用于药物发现定量结构活性关系建模的可解释手性感知图神经网络

Interpretable Chirality-Aware Graph Neural Network for Quantitative Structure Activity Relationship Modeling in Drug Discovery

Yunchao (Lance) Liu[1], Yu Wang[1], Oanh Vu[1], Rocco Moretti[1], Bobby Bodenheimer[1], Jens Meiler[1, 2], Tyler Derr[1]

[1]Vanderbilt University
[2]Leipzig University
{yunchao.liu, yu.wang.1, oanh.t.vu.2, rocco.moretti, bobby.bodenheimer, jens.meiler, tyler.derr}@vanderbilt.edu

图神经网络从化学结构预测生物活性方面取得了成功，但忽略了分子手性等重要化学信息，为了填补这一关键空白，我们提出了用于分子表示学习的分子核图神经网络(MolKGNN)，它具有分子-/构象不变性、手性感知和可解释性。

Abstract

In computer-aided drug discovery, quantitative structure activity relation models are trained to predict biological activity from chemical structure. Despite the recent success of applying graph neural network to this task, important chemical information such as molecular chirality is ignored. To fill this crucial gap, we propose Molecular-Kernel Graph Neural Network (MolKGNN) for molecular representation learning, which features SE(3)-/conformation invariance, chirality-awareness, and interpretability. For our MolKGNN, we first design a molecular graph convolution to capture the chemical pattern by comparing the atom's similarity with the learnable molecular kernels. Furthermore, we propagate the similarity score to capture the higher-order chemical pattern. To assess the method, we conduct a comprehensive evaluation with nine well-curated datasets spanning numerous important drug targets that feature realistic high class imbalance and it demonstrates the superiority of MolKGNN over other graph neural networks in computer-aided drug discovery. Meanwhile, the learned kernels identify patterns that agree with domain knowledge, confirming the pragmatic interpretability of this approach. Our code and supplementary material are publicly available at https://github.com/meilerlab/MolKGNN.

1 Introduction

Developing new drugs is time-consuming and expensive, e.g., it took cabozantinib, an oncologic drug, 8.8 years and $1.9 billion to get on the market (Prasad and Mailankody 2017). To assist this process, computer-aided drug discovery (CADD) has been widely used. In CADD, several mathematical and machine learning methods have been developed to model the Quantitative Structure Activity Relationship (QSAR) to predict the biological activity of molecules based on their geometric structures (Sliwoski et al. 2014).

Recently, Graph Neural Networks (GNNs) have successfully been applied in many fields, e.g., social networks and recommender systems (Zhou et al. 2020). As molecules can be essentially viewed as graphs with atoms as nodes and chemical bonds as edges, GNNs are naturally adopted to perform such graph classification, i.e., predicting the biological activity of molecules based on their geometric structures (Atz, Grisoni, and Schneider 2021). A typical GNN architecture for graph classification begins with an encoder extracting node representations by passing neighborhood information followed by pooling operations that integrate node representations into graph representations, which are fed into a classifier to predict graph classes (Zhou et al. 2020).

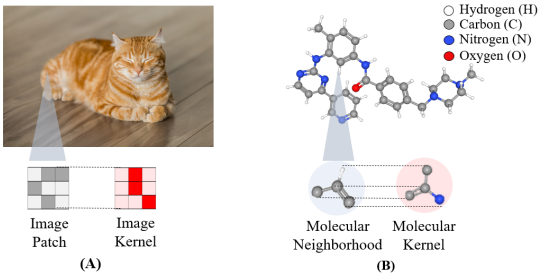

Figure 1: Analogy between (A) 2D image convolution and (B) 3D molecular convolution. In (A), the more similar the image patch is to the image kernel, the higher the output value. We design molecular convolution (B) to output a higher value if the molecular neighborhood is similar to the molecular kernel. The kernel provides the basis for our chirality calculation and has added benefit of interpretability.

如果分子邻域与分子核相似，则输出更高的值。

Despite the promise of GNN models applied to molecular representation learning, existing GNN models either blindly follow the message passing framework without considering molecular constraints on graphs (Feng et al. 2022), fail to integrate chirality (Schütt et al. 2017), or lack interpretability (Adams, Pattanaik, and Coley 2021). To fill this crucial gap, we develop a new GNN model named MolKGNN that features SE(3)/conformation invariance, chirality-awareness and provides a form of interpretability. Our main contributions can be summarized as follows:

- **Novel Interpretable Molecular Convolution:** We design a new convolution operation to capture chemical pattern of each atom by quantifying the similarity between the atom's neighboring subgraph and the learnable molecular kernel, which is inherently interpretable.
- **Better Chirality Characterization:** Rather than listing all permutations of neighbors for a chiral center, or using dihedral angles, the chirality calculation module in our design only needs a lightweight linear algebra calculation.

Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

根据分子的几何结构预测分子的生物活性

- **Comprehensive Evaluation in CADD:** We perform a comprehensive evaluation using well-curated datasets spanning numerous important drug targets (that feature realistic high class imbalance) and metrics that bias predicted active molecules for actual experimental validation. Ultimately, we demonstrate the superiority of MolKGNN over other GNNs in CADD.

2 Related Work

2.1 Extending Convolutions to the Graph Domain

Convolutional Neural Networks (CNN) have enjoyed much success on images. However, convolution fails to readily extend to graphs due to their irregular structures. Early efforts on GNNs focused on spectral convolution (Defferrard, Bresson, and Vandergheynst 2016; Kipf and Welling 2016). Later, spatial-based methods define graph convolution based on nodes' spatial relationship (Gilmer et al. 2017)

Vanilla GNN is known to have limited expressive power bounded by the Weisfeiler-Lehman (WL) graph isomorphism test (Xu et al. 2018) and hence have difficulty in finding substructures. On the other hand, graph kernels can take substructures into consideration by computing a similarity score among graph substructures. Recently, a strand of work extend GNNs by combining them with graph kernels to distinguish substructures (Cosmo et al. 2021; Feng et al. 2022). However, we argue that extending the expressive power to distinguish more substructures is not necessarily helpful with molecular representation learning. For example, (Cosmo et al. 2021) explicitly states their model could distinguish triangles. Nevertheless, although present in some drug molecules (Talele 2016), triangles are rare due to chemical instability. This can be verified by an empirical observation in the annual best-selling small molecule drugs posters[1] (McGrath, Brichacek, and Njardarson 2010). Moreover, learning a useful discrete structure in a differentiable way is challenging and hence the structure learning process in (Cosmo et al. 2021) uses random modification. This raises the question of whether it is worth the extra computational time associated with finding a structural similarity. Instead, our method identifies semantic similarity between a 1-hop molecular neighborhood and a molecule kernel (Figure. 3).

