKit to convert each SMILES string into molecular graph $\mathcal{G}_m = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V}$ and $\mathcal{E}$ represent the sets of atom nodes and bond edges. $v_i \in \mathcal{V}$ and

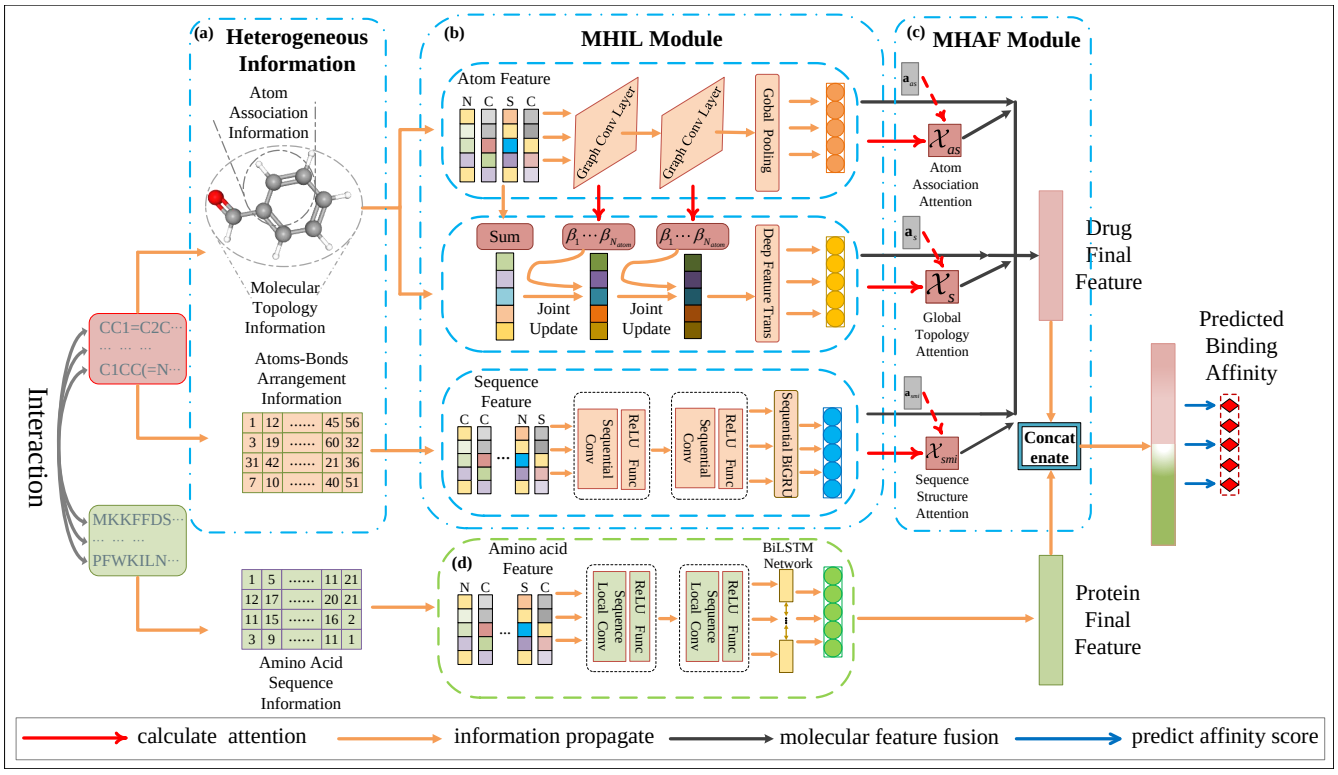


Fig. 2: The framework of proposed MolHF. (a) MolHF leverages three types of molecular heterogeneous information that are atom association, molecular topology and atoms-bonds arrangement information. (b) MHIL Module aims to project different types of heterogeneous properties into the common feature space, including three sub-modules corresponding to the embedding process for each attribute. (c) MHAF Module is used to calculate the attention weight for each heterogeneous information and fuse the embedded features according to coefficients. For example, $\mathcal{X}_{as}$ is the learned atom association attention and $\mathbf{a}_{as}$ denotes the attention vector for atom association information. (d) The feature of protein structure is extracted via unifying the sequential convolution and BiLSTM network.

$e_{i,j} \in \mathcal{E}, i, j \in \{1, 2, \cdots, N_{atom}\}$ denote the i-th atom and chemical bond between the i-th and the j-th atom. Each initial feature of node is comprised with five kinds of atomic properties and is represented in merged binary feature vector:

$$\mathbf{h} = [\mathbf{h_1} \oplus \mathbf{h_2} \oplus \mathbf{h_3} \oplus \mathbf{h_4} \oplus \mathbf{h_5}] \tag{3}$$

where $\oplus$ defines concatenation of vectors. The details of each atom properties are shown in Table1.

Table 1: Atom Properties Representation.

