Learning Generative Dynamics for 3D Molecule Generation via Sequential Neuro-Symbolic Constraints

Zhuoran Li, Xu Sun, Chang Wen Chen, Wanyu Lin*
Department of Computing
The Hong Kong Polytechnic University, Hong Kong SAR, China
zhuor5.li@connect.polyu.hk, dsai-xu.sun@connect.polyu.hk,
changwen.chen@polyu.edu.hk, wan-yu.lin@polyu.edu.hk

Abstract

Diffusion models have recently emerged as a powerful paradigm for generating 3D molecules. Despite showing remarkable promise, they learn atomic distributions implicitly from data, without explicitly encoding physicochemical rules into their generative dynamics. Consequently, generated samples can be statistically plausible yet chemically invalid. To address this problem, we propose Sequential Neuro-Symbolic Constrained Diffusion (SensDiff), a framework that integrates symbolic rules into generative modeling, enabling the model to internalize scientific priors during training. Central to SensDiff is the observation that the diffusion trajectory of 3D molecule generation exhibits a coarse-to-fine evolution: valid molecules arise by first mitigating macroscopic geometric conflicts and then refining structures toward microscopic chemical consistency through iterative denoising. Accordingly, SensDiff mirrors this dynamic by sequentially enforcing constraints with adaptive weighting during training, progressively steering generation toward valid molecular geometries. Moreover, SensDiff translates domain knowledge into interpretable generative controls, grounding opaque neural denoising in scientific principles. Experiments on molecular benchmarks and real-world constrained generation tasks confirm that SensDiff consistently improves chemical validity while adhering to physicochemical rules and user-specified targets. The code is publicly available at https://github.com/WanyuGroup/NeurIPS2026_SensDiff.

1 Introduction

Molecule generation is a central task across various scientific domains, such as drug discovery, materials engineering, and protein design [78, 72, 66]. Since molecules naturally exist in three-dimensional space and their properties are largely governed by spatial configuration, it is essential to capture accurate molecular geometry in generative modeling [74, 75, 49]. Despite considerable progress, producing 3D structures that are both geometrically realistic and chemically valid remains a critical challenge due to the high-dimensional, constraint-rich nature of molecular space [69, 64].

Deep generative models, particularly diffusion models, have been adapted to tackle the challenge of 3D molecule generation [30, 50]. Specifically, they gradually denoise atomic coordinates and types, enabling stable modeling of molecular distributions and delivering encouraging empirical results [20, 47, 52]. However, learning diffusion models solely from observed data tends to implicitly approximate physicochemical rules rather than internalize them within the generative process, inevitably yielding molecules that conform to the learned distribution but violate fundamental scientific principles.

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Motivated by these limitations, recent studies have attempted to equip diffusion models with scientific priors for 3D molecule generation. Among them, equivariant architectures [30, 42, 12, 69] enforce spatial symmetry by construction but acquire chemical feasibility only implicitly from data. Joint discrete-continuous methods [64, 55, 33, 31] co-denoise atomic features and coordinates in a unified process, yet their learned coupling offers no formal physicochemical assurance. Inference-time guidance [24, 5, 50, 38] injects external corrective signals during sampling without incorporating them into the training objective, and thus is not absorbed into the model parameters. Latent-space approaches [69, 49, 71] lower generative complexity via dimensionality reduction, but at the cost of structural fidelity. While these efforts have advanced the field, none systematically embeds scientific priors into the training-time generative dynamics of diffusion models. Consequently, constraint compliance remains an empirical outcome rather than a built-in guarantee.

To this end, we propose SensDiff, a diffusion-based 3D molecule generator that improves the physicochemical plausibility of generated molecules. SensDiff encodes domain knowledge as explicit symbolic constraints, enabling the model to directly learn specified scientific principles rather than inferring them through data fitting. Concretely, we cast symbolic rules governing 3D molecule generation as differentiable energy functions and incorporate them into the training objective. These energy constraints are activated sequentially in a coarse-to-fine curriculum: steric non-overlap first, then bond consistency, and finally valence validity. Such ordering mirrors the iterative structural refinement inherent in diffusion-based 3D molecule generation. In addition, we design a stage-wise weighting strategy to gradually enforce constraints and stabilize training. Specifically, it first linearly ramps up the weight of each newly activated constraint and then scales it in proportion to the degree of constraint satisfaction. We further extend this framework to the sampling stage, enforcing symbolic constraints during the reverse process of a pretrained diffusion model to steer generation toward chemically valid and geometrically realistic molecules without retraining.