2.2 Molecular Representation Learning

It is not surprising to see the application of GNN to molecules due to the ready interpretation of atoms as nodes and bonds as edges. Even the term *graph* (in the sense used in graph theory) was used for the first time to draw a relationship between mathematics and chemistry (Sylvester 1878). Several attempts have been made to leverage GNNs for molecular representation learning. In this paper, we classify them into four categories.

Models in the first category capture the 2D connectivity (i.e., molecular constitution). Examples include (Yang et al. 2019; Coley et al. 2017). Some molecular properties, especially pharmacological activities, are dependent on the chirality of molecules (H Brooks, C Guida, and G Daniel 2011). A chiral molecule cannot be superimposed on its own mirror image. For such tasks, molecules should be treated as non-invariant to reflection. Models in this second category are reflection-sensitive, or chirality-aware and sometimes called 2.5D QSAR models. Examples include (Liu et al. 2021; Pattanaik et al. 2020). However, molecules are not planar graphs but are 3D entities. Due to rotations around single bonds, molecules can display different conformations. Models in this third category, e.g., (Flam-Shepherd et al. 2021), take the dihedral angles of rotatable bonds into consideration to distinguish different conformations. As molecules exist as conformational ensembles, a fourth category (4D) encodes ensembles instead of individual conformations (Adams, Pattanaik, and Coley 2021).

[1] https://njardarson.lab.arizona.edu/content/top-pharmaceuticals-poster

3 Preliminaries and Problem Definition

In this section, we introduce all notations used throughout this paper and define the problem of QSAR modeling.

Notations We represent a molecule as an attributed and undirected graph $G = (\mathcal{V}^G, \mathcal{E}^G)$ where $\mathcal{V}^G, \mathcal{E}^G$ are the set of nodes (atoms) and edges (chemical bonds). Let $v \in \mathcal{V}^G$ denote the node v and $e_{vu} \in \mathcal{E}^G$ denote an edge between v and u. Moreover, we represent the node attribute matrix as $\mathbf{X}^G \in \mathbb{R}^{|\mathcal{V}^G| \times d_v}$ and edge attribute matrix as $\mathbf{E}^G \in \mathbb{R}^{|\mathcal{V}^G| \times |\mathcal{V}^G| \times d_e}$ where d_v, d_e are the dimension of node and edge features. Specifically, we let $\mathbf{X}^G_v$ be the attribute of node v and $\mathbf{E}^G_{vu}$ be the attribute of edge e_{vu}. The node coordinate matrix is represented as $\mathbf{P}^G \in \mathbb{R}^{|\mathcal{V}^G| \times 3}$ and $\mathbf{P}^G_v$ denotes the 3D coordinates of v. The graph topology is described by its adjacency matrix $\mathbf{A}^G \in \{1, 0\}^{|\mathcal{V}^G| \times |\mathcal{V}^G|}$ where $\mathbf{A}^G_{vu} = 1$ if $e_{vu} \in \mathcal{E}^G$, and $\mathbf{A}^G_{vu} = 0$ otherwise. Note that bond types are encoded as edge features. Furthermore, we denote the 1-hop neighborhood of v in G as $\mathcal{N}^G_v = \{u \in \mathcal{V}^G | (v, u) \in \mathcal{E}^G\}$.

Problem Definition Based on the above notations, we formulate QSAR modeling as a graph classification problem, which can be mathematically defined as: *Given a set of attributed molecule graphs* $\mathcal{G} = \{G_1, G_2, ..., G_N\}$ *with each molecule* $G_i = (\mathcal{V}^{G_i}, \mathcal{E}^{G_i}, \mathbf{X}^{G_i}, \mathbf{E}^{G_i}, \mathbf{P}^{G_i})$ *as defined above and its corresponding one-hot encoded label* $\mathbf{Y}_i$*, we aim to learn a graph encoder* $\mathcal{F}$ *and a classifier* $\mathcal{C} : \mathcal{C}(\mathcal{F}(\mathbf{X}^{G_i}, \mathbf{E}^{G_i}, \mathbf{A}^{G_i}, \mathbf{P}^{G_i})) \rightarrow \mathbf{Y}_i$ *that is well-predictive of the ground truth label* $\mathbf{Y}_i$ *of molecule* G_i*.*

4 Molecular-Kernel Graph Neural Network

In this section, we introduce the framework of our proposed MolKGNN. As shown in Figure 2, MolKGNN recursively performs molecular convolution and message aggregation to learn representations of each molecule. In molecular convolution, we design learnable molecular kernels to capture chemically-meaningful subgraph pattern of each node/atom. Specifically, we calculate the similarity scores of each atom with its neighborhood to the molecular kernels and treat the obtained score as new atom features, which essentially describes the distance of the atom's chemical properties to the patterns encoded in the kernels. Then in message aggregation, we leverage feature propagation to aggregate similarity

在二维图像中，卷积运算可以看作是计算图像patch与图像核之间的相似度。

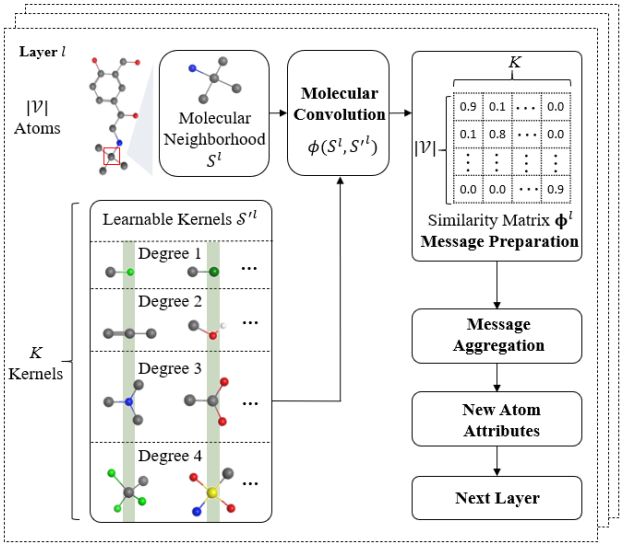

Figure 2: An overview of the proposed MolKGNN. For each atom $v \in \mathcal{V}$ of the molecule G, its 1-hop star-like molecular neighborhood subgraph S^l is convoluted with a set of K learnable kernels $\mathcal{S}'^l = \{S'^l_k\}_{k=1}^K$ to get its similarity vector and collectively for all nodes, after applying the same convolution as above, we end up with the similarity matrix $\mathbf{\Phi}^l \in \mathbb{R}^{|\mathcal{V}| \times K}$ in layer l. Specifically, $\mathbf{\Phi}_{ik} = \phi(S_{v_i}, S'_k)$ quantifies the similarity between the neighborhood subgraph around atom v_i and the k^{th} kernel (See Fig. 3 for more details of calculating $\phi(S_{v_i}, S'_k)$). The $\mathbf{\Phi}^l$ serves as the new atom attributes for the computation in the next layer $l+1$.

scores of neighborhoods to further capture chemical context of each atom. These two modules proceed alternatively to gradually enlarge the receptive field so that we can capture higher-order chemical patterns. Next, we introduce details of molecular convolution and message aggregation.

