Text-guided Molecule Generation with Conditional Discrete Graph Diffusion Model

Yang Yao
Tsinghua University
Beijing, China
yaoyang21@mails.tsinghua.edu.cn

Xin Wang*
Tsinghua University
Beijing, China
xin_wang@tsinghua.edu.cn

Yaofei Wu
Beijing University of Technology
Beijing, China
23027313@emails.bjut.edu.cn

Zeyang Zhang
Tsinghua University
Beijing, China
zy-zhang20@mails.tsinghua.edu.cn

Daixin Wang
Ant Group
Beijing, China
daixin.wdx@antgroup.com

Zhiqiang Zhang
Ant Group
Beijing, China
lingyao.zzq@antgroup.com

Hong Mei
Peking University
Beijing, China
meih@pku.edu.cn

Wenwu Zhu*
Tsinghua University
Beijing, China
wwzhu@tsinghua.edu.cn

Abstract

Text-guided molecule generation enables controlled molecular design from natural language descriptions and has broad applications in areas such as drug discovery. While recent methods have demonstrated promising capability in generating molecules that align well with textual descriptions, they often overlook the structural properties of the generated graphs. As a result, these approaches struggle to simultaneously ensure consistency with the input text and high structural quality of the generated molecules. In this paper, we propose a text-guided molecular graph generation framework that leverages the structural modeling power of graph diffusion models to achieve both strong alignment with textual descriptions and high-quality molecular structures. However, accomplishing this goal involves several key challenges: 1) how to align graph diffusion models with natural language instructions in order to generate molecular graphs with expected relational semantics from text, 2) how to directly optimize the quality of the generated molecular graphs without sacrificing fine-grained alignment with text-specific details. To tackle these challenges, we introduce Text-guided Conditional Discrete Graph Diffusion (TDGD), a discrete diffusion-based framework for generating molecular graphs from natural language descriptions. Our model incorporates a structure-aware cross-attention mechanism that aligns textual semantics with molecular structures by capturing relational semantics between textual descriptions and molecular structures. In addition, we propose a molecule structure consistency loss that explicitly enforces structural coherence during generation, leading to higher-quality and more consistent molecular graphs. Extensive experiments on ChEBI-20 and L+M-24 datasets demonstrate the effectiveness of our proposed **TDGD** model.

CCS Concepts

• **Computing methodologies** → **Artificial intelligence**.

Keywords

graph generation, discrete graph diffusion model

ACM Reference Format:
Yang Yao, Xin Wang, Yaofei Wu, Zeyang Zhang, Daixin Wang, Zhiqiang Zhang, Hong Mei, and Wenwu Zhu. 2026. Text-guided Molecule Generation with Conditional Discrete Graph Diffusion Model. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '26), August 09–13, 2026, Jeju Island, Republic of Korea.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3770855.3818981

1 Introduction

Molecule generation plays a crucial role in many scientific and industrial domains, particularly in drug discovery [10, 30], where designing molecules with desired properties is a fundamental yet challenging task. Traditional molecule generation approaches heavily rely on domain expertise and experimental conditions. However, the vast combinatorial space of potential molecule structures far exceeds human capabilities for exhaustive exploration. With the rapid advancements in computer science, researchers have increasingly leveraged machine learning techniques, such as autoregressive models [5, 14], variational autoencoders (VAEs) [4, 19], and diffusion models [13, 27], to facilitate molecule generation, achieving remarkable success. More recently, there has been growing interest in generating molecules that satisfy specific constraints, particularly those aligned with textual descriptions. This capability is highly valuable across multiple fields, as it enables the precise design of molecule structures with desired properties by specifying their characteristics in natural language.

*Corresponding author. Xin Wang and Wenwu Zhu are affiliated with Department of Computer Science and Technology, Beijing National Research Center for Information Science and Technology, Tsinghua University.

KDD '26, Jeju Island, Republic of Korea

ACM ISBN 979-8-4007-2259-2/2026/08
https://doi.org/10.1145/3770855.3818981

Molecules inherently exhibit non-Euclidean structural properties and can be naturally represented as graphs, where atoms and chemical bonds form complex topological relationships. Such graph-structured representations are essential for faithfully modeling molecular structures and have motivated the widespread adoption of graph neural networks (GNNs) in molecular tasks [1, 16, 17, 24, 32]. Recent studies [5, 22, 29, 34] have shown promising progress in generating molecules conditioned on textual descriptions. However, these methods primarily focus on semantic alignment between text and generated molecules, while largely neglecting the intrinsic structural properties of molecular graphs. As a result, they often struggle to generate molecules that are both well aligned with textual descriptions and structurally valid, especially when the input text involves complex or fine-grained constraints.

To address this limitation, we propose a text-guided molecular graph generation framework that leverages the structural modeling power of graph diffusion models to achieve both strong alignment with textual descriptions and high-quality molecular structures. However, accomplishing this goal involves several key challenges:

- How to align graph diffusion models to follow text descriptions to generate molecular graphs with expected *relational* semantics, where nodes are inter-dependent with edges in graphs.
- How to directly optimize the quality of generated molecular graphs without sacrificing fine-grained alignment with text-specific details, so as to accurately learn molecular graph distributions conditioned on textual descriptions.

To address these challenges, we propose the Text-guided Conditional Discrete Graph Diffusion model for Molecule Generation (**TDGD**), which is able to generate graphs conditioned on text descriptions. Specifically, we design a conditional discrete graph diffusion model for molecule generation. We propose a *structure-aware cross-attention mechanism* to guide the diffusion model to follow the text description by inherently capturing the relational semantics among texts and molecular structures. This mechanism effectively captures relational semantics between text and molecular structures, enabling structure-aware text-guided generation. Additionally, we propose a *molecule structure consistency loss* as a complement to the standard training objective of diffusion models, improving the consistency of generated molecular structures with the text descriptions. Extensive experiments on molecular datasets demonstrate the effectiveness of our proposed method.

The contributions of this paper are summarized as follows:

- We propose a text-guided conditional discrete graph diffusion framework for molecule generation, which directly models molecular graph structures during the diffusion process to improve structure-aware text-guided generation.
- We design the structure-aware cross-attention mechanism to capture relational semantics among texts and structures.
- We design the molecule structure consistency loss to directly optimize the quality of generated molecular graphs, improving the consistency of generated molecular structures.
- Extensive experiments on molecular datasets demonstrate the effectiveness of our proposed method. The detailed ablation studies verify the effectiveness of each module.

2 Preliminaries

To facilitate the understanding of our proposed method, we first present the problem formulation and review the background of discrete graph diffusion models, which serve as the theoretical foundation of our approach.

2.1 Problem Formulation

Given a dataset of paired text descriptions and molecular graphs $\mathcal{D} = \{(G_1, D_1), \ldots, (G_{|\mathcal{D}|}, D_{|\mathcal{D}|})\}$, where D_i is the text description of graph G_i. Our goal is to learn the conditional distribution $p(G|D)$, so that when given a text description D of a molecular graph, the model can generate the corresponding molecular graph G. Specifically, we aim to generate the atoms in the molecular graph G and the bonds between the atoms.

In addition to the atom types and bond types, formal charge is also an important property of atoms in molecules, which is the hypothetical charge assigned to an atom in a molecule, assuming that electrons in all chemical bonds are shared equally between atoms, regardless of relative electronegativity. However, existing graph diffusion models, which learn the distribution of molecular graphs and generate new molecular graphs, typically omit formal charge information. This omission limits the model's ability to accurately capture the overall molecular structure and its associated properties. To address this limitation, we explicitly incorporate formal charge as part of the molecular graph representation used in the diffusion model. We treat the formal charge as a categorical attribute that can take values from a predefined set.