To verify the effectiveness of SensDiff, we perform a comprehensive suite of evaluations, including (1) standard 3D molecule generation benchmarks; (2) constrained multi-objective molecule generation (*e.g.,* optimizing quantum-chemical and drug-likeness properties, with up to seven properties satisfied simultaneously); (3) out-of-distribution generalization of the learned scientific rules; (4) extension of the framework to sampling and fine-tuning; (5) transfer of the training strategy to various diffusion backbones; and (6) analysis of failure cases from a physicochemical perspective to guide model improvements. Crucially, we conduct extensive ablations on constraint enforcement mechanisms, weighting strategies, and training- and sampling-time schedules (*e.g.,* constraint activation curricula and epoch schedules), empirically validating each design choice. Our main contributions include:

- We introduce SensDiff, a neuro-symbolic constrained diffusion model for 3D molecule generation that produces molecules with enhanced geometric and physicochemical validity.
- We design a training strategy that mirrors the coarse-to-fine nature of diffusion molecular generation, enabling the model to internalize domain knowledge in its generative dynamics.
- We evaluate SensDiff through broad experiments, showing improved generation quality and constraint compliance, while ablations and failure case analysis validate each design choice.

2 Preliminaries

This section overviews the preliminaries, with details in Appendix B and notations in Appendix A.

Molecular Diffusion Models and Symbolic State. We adopt GEOLDM [69] as the diffusion backbone. A molecule with coordinates x and atom features h is encoded into a latent state $z = [z^{(x)} \| z^{(h)}]$, where $z^{(x)}$ and $z^{(h)}$ are the coordinate and feature channels. The clean latent state is $z_0 = z_{xh} = [z_{x\mu} \| z_{h\mu}]$. Diffusion defines a corrupted latent $z_t = \alpha_t z_0 + \sigma_t \epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$, and the reverse process is parameterized by $\epsilon_\theta(z_t, t)$. Let z_s denote the previous reverse-time latent state corresponding to z_t. From z_s, a soft decoder produces decoded coordinates $R = \{r_i\}_{i=1}^N$ and soft atom-type distributions A_i, while a bond posterior head predicts $B_{ij}^{(m)}$, where m denotes the bond type. The symbolic state (R, A, B) is the interface for evaluating and enforcing scientific constraints.

Symbolic Constraints. SensDiff maps physicochemical constraints over (R, A, B) to differentiable energies via a soft truth value $\kappa \in (0, 1)$ that quantifies constraint satisfaction, with energy $E = -(1/Z) \sum \omega \log(\kappa + \varepsilon)$, where ω is a soft antecedent (*i.e.*, precondition satisfaction)

weight, Z is a normalization factor, and ε is a numerical stability constant. This yields four energies, E_{clash}, $E_{\text{bond_len}}$, E_{angle}, and E_{val}, corresponding to steric non-overlap, bond-length consistency, bond-angle consistency, and valence validity. During training, they are combined as $\mathcal{L}_{\text{symbolic}}(e) = \lambda_c(e) \cdot E_{\text{clash}} + \lambda_b(e) \cdot E_{\text{bond_len}} + \lambda_a(e) \cdot E_{\text{angle}} + \lambda_v(e) \cdot E_{\text{val}}$, where e is the epoch and $\lambda_c(e),\ \lambda_b(e),\ \lambda_a(e),$ and $\lambda_v(e)$ are epoch-dependent constraint weights in the diffusion model's sequential training curriculum, designed to stabilize training when new constraints are introduced.

