Journal of Computer and Communications, 2025, 13(1), 152-168
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

MMGCF: Generating Counterfactual Explanations for Molecular Property Prediction via Motif Rebuild

Xiuping Zhang[1], Qun Liu[1], Rui Han[1,2]

[1]Chongqing Key Laboratory of Computational Intelligence, Chongqing University of Posts and Telecommunications, Chongqing, China
[2]College of Electronic and Information Engineering, Chongqing Open University, Chongqing, China
Email: liuqun@cqupt.edu.cn

How to cite this paper: Zhang, X.P., Liu, Q. and Han, R. (2025) MMGCF: Generating Counterfactual Explanations for Molecular Property Prediction via Motif Rebuild. *Journal of Computer and Communications*, **13**, 152-168.
https://doi.org/10.4236/jcc.2025.131011

Received: December 30, 2025
Accepted: January 28, 2025
Published: January 31, 2025

Abstract

Predicting molecular properties is essential for advancing for advancing drug discovery and design. Recently, Graph Neural Networks (GNNs) have gained prominence due to their ability to capture the complex structural and relational information inherent in molecular graphs. Despite their effectiveness, the "black-box" nature of GNNs remains a significant obstacle to their widespread adoption in chemistry, as it hinders interpretability and trust. In this context, several explanation methods based on factual reasoning have emerged. These methods aim to interpret the predictions made by GNNs by analyzing the key features contributing to the prediction. However, these approaches fail to answer critical questions: "How to ensure that the structure-property mapping learned by GNNs is consistent with established domain knowledge". In this paper, we propose MMGCF, a novel counterfactual explanation framework designed specifically for the prediction of GNN-based molecular properties. MMGCF constructs a hierarchical tree structure on molecular motifs, enabling the systematic generation of counterfactuals through motif perturbations. This framework identifies causally significant motifs and elucidates their impact on model predictions, offering insights into the relationship between structural modifications and predicted properties. Our method demonstrates its effectiveness through comprehensive quantitative and qualitative evaluations of four real-world molecular datasets.

Keywords

Interpretability, Causal Relationship, Counterfactual Explanation, Molecular Graph Generation

1. Introduction

Deep learning models [1] [2] are highly effective in capturing nonlinear relationships between molecular structures and their functions, enabling diverse applications in chemistry, such as quantum property calculations [3] [4], chemical property predictions [5] [6], and drug screening [7] [8]. Although Graph Neural Networks (GNNs) [9] outperform traditional descriptor-based models in molecular property prediction [10], interpretability challenges remain unresolved. After predictions are generated, researchers often aim to understand why certain compounds are predicted to exhibit specific properties.

In our study, existing studies [11]-[13] on predictability of explainable properties are primarily explanations based on factual reasoning. These methods identify key features, such as atoms and bonds in the molecular graph, which influence the model's outcomes. The selected key features are typically considered factual interpretations that provide sufficient information to yield the same predictions as the original instance.

However, we argue that these methods lack sufficient persuasiveness. One reason is that these methods fail to adequately account for the unique aspects of molecular structure, making it difficult to incorporate critical and complete substructures into their explanation graphs. In other words, fact-based explanations do not address the question: "For a specific molecule, what would happen to the molecular property prediction model's performance if we slightly modified the molecular structure?"

A Motivating Example. Using **Figure 1** as an example, the advantages of counterfactual analysis in explaining molecular property predictions are demonstrated, in comparison to methods based on factual reasoning.

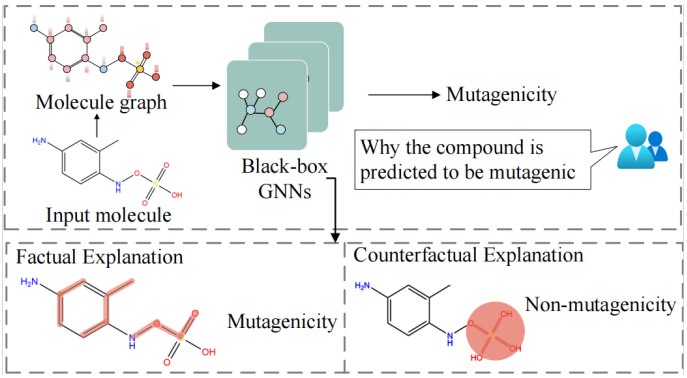

Figure 1. Illustration of factual explanations (bottom left) and counterfactual explanations (bottom right). The red highlights indicate the given explanations.

In a molecular property prediction task, a GNN models the molecular graph and classifies the input as mutagenic. Factual explanations identify specific edges in the molecular graph as key features contributing to mutagenicity. As shown in **Figure 1** (bottom left), factual explanations highlight the aromatic ring and other bonds in the molecular structure as critical contributors to the prediction. These

explainers focus on subgraphs that contain sufficient information to make the same predictions. However, the selected subgraphs may include redundant nodes or edges, making them neither compact nor complete.

In contrast, counterfactual explanations iteratively perturb chemically significant motifs, such as benzene rings and sulfonyl hydroxide (-SO3H), ultimately converging on the minimal modification required to alter the prediction outcome. As shown in **Figure 1** (bottom right), completely disrupting the sulfonyl hydroxide group, the molecule is rendered non-mutagenic, highlighting the critical role of this functional group. Such counterfactual explanations seek minimal perturbations to the input that result in a change in the prediction, thereby enabling differential inference before and after the perturbation to determine whether the structure-property mapping learned by GNNs aligns with established domain knowledge [14].

Our method and contribution. Recent advancements in counterfactual reasoning within artificial intelligence highlight the potential application of counterfactuals to GNN-based molecular property prediction. Most counterfactual methods rarely address chemical deep learning, as the complexity of molecular data poses significant challenges in generating counterfactual explanations.

To address these challenges, we propose MMGCF, a novel interpreter that integrates counterfactual reasoning with molecular motifs to enhance the interpretability of GNNs in predicting molecular properties. MMGCF aims to identify motifs that exhibit causal relationships with molecular properties, formulating the counterfactual perturbation problem as a differentiable motif generation task. Through motif perturbation, MMGCF establishes a framework for generating counterfactual explanations. Quantitative and qualitative analyses demonstrate that MMGCF provides reliable explanations for complex molecular data in classification and regression tasks, effectively capturing structural information compared to existing baselines.