Formally, we use $X = \{x_i\}$ to denote the nodes of the molecular graph, where x_i denotes atom type of the i-th atom (M_x types in total, e.g., C, H, O, N), $C = \{c_i\}$ to denote the formal charges of atoms, where c_i denotes the formal charge number of the i-th atom (M_c options in total, e.g., $-1, 0, +1, +2$), and use $E_k = \{e_{i,j}\}$ to denote the type of chemical bonds between atoms (M_e types in total, e.g., single bond, double bond). The absence of edges between nodes is represented with a special edge type. In this paper, we represent these categorical data as one-hot vectors, so we have $x_i \in \mathbb{R}^{M_x}$, $c_i \in \mathbb{R}^{M_c}$, $e_i \in \mathbb{R}^{M_e}$.

2.2 Discrete Graph Diffusion

Diffusion models as a class of generative models have recently gained popularity for their excellent performance in computer vision. Since the presence or absence of edges in graphs is naturally a binary and discretized property, and the kind of atoms and bonds in molecular graphs is also discrete, discrete-state graph diffusion models have grown popular for generating molecular graphs. We provide a brief introduction to the discrete graph diffusion models as follows.

The discrete graph diffusion model comprises a forward process and a reverse process. Formally, let $G = (X, E)$ denote a graph, where $X = \{x_i\}$ are the node types and $E = \{e_{i,j}\}$ are the edge types. The forward process follows a Markov chain, initiating from the original graph $G_0 = (X_0, E_0) = (X, E)$. At each step, noise is gradually introduced to both the node types X and edge types E, progressively corrupting the graph. This iterative process continues until the final state $G_T = (X_T, E_T)$, where the graph is fully transformed into noise. Specifically, in each step, the node types X

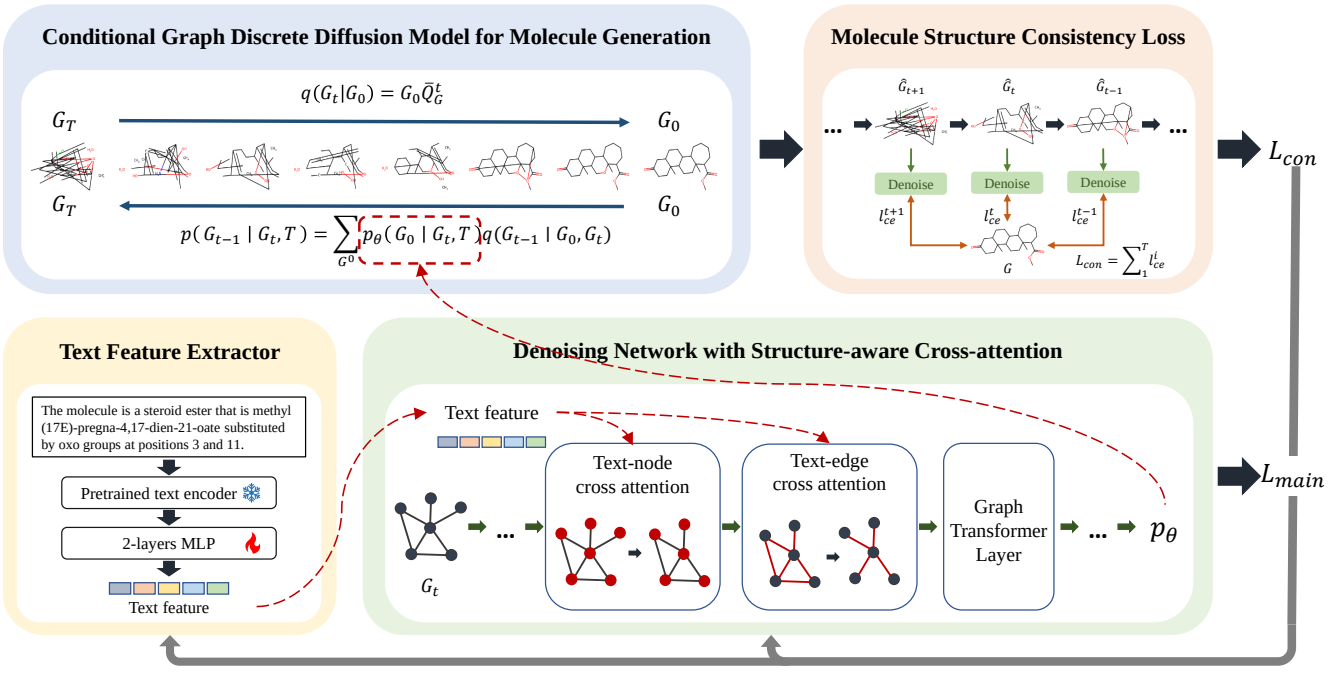


Figure 1: An overview of our method. (a) Upper-left: the conditional discrete graph diffusion model facilitates text-guided molecular graph generation. (b) Lower-left: a pretrained text encoder and a MLP-based projection module are responsible for extracting textual features, which are subsequently utilized by the denoising network. (c) Lower-right: the denoising network employs structure-aware cross-attention to refine noisy molecular graphs, utilizing the extracted text features as conditional information. (d) Upper-right: the molecule structure consistency loss quantifies the discrepancy between intermediate denoised results and the ground-truth molecular graph, providing a loss signal to optimize the quality of generated molecules.

and edge types E are modified according to predefined transition matrices Q_X^t and Q_E^t. Each node type and edge type is transformed independently, with a small probability of transforming into another type. This process can be formally described as follows:

$$\begin{aligned} q(x_t = j \mid x_{t-1} = i) &= Q_{X,ij}^t \\ q(e_t = j \mid e_{t-1} = i) &= Q_{E,ij}^t \end{aligned}, \tag{1}$$

where Q_X^t and Q_E^t are transition matrices for node types and edge types respectively.

Given the initial state G_0, we can derive a closed-form expression to directly sample G_T, which facilitates efficient sampling of noisy graphs for training. Specifically, by employing a one-hot encoding representation for the initial node types X_0 and edge types E_0, the evolution of the graph over time can be expressed as follows:

$$q(X_t \mid X_0) = X_0\bar{Q}_X^t, \quad \bar{Q}_X^t = Q_X^1 \ldots Q_X^t \tag{2}$$

$$q(E_t \mid E_0) = E_0\bar{Q}_E^t, \quad \bar{Q}_E^t = Q_E^1 \ldots Q_E^t. \tag{3}$$

The reverse process in a graph diffusion model is parameterized by a denoising neural network $p_\theta(G_0 \mid G_t)$, which is designed to remove noise from the noisy graph G_t. Each step of the reverse process is defined in terms of the transition probability $p(G_{t-1} \mid G_t)$, which can be expressed as a marginalization over the predicted clean graph:

$$p(G_{t-1} \mid G_t) = \sum_{G_0} p_\theta(G_0 \mid G_t) q(G_{t-1} \mid G_0, G_t), \tag{4}$$

where $q(G_{t-1} \mid G_0, G_t)$ can be computed from the definition of $q(G_t \mid G_0)$. With p_θ defined as a factorized distribution over the node types and edge types, Eq. (4) can be computed individually per node or edge type efficiently. The reverse process reconstructs the graph structures and generates novel graphs. The denoising network is optimized using the following objective function:

$$\begin{aligned} \mathcal{L} = \mathbb{E}_t \mathbb{E}_{G_0 \sim \mathcal{D}} \mathbb{E}_{G_t \sim q(G_t|G_0)} \\ \left[\lambda_X \mathrm{CE}(X_0, p_\theta(X_0 \mid G_t)) + \lambda_E \mathrm{CE}(E_0, p_\theta(E_0 \mid G_t))\right], \end{aligned} \tag{5}$$

where $\mathrm{CE}(\cdot)$ is the cross-entropy loss, and λ_X and λ_E are hyperparameters controlling the weighting of terms.