3 Method

Neuro-symbolic learning combines the expressiveness of neural networks with the interpretability of symbolic reasoning [26, 19, 27]. In this work, symbolic rules encode domain knowledge as differentiable energy constraints whose gradients provide supervisory signals to the denoising network, penalizing chemically invalid intermediate structures during training. We further propose a coarse-to-fine training strategy that sequentially activates the constraints with adaptive weighting. This aligns with the model's learning progression over training epochs, shaping the generative dynamics with physicochemical principles. As a result, these scientific rules are embedded in the model parameters and can generalize beyond finite training samples, rather than being implicitly memorized from data.

3.1 Differentiable Molecular Interface

A key challenge for SensDiff is that physicochemical rules are defined in 3D molecular space, whereas our diffusion evolves in a continuous latent space. To bridge the gap, we construct a differentiable interface that maps intermediate denoising states to 3D molecules. Let $z = [z^{(x)} \| z^{(h)}] \in \mathbb{R}^{B\times N\times D}$ denote the latent state, where $z^{(x)}$ and $z^{(h)}$ represent the coordinate and feature channels, and B, N, and D denote the batch size, padded number of atoms, and latent dimensionality, respectively. Given a reverse-time state z_s, we decode it into atomic coordinates and soft atom identities

$$(R, A) = \texttt{decode_soft}(z_s), \qquad A_i = m_i \cdot \texttt{softmax}(h_i^{\text{logit}}),$$

where $R = \{r_i\}_{i=1}^N$ denotes the decoded coordinates, A_i is the atom-type distribution for atom i, and $m_i \in \{0, 1\}$ is the node mask. For valid atoms with $m_i = 1$, we have $A_i \in \Delta^{C-1}$, where C is the number of atom types. For padded atoms with $m_i = 0$, the mask forces A_i to the zero vector. This soft decoding path is crucial: if atom identities are fixed via a hard $\arg\max$, the symbolic component becomes constant and yields zero gradients, so symbolic constraints cannot guide the denoising process. Using simplex-valued atom-type probabilities instead preserves uncertainty in atomic identity and keeps the latent-to-symbolic mapping entirely differentiable. Nevertheless, coordinates and soft atom identities alone cannot express bond or valence constraints. We therefore introduce a bond posterior head that infers a soft molecular graph from the decoded state, producing bond posteriors $B_{ij}^{(m)}$. For each atomic pair (i, j), we form the pair feature $\phi_{ij} = \left[|r_i - r_j|_2,\ A_i,\ A_j\right]$, map it to bond logits ℓ_{ij}, symmetrize via $\tilde{\ell}_{ij} = \frac{1}{2}(\ell_{ij} + \ell_{ji})$, and define the bond posteriors as

$$B_{ij}^{(m)} = \frac{\exp(\tilde{\ell}_{ij}^{(m)})}{\sum_{m'=0}^{M-1} \exp(\tilde{\ell}_{ij}^{(m')})},$$

where $m \in \{0, \ldots, M-1\}$ indexes bond classes, and M is the number of bond classes. This yields a probabilistic and differentiable relational representation of molecular bonds. Based on this, the symbolic state (R, A, B) produces the quantities needed to enforce steric, bonding, and valence constraints on intermediate denoising states. To link the symbolic state to empirical chemistry, we precompute statistics on the training data and interpolate them under soft atom identities. The non-bonded distance lower bound for atom types a and b is given by

$$d_{\min,ij} = \sum_{a=1}^{C}\sum_{b=1}^{C} A_{ia} A_{jb} d_{\min}^{ab}. \tag{1}$$

The bond-length statistics (mean μ and standard deviation σ) are computed as

$$\mu_{ijm}^{\text{bond}} = \sum_{a,b} A_{ia} A_{jb} \mu_{abm}^{\text{bond}}, \qquad \sigma_{ijm}^{\text{bond}} = \max\left(\sigma_{\min}^{\text{bond}}, \sum_{a,b} A_{ia} A_{jb} \sigma_{abm}^{\text{bond}}\right), \tag{2}$$

and the center-conditioned bond-angle statistics in cosine space are computed as

$$\mu^{\cos}_{j,mn} = \sum_{b=1}^{C} A_{jb}\mu^{\cos,\mathrm{fb}}_{b,mn}, \qquad \sigma^{\cos}_{j,mn} = \max\left(\sigma^{\text{angle}}_{\min}, \sum_{b=1}^{C} A_{jb}\sigma^{\cos,\mathrm{fb}}_{b,mn}\right), \tag{3}$$

