NGTM: Substructure-based Neural Graph Topic Model for Interpretable Graph Generation

Yuanxin Zhuang[1], **Dazhong Shen**[2], **Ying Sun**[1*]

[1] Artificial Intelligence Thrust, Hong Kong University of Science and Technology (Guangzhou)
[2] College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics
yzhuang436@connect.hkust-gz.edu.cn, shendazhong@nuaa.edu.cn, sunyinggilly@gmail.com

Abstract

Graph generation plays a pivotal role across numerous domains, including molecular design and knowledge graph construction. Although existing methods achieve considerable success in generating realistic graphs, their interpretability remains limited, often obscuring the rationale behind structural decisions. To address this challenge, we propose the Neural Graph Topic Model (NGTM), a novel generative framework inspired by topic modeling in natural language processing. NGTM represents graphs as mixtures of latent topics, each defining a distribution over semantically meaningful substructures, which facilitates explicit interpretability at both local and global scales. The generation process transparently integrates these topic distributions with a global structural variable, enabling clear semantic tracing of each generated graph. Experiments demonstrate that NGTM achieves competitive generation quality while uniquely enabling fine-grained control and interpretability, allowing users to tune structural features or induce biological properties through topic-level adjustments.

Introduction

Graph generation is a fundamental task with diverse applications—including molecular design (Tian et al. 2025) and relational data modeling (Sun et al. 2019, 2024). While generating valid and accurate graphs is essential, understanding the underlying generative process and explaining structure formation is equally crucial (Kumar et al. 2025). This significance is particularly evident in high-stakes applications like drug discovery (Zheng et al. 2024), where generation process interpretability builds trust, enables rational design, and supports downstream scientific analysis (Sun et al. 2021; Sun, Zhu, and Xiong 2025; Ji et al. 2025; Sun et al. 2025).

Despite its importance, interpretability remains a major challenge in existing graph generation models (Zhang et al. 2023; Liu et al. 2023). Sequential methods, such as the auto-regressive model Graph Generative Pre-trained Transformer (G2PT) (Chen et al. 2025) and the reinforcement learning-based model ExSelfRL (Wang and Zhu 2025), generate graphs step-by-step by incrementally adding nodes and edges. While their procedures are explicit, the underlying decisions depend heavily on hidden states, making semantic interpretation difficult. In contrast, models based on VAEs like DAVA (Hou et al. 2024) and GANs like ConfGAN (Xu et al. 2025) generate entire graphs from latent representations in a single forward pass, while diffusion-based models like ConStruct (Madeira et al. 2024) iteratively refine graphs from random noise. Although these methods achieve strong performance and capture complex distributions, their latent spaces are often opaque and unaligned with interpretable graph components. As a result, they struggle to provide insights into the reasoning behind generated graphs.

Graphs are often composed of key substructures that define their topology and functionality (Kengkanna and Ohue 2024). For example, molecular graphs contain recurring building blocks like functional groups or rings, which are essential to their structural diversity and chemical behavior (Jin, Barzilay, and Jaakkola 2018). Explicitly modeling these substructures improves the quality and diversity of generated graphs while uncovering hidden patterns that reveal new domain knowledge (Kong et al. 2022). Additionally, this approach enhances generation transparency, enables precise control over substructure inclusion, and ensures generated graphs align with real-world characteristics.

However, achieving interpretable graph generation remains challenging. Post hoc methods analyze latent features or learned representations after generation to identify meaningful patterns in graphs (Guo et al. 2020). While useful, these approaches do not influence the generative process itself, leaving the reasoning behind structural decisions untraceable. Some generative models construct molecular graphs from substructures (Kong et al. 2022). Although effective in improving validity and modularity, these methods have two key limitations: (1) they often struggle to automatically discover and organize previously unseen substructures into coherent semantic units; and (2) they often lack a unified, transparent narrative of the generation process, making it difficult to trace the relationships and contributions of individual substructures to the overall graph. Moreover, such reliance on predefined substructures can constrain generative flexibility and hinder the discovery of novel structures.

To address these challenges, we propose the Neural Graph Topic Model (NGTM), a novel interpretable generative framework inspired by topic modeling techniques in natural language processing (Blei, Ng, and Jordan 2003). NGTM

*Corresponding author.

models graphs as mixtures of latent topics, with each topic corresponding to a distinct distribution over interpretable graph substructures (e.g., rings, motifs, functional groups in molecular). By explicitly extracting reusable structural units and organizing them into semantically coherent topics, NGTM achieves clear semantic organization in graph generation. Based on the topics, the generation process in NGTM is transparent and traceable, comprising three clear steps: (1) sampling a topic mixture to define the semantic profile of the target graph; (2) sampling relevant substructures based on these sampled topics; and (3) assembling these substructures under learned structural constraints to form a coherent graph, thus integrating both local interpretability (via explicit substructures) and global interpretability (via topic mixtures) in a unified and transparent way.

Experiments reveal that NGTM not only achieves competitive generation quality but, more importantly, pioneers a new level of interpretability and controllability in graph generation. The learned topics align robustly with meaningful structural and functional classes, providing a clear semantic basis. By adjusting these topics, NGTM enables continuous and fine-grained control over key graph properties, allowing the generation process to be precisely steered toward desired outcomes. Our contributions are summarized as follows: (1) We propose the first transparent and interpretable graph generation framework applicable to general graph structures. (2) We introduce topic modeling principles to graph generation, enabling interpretable structural modeling. (3) Experiments show NGTM generates realistic graphs while providing clear traceability of the generation process.

Related Work

Graph Generation: Over the past decade, significant efforts have been dedicated to advancing graph generation techniques (Gong et al. 2025; Minello et al. 2024). These models have provided robust frameworks for generating realistic and diverse graphs. However, many of these methods function as black boxes, making it difficult to understand or interpret how specific graph structures are generated. Existing efforts toward interpretability in graph generation have primarily focused on disentanglement-based methods (Guo et al. 2020; Li et al. 2020; Du et al. 2022), which aim to learn latent factors aligned with controllable aspects of graphs, such as node or edge attributes. These approaches provide insight into global properties but often fail to explain the step-by-step structure-building process. In parallel, some methods explore substructure-based generation. JT-VAE (Jin, Barzilay, and Jaakkola 2018) assembles molecules from chemically valid fragments using junction trees, while PS-VAE (Kong et al. 2022) leverages frequently occurring subgraphs as building blocks. Although these methods enhance generation quality and modularity, they are not designed for interpretability and do not explain why or how substructures are selected and composed. In contrast, NGTM uniquely introduces a topic modeling perspective that organizes substructures into semantically meaningful topics. This provides not only local interpretability through explicit substructures but also global coherence through interpretable topic distributions. Our model offers a transparent generative narrative: each graph is generated by first sampling topic mixtures (explaining the "why"), then selecting substructures governed by these topics (explaining the "what"), and finally assembling them under structural guidance (explaining the "how").

