

MolSnap: Snap-Fast Molecular Generation with Latent Variational Mean Flow

Md Atik Ahamed¹, Qiang Ye², Qiang Cheng^{1, 3*}

¹Department of Computer Science

²Department of Mathematics

³Institute for Biomedical Informatics

University of Kentucky

{atikahamed,qye3,qiang.cheng}@uky.edu

Abstract

Molecular generation conditioned on textual descriptions is a fundamental task in computational chemistry and drug discovery. Existing methods often struggle to simultaneously ensure high-quality, diverse generation and fast inference. In this work, we propose a novel causality-aware framework that addresses these challenges through two key innovations. First, we introduce a Causality-Aware Transformer (CAT) that jointly encodes molecular graph tokens and text instructions while enforcing causal dependencies during generation. Second, we develop a Variational Mean Flow (VMF) framework that generalizes existing flow-based methods by modeling the latent space as a mixture of Gaussians, enhancing expressiveness beyond unimodal priors. VMF enables efficient one-step inference while maintaining strong generation quality and diversity. Extensive experiments on four standard molecular benchmarks demonstrate that our model outperforms state-of-the-art baselines, achieving higher novelty (up to 74.5%), diversity (up to 70.3%), and 100% validity across all datasets. Moreover, VMF requires only one number of function evaluation (NFE) during conditional generation and up to five NFEs for unconditional generation, offering substantial computational efficiency over diffusion-based methods.

1 Introduction

The ability to generate novel molecular structures conditioned on textual descriptions is a key enabler for modern drug discovery and materials science (Ma et al. 2021; Southey and Brunavs 2023). Traditional molecular design relies heavily on expert knowledge and brute-force screening of chemical libraries (Pyzer-Knapp et al. 2015), both of which are time-consuming and limited in their ability to explore the vast chemical space. Recent advances in deep generative modeling offer promising alternatives, enabling the automatic generation of diverse, valid molecules that satisfy user-defined criteria expressed in natural language (Bilodeau et al. 2022; Ilnicka and Schneider 2023).

Despite these advances, existing approaches face several core challenges. First, most methods treat molecule generation as a generic sequence-to-sequence or graph generation task, overlooking the causal dependencies that govern molecular assembly and property formation (Li et al. 2018; You et al. 2018). In real chemical systems, the emergence of

molecular properties is driven by specific causal relationships among structural motifs (Hajduk and Greer 2007). Second, leading models often rely on diffusion-based or multi-step processes that require many inference iterations, making them computationally expensive (Ho, Jain, and Abbeel 2020; Zhu, Xiao, and Honavar 2024). Third, flow-based models typically assume unimodal Gaussian priors (Madhawa et al. 2019; Zang and Wang 2020), which may fail to capture the inherently multimodal nature of molecular distributions, thereby limiting the diversity and quality of generated samples.

To overcome these limitations, we introduce a novel framework that combines causality-aware modeling with variational flow matching for efficient and expressive molecular generation. Motivated by the observation that structural features, such as functional groups, causally influence key molecular properties like reactivity, solubility, and bioactivity (Mandal, Mandal et al. 2009; Bickerton et al. 2012), our framework incorporates causal reasoning for both model architectures and generation dynamics.

Our approach builds on a shared latent space where molecular graphs and textual instructions are encoded to be causality aware through modality-specific converters, enabling seamless integration of structural, semantic information and causal relationships between molecular components. To validate this design, we conducted latent causality analysis on ChEBI-20 using Granger tests consistent with our causal formulation (Figure 3). As illustrated in Figure 1, numerous molecules exhibit strong causal links ($p < 0.001$), and nearly half show at least one link at $p < 0.1$. These findings confirm that the latent space captures structured molecular causality effectively leveraged by our attention mask. Building on this insight, we introduce the Causality-Aware Transformer (CAT), which enforces directional dependencies through masked attention to ensure causally coherent generation of molecular substructures.

In addition, we propose a variational mean flow (VMF) framework that extends traditional mean flow models (Geng et al. 2025). By modeling the latent space as a mixture of Gaussians, VMF captures the multimodal nature of molecule distributions and enables efficient one-step inference. This variational formulation promotes generation diversity while maintaining structural validity and semantic alignment with input instructions.

*Corresponding author

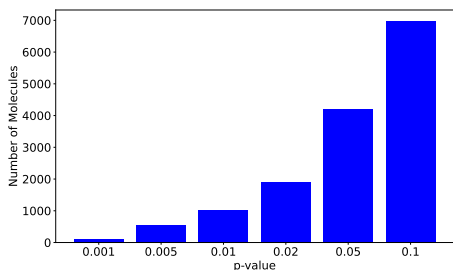


Figure 1: Causality analysis via Granger causality test verifying motivation for utilizing causality in our architecture.

In summary, our key contributions include:

- We introduce a causality-aware Transformer (CAT) that explicitly models dependencies among molecular graph tokens and text instruction tokens.
- We propose a variational mean flow (VMF) framework that generalizes standard flow matching and rectified flow models for molecule generation.
- VMF models the latent space as a mixture of Gaussians, improving diversity and capturing multimodal distributions more effectively than unimodal priors.
- Our model supports fast generation with only one or a few number of function evaluations (NFEs), significantly reducing inference cost compared to diffusion-based methods.

We validate our method on four standard molecular benchmarks, demonstrating that MolSnap achieves superior performance in terms of novelty, diversity, and validity, while also offering substantial gains in inference efficiency over state-of-the-art baselines.

2 Related Work

Molecular Representation and Generation. Molecular generation has progressed from SMILES-based sequence models (Weininger 1988; Bjerrum and Threlfall 2017; Gómez-Bombarelli et al. 2018; Kusner, Paige, and Hernández-Lobato 2017) to graph-based approaches that better capture molecular structures (Li et al. 2018; You et al. 2018; Liu et al. 2018; Jin, Barzilay, and Jaakkola 2018, 2020). While string-based methods face challenges with validity, graph-based models improve chemical correctness and interpretability but may lack diversity and scalability. Recent multimodal approaches leverage textual descriptions for molecule generation (Edwards, Zhai, and Ji 2021; Edwards et al. 2022; Schwaller et al. 2019; Born and Manica 2023; Fang et al. 2023), showing promise but often requiring slow multi-step inference and lacking explicit causal modeling.

Flow-Based Generative Models. Flow-based models enable exact likelihood estimation and stable training (Madhawa et al. 2019), with applications in molecular generation such as GraphNVP (Madhawa et al. 2019), MoFlow (Zang and Wang 2020), and GraphDF (Luo, Yan, and Ji 2021). Recent advances in flow matching and mean flow modeling (Geng et al. 2025) simplify training and improve efficiency.

However, most methods rely on unimodal Gaussian priors, limiting expressiveness for multimodal molecular distributions. Variational extensions like VRFM (Guo and Schwing 2025) improve diversity but are only explored in vision domains. Our work brings these advances to molecular generation while introducing causality-aware components.

Causality in Generative Modeling. Causal generative models (Deng et al. 2024) aim to uncover underlying mechanisms beyond mere correlations, which is essential for capturing functional group interactions and structure–property relationships in molecules (Hajduk and Greer 2007; Mandal, Mandal et al. 2009). However, prior work in causal representation learning has primarily focused on image domains and does not address the sequential nature of molecular assembly. Our causality-aware Transformer (CAT) fills this gap by enforcing autoregressive dependencies through masked attention, thereby enabling both structured generation and interpretability.