where $\sigma^{\text{angle}}_{\min} = \sigma^{\text{bond}}_{\min} = 0.03$ are lower bounds on the interpolated angle and bond-length standard deviations, used to stabilize standardized residuals. fb denotes a fallback version of the angle cosine statistics, indexed only by the central atom type and adjacent bond classes, to alleviate sparsity in the fully triplet-conditioned table. Accordingly, these interpolants turn discrete chemistry statistics into differentiable functions of z_s, allowing symbolic constraints to act directly on latent denoising states.

3.2 Physicochemical Rules as Symbolic Energy Constraints

The second step is to cast physicochemical rules as differentiable energies. The intuition here is to assign low energy to molecules that satisfy the rule and high energy to those that violate it. Let $\kappa \in (0, 1)$ denote a soft truth value measuring the degree to which a rule is satisfied. We map it to

$$E = -\frac{1}{Z}\sum \omega \log(\kappa + \varepsilon), \tag{4}$$

where ω is a soft antecedent (*i.e.,* precondition satisfaction) weight, Z is a normalization factor, and $\varepsilon = 10^{-8}$ ensures numerical stability. This assigns lower energy to strongly satisfied rules and larger penalties for more severe violations. We use the Pseudo-Huber penalty $\rho(u) = \sqrt{1+u^2} - 1$, which is quadratic near zero and linear in the tails [25], to avoid unstable gradients when noisy intermediate states deviate from chemically valid structures. We conduct extensive experiments (see Appendix I) and analyze failure cases from a physicochemical perspective (see Appendix M) in our preliminary study to determine which constraints to impose during training and in what order to apply them.

Steric Non-Overlap. The first rule encodes geometric chemistry, requiring non-bonded atom pairs to avoid spatial overlap. This is a physically fundamental rule, so it becomes reliable earliest during denoising. For pair (i, j), let $d_{ij} = \|r_i - r_j\|_2$. We define the steric soft truth as

$$\kappa^{\text{clash}}_{ij} = \sigma\left(\frac{d_{ij} - d_{\min,ij}}{\tau_c}\right), \qquad \tau_c = 0.05,$$

where $\sigma(\cdot)$ is the sigmoid function. This quantity is close to 1 when the distance exceeds the interpolated lower bound and approaches 0 as the atoms clash. Since this rule only applies to pairs believed to be non-bonded, its antecedent is the no-bond posterior $B^{(0)}_{ij}$, yielding energy as

$$E_{\text{clash}} = -\frac{1}{Z_c}\sum_{i<j} B^{(0)}_{ij} \log(\kappa^{\text{clash}}_{ij} + \varepsilon), \qquad Z_c = \sum_{i<j} B^{(0)}_{ij} + \varepsilon. \tag{5}$$

Bond Length. Bond length captures 1-hop geometry and enforces consistency among distance, atom type, and bond order. For each nonzero bond class $m \neq 0$, we define the standardized deviation

$$z^{\text{bond}}_{ijm} = \frac{d_{ij} - \mu^{\text{bond}}_{ijm}}{\sigma^{\text{bond}}_{ijm}}, \qquad \kappa^{\text{bond}}_{ijm} = \exp\big(-\rho(z^{\text{bond}}_{ijm})\big),$$

and the corresponding energy for bond length can be expressed as

$$E_{\text{bond_len}} = -\frac{1}{Z_b}\sum_{i<j}\sum_{m\neq 0} B^{(m)}_{ij} \log(\kappa^{\text{bond}}_{ijm} + \varepsilon), \qquad Z_b = \sum_{i<j}\sum_{m\neq 0} B^{(m)}_{ij} + \varepsilon. \tag{6}$$

