Journal of Medicinal Chemistry

Article

Cite This: J. Med. Chem. XXXX, XXX, XXX−XXX

pubs.acs.org/jmc

Pushing the Boundaries of Molecular Representation for Drug Discovery with the Graph Attention Mechanism

Zhaoping Xiong,[†,‡,§] Dingyan Wang,[‡,§] Xiaohong Liu,[†,‡] Feisheng Zhong,[‡,§] Xiaozhe Wan,[‡,§] Xutong Li,[‡,§] Zhaojun Li,[‡] Xiaomin Luo,[‡] Kaixian Chen,[†,‡] Hualiang Jiang,*[,†,‡] and Mingyue Zheng*[,‡]

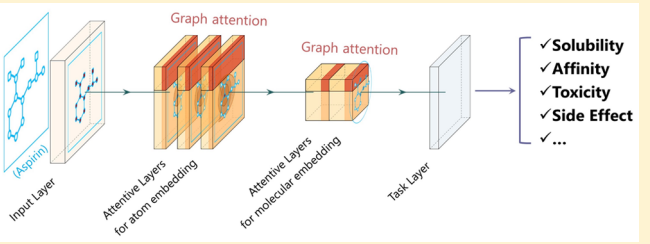
[†]Shanghai Institute for Advanced Immunochemical Studies, and School of Life Science and Technology, ShanghaiTech University, Shanghai 200031, China

[‡]Drug Discovery and Design Center, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, 555 Zuchongzhi Road, Shanghai 201203, China

[§]University of Chinese Academy of Sciences, No. 19A Yuquan Road, Beijing 100049, China

Supporting Information

ABSTRACT: Hunting for chemicals with favorable pharmacological, toxicological, and pharmacokinetic properties remains a formidable challenge for drug discovery. Deep learning provides us with powerful tools to build predictive models that are appropriate for the rising amounts of data, but the gap between what these neural networks learn and what human beings can comprehend is growing. Moreover, this gap may induce distrust and restrict deep learning applications in practice. Here, we introduce a new graph neural network architecture called Attentive FP for molecular representation that uses a graph attention mechanism to learn from relevant drug discovery data sets. We demonstrate that Attentive FP achieves state-of-the-art predictive performances on a variety of data sets and that what it learns is interpretable. The feature visualization for Attentive FP suggests that it automatically learns nonlocal intramolecular interactions from specified tasks, which can help us gain chemical insights directly from data beyond human perception.



INTRODUCTION

Efficient medicinal chemistry relies on associative reasoning and pattern recognition for molecular structures. However, the use of empirical "drug-likeness" rules and "privileged" chemical (sub)structures is failing because the low-hanging fruit has become scarcer. Even the most experienced medicinal chemists will have diverse preferences when prioritizing compounds to the next stage. The challenge for finding chemicals with favorable pharmacological, toxicological, and pharmacokinetic properties stems not only from the uncertainty of body biological systems but also from the intricate meanings of the information in chemical molecular systems, because humans are incapable of determining those properties directly from chemical structures. A molecular structure is usually composed of many-body interactions and complex electronic configurations, which make constructing their comprehensive representation a nontrivial issue. Given the rising amount of data and the complexity of chemical and biological systems, medicinal chemists have been working "at the edge of chaos" and in desperate need of augmented intelligence from AI.[1]

In the past few years, a large volume of data concerning the biological effects of chemical compounds has been accumulated and made publicly accessible.[1−6] In addition, an increasing number of large-scale high-quality quantum-chemical calculation results have been shared with the research community due to recent rapid advances in high performance computing (HPC). The increasing rate of data generation across all research disciplines related to drug discovery has provided unprecedented opportunities for understanding properties or actions that are useful for molecular design and for generating mechanistic hypotheses. Therefore, building machine learning models that can fit and predict big data from expensive biological assays and quantum-chemical calculations is of great interest. Many successful applications of machine learning methods have been reported[7−15] that outline the future of artificial intelligence in academia and industry. However, in contrast with fields such as image or voice recognition, generating suitable representations of chemical structures to extract the most relevant information regarding the properties of interest remains challenging. In this respect, molecular representation can be defined as a logical or mathematical procedure that transforms chemical information encoded within a molecular structure into a matrix of values. Molecular descriptors or "fingerprints" are frequently used for

Special Issue: Artificial Intelligence in Drug Discovery

Received: June 17, 2019
Published: August 13, 2019

DOI: 10.1021/acs.jmedchem.9b00959
J. Med. Chem. XXXX, XXX, XXX−XXX

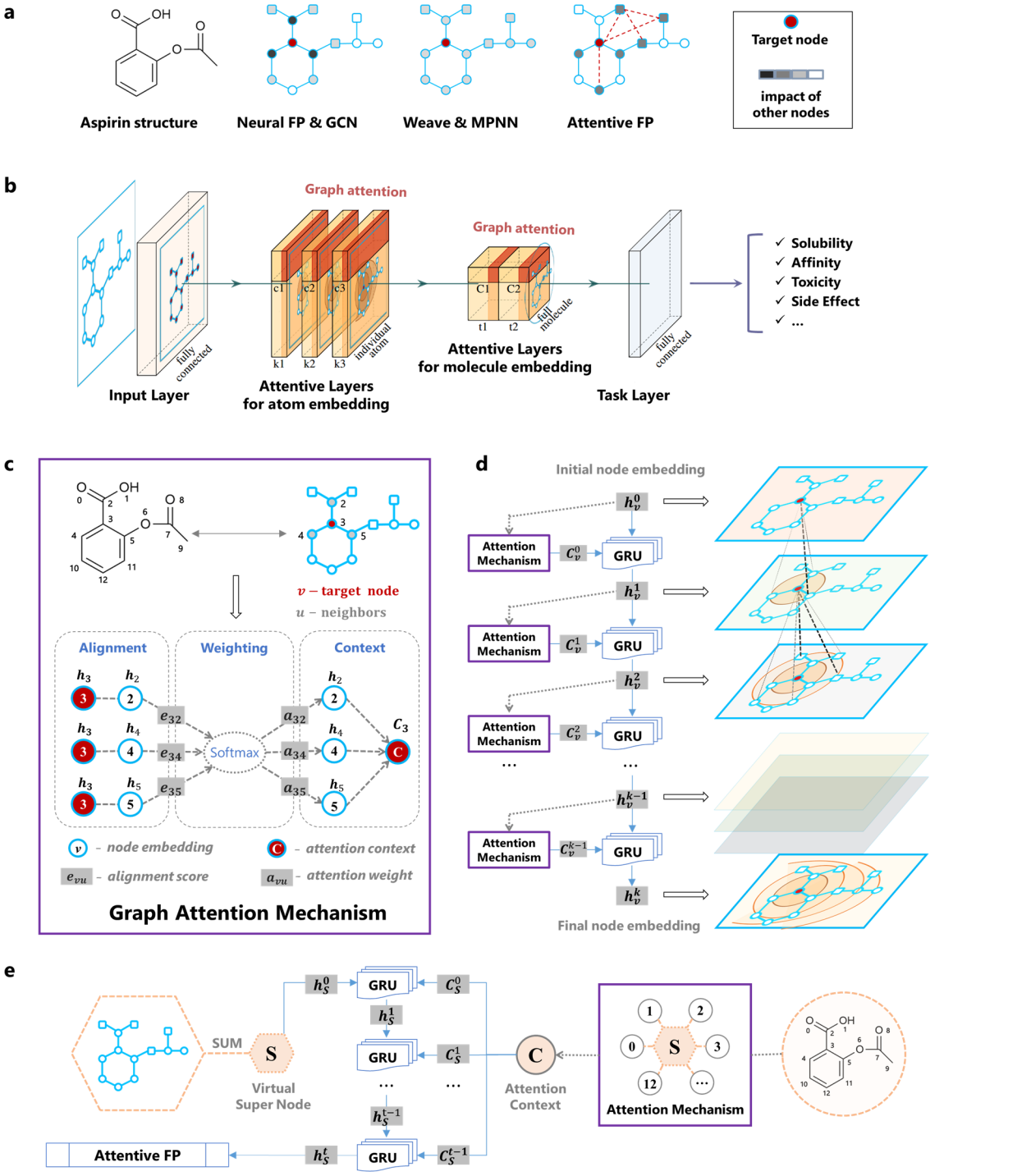


Figure 1. Molecular graph representation network architecture. (a) Compared to previous graph-based molecular representations, Attentive FP is more differentiable in evaluating the impact of neighbor atoms. (b) Overview of the Attentive FP network architecture. (c) Illustration of the graph attention mechanism on atom 3 in aspirin. (d) The framework that generates a state vector (embedding) for a target atom v. $\mathbf{h}_v^k$ and $\mathbf{C}_v^k$ are the state vector and attention context vector at time step k for atom $\mathbf{v}$, respectively. In higher time steps, target node embedding will include information from further nodes recursively. The darker dashed line implicates higher attention weight of the neighbor node. (e) Similar framework that generates the entire molecular graph embedding by assuming a super node connecting to all atoms in the molecular graph.