Atom Feature	Property Representation	Dimension
$\mathbf{h_1}$	Atom element	44
$\mathbf{h_2}$	Degree of the atom (direct bond)	11
$\mathbf{h_3}$	Total number of **H** bond to the atom	11
$\mathbf{h_4}$	Number of implicit **H** bound to the atom	11
$\mathbf{h_5}$	Whether the atom is aromatic	11

Then, inspired by the ability of GNN to propagate messages among nodes without modifying graph structure, we introduce three kinds of graph convolution neural network GCN [24], GAT [25], GIN [26] respectively. The process of extracting atom association feature is depicted as Local Graph Conv part in Fig 3. GCN first converts spatial graph into spectral space where all convolution operations are performed. Each convolution layer for molecular graph is denoted as follows:

$$\mathbf{Z}^l = \mathrm{ReLU}(\widetilde{\mathbf{D}}^{-\frac{1}{2}}\widetilde{\mathbf{A}}\widetilde{\mathbf{D}}^{-\frac{1}{2}}\mathbf{Z}^{l-1}\mathbf{W}^l_{GCN}) \tag{4}$$

where $\mathbf{Z}^l \in \mathbb{R}^{N_{atom} * d^l}$ is the hidden atomic feature matrix in the l-th layer. d^l is the hidden vector dimension in the l-th layer, $\mathrm{ReLU}(\cdot)$ denotes a non-linear activation function, $\widetilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ with $\mathbf{A}$ is the adjacency matrix of molecular graph and $\mathbf{I}$ is the identity matrix. $\widetilde{\mathbf{D}}$ is the graph diagonal degree matrix, $\mathbf{W}^l_{GCN}$ denotes the learnable parameter matrix in the l-th layer. GAT adds attention weights between nodes when aggregating neighborhood features:

$$\mathbf{h}^l_i = \mathrm{ELU}(\sum\nolimits_{\mathrm{j} \in \mathcal{N}_{(\mathrm{i})}} \alpha^l_{\mathrm{i,j}} \mathbf{W}^l_{GAT} \mathbf{h}^{l-1}_{\mathrm{j}}) \tag{5}$$

where $\mathbf{h}^l_i$ represents the learned node feature in the hidden layer, $\mathcal{N}_{(i)}$ is the set of one-hop neighborhood nodes of v_i. $\alpha^l_{i,j}$ represents the computed attention weights between v_i and v_j after using softmax$(\cdot)$ function to normalize. $\mathbf{W}^l_{GAT}$

defines the parameter matrix and ELU$(\cdot)$ is a non-linear activation function. GIN incorporates the deep transformation for the aggregated node features:

$$\mathbf{h}_i^l = \mathrm{BL}[\mathrm{MLP}((1+\omega_{GIN}^l)\mathbf{h}_i^{l-1} + \sum\nolimits_{j\in\mathcal{N}_{(i)}} \mathbf{h}_j^{l-1})] \quad (6)$$

where BL$[\cdot]$ represents batch normalization operation, MLP$(\cdot)$ is the multi-layer perceptron and w_{GIN}^l is the learnable parameter. In order to obtain feature representation of the whole molecular graph, a global pooling is performed as Equation 7:

$$\mathbf{h}_{as} = \mathrm{pooling}(\mathbf{h}_1^L, \mathbf{h}_2^L, \cdots, \mathbf{h}_{N_{atom}}^L) \quad (7)$$

where L defines the total layers of graph convolution, as is shorthand notation for atoms association.

As for the molecular topology learning, what we need is a global graph perspective. Thus, a global storage node v_s is initialized outside the given molecular graph, which is similar to [27]. Its feature $\mathbf{h}_s^0$ is obtained by Equation 8:

$$\mathbf{h}_s^0 = \sum_{i=1}^{N_{atom}} \mathbf{h}_i^0 \quad (8)$$

$\mathbf{h}_i^0$ defines the initial feature for each atom. During the iterations for the global node feature, a node-level attention mechanism is applied to make v_s focus on the atoms which have more important impacts:

$$\beta_{(i,s)} = \mathrm{softmax}[\mathbf{q}^T(\tanh(\mathbf{W}_a\mathbf{h}_i^l) * \tanh(\mathbf{W}_s\mathbf{h}_s^{l-1}))] \quad (9)$$

where $\beta_{(i,s)}$ is attention weight between v_i and v_s. $\mathbf{q}^T$ is attention vector in node-level. $\mathbf{W}_a$ and $\mathbf{W}_s$ are the respective feature transformation matrices of the atom nodes and the global node. softmax$[\cdot]$ is a normalized function and tanh$(\cdot)$ denotes a non-linear activated function. Furthermore, multi-head attention mechanism is applied to combine features under every attention space as follows:

$$\mathbf{h}_{comb}^l = \mathbf{W}_{comb}^l[\sum_{i=1}^{N_{atom}} \beta_{(i,s)}^1\mathbf{h}_i^l \oplus \cdots \oplus \sum_{i=1}^{N_{atom}} \beta_{(i,s)}^K\mathbf{h}_i^l] \quad (10)$$

where L defines total number of heads, $\mathbf{h}_{comb}^l$ is the weighted features combination from all K space. $\mathbf{W}_{comb}^l$ denotes the trainable matrix for the concatenated features from multiple attention spaces. After the calculation of multi-head attention mechanism, $\mathbf{h}_{comb}^l$ will be used to update the global molecular topology feature $\mathbf{h}_s^l$:

$$\mathbf{h}_s^l = \mathrm{MultiHead}[\mathbf{h}_s^{l-1}, (\mathbf{h}_1^l, \cdots, \mathbf{h}_{N_{atom}}^l)] + \mathbf{W}_{self}^l\mathbf{h}_s^{l-1} \quad (11)$$

$\mathbf{W}_{self}^l$ is the feature transformation matrix for v_s. The update process of molecular topology feature is described as Global Graph Topology in Fig 3.

