MoCDiff: Efficient Motif-Constrained Discrete Diffusion for Molecule Generation

Tasfia Nuzhat Ornee [*1] **Mohammad Jahid Ibna Basher** [1] **Siddhi Kanta Mishra** [1] **Ozlem Garibay** [1]
Ivan Garibay [1] **Niloofar Yousefi** [1]

Abstract

Molecular graph generation requires models that capture chemical structure while producing valid, diverse, and novel molecules within practical sampling budgets. We present MoCDiff, a Motif-Constrained masked Discrete Diffusion framework built on two complementary components: mSENT, a motif-aware graph-to-sequence tokenizer, and an optimized constrained sampler extending Constrained Discrete Diffusion (CDD). The mSENT tokenizer biases graph traversal toward chemically coherent substructures - ring systems, aromatic regions, and strongly coupled bond patterns - so that atoms sharing rigid chemical scaffolds appear at contiguous token positions rather than being scattered by syntax-driven ordering. The constrained sampler combines inexact augmented Lagrangian updates, adaptive penalty scheduling, lazy projection, and cache-based decode checks to concentrate feasibility enforcement near the final decoded molecule, where corrections carry useful chemical signal. Under a matched MDLM backbone on QM9, mSENT raises validity from 85.3% to 90.5% and uniqueness from 75.0% to 97.7% over standard SENT tokenization. On both QM9 and MOSES, the optimized sampler achieves 1.62× and 1.38× lower wall-clock time per candidate respectively, while improving accepted-sample throughput. Together, these results demonstrate that motif-aware serialization and efficient constraint enforcement are critical and complementary design choices for practical discrete diffusion over molecular graphs.

[1]Department of Industrial Engineering, University of Central Florida, Orlando, FL, United States. Correspondence to: Tasfia Nuzhat Ornee <tasfia.ornee@ucf.edu>.

Accepted at the 2026 Workshop on Generative and Agentic AI for Biology (ICML 2026)

1. Introduction

Generating molecules with desired chemical properties requires models that navigate a vast discrete space while producing outputs that are valid, novel, and unique after decoding (Polykovskiy et al., 2020).A natural route is to serialize molecular graphs into token sequences and apply sequence-based generative models, avoiding the complexity of graph native denoising (Vignac et al., 2022) (Yiming et al., 2025) (Zhang et al., 2025). Masked discrete diffusion (Sahoo et al., 2024) (Shi et al., 2025) provides the generative backbone, supporting parallel denoising and continuous-time noise schedules over discrete token vocabularies through standard Transformer architectures (Lee et al., 2025). The practical quality of this pipeline depends on two design choices: how the molecular graph is serialized into tokens, and how feasibility constraints are enforced at sampling time.

The token-order problem. Standard molecular string formats place chemically coupled atoms at distant token positions because their ordering is driven by syntax rules rather than chemical locality. In SMILES (Weininger, 1988), two atoms sharing a benzene ring may be separated by ten or more tokens when depth-first traversal visits a side chain between them. SELFIES (Krenn et al., 2020) encodes molecules through a validity-preserving grammar but inherits a similar traversal order; SAFE (Noutahi et al., 2024) segments by fragment attachment points rather than by chemical rigidity. In each case, the diffusion model must learn short-range chemical dependencies from long-range token correlations. SENT (Chen et al., 2025) provides a reversible, topology-preserving graph-to-sequence encoding, but its traversal order follows generic graph structure: atoms in the same ring system or aromatic core may still appear at non-contiguous positions, leaving the mismatch between molecular structure and token order unresolved.

The constraint-enforcement problem. Chemical validity and uniqueness can only be verified after decoding: they require RDKit sanitization and canonical SMILES deduplication on the fully reconstructed molecule, not on intermediate token distributions. Constrained Discrete Diffusion (CDD) (Cardei et al., 2025) addresses this by projecting

token-probability tensors toward a feasible set at each of the T reverse steps, adding a full decode-and-check pass per step. However, in early reverse steps most tokens are still masked, so decoded molecules are essentially random noise, correcting them wastes compute without improving the final output. Uniform projection across all steps is therefore unnecessarily expensive.

Our approach. We present MoCDiff, which addresses both design choices. mSENT modifies only the SENT traversal policy, biasing it toward chemically coherent substructures - specifically ring systems, aromatic cores, and atom pairs connected by double or triple bonds (collectively, motifs) - with the goal of keeping these rigid fragments contiguous in the token sequence. Because the SENT emission grammar and decoder are unchanged, exact graph decodability is preserved (Proposition B.1). An optimized constrained sampler reduces CDD overhead by skipping projection on early noisy states, using inexact augmented-Lagrangian inner solves, increasing penalties only for actively violated constraints, and caching decoded molecular checks. This concentrates correction effort near the final decoded molecule, where it has the greatest impact on chemical feasibility.
Experiments on QM9 (Ramakrishnan et al., 2014) and MOSES (Polykovskiy et al., 2020) evaluate the two contributions independently. Under a matched MDLM backbone and training budget, mSENT improves generation quality over standard SENT on QM9, raising validity by 5.2 percentage points and uniqueness by 22.7 points. Under the same trained denoiser, the MoCDiff sampler achieves 1.62× lower wall-clock time per candidate and higher accepted-sample throughput than CDD on QM9. We report these as controlled representation and efficiency results; published baselines trained at full scale remain stronger on several metrics. **Contributions:**

- **mSENT**: a motif-aware SENT traversal policy that retains full graph decodability while keeping chemically rigid substructures contiguous in the token sequence.
- **Optimized constrained sampler**: lazy projection, inexact ALM, and violation adaptive penalties, achieving 1.62× lower wall-clock time per candidate and 1.7× higher accepted-sample throughput over CDD on QM9.
- **Controlled experiments** showing mSENT raises QM9 validity from 85.3% to 90.5% under matched training, and the optimized sampler improves accepted-sample throughput on both QM9 and MOSES.

2. Related Works

Fragment based sequence representation. Noutahi et.al (Noutahi et al., 2024) introduced Sequential Attachment-based Fragment Embedding (SAFE), representing SMILES as an unordered sequence of interconnected fragment blocks, while preserving compatibility with existing SMILE parsers. The SAFE representation improved fragment-constrained design tasks such as linker design, scaffold decoration and motif-extension. However, SAFE still functions as a SMILES-compatible line notation, with its structure determined by the chosen fragmentation algorithm.

Discrete diffusion for molecular graph generation. Discrete diffusion models have been explored for graphs and molecular graph generation. Early score-based and continuous diffusion approaches modeled graph distributions through stochastic differential equations and permutation-invariant score matching (Niu et al., 2020; Jo et al., 2022). Models such as DiGress (Vignac et al., 2022), MG-DIFF (Zhang et al., 2025), SparseDiff (Yiming et al., 2025) generate molecules by denoising categorical atom and bond variables in node-edge space. The autoregressive and flow-based graph generators such as GraphAF (Shi et al., 2020) and GraphRNN (You et al., 2018) also demonstrated the effectiveness of structured graph generation for realistic molecular and graph topology modeling. Diffusion models have also been extended to molecular geometry generation and joint structural learning, including equivariant diffusion models for 3D molecular generation (Hoogeboom et al., 2022) and transformer-based molecular representation learning (Xu et al., 2023).
These methods primarily operate directly on graph structures or continuous latent graph representations. In contrast, our goal is to map molecular graphs to a reversible motif-aware sequence representation and perform discrete diffusion in that sequence space.

