Can Molecular Evolution Mechanism Enhance Molecular Representation?

Kun Li[1], Longtao Hu[1], Jiameng Chen[1], Hongzhi Zhang[1], Yida Xiong[1], Xiantao Cai[1*], Wenbin Hu[2, 1*], Jia Wu[3]

[1]School of Computer Science, Wuhan University, Wuhan, China
[2]Shenzhen Research Institute, Wuhan University, Shenzhen, China
[3]Department of Computing, Macquarie University, Sydney, Australia
{likun98, hlt_2003, jiameng.chen, zhanghongzhi, yidaxiong, caixiantao, hwb}@whu.edu.cn, jia.wu@mq.edu.au

Abstract

Molecular evolution is the process of simulating the natural evolution of molecules in chemical space to explore potential molecular structures and properties. The relationships between similar molecules are often described through transformations such as adding, deleting, and modifying atoms and chemical bonds, reflecting specific evolutionary paths. Existing molecular representation methods mainly focus on mining data, such as atomic-level structures and chemical bonds directly from the molecules, often overlooking their evolutionary history. Consequently, we aim to explore the possibility of enhancing molecular representations by simulating the evolutionary process. We extract and analyze the changes in the evolutionary pathway and explore combining it with existing molecular representations. Therefore, this paper proposes the molecular evolutionary network (MEvoN) for molecular representations. First, we construct the MEvoN using molecules with a small number of atoms and generate evolutionary paths utilizing similarity calculations. Then, by modeling the atomic-level changes, MEvoN reveals their impact on molecular properties. Experimental results show that the MEvoN-based molecular property prediction method significantly improves the performance of traditional end-to-end algorithms by approximately 33% on both the QM7 and QM9 datasets.

Code — https://github.com/DrugD/MEvoN

Introduction

Molecular evolution is the process of exploring potential molecular structures and properties by simulating the evolution of molecules in nature using structural mutations (e.g., substitutions, additions, deletions and isomerization) to make the molecules evolve in the chemical space (van Deursen and Reymond 2007; Lameijer et al. 2006). This concept is widely applied in molecular generation (Li et al. 2024b; Hu et al. 2025) and optimization (Xiong et al. 2026, 2025; Li et al. 2024a) for novel chemical structure discovery and potential active molecule identification. For example, by simulating Darwinian evolution through crossover and mutation, genetic algorithms (Fu et al. 2022; Devi et al. 2019) continuously optimize molecular structures to find optimal solutions in vast chemical space. Chemical space travel, proposed by Ruud van Deursen and Jean-Louis Reymond (van Deursen and Reymond 2007), combines molecular evolution with algorithms to efficiently obtain target molecules through multiple mutations that start with an initial molecule. For larger molecules, protein changes are interconnected within a multi-dimensional space through mechanisms such as mutations. In this scenario, random mutation and natural selection cooperate to shape the structure and function of larger molecules (Chen et al. 2025b,a; H and H. 2003). For instance, the ESM3 multimodal language model can simulate this natural evolutionary process (Hayes et al. 2025), significantly enhancing the model's analytical and inference capabilities. Hence, this evolutionary mechanism improves our understanding of the relationship between molecular structure and biological activity and leverages the structural variation patterns among similar molecules.

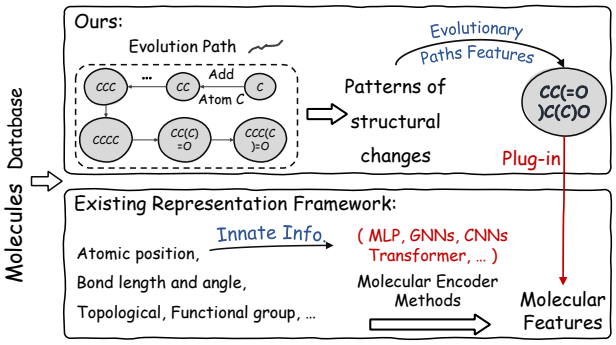

Figure 1: Molecular evolutionary network reveals the evolution pathways, providing a quantitative method for assessing the magnitude and direction of molecular property changes.

Molecular representation methods (Li et al. 2025c; Chen et al. 2025c) based on graphs (Yu et al. 2025a; Li et al. 2025d; Shalini and Mohan 2018) and sequences (Zhang et al. 2025) have been widely used in drug screening (Li et al. 2025a; Omodunbi et al. 2024; Moshkov et al. 2023; Vincent et al. 2022), materials science (Born and Manica 2023; Nazimuddin and Ali 2022), and molecular design (Yu et al. 2025b; Li et al. 2025b). During drug discovery,

*Corresponding author

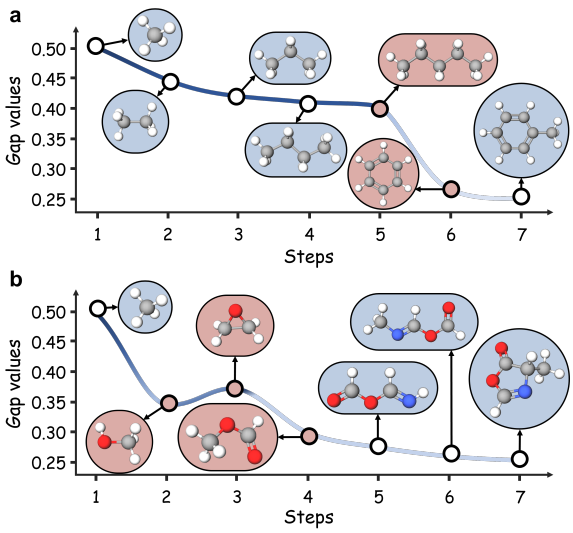

Figure 2: Evolutionary paths and molecular property changes for two molecules from the QM9 dataset. (a) corresponds to 'Cc1ccccc1', while (b) corresponds to 'CC1N=COC1=O'.

graph- and sequence-based methods (Sharma, Kumar, and Trivedi 2025; Li et al. 2024d,c; Sonti, Rukmini, and Munagala 2025) are used to screen potential candidates from large molecular datasets. As shown in Figure 1, existing molecular representation methods extract innate information from molecules (Wang et al. 2024b; Satorras, Hoogeboom, and Welling 2021; Schütt et al. 2017), such as atomic structures, chemical bonds, or other two-and three-dimensional features, often overlooking their historical evolutionary paths. This raises the question: **can we enhance molecular representations by simulating the molecule evolution process, and extracting and analyzing the structural changes within the evolutionary pathway?** By extracting and analyzing structural changes along evolution pathways, we may gain deeper insights into their influence on molecular representations, and integrating these findings with existing methods can enhance the overall understanding and depiction of molecular properties.