The third kind of heterogeneous information is atoms-bonds arrangement, which indicates the local connection between atoms and bonds. Based on data formats of the sequential structure, these symbols of atoms and bonds as

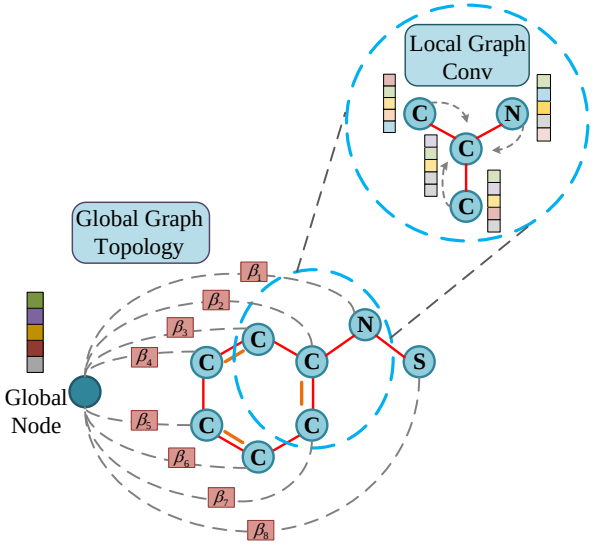


Fig. 3: The part of Global Graph Topology shows the definition of molecular global node and the update of global molecular feature. The part of Local Graph Conv describes the atom association feature learning based on Graph Convolution Neural Network.

specific integer characters are first randomly initialized as semantic vectors. Then a united sequential network is used for structure feature learning, which is comprised of a CNN network and BiGRU network. The CNN network is utilized to capture local chemical context feature and obtain the embedding feature vectors. For the consideration that the more distant components on the arrangement affect the other members, a BiGRU network is built to model the global sequential arrangement relationship which also intends to adapt the shorter length of molecular sequential representation. Finally, the global pooling operation is leveraged to scale the vectors atoms-bonds arrangement feature $\mathbf{h}_{smi}$.

3.2 Protein Structure Feature Learning

The raw data of protein usually consists of predefined token which represents the specific amino acid. We need to extract the local connection feature of protein sequential structure. However, each protein is composed of thousands of amino acids. Based on the previous works [28] and current task, we unify CNN network and BiLSTM network to overcome the challenge of long-range dependency relationships. First, The protein can be defined by a set of amino acid tokens, and these amino acids are randomly initialized into a token vector with fixed dimension. Then, a non-linear transformation layer is used on the whole protein embedding matrix. After that, the CNN network is adopted to extract the hidden features of the local protein sequence. Next, the BiLSTM network is utilized to capture the long-range dependency relationship among extracted hidden feature in different sequence positions and sequence order information. Finally, a

global pooling is applied to obtain final feature representation $\mathbf{h}_{pro}$ for protein.

3.3 MHAF Module

In this portion, the heterogeneous features with multiple types are fused via a feature-level heterogeneous attention. Since the diverse heterogeneous attributes have different contributions, we design a automatic gains selection module. As drug biochemical property are embedded to specific feature vectors, the MHAF Module is proposed to learn the weights of more important heterogeneous gains and fuse them into drug final feature representation. Giving the learned heterogeneous features $\mathbf{h}_{as}$, $\mathbf{h}_{s}$, $\mathbf{h}_{smi}$, the heterogeneous attention function is devised for these molecular features:

$$(\mathcal{X}_{as}, \mathcal{X}_{s}, \mathcal{X}_{smi}) = C_{att}(\mathbf{h}_{as}, \mathbf{h}_{s}, \mathbf{h}_{smi}) \tag{12}$$

where $\mathcal{X}_{as}$, $\mathcal{X}_{s}$, $\mathcal{X}_{sim}$ are the weighted coefficients corresponding to respective heterogeneous feature. $C_{att}(\cdot)$ denotes the attention score calculator. Take the process of calculating the atom association feature weight as example:

$$\mathcal{X}_{as} = \text{softmax}[\mathbf{a}_{as}^{T}\tanh(\mathbf{W}_{att}\mathbf{h}_{as}+\mathbf{b}_{att})] \tag{13}$$

where $\mathbf{W}_{att}$ and $\mathbf{b}_{att}$ are the learnable parameter matrix and bias vector. $\mathbf{a}_{as}$ defines the attention vector related to the a type of molecular heterogeneous feature. T is the vector transpose operation. softmax$[\cdot]$ function is used to normalize the multiple attention which ensures them distribute in interval [0,1] and the sum is 1. Other heterogeneous attention weights are calculated in a similar way.