3. Problem Setup

Let $\mathcal{G}$ be the space of molecular graphs, where each molecule is a labeled graph $G = (V, E)$ with atoms as nodes and bonds as edges. Given a molecular dataset $\mathcal{D} = \{G_i\}_{i=1}^N$, we convert each graph into a motif-aware SENT sequence:

$$z_i = T_{\mathrm{mSENT}}(G_i; m_i, \alpha),$$

where m_i denotes motif assignments and α controls the strength of motif-guided traversal. This gives a sequence dataset $\mathcal{Z} = \{z_i\}_{i=1}^N$, over which we train a masked discrete diffusion model $p_\theta(z)$. During generation, the model samples a sequence and decodes it back into a molecular graph:

$$\hat{z} \sim p_\theta(z), \qquad \hat{G} = T_{\mathrm{mSENT}}^{-1}(\hat{z}).$$

Because not every decoded sequence is chemically feasible or useful, we define a feasible generation set:

$$\mathcal{F}(\mathcal{D}, \mathcal{H}) = \left\{ z \in \mathcal{Z}^* : \begin{array}{l} g_j\big(T^{-1}_{\mathrm{mSENT}}(z); \mathcal{D}, \mathcal{H}\big) \le \tau_j, \\ j = 1, \dots, K \end{array} \right\}$$

Here, $\mathcal{H}$ is the set of molecules generated so far, each g_j measures violation of a constraint such as invalid chemistry, duplication, or overlap with the training set, and τ_j is the allowed tolerance for that constraint.

The ideal goal is to sample sequences from p_θ that lie in $\mathcal{F}$. However, enforcing constraints at every reverse diffusion step is expensive because correction may require decoding, graph reconstruction, chemical sanitization, and uniqueness or novelty checks. Therefore, we aim to apply corrections only at selected reverse timesteps to generate feasible molecules while reducing constraint-enforcement overhead.

4. Method

4.1. Overview

MoCDiff consists of three stages. First, *motif-aware tokenization* converts each molecular graph into a reconstructable sequence of motif-aware graph tokens. This step decomposes the molecule into chemically meaningful substructures and serializes them using the mSENT encoding, so that graph structure can be modeled as a discrete sequence.

Second, *masked discrete diffusion* trains an MDLM-style model (Sahoo et al., 2024) over these graph-token sequences. Starting from masked tokens, the model progressively denoises the sequence and generates candidate molecular graph sequences.

Third, *optimized constraint enforcement* corrects intermediate generated sequences only at selected reverse diffusion steps, rather than at every step, using adaptive projection to encourage chemical validity, novelty and uniqueness. The final corrected sequence is decoded back into a molecular graph.

4.2. Motif Aware SENT (mSENT)

In the original SENT algorithm, a molecule is represented as a labeled graph $G = (V, E)$, where V denotes atoms and E denotes chemical bonds. SENT converts this graph into a reconstructable sequence of graph tokens by encoding traversal steps, node indices, atom and bond types, and non-tree adjacency information. The proposed mSENT variant does not alter the SENT decoding rule; it only modifies the traversal policy used to construct the sequence.

Let $U_t \subseteq V$ denote the set of unvisited atoms at traversal step t, and let

$$\mathcal{N}_t(v) = \mathcal{N}_G(v) \cap U_t$$

be the set of unvisited neighbours of atom v, where $\mathcal{N}_G(v)$ denotes the graph neighbourhood of v in G. To bias the traversal toward chemically coherent regions, we define a *motif-aware traversal policy*

$$\mathbb{P}(v_{t+1} = u \mid v_t, U_t) = \frac{\mathbf{1}[u \in \mathcal{N}_t(v_t)] \exp(\alpha\, \mathbf{1}[\mu(u) = \mu(v_t)])}{\sum_{z \in \mathcal{N}_t(v_t)} \exp(\alpha\, \mathbf{1}[\mu(z) = \mu(v_t)])}$$

where $\mu : V \to \{1, \dots, K\}$ assigns each atom to a motif group and $\alpha > 0$ controls the strength of the within-motif preference. Larger values of α increase the probability of selecting an unvisited neighbour from the same motif as the current atom. In the limit $\alpha \to \infty$, the policy selects a same-motif neighbour whenever one is available, and crosses a motif boundary only when no same-motif unvisited neighbour is reachable from the current atom. This promotes grouping chemically bonded atoms into contiguous token spans while ensuring each transition follows a valid graph edge under SENT.

The mSENT preserves the standard SENT decoding guarantees because it modifies only the traversal policy while leaving the SENT emission grammar and decoder unchanged (Proposition B.1). The full proof is provided in Appendix B.5.

4.3. Continuous-Time Discrete Diffusion

We train the generative model over mSENT token sequences using a continuous-time masked discrete diffusion framework, following MDLM and related masked diffusion formulations (Shi et al., 2025). Let

$$x_0 = (x_0^{(1)}, \dots, x_0^{(L)}), \qquad x_0^{(i)} \in \mathcal{V}_{\mathrm{tok}},$$

denote an mSENT sequence. The forward process independently corrupts each token by replacing it with the absorbing token [MASK]. The marginal corruption distribution is

$$q(x_t^{(i)} \mid x_0^{(i)}) = \begin{cases} 1 - \alpha_t, & x_t^{(i)} = x_0^{(i)}, \\ \alpha_t, & x_t^{(i)} = [\mathrm{MASK}], \\ 0, & \text{otherwise}, \end{cases}$$

where

$$\alpha_t = 1 - \exp\left(- \int_0^t \beta(s)\, ds \right).$$

Thus, x_t is obtained by independently masking each original token with probability α_t.

4.4. Constrained Sampling

We use Constrained Discrete Diffusion (CDD) as a sampling-time correction mechanism. At reverse step t, the denoiser outputs

$$x'_t \in \Delta_{\mathrm{tok}} := \Delta^{L \times |\mathcal{V}_{\mathrm{tok}}|},$$

a token-probability tensor over the mSENT vocabulary. CDD projects this tensor toward the feasible set by solving

$$P_{\mathcal{C}}(x'_t) := \arg\min_{y \in \Delta_{\text{tok}}} \left\{ D_{\text{KL}}(x'_t \| y) : \ \arg\max(y) \in \mathcal{C} \right\}$$

We decompose feasibility as

$$\mathcal{C} = \mathcal{C}_{\text{soft}} \cap \mathcal{C}_{\text{hard}}.$$

The soft criteria are differentiable terms evaluated on relaxed token probabilities. In our implementation, $\mathcal{C}_{\text{soft}}$ includes property-surrogate scores and optional token-level regularizers, such as grammar, token-format, and motif-consistency scores when available. For criterion i,

$$g_i^{\text{rel}}(y) = h_i(\tilde{\phi}(y)), \qquad \Delta_i^{\text{rel}}(y) = \max\big(0,\ g_i^{\text{rel}}(y) - \tau_i\big)\,,$$

where $\tilde{\phi}(y)$ is the Gumbel-Softmax relaxation and h_i is a differentiable scorer. The hard checks are applied only after decoding: validity by RDKit sanitization, uniqueness by canonical-SMILES deduplication, and novelty against the training set.