molecular representations. To date, over 5000 molecular descriptors have been designed to characterize chemical meaning.[16] The conventional machine learning approaches for QSAR/QSPR have revolved around feature engineering for these molecular descriptors[17−21] in which the goal is to select a subset of the relevant descriptors for use in model construction. According to their raw input form, these molecular representations can be divided into graph-based and geometry-based representations. Graph-based representations take only the information concerning the topological arrangement of atoms as input, while geometry-based representations employ the molecular geometry information, including bond lengths, bond angles, and torsional angles. In addition to the molecular descriptors or fingerprints designed

DOI: 10.1021/acs.jmedchem.9b00959
J. Med. Chem. XXXX, XXX, XXX−XXX

by chemists, increasing numbers of molecular representations have been automatically generated by deep learning models from simple raw inputs. For example, there has been a surge in molecular representations learned from deep neural network models by fitting the quantum-chemical calculations to simple raw inputs.[22−25]

Although molecular representation would seem to benefit from a priori knowledge of a molecule's three-dimensional (3D) conformation, pragmatic considerations such as calculation cost, alignment invariance, and uncertainty in conformation generation limit the use of geometry-based representations. For example, for most drug discovery applications, the active conformation of a small molecule in a given binding process is usually unknown. In such cases, graph-based molecular representations are more suitable; however, the gaps between these two classes of molecular representations usually lack transferability and cannot predict properties interchangeably. Therefore, the question arises of whether a neural network architecture applied to a molecular graph might bridge this gap and make molecular representations more generalizable.

Molecular structures usually involve many-body interactions and complex electronic structures, but molecular graphs reduce the representation complexity, with nodes and edges representing atoms and bonds, respectively. The molecular graph assumes that the key interactions among nuclei and electrons in a molecule can be implicitly captured by a graph that provides a source of insight into the geometries, functions, and properties of the molecule. Recently, substantial progress has been made in designing neural network architectures that learn representations from graph structured data.[26−29] The underlying principle of these architectures is to learn a form of mapping (also called an embedding) of nodes and edges that fully captures the graph information, particularly for inferring relations between nodes. Compared with previous graph topology representation approaches, the recent neural network approaches are much more powerful at capturing the nonprominent patterns, and they require less feature engineering effort.

Figure 1a encapsulates the recent neural graph representation for molecules. Given a target node, the gray nodes indicate the probability of neighboring nodes impacting the target node in different graph-based molecular representation schemes. The darker the node color is, the greater the chance of it impacting the target node. For the Neural FP and GCN models,[26,27,29] the neighbor nodes' chances to influence the target decrease with topological distance during the recursive propagation procedure. In chemical molecules, atomic pairs that are topologically distant may also have significant interactions and hence affect the overall molecular properties. A desirable molecular graph representation framework should be capable of capturing the information contained among even distant atoms in a molecule, such as intramolecular hydrogen bonding. More recently, Weave[28] and MPNN[30] were proposed to construct virtual edges linking every pair of nodes in a molecule graph, meaning that any nodes, regardless of their distance to a target node, have an equal chance to exert influence, similar to the direct neighbors of the target node. Under these schemes, all atoms can influence one another with no distance limitations. For complex graphs such as social networks, these schemes work well for describing information that flows freely from node to node. However, for molecular graphs, intrinsic structures exist that are governed by physical laws, and the information flow among nodes is also constrained. Topologically adjacent nodes have a greater chance to impact each other, and they can form functional groups in some cases. In this sense, the Weave and MPNN representation schemes tended to make all the neighbors' impacts weak because of their averaging effect.

Here, we propose a new graph-based neural network architecture, Attentive FP, to represent molecules. Attentive FP not only characterizes the atomic local environment by propagating node information from nearby nodes to more distant ones but also allows for nonlocal effects at the intramolecular level by applying a graph attention mechanism. In this way, our resulting Attentive FP is powerful enough to capture the hidden critical linkage among any nodes efficiently under the premise of respecting the intrinsic molecule topological structure (Figure 1a). Benchmark tests of Neural FP and MoleculeNet are used here to perform an unbiased performance evaluation. Our Attentive FP achieves state-of-the-art results in modeling a wide range of molecular properties involving not only physical chemistry but also biophysics and physiology. More strikingly, it shows a predictive power comparable with recently reported geometry-based representations on the qm9 data sets—even without using molecular conformation data provided in advance. Furthermore, the visualization of the learned graph connections and node features agrees well with our intuition of chemical molecular structure and also reveals that Attentive FP can indeed extract nonlocal intramolecular interactions that are intractable for most graph-based representations.

METHODS

Graph Attention Mechanism. An attention mechanism allows a method to focus on task-relevant parts of a neural network. It has become a routine to apply the attention mechanism for tasks with sequence-structured data to allow the model to focus on the most relevant parts of the inputs and achieve a better prediction.[31] Recently, Velickovic and Bengio et al. proposed graph attention networks (GATs) that extend the attention mechanism to graph-structured data for node classification tasks.[32] The core idea of applying the attention mechanism to the graph is to obtain a context vector for the target node by focusing on its neighbors and local environment. The process can be categorized into three operations, (1) alignment, (2) weighting, and (3) context, as formulated below

Alignment:

$$\boldsymbol{e}_{vu} = leaky\_relu(\boldsymbol{W}\cdot[\boldsymbol{h}_v, \boldsymbol{h}_u]) \tag{1}$$

Weighting:

$$\boldsymbol{a}_{vu} = softmax(\boldsymbol{e}_{vu}) = \frac{\exp(\boldsymbol{e}_{vu})}{\sum_{u\in N(v)}\exp(\boldsymbol{e}_{vu})} \tag{2}$$

Context:

$$\boldsymbol{C}_v = elu\left(\sum_{u\in N(v)} \boldsymbol{a}_{vu}\cdot\boldsymbol{W}\cdot\boldsymbol{h}_u\right) \tag{3}$$

where $\boldsymbol{v}$ is the target node (a specific atom) and $\boldsymbol{h}_v$ is the state vector of node $\boldsymbol{v}$, as is $\boldsymbol{h}_u$ to the neighbor node (neighbor atom) $\boldsymbol{u}$. Here, *leaky_relu and elu*, variations of the *relu* nonlinear activation function, are used because they are more expressive and could consistently perform better by allowing a nonzero slope for the negative part of the *relu* function.[33] The procedures are as follows: (1) During the alignment operation, $[\boldsymbol{h}_v, \boldsymbol{h}_u]$ concatenates the state vectors of the target node and a neighbor node, followed by a linear transformation with a trainable weight matrix $\boldsymbol{W}$. $\boldsymbol{e}_{vu}$ is the output of the alignment

DOI: 10.1021/acs.jmedchem.9b00959
J. Med. Chem. XXXX, XXX, XXX−XXX

Table 1. Initial Atomic and Bond Features

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

operation for each target−neighbor pair. (2) During the weighting operation, e_{vu} is further normalized using the softmax function over the neighbor nodes, and it obtains a_{vu}, the importance (weight) of neighbor node u to target node v. (3) During the context operation, a linear transformation is performed on h_u, the state vectors of the neighbor nodes, followed by a weighted sum and a nonlinear activation function, which obtains C_v, the context vector of the target node v.

Graph Neural Networks. Neural networks typically introduce a nonlinear activation function after a linear transformation; this approach is capable of approximating any function but does not necessarily ensure the generalizability of the resulting model. Designing a feasible neural network that generalizes well in a specific domain is required to solve practical problems. Conceptually, graph neural networks (GNNs) extend neural network methods to process graph-structured data such as social networks, recommender systems, biological protein−protein interaction (PPI) networks, and molecular graph structures.[34] GNNs encompass an iterative procedure using recursive neural networks (RNNs) that agglomerates the "messages" of nodes from nearby to distant. As per the existing GNN architectures[35] which include a messaging phase and a readout phase, our Attentive FP is formulated as follows

Messaging:

$$C_v^{k-1} = \sum_{u \in N(v)} M^{k-1}(h_u^{k-1}, h_v^{k-1}) \tag{4}$$

Readout:

$$h_v^k = \mathrm{GRU}^{k-1}(C_v^{k-1}, h_v^{k-1}) \tag{5}$$

where h_v^k is the state vector of target node v after k iterations and $N(v)$ represents all of the neighbors of node v. In the messaging phase, the neighbor nodes aggregate information for the target node by applying the graph attention mechanism, M^{k-1}, which is referred to as the message function at iteration $k-1$. In the readout phase, GRU^{k-1} (gated recurrent unit), referred to as the update function at iteration $k-1$, takes the input of the previous state vector of the target node h_v^{k-1} and the "messages", attention context C_v^{k-1} from the neighbors and updates the previous state to the current state h_v^k.

