Briefings in Bioinformatics, 00(00), 2020, 1–10

doi: 10.1093/bib/bbaa266
Problem Solving Protocol

TrimNet: learning molecular representation from triplet messages for biomedicine

Pengyong Li†, Yuquan Li†, Chang-Yu Hsieh, Shengyu Zhang, Xianggen Liu, Huanxiang Liu, Sen Song and Xiaojun Yao

Corresponding authors: Xiaojun Yao, College of Chemistry and Chemical Engineering, Lanzhou University, Lanzhou 730000, China. Tel: +86-931-8912578. E-mail: xjyao@lzu.edu.cn; Sen Song, Laboratory for Brain and Intelligence and Department of Biomedical Engineering, Tsinghua University, Beijing 100084, China. and +86-10-62773357; E-mail: songsen@tsinghua.edu.cn
†These authors contributed equally to this work.

Abstract

Motivation: Computational methods accelerate drug discovery and play an important role in biomedicine, such as molecular property prediction and compound–protein interaction (CPI) identification. A key challenge is to learn useful molecular representation. In the early years, molecular properties are mainly calculated by quantum mechanics or predicted by traditional machine learning methods, which requires expert knowledge and is often labor-intensive. Nowadays, graph neural networks have received significant attention because of the powerful ability to learn representation from graph data. Nevertheless, current graph-based methods have some limitations that need to be addressed, such as large-scale parameters and insufficient bond information extraction. **Results:** In this study, we proposed a graph-based approach and employed a novel triplet message mechanism to learn molecular representation efficiently, named triplet message networks (TrimNet). We show that TrimNet can accurately complete multiple molecular representation learning tasks with significant parameter reduction, including the quantum properties, bioactivity, physiology and CPI prediction. In the experiments, TrimNet outperforms the previous state-of-the-art method by a significant margin on various datasets. Besides the few parameters and high prediction accuracy, TrimNet could focus on the atoms essential to the target properties, providing a clear interpretation of the prediction tasks. These advantages have established TrimNet as a powerful and useful computational tool in solving the challenging problem of molecular representation learning.
Availability: The quantum and drug datasets are available on the website of MoleculeNet: http://moleculenet.ai. The source code is available in GitHub: https://github.com/yvquanli/trimnet. **Contact:** xjyao@lzu.edu.cn, songsen@tsinghua.edu.cn

Pengyong Li is a PhD candidate in the Department of Biomedical Engineering at Tsinghua University. His research interests focus on graph neural network, drug discovery, artificial intelligence and bioinformatics.
Yuquan Li is a postgraduate student in College of Chemistry and Chemical Engineering at Lanzhou University. His interests mainly include representation learning, chemoinformatics and drug discovery.
Chang-Yu Hsieh obtained his PhD at the University of Ottawa. He works as a researcher in Tencent Quantum Lab. His main research interests include quantum computing and artificial intelligence.
Shengyu Zhang obtained his PhD in computer science at Princeton University. He works as a distinguished scientist in Tencent. His main research interests include quantum computing and artificial intelligence.
Xianggen Liu is a PhD candidate at Tsinghua University. He is interested in natural language process and bioinformatics.
Professor Huanxiang Liu received her PhD degree at Lanzhou University in 2005. She is a professor at Lanzhou University. Her research interests mainly include the misfolding and aggregation mechanism of amyloid-related proteins, drug resistance mechanism, structure-based drug design, etc.
Professor Sen Song received his PhD degree in biology at Brandeis University. He is a professor at Tsinghua University. His current research interests include computational neuroscience, neuroinformatics, artificial intelligence, comparative genomics and bioinformatics.
Professor Xiaojun Yao received his Ph.D. degree in chemoinformatics and theoretical chemistry at University Paris 7-Denis Diderot. He works as a professor of analytical chemistry and chemoinformatics at Lanzhou University. His current research interests include computer-aided molecular design, bioinformatics and computational biology.

Submitted: 4 August 2020; **Received (in revised form):** 11 September 2020

Key words: deep learning; molecular representation; molecular property; compound–protein interaction; computational method; graph neural networks

Introduction

Computational methods have been used in bioinformatics and cheminformatics studies for nearly three decades [8, 14, 16, 32, 34, 35, 37, 38, 42, 49], such as predicting molecular property and identifying the interactions between drugs/compounds and their targets' protein. In the early years, quantum mechanics [19], such as the density functional theory (DFT), were used to determine the molecular structure and calculate properties of interest for a molecule. However, the quantum computational method usually consumes tremendous computational resource and takes hours to days to calculate the molecular properties [37], which hinder their applications to the fields of high-throughput screening. For example, it would be unrealistic to estimate each compound for drug discovery, as the number of potential drug-like compounds is estimated up to 10^{64}. Several traditional machine learning methods are proposed to accelerate the in silico predictions for molecular properties [3, 18, 25, 49], such as support vector machines [7], decision trees [2, 45], random forest, k-nearest neighbors and naive Bayesian methods [10, 39]. These methods significantly shorten the prediction time and provide comparable performance for molecular search tasks [49]. Nevertheless, machine learning methods rely on manually extracted molecular features from molecular structures, and in the meantime, their performance still needs to be improved.

In recent years, deep learning (DL) [27] has achieved excellent performance in computer vision (CV) [17, 20] and natural language processing [9, 50] (NLP) and many DL methods [11, 21, 29, 43, 44] are employed to boost the performance of molecular property prediction. According to the forms of molecular representation, these methods can be divided into sequence-based and graph-based methods [26, 49]. Sequence-based methods usually employ convolutional neural networks (CNNs) or recurrent neural networks to deal with the molecular sequence representation, such as SMILES [52]. On the other hand, graph-based methods take the molecular graph (represented by atom features, bond features and adjacency matrix) as input and use graph neural networks (GNNs) [41] to accomplish the prediction task.

GNNs are the neural networks that perform features transformations based on graphs. It can capture non-Euclidean information and achieve impressive success in social networks, natural science, knowledge graphs and many other research areas [57, 62]. Compared with sequence-based methods, GNNs learn the representation over the molecular structure directly [23, 31, 40, 47, 48, 54, 60, 61]. MoleculeNet [56] has summarized the performances of sequence-based and graph-based methods on 17 molecular property datasets, including quantum mechanics, physical chemistry, biophysics and physiology datasets. The results show that graph-based models surpass conventional methods on 11/17 datasets. Typically, graph convolutional models and message passing neural network (MPNN) [13] achieved the state-of-the-art performance on most datasets. Attention mechanism [1, 15, 51], an effective method to compute representations based on the element importance score, has been widely applied to improve the performance in CV and NLP. Recently, several researchers [46, 55, 58] have introduced the attention mechanism into the MPNN architecture and achieved excellent performance on molecular property predictions. For example, Xiong *et al.* [58] developed attentive FP and achieved the state-of-the-art predictive performance on most datasets in MoleculeNet.