Following CDD, the relaxed projection minimizes

$$\mathcal{L}_t(y; x'_t, \lambda, \mu) = D_{\text{KL}}(x'_t \| y) + \sum_{i=1}^{m} \lambda_i \Delta_i^{\text{rel}}(y) + \sum_{i=1}^{m} \frac{\mu_i}{2} \left(\Delta_i^{\text{rel}}(y) \right)^2 .$$

MoCDiff keeps this objective but reduces sampling cost through inexact inner solves, conditionally adaptive penalty updates, lazy enforcement at selected reverse steps, and cached hard molecular checks. This concentrates feasibility pressure near the final decoded sequence rather than applying correction uniformly at every timestep.

4.4.1. Inexact ALM Subproblem Solving

Each CDD projection requires solving an augmented-Lagrangian subproblem. Solving this subproblem to convergence at every reverse diffusion step is unnecessary, especially when early reverse states are still highly uncertain. We therefore treat each projection as an inexact ALM update and accept an approximate solution y^{k+1} satisfying

$$\mathcal{L}_t(y^{k+1}; x'_t, \lambda^k, \mu^k) \le \inf_{y \in \Delta_{\text{tok}}} \mathcal{L}_t(y; x'_t, \lambda^k, \mu^k) + \varepsilon_k,$$

where $\varepsilon_k \ge 0$ controls the allowed subproblem error.

In practice, the infimum is not computed. Instead, we use computable stopping criteria: the inner loop stops when the maximum relaxed violation falls below a tolerance or when the relative improvement in $\mathcal{L}_t$ becomes small. This avoids over-solving intermediate denoising states while preserving constraint pressure near the final decoded sequence.

After the inexact primal update, we update multipliers using the hard decoded violation:

$$\lambda_i^{k+1} = \lambda_i^k + \mu_i^k \Delta_i^{\text{hard}}(y^{k+1}).$$

This hybrid update reflects the sampling objective: optimization is performed through a smooth relaxation, but molecular feasibility is ultimately evaluated after discrete decoding.

4.4.2. Conditionally Adaptive Penalty Updates

Standard CDD increases all penalty coefficients uniformly,

$$\mu_i \leftarrow \min(\alpha \mu_i, \mu_{\max}),$$

regardless of which criteria are still violated. MoCDiff instead uses conditionally adaptive penalty updates (CAPU): after hard decoding, only active violations receive increased penalty weight. Let

$$\Delta_{i,k}^{\text{hard}} = \Delta_i^{\text{hard}}(y^k)$$

be the hard decoded violation of criterion i at outer iteration k. For tolerance $\delta \ge 0$,

$$\mu_i^{k+1} = \begin{cases} \min(\alpha_i^k \mu_i^k, \mu_{\max}), & \Delta_{i,k}^{\text{hard}} > \delta, \\ \mu_i^k, & \text{otherwise}, \end{cases}$$

where

$$\alpha_i^k = 1 + (\alpha - 1) \frac{\Delta_{i,k}^{\text{hard}}}{\max_j \Delta_{j,k}^{\text{hard}} + \epsilon}.$$

This update keeps penalties unchanged for satisfied criteria and increases them proportionally to the relative violation magnitude. Clipping by $\mu_{\max}$ prevents excessive stiffness in later projection steps.

4.4.3. Lazy Constraint Enforcement

We use a stage-wise lazy enforcement schedule to avoid correcting highly corrupted intermediate sequences. Let $s \in \{0, \ldots, T-1\}$ index reverse sampling steps, with $s = 0$ the noisiest step, and define

$$\rho_s = \frac{s}{T-1}.$$

Projection is applied according to

$$I_{\text{enf}}(s) = \begin{cases} 0, & \rho_s < 0.2, \\ \mathbf{1}[s \equiv 0 \pmod 3], & 0.2 \le \rho_s < 0.7, \\ 1, & \rho_s \ge 0.7. \end{cases}$$

The final few reverse steps are always projected. Given the denoiser output x'_s, the sampler uses

$$\bar{x}_s = \begin{cases} \text{Proj}_{\text{iALM}}(x'_s), & I_{\text{enf}}(s) = 1, \\ x'_s, & I_{\text{enf}}(s) = 0, \end{cases}$$

and then performs the MDLM reverse update using $\bar{x}_s$. This schedule skips correction in early noisy states, applies periodic correction during intermediate denoising, and enforces feasibility more strongly near the final decoded molecule.

5. Experiments

Datasets and metrics. We evaluate on QM9 (Ramakrishnan et al., 2014), a benchmark of small organic molecules with up to nine heavy atoms, and MOSES (Polykovskiy et al., 2020), a drug-like molecular generation benchmark derived from ZINC. For QM9, we report validity, uniqueness, novelty, and atom stability. For MOSES, we report validity, uniqueness, novelty, filter score, FCD, SNN, and QED.

Validity, uniqueness, novelty, filter score, and atom stability are reported as percentages. FCD, SNN, and QED are reported as unitless scores. Lower FCD indicates that the generated molecular distribution is closer to the reference molecular distribution. A smaller SNN indicates reduced nearest-neighbor similarity to the reference set, suggesting lower memorization, while higher QED indicates greater estimated drug-likeness.

Model. We train a 4-layer, 8-head Transformer denoiser with hidden dimension 256 in the MDLM framework. The denoiser is trained on mSENT token sequences and learns to recover clean molecular sequences from masked inputs. This base model is evaluated first without any constraint enforcement to measure the quality of the learned mSENT-MDLM generator.

5.1. Main Results: Molecular Generation Quality

We evaluate generation quality separately from constrained-sampling efficiency. This distinction is important because unconstrained mSENT+MDLM isolates the effect of the proposed motif-aware tokenization under the same MDLM backbone, while MoCDiff refers to the complete framework with optimized constrained sampling enabled. Consequently, the comparison between SENT+MDLM and mSENT+MDLM evaluates the impact of motif-aware serialization under matched model capacity and training settings. For generation-quality evaluation of unconstrained mSENT+MDLM on QM9 and MOSES, all metrics are computed from 5000 generated molecular samples.

QM9 Generation Quality. Table 1 reports QM9 molecular generation quality under a matched MDLM backbone and training budget. Compared with SENT+MDLM, the proposed mSENT+MDLM improves validity from 85.3% to 90.5%, uniqueness from 75.0% to 97.7%, novelty from 73.3% to 78.1%, and atom stability from 88.7% to 94.7%.These improvements suggest that the motif-aware traversal policy produces graph-token sequences that are more structurally coherent and easier for masked discrete diffusion models to learn than standard SENT serialization.