Molecular Featurization. Node features first need to be defined before encoding a graph. Here, we use a total of nine types of atomic features and four types of bond features to characterize atoms and their local environment (Table 1). Most of these features are encoded in a one-hot fashion, except for formal charge and radical electron number, which are encoded as integers due to their additive nature. To create a one-hot encoding feature, all of the candidate categorical variables of the feature are listed and marked as either 1 or 0 (one-hot or all null) by their matches to those variables. For example, a vector of 16 bits is defined to encode atomic symbols, and a vector of 6 bits is defined to encode the hybridization state. Of note is that atomic chirality is encoded in three different bits: one indicates whether the atom is a chiral center, and the other two bits define whether it is in R- or S-form. Moreover, stereo types of double bonds are represented by a feature that distinguishes their potential E/Z configurations.

Attentive FP Network Architecture. Here, we propose a new graph neural network architecture for molecular representation, called Attentive FP, which introduces an attention mechanism for extracting nonlocal effects at the intramolecular level. This attention mechanism allows a method to focus on the most relevant parts of the inputs to achieve better prediction.[31] Figure 1 summarizes the architecture of the Attentive FP network: (1) We assume that a molecule, its bond features, and its atomic features are extracted with RDkit and encoded according to Table 1. Because the model is atom-centric, each atom has its own neighbor features that concatenate both neighboring atoms and the connecting bond features. Notably, the vectors of the atomic features and the neighboring atomic features do not have the same length; consequently, linear transformation and nonlinear activation were performed to unify the vector length. This procedure actually forms a fully connected layer and generates initial state vectors ("embeddings") for each atom and its neighbors. (2) Then, to include more information from the local environment, those initial state vectors are further embedded with stacked attentive layers for node embedding, allowing the atom to progressively aggregate "messages" from its neighborhoods using an attention mechanism allowing an atom to focus on the most relevant "messages" in its neighborhood. In each node embedding attentive layer, a new state vector is generated for each atom. After passing through several stacked attentive layers, the state vector includes more neighborhood information. (3) To combine the individual atom state vectors into a full-molecule state vector, we treat the entire molecule as a supervirtual node that connects every atom in a molecule and is embedded using the same atom embedding attention mechanism. This process is performed on stacked attentive layers for molecule embedding and generates a state vector for the whole molecule. (4) The final state vector is the learned representation that encodes structural information about the molecular graph, followed by a task-dependent layer for prediction. The entire network is trained in an end-to-end fashion, obtaining either a set of specific network weight parameters for a specific task or for multiple tasks simultaneously (Supplementary Table 3 summarizes the algorithm implementation pseudocode and the formulas for the Attentive FP neural network).

Attentive Layers on a Graph. The full network architecture for the attentive layers is illustrated in Figure 1d and e. As shown, the Attentive FP molecular representation scheme uses two stacks of attentive layers to extract information from the molecular graph. Specifically, one stack (with k layers) is for atom embedding (Figure 1c), and the other (with t layers) is for full-molecule embedding (Figure 1e). For the molecule embedding, all of the atom embeddings are aggregated by assuming a super virtual node that connects all the

DOI: 10.1021/acs.jmedchem.9b00959
J. Med. Chem. XXXX, XXX, XXX−XXX

atoms of the molecule. The attention mechanism is introduced in both the individual-atom embedding and the full-molecule embedding steps. For atom embedding, given a target atom v, its initial atom state vector h_v^0 is generated by a fully connected layer that includes only the initial atom and bond features. To better represent atom v, we introduce a graph attention mechanism at each layer that incorporates information from its neighborhoods $N(v)$. The graph attention mechanism in layer 0 takes the current state vector of target atom h_v^0 and its neighbors h_u^0 as inputs. That state vector is then subjected to alignment and weighting to obtain the attention context (C_v^0) of atom v. The output attention context is fed into the GRU (gated recurrent unit) together with the target atom's current state vector h_v^0, producing the updated state vector h_v^1 of atom v.

More intuitively, the graph attention mechanism in a single attentive layer is illustrated in Figure 1c. When applying attention to atom 3, the state vector of atom 3 is aligned with the state vector of its neighbors 2, 4, and 5, in which the features of connecting bonds have also been embedded by a fully connected layer. Then, the weight that measures how much attention we want to assign to the neighbors is calculated by a softmax function. Next, a weighted sum of the neighborhood information C_3 is obtained as the attention context vector of atom 3. These attention operations help the model to focus on task-relevant information from the local environment of the target atom. Lastly, C_3 (the attention context of atom 3) is fed into a GRU recurrent network unit together with the state vector h_3 of atom 3. The GRU used here is a variant of an LSTM (long short-term memory) recurrent network unit, which has shown good performance in retaining and filtering information using simplified update and reset gates[36] (Supplementary Figure 1). This scheme allows relevant information to be passed down without too much attrition,[37] which, in our case, means that the implicit linkages among distant atoms can still make a difference as long as they are related to the learning task. This property is what we desired to achieve for molecular representation.

Overall, in our well-designed attentive layer, the attention mechanism allows the target node to focus on the most related information from its neighbors, and the GRU recurrent network unit ensures that the information gets passed down to the neighborhoods efficiently during the update iterations. To perform molecule-level embedding, we assume a virtual supernode that connects to all of the atoms in the molecular graph; thus, the entire molecule can be embedded in the same fashion as the individual atoms (Figure 1e).

Data Sets and Benchmark Tests. For comparison purposes, we trained and tested our Attentive FP model with data sets that have previously been benchmarked. The first collection of data sets was benchmarked by Duvenaud et al.[26] and Kearnes el al.[28] and includes three different data sets spanning solubility, malaria bioactivity, and photovoltaic efficiency. The second collection of data sets was benchmarked by Wu et al.[30] in MoleculeNet. We tested all of the physical chemistry, biophysics, and physiology data sets collected by that paper except for the PDBbind data sets, which require a representation for the interaction between ligand and receptor and thus is out of the scope of this work (Supplementary Table 1). The third collection of data sets comes from quantum mechanical calculations. We tested our model on the largest qm9 data set[38,39]—the most comprehensive accessible quantum mechanical data set—which has been tested and benchmarked by previous models[30] (Supplementary Table 2).

The measurements in these data sets can be quantitative or qualitative. Generally, we built regression models for the quantitatively measured data sets and classification models for the qualitatively measured data sets. In addition, we adopted different performance metrics to use in the comparisons with previous benchmarks. Practically, regression models are evaluated by MAE (mean absolute error), MSE (mean-square error), or RMSE (root-mean-square error), and classification models are evaluated by AUC (area under the ROC (receiver operating characteristic) curve) or by the PRC (precision-recall curve). Here, we evaluated all of the classification models by the ROC, except for the models built on MUV (maximum unbiased validation) data sets, in which each task has 30 structure-distinct active compounds and each active compound has 500 structure-similar inactive compounds (the number of inactive compounds is 500 times that of the active compounds). These data sets are also evaluated using the PRC, which is a better metric of an algorithm's performance on highly imbalanced data sets.[40] The predictive model can be single-task or multitask. For the multitask models, we calculate the performance metrics for each individual task and report their average values.