In summary, the contributions of MMGCF are as follows:

- MMGCF constructs motif trees based on chemical rules, enabling in-depth exploration of molecular motifs while preserving chemical integrity.
- To the best of our knowledge, MMGCF is the first method in the field of chemistry to utilize motif-graph hierarchical information for counterfactual explanations.
- MMGCF generates explanations for the decisions made by GNN-based molecular property prediction models, which assist researchers in facilitating human interpretation through comparisons of input instances and corresponding counterfactual samples.
- MMGCF offers reliable explanations for complex molecular data in classification and regression tasks, effectively capturing structural information and surpassing existing baseline methods.

2. Related Work

2.1. Preliminaries of Graph Neural Networks

Deep graph networks leverage both the edge and node feature matrices to learn

graph representations. Various GNNs, such as GCN [15], GAT [16], GraphSAGE [17], and GIN [18], have gained significant attention for molecular property prediction.

In this paper, we employ Relational Graph Convolutional Networks (RGCN) [19], which extend GCN by incorporating edge features into the message-passing mechanism.

The propagation operation can be calculated as

$$h_v^{(l+1)} = update(h_v^{(l)}, h_{u\in N(v)}^{(l+1)}), \tag{1}$$

where is $h_v^{(l)}$ the output of the previous layer, and $h_{u\in N(v)}^{(l+1)}$ aggregate the information via

$$h^{(l+1)} = aggregate(W_r h_u^{(l)}, r \in R, u \in N_v^r, \tag{2}$$

Here, $h_v^{(l+1)}$ is the hidden vector of node v after $l+1$ iterations, N_v^r denotes the neighbors of node v under bond $r \in R$, where R denotes the set of bond types, and W is the weight for target node v.

2.2. Explainability in Graph Neural Networks

Existing methods [20]-[30] aim to decode the "black box" of GNNs by identifying important nodes or edges, such as text in tables, pixels in images, or nodes in graphs. These methods can be categorized into four types based on how importance scores are derived:

- **Gradient/Feature Perturbation-based Methods:** These methods [20] [21] are straightforward but have the significant drawback of being susceptible to gradient explosion.
- **Perturbation-based Methods:** These methods [22] [23] examine changes in predictions when input data is perturbed, allowing for the identification of important information.
- **Surrogate-based Methods:** These methods employ [12] [24] simple interpretable surrogate models to approximate the predictions of complex deep models in local regions of the input space.
- **Decomposition-based Methods:** These methods [25] analyze model parameters to reveal the relationships between features in the input space and the output predictions.

The aforementioned methods are instance-level explanation techniques.

Additionally, there are model-level explanation methods [26] [27] that generate novel graph structures tailored to specific properties. However, available generative model-based GNN interpreters have limited applicability, particularly in molecular graphs. As noted earlier, these methods rely on factual reasoning, where the important substructures they identify contain sufficient information to ensure that the output matches the original data. However, the identified substructures, while related to the output, may not necessarily represent the most compact structures, which constitutes a significant limitation of factual explanations.

2.3. Counterfactual Explanations

The concept of counterfactuals originates from mathematics and philosophy. It was first introduced by Kahneman and Miller in 1986 [28] and later defined by Woodward and Hitchcock [29] as an explanation of how differences in events or instances can lead to changes in outcomes. In other words, a counterfactual sample is one that is closest to the input but produces a different result. Counterfactual explanations are fundamentally grounded in the theory of causal manipulability [30], which focuses on manipulating outcomes through causal relationships. If a process is identified as a manipulation of an event, a causal relationship must exist.

Although both counterfactual and factual explanations can be used for predictive reasoning in GNNs, they fundamentally differ as they address distinct problems. CF-GNN-Explainer [31] introduces counterfactual reasoning to generate more compact yet crucial explanations for GNNs. CF^2 [32] integrates counterfactual and factual reasoning to extract sufficient and necessary GNN explanations. GEM [33] considers Granger causality to provide valid explanations. MEG [34] trains a reinforcement learning-based generator to produce counterfactual explanations.

Additionally, it is essential to distinguish between contrastive explanations, adversarial attacks [35], and counterfactual explanations. Adversarial attacks aim to deceive the model, while contrastive and counterfactual explanations primarily seek to explain predictions. Both adversarial attacks and counterfactual explanations involve small perturbations to the input. However, adversarial attack methods typically make minimal changes to the entire graph with the goal of degrading overall model performance. Contrastive explanations offer insights by comparing a specific prediction with other potential outcomes, helping users understand why the model made the current prediction instead of an alternative. In contrast, counterfactual explanations focus on how the model's predictions would change if the input data were altered.

3. Preliminary

Notations. Given a set of n molecular SMILES [36], we transform them using RDKit [37] into graphs $\mathcal{G} = \{G_1, G_2, \ldots, G_n\}$. Each molecular graph is represented as $G_i = (\boldsymbol{V}, \boldsymbol{E})$, consisting of a set of atoms $\boldsymbol{V}$ and a set of bonds $\boldsymbol{E}$. The atom feature matrix $A \in \mathbb{R}^{N \times K}$ represents the feature of each atom. Here, N denotes the total number of atoms, and K is the type of atoms. Additionally, an adjacency matrix $X \in \mathbb{R}^{c \times N \times N}$ captures the feature of bonds with $c \in N^+$ representing the bond types, $X(c,i,j) = X(c,j,i) = 1$ signifies a bond of type c between atom i and atom j. Thus, the structural information of each graph G_i is encoded in several matrices.

Motif tree. Our method decomposes molecules into multiple fragments, referred to as "motifs". These motifs are closely linked to the properties of the graph, which is shown to be of significant importance [38] [39]. The motifs are mapped into a motif tree by treating each motif as a node and the relative positional

relationships between motifs as edges.

Formally, for a given graph G, a motif tree $\mathcal{T}_G = (\mathcal{V}, \mathcal{E})$ with node set $\mathcal{V} = \{M_1, \ldots, M_k\}$ and edge set $\mathcal{E}$, where k is the number of motifs. Each motif $M_i = (V_i, E_i)$ is a subgraph of G, such that $\bigcup_i V_i = V$ and $\bigcup_i E_i \cup \mathcal{E} = E$. The motifs are non-overlapping, meaning $M_i \cap M_j = \varnothing \quad (i \neq j)$.