Table 1. QM9 molecular generation quality. All metrics are percentages. †: published full-training baselines, included for reference. SENT+MDLM and mSENT+MDLM use the same MDLM backbone and training budget.

Model	Valid↑	Uniq.↑	Novel↑	At.Stb↑
DiGress†	95.4	97.6	33.4	98.1
SENT-AR†	97.7	96.7	45.5	98.6
SENT+MDLM	85.3	75.0	73.3	88.7
mSENT+MDLM	90.5	97.7	78.1	94.7

However, these comparisons should be interpreted as controlled matched-compute evaluations rather than state-of-the-art claims. Published full-training baselines remain stronger on several QM9 metrics; for example, DiGress reports 95.4% validity, while SENT-AR reports 97.7%. Therefore, our primary claim is that mSENT provides consistent generation-quality improvements over standard SENT tokenization within the same MDLM training configuration.

MOSES generation quality. Table 2 reports MOSES molecular generation quality. The proposed mSENT+MDLM achieves strong validity (92.0%), high novelty (99.7%), and a high QED score (0.83), indicating that the generated molecules are chemically diverse and maintain favorable drug-like properties. In particular, the near-perfect novelty together with the lowest SNN score (0.45) suggests that the generated samples are less dominated by nearest-neighbor training examples and explore broader regions of chemical space beyond simple memorization.
However, this increased exploration is accompanied by a higher FCD score (2.12) compared to prior baselines, indicating lower global alignment with the reference MOSES molecular distribution. This behavior is consistent with the trade-off commonly observed in molecular generation benchmarks, where highly novel and structurally diverse samples often deviate from the training distribution and therefore yield higher distribution-matching distances.

Although the proposed motif-aware sequence representation improves novelty and diversity under unconstrained generation, the filter score (88.7%) remains below the strongest published baselines, which typically achieve near-perfect filter compliance. One possible explanation is that the unconstrained motif-aware diffusion process encourages exploration of chemical regions that are farther from the training distribution, as reflected by the near-perfect novelty and low

Table 2. MOSES molecular generation quality. †: published reference baselines. The mSENT+MDLM row reports unconstrained motif-aware sequence generation. Lower FCD indicates closer global alignment with the reference molecular distribution, while lower SNN indicates reduced nearest-neighbor similarity to the reference set.

Model	Valid↑	Uniq.↑	Novel↑	Filt.↑	FCD↓	SNN↓	QED↑
VAE†	97.7	99.8	69.5	99.7	0.57	0.58	–
JT-VAE†	100.0	100.0	99.9	97.8	1.00	0.53	–
GraphINVENT†	96.4	99.8	–	95.0	1.22	0.54	–
DiGress†	85.7	100.0	95.0	97.1	1.19	0.52	–
SENT-AR†	87.4	100.0	85.9	98.6	0.91	0.55	–
mSENT+MDLM	92.0	94.6	99.7	88.7	2.12	0.45	0.83

SNN score. While this improves diversity, it also increases the likelihood of generating chemically unusual or heuristically unfavorable structures that do not satisfy the MOSES filter criteria.

We therefore do not claim state-of-the-art MOSES performance. Instead, the results suggest that the proposed motif-aware serialization promotes broader chemical exploration under unconstrained generation, while additional constrained decoding, stronger chemical priors, or larger-scale training may be required to further improve distributional fidelity, uniqueness, and filter compliance.

5.2. Constrained Sampling: Efficiency Analysis

We next evaluate the full MoCDiff sampler. This experiment uses the same mSENT+MDLM backbone for both samplers and compares the original CDD projection routine against our optimized constrained sampler with inexact ALM solves, conditionally adaptive penalty updates, lazy enforcement, and cached hard molecular checks. Since the original CDD implementation is not publicly available, we re-implemented the baseline following the projection formulation and algorithmic description provided in the original paper (Cardei et al., 2025).

Both methods use the same soft validity proxy and hard validity/uniqueness acceptance criteria, so the comparison isolates the sampling-time efficiency contribution. Generation quality is evaluated using validity, uniqueness, and novelty. During constrained sampling, however, acceptance is defined only by validity and uniqueness constraints, while novelty is evaluated separately for unconstrained generation and is not enforced as an online acceptance criterion. For each experiment, 500 molecular candidates are generated for evaluation.

Table 3 reports sampling efficiency. On QM9, MoCDiff reduces total generation time from 621.33s to 383.42s, improving candidate throughput from 0.80 to 1.30 molecules/s and yielding a $1.62\times$ speedup over the baseline CDD sampler. The accepted-sample throughput also increases from 0.72 to 1.23 molecules/s while improving acceptance from 89.20% to 94.40%. On MOSES, MoCDiff reduces generation time from 531.22s to 385.83s, increasing candidate throughput from 0.94 to 1.30 molecules/s and achieving a $1.38\times$ speedup. Accepted-sample throughput improves from 0.83 to 1.18 molecules/s, while acceptance increases from 88.40% to 91.00%.

Table 3. Constrained-sampling efficiency on QM9 and MOSES. Time is total wall-clock time for 500 candidates; Cand/s is raw candidate throughput; Acc% and Acc/s are computed after hard validity and uniqueness checks.

Method	Data	Time (s)↓	Cand/s↑	Acc%↑	Acc/s↑	Speedup↑
CDD	QM9	621.33	0.80	89.20	0.72	1.00×
MoCDiff	QM9	**383.42**	**1.30**	**94.40**	**1.23**	**1.62×**
CDD	MOSES	531.22	0.94	88.40	0.83	1.00×
MoCDiff	MOSES	**385.83**	**1.30**	**91.00**	**1.18**	**1.38×**

The improvement comes from reducing redundant projection overhead during constrained reverse diffusion. The baseline CDD sampler applies stronger and more frequent constraint corrections across denoising steps, which increases optimization cost and can perturb partially formed molecular structures. In contrast, MoCDiff uses inexact ALM updates, adaptive penalties, and lazy constraint enforcement to avoid unnecessary projections and delay aggressive corrections until later denoising stages when token distributions become more stable. This reduces computational overhead while preserving feasible molecular trajectories, leading to both higher throughput and improved acceptance.

6. Conclusion

We presented MoCDiff, a motif-constrained masked discrete diffusion framework for molecular graph generation that jointly improves molecular sequence representation and constrained sampling efficiency. The proposed method combines mSENT, a motif-aware graph-to-sequence tokenizer for structurally coherent molecular serialization, with an optimized constrained diffusion sampler based on inexact ALM updates, adaptive penalties, and lazy constraint enforcement. Under a matched MDLM backbone, mSENT improves generation quality by better aligning molecular substructures with token order, while the optimized sampler reduces constrained-sampling overhead and improves accepted-sample throughput. Experiments on QM9 and MOSES demonstrate that motif-aware serialization and efficient constraint enforcement are complementary design choices for discrete diffusion over molecular graphs. Although the proposed framework improves validity, diversity, and sampling efficiency, performance remains below the strongest published baselines on several metrics, particularly under limited training and reduced projection budgets. Future work will explore larger-scale training, stronger property constraints, and property-guided molecular generation.