These attention weights can be interpreted as the importance of biochemical properties that different molecular heterogeneous information brings. The larger the attention weight, the more important the heterogeneous structure information. Finally, a heterogeneous feature selection process is performed according to different attention weights. In the selection process, heterogeneous features are combined by a sum operator after multiplying them with the corresponding attention weights, as shown in Fig. 2(c). After that, the drug final feature representation $\mathbf{h}_{mol}$ can be obtained.

3.4 Prediction for Affinity

The last stage is to integrate the learned features of drug and target which are used to predict the binding affinity. A multilayer Deep Neural Network is applied and the layer forward propagation is denoted as follows:

$$\mathbf{h}^{l} = \text{RELU}(\mathbf{W}_{pre}^{l}\mathbf{h}^{l-1} + \mathbf{b}_{pre}^{l}) \tag{14}$$

where $\mathbf{W}_{pre}^{l}$ and $\mathbf{b}_{pre}^{l}$ denote the deep transformation matrix and bias vector. When $l = 0$, $\mathbf{h}^{0}$ represents the concatenation of embedded drug and protein feature vector. In addition, we add the dropout operation following each layer to avoid overfitting in the predictive process. The Mean Squared Error (MSE) is used as loss function for training and optimizing.

4. Experiments Setting

In this paper, we propose an end-to-end deep learning model MolHF based on fusing molecular heterogeneous features on $\mathcal{H}$-DTA task.

4.1 Datasets

The performance of proposed method is evaluated on two benchmark datasets, Davis and KIBA. Davis [29] contains the kinase proteins and their relevant inhibitors (ligands) with respective dissociation constant (K_d) value. To ensure the stability of the values, the (K_d) values are converted into log space using Equation 15:

$$pK_d = -\log_{10}\frac{K_d}{1e9} \tag{15}$$

where the pK_d values are the drug-target affinity scores.

KIBA [30] regards preprocessed kinase inhibitor bioactivities as binding affinity values. SimBoost [11] filtered it to contain 229 unique proteins and 2111 unique drugs for a fair comparison. Table 2 summaries the detailed statics of the Davis and KIBA datasets.

Table 2: Statics of datasets.

Data Sets	Davis	KIBA
Num Drugs	68	2111
Num Targets	442	229
Train Pairs	25046	98545
Test Pairs	5010	19706
Affinity Range	5.0-10.8	0.0-17.2

4.2 Evaluation Metrics

To evaluate the performance of our method, two general evaluation metrics, Concordance Index (CI) and Mean Squared Error (MSE), are used that follow previous DTA works. CI calculates the consistency of predicted values with true values, i.e., it accesses the conformity level between predicted and true values. The larger predicted values with true values, the higher the prediction accuracy. In general, CI value between 0.50-0.70 is low accuracy; between 0.71-0.90 is moderate accuracy; greater than 0.90 is high accuracy, calculated as follows:

$$CI = \frac{1}{Q}\sum_{\delta_i > \delta_j} h(b_i > b_j) \tag{16}$$

where b_i is the larger affinity δ_i and b_j is the δ_j smaller affinity in predicted values. Q is a normalization constant an is a step function:

$$h(x) = \begin{cases} 1 & x > 0 \\ 0.5 & x = 0 \\ 0 & x < 0 \end{cases} \tag{17}$$

MSE quantifies the distance between predicted and true values. The smaller the MSE value, the closer the predicted value is to the true value:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(f_i - y_i)^2 \tag{18}$$

Another metric is used to evaluate the model ability of fitting the data in recent studies, which is R Squared (r_m^2):

$$r_m^2 = r^2 \times (1 - \sqrt{r^2 - r_0^2}) \tag{19}$$

where r^2 and r_0^2 represent the squared correlation coefficients with and without the intercept, respectively. The larger r^2 index indicates the better ability of fitting the data.

4.3 Parameter Setting

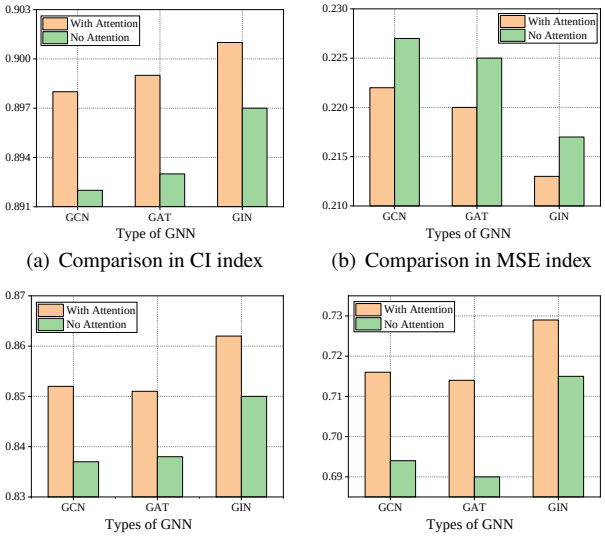
Experiments were conducted on Inspur heterogeneous cluster GPU:12 *32G Tesla V100s, memory 640G DDR2. The whole network framework was built with Pytorch, and Graph Convolution Network is implemented by PyTorch Geometric Framework. We set 1800 epochs for each training iteration and 0.0005 for learning rate. Since each of SMILES string or protein sequence has different lengths, the fixed lengths of SMILES and protein sequence are set as 85 and 1000 respectively in the experiment. And the parts larger than the fixed value are truncated, and the parts smaller than the fixed value are supplemented with 0.