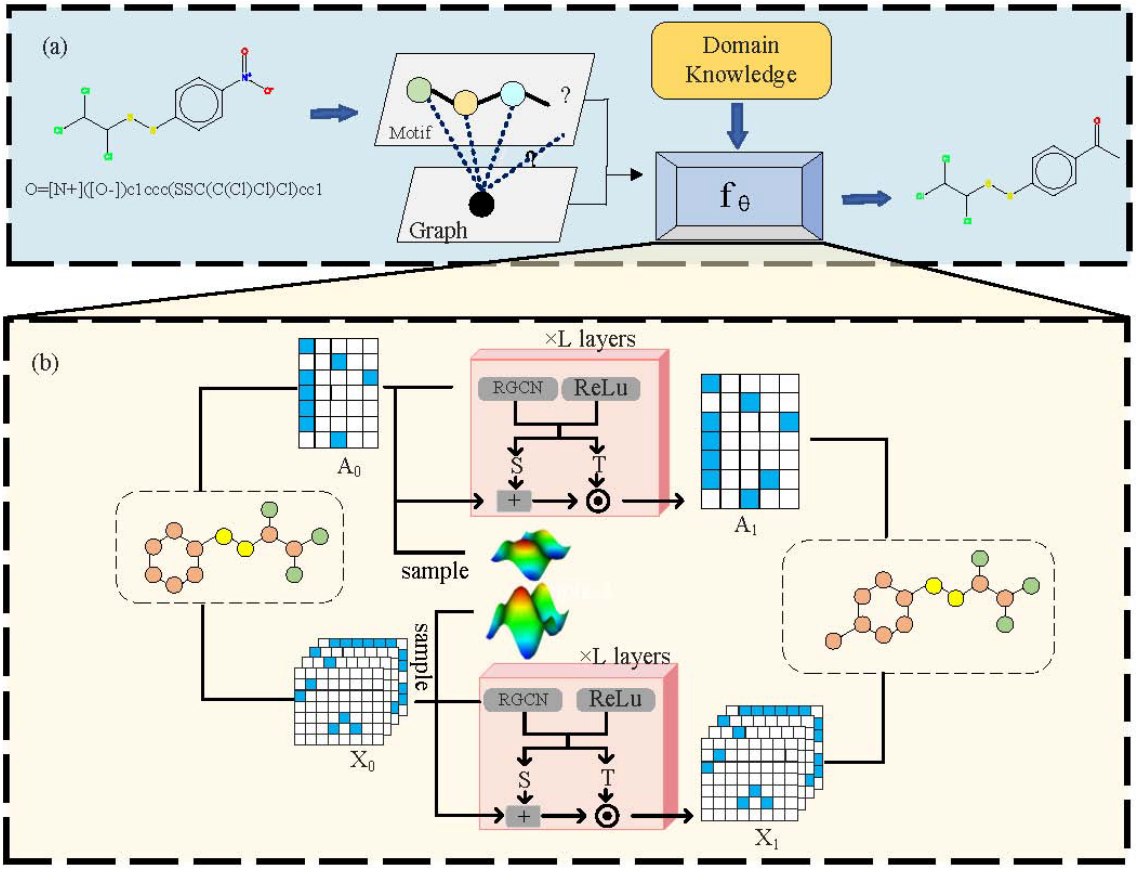

Figure 2. Overview of MMGCF. (a) A motif tree is constructed where each color represents a distinct motif. Motif generation is performed iteratively based on the motifs, graph structure, and domain knowledge. (b) The schematic diagram illustrates the generative procedure. The molecular graph G is defined by the atomic feature matrix A_i and the adjacency tensor X_i. RGCN with ReLU activation function is employed to obtain the mean and variance of the Gaussian distribution. By sampling from this Gaussian distribution and converting the results into discrete features, the domain knowledge evaluates the validity of the atoms and bonds, ultimately leading to the generation of new molecules. S and T represent the scale function and the transformation function, respectively.

4. Method: MMGCF

This section details MMGCF, which generates counterfactual explanations in molecular property prediction tasks. It consists of two key components: 1) chemically-guided motif tree construction. Each molecule is transformed into a motif tree. 2) Generation of counterfactual explanations. This component identifies explanations for predictions by perturbing specific motifs. We design a counterfactual generation framework and constraint functions to facilitate optimization. After optimization, MMGCF generates counterfactual explanations for molecular

predictions.

4.1. Chemically Guided Motif-Tree Construction

Given a molecule graph $G=(\boldsymbol{V},\boldsymbol{E})$, we convert it to a motif tree $\mathcal{T}_G=(\mathcal{V},\mathcal{E})$. Each node in the structure represents a topic, while the edges indicate the relative spatial relationships between motifs. The decision to convert the molecular structure from a graph to a tree structure is motivated by its advantages for the task of counterfactual data generation. A tree-like representation enables a clearer capture of hierarchical relationships and spatial distributions between topics, thereby providing robust support for subsequent analysis and processing.

We utilize the "Breaking of Interesting Chemical Substructures" (BRICS) algorithm [40] for bond cleavage, in which cleavage rules retain molecular fragments that possess significant structural and functional components. This method may yield overly large pieces due to its reliance on a limited set of reactions. To address this issue, we establish additional rules to ensure the generation of more effective molecular motifs.

Algorithm 1 outlines the construction of motif trees. V_1 and V_2 respectively contain non-ring bonds and rings of each fragment, both extracted using RDKit functions. If two bonds b_1 and b_2 share common atoms (*i.e.*, $b_1 \bigcap b_2 \neq \varnothing$) and the number of shared atoms exceeds three, these bonds are merged into a new motif. b_1 and b_2 are removed from V_1 and their intersection is added to V_1. Since the motif tree is not unique when atoms connect to multiple motifs, infinite weights are assigned to these edges in V_0 to ensure that the tree structure remains logical and free from cycles. Finally, we construct the motif tree as $\mathcal{T}_G=(\mathcal{V},\mathcal{E})$ using

Algorithm 1 Motif tree construction of molecule $G=(\mathbf{V},\mathbf{E})$

Creat fragment list $F=\{f_1,\ldots,f_n\}$ based on BRICS.
for f_i in F **do**:
  $V_1 \leftarrow$ the set of bonds $(u,v)\in E$ that do not belong to any rings.
  $V_2 \leftarrow$ the set of rings of G.
    for b_1,b_2 in V_1 **do**:
      if $b_1 \cap b_2 \neq \emptyset$ **and** $length(b_1 \cap b_2) > 3$ **then**
        $V_1 \leftarrow (b_1 \cap b_2); V_1 \setminus \{b_1,b_2\}$
      end if
    end for
    for r_1,r_2 in V_2 **do**
        if $length(r_1 \cap r_2) > 2$ **then** $V_2 \leftarrow (r_1 \cup r_2); V_2 \setminus \{r_1,r_2\}$
        end if
    end for
  end for
$V_0 \leftarrow$ Atoms being the intersection of three or more motifs in $V_1 \cup V_2$.
$\mathcal{V} \leftarrow V_0 \cup V_1 \cup V_2$.
$\mathcal{E} \leftarrow \{(m_i,m_j)\in \mathcal{V}\times\mathcal{V} \mid |m_i \cap m_j| > 0\}$.
for each $(m_i,m_j)\in\mathcal{E}$ **do**:
  if $m_i \in V_0$ **or** $m_j \in V_0$ **then** $W[i][j] \leftarrow \infty$
  else $W[i][j] \leftarrow 1$.
  end if
end for
Return The maximum spanning tree over molecule $\mathcal{T}_G=(\mathcal{V},\mathcal{E})$

the maximum spanning tree method, which maximizes connectivity or network capacity. Edges with infinite weights enforce the formation of a single connected spanning tree.