Bayesian Optimization for Hyper-Parameter Search. Tuning hyper-parameters is a challenging step in deep neural network modeling, especially for complex network architectures. Here, we use Bayesian optimization (BO) to find the appropriate sets of hyper-parameters due to its efficiency and much reduced time consumption when faced with the increasing flexibility of neural network architectures.[41,42] A BO search mimics the manual search approach using Gaussian process simulation. The intuition underlying BO is to select the next set of hyper-parameters to evaluate on the basis of past results, similar to an expert with domain expertise. Thus, BO focuses the search process on the most promising sets of hyper-parameters. We use the Python package pyGPGO[43] in this study. For each unrelated data set, we performed a new BO-based hyper-parameter search in which we optimized the following six hyper-parameters simultaneously: k (the number of attentive layers for atom embedding), t (the number of attentive layers for molecule embedding), fingerprint dimension, L2 weight decay, learning rate, and dropout rate. We used the Matern32 kernel as the covariance function and UCB (upper-confidence bound) as the acquisition strategy.[44]

Early Stopping. Early stopping[45] was used to avoid overfitting and reduce training time consumption. When searching hyper-parameters using BO, a training process is required to obtain the best performance that the current set of hyper-parameters can reach. In this training process, we set a maximum epoch of 800, and if the performance metric had not improved in 10 epochs on the training set and in 15 epochs on the validation set, the training process was terminated early. However, these two early termination criteria are empirical and could change on the basis of the data volume and tasks. Supplementary Figure 2 illustrates the rationale for early termination. Based on the current set of hyper-parameters, early stopping will return the favorable performance metric on the validation set, which is further fed back to the BO algorithm to search for the next set of hyper-parameters.

Training Protocol. Attentive FP was trained with the Pytorch[46] framework using the Adam[47] optimizer for gradient descent optimization. The best set of hyper-parameters for each category of tasks obtained from the previous Bayesian optimization process was used to train the most predictive model. We used MSELoss and CrossEntropyLoss, which measure mean-squared error and cross entropy as loss functions for the regression tasks and classification tasks, respectively. All of the models were trained until an early termination criterion was reached, indicating that the performance improvement had converged.

RESULTS AND DISCUSSION

Proof of Concept Experiments with the Data Sets Benchmarked by Duvenaud et al. Molecular representations are usually evaluated by their predictive performances. A good molecular representation is expected to extract intrinsic and useful information that improves the predictive performance on a variety of tasks. To evaluate the predictive performance of Attentive FP, we first tested it on a collection of three distinct data sets benchmarked by Duvenaud et al.,[26] whose molecular properties span solubility, malaria bioactivity, and photovoltaic efficiency. Because these data are quantitative, we built regression models using MSE as the evaluation metric as in previous research. BO is used to search for the hyper-parameters for each data set to minimize the MSE (exemplified in Supplementary Table 4). Using the best set of

DOI: 10.1021/acs.jmedchem.9b00959
J. Med. Chem. XXXX, XXX, XXX−XXX

Table 2. Prediction Performance on Three Various Tasks

data sets	solubility, $\log_{10}$(mol/L)	malaria bioactivity, $\log_e$(μmol/L)	photovoltaic efficiency (%)
sample variance	4.32 ± 0.45	1.52 ± 0.06	6.41 ± 0.08
*ECFP + linear layer	1.71 ± 0.13	1.13 ± 0.03	2.63 ± 0.09
*ECFP + neural net	1.40 ± 0.13	1.36 ± 0.10	2.00 ± 0.09
*Neural FP	0.52 ± 0.07	1.16 ± 0.03	1.43 ± 0.09
*Weave	0.46 ± 0.08	1.07 ± 0.06	1.10 ± 0.06
MPNN	0.35 ± 0.05	1.10 ± 0.11	1.03 ± 0.12
Attentive FP	**0.28 ± 0.03**	**0.99 ± 0.08**	**0.82 ± 0.07**

*The values of the starred models are taken from the ref 26.

Table 3. Predictive Performances on Data Sets Relevant to Drug Discovery

category	data sets	no. of compounds	metrics	splitting	*previous best	Attentive FP
physical chemistry	ESOL	1128	RMSE	random	MPNN: 0.58	**0.503**
	FreeSolv	643	RMSE	random	MPNN: 1.15	**0.736**
	Lipop	4200	RMSE	random	Neural FP: 0.655	**0.578**
bioactivity	MUV	93,127	PRC	random	MultiTask: 0.184	**0.221**
			ROC	random	GC: 0.775	**0.843**
	HIV	41,913	ROC	scaffold	SVM: 0.792	**0.832**
	BACE	1522	ROC	scaffold	RF: **0.867**	0.850
physiology or toxicity	BBBP	2053	ROC	scaffold	SVM: 0.729	**0.920**
	Tox21	8014	ROC	random	Neural FP: 0.829	**0.858**
	ToxCast	8615	ROC	random	Weave: 0.742	**0.805**
	SIDER	1427	ROC	random	RF: **0.684**	0.637
	ClinTox	1491	ROC	random	Weave: 0.832	**0.940**

*The values of the starred models are taken from the ref 30.

hyper-parameters, we performed three independent runs with different random seeds to train the model. The results are shown according to their originally reported forms for comparison.

Table 2 and Supplementary Figure 4 summarize the predictive performances of Attentive FP and previous models. Neural FP outperforms the simple neural network models that take ECFP as input features, which demonstrates the potential of graph neural networks in predictive tasks. However, as described in the original paper,[26] Neural FP has a few limitations, such as the information propagation across the molecular graph and an inability to distinguish stereoisomers. In contrast, Weave and MPNN models use an edge network to construct virtual links between every pair of atoms in a molecule. This approach helps to detect implicit interactions between distant atoms, and their predictive performance is considerably improved compared to Neural FP. To better distinguish the impacts of different atoms, a GRU (gated recurrent unit) was introduced in the MPNN model to control information flow during iteration, contributing to a further performance improvement. In our work, Attentive FP uses a simple scheme to extract atomic features and topological relationships in the molecular graph under the premise of distinguishing molecules using features that are as concise as possible. For example, Attentive FP distinguishes stereoisomers by adding chirality to atomic features and stereo to bond features but eliminates partial charge and ring size features that are empirical or can be derived easily from other features. More importantly, the introduction of the graph attention mechanism allows our model to focus on task-related information from neighborhoods, while the GRUs and the state update function are helpful in filtering out unrelated information during the iteration process. Together, these network architecture designs contribute to the state-of-the-art predictive performance that Attentive FP achieves on those data sets.

Predicting Different Bioactivities and Properties for Drug Discovery. While Attentive FP showed impressive results on the three benchmark data sets from Duvenaud et al., there are many more challenging molecular machine-learning tasks for drug discovery. These include (but are not limited to) bioactivity, physical chemistry, and quantum-mechanical properties.[2,6,20] In this section, we summarize the predictive results of Attentive FP on data sets covering a variety of molecular properties and bioactivities. As recommended previously,[30] HIV, BACE, and BBBP data sets were evaluated under scaffold splitting, and the rest of the data sets were randomly split, with the ratio of train:valid:test at 8:1:1. As described above, BO was used for the hyper-parameter search for each data set. Using the best set of hyper-parameters, three independent runs with different random seeds were performed to train the models, and the results are shown according to their originally reported forms for comparison purposes. (For more information, see Supplementary Table 6.)

Physical chemistry properties measured by water solubility, solvation free energy, and lipophilicity greatly influence the pharmacokinetic profile of a drug, such as its absorption and distribution in the human body. Desirable physical chemistry properties are the premise of a successful drug, and predicting those properties quickly in silico with high precision can significantly reduce the experimental cost of drug development. As reported in Table 3, Attentive FP achieves the lowest RMSE on all previous benchmarked physical chemistry data sets, including water solubility (ESOL), solvation free energy (FreeSolv), and lipophilicity (Lipop).

The bioactivity data summarized here describe either direct or indirect effects of chemical compounds toward different targets that are key to drug discovery. Vast amounts of such

DOI: 10.1021/acs.jmedchem.9b00959
J. Med. Chem. XXXX, XXX, XXX−XXX

Table 4. Predictive Performances on the qm9 Data Set Quantum Properties (MAE)

			geometry-based		graph-based			
task	unit	sample MAD	*CM	*DTNN	*ECFP	*GC	*MPNN	Attentive FP
mu	D	1.189	0.519	**0.244**	0.602	0.583	0.358	0.451
alpha	b^3	6.299	0.85	0.95	3.1	1.37	0.89	**0.492**
HOMO	hartree	0.016	0.00506	0.00388	0.0066	0.00716	0.00541	**0.00358**
LUMO	hartree	0.039	0.00645	0.00513	0.00854	0.00921	0.00623	**0.00415**
gap	hartree	0.040	0.0086	0.0066	0.01	0.0112	0.0082	**0.00528**
R2	b^2	202.017	46	**17**	125.7	35.9	28.5	26.839
ZPVE	hartree	0.026	0.00207	0.00172	0.01109	0.00299	0.00216	**0.00120**
U0	hartree	31.072	2.27	2.43	15.1	3.41	2.05	**0.898**
U	hartree	31.072	2.27	2.43	15.1	3.41	2	**0.893**
H	hartree	31.072	2.27	2.43	15.1	3.41	2.02	**0.893**
G	hartree	31.072	2.27	2.43	15.1	3.41	2.02	**0.893**
Cv	cal/mol/K	3.204	0.39	0.27	1.77	0.65	0.42	**0.252**

*The values of the starred models are taken from the ref 30.