Usually, most MPNN methods [13, 55] need to map the edge-feature vector into a matrix for applying linear transformation, which brings a large number of parameters, on graph nodes. Besides, most current attention-based graph methods [46, 55, 58] only aggregate the neighbor nodes' information to update node's representation, which may lead to insufficient edge information extraction. However, the bonds contain rich information about molecular scaffolds and conformers, which is essential for the molecular properties. Thus, a large amount of parameters and insufficient extraction of edge information may seriously hamper the models' application and performance.

To address these issues, we proposed a novel triplet message networks (TrimNet) to learn molecular representation efficiently. Specifically, this approach explicitly drops the matrix mapping of edge features and employs a triplet message mechanism to calculate message from atom-bond-atom information and update the hidden states of neural networks. In this way, TrimNet reduces the number of parameters and simultaneously improves the extraction of the edge information. We evaluate TrimNet on various molecular property predictive tasks, including quantum properties, bioactivity, physiology and compound–protein interaction (CPI). The experimental results indicate that TrimNet achieves the new state-of-the-art performance on a variety of datasets with a significant parameter reduction. In addition, we have explored the interpretation of TrimNet and found that TrimNet usually focuses on the atoms and the scaffolds essential for the target properties. These advantages indicate that TrimNet can serve as a powerful and useful computational tool for solving the challenging problem of molecular representation learning.

Methods

In this section, we introduce the datasets used in this study and elaborate on our newly proposed model in detail.

Dataset

We evaluated the TrimNet model for four different molecular properties, including quantum properties, bioactivity, physiology and CPI. The benchmark datasets used in this work include QM9, MUV, HIV, BACE, blood-brain barrier permeability (BBBP), Tox21, ToxCast, SIDER, ClinTox, Human and C.elegan. Human and C.elegan datasets are directly retrieved from Tsubaki *et al.* [47], and other datasets are downloaded from the website of MoleculNet [56].

QM9. For quantum property prediction tasks, we employed QM9 dataset for evaluation, which includes 12 calculated quantum properties for 134k molecules. We use the original dataset of MoleculeNet [56], in which the unit of targets $\epsilon HOMO, \epsilon LUMO, \Delta\epsilon, U0, U, H, G$ is *Hartree*, the unit of target μ is *Debye*, the unit of target α is $Bohr^3$, the unit of target $R2$ is $Bohr^2$ and the unit of target Cv is $cal/mol/K$.

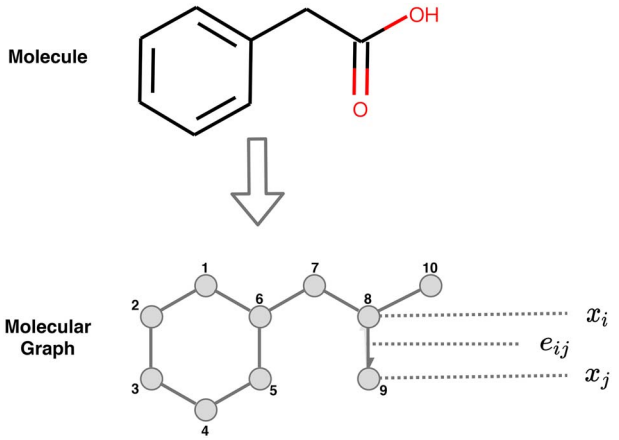

Figure 1. Molecule and molecular graph.

MUV is a subset of PubChem BioAssay by applying a refined nearest neighbor analysis. It contains 17 challenging tasks for 93 127 compounds, designed for validation of virtual screening techniques.

HIV dataset, introduced by the Drug Therapeutics Program AIDS Antiviral Screen, provides the ability to inhibit HIV replication for 41 127 compounds.

The BACE is a database that consists of binding results for a set of inhibitors of human β-secretase 1.

Tox21 contains qualitative toxicity measurements for 8014 compounds on 12 different targets, including stress response pathways and nuclear receptors.

ToxCast is another toxicity database providing toxicology data for compounds based on virtual screening. The processed collection in MoleculeNet contains qualitative results of 617 experiments on 8615 compounds.

SIDER is a database of marketed drugs and adverse drug reactions, which grouped into 27 system organ classes.

ClinTox dataset contains qualitative data of drugs approved by the FDA and those that have failed clinical trials for toxicity reasons.

Human and C.elegan, created by Liu *et al.* [30], include highly credible negative samples of compound–protein pairs by using a systematic screening framework. Positive samples of the datasets were retrieved from DrugBank 4.1 and Matador. We used a balanced dataset with 1:1 of positive and negative samples following Tsubaki *et al.* [47].

All molecules of datasets above are preprocessed into graphs with nodes features, edge features and adjacency matrix with RDKIT [24] as show in Figure 1. For QM9 dataset, we process all the molecules in this dataset into fully connected molecular graphs following Gilmer *et al.* [13]. It means that there is a connection between all pairs of atoms, even if there is no bond between them. For the rest dataset, the structure of molecule graph was encoded according to their original adjacency matrix and the node feature and edge features are identical to [58]. More detailed information about node features and edge features can be found in Supporting Information Tables 1 and 2.

Message phase

Gilmer *et al.* [13] have tried different message calculation functions and found that the edge network achieved the best performance. Nevertheless, this message calculation function has a large number of parameters and high computational cost, which restricts its applicability in a broader context. Hereby, we propose a triplet-attentive edge network as the message calculation function to boost the performance and reduce the computational cost. Our triplet-attentive edge network computes the attention score by a multi-head triplet attention mechanism and aggregates the neighbors' information (including neighboring nodes and edges) according to the attention. Specifically, given the node features $h^t = \{h_0^t, h_1^t, \ldots, h_N^t\}$ and edge features e^t at t time step, triplet-attentive edge network firstly map node pair h_i, h_j and edge hidden state e_{ij}^t to the same dimension D and then concatenate them into a triplet and feed into a feed-forward neural network according to

$$\tau_{i,j}^{t+1} = \text{LeakyReLU}(\boldsymbol{u}^T[\mathbf{W}_h \boldsymbol{h}_i^t || \mathbf{W}_e e_{ij}^t || \mathbf{W}_h \boldsymbol{h}_j^t]), \tag{1}$$