4.2. Generation of Counterfactual Explanations

A counterfactual generation framework is developed for GNN-based predictions of molecular properties. The central idea involves identifying minimal perturbations to the molecular graph that can reverse these predictions. This objective is achieved by addressing a counterfactual optimization challenge, which will be elaborated below.

The generative framework. For a given molecule graph G, it is first converted into a motif tree $\mathcal{T}_G = (\mathcal{V}, \mathcal{E})$, where the structural information of each motif $m \in \mathcal{V}$ is encoded as $(A^m, X^m) \subseteq (A, X)$. Since discrete data cannot be directly applied to continuous density models, a dequantization technique inspired by previous work [41] [42] is adopted. Specifically, noise from $U \in [0,1)$ is added to each atom A_i^m and each bond X_{ij}^m, mapping the discrete data (A^m, X^m) into continuous data (z^{A^m}, z^{X^m}).

$$z_i^{A^m} = A_i^m + u \quad u \sim U \in [0,1); z_{ij}^{A^m} = X_{ij}^m + u \quad u \sim U \in [0,1), \tag{3}$$

The continuous data (z^{A^m}, z^{X^m}) is then fed into the generative framework f_θ, which learns a bijection from the data space to a Gaussian distribution. A parameterized neural network is designed to guarantee that the learned data likelihood closely resembles a Gaussian distribution, comprising L layers of RGCN and ReLU activation functions. This network computes the mean μ_i^A, μ_{ij}^X and variance σ_i^A, σ_{ij}^X. Formally, the generated conditional distribution is defined as:

$$\begin{aligned} P(z^{A^m}) &= \mathcal{N}(\mu_i^A, (\sigma_i^A)^2), \quad \text{where } i \in \{1, \ldots, n\} \\ P(z^{X^m}) &= \mathcal{N}(\mu_{ij}^X, (\sigma_{ij}^X)^2), \quad \text{where } j \in \{1, \ldots, i-1\} \end{aligned}, \tag{4}$$

where n is the number of atoms in the given m.

Then, the data likelihood is modeled by using the change of variable formula [43]:

$$\begin{aligned} p(z_i^{A^m} \mid z^{A^m}) &= P(z^{A^m}) \left| det\left(\frac{\partial f_\theta(Z^{A^m}, A^m)}{\partial z^{A^m}}\right) \right| \\ p(z_{ij}^{X^m} \mid z^{X^m}) &= P(z^{X^m}) \left| det\left(\frac{\partial f_\theta(Z^{X^m}, X^m)}{\partial z^{X^m}}\right) \right|, \end{aligned} \tag{5}$$

The probability of the given $m = (A^m, X^m)$ can be calculated as:

$$p_m(m) = p_m(A^m, X^m) \approx p_m(A^m) p_m(X^m) = p(z_i^{A^m} \mid z^{A^m}) p(z_{ij}^{X^m} \mid z^{X^m}), \tag{6}$$

So we get the negative log-likelihood of f_θ:

$$\mathcal{L}_f = -(\log p(z_i^{A^m} \mid z^{A^m}) + \log p(z_{ij}^{X^m} \mid z^{X^m})), \tag{7}$$

To generate a new motif, random variables ζ^A and ζ^X from the Gaussian distribution and convert into discrete features:

$$A^m = \zeta^{A^m} \odot \sigma^{A^m} + \mu^{A^m}$$
$$X^m = \zeta^{X^m} \odot \sigma^{X^m} + \mu^{X^m}, \tag{8}$$

To generate new motif trees from the root node, DFS (Depth-First Search) or BFS (Breadth-First Search) approaches are utilized, with the motif ordering saved as τ.

The final step of our model is to reconstruct the molecular graph G from the motif tree. Our goal is to assemble the motifs into the correct molecular structure. Given the ordering τ, $\mathcal{L}_{recon}$ is used as the reconstruction loss in the generated molecule. It is defined as:

$$\mathcal{L}_{recon} = -(\log p(G) + \log p(\mathcal{V}^\tau, \mathcal{E}^\tau)), \tag{9}$$

Here, $\mathcal{V}^\tau$ represents the motifs in the specified arrangement and $\mathcal{E}^\tau$ denotes the concatenated edges between the motifs.

Counterfactual constraint. Leveraging the generation framework, any molecule is reconstructed based on perturbed motifs. MMGCF is employed to generate counterfactual explanations for molecular prediction tasks.

A counterfactual G^{cf} is specific to a given molecule G. For the molecule G, there is a predicted output $\Phi(G)$.

The counterfactual serves as an explanation for G, defined as the solution to the following constrained optimization problem:

$$C(G) = \arg\max \frac{1}{k}\sum_{i=1}^{k} 1\left(\Phi(G^{\mathrm{cf}}) \neq \Phi(G)\right) + \frac{\lambda}{k}\sum_{i=1}^{k} \mathrm{sim}(G^{\mathrm{cf}}, G), \tag{10}$$

where k is the total number of generated counterfactual samples, $sim(\cdot)$ is calculated as the similarity of G and G^{cf}, $1(\cdot)$ is an indicator function that outputs 1 when the input condition is true, otherwise, it outputs 0.

Equation (10) is defined for classification tasks. However, it must be apt for regression tasks. Instead of seeking transformations in the labels, we look for counterfactuals that lead to increases or decreases in predictions. In this context, a problem-specific hyperparameter Δ represents the value change.

$$C(G) = \arg\max \frac{1}{k}\sum_{i=1}^{k} 1\left(\Phi(G^{\mathrm{cf}}) - \Phi(G) \geq \Delta\right) + \frac{\lambda}{k}\sum_{i=1}^{k} \mathrm{sim}(G^{\mathrm{cf}}, G), \tag{11}$$

After being constrained, Equation (12) represents our final loss function.

$$L = \mathcal{L}_{recon} + \mathcal{L}_f + C(G), \tag{12}$$

5. Experimental and Result

In this section, we first introduce the datasets and the comparison baselines. Then, we report the main experimental results and the analyses.