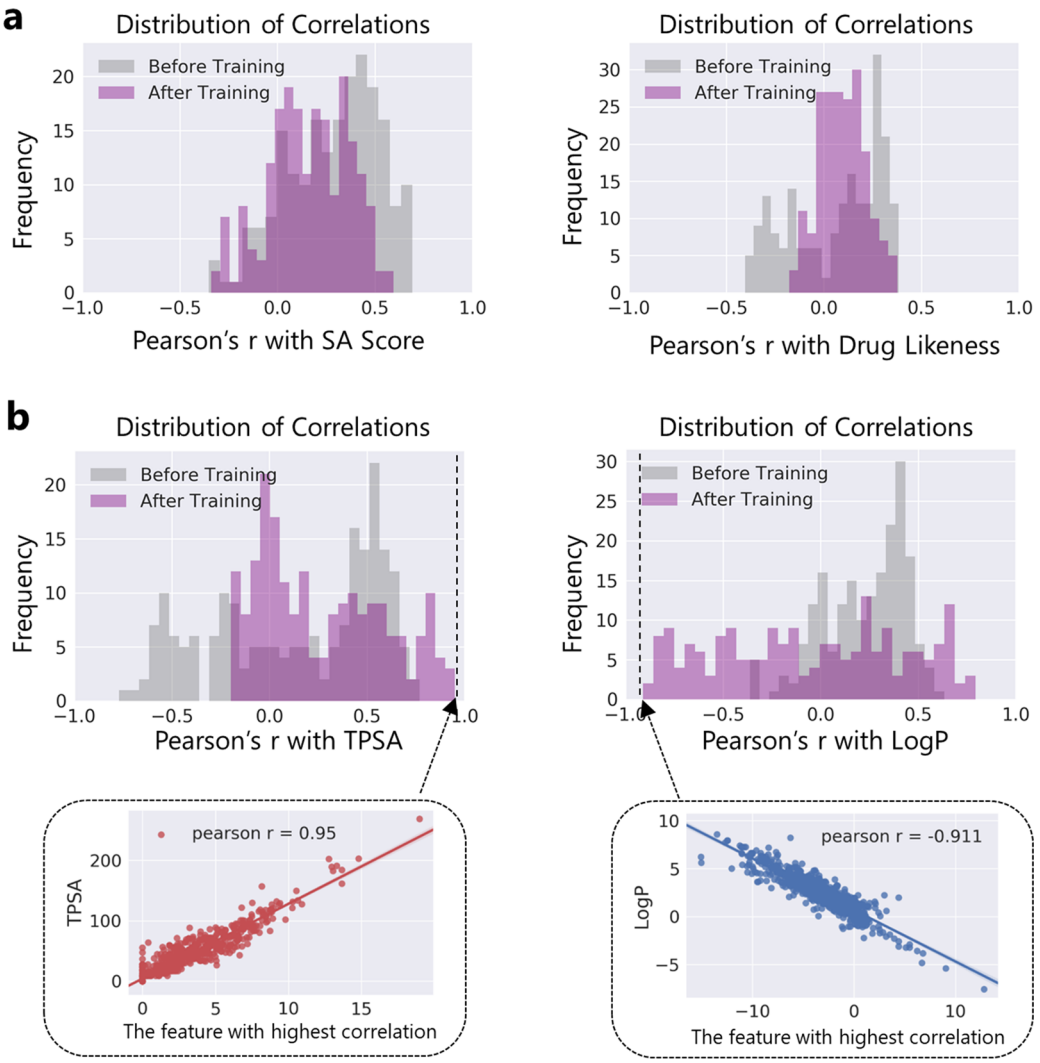


Figure 2. Comparing Attentive FP learned features with hand-crafted descriptors in the task of water solubility prediction. (a) The correlations of Attentive FP-learned features with two task-unrelated descriptors: SA (synthesis accessibility) score and drug likeness. The Pearson's *r* distributions do not change much before and after training, implying a weak relevance of the SA score and drug likeness to water solubility. (b) Conversely, the distribution of correlations to TPSA and LogP are skewed toward the extreme ends after training, suggesting the high relevance of TPSA and LogP to water solubility.

data have been accumulated in the public domain; thus, learning from these data provides a cost-effective way to perform drug candidate screening. Table 3 also presents the predictive performances of the classification models on

DOI: 10.1021/acs.jmedchem.9b00959
J. Med. Chem. XXXX, XXX, XXX−XXX

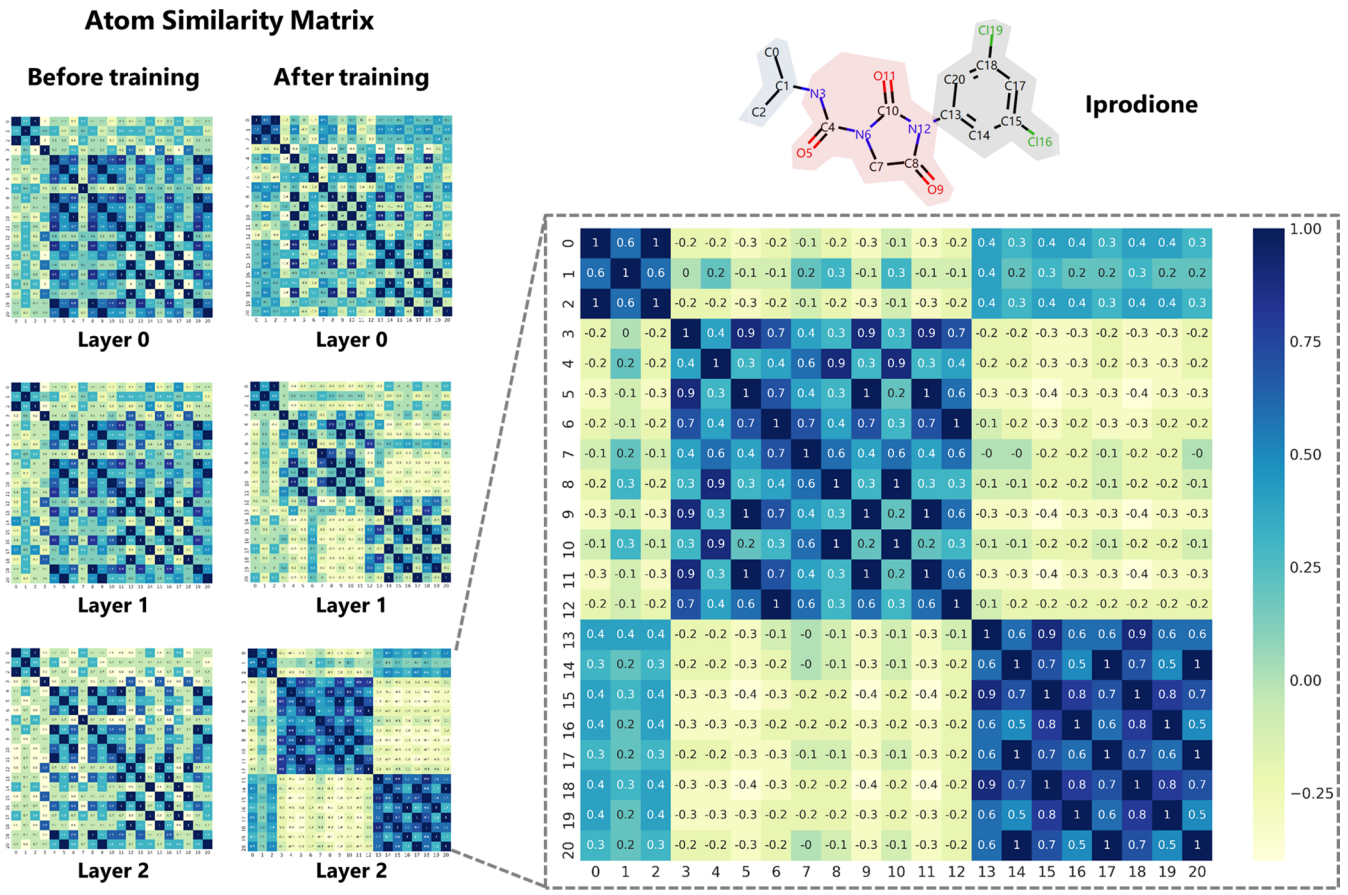


Figure 3. Heat maps of the atom similarity matrix for the compound Iprodione. The similarity scores are annotated in the corresponding squares and indicated by a color scheme. The atoms in Iprodione are automatically separated into three clusters during the learning process.

biophysics data sets (MUV, HIV, BACE). Attentive FP also shows noticeable improvements in terms of the ROC metric.