5. Result and Discussion

We conduct experiments to compare with recent baseline methods and answer several questions:
Q1: Does heterogeneous attention fusion benefits DTA prediction performance?
Q2: How do the different molecular heterogeneous features affect the prediction?
Q3: How does performance vary with the better embedding level of protein feature?

5.1 Effectiveness of Heterogeneous Attention Fusion

For Q1, several experiments are conducted to validate the effectiveness of proposed heterogeneous attention fusion in MHAF module. First, all three types of molecular heterogeneous feature are taken as the input of MHAF module. Second, different feature extraction methods of atom association, GCN, GAT, GIN, are performed respectively. Third, another group experiment that directly merges drug feature with concatenation is conducted as contrast. Fig. 4 shows the results using four metrics on Davis dataset. According to Fig. 4, the heterogeneous attention fusion in MHAF module has a better performance in contrast with that fusion type without attention. It is demonstrated that heterogeneous attention fusion can capture the more gainful component while alleviating possible redundancy among heterogeneous features.



(a) Comparison in CI index (b) Comparison in MSE index

(c) Comparison in Pearson index (d) Comparison in R Squared index

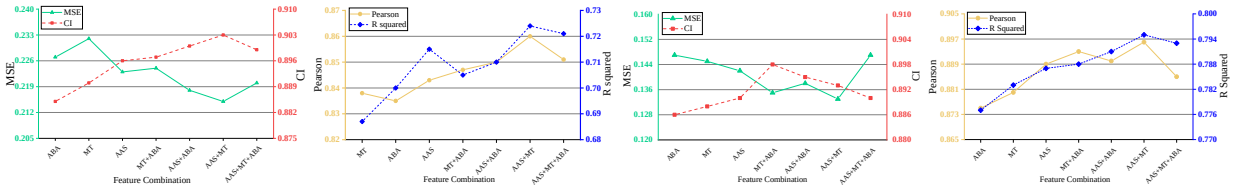
Fig. 4: Effectiveness of attention fusion on Davis Dataset.

In addition, we can also observe that graph convolution way defined by GIN has better adaptation to molecular structure on Davis dataset comparing to the others.

5.2 Effectiveness of Diverse Heterogeneous Features

To answer the Q2, we evaluate the effectiveness of incorporating different types of heterogeneous feature. It is worth thinking that there may be feature gain limit for the drug final representation with the number of molecular heterogeneous features gradually increase, i.e. when the number of heterogeneous features reaches a certain level, the effective information of drug final feature no longer increase. This causes the model prediction performance to no longer improve or even decline due to the possible information redundancy. So it may lead to a bad performance. Thus, we try different combinations of drug heterogeneity to build MolHF model, and adopt four metrics to explore the effectiveness of corresponding features or feature combinations. When implementing the evaluation on the Davis dataset, we use GAT to learn the molecular graph structure feature while GIN on KIBA dataset. Fig. 5a and Fig. 5b show the results in all four metrics on Davis. Fig. 5c and Fig. 5d show the results on KIBA. In Fig. 5, **ABA** denotes the usage of **A**toms-**B**onds **A**rrangement, **MT** represents the Molecular Topology and **AAS** is **A**toms **AS**sociation feature.

It can be seen that using the atom association feature achieve the best among all heterogeneous features. And this illustrates that the topological dependence between atoms in chemical space is more representative for small molecules. As the number of heterogeneous feature increases, prediction results have expected improvement compared with the single feature. It shows that the heterogeneous features provide more embedded views for molecular properties. Meanwhile, the combination of atom-association feature and molecular



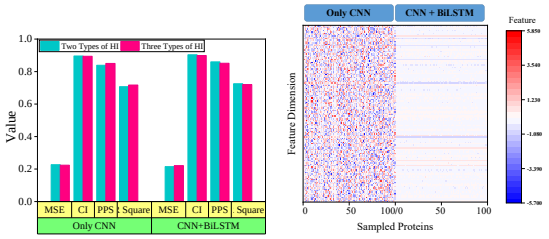
(a) Analysis in MSE and CI on Davis (b) Analysis in Pearson and R Squared on Davis (c) Analysis in MSE and CI on KIBA (d) Analysis in Pearson and R Squared on KIBA

Fig. 5: Effectiveness of diverse heterogeneous features

topology feature produces the best performance. The possible reason is that the comprehensive structure knowledge from local and global is learned from the internal and external perspectives of molecules. So the embedded features also contained the correlation between atoms (nodes) and bonds (edges). Furthermore, the fusion of atoms-bonds arrangement feature on the foundation of atom association and molecular topology feature does not lead to better performance, which demonstrated as we inferred before. It is may be that the effective drug embedding has been enough. However, incorporation of heterogeneous features has a substantial advantage over single learning for molecular attribute. In the following experiments, model will be performed using atom association feature and molecular topology feature.