Notably, similar molecules inherently contain rich information, and their property variations often follow certain trends. Therefore, we conduct an evolutionary analysis on molecules from the QM9 dataset, using the molecular orbital gap (i.e., HOMO-LUMO gap) as the target property. We focus on molecules with high similarity and a single atom difference. Figure 2 illustrates two molecular evolutionary paths: "Cc1ccccc1" and "CC1N=COC1=O," highlighting these trends. The property variation curves in Figure 2 (a) exhibit a sharp decrease between steps 5 and 6, when a carbon chain transforms into a benzene ring. This is because an increase in molecular size normally leads to a decrease in the energy gap. In addition, closing the carbon chain into a benzene ring introduces additional electron delocalization through its aromatic structure, significantly reducing the HOMO-LUMO energy gap (Cornil et al. 2001). This aligns with the general trend that aromatic molecules tend to have lower energy gaps (Pope and Swenberg 1999). Similarly, as shown in Figure 2 (b), significant changes in the evolutionary path of 'CC1N=COC1=O' occur between steps 2–3 and 3–4. These steps involve the formation and breaking of a 3-membered ring due to its high energy strain. Cleaving a 3-membered ring releases strain energy and alters the electronic structure, significantly impacting energy levels (Planells and Ferao 2020). By analyzing these mutation effects and patterns, we gain a deeper understanding of the relationship between molecular structure and properties.

Therefore, we explore the possibility of employing the phylogenetic analysis methods used in genomics and protein sequencing to construct a network that describes molecule evolution. This network can simulate atomic-level changes during molecular evolution. As a result, we propose the **m**olecular **evo**lutionary **n**etwork (MEvoN) for molecular representations. MEvoN regards molecules with fewer atoms as ancestral nodes and those with more as descendants. Thus, we can construct the evolutionary relationships by calculating the similarity between these two molecular node types. The MEvoN is formed by various evolutionary paths and molecular node sets, revealing the impact of atomic-level changes on molecular properties. Furthermore, we demonstrate the application of the MEvoN-based molecular property prediction method (MEvoN-MPP). The MEvoN-MPP method combines the evolutionary path- and label-aware modules to effectively capture the evolutionary information. The experimental results demonstrate that MEvoN-MPP, as a basic property prediction model plug-in, effectively integrates the molecule's evolutionary path information with the inherent features to enhance its representation. This paper's contributions are as follows:

- A novel molecular representation paradigm based on the evolutionary network is proposed. Evolutionary relationships are constructed by calculating the similarities between molecules, thus helping to analyze the influence of atomic-level changes on molecular properties.
- To integrate evolutionary information with molecular features, we propose the MEvoN-MPP method. Experiments on several datasets indicate that MEvoN-MPP improves molecular representation by an average of 33%.

Molecule Evolutionary Network

Notations. In this section, we systematically describe the method and principles for constructing the MEvoN. Figure 3 presents the MEvoN model's construction process. We formulate the MEvoN as a network representation $\mathcal{N} = (\mathcal{V}, \mathcal{E})$. The molecules are represented as the set of node $\mathcal{V}$, and $\mathcal{E}$ is the set of edges that connects the nodes. The set of nodes $\mathcal{V}$ contains N molecules, and the corresponding properties of each molecule are represented by $P_i \in \mathcal{P}$. Furthermore, each MEvoN is constructed from the molecular dataset with $\mathcal{M}$ representing the set of all the molecules from one dataset. Algorithmic complexity and pseudocode for MEvoN construction are presented in Appendix A.

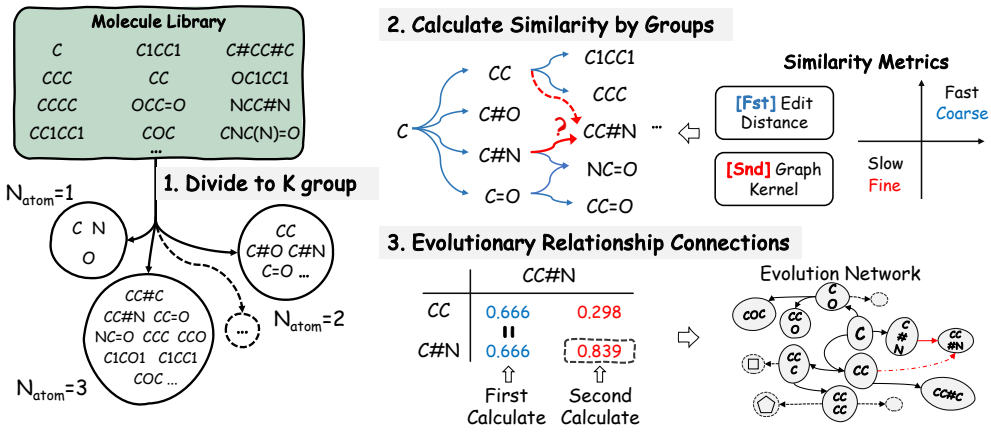

Figure 3: The MEvoN method's construction process. The steps are: 1) group molecules by atom count; 2) calculate inter-group similarity; and 3) determine evolutionary relationships based on multiple similarity measures.

Molecular Grouping

The MEvoN's construction aims to explore the evolutionary relationships among molecules by tracing their structural and compositional changes. First, to facilitate the MEvoN's construction, molecules are grouped according to their atomic compositions (see Figure 3). Thus, the number of atoms in each molecule serves as a distinguishing feature, organizing them into hierarchical groups. Molecules with fewer atoms form the base or initial network stages, while those with more appear later.

During the grouping process, the number of atoms $\mathrm{N}_{\text{atoms}}$ in each molecule is a positive integer. This constraint ensures that each molecule can be uniquely categorized according to its atomic count. The grouping process involves molecular iterations within the dataset (typically represented by SMILES strings). It also involves extracting the number of atoms $\mathrm{N}_{\text{atoms}}(\cdot)$ in each molecule, and categorizing them into the corresponding atom count groups G_k. These groups serve as the evolutionary network's initial levels. Formally, the molecules are grouped as:

$$G_k = \{m_i \mid \mathrm{N}_{\text{atoms}}(m_i) = k,\ m_i \in \mathcal{M}\}, \tag{1}$$

where m_i denotes a molecule and $\mathrm{N}_{\text{atoms}}(m_i)$ is the number of atoms in m_i.

The grouping process begins with the initial molecules, such as the molecule with one C atom. Molecules with the same number of atoms k are grouped into the set G_k, representing different stages of molecular evolution. The purpose of the molecular grouping is to simulate the gradual increase in atom count that is typically observed in natural molecule evolution. Therefore, evolutionary relationships are not constructed between molecules within the same group G_k.