Physical chemistry properties and biophysics bioactivities only indicate how likely small molecules are to have an effect on living bodies, while physiology and toxicity data sets represent the effects a molecule has in living bodies, such as blood brain barrier permeability (BBBP), adverse effects (SIDER), or toxicities (Tox21, Toxcast, ClinTox). The Attentive FP models outcompete previous models on the BBBP, SIDER, Tox21, Toxcast, and ClinTox data sets—the only exception is SIDER, on which the random forest model still results in a slightly higher performance.

Overall, Attentive FP achieves the new state-of-the-art performance on 10 out of 12 drug-discovery-related data sets, suggesting that it is a promising molecular representation scheme for drug discovery problems.

The values of the starred models in Table 4 are taken from the original papers[30] and can be reproduced in Deepchem. The best values are highlighted in bold. Data is randomly split into train, validation, and test sets with 80, 10, and 10% of whole data sets, respectively (training size ∼104k). The MAD (mean absolute deviation) of a sample can also be interpreted as the MAE by predicting directly from the mean value of the samples. (U0 here is referred to as internal energy rather than atomization energy that many papers referred to).

Furthermore, fitting the quantum mechanical calculations with machine learning methods has attracted considerable interest because of the huge computing costs involved when using DFT methods. We tested Attentive FP on the qm9 quantum mechanical data sets benchmarked by MoleculeNet, which include 12 calculated quantum properties for 134k stable small organic molecules composed of up to nine heavy atoms (C, O, N, and F). As shown in Table 4, our graph-based Attentive FP molecular representation scheme outcompetes other models on 10 out of 12 tasks in the qm9 data sets. DTNN showed the best results on task R2 (electronic spatial extent) and task mu (norm of the dipole moment). Note this test result is quite encouraging because Attentive FP achieves an overall performance comparable to that of the geometry-based models, indicating that Attentive FP implicitly learned 3D conformation-related information. Thus, the molecular representation by Attentive FP may provide a valuable workaround that addresses many of the problems involving explorations of the large conformational space of molecules. (For more information, see Supplementary Table 5.)

FEATURE VISUALIZATION AND INTERPRETATION

The models using Attentive FP achieved state-of-the-art performance on a variety of tests; therefore, it is worth exploring the interpretability issue. Interpretability matters for two main reasons: (1) first, the "black box" nature of a deep learning model makes it very difficult to map what a machine has learned (network connection weights) to scientific domain knowledge efficiently; (2) second, a deep learning model might reveal interesting patterns hidden under the data, and these might be similar to or distinct from existing chemical observations and intuitions.[7]

DOI: 10.1021/acs.jmedchem.9b00959
J. Med. Chem. XXXX, XXX, XXX−XXX

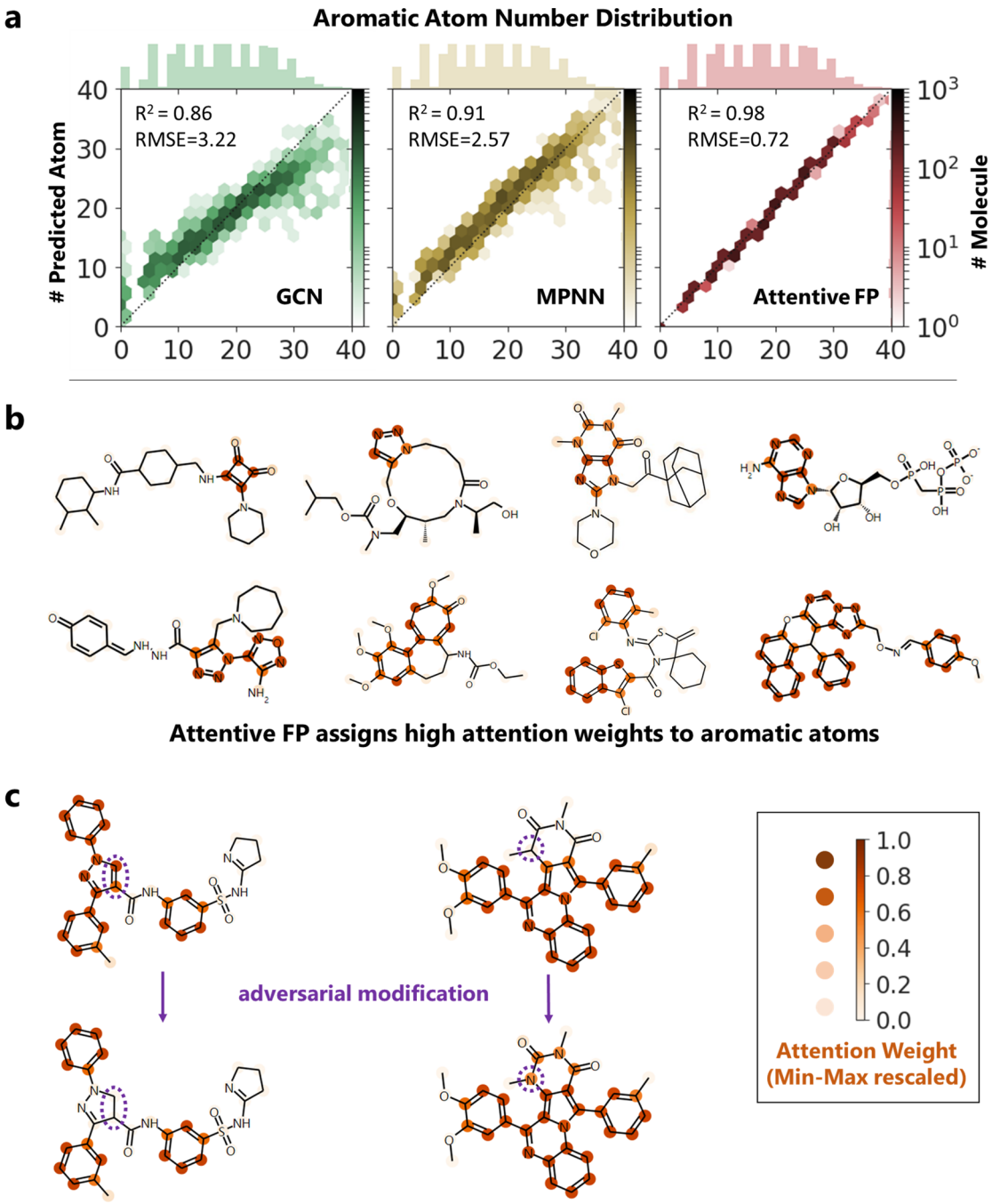


Figure 4. Learning an aromatic system. (a) The predictive performance of GCN, MPNN, and Attentive FP for predicting the numbers of aromatic atoms in molecules. The upper histogram indicates the distribution of the number of aromatic atoms. (b) The attention weights learned from the Attentive FP model are used to highlight the atoms and conform exactly to the well-defined aromaticity. (c) The molecules with adversarial modifications (a small bond or atom change that breaks or forms aromaticity) can be precisely distinguished.

Learning Water Solubility. First, we aim to explore why Attentive FPs achieve superior performances compared with those of previous models using conventional chemical descriptors as input. Therefore, we compared the automatically learned hidden features ("fingerprints") with hand-crafted chemical descriptors. Specifically, the learned feature for solubility prediction is a 200-dimensional embedding (vector), in which each dimension has its own chemical implication. Here, we calculated the Pearson's correlation coefficient between each feature dimension and the chemical descriptors, for example, the SA (synthesis accessibility) score or drug likeness. Figure 2a shows that the correlation distribution does not change significantly before and after training, implying that these chemical descriptors have only weak relevance to water solubility. In contrast, the correlation distributions of TPSA (topological polar surface area) and LogP (lipophilicity) skew toward extreme values after training (Figure 2b), indicating the high relevance of TPSA and LogP to water solubility. We can clearly observe a growing number of learned features that show high positive correlations with TPSA or high negative correlations with LogP, which conforms to the chemical intuition that TPSA is positively correlated and LogP is

DOI: 10.1021/acs.jmedchem.9b00959
J. Med. Chem. XXXX, XXX, XXX−XXX

negatively correlated with water solubility. As annotated, the most correlated hidden features with TPSA and LogP have Pearson's r values of 0.95 and −0.911, respectively.