Topic Modeling and Neural Topic Models: Topic modeling is a foundational NLP technique that uncovers latent semantic structures in text corpora (Shen et al. 2021a,b). Classical models like LDA (Blei, Ng, and Jordan 2003) treat documents as mixtures of topics, where each topic represents a distribution over words (Shen et al. 2018). Neural topic models, including NVDM (Miao, Yu, and Blunsom 2016) and contextualized topic models (Bianchi, Terragni, and Hovy 2020), extend these approaches using neural networks to learn more flexible topic representations. While widely adopted in textual domains, topic modeling has been underexplored for graph generation. Most prior work focuses on community detection or node classification (Xie et al. 2021) rather than generation. We argue that graphs exhibit naturally compositional structures—substructures like rings and chains resemble "words" while entire graphs resemble "documents" composed from semantic components. NGTM framework adopts this analogy by modeling graphs as mixtures of latent substructure topics, enabling both structural quality and interpretability in generation. To our knowledge, NGTM is the first to integrate topic modeling principles into graph generation.

Neural Graph Topic Model

In this section, we begin by outlining the NGTM generation process, followed by a description of the model architecture and the substructure assembly process. Finally, we detail the inference procedure and the loss function.

Substructure-based Graph Generation Process

Graphs can be viewed as compositions of structural motifs or substructures, each originating from a semantically meaningful latent topic. For example, molecular graphs typically contain interpretable substructures like rings, chains, and functional groups that frequently co-occur. To explicitly capture this structural diversity, we propose NGTM, a novel and interpretable generative approach that models graph structures as assemblies of substructures sampled from a set of latent topics. Specifically, we assume there are K latent topics, each topic defines a Gaussian distribution over latent embeddings of substructures, denoted as $\Phi = \mathcal{N}(\mu^k, \sigma^k)_{k=1}^{K}$. These distributions are parameterized and learned through a Conditional Variational Autoencoder (CVAE) (Sohn, Lee, and Yan 2015), with topics serving as conditioning variables during training. Such a design enables the generation of diverse and potentially novel substructures without relying on a predefined substructure vocabulary. Latent vectors sampled from these topic-specific distributions are decoded into adjacency matrices representing meaningful subgraphs. To govern the semantic composition, we introduce a latent variable drawn from $\mathcal{N}(\mu^\theta, \sigma^\theta)$, transformed via softmax into a topic proportion vector θ. This vector directs the topic selection for substructure generation, ensur-

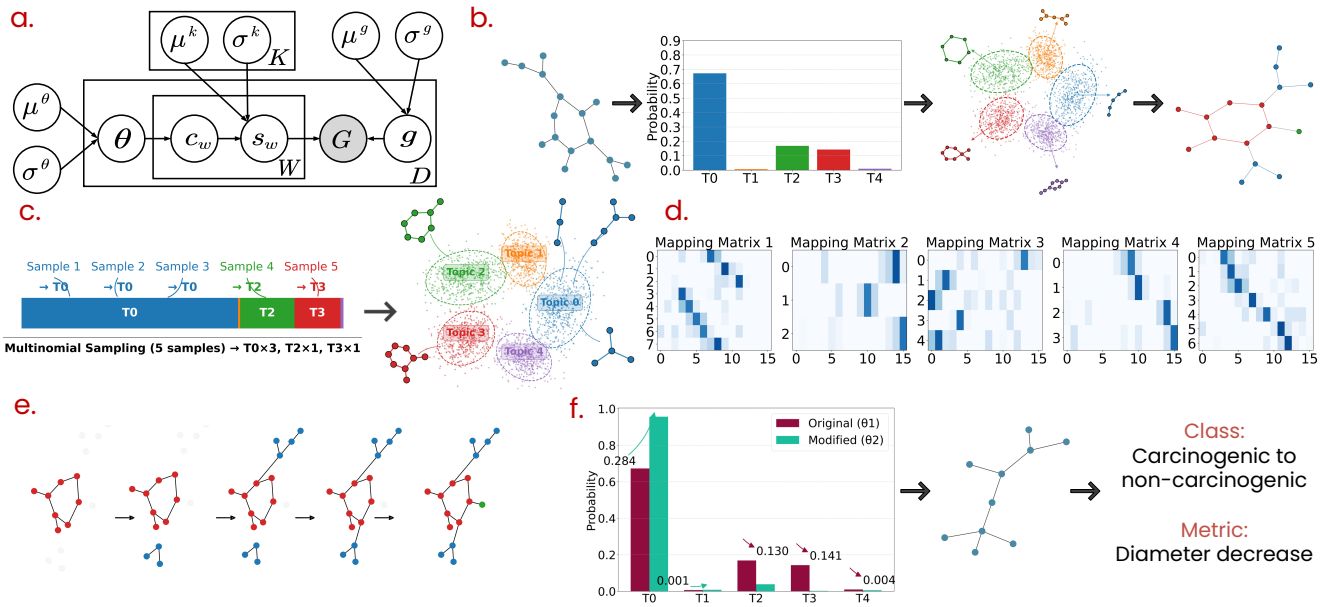


Figure 1: Overview of the NGTM framework for interpretable graph generation. (a) The probabilistic graphical model of NGTM. (b) Training phase: NGTM infers topic mixtures from real graphs and discovers semantically meaningful substructure topics, enabling reconstruction through interpretable latent factors. (c–e) Generation phase: (c) Multinomial sampling from θ assigns each substructure to a topic, and the corresponding latent vectors are decoded into interpretable substructures. (d) Mapping matrices determine how substructures are softly aligned and integrated into the growing graph. (e) Visualization of the sequential assembly process guided by global structure vectors. (f) Example of controllable generation: adjusting topic proportions modifies graph semantics.