4.1 Molecular Convolution

In 2D images, convolution operation can be regarded as calculating the similarity between the image patch and the image kernel. Larger output values indicate higher visual similarity patterns such as edges, strips, curves (Lin, Huang, and Wang 2021). Inspired by that, we design a molecular convolution that outputs higher values when a molecular neighborhood and kernels are more chemically similar (Figure 1). However, performing convolution on the irregular neighborhood subgraphs requires the learnable molecular kernels to have correspondingly different geometrical structures, which is computationally prohibitive. To handle this challenge, for each atom v of degree d in G, we only consider its 1-hop star-like neighborhood subgraph $S = (\mathcal{V}^S, \mathcal{E}^S)$ where $\mathcal{V}^S = \{v\} \cup \mathcal{N}_v^G$ and $\mathcal{E}^S = \{e_{vu} | u \in \mathcal{N}_v^G\}$. To make the molecular convolution feasible, we initialize the molecular kernel to also follow star-structure and denote it as $S' = (\mathcal{V}^{S'}, \mathcal{E}^{S'})$ where $\mathcal{V}^{S'} = \{v'\} \cup \mathcal{N}_{v'}^{S'}$ with v' being the central node without loss of generality and $\mathcal{E}^{S'} = \{e_{v'u'} | u' \in \mathcal{N}_{v'}^{S'}\}$. Let the learnable feature matrix

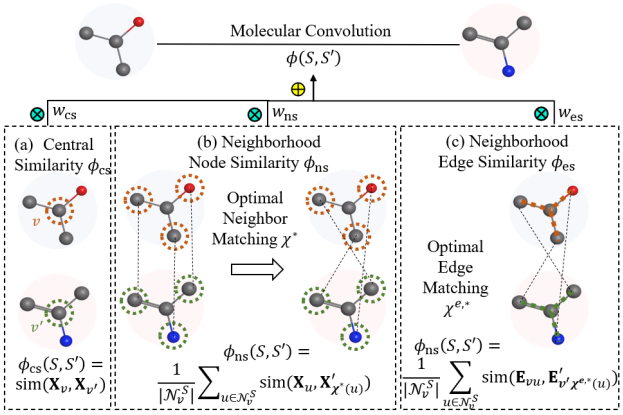

Figure 3: Illustration of the molecular convolution. The similarity between a neighborhood subgraph S and a kernel S' is quantified by $\phi(S, S')$. This similarity score is calculated as the combination of $\phi_{\text{cs}}, \phi_{\text{ns}}, \phi_{\text{es}}$, which quantify the similarity of center node, neighboring nodes, edges, respectively.

and edge feature matrix of S' be $\mathbf{X}^{S'} \in \mathbb{R}^{(d+1)\times d_n}$ and $\mathbf{E}^{S'} \in \mathbb{R}^{d\times d_e}$, respectively. Then we define the operation of molecular convolution between the atom v and the molecular kernel S' as quantifying the similarity ϕ between v's neighborhood subgraph S and the kernel S':

$$\phi(S, S') = w_{\text{cs}}\phi_{\text{cs}}(S, S') + w_{\text{ns}}\phi_{\text{ns}}(S, S') + w_{\text{es}}\phi_{\text{es}}(S, S'), \quad (1)$$

where $\phi_{\text{cs}}, \phi_{\text{ns}}, \phi_{\text{es}}$ quantify the similarity from three different aspects: the central similarity, neighborhood similarity, and edge similarity. We combine them together with learnable weights $w_{\text{cs}}, w_{\text{ns}}, w_{\text{es}} \in [0, 1]$ after softmax-normalization.

Central Similarity We first capture the chemical property of the atom v itself in S by computing its similarity to the central node v' in the kernel S':

$$\phi_{\text{cs}}(S, S') = \text{sim}(\mathbf{X}_v^S, \mathbf{X}_{v'}^{S'}), \quad (2)$$

where $\mathbf{X}_v^S, \mathbf{X}_{v'}^{S'}$ are attributes of the central atom v in the subgraph S and the central node v' in the kernel S'. The $\text{sim}(\cdot, \cdot)$ is the function measuring vector similarity and we use cosine similarity throughout this work.

Neighboring Node and Edge Similarity Besides the central node, the chemical property of an atom is also impacted by its neighborhood context, which motivates us to further quantify the similarity between 1) the neighboring nodes $\mathcal{N}_v^S$ in S and $\mathcal{N}_{v'}^{S'}$ in S', and 2) the edges $\mathcal{E}^S$ and $\mathcal{E}^{S'}$.

Before calculating ϕ_{ns}, ϕ_{es} between S and S', we face a matching problem. For example, in Figure 3, the node u_1 in S has more than one matching candidates, i.e., $\{u'_1, u'_2, u'_3\}$ in S'. Here we seek a bijective matching $\chi^* : \mathcal{N}_v^S \to \mathcal{N}_{v'}^{S'}$ such that the average attribute similarity between $u \in \mathcal{N}_v^S$ and $\chi^*(u) \in \mathcal{N}_{v'}^{S'}$ over all neighbors can be maximized:

$$\chi^* = \arg\max_{\chi} \frac{1}{|\mathcal{N}_v^S|} \sum_{u\in\mathcal{N}_v^S} \text{sim}(\mathbf{X}_u^S, \mathbf{X}_{\chi(u)}^{S'}). \quad (3)$$

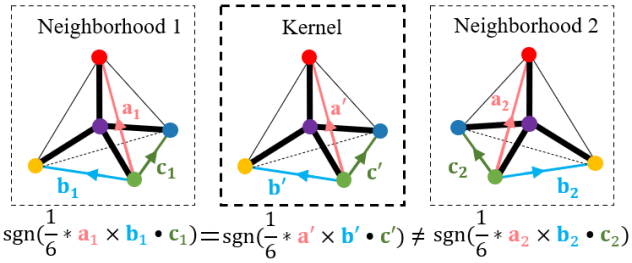

Figure 4: Illustration of chirality calculation. Corresponding nodes given by the optimal matching χ^* are of the same colors. $\text{sgn}(\cdot)$ is the sign function. We can distinguish mirror-imaged neighborhoods of two atoms by comparing the orientation of their corresponding tetrahedrons (i.e., the sign of the tetrahedral volume).