3 Methodology

We formally define the research problem in Section 2.1. To address this problem, we introduce the Text-guided Conditional Discrete Graph Diffusion (TDGD) model, which learns the conditional distribution $p_\theta(G|D)$ to generate a molecular graph G from a given textual description D. Our approach leverages a discrete diffusion process to model the complex structure of molecular graphs while conditioning on text representations to ensure semantic consistency.

The overall framework of our proposed method is depicted in Figure 1. We develop a **Conditional Discrete Graph Diffusion Model** designed to generate molecular graphs conditioned on a given textual description. To incorporate the text information into

Algorithm 1 Training TDGD

while not converged **do**
  Sample $t \sim \text{Uniform}(1, \ldots, T)$
  Sample (G_0, D) from the dataset
  Sample G_t using Eq. (6)
  if rand() < unconditional training probability **then**
    Compute $p_\theta(G_0 \mid G_t, D = \varnothing)$
    Optimize using $\mathcal{L}_{\text{main}}$ in Eq. (8)
  else
    Compute $p_\theta(G_0 \mid G_t, D)$
    Optimize using $\mathcal{L}_{\text{main}}$ in Eq. (8)
  end if
  if using molecule structure consistency loss **then**
    Sample $\hat{G}_T$ from the initial distribution
    for t' = T to 1 **do**
      Compute $p_\theta(G_0 \mid \hat{G}_{t'}, D)$
      Compute $\mathcal{L}_{\text{con}}$ in Eq. (14)
      Sample $\hat{G}_{t'-1}$ from $p(G_{t'-1} \mid \hat{G}_{t'}, D)$ using Eq. (7)
    end for
    Optimize using sum of $\mathcal{L}_{\text{con}}$ over t'
  end if
end while

Algorithm 2 Sampling from TDGD following the text description

Input: Text description D

Sample G_T from the initial distribution
for t = T to 1 **do**
  Compute conditional predictions $p_\theta(G_0 \mid G_t, D)$
  Compute unconditional predictions $p_\theta(G_0 \mid G_t, D = \varnothing)$
  Compute classifier-free guidance p_{cfg} using Eq. (15)
  Sample G_{t-1} from $p(G_{t-1} \mid G_t, D)$ using Eq. (7), using p_{cfg} in place of p_θ
end for
return molecular graph G_0

the model, we use a pretrained MolT5[5] encoder to extract the text features, followed by a two-layer MLP projection module that maps the text features into the desired dimensionality. The text features are utilized by the denoising network, which is a **graph transformer** [33] equipped with structure-aware cross-attention designed in Section 3.2. It enables the model to capture both the structural dependencies within the graph and the semantic alignment between the graph and the text. During the generation process, we incorporate classifier-free guidance to enhance the correspondence between the generated graph and the textual description, improving both fidelity and diversity. The training and sampling procedure for TDGD is presented in Algorithm 1 and 2 respectively.

3.1 Conditional Discrete Graph Diffusion Model for Molecule Generation

In our method, we design a conditional discrete graph diffusion model, which incorporates text features into the discrete graph diffusion model to achieve text-guided molecular graph generation.

As outlined in Section 2.1, we aim to generate the atom types, formal charges, and edge types of molecules. Therefore, we design the forward process of the conditional discrete graph diffusion model to operate on these components of molecules. Given the initial graph state $G_0 = (X_0, C_0, E_0)$, we can derive a closed-form expression that directly samples G_t at time t. The evolution of the graph over time is modeled as follows:

$$\begin{aligned} q(X_t \mid X_0) &= X_0 \bar{Q}_X^t, \quad \bar{Q}_X^t = Q_X^1 \ldots Q_X^t \\ q(C_t \mid C_0) &= C_0 \bar{Q}_C^t, \quad \bar{Q}_C^t = Q_C^1 \ldots Q_C^t \\ q(E_t \mid E_0) &= E_0 \bar{Q}_E^t, \quad \bar{Q}_E^t = Q_E^1 \ldots Q_E^t, \end{aligned} \tag{6}$$

where $X \in \mathbb{R}^{N \times M_X}$ represents the matrix of atom types, $E \in \mathbb{R}^{N \times N \times M_E}$ is the matrix of edge types, $C \in \mathbb{R}^{N \times M_C}$ denotes the matrix of formal charge numbers of the atoms, Q_X, Q_C, Q_E are predefined transition matrices, and t indicates the time step of the graph evolution.

To enable text-guided generation, we modify the denoising neural network $p_\theta(G_0 \mid G_t)$ in the discrete graph diffusion model by introducing conditional information based on the text description, $p_\theta(G_0 \mid G_t, D)$. This conditional network incorporates textual information D as an additional input using structure-aware cross-attention (described in Section 3.2), enabling it to learn a mapping that denoises G_t while adhering to the semantic constraints imposed by D. Then, the transition probability $p(G^{t-1} \mid G^t, D)$ in the reverse process can be expressed as a marginalization over the original graph distribution following the text description D of molecule:

$$p(G_{t-1} \mid G_t, D) = \sum_{G_0} p_\theta(G_0 \mid G_t, D) q(G_{t-1} \mid G_0, G_t), \tag{7}$$

which enables a gradual denoising procedure guided by D to reconstruct the underlying graph structure.

The conditional denoising network can be optimized using the following objective function:

$$\begin{aligned} \mathcal{L}_{\text{main}} = \mathbb{E}_t \mathbb{E}_{G_0 \sim \mathcal{D}} \mathbb{E}_{G_t \sim q(G_t \mid G_0)} [\lambda_X \text{CE}(X_0, p_\theta(X_0 \mid G_t, D)) \\ + \lambda_C \text{CE}(C_0, p_\theta(C_0 \mid G_t, D)) + \lambda_E \text{CE}(E_0, p_\theta(E_0 \mid G_t, D))], \end{aligned} \tag{8}$$

where $\text{CE}(\cdot)$ is the cross-entropy loss, and λ_X, λ_C and λ_E are hyperparameters controlling the weighting of terms.

3.2 Structure-aware Cross-attention Mechanism

To enable text-to-graph generation via diffusion models, it is essential to incorporate text features into the conditional denoising network. While standard approaches like affine conditioning and cross-attention have been used in diffusion models, they are not directly suited to graphs due to their unique structure. In particular, edges in graphs represent node relationships, so effective conditioning should ensure edge features are contextually dependent on node representations. Moreover, computing edge-level conditioning using cross-attention becomes inefficient, as the adjacency matrix scales quadratically with the number of nodes.

To address these challenges, we propose a structure-aware cross-attention mechanism that integrates sequential text features into the denoising network by aligning them with graph structures through structure-aware attention.

Specifically, we use the structure-aware cross-attention mechanism in the denoising network of the graph diffusion model, where the results of node attention and edge attention are added into the network using residual connections. Let X and E be the node and edge features in some layer of the denoising network, and C be the sequence of text features. Structure-aware cross-attention first computes the cross-attention between node features and text features as follows, and use the attention results for node conditioning:

$$Q = XW_Q, \quad K = CW_K, \quad V = CW_V, \tag{9}$$

$$A = \text{softmax}(\frac{QK^T}{\sqrt{d}}), \quad X_{\text{cond}} = AV, \tag{10}$$

where Q, K, and V are weight matrices for queries, keys, and values, d is the dimensionality of keys, A is the attention score for nodes, and X_{cond} is the node conditioning result.