ing semantic coherence in the resulting graph. Additionally, we introduce a global structural variable $g \sim \mathcal{N}(\mu^g, \sigma^g)$ to capture high-level topological properties such as density, connectivity, and overall layout. This global guidance ensures that the final assembled graph is coherent, realistic, and structurally sound. The NGTM generative framework is outlined in Fig. 1, where Fig. 1(a) shows the probabilistic graphical representation detailing the relationships between latent variables involved in the generation process. Fig. 1(b) illustrates the training phase, during which NGTM learns topic mixtures from observed graphs and organizes substructures into semantically coherent topics to support interpretable reconstruction. Fig. 1(c) to (e) depict the generation process: (c) shows how topic assignments are first drawn from the sampled topic proportion vector θ, and corresponding substructures are generated by sampling latent vectors from the assigned topic-specific distributions. (d) depicts the construction of soft mapping matrices for each substructure under the guidance of the global structural vector g, determining how substructures are integrated into the current graph. (e) visualizes the step-by-step assembly of the full graph by sequentially merging substructures according to the learned mappings and global structural constraints. Fig. 1(f) demonstrates controllable generation, where adjusting topic weights results in interpretable changes to graph-level properties, such as reducing diameter or altering predicted class labels. The full generative process of NGTM can be summarized in the following three stages:

1. Global and semantic initialization:

(a) Sample latent semantic vector $z^\theta \sim \mathcal{N}(\mu^\theta, \sigma^\theta)$ and derive topic proportion vector $\theta = \text{softmax}(z^\theta)$. (b) Sample global structural guidance vector $g \sim \mathcal{N}(\mu^g, \sigma^g)$.

2. Substructure generation loop (for $w = 1, \ldots, W$):

(a) Select topic $c_w \sim \text{Multinomial}(\theta)$. (b) Generate latent substructure vector $z_w \sim \mathcal{N}(\mu^{k=c_w}, \sigma^{k=c_w})$ and decode substructure s_w.

3. Final assembly: Combine substructures and global guidance: $G = f(s_1, s_2, \ldots, s_W, g)$.

NGTM Architecture

The NGTM framework is designed to learn topic distributions, substructure distributions, and a global structural prior from training data, while also learning how to assemble substructures into complete graphs. As illustrated in Fig. 2, NGTM consists of several key components: the Topic Encoder, Structure Encoder, Global Encoder, Structure Decoder, Node Position Encoder, and Mapping Network.

Encoder Module: NGTM uses three encoders to extract topic semantics, substructure-level patterns, and global structure context. The Topic Encoder and Global Encoder process the full graph G and output the parameters of corresponding Gaussian distributions via MLPs:

$$\mu_z^\theta, \sigma_z^\theta = \text{MLP}(\text{Enc}_{\text{Topic}}(G)), \mu_z^g, \sigma_z^g = \text{MLP}(\text{Enc}_{\text{Global}}(G)).$$

To model topic-specific substructure priors, NGTM employs Conditional VAEs (Sønderby et al. 2016). Each graph G is

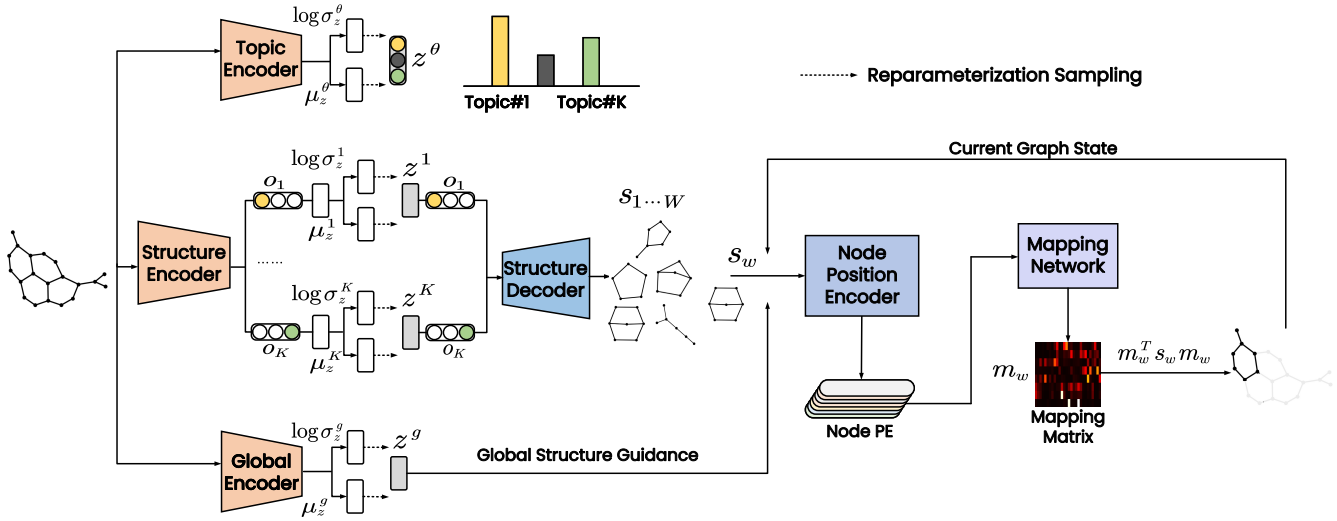


Figure 2: The Overview of NGTM Architecture.

encoded by the Structure Encoder and concatenated with a one-hot topic vector o_k. The result is passed through MLPs to produce: $\mu_z^k, \sigma_z^k = \text{MLP}(\text{Enc}_{\text{Structure}}(G) \oplus o_k)$. Here, the topic vector acts as a control signal, guiding the model to learn distinct distributions over structural motifs for each topic. The reparameterization trick (Kingma and Welling 2013) enables gradient-based optimization.

Decoder Module: Each sampled substructure vector $z^{k=c_w}$ is concatenated with its topic one-hot vector and decoded as follows: $s_w = \text{SD}(z^{k=c_w} \oplus o_{k=c_w})$. This conditioning ensures that each generated substructure is semantically linked to a specific topic, enhancing interpretability.

Substructure Assembly Module: Each substructure s_w consists of up to n nodes and is represented by an adjacency matrix A_{s_w}. The graph is constructed by iteratively assembling W substructures. At each step w, the current graph state G_{w-1}, the substructure s_w, and the global structure vector g are input into the Node Position Encoder to produce a Node Position Encoding : $p_w = \text{NPE}(s_w, G_{w-1}, g)$. This encoding is fed into the Mapping Network to generate a softmax-normalized mapping matrix: $m_w = \text{Softmax}(\text{MN}(p_w))$. The matrix m_w softly aligns nodes in s_w with positions in the current graph. The adjacency matrix is updated as: $A_w = A_{w-1} + m_w^T A_{s_w} m_w$. This soft assignment enables flexible integration of substructures while preserving a traceable correspondence between each motif and its position in the graph, reinforcing transparency throughout the generation process.