5.1. Datasets and Baselines

Datasets. To assess the effectiveness of MMGCF in explaining predictions made by GNN models, we conducted experiments on four molecular datasets: concerning aqueous solubility (ESOL), mutagenicity, HIV activity, and blood-brain

barrier permeability (BBBP).

Table 1. The performance of the consensus model and statistics of four molecule datasets. "#ave n" and "#ave e" represent the average number of nodes and edges per graph, respectively. "#graph" refers to the total number of graphs in the dataset.

Datasets	#ave n	#ave e	#graph	task	metric	performance
Mutagenicity	17.20	18.35	7672	classification	ROC-AUC	0.902
BBBP	24.42	26.46	1859	classification	ROC-AUC	0.901
HIV	26.89	28.98	4943	classification	ROC-AUC	0.804
ESOL	13.87	14.43	1128	regression	R^2	0.884

Table 1 provides statistics for all the datasets used and shows the performance of RGCN as a consensus model in different data test sets and tasks.

- Mutagenicity [44] [45] dataset includes 7672 compounds, each categorized as either mutagenic or non-mutagenic.
- BBBP [46] dataset Comprising 1859 compounds, this dataset is used to predict whether a molecule can penetrate the blood-brain barrier, forming a binary classification task.
- HIV [47] introduced by the Drug Therapeutics Program (DTP) AIDS Antiviral Screen, contains 41,122 compounds. Each compound is labeled as either active or inactive regarding HIV antiviral activity. A sample of 4943 molecules is selected from this dataset for our experiment."
- ESOL (aqueous solubility) [48] contains 1128 compounds and is commonly used for graph regression tasks. Aqueous solubility is one of the criteria for evaluating the absorption capacity of drug candidates.

Baselines. The following baselines were used for comparison:

1) Random: At each step, at most one motif is reconstructed for each graph being explained; 2) GNN-Explainer: GNN-Explainer [22] generates an edge mask to estimate the contribution of different edges to the model's prediction. For counterfactual generation, the removal of subgraphs identified as important is performed; 3) CF^2: CF^2 [32] integrates counterfactual and factual reasoning to produce an edge mask that estimates edge importance; 4) MEG: MEG [34] explicitly incorporates domain-specific knowledge from chemistry and generates counterfactuals using reinforcement learning.

5.2. Evaluation Metrics

Validity [49]. The metric reflects the proportion of effective counterfactual explanations, where an explanation is deemed effective if it leads to a significant change in the sample's prediction outcome.

$$\text{validity} = \frac{\#counterfactual}{\#molecules}, \tag{13}$$

The indicator serves as a critical measure of the practical applicability of our

method. For classification tasks, the proportion of counterfactuals reflecting the desired labels is evaluated. For regression tasks, the proportion of counterfactuals leading to an increase or decrease in the predicted values is assessed.

Fidelity [50]. The metric measures the faithfulness of explanations to the oracle by evaluating how well the counterfactual differentiates from the original graph with respect to the ground truth label.

$$fidelity = \frac{1}{N}\sum_{i\in|N|}\chi(\Phi(G_i, G_i^{\mathrm{cf}})) - \mathbf{1}(\Phi(G_i^{\mathrm{cf}}) = y_{G_i}), \tag{14}$$

Here, for classification tasks, $\chi(\Phi(G_i, G_i^{\mathrm{cf}})) = \Phi(G_i) - \Phi(G_i^{\mathrm{cf}})$, for regression tasks, $\chi(\Phi(G_i, G_i^{\mathrm{cf}})) = |\Phi(G_i) - \Phi(G_i^{\mathrm{cf}})|$, and y_G is the ground truth label.

Time [50]: The average time cost for generating counterfactual explanations for each instance (in seconds).

5.3. Quantitative Analysis

The results of three metrics across various models for different tasks are presented in **Table 2**. 1) Validity. The proposed method consistently outperforms various models by generating chemical data interpretations that emphasize effective substructures rather than individual atomic nodes or edges. Baseline methods fail to effectively capture structure-property relationships due to the complexity of chemical data. MEG primarily assesses whether perturbations, such as the addition or removal of atoms, are validated by valence electrons, whereas GNN-Explainer and CF^2 primarily generate edge masks. 2) Time. The time cost of MMGCF is relatively low compared to alternative methods. In contrast, MEG requires enumerating each perturbation operation, which significantly increases

Table 2. The performance (mean of ten repeated executions) of different models. The best results are highlighted in bold, while the runner-up results are underlined.

Metric	Model	Mutagenicity	BBBP	HIV	ESOL
Validity %	Random	63.99	54.61	55.27	95.91
	GNN-Explainer	41.62	66.15	65.27	96.45
	CF^2	37.91	96.61	83.33	95.80
	MEG	46.08	53.84	47.22	88.17
	MMGCF (DFS)	**100**	**99.57**	**99.70**	**100**
	MMGCF (BFS)	<u>100</u>	<u>99.56</u>	<u>99.41</u>	<u>100</u>
Time (s)	Random	**16.78**	28.69	31.47	33.15
	GNN-Explainer	49.32	**3.83**	**20.54**	**1.48**
	CF^2	83.38	18.93	62.05	18.78
	MEG	1403.83	1161.30	2047.90	427.56
	MMGCF (DFS)	<u>26.53</u>	<u>26.47</u>	<u>31.10</u>	<u>17.99</u>
	MMGCF (BFS)	27.49	26.64	33.62	18.13

its computational cost. Although GNN-Explainer incurs low time costs, existing methods often fail to adequately account for the unique characteristics of molecular structures. This limitation makes it difficult for these methods to include critical substructures in their explanation graphs.

Table 3 presents a performance comparison of various methods on the Fidelity metric across four real-world datasets. Specifically, MEG exhibits relatively poor performance on all datasets, as its counterfactuals rely on atomic-node-level perturbations, which fail to effectively guide the counterfactual samples in altering the model's predictions from the true labels. Although CF^2 and GNN-Explainer demonstrate strong performance on certain datasets, their overall results are inferior to those of the method proposed in this study. The superior performance of our method is attributed to its unique strategy of generating explanations based on individual motifs. Compared to other models, our method achieves an average improvement of approximately 45.92%, underscoring its significant performance advantages.