Inter-Group Similarity Calculation

The MEvoN method construction's core lies in calculating evolutionary distance or molecular similarity to establish the evolution pathways. The similarity between two molecules, m_i and m_j, denoted as $S(m_i, m_j)$, can be measured using various similarity metrics. The similarity value is between $[0, 1]$, where 1 indicates complete similarity and 0 implies no similarity.

- *Fingerprint-based similarity*: Fingerprint-based similarity methods (Wang et al. 2024a) represent molecules as binary fingerprint vectors, where each bit indicates the presence or absence of a specific structural feature within the molecule. The Tanimoto coefficient (Chung et al. 2019) is the most widely used similarity measure, quantifying the overlap between two binary fingerprints as follows:

$$S_{\text{fp}}(m_i, m_j) = \frac{|F(m_i) \cap F(m_j)|}{|F(m_i) \cup F(m_j)|}, \tag{2}$$

  where $F(m_i)$ and $F(m_j)$ represent the fingerprint sets of molecules m_i and m_j, respectively.
- *Graph-based similarity*: Graph-based similarity is determined by comparing the graphs using graph kernels, such as the Weisfeiler–Lehman graph kernel $\mathcal{S}_{\text{wl}}(m_i, m_j)$ (Shervashidze et al. 2011), described as follows:

$$\mathcal{S}_{\text{wl}}(m_i, m_j) = \mathrm{WL}(G(m_i), G(m_j)), \tag{3}$$

  where $G(m_i)$ and $G(m_j)$ represent the graph representations of molecules m_i and m_j, respectively.
- *Edit distance similarity*: The molecular edit distance $d_{\text{edit}}(m_i, m_j)$ measures the number of changes required to convert one molecule into another, where the changes correspond to atom insertions, deletions, and substitutions. The edit distance is given by:

$$\mathcal{D}_{\text{edit}}(m_i, m_j) = \min_{\Theta} \left(\sum_{opt \in \Theta} \mathrm{cost}(opt) \right), \tag{4}$$

  where, Θ represents the set of possible edit operations (i.e., insertions, deletions, and substitutions), and each opt is the cost associated with a specific operation. Thus, the edit distance similarity $\mathcal{S}_{\text{edit}}(m_1, m_2)$ is defined as:

$$\mathcal{S}_{\text{edit}}(m_1, m_2) = 1 - \frac{\mathcal{D}_{\text{edit}}(m_1, m_2)}{\max(\mathrm{len}(m_1), \mathrm{len}(m_2))}, \tag{5}$$

The similarity measure $S(m_i, m_j)$ ranges from 0 to 1. A higher value indicates greater similarity, capturing molecular structural differences and feature changes. To ensure consistency and clarity along the evolutionary pathway, the similarity calculation is only conducted when the inter-group condition $i < j$ met. Specifically, similarity is calculated only between $m_i \in G_i$ and $m_j \in G_j$. As a result, the inter-group similarity calculation effectively avoids redundant computations, ensuring that the evolutionary path progresses effectively within the hierarchical structure.

Establishing Evolutionary Relationships

After calculating the molecular similarities, valid edges $\mathcal{E}$ are added to the MEvoN to represent evolutionary relationships. These edges are formed based on the similarity values $S(m_i, m_j)$ between molecule pairs. Then, the predefined thresholds θ_1 and θ_2 are used to determine whether two molecules are evolutionarily related.

Consider two evolutionary groups, G_i and G_j, where $i < j$. This implies that G_i represents molecules from an earlier evolutionary stage, and G_j denotes those from a later phase. To calculate the similarity between G_i and G_j, each molecule $m_q \in G_j$ is compared with every $m_p \in G_i$, and the similarity between m_q and m_p is computed.

To establish evolutionary relationships, the $\mathcal{S}_{\text{edit/fp}}$ similarity function is used to calculate molecular similarity:

$$\text{Pair}^1 = \left\{(m_p, m_q) \mid \mathcal{S}_{\text{edit/fp}}(m_p, m_q) \geq \theta_1\right\}, \quad (6)$$

Thus, we obtain Pair^1 as the result of the first-stage screening. Let the maximum value in Pair^1 be denoted as $\text{Pair}^1_{\text{max}} = \left\{(m_p, m_q) \mid \mathcal{S}_{\text{edit/fp}}(m_p, m_q) = \text{Max}\left(\text{Pair}^1\right)\right\}$. In this scenario, $\mathcal{S}_{\text{edit/fp}}$ is the quickest and most efficient measure. However, a second round of similarity calculation is required to obtain a more precise evolutionary relationship, when the number of elements in $\text{Pair}^1_{\text{max}} > 1$.

In the second round, a more precise similarity calculation is conducted using the $\mathcal{S}_{\text{wl}}$ operation. This graph-based measure captures more intricate topological similarities between molecules, enhancing its ability to distinguish subtle structural differences, expressed as:

$$\text{Pair}^2 = \left\{(m'_p, m'_q) \mid \mathcal{S}_{\text{wl}}(m'_p, m'_q) \geq \theta_2, \right\}, \quad (7)$$

where $(m'_p, m'_q) \in \text{Pair}^1_{\text{max}}$ and Pair^2 is the final result. With the molecule pair Pair^2, a new set of nodes $\mathcal{V}'$ and edges $\mathcal{E}'$ can be added to $\mathcal{N}$:

$$\begin{cases} \mathcal{V}' = \left\{m'_p, m'_q \mid (m'_p, m'_q) \in \text{Pair}^2\right\}, \\ \mathcal{E}' = \left\{(m'_p, m'_q) \mid (m'_p, m'_q) \in \text{Pair}^2\right\}, \end{cases} \quad (8)$$

Thus, the evolutionary network $\mathcal{N} = (\mathcal{V} \cup \mathcal{V}', \mathcal{E} \cup \mathcal{E}')$ is updated by incorporating the new nodes and edges, where $\mathcal{V}$ and $\mathcal{E}$ represent the original set of nodes and edges, while $\mathcal{V}'$ and $\mathcal{E}'$ represent the newly added ones.