5.3 Influence of protein embedding feature

In this section, we explore how the embedding level of protein embedding influences the selection of heterogeneous gains. First, we add only CNN network for protein feature embedding as control group. Another experimental group uses the CNN+BiLSTM network to embed protein structure. Then, two types of molecular heterogeneous feature are used against three types of heterogeneous feature. The results are shown in Fig. 6a. When only using CNN network for protein embedding, three types of molecular heterogeneous feature fusion have a better performance, which is different from discovery when using CNN+BiLSTM network. When applying CNN+BiLSTM network, the model performs better and two types of heterogeneous feature are obviously best. It illustrates that poorer embedding for protein tends to have a worse impact on selecting molecular heterogeneous feature.



(a) The influence of protein feature embedding level. (b) Visualization of embedded protein feature values.

Fig. 6: Effectiveness of diverse heterogeneity combination.

To indicate that CNN+BiLSTM network has a better level for protein learning, we randomly sample 100 protein instances in test set and visualize the embedding values in Fig. 6b. The rows show the learned feature values in different dimensions and the columns represent the sampled proteins. Apparently, the protein features are learned via CNN+BiLSTM have centralized and uniform feature values and converge to 0. It is mainly because CNN+BiLSTM network extra captures the long-range dependency relationship. To sum up, better protein embedding method has positive impact on the selection of molecular heterogeneous features.

5.4 Method Comparison

In this section, we compare our method performance with recent state-of-the-art methods on DTA prediction task:

- DeepDTA (2018), which is a deep embedding method based on sequence using CNN.
- AttentionDTA (2019), adding an attention mechanism when combining the subsequence of drug and target.
- GANsDTA (2020), a semi-supervised method with extract sequence feature of drug and target using GAN.
- MATT_DTI (2021), using a multi-head attention mechanism when concatenating feature representation of drug and target.
- GraphDTA (2021), performing multi-type graph neural network for learning molecular graph structure feature.
- DeepGS (2020), concatenating learned molecular graph structure feature vector and sequence structure feature vector forfinal drug feature representation.

All of the baselines are performed and evaluated on the same datasets. The results are shown in Table3 and we have several observations. First, compared to DeepDTA and GANsDTA just using drug sequential feature, AttentionDTA and MATT_DTI that use the different attention mechanism have a greater performance on DTA prediction. It illustrates that the local sequential correlation and the similarity among drug-target pairs are captured by using attention mechanism. Second, the deep methods based on molecular graph structure learning like GraphDTA, outperform the methods using sequence structure feature overall. It can show that molecular graph structure can provide more biochemical information of self-structure.Third, MolHF achieves the best prediction performance compared with baseline methods on two datasets. Especially, compared with GraphDTA that achieve notable

results, our approach has 9% lower in MSE index, 0.013 improvement in CI index and about 6% improvement in R Squared index on Davis dataset. Meanwhile, it has 10% lower in MSE index, also 0.013 in CI index and more than 7% in R Squared index. All the results show that incorporation of molecular heterogeneous gains can enrich the information diversity for drug final feature representation.

Table 3: Compared with baseline methods

Datasets	Methods	MSE	CI	R Squared
Davis	DeepDTA	0.261	0.878	0.630
	AttentionDTA	0.222	0.886	0.676
	GANsDTA	0.276	0.881	0.653
	MATT_DTI	0.227	0.891	0.683
	GraphDTA	0.229	0.893	0.690
	DeepGS	0.252	0.880	0.686
	MolHF	**0.208**	**0.907**	**0.731**
KIBA	DeepDTA	0.194	0.863	0.673
	AttentionDTA	0.162	0.880	0.738
	GANsDTA	0.224	0.866	0.675
	MATT_DTI	0.150	0.889	0.756
	GraphDTA	0.139	0.891	0.741
	DeepGS	0.193	0.860	0.684
	MolHF	**0.125**	**0.904**	**0.810**

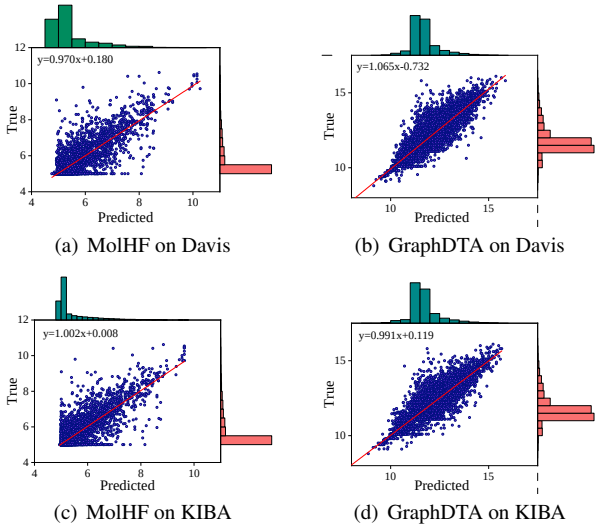
In addition, we add another metric PPC to show the robustness of our proposed method on prediction performance. GraphDTA has best performance among baselines, so we compare our model with GraphDTA in different graph convolution pattern. And the results can be seen in Table4. As a whole, MolHF has competitive prediction performance and fitting ability.