Conditional Molecular Generation. Conditional generation, whether guided by properties or text, is essential for controllable drug design (Westermayr et al. 2023; Schneuing et al. 2024). Property-conditioned VAEs (Gómez-Bombarelli et al. 2018), Transformer-based models (Christofidellis et al. 2023; Liu et al. 2023), and diffusion approaches (Zhu, Xiao, and Honavar 2024; Weiss et al. 2023) have made progress, but suffer from inference inefficiency. Models like Mol-Instructions (Fang et al. 2023) enable natural language control but do not capture causal structure-property relationships. Our framework uniquely combines causal modeling with variational mean flow, enabling fast, high-quality conditional generation with interpretable dynamics.

3 Methodology

In this section, we outline our methodology. Given a training set of molecule–text pairs, the goal is to learn a conditional generative model that produces valid, diverse molecules aligned with the text.

Each molecule is represented as a graph $g = (V, E)$, where $V = \{v_1, v_2, \dots, v_{|V|}\}$ is the set of atoms/nodes and E is the set of chemical bonds/edges. The atom information is stored in a matrix $N \in \mathbb{R}^{|V| \times D}$, where each row encodes properties such as atom type and chirality. The bonding structure is described by an adjacency tensor $A \in \mathbb{R}^{|V| \times |V| \times b}$, where each slice specifies the bond types. Our overall scheme is demonstrated in Figure 2.

Latent Converters

We refer to G and T as latent converters since they transform high-dimensional, unstructured graph and text data into compact latent representations. Specifically, G encodes a molecule graph g into a latent representation x , while T maps a conditional instruction i into a latent representation c . To ensure a shared latent space, we first apply contrastive alignment training, which enforces similarity between paired (g, i) representations and dissimilarity among unrelated pairs. Additionally, we train a graph encoder–decoder framework,

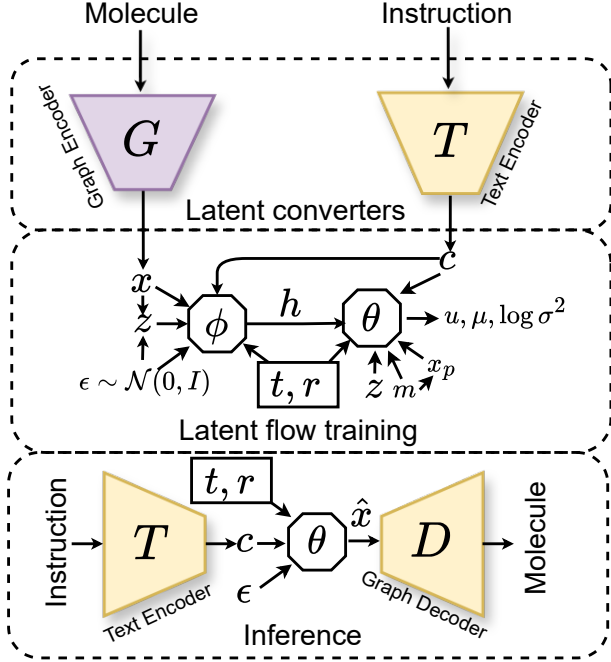


Figure 2: Overview of our method. Converters G and T map molecules and instructions into latents. Latent flow training learns generative dynamics in this space, while the inference module produces molecules conditioned on instructions.

where the decoder D reconstructs molecules from latent vectors. During inference, D is used to recover the final molecular graph from the predicted latent representation \hat{x} .

Latent Flow Training

Given the latent representations $x \leftarrow G(g)$ and $c \leftarrow T(i)$, we jointly train the networks θ and ϕ . The overall training process is outlined in Algorithm 1 for a single example. Our approach differs fundamentally from prior works: rather than relying on conventional diffusion models (Ho, Jain, and Abbeel 2020), we perform training directly in the latent space utilizing Mean Flow (Geng et al. 2025), VRFM (Guo and Schwing 2025), and CAT (Deng et al. 2024).

While the original Mean Flow was introduced for image data, we are the first to adapt and extend it to molecule-text modalities. This extension enables us for efficient 1-NFE molecular generation. While Mean Flow provides an efficient framework, we noticed that its latent space is modeled with unimodal gaussian distribution. However, this does not address the multi-modality issue identified in Variational Rectified Flow Matching (VRFM) (Guo and Schwing 2025).

To capture multi-modality, we study the use of a mixture model over velocities at each data-domain-time-domain location. Specifically, inspired by VRFM (Guo and Schwing 2025), we utilize a KL-Divergence term (Specific formulation is described later in this Section) as a regularization in our objective function together with Mean Flow based cost. Moreover, we go beyond a naive adaptation by integrat-

ing Mean Flow with Variational Rectified Flow Matching (VRFM) (Guo and Schwing 2025). VRFM was initially designed for image modalities, and our framework is the first to combine VRFM with Mean Flow for molecular generation, demonstrating a novel and effective integration.

To achieve this, we employ two core networks: a causality-aware Transformer (CAT), denoted as θ in Algorithm 1 and Algorithm 2, and a variational encoder ϕ . The encoder ϕ takes as input the condition c , noise ϵ , clean latent x , intermediate latent z , and timesteps t and r .

Its functionality is summarized by the following equations:

$$\mu, \log \sigma^2 = \phi(c, \epsilon, x, z, r, t), \quad (1)$$

$$\sigma = e^{\frac{1}{2} \log \sigma^2}, \quad (2)$$

$$h = \mu + \sigma \odot \epsilon, \quad \text{where } \epsilon \sim \mathcal{N}(0, I). \quad (3)$$

After obtaining the latent representation h from ϕ , we combine it with the partial clean input x_p , causal mask m , and with other inputs c, z, t, r . The roles of causal masking and the partial input x_p are discussed later in this subsection. The network θ processes these inputs to produce u , which represents the predicted velocity.

To compute both the instantaneous velocity u and its derivative with respect to time, we employ the Jacobian-vector product (JVP), which efficiently estimates directional derivatives. In our setting, JVP operates on the composite network (θ, ϕ) with input (z, r, t) and tangent vector $(v, 0, 1)$, as illustrated in Algorithm 1. where v is the target vector. Here, u represents the predicted velocity, while $\dot{u} = du/dt$ denotes its derivative with respect to the time variable t . With the help of v, t, r, \dot{u} , we compute the target velocity u_t as demonstrated in Algorithm 1.

In addition to the L_2 loss we also incorporate KL-Divergence loss \mathcal{L}_{KL} and Dispersive loss (Wang and He 2025) \mathcal{L}_{disp} and combine them with weight α and β respectively to formulate the final loss \mathcal{L} .

To regularize the latent space h produced by ϕ , we introduce a Kullback-Leibler (KL) divergence that aligns the latent distribution with a standard Gaussian prior $\mathcal{N}(0, I)$. Specifically, given μ and $\log \sigma^2$ from ϕ , the KL loss is computed as:

$$\mathcal{L}_{KL} = \frac{1}{2} \sum_j \left(e^{\log \sigma_j^2} + \mu_j^2 - 1 - \log \sigma_j^2 \right). \quad (4)$$

where the summation runs over latent dimensions j . This term encourages the latent variables to follow a unit Gaussian, improving stability and enabling a variational formulation.

To promote diversity in the latent space, we incorporate a dispersive loss \mathcal{L}_{disp} , defined as:

$$\mathcal{L}_{disp} = \log \left(\frac{1}{B^2} \sum_{b_1, b_2} \exp \left(-\frac{\|z_{b_1} - z_{b_2}\|_2^2}{\tau} \right) \right), \quad (5)$$

where z_{b_1} and z_{b_2} are latent representations in a batch of size B , and $\tau > 0$ is a temperature parameter controlling the sharpness of the distance penalty.