Molecular Property Prediction Using MEvoN

The evolutionary relationships between molecules provide valuable contextual information for understanding structure-property dependencies. By leveraging MEvoN, we can incorporate the molecules' evolutionary paths as auxiliary information, thereby enhancing representation. Therefore, we propose the MEvoN-MPP model, a MEvoN-based molecular property prediction method. MEvoN-MPP includes the path- (EvoP) and label-aware (EvoA) modules, and the molecular encoder (MolE). The EvoP module captures the evolutionary relationships between molecules, while the EvoA module leverages label information to weight the evolutionary paths, enabling the model to understand each molecule's evolutionary context more effectively. MolE encodes the structural features from molecular graphs and can utilize any deep learning model capable of encoding molecules, such as Graph Neural Networks (GNNs), Convolutional Neural Networks (CNNs), and Transformers. The steps of MEvoN-MPP are as follows:

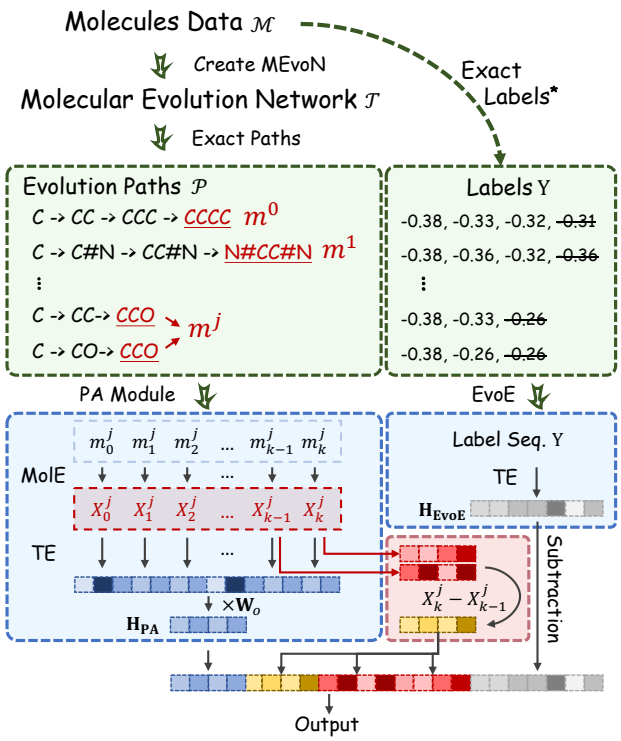

Figure 4: Overview of MEvoN-MPP, which employs the MEvoN to predict various molecular properties. The process includes evolutionary feature extraction and property prediction.

First, let $\mathcal{N}$ be the MEvoN constructed from a set of molecules $\mathcal{M}$ (see Section). For a single molecule m_i, we locate its position within the MEvoN and trace its evolutionary paths (denoted as $\mathcal{P}$). The evolutionary path $\mathcal{P}^i$ refers to the set of paths from the network's root node to the molecule $m_i \in \mathcal{M}$, which can be collected by the backtracking algorithm. Specifically, $\mathcal{P}^i$ is a path set, where each path $\mathcal{P}_k^i = \left(m_0^i, m_1^i, \ldots, m_{l-1}^i, m_l^i\right)$ represents an evolutionary path from the root to m_i. The length of path $\mathcal{P}_k^i$ denotes l, i.e., the number of molecules included in the path. Each molecule's graph features are extracted with MolE and serve as the initial path features for $\mathcal{P}_k^i$. This can be expressed as:

$$\mathbf{P}_k^i = \left[\text{MolE}(m_0^i), \ldots, \text{MolE}(m_l^i)\right], \quad (9)$$

Methods	MAE	Methods	MAE
• Pretrained Models			
GROVER	94.5(3.8)	MolCLR	66.8(2.3)
GEM	58.9(0.8)	Uni-Mol	**41.8(0.2)**
• Backbone Models with MEvoN-MPP			
GCN	122.9(2.2)	GIN	124.8(0.7)
+ MEvoN-MPP	45.9(3.4)	+ MEvoN-MPP	65.6(6.3)
(Improve.)	62.6%	(Improve.)	47.4%
Equiformer	58.3(7.4)	ViSNet	69.8(3.6)
+ MEvoN-MPP	55.5(5.8)	+ MEvoN-MPP	56.9(5.7)
(Improve.)	4.6%	(Improve.)	18.5%

Table 1: Comparison of MAE results for different models on QM7 dataset. The best results are **bolded**, and the second-best are underlined.

where $[\cdot]$ denotes element concatenation. Then, we obtain the path features $\mathbf{P}^i \in \mathbb{R}^{K\times L\times F}$, where K is the number of evolutionary paths $\mathcal{P}^i$, L is the maximum path length, and F is the feature dimension of each molecule obtained from the MolE. Thus, we encode the $\mathbf{P}^i$ as follows:

$$\begin{cases}\mathbf{H}_{\text{pos}} = \mathbf{W}_e\mathbf{P}_i + \mathbf{b}_e + \mathbf{E}_p, \\ \mathbf{H}_{\text{out}} = \text{TransformerEncoder}(\mathbf{H}_{\text{pos}}),\end{cases} \tag{10}$$

where $\mathbf{W}_e \in \mathbb{R}^{F\times D}$ is the weight matrix, $\mathbf{b}_e \in \mathbb{R}^D$ denotes the bias vector, and D represents the embedding dimension. To incorporate sequential dependencies, we incorporate learnable positional encoding $\mathbf{E}_p \in \mathbb{R}^{L\times D}$ into the embeddings. After that, the position embeddings $\mathbf{H}_{\text{pos}} \in \mathbb{R}^{K\times L\times D}$ are then passed through a Transformer encoder to capture the dependencies among molecules, producing the output sequence $\mathbf{H}_{\text{out}} \in \mathbb{R}^{K\times L\times D}$.

Subsequently, the final prediction is obtained by selecting the last hidden state $\mathbf{h}_{\text{last}} \in \mathbb{R}^{K\times 1\times D}$ from the output sequence, which is passed through a fully connected layer to produce the predicted molecular property $\mathbf{H}_{\text{EvoP}}$:

$$\mathbf{H}_{\text{EvoP}} = \mathbf{W}_o\mathbf{h}_{\text{last}} + \mathbf{b}_o, \tag{11}$$

where $\mathbf{W}_o \in \mathbb{R}^{D\times 1}$ is the weight matrix and $\mathbf{b}_o \in \mathbb{R}^K$ is the bias term.

The EvoL module's computation is similar to that of EvoP. The EvoL input is the molecular properties in $\mathcal{P}_k^i$, denoted as $\mathbf{Y}^i \in \mathbb{R}^L$. In this case, y_{l-1}^i and y_{l-2}^i represent the labels of the last and the second-to-last valid molecules of $\mathbf{Y}^i$, respectively. The label of each path's last valid molecule, y_{l-1}^i, is masked to prevent data leakage. After encoding with EvoL, the label path feature $\mathbf{H}_{\text{EvoL}}$ is obtained, which can be expressed as $\mathbf{H}_{\text{EvoL}} = \text{EvoL}(\mathbf{Y}^i)$.