where $\|$ represents concatenation, $\mathbf{W}_h \in \mathbb{R}^{F1\times D}$ and $\mathbf{W}_e \in \mathbb{R}^{F2\times D}$ are learnable weight matrices shared across all nodes and edges ($F1$ and $F2$ are the dimensions of initial node features and edge features), $\boldsymbol{u} \in \mathbb{R}^{3D}$ is also the learnable weight and LeakyReLU stands for the LeakyReLU nonlinear function. To facilitate the comparisons of coefficients across different nodes, we normalize them across all choices of neighbor j using the softmax function

$$\alpha_{i,j}^{t+1} = \text{softmax}(\tau_{i,j}^{t+1}) = \frac{e^{\tau_{i,j}^{t+1}}}{\sum_{j\in\mathcal{N}_i} e^{\tau_{i,j}^{t+1}}}, \tag{2}$$

where $\mathcal{N}_i$ is the set of neighbors for node i. Once obtained, the normalized attention coefficients and the neighboring node hidden states and edge hidden state are used to apply weighted summation operation, to derive the message $\boldsymbol{m}_i$ at each node:

$$\boldsymbol{m}_i^{t+1} = \sum_{j\in\mathcal{N}_i} \alpha_{i,j}^{t+1} \odot \mathbf{W}_h h_j^t \odot \mathbf{W}_e e_{ij}^t, \tag{3}$$

where $\odot$ represents element-wise multiplication. The attentive edge network also employs multi-head attention to stabilize the learning process of self-attention, that is, K independent attention mechanisms execute the transformation of Equation (3), and then their features are concatenated, resulting in the following output feature representation:

$$\boldsymbol{m}_i^{t+1} = ||_k^K \sum_{j\in\mathcal{N}_i} \alpha_{i,j}^{t+1,k} \odot \mathbf{W}_h^k h_j^t \odot \mathbf{W}_e^k e_{ij}^t, \tag{4}$$

where $\|$ represents concatenation, $\alpha_{i,j}^{t+1,k}$ are the normalized attention coefficients computed by the k-th attention mechanism and $\mathbf{W}_e^k$ is the corresponding weight matrix of input linear transformation. Note that, in this setting, the final returned output m_i^{t+1} will consist of KF features (rather than F) for each node. Then, TrimNet employs a gated recurrent unit (GRU) [6] as the vertex update functions to fuse the previously extracted message and the current integrated features and followed by a layer normalization to stabilize the hidden state dynamics in recurrent networks, which is computed by

$$h_i^{t+1} = LN(\mathbf{GRU}(h_i^t, m_i^{t+1})). \tag{5}$$

Note that h_i^{t+1} and h_i^t share the same dimension. The message phase integrate skip connection and performs the above

computations for T times recurrently to derive the final representation h_i^T for each node i.

Readout phase

Based on the final updated nodes' features $h^T = h_0^T, h_1^T, \ldots, h_N^T$, we employed Set2Set networks proposed by Vinyals et al. [51] as the readout function to produce a graph-level embedding. To be specific, Set2Set aggregates nodes features by different attention weights and concatenate the aggregated features with history information that is,

$$q_t = \mathrm{LSTM}(q_{t-1}^*), \tag{6}$$

$$\alpha_{i,t} = \mathrm{softmax}(h_i^T q_t), \tag{7}$$

$$r_t = \sum_{i=1}^{N} \alpha_{i,t} h_i^T, \tag{8}$$

$$q_t^* = q_t || r_t. \tag{9}$$

The Set2Set performs the above computations for T times recurrently to obtain the final graph representation q_T^*. Then, the derived representation q_T^* was fed into a feed-forward neural network to output the final prediction $\hat{y}$.

Loss function

Given the final prediction $\hat{y}$ and the target labels y, the training objective is to minimize the loss function. For prediction of quantum mechanical property, the L1 loss was adopted as the loss function

$$L = -\sum_{i=1}^{N} |\hat{y}_i - y_i|. \tag{10}$$

For the bioactivity and physiology prediction task, we employed a weighted focal loss to deal with the data imbalance problem

$$L = -\sum_{i=1}^{N} \alpha (1 - \hat{y}_i)^{\gamma} \log(\hat{y}_i), \tag{11}$$

where α is a weighting factor in balancing out the importance of positive and negative examples and γ is the focusing parameter to adjust the rate at which the easy examples are down-weighted. In our experiment, we set $\alpha = \frac{\text{No. of Negative samples}}{\text{No. of all samples}}$ and $\gamma = 2$.

For CPI, we used cross entropy loss function (following Tsubaki *et al.* [47]):

$$L = -\sum_{i=1}^{N} \log(\hat{y}_i). \tag{12}$$

Training details and hyperparameters

TrimNet were trained via the standard batch gradient descent method with the error back-propagation algorithm. Specifically, we used the optimization algorithm Adam [22] to update the parameters. Two regularization techniques, including the dropout and the weight decay, were employed to prevent the potential overfitting problem. We trained a separate model for each target on QM9 according to MPNN [13], and the rest datasets with multiple prediction tasks are fit jointly. Our model was implemented with the PyTorch and PyTorch geometric library [12]. The training and testing processes were conducted with V100 and TITAN Black graphic cards.

A grid search procedure was applied to obtain the optimal hyperparameters for the TrimNet. The hyperparameters tuning process involved the learning rate, the hidden size, the dropout rate, the number of attention heads and the number of iteration for the message phase. Finally, we reported the test performance based on the selected hyperparameters for different datasets in Supporting Information Table 3.

Results

The architecture of TrimNet

In this study, we proposed a DL approach employed a novel triplet message mechanism to learn molecular representation efficiently, named TrimNet. A molecule is represented by a graph structure ($\mathbf{G}$), which consists of atom features h_i and bond features e_{ij}. Formally, TrimNet has two phases of operation (the message phase and readout phase), as shown in Figure 2. The message phase contains a message calculation function $\mathbf{M}_t$ based on a multi-head triplet attention mechanism, a vertex update functions $\mathbf{U}_t$ and a layer normalization. The three components work sequentially to update the hidden state h_i^t at each node at each time step t. That is,

$$m_i^{t+1} = \sum_{j \in \mathcal{N}_i} \mathbf{M}_t(h_i^t, h_j^t, e_{ij}), \tag{13}$$

$$h_i^{t+1} = LN(U_t(h_i^t, m_i^{t+1})), \tag{14}$$

where $\mathcal{N}_i$ represents the neighbors of node i, e_{ij} denotes the edge between the node i and node j, LN is the layer normalization. The N message phase blocks are stacked by skip connections.

Given the final representation derived from the message phase, the readout phase leverage the readout function $\mathbf{R}$ to compute the feature vector for the graph

$$\hat{y} = \mathbf{R}(h_i^T | i \in \mathbf{G}). \tag{15}$$

More detailed information can be found in Method section.