Given that exhausting all $|\mathcal{N}_v^S|!$ possible matchings to find the optimal one is computationally infeasible, we significantly simplify this computation by constraining the searching space according to the inherent structure of molecules, which are: 1) node degrees in drug-like molecule graphs are usually less than 5, with most atoms having a degree of 1 and few nodes having a degree of 4 (Patrick 2013); 2) for nodes of degree 4, only 12 among the total 24 possible matchings are valid after considering chirality (Pattanaik et al. 2020) (See Section 4.4 for more details).

After we obtain the optimal node matching, the bijective edge matching can be trivially defined as: $\chi^{e,*} : \mathcal{E}^S \to \mathcal{E}^{S'}$ such that the edge $e_{vu} \in \mathcal{E}^S$ if and only if $e_{v'\chi^*(u)} \in \mathcal{E}^{S'}$. Then, with the defined node and edge matching $\chi^*, \chi^{e,*}$ as above, we compute ϕ_{ns} and ϕ_{es} as:

$$\phi_{\text{ns}} = \frac{1}{|\mathcal{N}_v^S|} \sum_{u \in \mathcal{N}_v^S} \text{sim}(\mathbf{X}_u^S, \mathbf{X}_{\chi^*(u)}^{S'}). \tag{4}$$

$$\phi_{\text{es}} = \frac{1}{|\mathcal{N}_v^S|} \sum_{u \in \mathcal{N}_v^S} \text{sim}(\mathbf{E}_{vu}^S, \mathbf{E}_{v'\chi^{e,*}(u)}^{S'}). \tag{5}$$

Chirality Characterization After we capture the chemical-informative pattern of the neighborhood subgraph around each atom by quantifying the above three different aspects of similarity, our model is still chirality-insensitive, i.e., it still cannot distinguish enantiomers (pairs of mirror-imaged molecules that are non-superimposable, like our left and right hands (McNaught, Wilkinson et al. 1997)) However, chirality is a key determinant of a molecule's biological activity (Sliwoski et al. 2012), which motivates us to characterize the chirality of the molecule in the next.

According to the rule of basic chemistry, chirality can only exist when the central atom has four unique neighboring substructures. Therefore, we only characterize the chirality for atom v where $|\mathcal{N}_v^S| = 4$ and its four neighboring substructures are different from each other. More specifically, given the neighborhood subgraph of an atom S forming the tetrahedron shown in Figure 4 where the four unique neighboring atoms are $\mathcal{N}_v^S = \{u_1, u_2, u_3, u_4\}$, we select u_1 without loss of generality as the anchor neighbor to define the three concurrent sides of the tetrahedron $\mathbf{a}^S = \mathbf{P}_{u_2}^S - \mathbf{P}_{u_1}^S, \mathbf{b}^S = \mathbf{P}_{u_3}^S - \mathbf{P}_{u_1}^S, \mathbf{c}^S = \mathbf{P}_{u_4}^S - \mathbf{P}_{u_1}^S$ and further calculate the tetrahedral volume of S as:

$$\xi^S = \frac{1}{6} * \mathbf{a}^S \times \mathbf{b}^S \cdot \mathbf{c}^S. \tag{6}$$

Similarly, we calculate $\xi^{S'}$ for the kernel S'. Since the sign of the tetrahedron volume of the molecule ξ^S defines its orientation (Sliwoski et al. 2012), if $\text{sgn}(\xi^S) = -\text{sgn}(\xi^{S'})$, with $\text{sgn}(\cdot)$ being the sign function, the four neighboring subgraph S, S' would be of opposite direction. Therefore, by treating the kernel as the anchor and comparing its direction with the ones of two neighborhood subgraphs as:

$$\phi(S, S') = \begin{cases} \phi(S, S'), & \text{if } \text{sgn}(\chi^S) = \text{sgn}(\chi^{S'}) \\ -\phi(S, S'), & \text{if } \text{sgn}(\chi^S) \neq \text{sgn}(\chi^{S'}), \end{cases} \tag{7}$$

we can distinguish two enantiomers and make the model chirality-sensitive.

After we encode the chirality into the similarity computation, our proposed molecular convolution could fully capture the chemical pattern of the atom in terms of its own property by $\phi_{\text{cs}}(S, S')$, its neighborhood property $\phi_{\text{ns}}(S, S'), \phi_{\text{es}}(S, S')$ and its chirality by the sign of $\phi(S, S')$. Since one kernel can only characterize one chemical pattern, we extend the above defined molecular convolution to the situation of multiple kernels in the next section.

4.2 Model Architecture

Suppose the set of K kernels at layer l be $\mathcal{S}'^l = \{S_k'^l\}_{k=1}^K$, we apply the proposed molecular convolution with the learnable molecular kernel $S_k'^l \in \mathcal{S}'^l$ over the node representation $\mathbf{H}^{l-1}$ at the previous layer $l-1$ to obtain the node similarity matrix at layer l as $\mathbf{\Phi}^l \in \mathbb{R}^{|\mathcal{V}| \times K}$:

$$\mathbf{\Phi}^l = \begin{bmatrix} \phi(S_{v_1}^{l-1}, S_1'^{l-1}) & \phi(S_{v_1}^{l-1}, S_2'^{l-1}) & \dots & \phi(S_{v_1}^{l-1}, S_K'^{l-1}) \\ \phi(S_{v_2}^{l-1}, S_1'^{l-1}) & \phi(S_{v_2}^{l-1}, S_2'^{l-1}) & \dots & \phi(S_{v_2}^{l-1}, S_K'^{l-1}) \\ \vdots & \vdots & \vdots & \vdots \\ \phi(S_{v_{|\mathcal{V}|}}^{l-1}, S_1'^{l-1}) & \phi(S_{v_{|\mathcal{V}|}}^{l-1}, S_2'^{l-1}) & \dots & \phi(S_{v_{|\mathcal{V}|}}^{l-1}, S_K'^{l-1}) \end{bmatrix}, \tag{8}$$

where $\phi(S_{v_i}^{l-1}, S_k'^{l-1})$ defines the similarity between the neighborhood subgraph around the atom v_i and the k^{th} kernel at layer $l-1$. We note that $\phi(S_{v_i}^{l-1}, S_k'^{l-1})$ is set to 0 if $S_{v_i}^{l-1}$ and $S_k'^{l-1}$ have different degrees so that back-propagation keeps the parameters in kernels of different degree untouched.