Then, structure-aware cross-attention computes the attention scores for edges based on the node attention results. For each edge (u, v), its attention score should be related to the attention scores of node u and v. We compute two scalar values $G_{1,uv}$ using a gating mechanism for each edge based on its features E_{uv}, which represents the influence of two endpoints u and v in the edge:

$$G_{1,uv} = \sigma(E_{uv}^T W_{G1}), G_{2,uv} = 1 - G_{1,uv}, \tag{11}$$

where W_{G1} is trainable weights, and σ is the logistic sigmoid function. The attention score of edge (u, v) is computed as a weighted mixture of the attention scores for node u and v:

$$A_{\text{edge},uv} = G_{1,uv}A_u + G_{2,uv}A_v, \tag{12}$$

where $A_{\text{edge},uv}$ is the attention score for edge (u, v), and A_u and A_v are attention scores for node u and v respectively.

Finally, the edge conditioning result is determined by mixing the attention values according to the edge attention scores:

$$E_{\text{cond},uv} = A_{\text{edge},uv}V. \tag{13}$$

In our method, we modify the graph transformer architecture of the denoising network, adding a structure-aware cross-attention module in each graph transformer layer.

By deriving the edge conditioning from the node conditioning, structure-aware cross-attention can utilize the text features in the conditional denoising network with relatively low computational costs, and captures the relational semantics among texts and structures more efficiently.

3.3 Molecule Structure Consistency Loss

In discrete graph diffusion models, the denoising network is trained to remove noise from corrupted graphs, but this objective does not directly optimize the quality of generated samples. In text-guided settings, where textual descriptions impose stricter structural constraints than unconditional generation, the standard loss fails to accommodate this increased complexity. To address this, we propose a novel loss function that explicitly enforces structural consistency between generated graphs and their textual descriptions.

In this section, we introduce our proposed Molecule Structure Consistency Loss. Specifically, given a dataset sample (G, D), where G represents the molecular graph and D is the corresponding textual description, we consider a diffusion sampling process $\hat{G}_T, \ldots, \hat{G}_0$ conditioned on D. Our objective is to ensure that the denoising network consistently predicts G_0 to be close to G at each step of the sampling process. Formally, we define the loss function as follows:

$$\begin{aligned}\mathcal{L}_{\text{con}} = \mathbb{E}_{\hat{G}_T,\ldots,\hat{G}_0} \sum_{t=1}^{T} \Big(\lambda_X \text{CE}(X, p_\theta(X_0 \mid \hat{G}_t, D)) + \\ \lambda_C \text{CE}(C, p_\theta(C_0 \mid \hat{G}_t, D)) + \lambda_E \text{CE}(E, p_\theta(E_0 \mid \hat{G}_t, D))\Big).\end{aligned} \tag{14}$$

where λ_X, λ_C, and λ_E are weighting terms.

This loss function in Eq. (14) shares a similar formulation with the main loss function in Eq. (8), as both aim to minimize the discrepancy between the denoising network's predictions and the ground-truth molecular graph. However, the key distinction lies in how the noisy data is obtained: instead of directly adding noise to the real molecular graph, the Molecule Structure Consistency Loss leverages trajectories from the sampling process to generate perturbed inputs. This design ensures that the denoising network learns to recover molecular structures under conditions that more closely align with the actual sampling dynamics.

3.4 Classifier-free Diffusion Guidance

Classifier-free guidance [11] is a widely used technique for conditional generation using diffusion models. It trains the denoising network in diffusion models for both unconditional and conditional inputs, and modifies the output of denoising network during sampling by scaling up the difference between the unconditional and conditional predictions.

For discrete graph diffusion models, since the denoising network predicts a discrete probability distribution rather than the mean of a Gaussian distribution as in continuous diffusion models, a novel classifier-free guidance method needs to be designed. The specific approach is outlined as follows:

$$\begin{aligned}p_{\text{cfg}}(G_0 \mid G_t, D) = \\ \frac{1}{Z} p_\theta(G_0 \mid G_t, D)^k p_\theta(G_0 \mid G_t, D = \varnothing)^{1-k}\end{aligned} \tag{15}$$

where k is the guidance scale, and Z is a normalization coefficient that ensures the sum of p_{cfg} equals 1. p_{cfg} can be computed by linearly mixing the pre-softmax logits produced by the denoising network according to k and then taking softmax.

4 Experiment

To evaluate the effectiveness of our proposed method, we designed comprehensive experiments on ChEBI-20 and L+M-24 datasets, which are two major datasets for the task of text-guided graph generation. Then, we perform a detailed ablation study to analyze the contributions of the key modules introduced in our framework. In the first part, we benchmark our approach against multiple baseline methods on the task of text-guided molecular graph generation using the ChEBI-20 dataset, assessing both generation quality and structural validity. In the second part, we extend this comparison to the L+M-24 dataset, further validating the generalization ability of our method across different molecular graph distributions. In the third part, we perform a detailed ablation study to analyze the contributions of the key modules introduced in our framework.

4.1 Text-guided Molecular Graph Generation on ChEBI-20 Dataset

Dataset. The ChEBI-20 [7] dataset consists of 33,010 pairs of molecules and descriptions, each containing text descriptions of more than 20 words. It was created by collecting ChEBI annotations from PubChem, with a focus on reducing noise and improving informativeness. The dataset is split into 80% training, 10% validation, and 10% test sets.

Baselines. We compare with the following methods:

- **MolT5** [5]: [leftmargin = 0.5cm] An encoder-decoder Transformer model initialized with a public checkpoint of T5, then pretrained on the combined dataset of C4 and ZINC, finally finetuned on the ChEBI-20 dataset. Divided into three models based on the number of trainable parameters: MolT5-Small, MolT5-Base and MolT5-Large.
- **3M-Diffusion** [34]: A latent multi-modal diffusion model that performs text-guided molecular graph generation in a shared graph latent space. The baseline aligns text embeddings and molecular graphs into a joint latent representation, and applies a conditional diffusion process to map text inputs to latent graph features, which are then decoded into molecular graphs.
- **UTGDiff** [29]: A unified text–graph diffusion framework that models text-guided graph generation within a shared diffusion process. The baseline uses a pretrained language-model-based transformer with attention bias as the denoising network to enable interactions between text tokens and graph components, while keeping the overall architecture simple and unified.
- **TGM-DLM** [9]: A text diffusion model for text-guided molecule generation, which iteratively refines SMILES token embeddings through a two-phase diffusion process—first optimizing from noise with text guidance and then correcting invalid molecular structures to ensure valid representations.

Metrics. We use the following metrics to evaluate our method and compare it with the baseline:

- Fingerprint-based molecule similarity metrics (MACCS FTS, RDK FTS, Morgan FTS): We measure the fingerprint Tanimoto similarity (FTS) between each generated molecule and the corresponding ground truth molecule. MACCS, RDK, and Morgan represent the three different extraction methods for molecular fingerprints. We consider these to be the primary metrics measuring the quality of generated graphs in our experiments.
- FCD: We measure the Fréchet ChemNet Distance (FCD) [23] between the generated molecules and ground truth molecules. It reflects the distance of chemical and biological information between the two sets of molecules.
- Text2Mol[7]: The metric evaluates the similarity between molecules and their textual descriptions by leveraging similarity between embeddings, with a trained multi-layer perceptron (MLP) model to rank molecule-description pairs.
- SMILES-based metrics (BLEU, Exact, Levenshtein): These metrics measure the similarity between the ground truth molecules and generated molecules using their SMILES representation. BLEU measures the n-gram overlap of the two strings, Exact measures the proportion of exact string matches, and Levenshtein measures the text editing distance. Since our method does not generate SMILES string directly, the metrics for our method are computed by converting both the ground truth and the generated molecules to their canonical SMILES representation.
- Validity: We report the validity of the generated molecules as measured by RDKit sanitization.