Table 4: The comparison of PPC index

Methods	Graph Convolution Pattern	Davis	KIBA
GraphDTA	GCN (*layer*=3)	0.825	0.891
	GAT (*layer*=2)	0.840	0.863
	GIN (*layer*=5)	0.848	0.879
MolHF	GCN (*layer*=3)	0.845	**0.905**
	GAT (*layer*=2)	0.851	0.875
	GIN (*layer*=5)	**0.856**	0.894

5.5 Visualization of prediction results

In this section, we visualize the predicted values of MolHF and GraphDTA and true affinity values on two datasets. Fig. 7a and Fig. 7b show the visualization comparison on Davis, and Fig. 7c and Fig. 7d depict on KIBA. The linear fitting function is measured to the predicted and true values to demonstrate the fitting ability of our approach. Generally, the closer the absolute value of the slope is to 1, the closer the bias value is to 0, indicating that the predicted value is closer to the original score and the better the model fits. It is obvious that the fitting function of MolHF is significantly closer to regression function with slope 1 than compared method, which illustrates that our proposed model has a stronger fitting ability.



(a) MolHF on Davis (b) GraphDTA on Davis

(c) MolHF on KIBA (d) GraphDTA on KIBA

Fig. 7: Visualization comparison of fitting performance.

6. Conclusion

In this paper, we define the drug-target affinity prediction on heterogeneity ($\mathcal{H}$-DTA) and present a deep prediction model to enrich the heterogeneous gains for drug feature learning. We propose a Molecular Heterogeneous Information Learning (MHIL) module to solve the issue of multi-attribute molecular heterogeneous information fusion and a Molecular Heterogeneous Attention Fusion (MHAF) module to extract more useful heterogeneous gains for drug feature. Extensive experiments on two benchmark datasets show our method has a significant improvement compared with the state-of-the-art methods. And the experiment using combination of heterogeneous information evaluates how incremental heterogeneous information stacking can influence molecular feature extraction.

Future work will first consider how various graph neural networks affect molecular feature learning. Second, the perturbation of molecular representation is brought by protein feature learning and further to consider the interaction between the chemical substructures included in both.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (61503273, 61702356), Industry-University Cooperation Education Program of the Ministry of Education and Shanxi Scholarship Council of China.

References

[1] Lu S, Ye Q and Singh D, "The SARS-CoV-2 nucleocapsid phosphoprotein forms mutually exclusive condensates with RNA and the membrane-associated M protein," Nature communications, vol. 12, pp. 1-15, January 2021.

[2] Abdool Karim S S, de Oliveira T. "New SARS-CoV-2 variants—clinical, public health, and vaccine implications," New Eng-