Inference and Optimization

We introduce latent variables $Z = \{z^\theta, z^1, \ldots, z^K, z^g\}$ to capture various aspects of the graph G. These latent variables are sampled from the distributions $\mathcal{N}(\mu_z^\theta, \sigma_z^\theta)$, $\mathcal{N}(\mu_z^k, \sigma_z^k)_{k=1}^K$ and $\mathcal{N}(\mu_z^g, \sigma_z^g)$, which are learned using the NGTM framework with parameters ϕ. To optimize the parameters ϕ, we employ variational inference (Blei, Kucukelbir, and McAuliffe 2017), which minimizes the Kullback-Leibler (KL) divergence. This is equivalent to maximizing the Evidence Lower Bound (ELBO):

$$\log p(G) \geq \text{ELBO} = \mathbb{E}_{q_\phi}\left[\log p(G \mid Z)\right] - D_{\text{KL}}[q_\phi(Z|G)||p(Z)]. \tag{1}$$

In this equation, $\mathbb{E}_{q_\phi}[\log p(G \mid Z)]$ represents the reconstruction term, assessing how effectively the latent variable Z reconstructs the graph G. The term $D_{\text{KL}}[q_\phi(Z|G)||p(Z)]$ measures the divergence between the variational approximation and the true prior distribution. We assume the latent variables $Z = \{z^\theta, z^1, \ldots, z^K, z^g\}$ are independent. Consequently, the variational posterior can be factorized as $q_\phi(Z \mid G) = q_\phi(z^\theta \mid G) \prod_{k=1}^K q_\phi(z^k \mid G) q_\phi(z^g \mid G)$. Similarly, the prior distribution can be factored as $p(Z) = p(z^\theta) \prod_{k=1}^K p(z^k) p(z^g)$. The KL divergence is then calculated as:

$$\begin{aligned} D_{\text{KL}}[q_\phi(Z|G)||p(Z)] &= D_{\text{KL}}[q_\phi(z^\theta|G)||p(z^\theta)] + \\ &\sum_{k=1}^K D_{\text{KL}}[q_\phi(z^k|G)||p(z^k)] + D_{\text{KL}}[q_\phi(z^g|G)||p(z^g)]. \end{aligned} \tag{2}$$

Loss Function Combining Equations 1 and 2, the loss function of NGTM consists of a reconstruction loss and $(K+2)$ KL divergence losses. We set the prior for both topic distribution and global structure distribution to a standard normal distribution $\mathcal{N}(0, I)$, so $\mathcal{L}_{\text{KL}}^\theta = D_{\text{KL}}[\mathcal{N}(\mu_z^\theta, \sigma_z^\theta) \parallel \mathcal{N}(0, I)]$ and $\mathcal{L}_{\text{KL}}^g = D_{\text{KL}}[\mathcal{N}(\mu_z^g, \sigma_z^g) \parallel \mathcal{N}(0, I)]$. The prior for the K substructure distributions is defined by the learnable parameters μ^k and σ^k. Each substructure distribution is weighted by the K-dimensional vector z^θ, resulting in the following KL divergence term: $\mathcal{L}_{\text{KL}}^K = \sum_{k=1}^K z_k^\theta D_{\text{KL}}[\mathcal{N}(\mu_z^k, \sigma_z^k) \parallel \mathcal{N}(\mu^k, \sigma^k)]$. For the reconstruction loss, we employ the micro-macro loss (Zahirnia et al. 2022). The micro loss captures local properties of the training graph, such as node-level or pairwise properties. While the macro loss captures higher-order graph properties represented by graph statistics. These statistics include

the distribution of node degrees (the number of connections each node has), the number of triangles (specific patterns of three interconnected nodes), and s-step transition probabilities (the likelihood of reaching one node from another in a given number of steps). This loss method reduces dependence on specific node ordering by utilizing permutation-invariant graph statistics. Additionally, we apply orthogonality regularization on the means of the substructure distributions: $\mathcal{L}_{\text{ortho}} = \sum_{i=1}^{K} \sum_{j=1, j\neq i}^{K} ((\mu_z^{k_i})^\top \mu_z^{k_j})^2$. This ensures that the K topic distributions are as distinct from each other as possible. NGTM is trained using a weighted sum of the aforementioned loss functions.

Experiment

Baselines

We compare our approach against a range of state-of-the-art graph generation models from different generative paradigms: **(1) GraphVAE**(Simonovsky and Komodakis 2018), **(2) GraphVAE-MM**(Zahirnia et al. 2022), **(3) GraphRNN**(You et al. 2018), **(4) GRAN**(Liao et al. 2019), **(5) BiGG**(Dai et al. 2020), **(6) DiGress**(Vignac et al. 2022), **(7) G2PT**(Chen et al. 2025), and **(8) ConStruct**(Madeira et al. 2024). Detailed descriptions of these baselines are provided in the Appendix. **(9) NGTM Variants:** We further compare against ablated versions of our NGTM model: (a) **NGTMw/oGE**: This variant removes the global encoder, assessing the importance of global structural guidance. (b) **NGTMParallel**: In this version, the Substructure Assembly Module performs simultaneous aggregation of all substructures rather than sequential composition.

Datasets

We utilize one synthetic dataset and three real-world datasets: **(1) Lobster** (Golomb 1996): This dataset includes 100 synthetic graphs with $10 \leq |V| \leq 100$. The average number of nodes is 52. These graphs are trees where each node is at most 2 hops away from a backbone path. **(2) MUTAG** (Debnath et al. 1991): MUTAG is a dataset with $10 \leq |V| \leq 28$ comprising 188 mutagenic aromatic and heteroaromatic nitro compounds. The average number of nodes is 18. **(3) PTC** (Toivonen et al. 2003; Xu et al. 2018): PTC is a dataset of 344 chemical compounds with $4 \leq |V| \leq 103$ that report the carcinogenicity of male and female rats. The average number of nodes is 26. **(4) Ogbg-molbbbp** (Hu et al. 2020): This dataset consists of 2039 real-world molecular graphs with $2 \leq |V| \leq 132$. The average number of nodes is 23. **Train/Test Split.** We follow the same protocol as (Dai et al. 2020; Zahirnia et al. 2022) and create random 80% and 20% splits of the graphs for training and testing, respectively. Additionally, 20% of the training data in each split is used as the validation set.