The above molecular kernel convolution can only capture the chemical pattern embedded in the 1-hop neighborhood around each atom. To further discover the chemical pattern embedded in the multi-hop neighborhood, we leverage the message-passing and directly aggregate the calculated neighborhood similarity $\mathbf{\Phi}^l$ as:

$$\mathbf{H}^l = \mathbf{A}\mathbf{\Phi}^l. \tag{9}$$

After recursively alternating between the molecular convolution and the message-passing L layers, the final atom representation $\mathbf{H}^L$ describes the chemical pattern up to L hops away of each atom. We further apply a readout function to

integrate node presentations into the graph representation $\mathbf{G}$ for each graph G as:

$$\mathbf{G} = \text{READOUT}(\{\mathbf{H}_i^L | v_i \in \mathcal{V}\}) \tag{10}$$

Here we employ global-sum pooling as our READOUT function, which adds all nodes' representations.

4.3 Model Optimization

From now, let the graph representation of G_i be $\mathbf{G}_i$ and so we apply the above process to get the representations for all N labeled graphs in $\mathcal{G}$. Then given the one-hot encoded label matrix $\mathbf{Y} \in \mathbb{R}^{N \times C}$, the overall objective function of MolKGNN is formally defined as:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} \mathbf{Y}_{ic} \log \hat{\mathbf{Y}}_{ic}, \tag{11}$$

where $\hat{\mathbf{Y}} = \sigma(f(\mathbf{G}))$ and $f(\cdot)$ is a classifier, e.g., Multi-Layer Perceptron followed by a softmax normalization σ.

4.4 Model Computational Complexity

It may seem to be formidable to enumerate all possible matchings described in Section 4.1. However, most nodes only have one neighbor (e.g., hydrogen, fluorine, chlorine, bromine and iodine). Take AID 1798 for example, 49.03%, 6.12%, 31.08% and 13.77% nodes are with one, two, three and four neighbors among all nodes, respectively. For nodes with four neighbors, only 12 out of 24 matchings need to be enumerated because of chirality (Pattanaik et al. 2020).

Since the adjacency matrix of molecular graphs is sparse, most GNNs incur a time complexity of $\mathcal{O}(|\mathcal{E}|)$. And as analyzed above, the permutation is bounded by up to four neighbors (12 matchings). Thus, finding the optimal matching has a time complexity of $\mathcal{O}(1)$. The calculation of molecular convolution is linear to the number of K kernels and hence has a time complexity of $\mathcal{O}(K)$. Overall, our method takes a computation time of $\mathcal{O}(|\mathcal{E}|K)$.

5 Experiments

The experiments discussed in this section answer the following research questions:

- RQ1: How does MolKGNN compare against existing methods in a realistic drug discovery benchmark?
- RQ2: Do the learned kernels of MolKGNN align with domain knowledge and provide further interpretability?
- RQ3: Does MolKGNN possess the expressiveness for distinguishing chiral molecules?

5.1 A Realistic Drug Discovery Scenario

High-throughput screening (HTS) is the use of automated equipment to rapidly screen thousands to millions of molecules for the biological activity of interest in the early drug discovery process (Bajorath 2002). However, this brute-force approach has low hit rates, typically around 0.05%-0.5%. QSAR models are trained on the results of HTS experiments in order to screen additional molecules virtually and prioritize these for acquisition (Mueller et al. 2010). Dataset challenges for constructing QSAR models include imbalance (many more inactive molecules) and containing false positives/negatives (Baell and Holloway 2010). Thus, for developing QSAR methods, curated high-quality datasets are needed. Unfortunately, too often small, uncurated, or unrealistic datasets are used, e.g., the commonly used ogbg-molpcba (Hu et al. 2020), which had no preprocessing to remove potential experimental artifacts (Ramsundar et al. 2015). Another commonly-used large molecule dataset is OGB-LSC PCQM4Mv2 ($\sim$ 3M molecules) (Hu et al. 2021), but is of molecular properties instead of biological activities. Lastly, when assessing the performance of QSAR models, a metric that bias the molecules with the highest predicted activities is of interest as ultimately only these will be acquired or synthesized and tested.

Protein Target	Total # Graphs	# Active Labels	Avg. # of Nodes (Edges)
GPCR			
· 435008: Orexin1 Receptor	218,156	233	45.14 (94.37)
·1798: M1 Muscarinic Receptor Agonists	61,832	187	43.60 (91.37)
·435034: M1 Muscarinic Receptor Antagonists	61,755	362	43.61 (91.41)
Ion Channel			
·1843: Potassium Ion Channel Receptor Kir2.1	301,490	172	44.41 (92.81)
·2258: KCNQ2 Potassium Channel	302,402	213	44.44 (92.88)
·463087: Cav3 T-type Calcium Channels	100,874	703	43.75 (91.57)
Transporter			
·488997: Choline Transporter	302,303	252	44.46 (92.90)
Kinase			
·2689: Serine/Threonine Kinase 33	319,789	172	44.85 (93.70)
Enzyme			
·485290: Tyrosyl-DNA Phosphodiesterase	341,304	281	46.13 (96.50)

Table 1: Statistics of datasets used in the experiment. Datasets are identified by their PubChem Assay ID (AID).

5.2 Datasets

PubChem (Kim et al. 2021) is a database supported by National Institute of Health (NIH) that contains biological activities for millions of drug-like molecules, often from HTS experiments. However, the raw primary screening data from PubChem have a high false positive rate (Butkiewicz et al. 2017, 2013). We benchmark our model using nine high-quality HTS experiments from PubChem that cover all important protein classes for drug discovery (Butkiewicz et al. 2017, 2013), which are summarized in Table 1. The datasets feature in the large data size, highly imbalanced labels, and diverse protein targets.