Impact Statement

This work investigates motif-aware discrete diffusion for molecular graph generation, with potential applications in AI-driven molecular design, drug discovery, and scientific generative modeling. By improving the structural coherence of graph-to-sequence representations and reducing the computational overhead of constrained sampling, the proposed framework may support more efficient exploration of chemical space and accelerate early-stage molecular generation workflows.

The methods presented in this paper are intended for research purposes in machine learning and computational molecular design. While no immediate harmful societal impacts are anticipated, generative molecular models may potentially be misused for the design of hazardous or biologically harmful compounds if deployed without appropriate safeguards. In this work, experiments are limited to standard benchmark datasets and do not involve optimization toward toxic, harmful, or controlled chemical agents.

More broadly, this work aims to contribute to the development of efficient and reliable generative modeling methods for scientific applications, while encouraging responsible use of AI systems in molecular generation research.

References

Cardei, M., Christopher, J. K., Hartvigsen, T., Kailkhura, B., and Fioretto, F. Constrained discrete diffusion. *arXiv preprint arXiv:2503.09790*, 2025.

Chen, D., Krimmel, M., and Borgwardt, K. Flatten graphs as sequences: Transformers are scalable graph generators. *arXiv preprint arXiv:2502.02216*, 2025.

Hoogeboom, E., Satorras, V. G., Vignac, C., and Welling, M. Equivariant diffusion for molecule generation in 3d. In *International Conference on Machine Learning (ICML)*, 2022.

Jo, J., Lee, S., and Hwang, S. J. Score-based generative modeling of graphs via the system of stochastic differential equations. In *International Conference on Machine Learning (ICML)*, 2022.

Krenn, M., Häse, F., Nigam, A., Friederich, P., and Aspuru-Guzik, A. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1(4):045024, 2020.

Lee, S., Kreis, K., Veccham, S. P., Liu, M., Reidenbach, D., Peng, Y., Paliwal, S., Nie, W., and Vahdat, A. Genmol: A drug discovery generalist with discrete diffusion. *arXiv preprint arXiv:2501.06158*, 2025.

Niu, C., Song, Y., et al. Permutation invariant graph generation via score-based generative modeling. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

Noutahi, E., Gabellini, C., Craig, M., Lim, J. S., and Tossou, P. Gotta be safe: a new framework for molecular design. *Digital Discovery*, 3(4):796–804, 2024.

Polykovskiy, D., Zhebrak, A., Sanchez-Lengeling, B., Golovanov, S., Tatanov, O., Belyaev, S., Kurbanov, R., Artamonov, A., Aladinskiy, V., Veselov, M., Kadurin, A., Johansson, S., Chen, H., Nikolenko, S., Aspuru-Guzik, A., and Zhavoronkov, A. Molecular sets (moses): A benchmarking platform for molecular generation models. *Frontiers in Pharmacology*, 11:565644, 2020.

Ramakrishnan, R., Dral, P. O., Rupp, M., and von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1:140022, 2014.

Sahoo, S. S., Arriola, M., Schiff, Y., Gokaslan, A., Marroquin, E., Chiu, J. T., Rush, A. M., and Kuleshov, V. Simple and effective masked diffusion language models. In *Advances in Neural Information Processing Systems*, 2024.

Shi, C. et al. Graphaf: A flow-based autoregressive model for molecular graph generation. In *International Conference on Learning Representations (ICLR)*, 2020.

Shi, J., Han, K., Wang, Z., Doucet, A., and Titsias, M. K. Simplified and generalized masked diffusion for discrete data. In *International Conference on Learning Representations*, 2025.

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

Weininger, D. 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, 1988.

Xu, M. et al. One transformer can understand both 2d 3d molecular data. In *International Conference on Learning Representations (ICLR)*, 2023.

Yiming, Q., Vignac, C., and Frossard, P. Sparsediff: Sparse discrete diffusion for scalable graph generation. *Transactions on Machine Learning Research*, 2025.

You, J., Ying, Z., Ren, X., Hamilton, W., and Leskovec, J. Graphrnn: Generating realistic graphs with deep autoregressive models. In *International Conference on Machine Learning (ICML)*, 2018.

Zhang, X., Wang, S., Fang, Y., and Zhang, Q. Mg-diff: A novel molecular graph diffusion model for molecular generation and optimization. *Plos one*, 20(10):e0331450, 2025.

A. Reproducibility Details

This appendix provides the implementation details and hyperparameters used in our experiments. The main paper reports only the essential setup. Unless stated otherwise, all experiments use the same MDLM backbone and differ only in the tokenizer or sampling-time correction strategy.

A.1. Tokenizer and Data Configuration

We use SENT or motif-aware SENT tokenization depending on the experiment. For mSENT, motif guidance is enabled with $\alpha = 2.0$. We do not use a hard same-motif-first rule; instead, motif information biases traversal while keeping the emitted sequence compatible with the original SENT decoding format. The tokenizer appends an end-of-sequence token and reserves an additional token id for $[\mathrm{MASK}]$.

A.2. Training Configuration

The denoiser is trained with a BERT-style masked language model backbone inside the MDLM framework. Training uses pretokenized graph-token sequences and attention masks constructed from non-padding positions. We train with PyTorch Lightning on a single GPU.

Table 4. Model and training configuration.

Parameter	Value
Backbone	BERT-style masked LM
Transformer layers	4
Attention heads	8
Hidden dimension	256
Feed-forward dimension	1024
Dropout	0.2
Maximum sequence length	256
Optimizer	AdamW
Learning rate	3×10^{-4}
Weight decay	0.01
Random seed	42

A.3. Sampling Configuration

For unconstrained generation, samples are drawn directly from the trained MDLM model and decoded into molecular graphs. All unconstrained sampling experiments use temperature 0.8, randomness 0.5, chunk size 8, and 5000 generated samples. QM9 uses masked length 32, while MOSES uses masked length 64 to account for longer molecular sequences. Generation quality is evaluated using validity, uniqueness, novelty, filtering score, FCD, SNN, and QED.

For constrained generation, we evaluate sampling efficiency and constrained acceptance behavior using 500 generated molecular candidates. In these experiments, hard acceptance requires validity and uniqueness, while novelty is reported separately for unconstrained generation quality evaluation and is not enforced as an online acceptance criterion. The original CDD baseline applies augmented-Lagrangian projection during sampling. Since the original CDD implementation is not publicly available, we re-implemented the baseline following the projection formulation and algorithmic description provided in the original paper (Cardei et al., 2025). MoCDiff modifies the sampling-time correction strategy through inexact ALM solves, conditionally adaptive penalty updates, lazy enforcement, and cached hard molecular checks.