Implementation Details

Training is performed using the Adam optimizer with a learning rate of 0.0003. The model is trained for up to 20,000 epochs, and the version with the best validation performance is selected for testing. We fix the number of topics $K = 10$, the number of sampled substructures $W = 30$, and set the substructure size n to the average number of nodes in each dataset. We fix the number of topics $K = 10$, the number of sampled substructures $W = 30$, and set the substructure size n to the average number of nodes in each dataset. Although these hyperparameters may not yield the best possible reconstruction accuracy, they provide a balance between interpretability and generation quality in our experiments. The NGTM framework is composed of modular components: a topic encoder, structure encoder, and global encoder—each implemented with four GCN layers, followed by Layer Normalization and average pooling. The structure decoder consists of three linear layers. For substructure assembly, we employ three MultiheadAttention modules to compute attention between the substructure, the current graph, and the global structure vector. The mapping network includes three linear layers with LayerNorm. All hidden dimensions are set to 256.

Evaluation Metrics

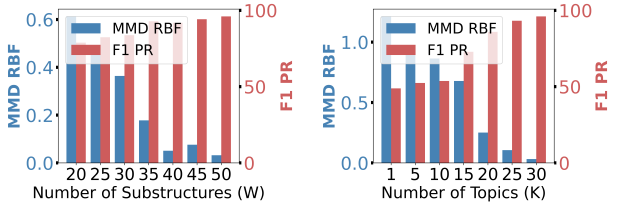
We evaluate the effectiveness of NGTM through both qualitative and quantitative metrics to comprehensively assess the diversity and realism of the generated graphs. We use two types of metrics to measure the distributional distance between generated and test graphs: (a) **GNN-based Metrics:** These metrics utilize a task-agnostic Graph Neural Network (Random-GNN) (Thompson et al. 2022) to extract graph representations. We evaluate the discrepancies using F1 PR, which measures the diversity of the generated graphs, and MMD RBF, which assesses their realism based on structural and semantic alignment. (b) **Statistics-based Metrics:** This approach directly compares structural statistics such as degree distributions, orbit counts, clustering coefficients, spectral features, and graph diameter (Liao et al. 2019; Zahirnia et al. 2022). Following O'Bray et al. (O'Bray et al. 2021), we also report ideal scores obtained from a 50/50 data split, which serve as a lower bound for distributional distance. These combined metrics provide a comprehensive view of the model's performance.

Generation Quality Evaluation

We conducted experiments comparing NGTM with several graph generation models across four representative datasets, spanning synthetic, chemical, and bioinformatics domains. As shown in Table 1, NGTM consistently achieves results that rival or outperform the strongest baselines, despite explicitly emphasizing interpretability—a quality often seen as coming at the cost of generation fidelity. Notably, our evaluations show that NGTM is capable of robustly capturing both local structural motifs and global topological characteristics, essential for realistic graph synthesis. This is particularly significant in chemically relevant scenarios such as MUTAG, PTC, and Ogbg-molbbbp datasets, where generating accurate molecular structures directly impacts real-world outcomes, including drug design and toxicity prediction. NGTM-generated graphs exhibit clear and consistent chemical substructures that align well with known domain knowledge, demonstrating its strong practical potential in scientific contexts.

Method	MUTAG							Lobster						
	Deg. ↓	Clus. ↓	Orbit ↓	Spect ↓	Diam. ↓	MMD ↓	F1 ↑	Deg. ↓	Clus. ↓	Orbit ↓	Spect ↓	Diam. ↓	MMD ↓	F1 ↑
50/50 split	3e-4	0	2e-5	0.007	0.002	0.027	97.96	8e-4	0	0.003	0.004	0.023	0.04	98.99
GraphRNN	0.007	0.266	0.001	0.070	0.728	0.832	52.81	0.004	0	0.039	0.044	0.376	0.855	62.33
GRAN	5e-4	0.011	0.005	0.059	0.645	0.276	90.87	0.006	0.475	0.412	0.039	0.502	0.221	45.39
BiGG	0.006	0	0.003	0.043	0.308	0.521	97.93	5e-4	0	**0.001**	0.017	0.009	0.127	95.19
GraphVAE	0.006	0.122	0.007	0.024	0.051	0.109	69.79	0.094	0.866	0.405	0.071	0.137	0.415	69.54
GraphVAE-MM	0.001	0	5e-4	0.017	0.021	0.061	89.58	4e-4	0	0.008	0.019	0.175	0.136	100
DiGress	0.002	0.015	7e-4	0.020	0.060	0.085	95.23	0.001	0.045	0.006	0.030	0.038	0.148	90.44
G2PT	0.004	0.052	0.003	0.029	0.087	0.132	97.66	0.003	0.076	0.005	0.034	0.051	0.193	95.12
ConStruct	0.001	0	0.001	0.018	0.020	0.059	94.12	**2e-4**	0	0.004	0.015	0.010	**0.097**	91.36
$\text{NGTM}_{\text{w/oGE}}$	0.005	0.041	9e-4	0.023	0.079	0.082	96.73	0.002	0.071	0.009	0.035	0.027	0.139	98.75
$\text{NGTM}_{\text{Parallel}}$	0.006	0	4e-4	0.018	0.050	0.068	98.67	0.003	0.021	0.012	0.016	0.017	0.159	100
NGTM	**4e-4**	**0**	**3e-4**	**0.015**	**0.015**	**0.053**	**98.72**	3e-4	**0**	0.003	**0.012**	**0.007**	0.102	**100**
	PTC							**Ogbg-molbbbp**						
50/50 split	5e-5	3e-4	7e-5	0.003	0.008	0.022	96.73	5e-4	4e-5	8e-5	8e-4	7e-4	0.001	97.98
GraphRNN	0.007	0.004	0.006	0.071	0.417	0.816	33.15	0.003	0.001	9e-4	0.126	0.553	1.361	94.73
GRAN	0.020	0.115	0.009	0.038	0.179	0.164	78.84	0.006	0.376	0.011	0.044	0.359	0.354	93.96
BiGG	**3e-4**	0.004	**9e-5**	0.019	0.027	0.056	95.89	0.002	0.001	9e-4	0.009	0.031	0.059	96.22
GraphVAE	0.223	0.729	0.537	0.035	0.204	0.613	42.31	0.031	0.562	0.048	0.019	0.061	0.313	57.38
GraphVAE-MM	0.033	5e-4	0.002	0.022	0.038	0.055	82.39	0.002	0.007	9e-4	**0.004**	0.031	0.047	92.96
DiGress	0.012	0.008	0.007	0.026	0.098	0.113	89.85	0.003	0.005	0.001	0.010	0.042	0.073	93.67
G2PT	0.022	0.016	0.010	0.033	0.079	0.104	93.70	0.002	0.008	6e-4	0.013	0.048	0.092	**96.85**
ConStruct	0.005	0.001	0.003	0.021	0.032	0.051	94.73	0.001	**5e-4**	0.001	0.006	0.030	0.052	92.18
$\text{NGTM}_{\text{w/oGE}}$	0.029	0.008	0.027	0.053	0.142	0.178	90.82	0.004	0.007	9e-4	0.033	0.097	0.271	85.39
$\text{NGTM}_{\text{Parallel}}$	0.056	7e-4	0.014	0.067	0.398	0.367	83.71	0.001	0.009	7e-4	0.026	0.245	0.289	79.24
NGTM	0.002	**2e-4**	0.003	**0.017**	**0.022**	**0.046**	**96.03**	**0.001**	0.004	**5e-4**	0.008	**0.029**	**0.032**	96.38

Table 1: Comparison of NGTM and baselines.