Results and Analysis. From Table 1, we can observe that: (1) Compared to MolT5, our approach consistently achieves superior performance across molecular metrics, indicating that it captures molecular structures more effectively than language-model-based methods when generating molecular graphs from textual descriptions. (2) When evaluated against diffusion-based methods, our method reaches superior or comparable performance on all three FTS metrics, which are highly correlated with molecular structure and properties. This indicates that, compared to text-based diffusion models, our method generates molecular graphs that more accurately align with the structural information described in the text. (3) By evaluating the generated molecular graphs in SMILES representations, our approach demonstrates superior performance on SMILES-based metrics.

4.2 Text-guided Molecular Graph Generation on L+M-24 Dataset

Dataset. The L+M-24 [6] dataset consists of 200,615 molecule-description pairs derived from four key categories: Biomedical, Light and Electricity, Human Interaction and Organoleptics, and Agriculture and Industry. It includes functional and compositional properties, designed to test model performance on abstract, functional, and compositional tasks.

Baselines. We compare with MolT5, 3M-Diffusion, UTGDiff, as well as the following method.

- **Lang2Mol-Diff** [20]: An enhanced version of TGM-DLM [9] tailored for the L+M-24 benchmark. This baseline follows the diffusion-based text-guided molecule generation paradigm of TGM-DLM, and introduces minor adaptations to the text conditioning and training configuration to better handle the characteristics of L+M-24, while preserving the original model architecture and generation process.

Metrics. We use the following metrics to evaluate our method and compare it with the baseline: MACCS FTS, RDK FTS, Morgan FTS, FCD, BLEU, Exact, Levenshtein, Validity.

Results and Analysis. Compared to ChEBI-20, the L+M-24 dataset is constructed from a broader range of sources and contains more diverse textual descriptions of molecules, making molecular generation significantly more challenging. This increased difficulty is reflected in the near-zero exact match scores across all methods. Nevertheless, our approach maintains strong performance on L+M-24. As shown in Table 2, our experimental results indicate that: (1) Compared to language model-based methods, our approach consistently achieves superior performance on molecular evaluation metrics. (2) Compared to diffusion-based methods, our method outperforms them across all reported metrics, demonstrating its enhanced capability in generating molecular graph structures from complex textual descriptions.

Table 1: The result of text-guided molecule generation on ChEBI-20 dataset. "MACCS FTS", "RDK FTS", and "Morgan FTS" are fingerprint-based molecule similarity metrics, "FCD" and "Text2Mol" are embedding-based metrics, and "Rank" is the average ranking among listed methods over molecular metrics. "BLEU", "Exact", and "Levenshtein" metrics computed using SMILES strings, which are only listed for reference and do not reflect the performance of molecule generation. "Validity" is the proportion of generated molecules that are chemically valid. ↑ indicates higher is better, ↓ indicates lower is better, and Bold highlights the best score. ⋆ FCD values of baselines may differ from those listed in other works due to subtle differences in the way the metric is computed. We try to ensure a consistent evaluation protocol across all baselines.

Model	Molecular metrics						SMILES-based metrics			
	MACCS FTS↑	**RDK FTS↑**	**Morgan FTS↑**	**FCD↓ ⋆**	**Text2Mol↑**	**Rank↓**	BLEU↑	Exact↑	Levenshtein↓	Validity↑
Ground Truth	1.000	1.000	1.000	0.0	0.609	—	1.000	1.000	0.0	1.0
MolT5-Small	0.703	0.568	0.517	2.49	0.482	7.2	0.755	0.079	25.988	0.721
MolT5-Base	0.721	0.588	0.529	2.18	0.496	6.0	0.769	0.081	24.458	0.772
MolT5-Large	0.834	0.746	0.684	1.20	0.554	4.6	**0.854**	0.311	16.071	0.905
3M-Diffusion	0.557	0.380	0.302	2.35	0.416	7.8	0.507	0.003	20.480	0.595
UTGDiff	0.885	**0.795**	0.724	1.34	0.587	2.4	0.773	0.374	14.671	0.893
TGM-DLM$_{w/o\ corr}$	0.874	0.771	0.722	0.89	**0.589**	2.4	0.828	0.242	16.897	0.789
TGM-DLM	0.854	0.739	0.688	**0.77**	0.581	3.6	0.826	0.242	17.003	0.871
Ours	**0.892**	0.794	**0.757**	1.08	0.584	**2.0**	0.738	**0.460**	**13.755**	**1.000**

Table 2: The result of text-guided molecule generation on L+M-24 dataset. "MACCS FTS", "RDK FTS", and "Morgan FTS" are fingerprint-based molecule similarity metrics, "FCD" is an embedding-based metric, and "Rank" is the average ranking among listed methods over molecular metrics. "BLEU", "Exact", and "Levenshtein" metrics computed using SMILES strings, which are only listed for reference and do not reflect the performance of molecule generation. "Validity" is the proportion of generated molecules that are chemically valid. ↑ indicates higher is better, ↓ indicates lower is better, and Bold highlights the best score.

Model	Molecular metrics					SMILES-based metrics			
	MACCS FTS↑	**RDK FTS↑**	**Morgan FTS↑**	**FCD↓**	**Rank↓**	BLEU↑	Exact↑	Levenshtein↓	Validity↑
Ground Truth	1.000	1.000	1.000	0.0	—	1.000	1.000	0.00	1.0
MolT5-Small	0.642	0.581	0.374	NaN	4.50	0.566	0.00	56.34	0.805
MolT5-Base	0.760	0.652	0.475	NaN	3.25	0.684	0.00	44.79	1.000
MolT5-Large	0.757	0.650	0.395	17.52	2.75	0.564	0.00	55.40	0.994
3M-Diffusion	0.215	0.120	0.073	22.91	6.25	0.093	0.00	71.63	0.408
UTGDiff	0.481	0.357	0.234	6.49	4.75	0.583	0.00	45.22	0.617
Lang2Mol-Diff	0.606	0.332	0.328	38.09	5.25	0.543	0.00	55.87	1.000
Ours	**0.781**	**0.719**	**0.560**	**4.67**	**1.00**	**0.814**	0.00	**29.38**	**1.000**

4.3 Ablation Study

We conducted ablation experiments to explore the effects of different settings. All ablation experiments are performed on the ChEBI-20 dataset. Here, we focus on the molecule similarity metrics, namely the FTS metrics as well as FCD.

4.3.1 Effect of Different Text Conditioning Methods. We compare our structure-aware cross-attention mechanism with other text conditioning methods for diffusion models.

Compared Methods. We compare the following methods:

- Affine: The features of the last token in the text description are inserted into the denoising network using feature-wise affine transformations, also known as feature-wise linear modulation (FiLM).
- Cross-attention: The text features are inserted into the denoising network using cross-attention between node features and text features. The edge features are not modified directly.

Table 3: The result of different text conditioning methods on ChEBI-20 dataset. "MACCS FTS", "RDK FTS", and "Morgan FTS" are fingerprint-based molecule similarity metrics, "FCD" is embedding-based metric. ↑ indicates higher is better, ↓ indicates lower is better, and Bold highlights the best score.

Model	MACCS FTS↑	RDK FTS↑	Morgan FTS↑	FCD↓
Affine	0.833	0.726	0.653	2.16
Cross-attention	0.856	0.747	0.701	1.64
Ours	**0.892**	**0.794**	**0.757**	**1.08**

Table 4: Ablation study results of molecule structure consistency loss on ChEBI-20 dataset.

Model	MACCS FTS↑	RDK FTS↑	Morgan FTS↑	FCD↓
Not using loss	0.866	0.760	0.720	1.28
Using loss	**0.892**	**0.794**	**0.757**	**1.08**

Table 5: Ablation study results of generating formal charges on ChEBI-20 dataset.