Molecular quantum properties prediction

Quantum mechanics is essential for understanding molecular structure and properties, which constitute a cornerstone of chemistry and material discovery. However, the traditional DFT method has high computational costs, so many DL methods, such as MPNN [13] and attentive FP [58], have been developed to accelerate the molecular property prediction and achieved a impressive performance. Here, we focus on the QM9 dataset [36], which consists of 12 quantum properties of 134k molecules calculated by the DFT method to evaluate our TrimNet model.

To achieve a fair comparison, we feed the same features as attentive FP into our TrimNet (denoted as TrimNet^a). Table 1 shows that TrimNet significantly decreased the mean absolute error (MAE). Compared with attentive FP, TrimNet decreased the MAE on six properties ($\mu, ZPVE, U0, U, H, G$) by more than 80% and on five properties ($\alpha, \epsilon HOMO, \epsilon LUMO, \Delta\epsilon, Cv$) prediction tasks by more than 17%. Finally, our model achieved the new state-of-the-art on 12 out of 12 targets. Compared with attentive FP and MPNN, TrimNet focuses more on utilizing the edge information, as quantum properties depend sensitively on bonds,

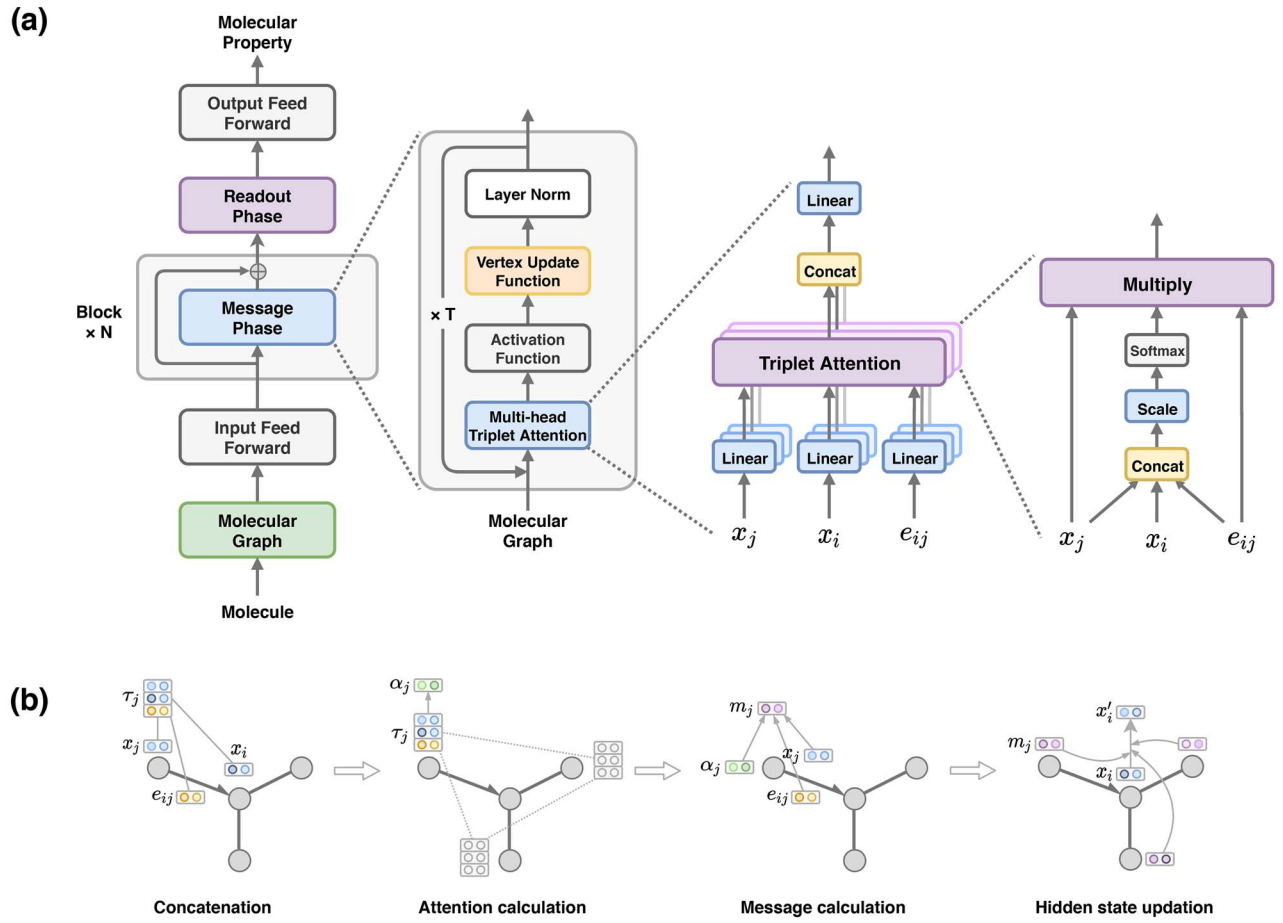

Figure 2. The architecture of TrimNet.

scaffolds and conformers. Besides, Section 3.7 has demonstrated the importance of edge features as inputs. Therefore, emphasizing the edge information may bring excellent performance for Triplet.

To further boost the performance, we feed another set of features into TrimNet (denoted as TrimNet^b). Specifically, aiming to utilize the edge information more fully, we added more edge features, including distances, angles and some electronic features of atom. As a result, TrimNet^b outperformed TrimNet^a on 9 out of 12 tasks in terms of MAE. Besides, it should be noted that the performance can be improved as the depth increased (Section 3.7), but considering the computation cost and training time, we employed the TrimNet model with a depth of three blocks to perform the ablation studies on edge information and layer normalization in Section 3.7. Overall, these results indicate that TrimNet provides a promising representation learning ability for predicting quantum mechanical properties.

Bioactivity and physiology prediction

While TrimNet achieved the state-of-the-art performance on the prediction of quantum mechanical properties, we also applied our model to tasks relevant for drug discovery. There are many fundamental tasks for drug discovery, including the prediction of bioactivity and physiology. To evaluate our model, we picked several benchmark datasets from MoleculeNet (for more details, see Methods). As recommended by MoleculeNet [56], we split HIV, BACE and BBBP datasets by scaffold splitting. The remaining datasets were randomly split according to the molecular substructure, with 8:1:1 for training, validation and test. The node and edge features of molecular graph adopted in this work are identical to that of attentive FP, in order to provide a fair comparison.