(a) W (b) K

Figure 3: Impact of parameters on the PTC dataset. (a) Effect of substructure count (W). (b) Effect of topic count (K).

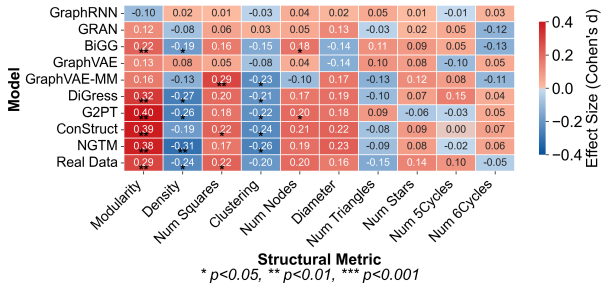


Figure 4: Comparison of Structural Metric Effect Sizes Between Real and Generated Graphs across Models.

Further analysis through ablation studies highlights crucial insights into NGTM's design. Removing the global structural guidance module ($NGTM_{\text{w/oGE}}$) markedly degraded the coherence and realism of generated graphs. This observation clearly illustrates that maintaining high-level global structure constraints significantly enhances the model's ability to produce realistic and semantically coherent graphs. Similarly, replacing sequential substructure assembly with parallel integration ($NGTM_{\text{Parallel}}$) adversely affected structural consistency, suggesting that our sequential assembling of semantic components has been critical for preserving complex structural dependencies.

Parameter Sensitivity

Fig. 3 illustrates how varying key parameters affects graph generation performance on the PTC dataset. The results reveal consistent trends across both diversity (F1 PR) and realism (MMD RBF) metrics. (a) Substructure Count W: Increasing the number of sampled substructures significantly enhances both graph realism and diversity. MMD RBF decreases sharply while F1 PR approaches saturation as W increases from 20 to 50. This suggests that denser substructure sampling enables NGTM to better capture nuanced structural variations. (b) Topic Count K: Expanding the number of latent topics from 1 to 30 results in a marked improvement in performance. As K grows, the model gains greater flexibility in organizing substructures into semantically coherent patterns. These trends collectively indicate that a careful balance of these parameters is essential not only for optimizing the expressiveness and quality of the generated graphs, but also for maintaining a trade-off between performance and interpretability.

Class-wise Structural Analysis

To evaluate whether NGTM captures semantically meaningful structural differences, we examine how well it distinguishes carcinogenic from non-carcinogenic graphs using

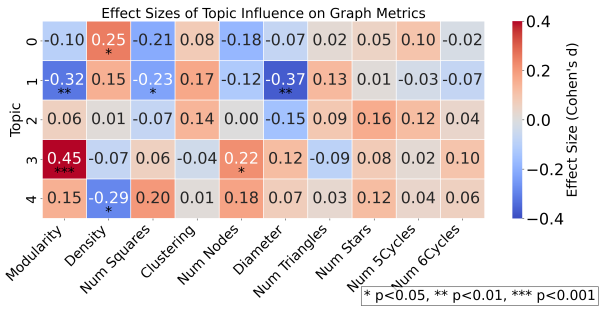
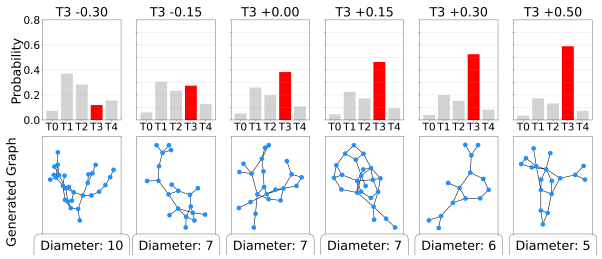


Figure 5: Topic-wise impacts on structural metrics evaluated through systematic topic weight manipulation.

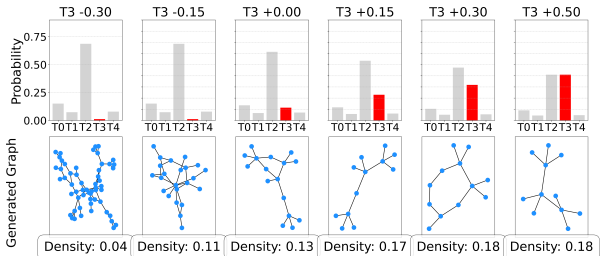
the PTC dataset. We analyze ten structural metrics including modularity, density, clustering coefficient, diameter, and cycle counts, which are critical for understanding molecular carcinogenic behavior (Gonzalez, Holder, and Cook 2001; Swamidass et al. 2005). We compute Cohen's d effect sizes for each metric to quantify class separation: larger absolute values indicate stronger differentiation between carcinogenic and non-carcinogenic graphs. Fig. 4 shows the results across all models. Real PTC data (bottom row) exhibits strong structural differentiation, with carcinogenic graphs showing higher modularity and more 4-cycles but lower density and clustering—consistent with known biological patterns (Blinova et al. 2003). Among all generative models, NGTM best replicates the effect size patterns observed in real data, closely matching both direction and magnitude of class-specific structural trends. NGTM is the only model that consistently captures the full profile of structural effects, including significant separation on key metrics like modularity, density, and 4-cycles. This demonstrates that NGTM not only generates realistic graphs but also preserves meaningful structural variations associated with biological classes, supporting the model's interpretability and semantic alignment with domain knowledge.