Algorithm 1: Training

```

1:  $\epsilon \sim \mathcal{N}(0, I)$ 
2: Sample  $t, r$  % Uniform or log-normal
3:  $z \leftarrow (1 - t) \cdot x + t \cdot \epsilon$ 
4:  $v \leftarrow \epsilon - x$ 
5:  $(u, \dot{u}, \mu, \log \sigma^2) \leftarrow \text{jvp}((\theta, \phi), (z, r, t), (v, 0, 1))$ 
   %  $h, \mu, \log \sigma^2 \leftarrow \phi \leftarrow (c, \epsilon, x, z, t, r)$ 
   %  $u \leftarrow \theta \leftarrow (c, h, x_p, z, m, t, r)$ 
6:  $u_t \leftarrow v - (t - r) \cdot \dot{u}$  % target
7:  $\mathcal{L} \leftarrow \|u - \text{stopgrad}(u_t)\|^2 + \alpha \mathcal{L}_{KL} + \beta \mathcal{L}_{disp}$ 

```

Causality integration. Motivated by the causality analysis in Figure 1, we integrate a causal attention mechanism into θ , making it a *causality-aware Transformer* (CAT). The inputs to θ are concatenated along the token dimension, enabling dynamic context modeling during training and decoder-like behavior during inference, where only ϵ and c are required.

As shown in Figure 3, causal dependencies are enforced via an attention mask m . The clean latent x and noisy latent z are partitioned into groups (e.g., g_1, g_2, g_3), where x_p refers to the partial clean tokens (e.g., the first two groups). All tokens of x_p and z attend to the condition c and latent encoding h , but the current group x_{p,g_i} only attends up to its preceding groups $x_{p,g_{i-1}}$. For z , temporal information (t, r) is incorporated, and causal masking ensures that each group depends only on earlier groups of x_p . This setup resembles autoregressive modeling in language tasks, where future tokens are masked to ensure strictly causal learning.

During inference, the clean latent tokens x and x_p are unavailable, as they are derived from the original data. Therefore, θ is designed to work solely with ϵ and c at test time. By appending all inputs along the sequence dimension during training, the model learns to rely on the dynamic combination of noise ϵ and textual condition c for reconstruction. Consequently, θ naturally functions in a decoder-only manner during inference, where the learned causal dependencies allow it to sequentially generate the latent representation without requiring x or x_p .

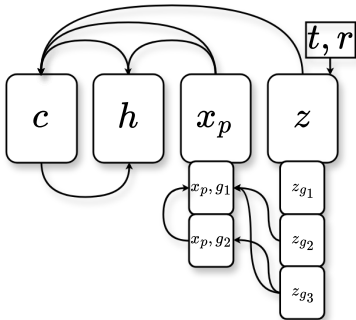


Figure 3: Causality-aware grouping of latent tokens.

Inference

During inference, we start with a Gaussian noise sample $\epsilon \sim \mathcal{N}(0, I)$ and iteratively update the latent variable by predicting the velocity $u = \theta(c, \epsilon, t, r)$ using the network θ ,

then adjusting the latent as $\hat{x} \leftarrow \hat{x} - h \cdot u$ at each step, where $h = t - r$. For conditional generation, we leverage a one-step (1-NFE) inference strategy (Algorithm 2), which makes the process significantly faster compared to other transformer variants that typically requires many or even 1000 NFES.

To control the balance between conditional and unconditional signals, we employ classifier-free guidance (CFG). Specifically, we compute both a conditional output $u_{\text{cond}} = \theta(c, \epsilon, t, r)$ and an unconditional output $u_{\text{uncond}} = \theta(c_{\text{null}}, \epsilon, t, r)$, where c_{null} represents a null condition. The final guided velocity is then represented by equation 6:

$$u_{\text{final}} = w u_{\text{cond}} + (1 - w) u_{\text{uncond}} \quad (6)$$

where w is the guidance scale. For unconditional generation, c is replaced with c_{null} , and θ operates solely on the noise input ϵ to produce valid molecular structures. Since our latent space is regularized by a KL divergence term to follow a standard Gaussian prior, we require only a single ϵ sample, without needing multiple independent noise components.

Algorithm 2: Inference (1-NFE)

```

1:  $\epsilon \sim \mathcal{N}(0, I)$ 
2:  $\hat{x} \leftarrow \epsilon - \theta(c, \epsilon, r = 0, t = 1)$ 

```

Flow-based Models as Special Cases

Our Variational Mean Flow (VMF) framework generalizes several existing flow-based generative models, including Mean Flow (MF) (Geng et al. 2025), Flow Matching (FM) (Lipman et al. 2022), and Rectified Flow Matching (RFM) (Liu, Gong, and Liu 2022). Specifically, VMF extends MF by incorporating a variational inference approach with a mixture-of-Gaussians prior, capturing multimodal latent distributions more effectively than the standard unimodal Gaussian used in MF. Flow Matching becomes a special case within VMF when adopting a unimodal prior and bypassing variational inference, simplifying the model to direct velocity matching (Lipman et al. 2022). Similarly, RFM can be viewed as a variant of VMF where rectification adjustments are implicitly modeled by the variational posterior distribution, allowing corrections and improved stability during inference (Liu, Gong, and Liu 2022). By integrating these frameworks, VMF significantly enhances expressiveness, resulting in improved molecular generation quality, greater diversity, and more efficient inference.

4 Experiments

In this section, we describe the experimental setup and summarize the evaluation results.

Datasets. We conduct experiments on four molecular datasets: PubChem (Liu et al. 2023), ChEBI-20 (Edwards, Zhai, and Ji 2021), PCDes (Zeng et al. 2022), and MoMu (Su et al. 2022). Following prior work on molecular structure generation (Irwin et al. 2012; Blum and Reymond 2009; Rupp et al. 2012; Zhu, Xiao, and Honavar 2024), we limit our analysis to molecules with fewer than 30 atoms, consistent with standard practice (Ramakrishnan et al. 2014; Polykovskiy

Methods	ChEBI-20					PubChem				
	Sim.	Nov.	Div.	Val.	Ove.	Sim.	Nov.	Div.	Val.	Ove.
MolT5-small	73.32	31.43	17.22	78.27	50.06	68.36	20.63	9.32	78.86	44.29
MolT5-base	80.75	32.83	17.66	84.63	53.97	73.85	21.86	9.89	79.88	46.37
MolT5-large	96.88	12.92	11.20	98.06	54.77	91.57	20.85	9.84	95.18	54.36
ChemT5-small	96.22	13.94	13.50	96.74	55.10	89.32	20.89	13.10	93.47	54.19
ChemT5-base	95.48	15.12	13.91	97.15	55.42	89.42	22.40	13.98	92.43	54.56
Mol-Instruction	65.75	32.01	26.50	77.91	50.54	23.40	37.37	27.97	71.10	39.96
3M-Diffusion	87.09	55.36	34.03	100.0	69.12	87.05	64.41	33.44	100.0	71.22
Ours-MF	84.85	58.80	49.37	100.0	73.26	79.90	67.80	52.30	100.0	75.00
Ours-VMF	76.54	70.77	59.78	100.0	76.77	75.23	70.99	57.29	100.0	75.88
Ours-VMFD	77.02	69.33	58.08	100.0	76.11	78.55	68.67	54.69	100.0	75.48
Ours-MFD	83.68	59.26	60.34	100.0	75.82	80.21	68.05	49.30	100.0	74.39

Table 1: Quantitative comparison of conditional generation on ChEBI-20 and PubChem. Our method significantly outperforms SOTA baselines in novelty (Nov.), diversity (Div.), and validity (Val.), while maintaining strong similarity (Sim.) and excelling in overall (Ove.) performance. Results are percentages (higher is better).