Table 2 summarizes the performance of TrimNet and the previous best models on the drug-related benchmark dataset. MUV, HIV and BACE describe the effects of molecules toward different targets, which is fundamental for virtual screening in drug discovery. Our model achieves the state-of-the-art performance on MUV and BACE dataset in terms of the ROC metric. The physiology datasets indicate the effects of chemical compounds in living bodies, including toxicities (Tox21, Toxcast, ClinTox), BBBP and adverse effects (SIDER). TrimNet outperforms the previous best models on the Tox 21, SIDER and ClinTox datasets. Overall, TrimNet achieves the state-of-art performance on six out of eight datasets relevant to drug discovery. These results convincingly reveal TrimNet's potential to master molecular representation learning for drug discovery.

Prediction of CPIs

Identifying interactions between compounds and proteins play an essential role in virtual screening for drug discovery. Various machine learning and DL methods have been developed and achieved excellent performance for CPI prediction [5, 33, 53].

Table 1. Comparison of prediction mean absolute error on QM9dataset

Model	Name	ECFP*	CM*	DTNN*	MPNN*	Attentive FP*	TrimNeta	TrimNetb
	Depth	–	–	–	1	2	3	3
	Params	–	–	c–	–	1683k	56k	**56k**
TASKS	μ	0.602	0.519	0.244	0.358	0.451	0.414	**0.0741**
	α	3.1	0.85	0.95	0.89	0.492	0.299	**0.216**
	$\epsilon HOMO$	0.0066	0.00506	0.00388	0.00541	0.00358	0.00290	**0.00226**
	$\epsilon LUMO$	0.00854	0.00645	0.00513	0.00623	0.00415	0.00300	**0.00192**
	$\Delta\epsilon$	0.01	0.0086	0.0066	0.0082	0.00528	0.00433	**0.00336**
	R2	125.7	46	17	28.5	26.839	24.98	**2.178**
	ZPVE	0.01109	0.00207	0.00172	0.00216	0.00120	0.000233	**0.000140**
	U0	15.1	2.27	2.43	2.05	0.898	**0.0681**	0.0927
	U	15.1	2.27	2.43	2	0.893	**0.0631**	0.0861
	H	15.1	2.27	2.43	2.02	0.893	**0.0638**	0.0774
	G	15.1	2.27	2.43	2.02	0.893	0.0791	**0.0717**
	Cv	1.77	0.39	0.27	0.42	0.252	0.118	**0.0715**

[*] The results of these models are taken from attentive FP [58]. [a] This model uses the same features as attentive FP [58]. [b] This model uses another set of features and use fully connected molecular graphs with hydrogenation added. Details of the features can be found in Support Information Table 1.

Table 2. The performance of TrimNet models on the drug discovery-related dataset

Category	Dataset	Compounds	Tasks	Task type	Split	Metric	Previous best	TrimNet
Bioactivity	MUV	93127	17	Classification	Random	AUC	Graphconv+dummy: 0.845[a]	**0.851**
	HIV	41127	1	Classification	Scaffold	AUC	Attentive FP: 0.832	0.804
	BACE	1513	1	Classification	Scaffold	AUC	RF: 0.867[b]	**0.878**
Physiology	BBBP	2053	1	Classification	Scaffold	AUC	Attentive FP: 0.920	0.850
	Tox21	7831	12	Classification	Random	AUC	Attentive FP: 0.858	**0.860**
	ToxCast	8575	617	Classification	Random	AUC	Attentive FP: 0.805	0.777
	SIDER	1427	27	Classification	Random	AUC	Attentive FP: 0.637	**0.657**
	ClinTox	1478	2	Classification	Random	AUC	Attentive FP: 0.940	**0.948**

[a] [a] This model are referenced from Li *et al.* [29]. [b] [b] This model are referenced from attentive FP. [58].

Table 3. Performance on CPI prediction

Dataset	Model	Precision	Recall	AUC
Human	GNN-CNN	0.923	0.918	0.970
	TrimNet-CNN	0.918	**0.953**	**0.974**
C.elegans	GNN-CNN	0.938	0.929	0.978
	TrimNet-CNN	**0.946**	**0.945**	**0.987**

Tsubaki *et al.* [47] proposed a novel framework that employed GNN and CNN for CPI predictions to learn the representation of compounds and protein sequence, which significantly outperformed existing methods. Here, we adapted their algorithm and replaced their GNN part with our TrimNet to evaluate the effectivity of TrimNet on CPI prediction. To ensure a fair comparison, we experimented with the same parameter setting and the same dataset. Table 3 shows that our model outperforms Tsubaki's model on the Human and *C.elegans* datasets. The results show that replacing GNN with TrimNet make the DTI prediction more accurate.

TrimNet uses fewer parameters in comparison to previous best models

Besides the impressive performance on molecular representation learning, another remarkable achievement of TrimNet is the significant reduction of the number of the parameters. The original MPNN proposed by Gilmer *et al.* [13] and recent attentive FP [58] has achieved excellent results on quantum property prediction but suffers from a large number of parameters, which limited their applications in a broader context. Table 1 shows that our TrimNet model uses only 1/30 of attentive FP parameters in quantum property prediction tasks. And on drug discovery-related datasets, TrimNet also significantly reduced the parameters, as shown in Figure 3. The parameters of MPNN and attentive FP mainly come from the message function, in which the edge vector needs to be mapped to a $D \times D$ matrix (D represents the hidden size). In contrast, we employed the triplet attention and element-wise multiplication (Equations (2) and (3)) to avoid matrix mapping and decrease the parameters effectively. Due to less memory consumption, this lightweight advantage clearly implies that TrimNet should be more readily applicable to large graph databases.

TrimNet provides interpretability

As TrimNet performs well on a variety of task predictions, we investigate the model's interpretability, that is, how it arrives at a successful prediction. Usually, many DL methods tend to behave like a black box, the lack of explanation limits their applications, especially in some medical applications. Aiming to rationalize the TrimNet model, we visualized some molecules based on atomic attention weights. The readout phase of TrimNet, Set2Set [51] network, learns the final molecular representation by aggregating the atoms' information recurrently with atomic attention

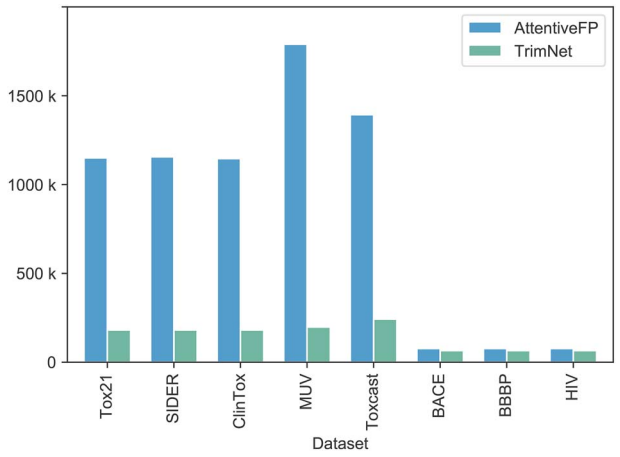

Figure 3. The total number of parameters of TrimNet and previous best model.