Topic Manipulation on Graph Structure

We conduct a controlled experiment to evaluate whether learned topics meaningfully control graph-level properties. We systematically adjust topic weights in the mixture vector θ: for each topic T_i, we vary its normalized weight from -0.30 to +0.50 while proportionally rescaling remaining topics to maintain a valid distribution. For each setting, we generate 300 graphs and compute structural metrics, quantifying changes using Cohen's d effect sizes. Fig. 5 shows effect sizes across ten structural metrics, where warmer colors indicate positive shifts and asterisks denote statistical significance. Topic 3 exhibits the strongest influence, substantially increasing modularity along with moderate increases in node count and diameter. This profile matches carcinogenic graphs, which are typically large, modular, and sparse (Swamidass et al. 2005; Gonzalez, Holder, and Cook 2001), suggesting Topic 3 encodes carcinogenic-like motifs. Conversely, Topic 1 reduces modularity and diameter, promoting compact, highly connected structures characteristic of non-carcinogenic graphs (Blinova et al. 2003). Other topics show localized effects: Topic 0 increases density and 4-cycles while reducing clustering, consistent with heteroaromatic scaffolds; Topic 4 decreases density but increases clustering and 6-cycles, resembling polycyclic natural products; Topic 2 shows minimal effects, indicating a neutral background role. These results demonstrate that NGTM topics capture meaningful structural patterns and enable interpretable control over graph generation.



(a) Reducing Graph Diameter via Topic 1 Adjustment



(b) Increasing Graph Density via Topic 3 Adjustment

Figure 6: Structural changes induced by progressively increasing the contribution of Topic 1 and Topic 3.

Fig. 6 demonstrates concrete examples of topic manipulation in NGTM. Increasing Topic 1 weight progressively reduces graph diameter from 8 to 4 nodes, confirming its association with compact topologies (Fig. 6(a)). Conversely, increasing Topic 3 weight steadily increases density from 0.04 to 0.18, reflecting increasingly dense and highly connected topologies (Fig. 6(b)). These results confirm that NGTM's learned topics serve as interpretable and controllable factors that systematically influence graph properties. The observed trends align with class-specific structural patterns in real PTC graphs, reinforcing NGTM's semantic interpretability.

Conclusion

We introduced NGTM, a framework that models graphs as mixtures of latent substructure topics, enabling interpretable and transparent graph generation. Experiments demonstrate that NGTM achieves competitive generation quality while providing clear interpretations of structural components, enabling users to understand and control graph generation for applications requiring transparency. NGTM establishes a promising direction for graph generative models that produce realistic, explainable, and controllable structures.

Acknowledgments

This work is partly supported by the National Natural Science Foundation of China (No. 62306255, 62406141), China Postdoctoral Science Foundation (No. GZC20252740), the Guangdong Basic and Applied Basic Research Foundation (No. 2024A1515011839), Jiangsu Funding Program for Excellent Postdoctoral Talent, National Natural Science Foundation of China.

References

Bianchi, F.; Terragni, S.; and Hovy, D. 2020. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. *arXiv:2004.03974*.

Blei, D. M.; Kucukelbir, A.; and McAuliffe, J. D. 2017. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518): 859–877.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan): 993–1022.

Blinova, V.; Dobrynin, D.; Finn, V. K.; Kuznetsov, S. O.; and Pankratova, E. 2003. Toxicology analysis by means of the JSM-method. *Bioinformatics*, 19(10): 1201–1207.

Chen, X.; Wang, Y.; He, J.; Du, Y.; Hassoun, S.; Xu, X.; and Liu, L.-P. 2025. Graph Generative Pre-trained Transformer. *arXiv:2501.01073*.

Dai, H.; Nazi, A.; Li, Y.; Dai, B.; and Schuurmans, D. 2020. Scalable Deep Generative Modeling for Sparse Graphs. *International Conference on Machine Learning,International Conference on Machine Learning*.

Debnath, A. K.; Lopez de Compadre, R. L.; Debnath, G.; Shusterman, A. J.; and Hansch, C. 1991. Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. correlation with molecular orbital energies and hydrophobicity. *Journal of medicinal chemistry*, 34(2): 786–797.

Du, Y.; Guo, X.; Shehu, A.; and Zhao, L. 2022. Interpretable molecular graph generation via monotonic constraints. In *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*, 73–81. SIAM.

Golomb, S. W. 1996. *Polyominoes: puzzles, patterns, problems, and packings*, volume 16. Princeton University Press.

Gong, Z.; Shen, S.; Meng, C.; and Sun, Y. 2025. Exploring Hypergraph Condensation via Variational Hyperedge Generation and Multi-Aspectual Amelioration. In *Proceedings of the ACM on Web Conference 2025*, 1248–1260.

Gonzalez, J.; Holder, L.; and Cook, D. J. 2001. Application of graph-based concept learning to the predictive toxicology domain. In *Proceedings of the Predictive Toxicology Challenge Workshop*.

Guo, X.; Zhao, L.; Qin, Z.; Wu, L.; Shehu, A.; and Ye, Y. 2020. Interpretable deep graph generation with node-edge co-disentanglement. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 1697–1707.

Hou, D.; Gao, C.; Li, X.; and Wang, Z. 2024. DAG-Aware Variational Autoencoder for Social Propagation Graph Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 8508–8516.

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

Ji, Y.; Sun, Y.; Zhang, Y.; Wang, Z.; Zhuang, Y.; Gong, Z.; Shen, D.; Qin, C.; Zhu, H.; and Xiong, H. 2025. A comprehensive survey on self-interpretable neural networks. *arXiv preprint arXiv:2501.15638*.

Jin, W.; Barzilay, R.; and Jaakkola, T. 2018. Junction tree variational autoencoder for molecular graph generation. In *International conference on machine learning*, 2323–2332. PMLR.

Kengkanna, A.; and Ohue, M. 2024. Enhancing property and activity prediction and interpretation using multiple molecular graph representations with MMGX. *Communications Chemistry*, 7(1): 74.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv:1312.6114*.

Kong, X.; Huang, W.; Tan, Z.; and Liu, Y. 2022. Molecule generation by principal subgraph mining and assembling. *NeurIPS*, 35: 2550–2563.

Kumar, A.; Hora, H.; Rohilla, A.; Kumar, P.; and Gautam, R. 2025. Explainable Artificial Intelligence (XAI) for Healthcare: Enhancing Transparency and Trust. In *International Conference on Cognitive Computing and Cyber Physical Systems*, 295–308. Springer.

Li, J.; Yu, J.; Li, J.; Zhang, H.; Zhao, K.; Rong, Y.; Cheng, H.; and Huang, J. 2020. Dirichlet graph variational autoencoder. *NeurIPS*, 33: 5274–5283.