PubChem AID	MolKGNN (ours)	SchNet	SphereNet	DimeNet++	ChiRo	KerGNN
435008	0.255 ± 0.014	0.187 ± 0.027	0.215 ± 0.024	0.203 ± 0.047	0.168 ± 0.019	0.147 ± 0.015
1798	0.174 ± 0.029	0.195 ± 0.025	0.196 ± 0.035	0.208 ± 0.035	0.165 ± 0.040	0.078 ± 0.042
435034	0.227 ± 0.022	0.246 ± 0.020	0.230 ± 0.034	0.235 ± 0.044	0.211 ± 0.023	0.179 ± 0.045
1843	0.362 ± 0.033	0.358 ± 0.037	0.258 ± 0.048	0.284 ± 0.034	0.326 ± 0.010	0.292 ± 0.027
2258	0.301 ± 0.028	0.240 ± 0.037	0.380 ± 0.037	0.340 ± 0.032	0.251 ± 0.010	0.195 ± 0.020
463087	0.390 ± 0.056	0.332 ± 0.022	0.399 ± 0.011	0.389 ± 0.026	0.258 ± 0.019	0.150 ± 0.011
488997	0.303 ± 0.027	0.319 ± 0.017	0.309 ± 0.029	0.315 ± 0.011	0.193 ± 0.029	0.081 ± 0.023
2689	0.415 ± 0.020	0.324 ± 0.020	0.401 ± 0.016	0.367 ± 0.049	0.351 ± 0.048	0.264 ± 0.017
485290	0.498 ± 0.015	0.333 ± 0.047	0.450 ± 0.039	0.463 ± 0.040	0.295 ± 0.068	0.223 ± 0.026
Average	0.325	0.282	0.315	0.312	0.247	0.179
Avg. Rank	**2.333**	3.222	2.556	2.556	4.556	5.778

Table 2: Comparison of $\text{logAUC}_{[0.001,0.1]}$ between models. The performance is better when the value is higher. Reported are the mean values over five runs, with standard deviation. Avg. Rank is the average model rank among all AIDs. This is our main result since it uses a domain-related metric.

PubChem AID	MolKGNN (ours)	SchNet	SphereNet	DimeNet++	ChiRo	KerGNN
435008	0.836 ± 0.012	0.820 ± 0.009	0.794 ± 0.026	0.787 ± 0.028	0.797 ± 0.015	0.806 ± 0.017
1798	0.721 ± 0.027	0.707 ± 0.007	0.655 ± 0.025	0.649 ± 0.028	0.683 ± 0.052	0.663 ± 0.041
435034	0.816 ± 0.028	0.838 ± 0.009	0.836 ± 0.014	0.834 ± 0.019	0.822 ± 0.017	0.821 ± 0.016
1843	0.879 ± 0.025	0.896 ± 0.012	0.875 ± 0.021	0.857 ± 0.011	0.881 ± 0.010	0.906 ± 0.020
2258	0.806 ± 0.019	0.792 ± 0.020	0.801 ± 0.042	0.821 ± 0.025	0.782 ± 0.018	0.766 ± 0.024
463087	0.895 ± 0.003	0.910 ± 0.005	0.904 ± 0.005	0.902 ± 0.009	0.891 ± 0.004	0.859 ± 0.009
488997	0.866 ± 0.018	0.831 ± 0.012	0.822 ± 0.017	0.839 ± 0.023	0.817 ± 0.019	0.757 ± 0.044
2689	0.906 ± 0.019	0.905 ± 0.021	0.867 ± 0.021	0.832 ± 0.016	0.919 ± 0.017	0.912 ± 0.013
485290	0.866 ± 0.012	0.893 ± 0.011	0.879 ± 0.021	0.884 ± 0.016	0.816 ± 0.015	0.853 ± 0.009
Avgerage	0.843	0.844	0.826	0.823	0.823	0.816
Avg. Rank	2.889	**2.111**	3.778	3.889	4.000	4.222

Table 3: Comparison of AUC between models. The performance is better when the value is higher. Reported are the mean values over five runs, with standard deviation. Avg. Rank is the average model rank among all AIDs. This result is measured by a general metric. Together with the result in Table 2, we show that a well-performing model measured by a general metric could potentially perform badly in the application-related metric.

5.3 Baselines

We benchmark our method, **MolKGNN**, in comparison to five other methods. **SchNet** (Schütt et al. 2017), **SphereNet** (Liu et al. 2021), **DimeNet++** (Gallicchio and Micheli 2020), **ChIRo** (Adams, Pattanaik, and Coley 2021) and **KerGNN** (Feng et al. 2022). The first four are GNNs for molecular representation learning. The last one is a GNN that is architecturally similar to ours. Another similar work (Cosmo et al. 2021) is excluded from benchmarking due to no publicly available code at the time of writing.

5.4 Evaluation Metrics

Two metrics are used to evaluate our methods specifically, $\text{logAUC}_{[\mathbf{0.001,0.1}]}$, AUC :

- Logarithmic Receiver-Operating-Characteristic Area Under the Curve with the False Positive Rate in [0.001, 0.1] (**logAUC**$_{[\mathbf{0.001,0.1}]}$): Ranged logAUC (Mysinger and Shoichet 2010) is used because only a small percentage of molecules predicted with high activity can be selected for experimental tests in consideration of cost in a real-world drug campaign (Butkiewicz et al. 2017). This high decision cutoff corresponds to the left side of the Receiver-Operating-Characteristic (ROC) curve, i.e., those False Positive Rates (FPRs) with small values. Also, because the threshold cannot be predetermined, the area under the curve is used to consolidate all possible thresholds within a certain FPR range. Finally, the logarithm is used to bias towards smaller FPRs. Following prior work (Mendenhall and Meiler 2016; Golkov et al. 2020), we choose to use $\text{logAUC}_{[0.001,0.1]}$. A perfect classifier achieves a $\text{logAUC}_{[0.001,0.1]}$ of 1, while a random classifier reaches a $\text{logAUC}_{[0.001,0.1]}$ of around 0.0215, as shown below:
$$\frac{\int_{0.001}^{0.1} x \mathrm{d}\log_{10} x}{\int_{0.001}^{0.1} 1 \mathrm{d}\log_{10} x} = \frac{\int_{-3}^{-1} 10^u \mathrm{d}u}{\int_{-3}^{-1} 1 \mathrm{d}u} \approx 0.0215$$
- Receiver-Operating-Characteristic Area Under the Curve (**AUC**): We include AUC since this has historically been used as a general evaluation metric for graph classification (Wu et al. 2018). Comparison with AUC also highlights the fact that overall performance/ranking of methods according to AUC may not align well with that of the domain specific evaluation metric, i.e., $\text{logAUC}_{[0.001,0.1]}$.

5.5 Training Details

The datasets are split into 80%/10%/10% for training, validation and testing, respectively. Due to the large size of the datasets and limited computation resources, we reduce our training sets. We use 10,000 randomly selected inactive-labeled samples while keeping all the active-labeled samples

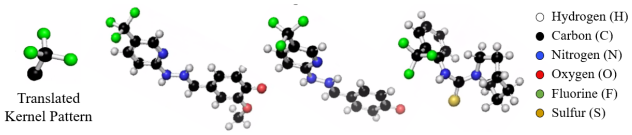

Figure 5: Visualization of a learned kernel of MolKGNN when trained on AID 2689.

for training , following (Mendenhall and Meiler 2016). The validation and testing sets remain the same. More training details can be found in the supplementary material.