For path $\mathcal{P}_k^i$, the last two valid molecules, m_{l-2}^i and m_{l-1}^i, serve as input to the MolE. MEvoN-MPP predicts the property changes caused by the molecular pair, learning their evolution patterns—specifically, the property changes arising from the addition of atoms and chemical bonds at different positions. Then, the Evo and the Mol branches are used to predict the property changes caused by the molecular pair (m_{l-2}^i, m_{l-1}^i) and the properties of m_{l-1}^i, respectively.

Dataset	Molecules	MEvoN		Max Path
		Edges	Nodes	
QM7	6832	9095	6832	110
QM8	21766	27068	21766	46
QM9	133885	165790	133330	253

Table 2: Overview of MEvoN construction on the QM7 (Rupp et al. 2012), QM8 (Ruddigkeit et al. 2012), and QM9 (Ramakrishnan et al. 2014) datasets. 'Max Path' refers to the maximum number of distinct paths originating from a single carbon atom 'C' and connecting to a specific molecule.

In the MolE branch, property prediction is performed directly on the feature X_1 extracted by the MolE, and the output is denoted as y_1^{pred}. The molecular representations X_1 and X_2 can be expressed as:

$$X_1 = \text{MolE}(m_{l-1}^i), \quad X_2 = \text{MolE}(m_{l-2}^i), \tag{12}$$

In the Evo branch, the molecular evolutionary pair's features X_1 and X_2 are extracted using the MolE. Then, the difference between these features $\Delta X = X_2 - X_1$ is computed. Subsequently, the difference feature ΔX is concatenated with the evolutionary path features $\mathbf{H}_{\text{EvoP}}$ and $\mathbf{H}_{\text{EvoL}}$ extracted by the EvoP and EvoL modules. This can be expressed as:

$$\begin{cases}y_1^{\text{pred}} = \mathcal{F}(X_1), \\ y_2^{\text{pred}} = \mathcal{F}\left([\Delta X, X_1, X_2, \mathbf{H}_{\text{EvoP}}, \mathbf{H}_{\text{EvoL}}]\right),\end{cases} \tag{13}$$

where the multilayer perceptron is denoted as $\mathcal{F}(\cdot)$. Finally, the loss function is defined as:

$$\mathcal{L} = \alpha \cdot \text{MSE}(y_1^{\text{pred}}, y_{l-1}^i) + \beta \cdot \text{MSE}(y_2^{\text{pred}}, y_{l-2}^i - y_{l-1}^i). \tag{14}$$

where α and β are hyperparameters for loss weights and the $\text{MSE}(\cdot)$ stands for mean squared error.

Experiments

This study focuses on MEvoN-based molecular property prediction utilizing the QM7 (Rupp et al. 2012) and QM9 (Ramakrishnan et al. 2014) datasets, which provide extensive quantum chemical properties for molecular modeling and property prediction. The datasets were randomly split into training, validation, and test sets with a ratio of 8:1:1. For the QM7 experiments, seeds 38–42 were used, while for QM9, random seed 42 was employed. These regression tasks apply the mean absolute error (MAE) used as the performance metric. The default values of the loss weights α and β are both 1.

Validating MEvoN

To validate the MEvoN model's effectiveness, we constructed three evolutionary networks based on the QM7, QM8, and QM9 datasets. Table 2 shows the number of molecules in the datasets, the number of edges and nodes in MEvoN, and the maximum number of paths per molecule. The networks were constructed with a similarity threshold

Methods	Property	$\varepsilon_{\text{HOMO}}$	$\varepsilon_{\text{LUMO}}$	$\Delta\varepsilon$	ZPVE	μ	α	$\langle R \rangle^2$	C_V
	Unit	eV	eV	eV	eV	D	bohr^3	bohr^2	cal/mol K
GCN	Original	0.2539	0.1336	0.5928	0.3273	0.7943	2.2469	101.7469	1.5261
	Mol-branch	**0.1419**	**0.1207**	0.2605	0.0649	0.6134	2.5860	68.2874	0.9806
	Evo-branch	0.1453	0.1373	**0.2579**	**0.0322**	**0.5825**	**0.7621**	**46.5656**	**0.3432**
	(Improve.)	44.11%	9.64%	56.50%	90.16%	26.67%	66.08%	54.23%	77.51%
GIN	Original	0.2174	0.1247	0.2916	0.1622	0.4981	2.0459	77.8414	1.0610
	Mol-branch	0.1662	**0.1207**	0.2270	0.1458	0.5087	2.5720	56.3729	3.0892
	Evo-branch	**0.1662**	0.1267	**0.2265**	**0.0572**	**0.4893**	**0.8115**	**35.2868**	**0.5370**
	(Improve.)	23.57%	3.19%	22.32%	64.75%	1.77%	60.34%	54.67%	49.39%
SchNet	Original	0.0772	0.0586	0.1066	0.0053	0.0972	0.1813	1.6577	0.0625
	Mol-branch	0.0538	**0.0565**	**0.0809**	**0.0044**	0.0646	0.1368	**1.3347**	**0.0547**
	Evo-branch	**0.0534**	0.0578	0.0820	0.0064	**0.0644**	**0.1271**	1.5003	0.0557
	(Improve.)	30.73%	3.61%	24.06%	17.60%	33.74%	29.92%	19.48%	10.88%
ComENet	Original	0.0924	0.0638	0.1232	0.0068	0.1034	0.2997	2.2417	0.1200
	Mol-branch	0.0568	0.0533	**0.0862**	**0.0029**	0.0833	0.2268	2.1444	0.0918
	Evo-branch	**0.0559**	**0.0530**	0.0862	0.0107	**0.0825**	**0.2151**	**2.0802**	**0.0892**
	(Improve.)	38.56%	16.45%	30.00%	57.69%	19.44%	28.23%	4.34%	25.69%
Equiformer	Original	0.0263	0.0236	0.0492	0.0019	0.0183	0.0801	0.6542	0.0349
	Mol-branch	**0.0223**	0.0233	0.0464	**0.0005**	0.0097	**0.0564**	**0.2895**	0.0283
	Evo-branch	0.1156	**0.0227**	**0.0439**	0.0016	**0.0087**	0.0570	0.5137	**0.0277**
	(Improve.)	15.21%	3.81%	10.30%	73.68%	52.46%	29.59%	55.75%	20.63%
ViSNet	Original	0.0322	0.0258	0.0515	0.0014	0.0191	0.1350	0.8604	0.0289
	Mol-branch	**0.0272**	**0.0252**	**0.0501**	**0.0009**	**0.0184**	**0.0719**	**0.5444**	**0.0271**
	Evo-branch	0.0295	0.0253	0.0530	0.0017	0.0186	0.0745	0.8538	0.0334
	(Improve.)	15.53%	1.83%	2.72%	35.71%	3.66%	46.74%	36.73%	6.23%

Table 3: Comparison of MAE results for different models on QM9 dataset. The best results are **bolded**.

set to 0.3. Notably, 555 molecules from QM9 were excluded from the network because F-containing molecules (a total of 446) exhibited low similarity with most of the others, making it challenging to establish evolutionary relationships. Additionally, some N-containing cyclic structures showed low similarity to the main molecules and were also excluded.