Model	MACCS FTS↑	RDK FTS↑	Morgan FTS↑	FCD↓
No formal charge	0.871	0.793	0.706	1.44
With formal charge	**0.892**	**0.794**	**0.757**	**1.08**

- Ours: The text features are inserted into the denoising network using our proposed structure-aware cross-attention mechanism.

Results and Analysis. The experimental results are shown in Table 3. We can find that structure-aware cross-attention achieves the best performance among compared methods in terms of molecular similarity to the ground truth. This indicates that our proposed method can better incorporate text features into graph diffusion models.

4.3.2 Effect of Molecule Structure Consistency Loss. We conducted experiments to verify the effectiveness of the molecule structure consistency loss proposed in Section 3.3.

Compared methods. We compare the following methods:

- Not using loss: The model is optimized using only denoising loss L_{main} in Eq. (8).
- Using loss: The model is optimized using only denoising loss L_{main} in Eq. (8) and molecule structure consistency loss L_{con} in Eq. (14).

Results and Analysis. The experimental results are shown in Table 4. It can be observed that incorporating the proposed loss consistently outperforms the variant without this loss across all evaluation metrics. This improvement suggests that the molecule structure consistency loss provides additional and effective training signals beyond the standard denoising objective, encouraging the model to preserve structural coherence throughout the diffusion process. As a result, the denoising network is able to learn more accurate and stable molecular graph structures that better align with the underlying ground-truth molecules.

4.3.3 Effect of Atom's Formal Charge. We conducted experiments to verify the effectiveness of the atom's formal charge added in conditional discrete graph diffusion model in Section 3.1.

Compared Methods. We compare the following methods:

- No formal charge: The model only takes into account the atom types and chemical bond types.
- With formal charge : The model takes into account the atom types, chemical bond types and formal charge number of atoms.

Results and Analysis. The experimental results are reported in Table 5. We observe that explicitly incorporating formal charge information consistently improves model performance across all evaluation metrics. This suggests that formal charge provides complementary structural cues beyond atom and bond types, enabling the model to better capture fine-grained molecular structure. As a result, considering formal charge in the graph diffusion process contributes to more accurate molecular graph generation and improved alignment with the target molecular structures.

5 Related Work

In this section, we review related work on diffusion-based graph generation and text-guided molecule generation.

5.1 Diffusion-based Graph Generation

Diffusion models have achieved great success in the field of computer vision. Recently, some researchers[2, 12–14, 18, 27, 31] have used diffusion models to solve graph generation tasks. For example, EDP-GNN [21] is the first work using Score Matching with Langevin Dynamics (SMLD) [25] diffusion model to generate graphs, which learns the score function of the adjacency matrices distributions of the graphs. GDSS [13] proposes a graph generation method using continuous-time diffusion models [26], which models the joint distribution of the nodes and edges through stochastic differential equations (SDEs). DiGress [27] uses a diffusion model over discrete data space for graph generation, and additionally preserves the marginal distribution of node and edge types and incorporates auxiliary graph-theoretic features. These methods have demonstrated excellent performance on the task of graph generation.

In order to generate graphs that match specific requirements, conditional graph generation[2, 18, 27] has received attention in recent years. For example, DiGress [27] uses classifier guidance to perform graph generation guided by several graph-level properties, like the dipole moment and highest occupied molecular orbit of molecular graphs.

5.2 Text-guided Molecule Generation

Molecules can be represented in both text and graph forms. Most existing text-guided molecular generation methods[3, 5, 8, 9, 20, 22] primarily focus on generating molecular sequences such as SMILES[28] and SELFIES[15]. These methods can generally be categorized into autoregressive language models and text-based diffusion models. Approaches like MolT5[5], Text+ChemT5[3], and BioT5[22] employ an encoder-decoder Transformer architecture similar to the T5 model.

To address the adaptability limitations imposed by the fixed generation order in autoregressive models, recent methods, such as TGM-DLM[9] and Lang2Mol-Diff[20], have introduced two-stage text-based diffusion models for molecular generation. In these models, the first stage iteratively optimizes the latent embeddings from random noise under textual guidance, while the second stage refines invalid SMILES/SELFIES strings in an unsupervised manner to produce valid molecular representations. Additionally, some studies[34] have explored the use of diffusion models for latent space sampling, followed by graph-based decoders to reconstruct molecular graphs. More recently, a unified text–graph diffusion

framework [29] integrates text conditioning and molecular graph generation within a single diffusion process by employing a pre-trained language-model-based transformer with attention bias, enabling joint modeling of text tokens and graph structures.

Although existing diffusion-based methods have shown strong performance in text-guided molecular generation, many of them incorporate molecular structures in an indirect manner, such as through latent spaces or unified token representations. While recent text–graph diffusion models begin to jointly model text and molecular graphs, effectively capturing fine-grained structural dependencies and directly enforcing structural consistency during generation remain challenging. In this work, we investigate conditional discrete graph diffusion models to more explicitly leverage molecular structural information for text-guided molecular graph generation.

6 Conclusion

In this paper, we introduced the Text-guided Conditional Discrete Graph Diffusion (TDGD) model to leverage graph discrete diffusion models for text-guided molecule generation. To address key challenges in aligning molecular structures with text descriptions, we proposed a structure-aware cross-attention mechanism to enhance the model's ability to capture relational semantics and a molecule structure consistency loss to improve generation quality. Extensive experiments on the ChEBI-20 and L+M-24 datasets demonstrate the effectiveness of our proposed TDGD in generating molecular graphs that faithfully adhere to textual constraints. Our findings demonstrate the potential of conditional discrete graph diffusion models in advancing controlled molecule generation.

Limitations and Ethical Considerations

Our method relies on the quality and coverage of text–molecule data pairs, which may limit generalization to out-of-distribution descriptions. In addition, while the model captures overall semantic alignment, achieving fine-grained control over detailed chemical constraints from text remains challenging. The computational cost of diffusion models is also relatively high compared to simpler generative approaches. While our method is intended to support beneficial applications such as drug discovery, it could potentially be misused for generating harmful or unsafe chemical structures. Therefore, generated molecules should not be directly used without proper domain-specific validation, and practical deployment should incorporate safety screening and expert oversight.

Acknowledgement

This work was supported by the National Key Research and Development Program of China No.2023YFF1205001.

References

[1] Haibo Chen, Xin Wang, Zeyang Zhang, Haoyang Li, Ling Feng, and Wenwu Zhu. 2025. AutoGFM: Automated Graph Foundation Model with Adaptive Architecture Customization. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025 (Proceedings of Machine Learning Research)*, Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (Eds.). PMLR / OpenReview.net. https://proceedings.mlr.press/v267/chen25bp.html

[2] Xiaohui Chen, Jiaxing He, Xu Han, and Li-Ping Liu. 2023. Efficient and Degree-Guided Graph Generation via Discrete Diffusion Modeling. *CoRR* abs/2305.04111 (2023). arXiv:2305.04111 doi:10.48550/ARXIV.2305.04111

[3] Dimitrios Christofidellis, Giorgio Giannone, Jannis Born, Ole Winther, Teodoro Laino, and Matteo Manica. 2023. Unifying Molecular and Textual Representations via Multi-task Language Modelling. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 6140–6157. https://proceedings.mlr.press/v202/christofidellis23a.html

[4] Hanjun Dai, Yingtao Tian, Bo Dai, Steven Skiena, and Le Song. 2018. Syntax-Directed Variational Autoencoder for Structured Data. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net. https://openreview.net/forum?id=SyqShMZRb

[5] Carl Edwards, Tuan Manh Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. 2022. Translation between Molecules and Natural Language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, 375–413. doi:10.18653/V1/2022.EMNLP-MAIN.26

[6] Carl Edwards, Qingyun Wang, Lawrence Zhao, and Heng Ji. 2024. L+ M-24: Building a Dataset for Language+ Molecules@ ACL 2024. *CoRR* (2024).