5.6 Result

From Table 2, we can see MolKGNN achieves superior results in recovering the active molecules with a high decision threshold. This demonstrates the applicability of MolKGNN in a real-world scenario. KerGNN has the worst performance, which aligns with our arguments in Section 2.1 that semantic similarity is more useful than structural in drug discovery. Moreover, we find MolKGNN also performs on par with other GNN in terms of AUC (see Table 3), which demonstrates its potential applicability beyond drug discovery in a general setting. It is worth noting that different rankings of models are observed in the two tables. This demonstrates that a generally good performing model measured by AUC could potentially perform badly in a specific false positive rate region. Additionally, it highlights the ability of the proposed model to perform well in the application-related metric and indicates its practical significance. Finally the results prompt us to wonder if 3D model can indeed process more information than 2.5D model. The supplementary material contains a discussion on 2.5D vs. 3D models.

5.7 Investigation of Interpretability

We train an autoencoder-like architecture for interpreting kernels. The encoder is the same as the one used in MolKGNN to convert node features into the node embedding via batch normalization. The decoder converts the node embedding back to the corresponding atomic number. This encoder can be used to translate the node embedding in the kernels into atomic numbers. Currently we only examine the first layer and the node attributes of the kernels, but our kernels offer the potentials for retrieving more complicated pattern and we leave the investigation of that for future works.

In Figure 5, the learned pattern shows a center atom of carbon surrounded by three fluorine and another carbon. Examining the training set reveals several molecules displaying this pattern. This highlights the interpretability of our model and the finding corresponds to the domain knowledge: the pattern is known as the trifluoromethyl group in medicinal chemistry and has been used in several drugs (Yale 1958).

5.8 Ability to Distinguish Chirality

We use the CHIRAL1 dataset (Pattanaik et al. 2020) that contains 102,389 enantiomer pairs for a single 1,3-dicyclohexylpropane skeletal scaffold with one chiral center. The data is labeled as R or S stereocenter and we use

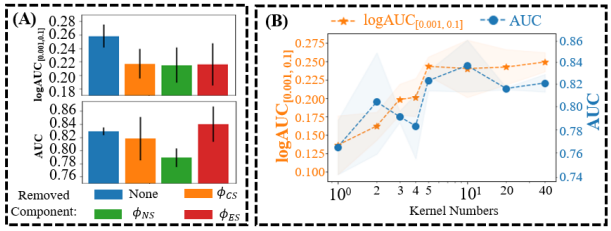

Figure 6: (A) Ablation study result for $\phi(S, S')$ components using AID 435008. Reported are average values over three runs, with standard deviation. (B) Performance for different kernel numbers using AID 435008. The number shown is applied to kernels of all degrees. Results are average values over three runs, with standard deviation.

accuracy to evaluate the performance. For comparison, we use GCN (Kipf and Welling 2016) and a modified version of our model, MolKGNN-NoChi, that removes the chirality calculation module. Our experiments observed GCN and MolKGNN-NoChi achieve 50% accuracy while MolKGNN achieves nearly 100%, which empirically demonstrates our proposed method's ability to distinguish chiral molecules.

5.9 Ablation Study

Component of $\phi(S, S')$ We conduct an ablation study on the three compnents of $\phi(S, S')$, i.e., ϕ_{cs}, ϕ_{ns}, ϕ_{es}. From Figure 6(A), we observe that the removal of any of the components has a negative impact on $\text{logAUC}_{[0.001,0.1]}$. The impact is bigger for $\text{logAUC}_{[0.001,0.1]}$ than AUC in terms of the percentage of performance change. Note that in some cases such as the removal of ϕ_{es}, there is an increase in performance according to AUC, but this would significantly hinder the $\text{logAUC}_{[0.001,0.1]}$ metric.

Kernel Number An ablation study is conducted to study the impact of kernel numbers (Figure 6(B)). When the number of kernels is too small (kernel per degree< 5), it greatly impacts the performance. However, once it is large enough to a certain point, a larger number of kernels has little impact on the performance.

6 Conclusion

In this work, we introduce a new GNN model, MolKGNN, to address the QSAR modeling problem. MolKGNN utilizes a newly-designed molecular convolution, where a molecular neighborhood is compared with a kernel and outputs a similarity score. Well-curated datasets that consist of experimental HTS data from diverse protein target classes are used for the evaluation. The highly-imbalanced datasets highlight the scarcity of positive signals in this real-world problem (Wang et al. 2022). Evaluation using domain-related ($\text{logAUC}_{[0.001,0.1]}$) metric demonstrates the practical value of MolKGNN in drug discovery while performance measured by a general metric (AUC) is also provided for comparison. Moreover, this paper provides a theoretical justification and experimental demonstration that MolKGNN is able to distinguish chiral molecules while providing interpretability for its results.

Acknowledgements

This work was supported through NIDA R01 DA046138. JM acknowledges funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through SFB1423 and SPP 2363. JM is supported by a Humboldt Professorship of the Alexander von Humboldt Foundation. BB acknowledges funding by the National Science Foundation under grant 1763966. YL acknowledges the support from the NVIDIA Academic Hardware Grant Program.

References

Adams, K.; Pattanaik, L.; and Coley, C. W. 2021. Learning 3D Representations of Molecular Chirality with Invariance to Bond Rotations. *arXiv preprint arXiv:2110.04383*.

Atz, K.; Grisoni, F.; and Schneider, G. 2021. Geometric deep learning on molecular representations. *Nature Machine Intelligence*, 3(12): 1023–1032.

Baell, J. B.; and Holloway, G. A. 2010. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *Journal of medicinal chemistry*, 53(7): 2719–2740.

Bajorath, J. 2002. Integration of virtual and high-throughput screening. *Nature Reviews Drug Discovery*, 1(11): 882–894.

Butkiewicz, M.; Lowe Jr, E. W.; Mueller, R.; Mendenhall, J. L.; Teixeira, P. L.; Weaver, C. D.; and Meiler, J. 2013. Benchmarking ligand-based virtual High-Throughput Screening with the PubChem database. *Molecules*, 18(1): 735–756.

Butkiewicz, M.; Wang, Y.; Bryant, S. H.; Lowe Jr, E. W.; Weaver, D. C.; and Meiler, J. 2017. High-throughput screening assay datasets from the pubchem database. *Chemical informatics (Wilmington, Del.)*, 3(1).