For fairness, both samplers share the same main ALM settings: $\lambda_0 = 0.0$, $\mu_0 = 1.0$, $\mu_{\max} = 1000.0$, and Gumbel-Softmax temperature 1.0. For computational tractability, the constrained-sampling experiments use a reduced ALM budget of 10 outer iterations and 5 inner gradient steps per projection. This budget is smaller than the full projection setting used in the original CDD molecular experiments. We leave larger-budget evaluation and systematic budget–quality tradeoff analysis to future work.

Algorithm 1 Motif-aware SENT tokenization (mSENT)

Require: Molecular graph $G = (V, E, \ell_V, \ell_E)$, motif assignment $\mu : V \to \{1, \ldots, K\}$, motif bias $\alpha > 0$
Ensure: Token sequence x and optional aligned metadata (m_i, s_i, r_i)

$U \leftarrow V$, $x \leftarrow [\,]$, $q \leftarrow 0$
Choose an initial atom $v \in U$; $U \leftarrow U \setminus \{v\}$
$q \leftarrow q + 1$
Start segment q and emit the labelled SENT tuple for $(v, \emptyset)$
while $U \neq \emptyset$ **do**
  $C \leftarrow \mathcal{N}_G(v) \cap U$
  if $C \neq \emptyset$ **then**
    Sample $u \in C$ according to

$$\Pr(u \mid v, U) = \frac{\exp(\alpha\, \mathbf{1}[\mu(u) = \mu(v)])}{\sum_{z \in C} \exp(\alpha\, \mathbf{1}[\mu(z) = \mu(v)])}.$$

    $A_u \leftarrow (\mathcal{N}_G(u) \setminus \{v\}) \cap (V \setminus U)$
    Emit the traversal bond token $\ell_E(v, u)$
    Emit the labelled SENT tuple for (u, A_u)
    Record optional metadata for emitted tokens using motif id $\mu(u)$, segment id q, and role type
    $U \leftarrow U \setminus \{u\}$
    $v \leftarrow u$
  else
    Emit segment-break token `[RESET]`
    Choose a new start atom $v \in U$; $U \leftarrow U \setminus \{v\}$
    $q \leftarrow q + 1$
    $A_v \leftarrow \mathcal{N}_G(v) \cap (V \setminus U)$
    Emit the labelled SENT tuple for (v, A_v)
    Record optional metadata for emitted tokens using motif id $\mu(v)$, segment id q, and role type
  end if
end while
Prepend `[SOS]` and append `[EOS]` to x
Assign role labels to token positions:

$$r_i \in \{\texttt{special}, \texttt{syntax}, \texttt{interior}, \texttt{interface}\}$$

return x and optional aligned metadata (m_i, s_i, r_i)

B. Additional Details on Motif-Aware SENT

This appendix provides the full motif-aware SENT (mSENT) serialization procedure, together with optional token-aligned metadata and a simple decodability statement.

B.1. mSENT Serialization Algorithm

Algorithm 1 summarizes the mSENT construction procedure. The key difference from standard SENT is the traversal policy: among unvisited neighbours, mSENT biases the next step toward atoms belonging to the same motif. The emitted token format remains the original SENT format, so the standard SENT decoder can still be used.

B.2. Motif Partitioning

Let $G = (V, E)$ be a molecular graph with atom and bond labels. We construct a motif partition by selecting chemically rigid or strongly coupled bonds. The selected edge set is defined as

$$E_{\text{sel}} = \{(u, v) \in E : \text{InRing}(u, v) \vee \text{BondType}(u, v) \neq \text{SINGLE}\}.$$

where $\mathrm{InRing}(u, v)$ indicates whether the bond (u, v) belongs to a ring. Edges incident to hydrogen atoms are excluded when constructing E_{sel}, so that motif blocks are determined primarily by the heavy-atom scaffold.

The connected components of the selected subgraph

$$G_{\mathrm{sel}} = (V, E_{\mathrm{sel}})$$

define multi-atom motif blocks. Atoms not connected to any selected edge, including hydrogens, are assigned singleton motifs. This induces a partition

$$\mathcal{B} = \{B_1, \ldots, B_M\}, \qquad \bigsqcup_{j=1}^{M} B_j = V,$$

and a motif assignment function

$$\mu : V \to \{0, \ldots, M-1\}.$$

This partition groups ring systems, aromatic regions, and atoms connected by double or triple bonds into shared motif blocks, while leaving flexible single-bond aliphatic regions as smaller units or singleton motifs. The resulting motif assignment is used only to bias the traversal order; the underlying SENT encoding remains reversible because no graph connectivity or label information is removed.

B.3. Optional Token-Aligned Metadata

The mSENT tokenizer may optionally return metadata aligned with the token sequence. For each token position i, we store

$$m_i = \text{motif id}, \qquad s_i = \text{motif-span id}, \qquad r_i = \text{role label}.$$

These annotations are not used as model inputs in the present work; the diffusion model is trained only on token identities. We retain them for diagnostic analysis and possible future extensions, such as motif-aware masking, span-level conditioning, or interface-specific constraint enforcement.

The optional role label is defined as

$$r_i \in \{\texttt{special}, \texttt{syntax}, \texttt{interior}, \texttt{interface}\}.$$

The label `special` is assigned to global control tokens such as [`SOS`], [`EOS`], and [`PAD`]. The label `syntax` is assigned to SENT grammar tokens such as segment breaks and neighbourhood delimiters. The label `interior` is assigned to atom or bond tokens whose associated atoms lie within the same motif, whereas `interface` is assigned to positions corresponding to cross-motif attachments.

B.4. Ablation on Motif-Bias Strength

We evaluate the effect of the motif-bias parameter α in mSENT tokenization on QM9. The parameter α controls how strongly the traversal policy prefers transitions between atoms belonging to the same motif. When $\alpha = 0$, the tokenizer reduces to a traversal policy without explicit motif preference, while larger values increasingly encourage motif-contiguous token ordering.

For analysis, we report the fraction of transitions occurring within the same motif (Same-motif Trans.), the average length of consecutive within-motif traversal segments (Motif Run Len.), the number of traversal resets (Reset Count), and the average resulting token sequence length (Seq. Len.).

Table 5 shows that increasing α consistently improves motif locality in the generated traversal sequences. The same-motif transition fraction increases from 0.2335 at $\alpha = 0$ to 0.2929 at $\alpha = 4.0$, while the average motif run length increases from 1.310 to 1.416. These results indicate that stronger motif bias produces more contiguous motif-aware token spans, improving local chemical coherence in the serialized sequence.

The improvement begins to saturate beyond $\alpha = 2.0$. Increasing α from 2.0 to 4.0 yields only marginal gains in same-motif transitions and motif run length, while reset count and sequence length remain nearly unchanged across all settings. This suggests that motif-aware traversal reorganizes token locality without substantially increasing serialization complexity or traversal instability.

Based on this tradeoff, we use $\alpha = 2.0$ as the default setting throughout the experiments. This value captures most of the motif-locality improvement while avoiding unnecessarily strong traversal bias.