[7] Carl Edwards, ChengXiang Zhai, and Heng Ji. 2021. Text2Mol: Cross-Modal Molecule Retrieval with Natural Language Queries. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, 595–607. doi:10.18653/V1/2021.EMNLP-MAIN.47

[8] Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. 2024. Mol-Instructions: A Large-Scale Biomolecular Instruction Dataset for Large Language Models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net. https://openreview.net/forum?id=Tlsdsb6l9n

[9] Haisong Gong, Qiang Liu, Shu Wu, and Liang Wang. 2024. Text-guided molecule generation with diffusion language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 109–117.

[10] Nate Gruver, Samuel Stanton, Nathan C. Frey, Tim G. J. Rudner, Isidro Hötzel, Julien Lafrance-Vanasse, Arvind Rajpal, Kyunghyun Cho, and Andrew Gordon Wilson. 2023. Protein Design with Guided Discrete Diffusion. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (Eds.). http://papers.nips.cc/paper_files/paper/2023/hash/29591f355702c3f4436991335784b503-Abstract-Conference.html

[11] Jonathan Ho and Tim Salimans. 2021. Classifier-Free Diffusion Guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.

[12] Han Huang, Leilei Sun, Bowen Du, and Weifeng Lv. 2023. Conditional Diffusion Based on Discrete Graph Structures for Molecular Graph Generation. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, Brian Williams, Yiling Chen, and Jennifer Neville (Eds.). AAAI Press, 4302–4311. doi:10.1609/AAAI.V37I4.25549

[13] Jaehyeong Jo, Seul Lee, and Sung Ju Hwang. 2022. Score-based Generative Modeling of Graphs via the System of Stochastic Differential Equations. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (Eds.). PMLR, 10362–10383. https://proceedings.mlr.press/v162/jo22a.html

[14] Lingkai Kong, Jiaming Cui, Haotian Sun, Yuchen Zhuang, B. Aditya Prakash, and Chao Zhang. 2023. Autoregressive Diffusion Model for Graph Generation. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 17391–17408. https://proceedings.mlr.press/v202/kong23b.html

[15] Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alán Aspuru-Guzik. 2020. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Mach. Learn. Sci. Technol.* 1, 4 (2020), 45024. doi:10.1088/2632-2153/ABA947

[16] Haoyang Li, Xin Wang, Zeyang Zhang, Haibo Chen, Ziwei Zhang, and Wenwu Zhu. 2024. Disentangled Graph Self-supervised Learning for Out-of-Distribution Generalization. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024 (Proceedings of Machine Learning Research)*, Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (Eds.). PMLR / OpenReview.net, 28890–28904. https://proceedings.mlr.press/v235/li24br.html

[17] Haoyang Li, Xin Wang, Xueling Zhu, Weigao Wen, and Wenwu Zhu. 2025. Disentangling Invariant Subgraph via Variance Contrastive Estimation under Distribution Shifts. In *Forty-second International Conference on Machine Learning, ICML*

2025, Vancouver, BC, Canada, July 13-19, 2025 (Proceedings of Machine Learning Research), Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (Eds.). PMLR / OpenReview.net. https://proceedings.mlr.press/v267/li25cv.html

[18] Gang Liu, Jiaxin Xu, Tengfei Luo, and Meng Jiang. 2024. Graph Diffusion Transformers for Multi-Conditional Molecular Generation. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (Eds.). http://papers.nips.cc/paper_files/paper/2024/hash/0f6931a9e339a012a9909306d7c758b4-Abstract-Conference.html

[19] Changsheng Ma and Xiangliang Zhang. 2021. GF-VAE: a flow-based variational autoencoder for molecule generation. In *Proceedings of the 30th ACM international conference on information & knowledge management*. 1181–1190.

[20] Nguyen Nguyen, Nhat Truong Pham, Duong Tran, and Balachandran Manavalan. 2024. Lang2Mol-Diff: A Diffusion-Based Generative Model for Language-to-Molecule Translation Leveraging SELFIES Representation. In *Proceedings of the 1st Workshop on Language+ Molecules (L+ M 2024)*. 128–134.

[21] Chenhao Niu, Yang Song, Jiaming Song, Shengjia Zhao, Aditya Grover, and Stefano Ermon. 2020. Permutation Invariant Graph Generation via Score-Based Generative Modeling. In *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy] (Proceedings of Machine Learning Research, Vol. 108)*, Silvia Chiappa and Roberto Calandra (Eds.). PMLR, 4474–4484. http://proceedings.mlr.press/v108/niu20a.html

[22] Qizhi Pei, Wei Zhang, Jinhua Zhu, Kehan Wu, Kaiyuan Gao, Lijun Wu, Yingce Xia, and Rui Yan. 2023. BioT5: Enriching Cross-modal Integration in Biology with Chemical Knowledge and Natural Language Associations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, 1102–1123. doi:10.18653/V1/2023.EMNLP-MAIN.70

[23] Kristina Preuer, Philipp Renz, Thomas Unterthiner, Sepp Hochreiter, and Günter Klambauer. 2018. Fréchet ChemNet Distance: A Metric for Generative Models for Molecules in Drug Discovery. *J. Chem. Inf. Model.* 58, 9 (2018), 1736–1741. doi:10.1021/ACS.JCIM.8B00234

[24] Yijian Qin, Xin Wang, Ziwei Zhang, Hong Chen, and Wenwu Zhu. 2023. Multi-task Graph Neural Architecture Search with Task-aware Collaboration and Curriculum. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (Eds.). http://papers.nips.cc/paper_files/paper/2023/hash/4e839c9c398c58c878a394633b806ccd-Abstract-Conference.html

[25] Yang Song and Stefano Ermon. 2019. Generative Modeling by Estimating Gradients of the Data Distribution. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 11895–11907. https://proceedings.neurips.cc/paper/2019/hash/3001ef257407d5a371a96dcd947c7d93-Abstract.html

[26] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. https://openreview.net/forum?id=PxTIG12RRHS

[27] Clément Vignac, Igor Krawczuk, Antoine Siraudin, Bohan Wang, Volkan Cevher, and Pascal Frossard. 2023. DiGress: Discrete Denoising diffusion for graph generation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. https://openreview.net/pdf?id=UaAD-Nu86WX

[28] David Weininger. 1988. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* 28, 1 (1988), 31–36. doi:10.1021/CI00057A005

[29] Yuran Xiang, Haiteng Zhao, Chang Ma, and Zhi-Hong Deng. 2024. Instruction-Based Molecular Graph Generation with Unified Text-Graph Diffusion Model. *CoRR* abs/2408.09896 (2024). arXiv:2408.09896 doi:10.48550/ARXIV.2408.09896

[30] Minkai Xu, Alexander S. Powers, Ron O. Dror, Stefano Ermon, and Jure Leskovec. 2023. Geometric Latent Diffusion Models for 3D Molecule Generation. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 38592–38610. https://proceedings.mlr.press/v202/xu23n.html

[31] Zhe Xu, Ruizhong Qiu, Yuzhong Chen, Huiyuan Chen, Xiran Fan, Menghai Pan, Zhichen Zeng, Mahashweta Das, and Hanghang Tong. 2024. Discrete-state Continuous-time Diffusion for Graph Generation. *CoRR* abs/2405.11416 (2024). arXiv:2405.11416 doi:10.48550/ARXIV.2405.11416

[32] Yang Yao, Xin Wang, Yijian Qin, Ziwei Zhang, Wenwu Zhu, and Hong Mei. 2024. Data-Augmented Curriculum Graph Neural Architecture Search under Distribution Shifts. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (Eds.). AAAI Press, 16433–16441. doi:10.1609/AAAI.V38I15.29580

[33] Seongjun Yun, Minbyul Jeong, Raehyun Kim, Jaewoo Kang, and Hyunwoo J. Kim. 2019. Graph Transformer Networks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 11960–11970. https://proceedings.neurips.cc/paper/2019/hash/9d63484abb477c97640154d40595a3bb-Abstract.html

[34] Huaisheng Zhu, Teng Xiao, and Vasant G Honavar. 2024. 3M-Diffusion: Latent Multi-Modal Diffusion for Language-Guided Molecular Structure Generation. In *First Conference on Language Modeling*.