Coley, C. W.; Barzilay, R.; Green, W. H.; Jaakkola, T. S.; and Jensen, K. F. 2017. Convolutional embedding of attributed molecular graphs for physical property prediction. *Journal of chemical information and modeling*, 57(8): 1757–1772.

Cosmo, L.; Minello, G.; Bronstein, M.; Rodolà, E.; Rossi, L.; and Torsello, A. 2021. Graph kernel neural networks. *arXiv preprint arXiv:2112.07436*.

Defferrard, M.; Bresson, X.; and Vandergheynst, P. 2016. Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems*, 29.

Feng, A.; You, C.; Wang, S.; and Tassiulas, L. 2022. Kergnns: Interpretable graph neural networks with graph kernels. *ArXiv Preprint: https://arxiv. org/abs/2201.00491*.

Flam-Shepherd, D.; Wu, T. C.; Friederich, P.; and Aspuru-Guzik, A. 2021. Neural message passing on high order paths. *Machine Learning: Science and Technology*, 2(4): 045009.

Gallicchio, C.; and Micheli, A. 2020. Fast and deep graph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 3898–3905.

Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; and Dahl, G. E. 2017. Neural message passing for quantum chemistry. In *International conference on machine learning*, 1263–1272. PMLR.

Golkov, V.; Becker, A.; Plop, D. T.; Čuturilo, D.; Davoudi, N.; Mendenhall, J.; Moretti, R.; Meiler, J.; and Cremers, D. 2020. Deep Learning for Virtual Screening: Five Reasons to Use ROC Cost Functions. *arXiv preprint arXiv:2007.07029*.

H Brooks, W.; C Guida, W.; and G Daniel, K. 2011. The significance of chirality in drug design and development. *Current topics in medicinal chemistry*, 11(7): 760–770.

Hu, W.; Fey, M.; Ren, H.; Nakata, M.; Dong, Y.; and Leskovec, J. 2021. Ogb-lsc: A large-scale challenge for machine learning on graphs. *arXiv preprint arXiv:2103.09430*.

Hu, W.; Fey, M.; Zitnik, M.; Dong, Y.; Ren, H.; Liu, B.; Catasta, M.; and Leskovec, J. 2020. Open graph benchmark: Datasets for machine learning on graphs. *Advances in neural information processing systems*, 33: 22118–22133.

Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; et al. 2021. PubChem in 2021: new data content and improved web interfaces. *Nucleic acids research*, 49(D1): D1388–D1395.

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

Lin, Z.-H.; Huang, S. Y.; and Wang, Y.-C. F. 2021. Learning of 3d graph convolution networks for point cloud analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Liu, Y.; Wang, L.; Liu, M.; Lin, Y.; Zhang, X.; Oztekin, B.; and Ji, S. 2021. Spherical message passing for 3d molecular graphs. In *International Conference on Learning Representations*.

McGrath, N. A.; Brichacek, M.; and Njardarson, J. T. 2010. A graphical journey of innovative organic architectures that have improved our lives. *Journal of chemical education*, 87(12): 1348–1349.

McNaught, A. D.; Wilkinson, A.; et al. 1997. *Compendium of chemical terminology*, volume 1669. Blackwell Science Oxford.

Mendenhall, J.; and Meiler, J. 2016. Improving quantitative structure–activity relationship models using Artificial Neural Networks trained with dropout. *Journal of computer-aided molecular design*, 30(2): 177–189.

Mueller, R.; Rodriguez, A. L.; Dawson, E. S.; Butkiewicz, M.; Nguyen, T. T.; Oleszkiewicz, S.; Bleckmann, A.; Weaver, C. D.; Lindsley, C. W.; Conn, P. J.; et al. 2010. Identification of metabotropic glutamate receptor subtype 5 potentiators using virtual high-throughput screening. *ACS chemical neuroscience*, 1(4): 288–305.

Mysinger, M. M.; and Shoichet, B. K. 2010. Rapid context-dependent ligand desolvation in molecular docking. *Journal of chemical information and modeling*, 50(9): 1561–1573.

Patrick, G. L. 2013. *An introduction to medicinal chemistry*. Oxford university press.

Pattanaik, L.; Ganea, O.-E.; Coley, I.; Jensen, K. F.; Green, W. H.; and Coley, C. W. 2020. Message passing networks for molecules with tetrahedral chirality. *arXiv preprint arXiv:2012.00094*.

Prasad, V.; and Mailankody, S. 2017. Research and development spending to bring a single cancer drug to market and revenues after approval. *JAMA internal medicine*, 177(11): 1569–1575.

Ramsundar, B.; Kearnes, S.; Riley, P.; Webster, D.; Konerding, D.; and Pande, V. 2015. Massively multitask networks for drug discovery. *arXiv preprint arXiv:1502.02072*.

Schütt, K.; Kindermans, P.-J.; Sauceda Felix, H. E.; Chmiela, S.; Tkatchenko, A.; and Müller, K.-R. 2017. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in neural information processing systems*, 30.

Sliwoski, G.; Kothiwale, S.; Meiler, J.; and Lowe, E. W. 2014. Computational methods in drug discovery. *Pharmacological reviews*, 66(1): 334–395.

Sliwoski, G.; Lowe Jr, E. W.; Butkiewicz, M.; and Meiler, J. 2012. BCL:: EMAS—enantioselective molecular asymmetry descriptor for 3D-QSAR. *Molecules*, 17(8): 9971–9989.

Sylvester, J. J. 1878. Chemistry and algebra. *Nature*, 17(432): 284.

Talele, T. T. 2016. The "cyclopropyl fragment" is a versatile player that frequently appears in preclinical/clinical drug molecules. *Journal of medicinal chemistry*, 59(19): 8712–8756.

Wang, Y.; Zhao, Y.; Shah, N.; and Derr, T. 2022. Imbalanced graph classification via graph-of-graph neural networks. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2067–2076.

Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; and Pande, V. 2018. MoleculeNet: a benchmark for molecular machine learning. *Chemical science*, 9(2): 513–530.

Xu, K.; Hu, W.; Leskovec, J.; and Jegelka, S. 2018. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*.

Yale, H. L. 1958. The trifluoromethyl group in medical chemistry. *Journal of Medicinal Chemistry*, 1(2): 121–133.

Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; et al. 2019. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, 59(8): 3370–3388.

Zhou, J.; Cui, G.; Hu, S.; Zhang, Z.; Yang, C.; Liu, Z.; Wang, L.; Li, C.; and Sun, M. 2020. Graph neural networks: A review of methods and applications. *AI Open*, 1: 57–81.