A detailed analysis of the QM9 dataset revealed that the mutations could be categorized into 14 types: adding C, N, and O atoms; increasing or decreasing bond order; forming or breaking 3/4/5/6-membered rings; and changing ring types. We observed the influence of different mutations on molecular relationships. Based on SHapley Additive exPlanations (SHAP) analysis (Lundberg and Lee 2017), we built a simplified regression model to quantify the impact of various mutations on molecular properties, using HOMO-LUMO gap (GAP) as an example. The results, shown in Figure 5, reveal regular patterns in the evolutionary modifications' effects on molecular properties. For instance, adding 5- and 6-membered rings, and introducing O and N atoms, generally decreases the GAP. Meanwhile, adding a 3-membered ring tends to increase the GAP. The effects of increasing or decreasing bond order are mutually exclusive, with an increase leading to GAP reduction. Furthermore, adding a single C atom has a minimal effect. Details of the analysis on property variation under structural changes are provided in Appendix B.4.

These patterns are consistent with the theoretical quantum chemistry findings in various studies (Pope and Swenberg 1999), providing strong evidence for the MEvoN method's capacity and significance in capturing molecular evolutionary relationships.

Molecular Property Prediction Using MEvoN

To validate the effectiveness of our proposed MEvoN-MPP method across various molecular encoding architectures, we conducted property prediction experiments using the QM7 and QM9 datasets. For the QM9 experiments, we selected eight commonly used properties as the targets following the MolCLR method (Wang et al. 2022). The baselines included SOTA pre-training methods (GEM (Fang et al. 2022), Uni-Mol (Zhou et al. 2023), etc.) and SOTA backbone methods (Equiformer (Liao and Smidt 2023), VisNet (Wang et al. 2024c), etc.). Details of baseline models, including backbone architectures and pre-training strategies, are provided in Appendix B.

The results on the QM9 dataset, shown in Table 3, demonstrate that MEvoN effectively enhances molecular representations by an average of 32.3%. The QM7 results, shown in Table 1, indicate that our GCN-based model could compete with SOTA pre-training methods like GEM and Uni-Mol, with a 33.2% improvement. For example, Uni-Mol is pre-trained on a database containing 19 million molecules and 209 million conformations, whereas our method is trained and evaluated only on the QM7 dataset (with less than 7,000 molecules).

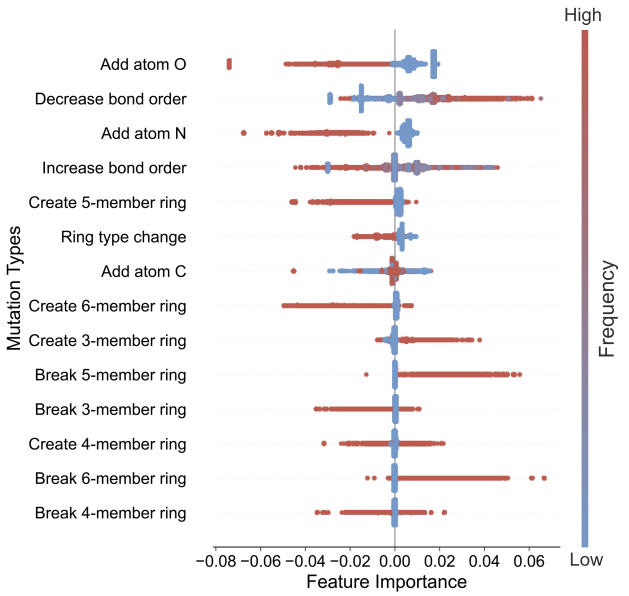

Figure 5: The strength and direction of different mutation types' effects. The color indicates the frequency of mutation occurrence, and the x-axis indicates its effect on the GAP value.

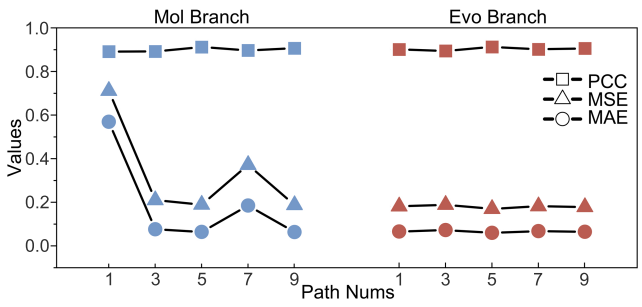

Figure 6: Hyperparameter experiment exploring the impact of different path numbers on evolutionary path representation.

Ablation Experiments

To investigate the contributions of different modules in MEvoN-MPP, we conducted an ablation study to evaluate the EvoP and EvoL modules' ability to leverage MEvoN's molecular feature representations and impact exploration capacity effectively. The ablation study's results are presented in Table 4. When only MolE was used for prediction, the MAE was 0.2916. Introducing the EvoP or EvoL modules individually increased the MAE to approximately 0.98 and 0.40, respectively. This indicates that these modules are not effective when applied independently. However, when the EvoP and EvoL modules are applied together, model performance improved by approximately 22.3% compared to using only MolE. This demonstrates that the collaboration between the EvoP and EvoL modules significantly enhances the model's ability to predict molecular properties. Furthermore, the two modules effectively integrate the molecule's local and global features by combining path and label sequence encoding. This fusion mechanism enables the model to focus on structural changes at key positions while capturing the molecule's overall evolutionary trends, leading to a more comprehensive and precise representation of molecular features.

EvoP	EvoL	MolE	Result(MAE)	
			Mol-branch	Evo-branch
✗	✗	✓	0.2916	-
✓	✗	✓	0.9809	0.5187
✗	✓	✓	0.4051	0.3521
✓	✓	✓	**0.2270**	**0.2265**

Table 4: Ablation study of MEvoN-MPP on the QM9 dataset for the GAP property. The best results are bolded.