Methods	PCDes					MoMu				
	Sim.	Nov.	Div.	Val.	Ove.	Sim.	Nov.	Div.	Val.	Ove.
MolT5-small	64.84	24.91	9.67	73.96	43.35	16.64	97.49	29.95	60.19	51.07
MolT5-base	71.71	25.85	10.50	81.92	47.50	19.76	97.78	29.98	68.84	54.09
MolT5-large	88.37	20.15	9.49	96.48	53.62	25.07	97.47	30.33	90.40	60.82
ChemT5-small	86.27	23.28	13.17	93.73	54.11	23.25	96.97	30.04	88.45	59.68
ChemT5-base	85.01	25.55	14.08	92.93	54.39	23.40	97.65	30.07	87.61	59.68
Mol-Instruction	60.86	35.60	24.57	79.19	50.06	14.89	97.52	30.17	68.32	52.73
3M-Diffusion	81.57	63.66	32.39	100.0	69.41	24.62	98.16	37.65	100.0	65.11
Ours-MF	73.49	69.83	51.25	100.0	73.64	24.61	97.90	59.82	100.0	70.58
Ours-VMF	72.09	74.50	61.61	100.0	77.05	24.71	97.94	70.29	100.0	73.24
Ours-VMFD	69.00	73.86	55.25	100.0	74.53	24.71	97.49	62.50	100.0	71.18
Ours-MFD	72.89	72.11	54.96	100.0	74.99	26.78	97.34	63.18	100.0	71.83

Table 2: Quantitative comparison of conditional generation on PCDes and MoMu. Our method surpasses SOTA baselines in novelty (Nov.), diversity (Div.), and validity (Val.), while maintaining strong similarity (Sim.) and consistently leading in overall (Ove.) scores. Results are in percentages (higher is better).

Methods	ChEBI-20						PubChem					
	Uni.	Nov.	KL	FCD	Val.	Ove.	Uni.	Nov.	KL	FCD	Val.	Ove.
CharRNN	72.46	11.57	95.21	75.95	98.21	70.68	63.28	23.47	90.72	76.02	94.09	69.52
VAE	57.57	47.88	95.47	74.19	63.84	67.79	44.45	42.47	91.67	55.56	94.10	65.65
AAE	1.23	1.23	38.47	0.06	1.35	8.47	2.94	3.21	39.33	0.08	1.97	9.51
LatentGAN	66.93	57.52	94.38	76.65	73.02	73.70	52.00	50.36	91.38	57.38	53.62	60.95
BwR	22.09	21.97	50.59	0.26	22.66	23.51	82.35	82.34	45.53	0.11	87.73	59.61
HierVAE	82.17	72.83	93.39	64.32	100.0	82.54	75.33	72.44	89.05	50.04	100.0	77.37
PS-VAE	76.09	74.55	83.16	32.44	100.0	73.25	66.97	66.52	83.41	14.41	100.0	66.26
3M-Diffusion	83.04	70.80	96.29	77.83	100.0	85.59	85.42	81.20	92.67	58.27	100.0	83.51
Ours	80.74	71.75	94.33	76.20	100.0	84.60	91.64	87.66	92.55	61.07	100.0	86.58

Table 3: Quantitative comparison of unconditional generation. Results of Uniq (Uni.), KL Div (KL), and FCD on ChEBI-20 and PubChem, which refer to Uniqueness, KL Divergence, and Fréchet ChemNet Distance, respectively. Results are presented in percentage values. A higher number indicates a better generation quality.

Dataset	Training	Validation	Test
ChEBI-20	15,409	1,971	1,965
PubChem	6,912	571	1,162
PCDes	7,474	1,051	2,136
MoMu	7,474	1,051	4,554

Table 4: Dataset statistics for training, validation, and test.

et al. 2020; Brown et al. 2019). For instance, datasets like QM9 contain molecules with up to 9 heavy atoms (Ramakrishnan et al. 2014), while ZINC-based datasets such as ZINC-250K and MOSES constrain compounds to a similar size range (Irwin et al. 2012; Polykovskiy et al. 2020). The GuacaMol benchmark likewise targets lead-like molecules in this range (Brown et al. 2019). Dataset statistics are shown in Table 4. Note that PCDes and MoMu share the same training/validation splits but differ in test sets, and all comparisons are reported on the respective test sets.

Implementation details. This section outlines our implementation for conditional and unconditional molecule generation, along with baseline models for comparison.

For instruction-guided/conditional generation, our model is benchmarked against MolT5 (Edwards et al. 2022), ChemT5 (Christofidellis et al. 2023), Mol-Instruction (Fang et al. 2023), and 3M-Diffusion (Zhu, Xiao, and Honavar 2024), considering multiple variants of MolT5 (small, base, large) and ChemT5 (small, base). For the unconditional scenario, we compare with prominent approaches: CharRNN (Segler et al. 2018), VAE (Kingma and Welling 2013), AAE (Makhzani et al. 2015), LatentGAN (Prykhodko et al. 2019), BwR (Diamant et al. 2023), HierVAE (Jin, Barzilay, and Jaakkola 2020), PS-VAE (Kong et al. 2022), and 3M-Diffusion (Zhu, Xiao, and Honavar 2024). To ensure fair comparison, we use the same train, validation, and test splits as 3M-Diffusion and report its results from the original paper to avoid discrepancies from environment or hyper-parameters.

Our architecture uses GIN (Hu et al. 2019) as the graph encoder for molecular structures, SciBERT (Beltagy, Lo, and Cohan 2019) for text encoding, and HierVAE (Jin, Barzilay, and Jaakkola 2020) as the graph decoder. The converted latent representation is utilized via our proposed VMF training, where θ utilizes transformer architecture, maintaining causal dependencies, and ϕ uses MLPs. We set sampling steps to 1-NFE for conditional generation and 3/5-NFE for unconditional generation. The model is optimized using the Adam optimizer (Jimmy and Diederik 2014) with a learning rate of 0.001 and trained for 1000 epochs.

Classifier-free guidance (Ho and Salimans 2022) is implemented by randomly dropping conditional embeddings with 0.1 probability during training and inference; for unconditional generation, all conditional inputs are removed. Models are implemented in PyTorch (Paszke et al. 2019) and trained on NVIDIA A100 GPUs. Code is provided in the supplementary files.

Performance Metrics. We evaluate our model on both conditional (text-guided) and unconditional molecule generation

tasks. For conditional generation, we follow standard protocols (Edwards et al. 2022; Christofidellis et al. 2023; Fang et al. 2023; Zhu, Xiao, and Honavar 2024) and assess: (i) *Similarity*, the proportion of generated molecules matching the ground truth with MACCS (Durant et al. 2002) cosine similarity with a threshold of 0.5; (ii) *Novelty*, the fraction of generated molecules with $f(G, \hat{G}) < 0.8$, indicating that the generated molecules differ from the references; (iii) *Diversity*, defined as the average pairwise distance $1 - f(\cdot, \cdot)$ among valid molecules (where f is the MACCS similarity and a molecule is considered valid if $f(G, \hat{G}) \geq 0.5$); and (iv) *Validity*, the percentage of chemically valid molecules.