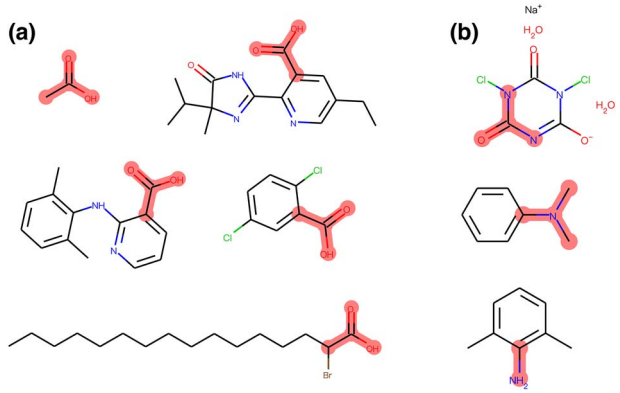

Figure 4. The atomic groups which model pay attention to.

weights. As an indispensable part of the network, atomic attention weights in Set2Set may provide interpretability from input molecules to output properties.

We extract the atomic attention in the Set2Set network for molecular visualization. To be specific, we detect the max-weight atom according to the attention and highlight the atom and its first-level neighbors (on the molecules drawn using RDKIT). Because of the message passing mechanism according to Section 2.2, its first-level neighbors contribute significantly and are considered for highlighting. Finally, in the toxicity prediction task on ClinTox and Tox21 datasets, we found that TrimNet model may pay more attention to atomic groups that may cause toxicity, such as aniline scaffold or carboxylic acid, as shown in Figure 4A and B. These visualization results show that our model detects essential atomic groups with a clear interpretability.

Ablation studies

Edge information. In the attentive edge network of our TrimNet, edge features are employed to calculate the attention weights, then element-wise multiplied by neighbor node representation and aggregated by the obtained weight. To evaluate the importance of edge features, we removed all edge features (using only node features) from the inputs and compared their performance. The experimental results indicated that the performance deteriorates dramatically after deleting the edge features, as shown in Table 4. To further validate the role of edge information, we only removed the edge representation in the element-wise multiplication (Equation (3)) but kept it in the process of attention score calculation (Equations (1) and (2)). The modified TrimNet again suffers a drop in performance. These results indicate that edge features play essential roles in the tested molecular properties.

Layer normalization. In TrimNet, we creatively employed layer normalization in every step of the message passing phase of the model to reduce gradients vanishment and explosion (Equation (5)). To the best of our knowledge, this is the first time layer normalization has been incorporated into MPNNs. Here, the layer normalization was removed to study the effect that this had on the performance of TrimNet. Table 4 shows that MAEs of TrimNet on the QM9 dataset incur significant increase when layer normalization are removed. In fact, the MAE rises more than 90% for five tasks, which suggests that the layer normalization is a valuable addition for the MPNN architecture.

Depth. Here, we investigate the effects of the depth of the message passing phase on TrimNet's performance. Table 5 displays the performance comparison of triplet with different depths on the QM9 dataset. As the depth increased, the performance improved. This positive trend of continuing improvement was observed on half of all tasks even when the depth increased to 10. These results indicate that TrimNet with more layers has a greater learning capacity, while most of GNNs are limited to very shallow models, usually no deeper than two or three layers [28, 59, 62]. TrimNet's success may be attributed to the skip connection adopted in the message phase to alleviate the over-smoothing and the vanishing gradient problem in GNNs [4, 63].

Discussion

In this paper, we show that TrimNet achieved impressive performance on molecular representation learning tasks with significant parameter reduction. Compared with the previous state-of-the-art attentive FP and MPNN, TrimNet explicitly dropped the matrix mapping of the edge information and employed the triplet-attentive edge network we proposed to reduce the parameters and enhance the edge information extraction. The edge features element-wise dot is the key component of triplet-attentive edge network, which may make up for the edge information loss caused by dropping the matrix mapping. Besides, we found that the layer normalization is necessary for the vertex representation update. However, although we reduced the parameters, TrimNet has nearly the same training time as MPNN when they have the same hyperparameters setting. That means TrimNet could not reduce the computation complexity (time) compared with MPNN. Nevertheless, the current computation time of TrimNet can fully satisfy the chemists' demands for chemistry discovery, and we believe that this is unlikely to become a bottleneck in the applications of our method.

Conclusions

In this work, we have proposed an DL approach employed a novel triplet message mechanism to learn molecular representation efficiently, named TrimNet. In particular, TrimNet achieves the new state-of-the-art performance on a variety of molecular properties, including quantum properties, bioactivity and physiology. In addition to the higher predictive accuracy and fewer parameters, TrimNet provides an interpretable learning with attention weights naturally focusing on crucial atoms and

Table 4. Ablation studies on QM9 dataset

Model		No norm[a]	No edge[b]	No edge multiply[c]	TrimNet
TASKS	μ	0.0726	0.661	0.0930	**0.0741**
	α	0.292	0.833	0.260	**0.216**
	$\epsilon HOMO$	0.00222	0.00678	0.00266	**0.00226**
	$\epsilon LUMO$	0.00197	0.00787	0.00250	**0.00192**
	$\Delta\epsilon$	0.00325	0.00990	0.00416	**0.00336**
	R2	40.824	83.734	9.510	**2.178**
	ZPVE	0.000150	0.000640	0.000230	**0.000140**
	U0	0.1871	0.1096	**0.0684**	0.0927
	U	0.1557	0.1018	**0.0715**	0.0861
	H	0.1987	0.0792	0.0861	**0.0774**
	G	0.1912	0.0875	0.0769	**0.0717**
	Cv	0.0876	0.4413	0.0815	**0.0715**

[a] Remove layer normalization in the message phase. [b] Remove edge feature using in the message phase. [c] Use $\alpha * x_j$ instead of $\alpha * e_{ij} * x_j$ (Equation (3)) in the message phase.