Learning the Hidden Environment. In addition to the molecule-level learned features, each atom has its own state vector in each hidden layer for node embedding. To investigate how the atom state vectors evolved during the learning process, we obtained the similarity coefficient between atom pairs by calculating the Pearson correlation coefficient for those state vectors. Then, we plotted heat maps of the atom similarity matrix for the compound to observe the pattern changes. Taking the molecule structure of Iprodione as an example (Figure 3), before training, the visual pattern in the heat maps of the similarity matrix shows similar levels of chaos across different layers. After training, however, the higher order layer shows a distinct pattern in a specific order. Zooming in on the heat map of layer 2, we find that the atoms in Iprodione are clearly separated into three clusters—an isopropyl group (atoms 0−2), a dioxoimidazolidine−carboxamide linkage (atoms 3−12, highlighted in pink), and a dichlorophenyl group (atoms 13−20)—which strongly agrees with our chemical intuition regarding the Iprodione structure. Moreover, this pattern clearly suggests that Attentive FP has learned a representation related to molecular solubility. For example, both the isopropyl and dichlorophenyl groups are hydrophobic and have low polarity, and the correlation between atoms of these two moieties tends to be positive. In contrast, their correlation with the dioxoimidazolidine−carboxamide group, the flexible and polar moiety, is low and negative. Another interesting finding is that atom N3 shows a higher correlation with O11 (Pearson's $r \approx 0.9$) than with its nearer neighbor N6 (Pearson's $r \approx 0.7$). This result is counterintuitive because, from a graph representation perspective, two nodes with similar node features and topologically close to each other (i.e., N3 and N6 in this case) should also share a higher similarity in the "embedded" hidden space. Revisiting the chemical structure of Iprodione, we infer that the high correlation between N3 and O11, which do not have a similar chemical environment, may represent the presence of intramolecular hydrogen bonds between these two atoms. Clearly, the formation of intramolecular hydrogen bonds has been proven to contribute directly to solubility. This observation demonstrates that Attentive FP has indeed successfully extracted relevant information by learning from a specific task, and it also highlights the advantage of the attention mechanism introduced here for capturing nonlocal effects among atoms. (For more examples, see Supplementary Figure 2)

LEARNING AROMATICITY

To further explore how Attentive FP can learn the nonlocal effects within chemical structures, we constructed a task to predict the number of aromatic atoms in a molecule, which is slightly different from the study of Matlock et al.[48] that learns aromaticity and conjugated systems under a series of supervision tasks. In our setting, the learning is more challenging from the standpoint of machine intelligence, as it is only supervised by one integer per molecule. A total of 3945 molecules with 0−40 aromatic atoms were sampled from the PubChem BioAssay data set for this analysis. All bond features and all atom aromatic features were excluded from the molecular featurization procedure to eliminate any prior knowledge of aromaticity; i.e., the featurization process generated only 38 bits for each atom and no bits for bonds. We also compared our Attentive FP with GCN and MPNN implemented in Deepchem 1.3.1[49] on the same learning task. As shown in Figure 4, Attentive FP outperforms GCN and MPNN, achieving smaller RMSE and higher R^2 values. Even more interesting, the Attentive FP model precisely assigns high attention weights to the aromatic atoms and low attention weights to the nonaromatic atoms, and this assignment is robust to adversarial modification of molecules. As shown in Figure 4c, a small bond or atom change that disrupts conjugated **pi** systems can be accurately recognized. These observations suggest that the attention weight of Attentive FP at the atom level indeed has chemical implications that can be easily interpreted as an aromaticity property in this case. For more intricate problems, attention weight may also be taken as hints for discovering new knowledge.

CONCLUSION

An ambitious goal for drug design is to read properties directly from chemical structure; however, it remains an open question as to what extent and how accurately information can be extracted. Other related tasks, such as reaction outcome and yield predictions, retrosynthesis analysis, and synthetic planning, can also gain essential benefits from better molecular representations for property prediction. Molecular representation with a deep learning approach provides a viable option, which can not only help establish a predictive model for molecular properties but also recreate knowledge from existing data and even form new theories to describe chemical systems.[50] In this direction, much more effort is still required not only to improve the predictive power of the resulting model but also to interpret the model rather than simply accepting the "black box" results.

In this work, we proposed Attentive FP, a small molecule representation framework based on a graph neural network. The adoption of graph attention mechanisms at both the atom and molecule levels allows this new representation framework to learn both local and nonlocal properties of a given chemical structure. Accordingly, it captures subtle substructure patterns such as intramolecular hydrogen bonding and aromatic systems, contributing to its excellent learning capability for a wide range of different molecular properties. Moreover, inverting the Attentive FP model by extracting the hidden layers or attention weights provides access to the model's interpretation, which will help chemists gain insights into the skyrocketing volume and complexity of drug discovery data.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jmedchem.9b00959.

More detailed model tests and validations (PDF)

AUTHOR INFORMATION

Corresponding Authors

*E-mail: hljiang@simm.ac.cn. Phone: +86-21-50806600-1303.
*E-mail: myzheng@simm.ac.cn. Phone: +86-21-50806600-1308.

ORCID

Xiaomin Luo: 0000-0003-0426-3417
Mingyue Zheng: 0000-0002-3323-3092

DOI: 10.1021/acs.jmedchem.9b00959
J. Med. Chem. XXXX, XXX, XXX−XXX

Author Contributions

Z.X., M.Z., K.C., and H.J. conceived and designed the algorithm and visualization as well as wrote the paper. D.W., X.L., F.Z., X.W., Z.L., X.L., and X.L. performed the computational experiments with benchmarked data sets. All authors read and approved the final manuscript.

Notes

The authors declare no competing financial interest.

Feature visualization, instructions, and code for Attentive FP are available at https://github.com/OpenDrugAI/AttentiveFP.

■ ACKNOWLEDGMENTS

We gratefully acknowledge financial support from the National Natural Science Foundation of China (81773634 to M.Z. and 81430084 to K.C.), National Science & Technology Major Project "Key New Drug Creation and Manufacturing Program", China (Number: 2018ZX09711002 to H.J.), and "Personalized Medicines—Molecular Signature-based Drug Discovery and Development", Strategic Priority Research Program of the Chinese Academy of Sciences (XDA12050201 to M.Z.), and the open fund of state key laboratory of Pharmaceutical Biotechnology, Nanjing University, China (KF-GN-201706 to H.J.).

■ ABBREVIATIONS USED

FP, fingerprint; GAT, graph attention network; HPC, high performance computing; GRU, gated recurrent unit; GNN, graph neural network; GCN, graph convolutional network; MPNN, message passing neural network; DTNN, deep tensor neural network; CM, coulomb matrix; SA, synthesis accessibility; TPSA, topological polar surface area; BO, Bayesian optimization; MAD, mean absolute deviation; MAE, mean absolute error; MSE, mean-squared error; RMSE, root-mean-squared error; AUC, area under curve; ROC, receiver operating characteristics curve; PRC, precise-recall curve

■ REFERENCES

(1) Schneider, G. Mind and Machine in Drug Design. *Nat. Mach. Intell.* **2019**, *1* (3), 128−130.

(2) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Davies, M. The ChEMBL Database in 2017. *Nucleic Acids Res.* **2017**, *45* (D1), D945−D954.

(3) Huang, R.; Xia, M.; Sakamuru, S.; Zhao, J.; Shahane, S. A.; Attene-Ramos, M.; Zhao, T.; Austin, C. P.; Simeonov, A. Modelling the Tox21 10 K Chemical Profiles for in Vivo Toxicity Prediction and Mechanism Characterization. *Nat. Commun.* **2016**, *7*, 10425.

(4) Kuhn, M.; Letunic, I.; Jensen, L. J.; Bork, P. The SIDER Database of Drugs and Side Effects. *Nucleic Acids Res.* **2016**, *44* (D1), D1075−D1079.

(5) Bolton, E. E.; Wang, Y.; Thiessen, P. A.; Bryant, S. H. PubChem: Integrated Platform of Small Molecules and Biological Activities. *Annu. Rep. Comput. Chem.* **2008**, *4*, 217.

(6) Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Zhou, Z.; Han, L.; Karapetyan, K.; Dracheva, S.; Shoemaker, B. A.; Bolton, E. PubChem's BioAssay Database. *Nucleic Acids Res.* **2012**, *40* (D1), D400−D412.

(7) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine Learning for Molecular and Materials Science. *Nature* **2018**, *559* (7715), 547−555.

(8) Altae-Tran, H.; Ramsundar, B.; Pappu, A. S.; Pande, V. Low Data Drug Discovery with One-Shot Learning. *ACS Cent. Sci.* **2017**, *3* (4), 283−293.