For unconditional generation, we rely on the GuacaMol benchmarks (Brown et al. 2019) and report: (i) *Uniqueness*, the ratio of distinct molecules; (ii) *Novelty*, the proportion of molecules not present in the training set; (iii) *KL Divergence*, which quantifies the distributional similarity between generated and training molecules; and (iv) *Fréchet ChemNet Distance* (FCD) (Preuer et al. 2018), which measures feature-level alignment using a ChemNet encoder. All metrics are normalized to [0, 1] and reported as percentages, where higher is better.

Individual metrics often show trade-offs. As shown in Figure 4, raising the similarity threshold decreases similarity and diversity but boosts novelty, revealing metric tension. To address this, we introduce an (v) *Overall* metric, calculated as the average of all metrics, for balanced evaluation, preventing models from being favored for excelling in only one metric.

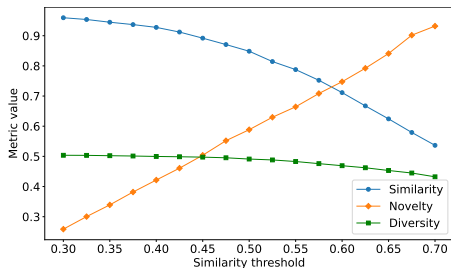


Figure 4: Tradeoff curve demonstrating result variance.

Result analysis. We evaluate our proposed variants: Mean Flow (MF), Variational Mean Flow (VMF), Mean Flow with Dispersive Loss (MFD), and Variational Mean Flow with Dispersive Loss (VMFD) against state-of-the-art (SOTA) models on four benchmark datasets. Results are summarized in Tables 1, 2, and 3.

In the conditional generation setting (Tables 1 and 2), our models consistently outperform baselines such as MolT5, ChemT5, and 3M-Diffusion in novelty, diversity, and validity. While MolT5-large achieves high similarity (e.g., 96.88% on ChEBI-20), it suffers from low novelty and diversity due to overfitting. In contrast, VMF and VMFD maintain strong similarity while significantly improving diversity (up to 61.61% on PCDes) and novelty (above 70% across datasets), leading to the highest overall (Ove.) scores: 77.05% on PCDes and 75.88% on PubChem. MFD also highlights the benefit of dispersive loss, boosting diversity (e.g., 60.34% on ChEBI-20).

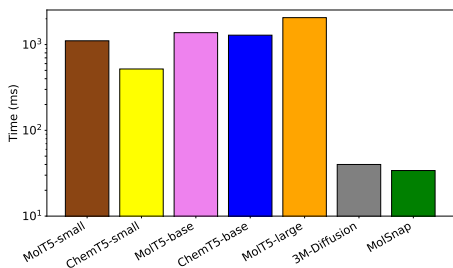


Figure 5: Inference time comparison

For the unconditional generation task (Table 3), our MF variant achieves competitive or superior performance compared to both VAE-based and diffusion-based baselines. On ChEBI-20, MF reaches 84.60% overall, closely matching or exceeding 3M-Diffusion, with perfect validity (100%). On PubChem, MF attains 91.64% uniqueness and 87.66% novelty, outperforming all baselines by a notable margin. Due to space limits, visual case studies are provided in the Appendix.

Figure 5 compares inference times across models, showing that our method (MolSnap) is over an order of magnitude faster than 3M-Diffusion and 10–50 \times faster than transformer-based models like MolT5 and ChemT5. This speedup comes from our novel flow-based design, which requires far fewer function evaluations than diffusion-based methods.

These results highlight two key strengths of our approach: (i) the integration of variational modeling with Mean Flow enhances the exploration of chemical space, leading to higher novelty and diversity; and (ii) the proposed flow-based training framework enables faster sampling while maintaining similarity, with fewer function evaluations required compared to diffusion-based models.

Our framework demonstrates strong generalization across both conditional and unconditional tasks, outperforming SOTA models in most metrics and achieving the best overall balance between similarity, novelty, diversity, and validity.

5 Ablation

In this section, we present ablation studies on key components of our framework. While Tables 1 and 2 report results for multiple variants, we further analyze additional factors affecting performance. We report similarity, diversity, and novelty metrics, as validity remains unchanged.

Adaptive vs. Regular L_2 Loss

Our framework uses the standard L_2 /Mean Squared Error (MSE) loss. Mean Flow (Geng et al. 2025) proposed an adaptive L_2 loss, which we incorporated to evaluate its effect. Figure 6a presents a comparison on the PubChem dataset, showing that the regular L_2 loss outperforms adaptive L_2 in similarity and diversity metrics.

Sampling strategy for t, r

We also evaluated uniform and log-normal sampling strategies for t and r , following Mean Flow (Geng et al. 2025). Our experiments reveal that each strategy offers trade-offs across metrics: uniform sampling tends to improve novelty

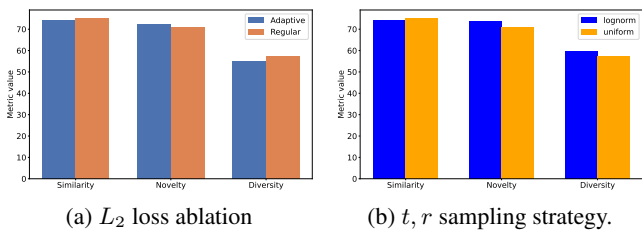


Figure 6: Ablation on loss and time sampling strategy.

and diversity, while log-normal sampling better preserves similarity, as shown in Figure 6b for PubChem.

Effect of Classifier-free guidance

We analyze the impact of classifier-free guidance (CFG) on conditional generation. Figure 7 shows how similarity, novelty, and diversity vary with different CFG scales. Higher unconditional probability boosts novelty and diversity by encouraging chemical space exploration but reduces similarity as the model becomes less condition-aware. A moderate CFG scale offers a balanced trade-off, generating relevant and diverse molecules.

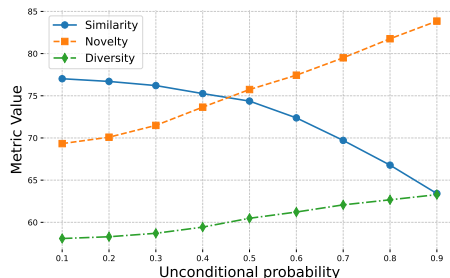


Figure 7: Performance variation with CFG.

6 Conclusion

We proposed a novel framework for conditional molecular generation that combines a causality-aware Transformer (CAT) with a variational mean flow (VMF) approach. CAT captures the causal dependencies in molecular assembly, while VMF models the latent space as a Gaussian mixture, enabling better representation of multimodal molecular distributions. Extensive experiments on four molecular benchmarks show that our method outperforms state-of-the-art models in novelty, diversity, and validity. VMF consistently achieves 100% validity and improves diversity (up to 70.3%) and novelty (up to 74.5%) while enabling fast inference with only 1–5 number of function evaluations, offering significant computational advantages over diffusion-based methods. Our integration of causality modeling with variational flow matching offers a compelling path forward for efficient and interpretable molecular generation.

Limitations and Future Work. Future directions include modeling more complex molecular interactions beyond token-level causality, extending the framework to larger molecular structures, and integrating 3D structural information to enhance expressiveness for structure-based drug design.