land Journal of Medicine, vol. 384, pp. 1866-1868, May 2021.
[3] Bagherian M, Sabeti E and Wang K, "Machine learning approaches and databases for prediction of drug–target interaction: a survey paper," Briefings in bioinformatics, vol. 22, pp. 247-269, January 2021.
[4] Chan H C S, Shan H and Dahoun T, "Advancing drug discovery via artificial intelligence," Trends in pharmacological sciences, vol. 40, pp. 801-801, October 2019.
[5] Peska L, Buza K and Koller J, "Drug-target interaction prediction: A Bayesian ranking approach," Computer methods and programs in biomedicine, vol. 152, pp. 15-21, December 2017.
[6] Trott O, Olson A J, "AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading," Journal of computational chemistry, vol. 31, pp. 455-461, January 2010.
[7] Perlman L, Gottlieb A and Atias N, "Combining drug and gene similarity measures for drug-target elucidation," Journal of computational biology, vol. 18, pp. 133-145, February 2011.
[8] Wang M, Tang C and Chen J, "Drug-target interaction prediction via dual Laplacian graph regularized matrix completion," BioMed Research International, vol. 2018, 2018.
[9] He T, Heidemeyer M and Ban F, "SimBoost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines," Journal of cheminformatics, vol. 9, pp. 1-14, April 2017.
[10] Zhang W, Chen Y and Li D, "Drug-target interaction prediction through label propagation with linear neighborhood information," Molecules, vol. 22, pp. 2056, December 2017.
[11] Liu Y, Wu M and Miao C, "Neighborhood regularized logistic matrix factorization for drug-target interaction prediction," PLoS computational biology, vol. 12, February 2016.
[12] Ding Y, Tang J and Guo F, "Identification of drug-target interactions via multiple information integration," Information Sciences, vol. 418, pp. 546-560, December 2017.
[13] Matsumoto S, Ishida S and Araki M, "Extraction of protein dynamics information from cryo-EM maps using deep learning," Nature Machine Intelligence, vol. 3, February 2021.
[14] Callaway E, "'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures," Nature, vol. 588, pp. 203-204, December 2020.
[15] Mengying Sun, Sendong Zhao, Coryandar Gilvary, Olivier Elemento,Jiayu Zhou and Fei Wang, "Graph convolutional networks for computational drug development and discovery," Briefings in bioinformatics, vol. 21, pp. 919-935, May 2020.
[16] Huang K, Xiao C and Glass L M, "MolTrans: Molecular Interaction Transformer for drug–target interaction prediction," Bioinformatics, vol. 37, pp. 830-836, March 2021.
[17] Öztürk Hakime, zgür Arzucan and Elif O, "DeepDTA: Deep Drug-Target Binding Affinity Prediction," Bioinformatics, vol. 34, pp. 821-829, September 2018.
[18] Karimi M, Wu D and Wang Z, "DeepAffinity: Interpretable Deep Learning of Compound-Protein Affinity through Unified Recurrent and Convolutional Neural Networks," Bioinformatics, vol. 35, pp. 3329-3338, September 2019.
[19] Zhao L, Wang J and Pang L, "GANsDTA: Predicting Drug-Target Binding Affinity Using GANs," Frontiers in Genetics, vol. 10, January 2020.
[20] Gao K Y, Fokoue A and Luo H, "Interpretable Drug Target Prediction Using Deep Neural Representation,"// Twenty-Seventh International Joint Conference on Artificial Intelligence IJCAI-18. 2018.
[21] Tsubaki M, Tomii K and Sese J, "Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences," Bioinformatics, vol. 35, pp. 309-318, January 2019.
[22] Nguyen T, Le H and Quinn T P, "GraphDTA: Predicting drug–target binding affinity with graph neural networks," Bioinformatics, vol. 37, pp. 1140-1147, May 2021.
[23] Lin X, Zhao K and Xiao T, "DeepGS: Deep representation learning of graphs and sequences for drug-target binding affinity prediction,"// 24th European Conference on Artificial Intelligence (ECAI) 2020.
[24] Kip F T N and Welling M, "Semi-Supervised Classification with Graph Convolutional Networks," // 5th International Conference on Learning Representations (ICLR), 2017.
[25] Velikovi P, Cucurull G and Casanova A, "Graph Attention Networks," // 6th International Conference on Learning Representations (ICLR), 2018.
[26] Xu K, Hu W and Leskovec J, "How Powerful are Graph Neural Networks ? ," // 7th International Conference on Learning Representations (ICLR), 2019.
[27] Ishiguro K, Maeda S I and Koyama M, "Graph Warp Module: an Auxiliary Module for Boosting the Power of Graph Neural Networks," 2019.
[28] Wei L, Ye X and Xue Y, "ATSE: a peptide toxicity predictor by exploiting structural and evolutionary information based on graph neural network and attention mechanism," Briefings in Bioinformatics, April 2021.
[29] Davis M I, Hunt J P and Herrgard S, "Comprehensive analysis of kinase inhibitor selectivity," NATURE BIOTECHNOLOGY, vol. 29, pp. 1046-U124, November 2011.
[30] Tang J, Szwajda A and Shakyawar S, "Making Sense of Large-Scale Kinase Inhibitor Bioactivity Data Sets: A Comparative and Integrative Analysis," Journal of Chemical Information and Modeling, vol. 54, pp. 735-743, March 2014.

Runze Wang received the B.S. degree from Qufu Normal University, Qufu, China in 2019. He is currently a M.S. candidate at Taiyuan University of Technology from 2019. His research interests include graph data mining and bio-feature recognition.

Zehua Zhang received the Ph.D. degree from Tongji University, Shanghai, China in 2014. He is currently an associate professor at Taiyuan University of Technology. His research interests include soft computing and machine learning, bio-feature recognition and applications, social networks, complex network pattern analysis.

Yueqin Zhang received the M.S. degree from University of Science and Technology Beijing, Beijing, China in 1997. She is currently a professor at Taiyuan University of Technology. Her research interests include Data mining, intelligent information processing, software development techniques and applications.

Zhongyuan Jiang received the Ph.D. degree from Beijing Jiaotong University, Beijing, China in 2013. He is currently an associate professor of School of Cyber Engineering, Xidian University. His research interests include privacy preserving, social computing, urban computing, and network functions virtualization.

Shilin Sun received the B.S.degree from Xinjiang University, Urumqi, China in 2019. He is currently a M.S. candidate at Taiyuan University of Technology from 2020. His research interests include graph data mining in bioinformatics.

Guixiang Ma received the PhD degree in Computer Science from University of Illinois at Chicago in 2019. She is currently an AI Research Scientist at Intel Labs. Her research interests include machine learning, data mining, graph representation learning and their applications in various domains.