This improvement underscores the importance of attributing each motif to the target property. Domain experts are particularly interested in understanding how case-based differential reasoning supports counterfactual predictions for specific molecules. In such cases, the model being explained may interpret similar structures as fundamentally different with respect to the predicted property. Counterfactual molecules generated through this approach enable experts to evaluate whether the structure-to-function mapping learned by the model aligns with established domain knowledge, at least within the immediate neighborhood of the molecules under investigation.

Table 3. Fidelity on four real datasets.

Metric	Model	Mutagenicity	BBBP	HIV	ESOL
Fidelity %	Random	63.19	53.16	54.23	96.95
	GNN-Explainer	35.00	86.15	77.77	95.15
	CF^2	26.03	96.15	88.87	93.48
	MEG	43.86	53.36	46.30	87.36
	MMGCF(DFS)	99.45	98.23	100	100
	MMGCF(BFS)	99.47	98.33	100	100

5.4. Qualitative Analysis

Figure 3 provides a qualitative evaluation of four methods applied to four datasets, with actual molecules from the datasets included for reference. The results demonstrate that MMGCF produces more meaningful motif-based explanations than the other methods. Moreover, it is shown that GNN-explainer, CF^2, and MEG lack this ability.

MEG is relatively ambiguous in identifying true causal substructures. This ambiguity stems from its process of generating explanations by adding or

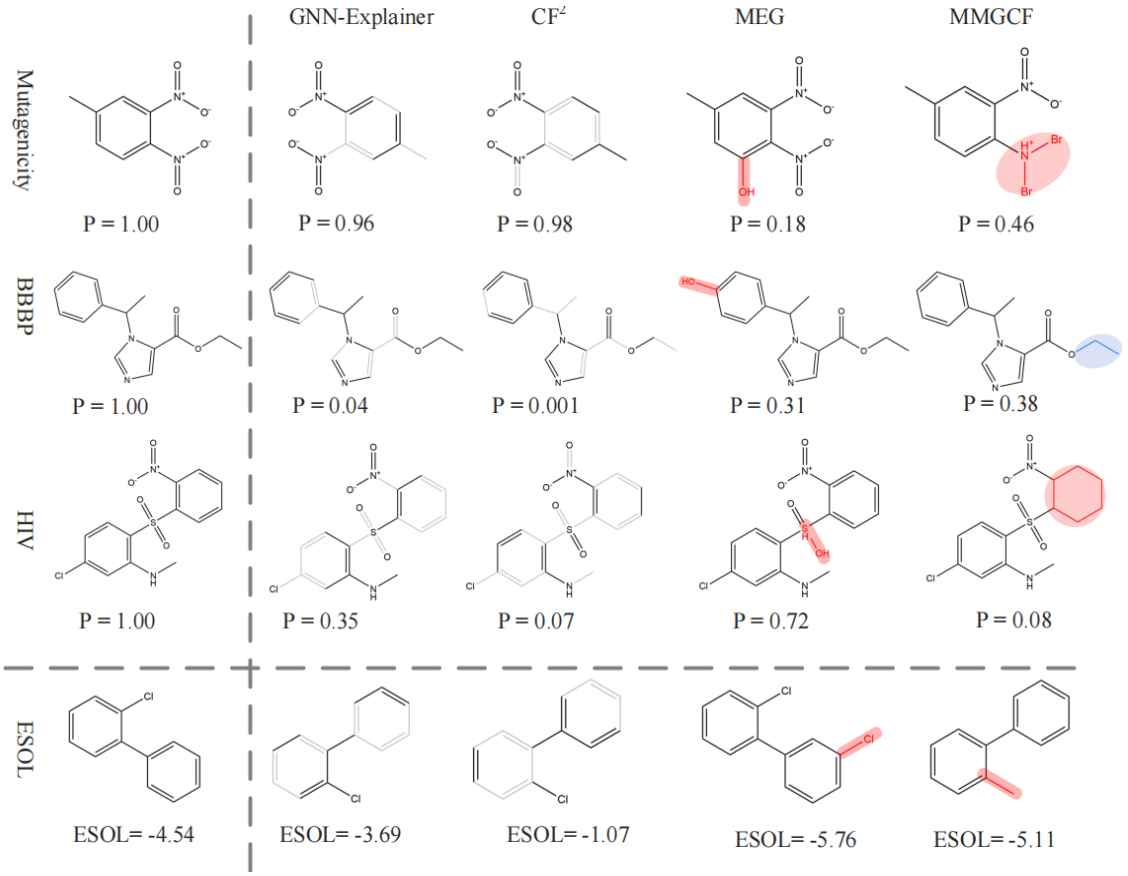

Figure 3. Qualitative Analysis. The leftmost column displays the instances, where the blurred edges represent important features identified by GNN-Explainer or CF2. For MEG and MMGCF, red highlights indicate added or modified structures, while blue highlights denote deletions. The probability shown beneath each graph or subgraph represents the likelihood of classification into the "Mutagenic," "BBBP," or "HIV activity" class, respectively. ESOL evaluates the absorption capacity of drug candidates.

removing nodes or edges, without assessing whether the perturbation is interpretable, whether the molecule loses its original properties due to structural disruption, or if other underlying factors are involved. As a result, the explanation lacks clarity and precision.

When focusing on the GNN-Explainer and CF^2, it is consistently observed across all four datasets, for both classification and regression tasks, that these models fail to capture complete and compact explanatory subgraphs. This limitation stems from the fact that GNN-Explainer and CF^2 provide sample-level explanations by masking edges of lower importance within the graph. The GNN-Explainer, CF^2, and MEG methods are ineffective at identifying complex substructures, leading to less effective explanations. In contrast, our method generates more meaningful, accurate, and structurally coherent explanations.

6. Conclusion

In this work, we proposed MMGCF, a novel counterfactual explanation framework tailored for GNN-based molecular property prediction. By leveraging

chemically guided motif tree construction and counterfactual reasoning, MMGCF bridges the gap between machine learning predictions and domain-specific chemical insights. Unlike existing approaches, which often lack chemical relevance or generalizability, MMGCF introduces a hierarchical motif-based framework that ensures interpretability while preserving molecular structural integrity. Through extensive experiments on four benchmark datasets, MMGCF demonstrated superior performance in generating reliable, causally meaningful explanations compared to state-of-the-art methods. It was particularly effective in identifying critical molecular motifs, providing chemists with actionable insights into the relationship between structural modifications and property variations. Future research can explore extending MMGCF to other graph-based domains, incorporating more sophisticated diversity constraints.