References

- Beltagy, I.; Lo, K.; and Cohan, A. 2019. SciBERT: A pre-trained language model for scientific text. *arXiv preprint arXiv:1903.10676*.
- Bickerton, G. R.; Paolini, G. V.; Besnard, J.; Muresan, S.; and Hopkins, A. L. 2012. Quantifying the chemical beauty of drugs. *Nature Chemistry*, 4(2): 90–98.
- Bilodeau, C. L.; Jin, W.; Jaakkola, T. S.; Barzilay, R.; and Jensen, K. F. 2022. Generative models for molecular discovery: Recent advances and challenges. *WIREs Computational Molecular Science*, 12(4): e1608.
- Bjerrum, E. J.; and Threlfall, R. 2017. Molecular generation with recurrent neural networks (RNNs). *arXiv preprint arXiv:1705.04612*.
- Blum, L. C.; and Reymond, J.-L. 2009. 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *Journal of the American Chemical Society*, 131(25): 8732–8733.
- Born, J.; and Manica, M. 2023. Regression Transformer enables concurrent sequence regression and generation for molecular language modelling. *Nature Machine Intelligence*, 5(4): 432–444.
- Brown, N.; Fiscato, M.; Segler, M. H.; and Vaucher, A. C. 2019. GuacaMol: benchmarking models for de novo molecular design. *Journal of Chemical Information and Modeling*, 59(3): 1096–1108.
- Christofidellis, D.; Giannone, G.; Born, J.; Winther, O.; Laino, T.; and Manica, M. 2023. Unifying molecular and textual representations via multi-task language modelling. *arXiv preprint arXiv:2301.12586*.
- Deng, C.; Zh, D.; Li, K.; Guan, S.; and Fan, H. 2024. Causal Diffusion Transformers for Generative Modeling. *arXiv preprint arXiv:2412.12095*.
- Diamant, N. L.; Tseng, A. M.; Chuang, K. V.; Biancalani, T.; and Scalia, G. 2023. Improving graph generation by restricting graph bandwidth. In *International Conference on Machine Learning*, 7939–7959. PMLR.
- Durant, J. L.; Leland, B. A.; Henry, D. R.; and Nourse, J. G. 2002. Reoptimization of MDL keys for use in drug discovery. *Journal of Chemical Information and Computer Sciences*, 42(6): 1273–1280.
- Edwards, C.; Lai, T.; Ros, K.; Honke, G.; Cho, K.; and Ji, H. 2022. Translation between molecules and natural language. *arXiv preprint arXiv:2204.11817*.
- Edwards, C.; Zhai, C.; and Ji, H. 2021. Text2mol: Cross-modal molecule retrieval with natural language queries. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 595–607.
- Fang, Y.; Liang, X.; Zhang, N.; Liu, K.; Huang, R.; Chen, Z.; Fan, X.; and Chen, H. 2023. Mol-Instructions: A Large-Scale Biomolecular Instruction Dataset for Large Language Models. *arXiv preprint arXiv:2306.08018*.
- Geng, Z.; Deng, M.; Bai, X.; Kolter, J. Z.; and He, K. 2025. Mean flows for one-step generative modeling. *arXiv preprint arXiv:2505.13447*.
- Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; and Aspuru-Guzik, A. 2018. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, 4(2): 268–276.
- Guo, P.; and Schwing, A. 2025. Variational Rectified Flow Matching. In *Forty-second International Conference on Machine Learning*.
- Hajduk, P. J.; and Greer, J. 2007. A decade of fragment-based drug design: strategic advances and lessons learned. *Nature Reviews Drug Discovery*, 6(3): 211–219.
- Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.
- Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.
- Hu, W.; Liu, B.; Gomes, J.; Zitnik, M.; Liang, P.; Pande, V.; and Leskovec, J. 2019. Strategies for pre-training graph neural networks. *arXiv preprint arXiv:1905.12265*.
- Ilnicka, A.; and Schneider, G. 2023. Designing molecules with autoencoder networks. *Nature Computational Science*, 3(11): 922–933.
- Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; and Coleman, R. G. 2012. ZINC: a free tool to discover chemistry for biology. *Journal of Chemical Information and Modeling*, 52(7): 1757–1768.
- Jimmy, B.; and Diederik, P. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv: 1412.6980*, 2014.
- 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.
- Jin, W.; Barzilay, R.; and Jaakkola, T. 2020. Hierarchical generation of molecular graphs using structural motifs. In *International Conference on Machine Learning*, 4839–4848. PMLR.
- Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.
- Kong, X.; Huang, W.; Tan, Z.; and Liu, Y. 2022. Molecule generation by principal subgraph mining and assembling. *Advances in Neural Information Processing Systems*, 35: 2550–2563.
- Kusner, M. J.; Paige, B.; and Hernández-Lobato, J. M. 2017. Grammar variational autoencoder. In *International Conference on Machine Learning*, 1945–1954. PMLR.
- Li, Y.; Vinyals, O.; Dyer, C.; Pascanu, R.; and Battaglia, P. 2018. Learning deep generative models of graphs. *arXiv preprint arXiv:1803.03324*.
- Lipman, Y.; Chen, R. T. Q.; Ben-Hamu, H.; Nickel, M.; and Le, M. 2022. Flow Matching for Generative Modeling. *arXiv preprint arXiv:2210.02747*. Presented at ICLR 2023.
- Liu, Q.; Allamanis, M.; Brockschmidt, M.; and Gaunt, A. 2018. Constrained graph variational autoencoders for