A More Experiment Results

More Ablation Study. We provide additional ablation study in Table 6 and Table 7.

From Table 6, it can be observed that the quality of generated molecules is lower when the number of diffusion steps is reduced, with the effect being more pronounced when the number of steps is very small, matching the general phenomenon in diffusion models.

Table 6: Comparison of different diffusion steps

Steps	MACCS FTS↑	RDK FTS↑	Morgan FTS↑
100	0.750	0.580	0.501
250	0.882	0.771	0.728
500	**0.892**	**0.794**	**0.757**

From Table 7, we investigate the effect of different CFG scale, with cfg = 1.0 meaning no classifier-free guidance. Using classifier guidance (cfg = 1.5) leads to improved performance over not using it (cfg = 1.0), but a high CFG scale reduces the model performance.

Table 7: Comparison of different CFG scale

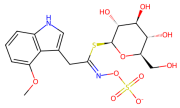
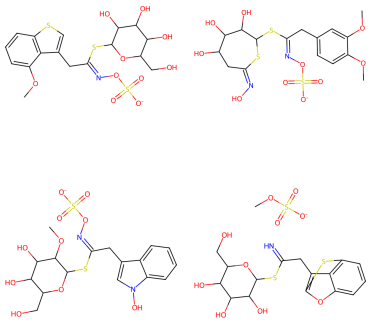
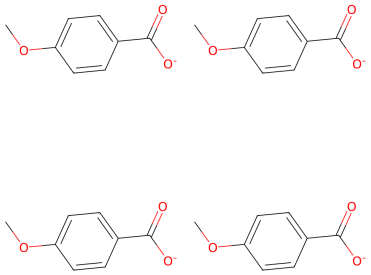
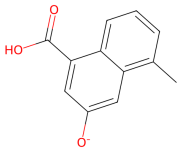
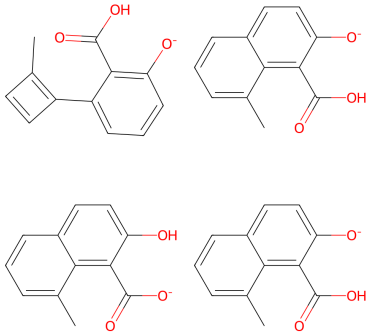
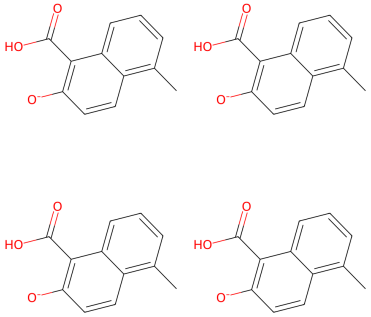
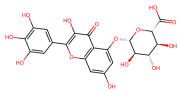
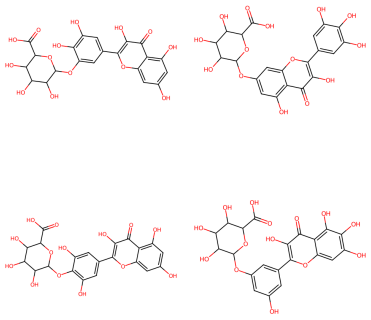
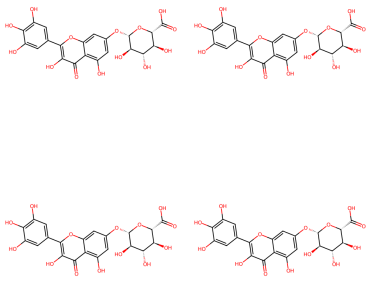
cfg	MACCS FTS↑	RDK FTS↑	Morgan FTS↑
1.0	0.870	0.764	0.723
1.5	**0.892**	**0.794**	**0.757**
2.0	0.879	0.789	0.744

Visualization of Generated Molecules. In order to better illustrate the quality of generated graphs of our method, we list in Table 8 some comparison results with MolT5. It can be seen from the table that our method can generate diverse results while maintaining the basic molecular structure, while the diversity and accuracy of MolT5 is generally worse. This indicates that our method has more potential in the application of text-to-molecule generation.

B Experimental Setting

Our code is based on DiGress, with modifications to allow text-guided graph generation. The information of the dataset is shown in Table 9. All experiments are performed with six NVIDIA RTX

Table 8: Visualization of generated molecules by our method and MolT5.

Text description	Ground truth	Our method	MolT5
The molecule is an indolylmethylglucosinolate that is the conjugate base of 4-methoxyglucobrassicin, obtained by deprotonation of the sulfo group. It is a conjugate base of a 4-methoxyglucobrassicin.			
The molecule is a member of the class of naphthoates that is 1-naphthoate substituted at positions 3 and 5 by hydroxy and methyl groups respectively; major species at pH 7.3. It has a role as a bacterial metabolite. It is a conjugate base of a 3-hydroxy-5-methyl-1-naphthoic acid.			
The molecule is a myricetin O-glucuronide that is myricetin with a beta-D-glucosiduronic acid residue attached at the 5-position. It has a role as a metabolite. It is a myricetin O-glucuronide, a pentahydroxyflavone, a member of flavonols and a monosaccharide derivative.			

4090 GPUs. Training our diffusion model takes about 2 days on ChEBI-20 and 7 days on L+M-24 datasets.

Table 9: Dataset info

Dataset	Total number	Training set	Validation set	Test set
ChEBI-20	33,010	26,407	3,301	3,300
L+M-24	200,615	126,864	33,696	21,805

Hyperparameters. The model training parameters on the ChEBI-20 dataset are shown in Table 10. The model training parameters on the L+M-24 dataset are shown in Table 11. We determine the number of layers and the dimension sizes, as well as λ_X and λ_E by following DiGress. λ_C is set to the same as λ_X as they both denote the weighting of node-level loss values. The batch size set to the largest value that fits in VRAM for our GPU. The unconditional training probability is set to 0.2, a common value in diffusion models for images, and was not specifically tuned.

Table 10: Hyperparameter setting on ChEBI-20 dataset

Hyperparameter	Value
epoch	1000
batch size	76
learning rate	2e-4
unconditional probability	0.2
λ_X	1
λ_C	1
λ_E	5
num of layers	9
hidden mlp dims	[X': 256, 'E': 128, 'y': 256]
hidden dims	['dx': 256, 'de': 64, 'dy': 128, 'n_head': 8, 'dim_ffX': 256, 'dim_ffE': 128, 'dim_ffy': 256]
text feature dims	256

Table 11: Hyperparameter setting on L+M-24 dataset

Hyperparameter	Value
epoch	500
batch size	22
learning rate	2e-4
unconditional probability	0.2
λ_X	1
λ_C	1
λ_E	5
num of layers	9
hidden mlp dims	[X': 256, 'E': 128, 'y': 256]
hidden dims	['dx': 256, 'de': 64, 'dy': 128, 'n_head': 8, 'dim_ffX': 256, 'dim_ffE': 128, 'dim_ffy': 256]
text feature dims	256

C Computational Costs

Compared to unconditional graph diffusion models, our method only requires a small increase in training and inference time. For inference, our method, with the addition of text feature extraction and structure-aware cross-attention, only increased the inference time by about 30%.