Hyperparameter Experiments

In our MEvoN model, each molecule typically has more than one evolutionary path. To investigate the impact of embedding different evolutionary path numbers on molecular representation, we conducted hyperparameter experiments. As shown in Figure 6, we evaluated the $\varepsilon_{\mathrm{HOMO}}$ property on the QM9 dataset with path numbers K set to 1, 3, 5, 7, and 9, using four regression metrics. To ensure fair comparisons, the training epoch was fixed at 150 across all experiments. The results indicate that K has a significant effect on molecular representation. Specifically, when K is set between 3 and 5, the model demonstrates stable performance, with few errors and high prediction accuracy during regression tasks. Conversely, insufficient or excessive paths lead to instability. Obtaining insufficient paths may result in a failure to capture the diversity and complexity of molecular evolution, reducing prediction accuracy. In contrast, an excessive number of paths increases computational complexity, slowing model convergence and resulting in decreased performance during comparative evaluations.

Conclusion

This paper introduces a novel molecular representation method based on the MEvoN. By simulating the evolutionary pathway from ancestral to current structures, MEvoN captures dynamic, multi-level features reflecting molecular structural changes. When combined with traditional encoding methods, MEvoN enhances molecular representation for downstream tasks. To validate its effectiveness, we applied MEvoN to molecular property prediction tasks, experimenting on eight sub-tasks from the QM7 and QM9 datasets and using four encoding methods. The results demonstrate a 32.3% average performance improvement. Therefore, the MEvoN effectively captures structural variations, deepening our understanding of the relationship between molecular evolution and properties, with promising applications in drug discovery and materials optimization.

Acknowledgments

This work was supported in part by the Natural Science Foundation of China (No. 62476203), Key Project of Traditional Chinese Medicine Joint Fund of Hubei Provincial Natural Science Foundation (No.2025AFD47), Hubei Province Science and Technology Innovation Plan Project (No.2025BCB035), the Guangdong Provincial Natural Science Foundation General Project (No. 2025A1515012155), the Shenzhen Natural Science Foundation Project (No. JCYJ20250604122534006)

References

Born, J.; and Manica, M. 2023. Regression transformer enables concurrent sequence regression and generation for molecular language modelling. *Nature Machine Intelligence*, 5(4): 432–444.

Chen, J.; Cai, X.; Wu, J.; and Hu, W. 2025a. Antibody Design and Optimization with Multi-scale Equivariant Graph Diffusion Models for Accurate Complex Antigen Binding. In Kwok, J., ed., *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, 2722–2730. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Chen, J.; Xiong, Y.; Li, K.; Zhang, H.; Cai, X.; Hu, W.; and Wu, J. 2025b. FP-AbDiff: Improving Score-based Antibody Design by Capturing Nonequilibrium Dynamics through the Underlying Fokker-Planck Equation. *arXiv preprint arXiv:2511.03113*.

Chen, M.; Gong, X.; Pan, S.; Wu, J.; Lin, F.; Du, B.; and Hu, W. 2025c. Unified Knowledge-Guided Molecular Graph Encoder with multimodal fusion and multi-task learning. *Neural Networks*, 184: 107068.

Chung, N. C.; Miasojedow, B.; Startek, M.; and Gambin, A. 2019. Jaccard/Tanimoto similarity test and estimation methods for biological presence-absence data. *BMC bioinformatics*, 20(Suppl 15): 644.

Cornil, J.; Beljonne, D.; Calbert, J.-P.; and Brédas, J.-L. 2001. Interchain interactions in organic π-conjugated materials: impact on electronic structure, optical response, and charge transport. *Advanced materials*, 13(14): 1053–1067.

Devi, R. V.; Sathya, S. S.; Kumar, N.; and Coumar, M. S. 2019. Multi-objective monkey algorithm for drug design. *International Journal of Intelligent Systems and Applications*, 11(3): 31.

Fang, X.; Liu, L.; Lei, J.; He, D.; Zhang, S.; Zhou, J.; Wang, F.; Wu, H.; and Wang, H. 2022. Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence*, 4(2): 127–134.

Fu, T.; Gao, W.; Coley, C.; and Sun, J. 2022. Reinforced genetic algorithm for structure-based drug design. *Advances in Neural Information Processing Systems*, 35: 12325–12338.

H, W. K.; and H., L. W. 2003. Molecular evolution meets the genomics revolution. *Nature genetics*, 33(3): 255–265.

Hayes, T.; Rao, R.; Akin, H.; Sofroniew, N. J.; Oktay, D.; Lin, Z.; Verkuil, R.; Tran, V. Q.; Deaton, J.; Wiggert, M.; Badkundri, R.; Shafkat, I.; Gong, J.; Derry, A.; Molina, R. S.; Thomas, N.; Khan, Y. A.; Mishra, C.; Kim, C.; Bartie, L. J.; Nemeth, M.; Hsu, P. D.; Sercu, T.; Candido, S.; and Rives, A. 2025. Simulating 500 million years of evolution with a language model. *Science*, 0(0): eads0018.

Hu, C.; Li, K.; Hu, L.; Xiong, Y.; Cai, X.; and Hu, W. 2025. Collaborative Drug Design Based on A Drug-Drug Interaction-Guided Diffusion Model. In *2025 28th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, 1374–1380. IEEE.

Lameijer, E.-W.; Kok, J. N.; Bäck, T.; and IJzerman, A. P. 2006. The molecule evoluator. An interactive evolutionary algorithm for the design of drug-like molecules. *Journal of chemical information and modeling*, 46(2): 545–552.

Li, K.; Cai, X.; Wu, J.; Du, B.; and Hu, W. 2024a. Fragment-Masked Molecular Optimization. *arXiv preprint arXiv:2408.09106*.

Li, K.; Gong, X.; Pan, S.; Wu, J.; Du, B.; and Hu, W. 2024b. Regressor-free molecule generation to support drug response prediction. *arXiv preprint arXiv:2405.14536*.

Li, K.; Gong, X.; Wu, J.; and Hu, W. 2024c. Contrastive Learning Drug Response Models from Natural Language Supervision. In Larson, K., ed., *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, 2126–2134. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Li, K.; Liu, W.; Luo, Y.; Cai, X.; Wu, J.; and Hu, W. 2024d. Zero-shot Learning for Preclinical Drug Screening. In Larson, K., ed., *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, 2117–2125. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Li, K.; Wu, Z.; Wang, S.; Wu, J.; Pan, S.; and Hu, W. 2025a. Drugpilot: Llm-based parameterized reasoning agent for drug discovery. *arXiv preprint arXiv:2505.13940*.

Li, K.; Wu, Z.; Xiong, Y.; Zhang, H.; Hu, L.; Liu, Z.; Zeng, J.; Wu, W.; Chen, M.; Chen, J.; et al. 2025b. BSL: A Unified and Generalizable Multitask Learning Platform for Virtual Drug Discovery from Design to Synthesis. *arXiv preprint arXiv:2508.01195*.