Liao, R.; Li, Y.; Song, Y.; Wang, S.; Hamilton, W.; Duvenaud, D. K.; Urtasun, R.; and Zemel, R. 2019. Efficient graph generation with graph recurrent attention networks. *NeurIPS*, 32.

Liu, C.; Fan, W.; Liu, Y.; Li, J.; Li, H.; Liu, H.; Tang, J.; and Li, Q. 2023. Generative diffusion models on graphs: Methods and applications. *arXiv:2302.02591*.

Madeira, M.; Vignac, C.; Thanou, D.; and Frossard, P. 2024. Generative modelling of structurally constrained graphs. *NeurIPS*, 37: 137218–137262.

Miao, Y.; Yu, L.; and Blunsom, P. 2016. Neural variational inference for text processing. In *International conference on machine learning*, 1727–1736. PMLR.

Minello, G.; Bicciato, A.; Rossi, L.; Torsello, A.; and Cosmo, L. 2024. Graph Generation via Spectral Diffusion. *arXiv:2402.18974*.

O'Bray, L.; Horn, M.; Rieck, B.; and Borgwardt, K. 2021. Evaluation Metrics for Graph Generative Models: Problems, Pitfalls, and Practical Solutions. *Learning,Learning*.

Shen, D.; Qin, C.; Wang, C.; Dong, Z.; Zhu, H.; and Xiong, H. 2021a. Topic modeling revisited: A document graph-based neural network perspective. *Advances in neural information processing systems*, 34: 14681–14693.

Shen, D.; Qin, C.; Wang, C.; Zhu, H.; Chen, E.; and Xiong, H. 2021b. Regularizing variational autoencoder with diversity and uncertainty awareness. *IJCAI 21*.

Shen, D.; Zhu, H.; Zhu, C.; Xu, T.; Ma, C.; and Xiong, H. 2018. A joint learning approach to intelligent job interview assessment. In *IJCAI*, volume 18, 3542–3548.

Simonovsky, M.; and Komodakis, N. 2018. GraphVAE: Towards Generation of Small Graphs Using Variational Autoencoders. *Le Centre pour la Communication Scientifique Directe - HAL - Diderot,Le Centre pour la Communication Scientifique Directe - HAL - Diderot*.

Sohn, K.; Lee, H.; and Yan, X. 2015. Learning structured output representation using deep conditional generative models. *NeurIPS*, 28.

Sønderby, C. K.; Raiko, T.; Maaløe, L.; Sønderby, S. K.; and Winther, O. 2016. Ladder variational autoencoders. *NeurIPS*, 29.

Sun, Y.; Ji, Y.; Zhu, H.; Zhuang, F.; He, Q.; and Xiong, H. 2025. Market-aware long-term job skill recommendation with explainable deep reinforcement learning. *ACM Transactions on Information Systems*, 43(2): 1–35.

Sun, Y.; Zhu, H.; Qin, C.; Zhuang, F.; He, Q.; and Xiong, H. 2021. Discerning decision-making process of deep neural networks with hierarchical voting transformation. *Advances in Neural Information Processing Systems*, 34: 17221–17234.

Sun, Y.; Zhu, H.; Wang, L.; Zhang, L.; and Xiong, H. 2024. Large-scale online job search behaviors reveal labor market shifts amid COVID-19. *Nature Cities*, 1(2): 150–163.

Sun, Y.; Zhu, H.; and Xiong, H. 2025. Toward Faithful Neural Network Intrinsic Interpretation With Shapley Additive Self-Attribution. *IEEE Transactions on Neural Networks and Learning Systems*.

Sun, Y.; Zhuang, F.; Zhu, H.; Song, X.; He, Q.; and Xiong, H. 2019. The impact of person-organization fit on talent management: A structure-aware convolutional neural network approach. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 1625–1633.

Swamidass, S. J.; Chen, J.; Bruand, J.; Phung, P.; Ralaivola, L.; and Baldi, P. 2005. Kernels for small molecules and the prediction of mutagenicity, toxicity and anti-cancer activity. In *ISMB (Supplement of Bioinformatics)*, 359–368.

Thompson, R.; Knyazev, B.; Ghalebi, E.; Kim, J.; and Taylor, G. W. 2022. On evaluation metrics for graph generative models. *arXiv:2201.09871*.

Tian, Y.; Chen, Z.; Wang, Y.; Lv, L.; Li, H.; Lin, Z.; and Yuan, L. 2025. Leveraging Domain Motif Assembler for Multi-objective, Multi-domain and Explainable Molecular Design.

Toivonen, H.; Srinivasan, A.; King, R. D.; Kramer, S.; and Helma, C. 2003. Statistical evaluation of the predictive toxicology challenge 2000–2001. *Bioinformatics*, 19(10): 1183–1193.

Vignac, C.; Krawczuk, I.; Siraudin, A.; Wang, B.; Cevher, V.; and Frossard, P. 2022. Digress: Discrete denoising diffusion for graph generation. *arXiv:2209.14734*.

Wang, J.; and Zhu, F. 2025. ExSelfRL: An exploration-inspired self-supervised reinforcement learning approach to molecular generation. *Expert Systems with Applications*, 260.

Xie, Q.; Zhu, Y.; Huang, J.; Du, P.; and Nie, J.-Y. 2021. Graph neural collaborative topic model for citation recommendation. *ACM Transactions on Information Systems (TOIS)*, 40(3): 1–30.

Xu, C.; Deng, X.; Lu, Y.; and Yu, P. 2025. Generation of molecular conformations using generative adversarial neural networks. *Digital Discovery*, 4(1): 161–171.

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

You, J.; Ying, R.; Ren, X.; Hamilton, W.; and Leskovec, J. 2018. GraphRNN: Generating Realistic Graphs with Deep Auto-regressive Models. *International Conference on Machine Learning,International Conference on Machine Learning*.

Zahirnia, K.; Schulte, O.; Naddaf, P.; and Li, K. 2022. Micro and Macro Level Graph Modeling for Graph Variational Auto-Encoders.

Zhang, M.; Qamar, M.; Kang, T.; Jung, Y.; Zhang, C.; Bae, S.-H.; and Zhang, C. 2023. A survey on graph diffusion models: Generative ai in science for molecule, protein and material. *arXiv:2304.01565*.

Zheng, L.; Shi, F.; Peng, C.; Xu, M.; Fan, F.; Li, Y.; Zhang, L.; Du, J.; Wang, Z.; Lin, Z.; et al. 2024. Application scenario-oriented molecule generation platform developed for drug discovery. *Methods*, 222: 112–121.