Table 5. Sensitivity of mSENT tokenization to the motif-bias strength α on QM9. Larger values encourage stronger motif-contiguous traversal behavior.

α	Same-motif Trans. ↑	Motif Run Len. ↑	Reset Count	Seq. Len.
0.0	0.2335	1.310	7.534	89.83
0.5	0.2557	1.350	7.581	90.31
1.0	0.2607	1.359	7.557	90.26
2.0	0.2890	1.408	7.582	90.84
3.0	0.2915	1.413	7.583	90.86
4.0	0.2929	1.416	7.577	90.83

B.5. Proof of Proposition B.1

Proposition B.1 (mSENT Decodability)**.** *Under the standard SENT emission and decoding rules, the mSENT tokenization procedure produces a valid SENT sequence. Ignoring optional motif metadata, the standard SENT decoder reconstructs a labelled molecular graph isomorphic to the original graph G.*

Proof. Let

$$G = (V, E, \ell_V, \ell_E)$$

be a labelled molecular graph, where ℓ_V and ℓ_E denote atom and bond labels. Let

$$x = \mathrm{mSENT}(G; \mu, \alpha)$$

be the token sequence produced by the mSENT procedure for a motif assignment $\mu : V \to \{1, \dots, K\}$ and motif-bias parameter $\alpha > 0$.

Let $U_t \subseteq V$ denote the set of unvisited atoms at step t, and let v_t be the current atom. mSENT selects the next atom from

$$N_G(v_t) \cap U_t.$$

Therefore, for every traversal transition $v_t \to v_{t+1}$, we have

$$v_{t+1} \in N_G(v_t),$$

which implies

$$(v_t, v_{t+1}) \in E.$$

Hence every traversal edge emitted by mSENT is a valid edge of G.

For each newly visited atom v, mSENT emits a causal neighborhood set of the form

$$A_v \subseteq N_G(v) \cap V_{<v},$$

where $V_{<v}$ denotes the set of atoms visited before v. Therefore, for every $u \in A_v$,

$$(u, v) \in E.$$

Thus every neighborhood edge emitted by mSENT is also a valid edge of G. Consequently, the decoded edge set E' satisfies

$$E' \subseteq E.$$

We now show the reverse inclusion. Let $(u, v) \in E$. Without loss of generality, suppose u is visited before v. If v is reached directly from u during traversal, then (u, v) is emitted as a traversal edge. Otherwise, when v is emitted, u is already a previously visited neighbor of v. By the SENT causal neighborhood emission rule, u is included in the neighborhood set A_v. Hence (u, v) is emitted as a neighborhood edge. Therefore every edge of G appears in the emitted SENT sequence, and

$$E \subseteq E'.$$

Combining both inclusions gives

$$E' = E.$$

The same argument applies to labels. For every emitted atom v, mSENT outputs the original atom label $\ell_V(v)$, and for every emitted edge (u, v), it outputs the original bond label $\ell_E(u, v)$. Hence the decoded labelled graph preserves both atom and bond labels.

The SENT tokenization reindexes vertices according to their first occurrence in the sequence. Let

$$\pi : V \to V'$$

denote this first-occurrence reindexing map. Since each atom is visited exactly once as a newly introduced vertex, π is a bijection. From $E' = E$, we obtain

$$(u, v) \in E \quad \Longleftrightarrow \quad (\pi(u), \pi(v)) \in E'.$$

Moreover,

$$\ell_V(v) = \ell'_V(\pi(v)), \qquad \ell_E(u, v) = \ell'_E(\pi(u), \pi(v)).$$

Thus π is a labelled graph isomorphism between G and the graph decoded from x.

Finally, the optional motif metadata is not an argument of the standard SENT decoder. Removing this metadata leaves the SENT token sequence unchanged. Therefore it does not affect the decoded graph.

Hence the mSENT tokenization procedure produces a valid SENT sequence and decodes to a labelled molecular graph isomorphic to G. □

C. Sampling Efficiency Ablation

Table 6 presents the ablation study of the proposed constrained sampling optimizations using 200 generated molecular candidates. Due to the higher computational cost of repeated constrained sampling runs, the ablation study uses a smaller evaluation set than the main constrained-sampling efficiency benchmark in Table 3, which uses 500 generated molecular candidates. The baseline CDD sampler achieves 95.00% acceptance with a runtime of 248.72s and a candidate throughput of 0.81 molecules/s. Replacing the exact projection updates with only inexact ALM reduces runtime to 220.61s and improves throughput to 0.90 molecules/s, but slightly decreases acceptance to 92.50%. This suggests that approximate optimization alone can reduce projection overhead, but without additional stabilization mechanisms it may weaken constrained trajectory consistency during reverse diffusion.

In contrast, the full MoCDiff sampler, which combines inexact ALM, conditionally adaptive penalty updates (CAPU), and lazy constraint enforcement, substantially reduces runtime from 248.72s to 148.15s while also improving acceptance from 95.00% to 96.50%. MoCDiff achieves the highest candidate throughput (1.34 molecules/s) and the highest accepted molecules per second (1.29), corresponding to approximately $1.68\times$ higher accepted throughput than the baseline CDD sampler. These results suggest that the efficiency improvement emerges from the coordinated interaction of all three components: lazy enforcement reduces unnecessary projection operations, inexact ALM lowers the optimization cost per projection step, and CAPU stabilizes approximate constrained optimization under reduced enforcement frequency. Together, these components enable substantially faster constrained sampling while preserving stable molecular validity and acceptance quality.

Table 6. Ablation study of the proposed constrained sampling optimizations on QM9 using 200 generated molecular candidates. Acceptance is defined as valid and unique molecules.

Configuration	Inexact ALM	CAPU	Lazy	Time (s)↓	Cand/s↑	Acc%↑	Acc/s↑
CDD baseline	×	×	×	248.72	0.81	95.00	0.76
+ Inexact only	✓	×	×	220.61	0.90	92.50	0.83
Full MoCDiff	✓	✓	✓	148.15	1.34	96.50	1.29