Acknowledgements

This work is supported by the State Key Program of the National Nature Science Foundation of China (61936001), the key cooperation project of Chongqing Municipal Education Commission (HZ2021008), the Natural Science Foundation of Chongqing (cstc2019jcyj-cxttX0002, cstc2021ycjhbgzxm0013).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Gao, H., Wang, Z. and Ji, S. (2018) Large-Scale Learnable Graph Convolutional Networks. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, London, 19-23 August 2018, 1416-1424. https://doi.org/10.1145/3219819.3219947

[2] Liu, Q., Tan, H.S., Zhang, Y.M. and Wang, G.Y. (2022) Dynamic Heterogeneous Network Representation Method Based on Meta-Path. *Acta Electronica Sinica*, **50**, 1830-1839. https://doi.org/10.12263/DZXB.20211288

[3] Wasielewski, M.R., Forbes, M.D.E., Frank, N.L., Kowalski, K., Scholes, G.D., Yuen-Zhou, J., *et al.* (2020) Exploiting Chemistry and Molecular Systems for Quantum Information Science. *Nature Reviews Chemistry*, **4**, 490-504. https://doi.org/10.1038/s41570-020-0200-5

[4] Wu, W., Zhu, J., Yao, Y. and Lan, Y. (2024) Can Molecular Quantum Computing Bridge Quantum Biology and Cognitive Science? *Intelligent Computing*, **3**, Article ID: 0072. https://doi.org/10.34133/icomputing.0072

[5] Fang, X., Liu, L., Lei, J., He, D., Zhang, S., Zhou, J., *et al.* (2022) Geometry-enhanced Molecular Representation Learning for Property Prediction. *Nature Machine Intelligence*, **4**, 127-134. https://doi.org/10.1038/s42256-021-00438-4

[6] Gong, X., Liu, M., Sun, H., Li, M. and Liu, Q. (2022) HS-DTI: Drug-Target Interaction Prediction Based on Hierarchical Networks and Multi-Order Sequence Effect. 2022 *IEEE International Conference on Bioinformatics and Biomedicine* (*BIBM*), Las Vegas, 6-8 December 2022, 322-327. https://doi.org/10.1109/bibm55620.2022.9994908

[7] Bongini, P., Bianchini, M. and Scarselli, F. (2021) Molecular Generative Graph

Neural Networks for Drug Discovery. *Neurocomputing*, **450**, 242-252. https://doi.org/10.1016/j.neucom.2021.04.039

[8] Jiménez-Luna, J., Grisoni, F., Weskamp, N. and Schneider, G. (2021) Artificial Intelligence in Drug Discovery: Recent Advances and Future Perspectives. *Expert Opinion on Drug Discovery*, **16**, 949-959. https://doi.org/10.1080/17460441.2021.1909567

[9] Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C. and Yu, P.S. (2021) A Comprehensive Survey on Graph Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, **32**, 4-24. https://doi.org/10.1109/tnnls.2020.2978386

[10] Rao, J., Zheng, S., Lu, Y. and Yang, Y. (2022) Quantitative Evaluation of Explainable Graph Neural Networks for Molecular Property Prediction. *Patterns*, **3**, Article ID: 100628. https://doi.org/10.1016/j.patter.2022.100628

[11] Luo, D., Cheng, W., Xu, D., Yu, W., Zong, B., *et al.* (2020) Parameterized Explainer for Graph Neural Network. *Advances in Neural Information Processing Systems*, **33**, 19620-19631.

[12] Huang, Q., Yamada, M., Tian, Y., Singh, D. and Chang, Y. (2023) GraphLIME: Local Interpretable Model Explanations for Graph Neural Networks. *IEEE Transactions on Knowledge and Data Engineering*, **35**, 6968-6972. https://doi.org/10.1109/tkde.2022.3187455

[13] Yuan, H., Yu, H., Wang, J., Li, K. and Ji, S. (2021) On Explainability of Graph Neural Networks via Subgraph Explorations. *ReScience*, **9**, Article 41.

[14] Roese, N.J. (1997) Counterfactual Thinking. *Psychological Bulletin*, **121**, 133-148. https://doi.org/10.1037/0033-2909.121.1.133

[15] Kipf, T.N. and Welling, M. (2016) Semi-Supervised Classification with Graph Convolutional Networks. arXiv: 1609.02907.

[16] Velickovic, P., Cucurull, G., Casanova, A., *et al.* (2017) Graph Attention Networks. https://arxiv.org/abs/1710.10903

[17] Hamilton, W., Ying, Z. and Leskovec, J. (2017) Inductive Representation Learning on Large Graphs. 31*st Conference on Neural Information Processing Systems* (*NIPS* 2017), Long Beach, 4-7 December 2017, 30 p.

[18] Xu, K., Hu, W., Leskovec, J. and Jegelka, S. (2018) How Powerful Are Graph Neural Networks? arXiv: 1810.00826.

[19] Schlichtkrull, M., Kipf, T.N., Bloem, P., van den Berg, R., Titov, I. and Welling, M. (2018) Modeling Relational Data with Graph Convolutional Networks. In: Gangemi, A., *et al.*, Eds., The Semantic Web, Springer, 593-607. https://doi.org/10.1007/978-3-319-93417-4_38

[20] Baldassarre, F. and Azizpour, H. (2019) Explainability Techniques for Graph Convolutional Networks. arXiv: 1905.13686.

[21] Pope, P.E., Kolouri, S., Rostami, M., Martin, C.E. and Hoffmann, H. (2019) Explainability Methods for Graph Convolutional Neural Networks. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 10764-10773. https://doi.org/10.1109/cvpr.2019.01103

[22] Ying, Z., Bourgeois, D., You, J., Zitnik, M. and Leskovec, J. (2019) GNNExplainer: Generating Explanations for Graph Neural Networks. 33*rd Conference on Neural Information Processing Systems* (*NeurIPS* 2019), Vancouver, 8-14 December 2019, 32 p.

[23] Wang, X., Wu, Y., Zhang, A., *et al.* (2021) Towards Multi-Grained Explainability for Graph Neural Networks. *Advances in Neural Information Processing Systems*, **34**, 18446-18458.

[24] Duval, A. and Malliaros, F.D. (2021) Graphsvx: Shapley Value Explanations for Graph Neural Networks. In: Oliver, N., Pérez-Cruz, F., Kramer, S., Read, J. and Lozano, J.A., Eds., *Machine Learning and Knowledge Discovery in Databases*, Springer, 302-318. https://doi.org/10.1007/978-3-030-86520-7_19

[25] Schnake, T., Eberle, O., Lederer, J., Nakajima, S., Schutt, K.T., Muller, K., *et al.* (2022) Higher-order Explanations of Graph Neural Networks via Relevant Walks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **44**, 7581-7596. https://doi.org/10.1109/tpami.2021.3115452

[26] Yuan, H., Tang, J., Hu, X. and Ji, S. (2020) XGNN: Towards Model-Level Explanations of Graph Neural Networks. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 6-10 July 2020, 430-438. https://doi.org/10.1145/3394486.3403085

[27] Wang, X. and Shen, H.W. (2022) Gnninterpreter: A Probabilistic Generative Model-Level Explanation for Graph Neural Networks. arXiv: 2209.07924.