Li, K.; Xiong, Y.; Zhang, H.; Cai, X.; Wu, J.; Du, B.; and Hu, W. 2025c. Graph-Structured Small Molecule Drug Discovery Through Deep Learning: Progress, Challenges, and Opportunities. In *2025 IEEE International Conference on Web Services (ICWS)*, 1033–1042.

Li, K.; Zeng, Y.; Xiong, Y.-d.; Wu, H.-c.; Fang, S.; Qu, Z.-y.; Zhu, Y.; Du, B.; Gao, Z.-b.; and Hu, W.-b. 2025d. Contrastive learning-based drug screening model for GluN1/GluN3A inhibitors. *Acta Pharmacologica Sinica*, 1–13.

Liao, Y.-L.; and Smidt, T. 2023. Equiformer: Equivariant Graph Attention Transformer for 3D Atomistic Graphs. In *The Eleventh International Conference on Learning Representations*.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st*

International Conference on Neural Information Processing Systems, NIPS'17, 4768–4777. Curran Associates Inc. ISBN 9781510860964.

Moshkov, N.; Becker, T.; Yang, K.; Horvath, P.; Dancik, V.; Wagner, B. K.; Clemons, P. A.; Singh, S.; Carpenter, A. E.; and Caicedo, J. C. 2023. Predicting compound activity from phenotypic profiles and chemical structures. *Nature communications*, 14(1): 1967.

Nazimuddin, A.; and Ali, M. S. 2022. Application of differential geometry on a chemical dynamical model via flow curvature method. *Int. J. Mathem. Sci. Comp.(IJMSC)*, 8(1): 18–27.

Omodunbi, T. O.; Alilu, G. E.; Obohwemu, K. O.; and Ikono, R. N. 2024. Enhancing drug recommender system for peptic ulcer treatment. *Int. J. Inf. Technol. Comput. Sci*, 16(6): 15–26.

Planells, A. R.; and Ferao, A. E. 2020. Accurate ring strain energy calculations on saturated three-membered heterocycles with one group 13–16 element. *Inorganic Chemistry*, 59(16): 11503–11513.

Pope, M.; and Swenberg, C. E. 1999. *Electronic Processes in Organic Crystals and Polymers*. Oxford University Press.

Ramakrishnan, R.; Dral, P. O.; Rupp, M.; and Von Lilienfeld, O. A. 2014. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data*, 1(1): 1–7.

Ruddigkeit, L.; Van Deursen, R.; Blum, L. C.; and Reymond, J.-L. 2012. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *Journal of chemical information and modeling*, 52(11): 2864–2875.

Rupp, M.; Tkatchenko, A.; Müller, K.-R.; and von Lilienfeld, O. A. 2012. Fast and accurate modeling of molecular atomization energies with machine learning. *Physical Review Letters*, 108: 058301.

Satorras, V. G.; Hoogeboom, E.; and Welling, M. 2021. E(n) Equivariant Graph Neural Networks. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 9323–9332. PMLR.

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.

Shalini, R.; and Mohan, R. 2018. Drugs relationship discovery using hypergraph. *Int. J. Inf. Technol. Comput. Sci*, 10: 54–63.

Sharma, K.; Kumar, S.; and Trivedi, R. S. 2025. Diffuse, sample, project: plug-and-play controllable graph generation. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Shervashidze, N.; Schweitzer, P.; Van Leeuwen, E. J.; Mehlhorn, K.; and Borgwardt, K. M. 2011. Weisfeiler-lehman graph kernels. *Journal of Machine Learning Research*, 12(9).

Sonti, N.; Rukmini, M.; and Munagala, V. 2025. A Novel Approach for Enhancing COVID-19 Diagnosis Accuracy through Graph Neural Networks Using Respiratory Sound Data. *International Journal of Image, Graphics and Signal Processing (IJIGSP)*, 2: 47–61.

van Deursen, R.; and Reymond, J.-L. 2007. Chemical space travel. *ChemMedChem: Chemistry Enabling Drug Discovery*, 2(5): 636–640.

Vincent, F.; Nueda, A.; Lee, J.; Schenone, M.; Prunotto, M.; and Mercola, M. 2022. Phenotypic drug discovery: recent successes, lessons learned and new directions. *Nature Reviews Drug Discovery*, 21(12): 899–914.

Wang, J.; Zhang, L.; Sun, J.; Yang, X.; Wu, W.; Chen, W.; and Zhao, Q. 2024a. Predicting drug-induced liver injury using graph attention mechanism and molecular fingerprints. *Methods*, 221: 18–26.

Wang, Y.; Wang, J.; Cao, Z.; and Barati Farimani, A. 2022. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3): 279–287.

Wang, Y.; Wang, T.; Li, S.; He, X.; Li, M.; Wang, Z.; Zheng, N.; Shao, B.; and Liu, T.-Y. 2024b. Enhancing geometric representations for molecules with equivariant vector-scalar interactive message passing. *Nature Communications*, 15(1): 313.

Wang, Y.; Wang, T.; Li, S.; He, X.; Li, M.; Wang, Z.; Zheng, N.; Shao, B.; and Liu, T.-Y. 2024c. Enhancing geometric representations for molecules with equivariant vector-scalar interactive message passing. *Nature Communications*, 15(1): 313.

Xiong, Y.; Chen, J.; Li, K.; Zhang, H.; Cai, X.; and Hu, W. 2025. Hierarchical Bayesian Flow Networks for Molecular Graph Generation. *arXiv preprint arXiv:2510.10211*.

Xiong, Y.; Li, K.; Chen, J.; Zhang, H.; Lin, D.; Che, Y.; and Hu, W. 2026. Text-guided multi-property molecular optimization with a diffusion language model. *Information Fusion*, 127: 103907.

Yu, J.; Wu, Z.; Lu, J.; Wang, T.; and Wang, H. 2025a. A Centrality-based Graph Learning Framework. In Kwok, J., ed., *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, 3588–3596. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Yu, J.; Zheng, Y.; Koh, H. Y.; Pan, S.; Wang, T.; and Wang, H. 2025b. Collaborative expert llms guided multi-objective molecular optimization. *arXiv preprint arXiv:2503.03503*.

Zhang, H.; Liu, Z.; Meng, K.; Chen, J.; Wu, J.; Du, B.; Lin, D.; Che, Y.; and Hu, W. 2025. Zero-Shot Learning with Subsequence Reordering Pretraining for Compound-Protein Interaction. arXiv:2507.20925.

Zhou, G.; Gao, Z.; Ding, Q.; Zheng, H.; Xu, H.; Wei, Z.; Zhang, L.; and Ke, G. 2023. Uni-Mol: A Universal 3D Molecular Representation Learning Framework. In *The Eleventh International Conference on Learning Representations*.