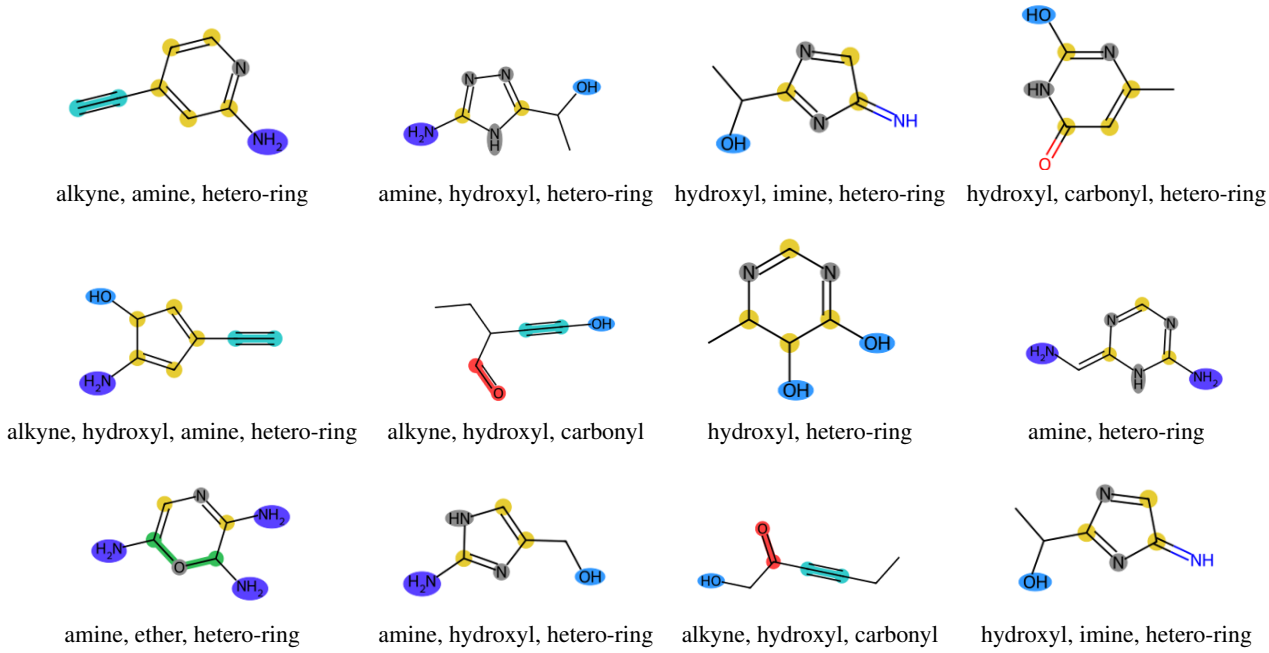

Figure 1. Representative molecules generated from a QM9 training checkpoint using the proposed mSENT representation with an MDLM backbone. The samples exhibit diverse combinations of functional groups and heterocyclic motifs, including alkynes, hydroxyl groups, carbonyl groups, ethers, amines, imines, and nitrogen-containing rings.

D. Qualitative motif diversity analysis

To further analyze the effect of motif-aware sequence tokenization, we visualize representative molecules generated from QM9 and MOSES training checkpoints in Figures 1 and 2. For better qualitative interpretation, we additionally highlight chemically meaningful motifs, including alkynes, carbonyl groups, hydroxyl groups, amines, amides, ethers, halogens, sulfur-containing groups, aromatic rings, and nitrogen-containing hetero-rings.

The QM9 samples exhibit structurally valid small molecules with diverse functional-group compositions, including oxygen-rich chains, nitrogen-containing heterocycles, mixed ring and non-ring topologies, and combinations of carbonyl, hydroxyl, nitrile, and amine motifs. Although some repeated scaffold patterns remain visible, particularly among alkyne-containing structures, the generated molecules still show substantial motif-level variation rather than collapsing to a single repeated topology.

The MOSES samples further demonstrate that the proposed motif-aware serialization generalizes to larger and more drug-like molecular structures. The generated molecules preserve coherent aromatic systems, amide linkers, hetero-rings, ether groups, halogenated substituents, and sulfur-containing pharmacophore motifs. Compared with the QM9 samples, the MOSES generations exhibit richer scaffold-level diversity and more medicinal-chemistry-like motif compositions, including multi-ring structures and mixed functionalized linkers.

Overall, the qualitative results suggest that motif-aware sequence serialization preserves chemically coherent substructures during discrete diffusion sampling and encourages compositionally meaningful motif combinations across both small-molecule and drug-like molecular regimes. At the same time, some recurring scaffold patterns remain visible among frequently occurring motifs, indicating that additional diversity-aware decoding or larger-scale training may further improve global scaffold coverage.

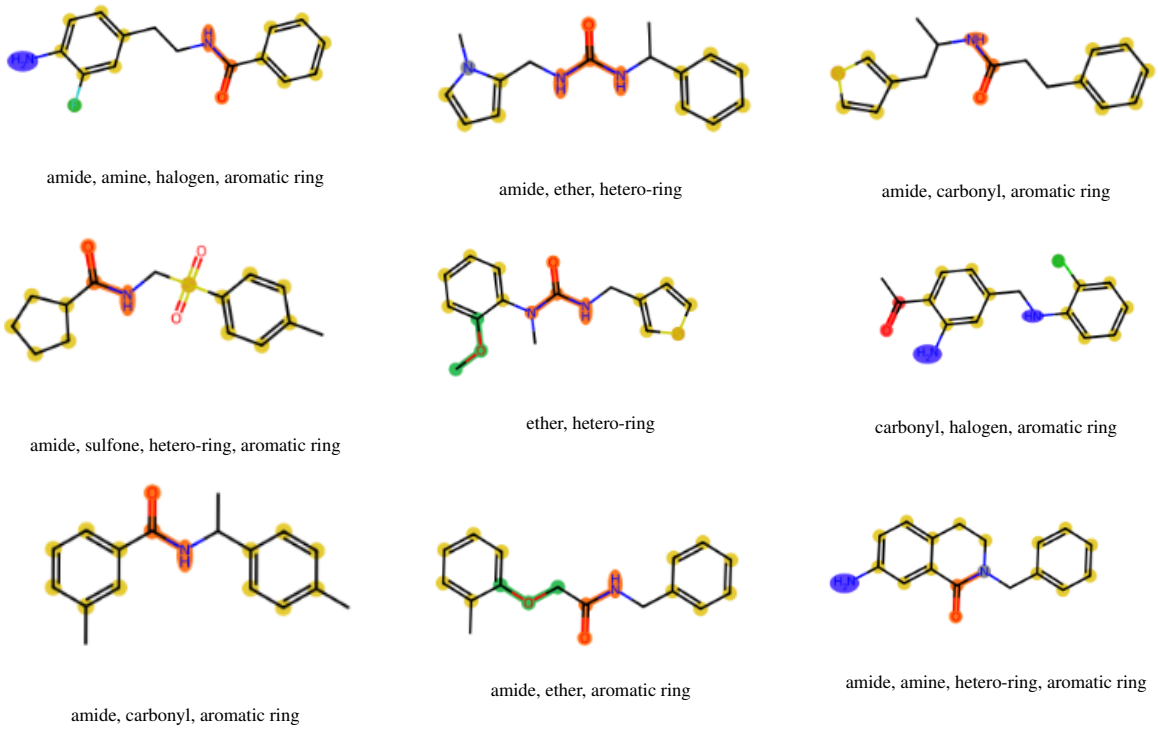

Figure 2. Representative molecules generated from the MOSES checkpoint using the proposed motif-aware sequence tokenization. Highlighted motifs include amides, aromatic rings, hetero-rings, ethers, carbonyl groups, halogenated substituents, sulfur-containing groups, and amine motifs. The generated samples exhibit structurally coherent and drug-like scaffold compositions across multiple motif families.