[28] Kahneman, D. and Miller, D.T. (1986) Norm Theory: Comparing Reality to Its Alternatives. *Psychological Review*, **93**, 136-153. https://doi.org/10.1037//0033-295x.93.2.136

[29] Woodward, J. and Hitchcock, C. (2003) Explanatory Generalizations, Part I: A Counterfactual Account. *Noûs*, **37**, 1-24. https://doi.org/10.1111/1468-0068.00426

[30] Pearl, J. (2009) Causality. 2nd Edition, Cambridge University Press. https://doi.org/10.1017/cbo9780511803161

[31] Lucic, A., Ter Hoeve, M.A., Tolomei, G., *et al.* (2022) CF-GNNExplainer: Counterfactual Explanations for Graph Neural Networks. *International Conference on Artificial Intelligence and Statistics*, Valencia, 28-30 March 2022, 4499-4511

[32] Tan, J., Geng, S., Fu, Z., Ge, Y., Xu, S., Li, Y., *et al.* (2022) Learning and Evaluating Graph Neural Network Explanations Based on Counterfactual and Factual Reasoning. *Proceedings of the ACM Web Conference* 2022, Lyon, 25-29 April 2022, 1018-1027. https://doi.org/10.1145/3485447.3511948

[33] Lin, W., Lan, H. and Li, B. (2021) Generative Causal Explanations for Graph Neural Networks. *International Conference on Machine Learning*, 18-24 July 2021, 6666-6679.

[34] Numeroso, D. and Bacciu, D. (2021) MEG: Generating Molecular Counterfactual Explanations for Deep Graph Networks. 2021 *International Joint Conference on Neural Networks* (*IJCNN*), Shenzhen, 18-22 July 2021, 1-8. https://doi.org/10.1109/ijcnn52387.2021.9534266

[35] Sun, L., Dou, Y., Yang, C., Zhang, K., Wang, J., Yu, P.S., *et al.* (2022) Adversarial Attack and Defense on Graph Data: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, **35**, 7693-7711. https://doi.org/10.1109/tkde.2022.3201243

[36] Weininger, D. (1988) SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *Journal of Chemical Information and Computer Sciences*, **28**, 31-36. https://doi.org/10.1021/ci00057a005

[37] Bento, A.P., Hersey, A., Félix, E., Landrum, G., Gaulton, A., Atkinson, F., *et al.* (2020) An Open Source Chemical Structure Curation Pipeline Using RDKit. *Journal of Cheminformatics*, **12**, Article No. 51. https://doi.org/10.1186/s13321-020-00456-1

[38] Jin, W., Barzilay, R. and Jaakkola, T. (2018) Junction Tree Variational Autoencoder for Molecular Graph Generation. *International Conference on Machine Learning*, Stockholm, 10-15 July 2018, 2323-2332.

[39] Zhang, Z., Liu, Q., *et al.* (2021) Motif-Based Graph Self-Supervised Learning for

Molecular Property Prediction. *Advances in Neural Information Processing Systems*, **34**, 15870-15882.

[40] Degen, J., Wegscheid-Gerlach, C., Zaliani, A. and Rarey, M. (2008) On the Art of Compiling and Using 'Drug-Like' Chemical Fragment Spaces. *ChemMedChem*, **3**, 1503-1507. https://doi.org/10.1002/cmdc.200800178

[41] Ho, J., Chen, X., Srinivas, A., Duan, Y. and Abbeel, P. (2019) Flow++: Improving Flow-Based Generative Models with Variational Dequantization and Architecture Design. *International Conference on Machine Learning*, Long Beach, 10-15 June 2019, 2722-2730.

[42] Shi, C., Xu, M., Zhu, Z., *et al.* (2020) Graphaf: A Flow-Based Autoregressive Model for Molecular Graph Generation. arXiv: 2001.09382.

[43] Zhu, Y., Ouyang, Z., Liao, B., Wu, J., Wu, Y., Hsieh, C., *et al.* (2023) MolHF: A Hierarchical Normalizing Flow for Molecular Graph Generation. *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, Macao, 19-25 August 2023, 5002-5010. https://doi.org/10.24963/ijcai.2023/556

[44] Kazius, J., McGuire, R. and Bursi, R. (2004) Derivation and Validation of Toxicophores for Mutagenicity Prediction. *Journal of Medicinal Chemistry*, **48**, 312-320. https://doi.org/10.1021/jm040835a

[45] Wu, Z., Wang, J., Du, H., Jiang, D., Kang, Y., Li, D., *et al.* (2023) Chemistry-intuitive Explanation of Graph Neural Networks for Molecular Property Prediction with Substructure Masking. *Nature Communications*, **14**, Article No. 2585. https://doi.org/10.1038/s41467-023-38192-3

[46] Sakiyama, H., Fukuda, M. and Okuno, T. (2021) Prediction of Blood-Brain Barrier Penetration (BBBP) Based on Molecular Descriptors of the Free-Form and In-Blood-Form Datasets. *Molecules*, **26**, Article 7428. https://doi.org/10.3390/molecules26247428

[47] Wu, Z., Ramsundar, B., Feinberg, E.N., Gomes, J., Geniesse, C., Pappu, A.S., *et al.* (2018) MoleculeNet: A Benchmark for Molecular Machine Learning. *Chemical Science*, **9**, 513-530. https://doi.org/10.1039/c7sc02664a

[48] Ramsundar, B., Eastman, P., *et al.* (2019) Deep Learning for the Life Sciences: Applying Deep Learning to Genomics, Microscopy, Drug Discovery, and More. O'Reilly Media, Inc.

[49] Ma, J., Guo, R., Mishra, S., Zhang, A. and Li, J. (2022) Clear: Generative Counterfactual Explanations on Graphs. *Advances in Neural Information Processing Systems*, **35**, 25895-25907.

[50] Prado-Romero, M.A., Prenkaj, B., Stilo, G. and Giannotti, F. (2024) A Survey on Graph Counterfactual Explanations: Definitions, Methods, Evaluation, and Research Challenges. *ACM Computing Surveys*, **56**, 1-37. https://doi.org/10.1145/3618105