(9) Gupta, A.; Zou, J. Feedback GAN for DNA Optimizes Protein Functions. *Nat. Mach. Intell.* **2019**, *1* (2), 105−111.

(10) Liu, B.; Ramsundar, B.; Kawthekar, P.; Shi, J.; Gomes, J.; Luu Nguyen, Q.; Ho, S.; Sloane, J.; Wender, P.; Pande, V. Retrosynthetic Reaction Prediction Using Neural Sequence-to-Sequence Models. *ACS Cent. Sci.* **2017**, *3*, 1103.

(11) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. *Nature* **2018**, *555*, 604.

(12) Carrasquilla, J.; Torlai, G.; Melko, R. G.; Aolita, L. Reconstructing Quantum States with Generative Models. *Nat. Mach. Intell.* **2019**, *1* (3), 155−161.

(13) Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. Predicting Reaction Performance in C-N Cross-Coupling Using Machine Learning. *Science* **2018**, *360* (6385), 186−190.

(14) Lake, B. M.; Salakhutdinov, R.; Tenenbaum, J. B. Human-Level Concept Learning through Probabilistic Program Induction. *Science* **2015**, *350*, 1332.

(15) Neftci, E. O.; Averbeck, B. B. Reinforcement Learning in Artificial and Biological Systems. *Nat. Mach. Intell.* **2019**, *1* (3), 133−143.

(16) Dragon 7.0. https://chm.kode-solutions.net/products_dragon.php (accessed Jan 8, 2019).

(17) Gedeck, P.; Rohde, B.; Bartels, C. QSAR - How Good Is It in Practice? Comparison of Descriptor Sets on an Unbiased Cross Section of Corporate Data Sets. *J. Chem. Inf. Model.* **2006**, *46*, 1924.

(18) Miyao, T.; Kaneko, H.; Funatsu, K. Inverse QSPR/QSAR Analysis for Chemical Structure Generation (from y to x). *J. Chem. Inf. Model.* **2016**, *56*, 286.

(19) Braga, R. C.; Alves, V. M.; Silva, M. F. B.; Muratov, E.; Fourches, D.; Tropsha, A.; Andrade, C. H. Tuning HERG Out: Antitarget QSAR Models for Drug Development. *Curr. Top. Med. Chem.* **2014**, *14* (11), 1399−1415.

(20) Karelson, M.; Lobanov, V. S.; Katritzky, A. R. Quantum-Chemical Descriptors in QSAR/QSPR Studies. *Chem. Rev.* **1996**, *96* (3), 1027−1044.

(21) Gavaghan, C. L.; Arnby, C. H.; Blomberg, N.; Strandlund, G.; Boyer, S. Development, Interpretation and Temporal Evaluation of a Global QSAR of HERG Electrophysiology Screening Data. *J. Comput.-Aided Mol. Des.* **2007**, *21* (4), 189−206.

(22) Schütt, K. T.; Arbabzadah, F.; Chmiela, S.; Müller, K. R.; Tkatchenko, A. Quantum-Chemical Insights from Deep Tensor Neural Networks. *Nat. Commun.* **2017**, *8*, 13890.

(23) Huo, H.; Rupp, M. Unified Representation of Molecules and Crystals for Machine Learning. 2017, arXiv:1704.06439. arXiv.org e-Print archive. https://arxiv.org/abs/1704.06439.

(24) Lubbers, N.; Smith, J. S.; Barros, K. Hierarchical Modeling of Molecular Energies Using a Deep Neural Network. *J. Chem. Phys.* **2018**, *148* (24), 241715.

(25) Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. SchNet - A Deep Learning Architecture for Molecules and Materials. *J. Chem. Phys.* **2018**, *148* (24), 241722.

(26) Duvenaud, D.; Maclaurin, D.; Aguilera-Iparraguirre, J.; Gómez-Bombarelli, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints. 2015, arXiv:1509.09292. arXiv.org e-Print archive. https://arxiv.org/abs/1509.09292.

(27) Coley, C. W.; Barzilay, R.; Green, W. H.; Jaakkola, T. S.; Jensen, K. F. Convolutional Embedding of Attributed Molecular Graphs for Physical Property Prediction. *J. Chem. Inf. Model.* **2017**, *57*, 1757.

(28) Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular Graph Convolutions: Moving Beyond Fingerprints. *J. Comput.-Aided Mol. Des.* **2016**, *30*, 595.

(29) Zhou, Z.; Li, X. Graph Convolution: A High-Order and Adaptive Approach. 2017, arXiv:1706.09916. arXiv.org e-Print archive. https://arxiv.org/abs/1706.09916.

(30) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: A Benchmark for Molecular Machine Learning. *Chem. Sci.* **2018**, *9* (2), 513−530.

DOI: 10.1021/acs.jmedchem.9b00959
J. Med. Chem. XXXX, XXX, XXX−XXX

(31) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. 2017, arXiv:1706.03762. arXiv.org e-Print archive. https://arxiv.org/abs/1706.03762.
(32) Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; Bengio, Y. Graph Attention Networks. 2017, arXiv:1710.10903. arXiv.org e-Print archive. https://arxiv.org/abs/1710.10903.
(33) Xu, B.; Wang, N.; Chen, T.; Li, M. Empirical Evaluation of Rectified Activations in Convolutional Network. 2015, arXiv:1505.00853. arXiv.org e-Print archive. https://arxiv.org/abs/1505.00853.
(34) Hamilton, W. L.; Ying, R.; Leskovec, J. Representation Learning on Graphs: Methods and Applications. *IEEE Data Eng. Bull.* **2017**, *40*, 52−74.
(35) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. 2017, arXiv:1704.01212. arXiv.org e-Print archive. https://arxiv.org/abs/1704.01212.
(36) Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. 2014, arXiv:1412.3555. arXiv.org e-Print archive. https://arxiv.org/abs/1412.3555.
(37) Ravanelli, M.; Brakel, P.; Omologo, M.; Bengio, Y. Light Gated Recurrent Units for Speech Recognition. 2018, arXiv:1803.10225. arXiv.org e-Print archive. https://arxiv.org/abs/1803.10225.
(38) Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864.
(39) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum Chemistry Structures and Properties of 134 Kilo Molecules. *Sci. Data* **2014**, *1*, 140022.
(40) Davis, J.; Goadrich, M. The Relationship Between Precision-Recall and ROC Curves. In *Proceedings of the 23rd International Conference on Machine Learning*; New York, 2006; pp 233−240.
(41) Snoek, J.; Larochelle, H.; Adams, R. P. Practical Bayesian Optimization of Machine Learning Algorithms. In *Advances in neural information processing systems*; Lake Tahoe, NV, 2012; pp 2951−2959.
(42) Murugan, P. Hyperparameters Optimization in Deep Convolutional Neural Network/Bayesian Approach with Gaussian Process Prior. 2017, arXiv:1712.07233. arXiv.org e-Print archive. https://arxiv.org/abs/1712.07233.
(43) pyGPGO. https://github.com/hawk31/pyGPGO (accessed May 18, 2018).
(44) Contal, E.; Buffoni, D.; Robicquet, A.; Vayatis, N. Parallel Gaussian Process Optimization with Upper Confidence Bound and Pure Exploration. In *Machine Learning and Knowledge Discovery in Databases*; Blockeel, H., Kersting, K., Nijssen, S., Železný, F., Eds.; Springer: Berlin, Heidelberg, 2013; pp 225−240.
(45) Yao, Y.; Rosasco, L.; Caponnetto, A. On Early Stopping in Gradient Descent Learning. *Constr. Approx.* **2007**, *26* (2), 289−315.
(46) Pytorch. https://github.com/pytorch/pytorch (accessed Oct 8, 2017).
(47) Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. 2014, arXiv:1412.6980. arXiv.org e-Print archive. https://arxiv.org/abs/1412.6980.
(48) Matlock, M. K.; Dang, N. L.; Swamidass, S. J. Learning a Local-Variable Model of Aromatic and Conjugated Systems. *ACS Cent. Sci.* **2018**, *4* (1), 52−62.
(49) Ramsundar, B.; Eastman, P.; Leswing, K.; Walters, P.; Pande, V. *Deep Learning for the Life Sciences*; O'Reilly Media, Shroff Publishers & Distributors Pvt. Ltd: Mumbai, India, 2019.
(50) Zhou, Q.; Tang, P.; Liu, S.; Pan, J.; Yan, Q.; Zhang, S.-C. Learning Atoms for Materials Discovery. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115* (28), E6411−E6417.

DOI: 10.1021/acs.jmedchem.9b00959
J. Med. Chem. XXXX, XXX, XXX−XXX