Table 5. The performances of TrimNet with different depth on QM9 dataset

Model information	Depth Total params	1 28k	3 56k	5 85k	7 114k	10 157k
TASKS	μ	0.0889	0.0741	0.0697	0.0675	0.0705
	α	0.255	0.216	0.176	0.177	0.170
	$\epsilon HOMO$	0.00266	0.00226	0.00219	0.00203	0.00213
	$\epsilon LUMO$	0.00235	0.00192	0.00190	0.00180	0.00174
	$\Delta\epsilon$	0.00384	0.00336	0.00310	0.00313	0.00301
	R2	3.771	2.178	1.826	1.584	2.830
	ZPVE	0.000173	0.000140	0.000140	0.000150	0.000130
	U0	0.1021	0.0927	0.0624	0.0485	0.0422
	U	0.0927	0.0861	0.0663	0.0448	0.0511
	H	0.0928	0.0774	0.0593	0.0465	0.0490
	G	0.0923	0.0717	0.0611	0.0530	0.0418
	Cv	0.0918	0.0715	0.0663	0.0647	0.0675

substructures responsible for the target properties. These results have established TrimNet as a powerful and useful computational tool in solving the challenge of molecular representation learning.

Key points

- TrimNet provides a new perspective from triplet messages for molecular representation learning. It employs a triplet message mechanism to calculate message from atom-bond-atom information and updates the hidden states of neural networks.
- TrimNet can significantly outperform current state-of-the-art methods on multiple molecular representation learning tasks, including molecular property predictions (quantum properties, bioactivity, physiology) and compound–protein interaction identification.
- Compared with the previous best method attentive FP, TrimNet reduced the parameters, even by more than an order of magnitude.
- TrimNet provides interpretability of molecular representation learning with attention weights naturally focusing on crucial atoms and substructures responsible for the target properties.

Supplementary data

Supplementary data are available at *Briefings in Bioinformatics* online.

Data Availability

The quantum and drug datasets are available on the website of MoleculeNet: http://moleculenet.ai. The source code is available in GitHub: https://github.com/yvquanli/trimnet.

Acknowledgments

The authors thank attentive FP's authors for help.

Funding

This work was supported in part by funds from the National Natural Science Foundation of China (21775060, 61872216, 61472205, 81630103, 31871071 and 61836004), the Turing AI Institute of Nanjing and the Beijing Brain Science Special Project (Z181100001518006).

References

1. Bahdanau D, Cho K, Bengio Y, et al. Neural machine translation by jointly learning to align and translate. In: *International Conference on Learning Representations*, Banff, Canada: ICLR Press, 2015.
2. Breiman L. Random forests. *Mach Learn* 2001; **45**(1): 5–32.
3. Butler KT, Davies DW, Cartwright H, *et al*. Machine learning for molecular and materials science. *Nature* 2018; **559**(7715): 547–55.
4. Chen D, Lin Y, Li W, Et al. Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. In: *AAAI Conference on Artificial Intelligence, New York*, Palo Alto, CA, USA: AAAI Press, 2020, 3438–3445.
5. Chen R, Liu X, Jin S, *et al*. Machine learning for drug-target interaction prediction. *Molecules* 2018; **23**(9): 2208.
6. Chung J, Gulcehre C, Cho KH *et al*. Empirical evaluation of gated recurrent neural networks on sequence modeling. In: arXiv preprint, arXiv:1412.3555, 2014. https://axiv.org/abs/1412.3555
7. Cortes C, Vapnik V. Support-vector networks. *Mach Learn* 1995; **20**(3): 273–97.
8. Curtarolo S, Hart GLW, Nardelli MB, *et al*. The high-throughput highway to computational materials design. *Nat Mater* 2013; **12**(3): 191–201.
9. Devlin J, Chang MW, Lee K, *et al*. BERT: pre-training of deep bidirectional transformers for language understanding. In: *The North American Chapter of the Association for Computational Linguistics, Minneapolis, Minnesota*, Stroudsburg PA1, USA: Association for Computational Linguistics, 2019, Vol. **1**, 4171–4186.
10. Duda, R. O., P. E. Hart, and D. G. Stork *Pattern Classification*. New York: John Wiley & Sons, 2012.
11. Feinberg EN, Sur D, Wu Z, *et al*. PotentialNet for molecular property prediction. *ACS Cent Sci* 2018; **4**(11): 1520–30.
12. Fey M, Lenssen JE. Fast graph representation learning with PyTorch Geometric. In: *ICLR Workshop on Representation Learning on Graphs and Manifolds*, New Orleans, Louisiana, 2019. ICLR Press.
13. Gilmer J, Schoenholz SS, Riley P, *et al*. Neural message passing for quantum chemistry. In :*International Conference on Machine Learning*, Sydney, Australia, 2017, 1263–1272. ICML Press.
14. Goh GB, Hodas NO, Vishnu A. Deep learning for computational chemistry. *J Comput Chem* 2017; **38**(16): 1291–307.
15. Graves A, Wayne G, Danihelka I *et al*. Neural Turing machines. In: arXiv preprint, arXiv:1410.5401, 2014. https://arxiv.org/abs/1410.5401
16. Hachmann J, Olivares-Amaya R, Atahan-Evrenk S, *et al*. The Harvard clean energy project: large-scale computational screening and design of organic photovoltaics on the world community grid. *J Phys Chem Lett* 2011; **2**(17): 2241–51.
17. He K, Zhang X, Ren S, *et al*. Deep residual learning for image recognition. In: *Computer Vision and Pattern Recognition*, Las Vegas, NV, 2016, pp. 770–778. IEEE.
18. Hessler G, Baringhaus K-H. Artificial intelligence in drug design. *Molecules* 2018; **23**(10): 2520.
19. Hohenberg P, Kohn W. Inhomogeneous electron gas. *Phys Rev* 1964; **136**(3B): B864.
20. Deng J, Dong W, Socher R, *et al*. Image Net: a large-scale hierarchical image database. In: *Computer Vision and Pattern Recognition*, Miami, FL, 2009, 248–55. IEEE.
21. Kearnes S, McCloskey K, Berndl M, *et al*. Molecular graph convolutions: moving beyond fingerprints. *J Comput Aided Mol Des* 2016; **30**(8): 595–608.
22. Kingma DP, Ba JL, Adam: a method for stochastic optimization. In: *International Conference on Learning Representations*, San Diego, CA, 2015. ICLR Press
23. Klicpera J, Groß J, Günnemann S *et al*. Directional message passing for molecular graphs. In: *International Conference on Learning Representations*, Addis Ababa, Ethiopia, 2020. ICLR Press
24. RDKIT: Open-source cheminformatics. In 2006. https:/ww.rdkit.org/
25. Lavecchia A. Machine-learning approaches in drug discovery: methods and applications. *Drug Discov Today* 2015; **20**(3): 318–31.
26. Lavecchia A. Deep learning in drug discovery: opportunities, challenges and future prospects. *Drug Discov Today* 2019; **24**(10): 2017–32.
27. Lecun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015; **521**(7553): 436–44.
28. Li G, Muller M, Thabet A *et al*. Deep GCNs: can GCNs goas deep as CNNs? In: *International Conference on Computer Vision*, Seoul, Korea, 2019, 9267–76. IEEE.
29. Li J, Deng C, He X *et al*. Learning graph-level representation for drug discovery. In: arXiv preprint, arXiv:1709.03741, 2017. https://axiv.or./abs/1709.03741.
30. Liu H, Sun J, Guan J, *et al*. Improving compound-protein interaction prediction by building up highly credible negative samples. *Bioinformatics* 2015; **31**(12): i221–9.
31. Liu K, Sun X, Jia L, *et al*. Chemi-net: a molecular graph convolutional network for accurate drug property prediction. *Int J Mol Sci* 2019; **20**(14): 3389.
32. Mater AC, Coote ML. Deep learning in chemistry. *J Chem Inf Model* 2019; **59**(6): 2545–59.
33. Mousavian Z, Masoudi-Nejad A. Drug–target interaction prediction via chemogenomic space: learning-based methods. *Expert Opin Drug Metab Toxicol* 2014; **10**(9): 1273–87.
34. Nørskov JK, Bligaard T, Rossmeisl J, *et al*. Towards the computational design of solid catalysts. *Nat Chem* 2009; **1**(1): 37–46.
35. Pyzer-Knapp EO, Suh C, Gómez-Bombarelli R, *et al*. What is high-throughput virtual screening? A perspective from organic materials discovery. *Annu Rev Mat Res* 2015; **45**(1): 195–216.
36. Ramakrishnan R, Dral PO, Rupp M, *et al*. Quantum chemistry structures and properties of 134 kilo molecules. *Sci Data* 2014; **1**:140022.
37. Ramakrishnan R, Dral PO, Rupp M, *et al*. Big data meets quantum chemistry approximations: the Δ-machine learning approach. *J Chem Theory Comput* 2015; **11**(5): 2087–96.
38. Rifaioglu AS, Atas H, Martin MJ, *et al*. Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases. *Brief. Bioinform.* **20** (2019): 1878–912.
39. Rogers D, Brown RD, Hahn M. Using extended-connectivity fingerprints with Laplacian-modified Bayesian analysis in high-throughput screening follow-up. *J Biomol Screen* 2005; **10**(7): 682–6.
40. Ryu S, Kwon Y, Kim WY. A Bayesian graph convolutional network for reliable prediction of molecular properties with uncertainty quantification. *Chem Sci* 2019; **10**(36): 8438–46.
41. Scarselli F, Gori M, Tsoi AC, *et al*. The graph neural network model. *IEEE Trans Neural Netw* 2009; **20**(1): 61–80.
42. Schneider G. Automating drug discovery. *Nat Rev Drug Discov* 2018; **17**(2): 97–113.