- molecule design. *Advances in Neural Information Processing Systems*, 31.
- Liu, X.; Gong, C.; and Liu, Q. 2022. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. *arXiv preprint arXiv:2209.03003*. Introduces “Rectified Flow” for straight-path transport mappings in generative modeling.
- Liu, Z.; Li, S.; Luo, Y.; Fei, H.; Cao, Y.; Kawaguchi, K.; Wang, X.; and Chua, T.-S. 2023. Molca: Molecular graph-language modeling with cross-modal projector and uni-modal adapter. *arXiv preprint arXiv:2310.12798*.
- Luo, Y.; Yan, K.; and Ji, S. 2021. Graphdf: A discrete flow model for molecular graph generation. In *International Conference on Machine Learning*, 7192–7203. PMLR.
- Ma, Y.-S.; Xin, R.; Yang, X.-L.; Shi, Y.; Zhang, D.-D.; Wang, H.-M.; Wang, P.-Y.; Liu, J.-B.; Chu, K.-J.; and Fu, D. 2021. Paving the way for small-molecule drug discovery. *American Journal of Rranslational Research*, 13(3): 853.
- Madhawa, K.; Ishiguro, K.; Nakago, K.; and Abe, M. 2019. Graphnvp: An invertible flow model for generating molecular graphs. *arXiv preprint arXiv:1905.11600*.
- Makhzani, A.; Shlens, J.; Jaitly, N.; Goodfellow, I.; and Frey, B. 2015. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*.
- Mandal, S.; Mandal, S. K.; et al. 2009. Rational drug design. *European Journal of Pharmacology*, 90–100.
- Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32.
- Polykovskiy, D.; Zhebrak, A.; Sanchez-Lengeling, B.; Golovanov, S.; Tatanov, O.; Belyaev, S.; Kurbanov, R.; Artamonov, A.; Aladinskiy, V.; Veselov, M.; et al. 2020. Molecular sets (MOSES): a benchmarking platform for molecular generation models. *Frontiers in Pharmacology*, 11: 565644.
- Preuer, K.; Renz, P.; Unterthiner, T.; Hochreiter, S.; and Klambauer, G. 2018. Fréchet ChemNet distance: a metric for generative models for molecules in drug discovery. *Journal of Chemical Information and Modeling*, 58(9): 1736–1741.
- Prykhodko, O.; Johansson, S. V.; Kotsias, P.-C.; Arús-Pous, J.; Bjerrum, E. J.; Engkvist, O.; and Chen, H. 2019. A de novo molecular generation method using latent vector based generative adversarial network. *Journal of Cheminformatics*, 11: 1–13.
- Pyzer-Knapp, E. O.; Suh, C.; Gómez-Bombarelli, R.; Aguilera-Iparraguirre, J.; and Aspuru-Guzik, A. 2015. What is high-throughput virtual screening? A perspective from organic materials discovery. *Annual Review of Materials Research*, 45: 195–216.
- Ramakrishnan, R.; Dral, P. O.; Rupp, M.; and Von Lilienfeld, O. A. 2014. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1(1): 1–7.
- Rupp, M.; Tkatchenko, A.; Müller, K.-R.; and Von Lilienfeld, O. A. 2012. Fast and accurate modeling of molecular atomization energies with machine learning. *Physical Review Letters*, 108(5): 058301.
- Schneuing, A.; Harris, C.; Du, Y.; Didi, K.; Jamasb, A.; Igashov, I.; Du, W.; Gomes, C.; Blundell, T. L.; Lio, P.; Welling, M.; Bronstein, M.; and Correia, B. 2024. Structure-based drug design with equivariant diffusion models. *Nature Computational Science*, 4(12): 899–909.
- Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; and Lee, A. A. 2019. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS Central Science*, 5(9): 1572–1583.
- Segler, M. H.; Kogej, T.; Tyrchan, C.; and Waller, M. P. 2018. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Central Science*, 4(1): 120–131.
- Southey, M. W.; and Brunavs, M. 2023. Introduction to small molecule drug discovery and preclinical development. *Frontiers in Drug Discovery*, 3: 1314077.
- Su, B.; Du, D.; Yang, Z.; Zhou, Y.; Li, J.; Rao, A.; Sun, H.; Lu, Z.; and Wen, J.-R. 2022. A molecular multimodal foundation model associating molecule graphs with natural language. *arXiv preprint arXiv:2209.05481*.
- Wang, R.; and He, K. 2025. Diffuse and Disperse: Image Generation with Representation Regularization. *arXiv preprint arXiv:2506.09027*.
- Weininger, D. 1988. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1): 31–36.
- Weiss, T.; Yanes, E. M.; Chakraborty, S.; Cosmo, L.; Bronstein, A. M.; and Gershoni-Poranne, R. 2023. Guided diffusion for inverse molecular design. *Nature Computational Science*, 3(10): 873–882.
- Westermayr, J.; Gilkes, J.; Barrett, R.; and Maurer, R. J. 2023. High-throughput property-driven generative design of functional organic molecules. *Nature Computational Science*, 3(2): 139–148.
- You, J.; Liu, B.; Ying, Z.; Pande, V.; and Leskovec, J. 2018. Graph convolutional policy network for goal-directed molecular graph generation. *Advances in Neural Information Processing Systems*, 31.
- Zang, C.; and Wang, F. 2020. Moflow: an invertible flow model for generating molecular graphs. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 617–626.
- Zeng, Z.; Yao, Y.; Liu, Z.; and Sun, M. 2022. A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals. *Nature Communications*, 13(1): 862.
- Zhu, H.; Xiao, T.; and Honavar, V. G. 2024. 3M-Diffusion: Latent Multi-Modal Diffusion for Language-Guided Molecular Structure Generation. In *First Conference on Language Modeling*.

APPENDIX

A Attention mask formulation

Algorithm 3: Split Integer with Exponential Decay

```

1: Input:  $S$  (Sample length),  $\alpha$  (decay factor)
2: Output: result (split sizes), cumsum (cumulative steps)
3: if  $\alpha = 1.0$  then
4:    $N \sim \text{UniformInt}(1, S)$ 
5: else
6:    $\text{base} \leftarrow \frac{1-\alpha}{1-\alpha^S}$ 
7:    $p[i] \leftarrow \text{base} \cdot \alpha^i, \quad i = 0 \dots S-1$ 
8:    $N \sim \text{Categorical}(1 \dots S, p)$ 
9: end if
10:  $\text{cumsum} \leftarrow [0] \cup \text{sorted}(\text{Sample}(1 \dots S-1, N-1)) \cup [S]$ 
11:  $\text{result}[i] \leftarrow \text{cumsum}[i+1] - \text{cumsum}[i], \quad i = 0 \dots \text{len}(\text{cumsum}) - 2$ 
12: return result, cumsum

```

Algorithm 4: Construct Attention Mask, m

```

1: Input: sample_len, cond_len, latent_len, split_sizes, cumsum
2: Output: Attention mask  $m$ 
3:  $\text{cond\_len} \leftarrow \text{cond\_len} + \text{latent\_len}$ 
4:  $\text{visible\_len} \leftarrow \text{sample\_len} - \text{split\_sizes}[-1]$ 
5:  $\text{ctx\_len} \leftarrow \text{cond\_len} + \text{visible\_len}$ 
6:  $\text{seq\_len} \leftarrow \text{ctx\_len} + \text{sample\_len}$ 
7: Initialize  $m \leftarrow \mathbf{1}_{\text{seq\_len} \times \text{seq\_len}}$ 
8:  $m[:, : \text{cond\_len}] \leftarrow 0$ 
9: Initialize  $T_1, T_2, T_3$  as all-ones triangular matrices
10: for  $i = 0$  to  $|\text{split\_sizes}| - 2$  do
11:    $T_1[\text{cumsum}[i] : \text{cumsum}[i+1], 0 : \text{cumsum}[i+1]] \leftarrow 0$ 
12:    $T_2[\text{cumsum}[i+1] : \text{cumsum}[i+2], 0 : \text{cumsum}[i+1]] \leftarrow 0$ 
13: end for
14: for  $i = 0$  to  $|\text{split\_sizes}| - 1$  do
15:    $T_3[\text{cumsum}[i] : \text{cumsum}[i+1], \text{cumsum}[i] : \text{cumsum}[i+1]] \leftarrow 0$ 
16: end for
17:  $m[\text{cond\_len} : \text{ctx\_len}, \text{cond\_len} : \text{ctx\_len}] \leftarrow T_1$ 
18:  $m[\text{ctx\_len} : \text{cond\_len} + \text{visible\_len}] \leftarrow T_2$ 
19:  $m[\text{ctx\_len} : \text{ctx\_len} + \text{sample\_len}] \leftarrow T_3$ 
20: return  $m[\text{None}, \text{None}, :, :]$ 

```

We construct the attention mask m in two steps, using Algorithm 3 and Algorithm 4. First, Algorithm 3 generates the autoregressive step structure by returning `split_sizes` and their cumulative indices (`cumsum`). These are then used by Algorithm 4 to construct the final attention mask m , which enforces directional and grouped attention among visible and

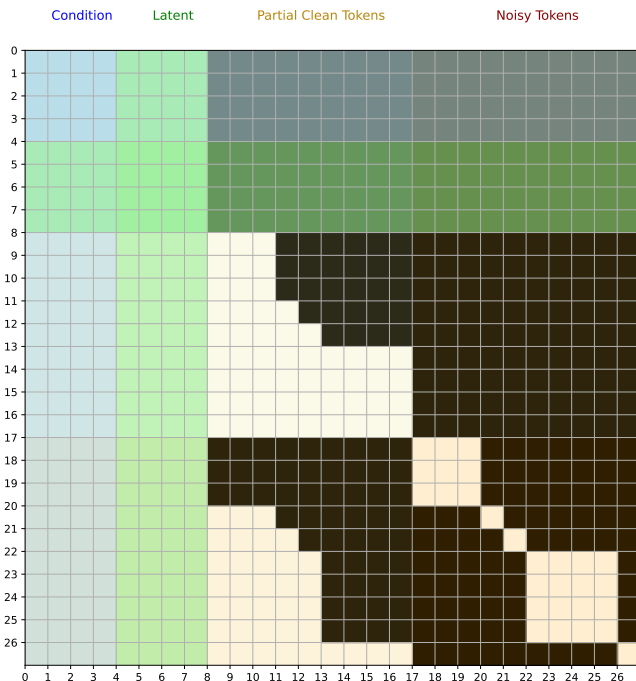


Figure 8: Example of an attention mask with token **lengths**: condition tokens (c) = 4, latent tokens (h) = 4, partial clean tokens (x_p) = 9, and noisy tokens (z) = 10. Black shaded regions indicate blocked attention, while the lighter cells represent allowable attention.

noisy tokens, while allowing free attention from conditional and latent tokens. An illustration of a resulting attention mask is shown in Figure 8.