43. Schütt KT, Kindermans PJ, Sauceda HE *et al*. Sch Net: a continuous-filter convolutional neural network for modeling quantum interactions. In: *Advances in Neural Information Processing Systems*, Long Beach, CA, 2017, 992–1002. Neur IPS Press.
44. Schütt KT, Arbabzadah F, Chmiela S, *et al*. Quantum-chemical insights from deep tensor neural networks. *Nat Commun* 2017; **8**:6–13.
45. Svetnik V, Liaw A, Tong C, *et al*. Random forest: a classification and regression tool for compound classification and QSAR modeling. *J Chem Inf Comput Sci* 2003; **43**(6): 1947–58.
46. Tang B, Kramer ST, Fang M, *et al*. A self-attention based message passing neural network for predicting molecular lipophilicity and aqueous solubility. *J Chem* 2020; **12**(1): 1–9.
47. Tsubaki M, Tomii K, Sese J. Compound-protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics* 2019; **35**(2): 309–18.
48. Unke OT, Meuwly M. PhysNet: a neural network for predicting energies, forces, dipole moments, and partial charges. *J Chem Theory Comput* 2019; **15**(6): 3678–93.
49. Vamathevan J, Clark D, Czodrowski P, *et al*. Applications of machine learning in drug discovery and development. *Nat Rev Drug Discov* 2019; **18**(6): 463–77.
50. Vaswani A, Shazeer N, Parmar N, Attention is all you need. In: *Advances in Neural Information Processing Systems*, Long Beach, CA, 2017, 5999–6009. Neur IPS Press
51. Vinyals O, Bengio S, Kudlur M *et al*. Order matters: sequence to sequence for sets. In: *International Conference on Learning Representations*, San Juan, Puerto Rico, 2016. ICLR Press.
52. Weininger D. Smiles, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci* 1988; **28**(1): 31–6.
53. Wen M, Zhang Z, Niu S, *et al*. Deep-learning-based drug–target interaction prediction. *J Proteome Res* 2017; **16**(4): 1401–9.
54. Winter R, Montanari F, Noé F, *et al*. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chem Sci* 2019; **10**(6): 1692–701.
55. Withnall M, Lindelöf E, Engkvist O, *et al*. Building attention and edge message passing neural networks for bioactivity and physical-chemical property prediction. *J Chem* 2020; **12**(1): 1–18.
56. Wu Z, Ramsundar B, Feinberg EN, *et al*. MoleculeNet: a benchmark for molecular machine learning. *Chem Sci* 2018; **9**(2): 513–30.
57. Wu, Z., S. Pan, F. Chen, *et al*. "A comprehensive survey on graph neural networks." *IEEE Trans Neural Netw Learn Syst*. (2020): 1–21.
58. Xiong Z, Wang D, Liu X, *et al*. Pushing the boundaries of molecular representation for drug discovery with graph attention mechanism. *J Med Chem* 2019. page acs.jmedchem.9b00959.
59. Xu K, Li C, Tian Y *et al*. Representation learningon graphs with jumping knowledge networks. In: *International Conference on Machine Learning*, Stockholm, Sweden, 2018, Vol. **12**, 8676–8685. ICML Press.
60. Yang K, Swanson K, Jin W, *et al*. Analyzing learned molecular representations for property prediction. *J Chem Inf Model* 2019; **59**(8): 3370–88.
61. Zhang Y, Lee AA. Bayesian semi-supervised learning for uncertainty-calibrated prediction of molecular properties and active learning. *Chem Sci* 2019; **10**(35): 8154–63.
62. Zhang Z, Cui P, Zhu W. Deep learning on graphs: a survey. *IEEE Trans Knowl Data Eng* 2020; **14**(8): 1–1.
63. Zhao L, Akoglu L. PairNorm: tackling oversmoothing in GNNs. 2019. arXiv preprint arXiv:1909.12223.