B Conditional samples

Figure 9 presents visual results of text-conditional molecule generation on the ChEBI-20 dataset. Given detailed molecular descriptions, our model generates chemically meaningful structures that closely align with the input text. The examples include glycosides, chiral diols (enantiomers), substituted benzimidazoles, methionine derivatives with long aliphatic chains, and hydroxylated fatty acids. In all cases, the generated structures capture key functional groups and molecular backbones, demonstrating strong alignment with the semantics of the input descriptions.

C Case Studies

We present visual comparisons of molecules generated under different conditions. Table 5 shows generation results under the anthocyanidin cation condition, where MolSnap consistently produces structures with higher similarity to the ground truth compared to baselines. In Table 6, molecules are sampled based on the drug-likeness condition, with top samples selected by QED scores. MolSnap yields chemically diverse and property-aligned molecules with higher QED, demonstrating its superior conditional generation capability.

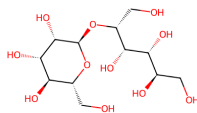
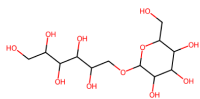
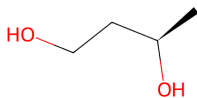
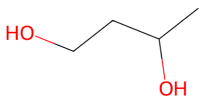
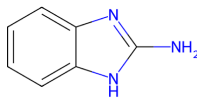
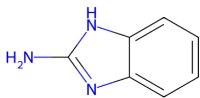
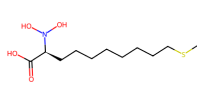
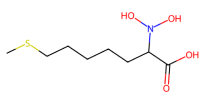
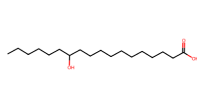
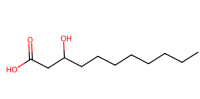
Description	Ground Truth	Generated
The molecule is a glycosyl alditol that is D-mannitol in which the hydroxy group at position 2 has been converted into the corresponding alpha-D-mannopyranoside. It derives from a D-mannitol and an alpha-D-mannose.		
The molecule is a butane-1,3-diol of R-configuration. It is an enantiomer of a (S)-butane-1,3-diol. It derives from a hydride of a butane.		
The molecule is a member of the class of benzimidazoles that is benzimidazole in which the hydrogen at position 2 is replaced by an amino group. It has a role as a marine xenobiotic metabolite.		
The molecule is an N,N-dihydroxy-L-polyhomomethionine in which there are eight methylene groups between the alpha-carbon and sulfur atoms. It is a N,N-dihydroxy-L-polyhomomethionine and a N,N-dihydroxyhexahomomethionine. It is a conjugate acid of a N,N-dihydroxy-L-hexahomomethioninate.		
The molecule is a hydroxy fatty acid that is stearic acid bearing a hydroxy substituent at position 12. It has a role as a plant metabolite and a bacterial xenobiotic metabolite. It is a hydroxyoctadecanoic acid and a secondary alcohol. It is a conjugate acid of a 12-hydroxyoctadecanoate.		

Figure 9: Text conditional molecule generation from ChEBI-20 dataset.

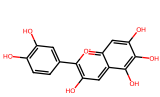
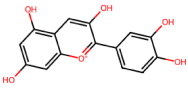
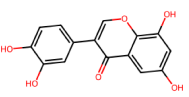
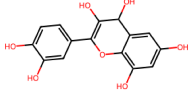
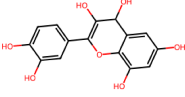
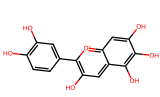
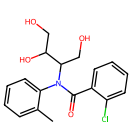
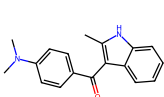
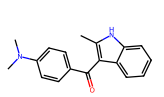
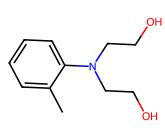
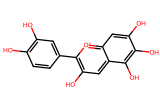
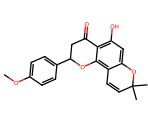
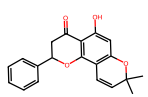
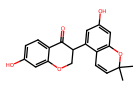
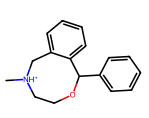
Condition: The molecule is an anthocyanidin cation.					
MolSnap					
	Reference	Sim: 1.0	Sim: 0.93	Sim: 0.90	Sim: 0.90
MolT5-large					
	Reference	Sim: 0.74	Sim: 0.74	Sim: 0.74	Sim: 0.73
3M-Diffusion					
	Reference	Sim: 0.87	Sim: 0.85	Sim: 0.83	Sim: 0.81

Table 5: Comparison of molecules generated by MolSnap, 3M-Diffusion, and MolT5-large under the anthocyanidin cation condition. Higher similarity (Sim.) denotes better alignment with the reference.

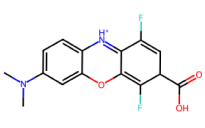
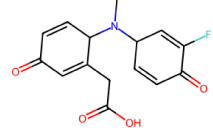
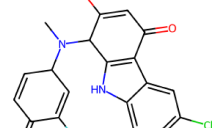
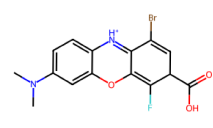
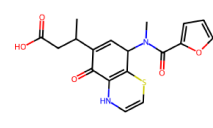
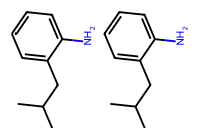
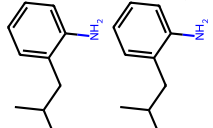
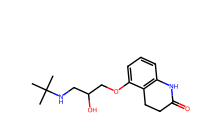
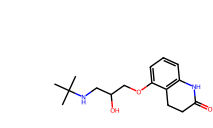
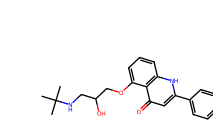
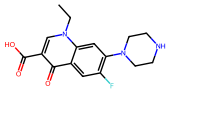
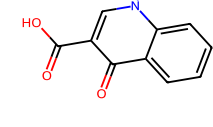
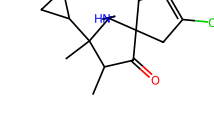
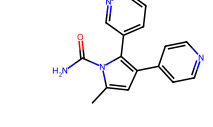
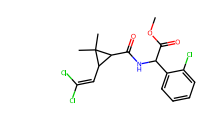
Condition: This molecule is like a drug.					
MolSnap	 QED: 0.86	 QED: 0.84	 QED: 0.84	 QED: 0.83	 QED: 0.82
MolT5-large	 QED: 0.79	 QED: 0.79	 QED: 0.77	 QED: 0.77	 QED: 0.63
3M-Diffusion	 QED: 0.89	 QED: 0.83	 QED: 0.80	 QED: 0.78	 QED: 0.76

Table 6: Comparison of molecules generated by MolSnap, 3M-Diffusion, and MolT5-large under the drug likeliness condition. Top molecules are selected based on desired properties, with drug likeliness measured by QED (higher values indicate better drug likeliness).