Bond Angle. Bond angle captures 2-hop local coordination geometry and therefore becomes meaningful later than bond length. Directly working with the angle itself would require $\arccos$, whose derivative becomes unstable near 0° and 180°. We instead work in cosine space

$$c_{ijk} = \frac{\langle r_i - r_j,\ r_k - r_j\rangle}{\|r_i - r_j\|_2\,\|r_k - r_j\|_2 + \varepsilon},$$

where j denotes the center atom, and i and k are candidate neighbors of j. Since enumerating all atom triplets scales as $\mathcal{O}(N^3)$, we restrict evaluation to the top-K_{nbr} likely neighbors of each center

atom, selected according to $p_{ji}^{\text{bond}} = \sum_{m\neq 0} B_{ji}^{(m)}$. Let $\mathcal{N}_{K_{\text{nbr}}}(j)$ be the corresponding top-K_{nbr} set, with $K_{\text{nbr}} = 4$. For $i, k \in \mathcal{N}_{K_{\text{nbr}}}(j)$ and $i < k$, and for nonzero bond classes m and n, we have

$$z_{ijkmn}^{\text{angle}} = \frac{c_{ijk} - \mu_{j,mn}^{\cos}}{\sigma_{j,mn}^{\cos}}, \qquad \kappa_{ijkmn}^{\text{angle}} = \exp\big(-\rho(z_{ijkmn}^{\text{angle}})\big),$$

and the corresponding energy for bond angle can be expressed as

$$E_{\text{angle}} = \frac{\sum_j \sum_{\substack{i<k \\ i,k\in\mathcal{N}_{K_{\text{nbr}}}(j)}} \sum_{m\neq 0}\sum_{n\neq 0} B_{ji}^{(m)} B_{jk}^{(n)} \left[-\log(\kappa_{ijkmn}^{\text{angle}} + \varepsilon)\right]}{\sum_j \sum_{\substack{i<k \\ i,k\in\mathcal{N}_{K_{\text{nbr}}}(j)}} \sum_{m\neq 0}\sum_{n\neq 0} B_{ji}^{(m)} B_{jk}^{(n)} + \varepsilon}. \tag{7}$$

The product $B_{ji}^{(m)} B_{jk}^{(n)}$ serves as a two-edge antecedent, so the angle rule is activated only when both incident (neighboring) bonds are plausible. Together, $E_{\text{bond_len}}$ and E_{angle} enforce bond consistency.

Valence Validity. The final rule enforces valence validity. Let ν_m denote the bond-order weight of bond class m, with $\nu_0 = 0$, $\nu_1 = 1$, $\nu_2 = 2$, $\nu_3 = 3$, and $\nu_{\text{aroma}} = 1.5$. The soft valence of atom i is

$$v_i^{\text{soft}} = \sum_{j\neq i} \sum_{m=0}^{M-1} \nu_m B_{ij}^{(m)}.$$

This quantity is the expected total incident bond order under the soft bond posterior. To determine whether it is chemically valid, we compare it against the admissible valence set $\mathcal{V}(a)$ for each atom type a, weighted by the empirical prior $\pi_{a,v}$, and define the valence soft truth as

$$\kappa_i^{\text{val}} = \sum_{a=1}^{C} A_{ia} \sum_{v\in\mathcal{V}(a)} \pi_{a,v} \exp\left(-\frac{(v_i^{\text{soft}} - v)^2}{2\sigma_v^2}\right), \qquad \sigma_v = 0.15.$$

Notably, when an atom type permits multiple valid valences, κ_i^{val} remains high near any admissible value. The corresponding energy for valence validity can be expressed as

$$E_{\text{val}} = -\frac{1}{N_{\text{valid}}} \sum_i m_i \log(\kappa_i^{\text{val}} + \varepsilon), \qquad N_{\text{valid}} = \sum_i m_i. \tag{8}$$

3.3 Coarse-to-Fine Curriculum with Adaptive Loss Weighting

SensDiff aligns symbolic guidance with the evolution of diffusion-based molecule generation. As structure emerges, bond length becomes meaningful once local connectivity is plausible, and bond angle follows as a 2-hop signal that requires two credible incident bonds. Valence is most downstream, since it aggregates bond order over a neighborhood and amplifies uncertainty in the bond posterior.

Our training pipeline consists of one preparation stage followed by four curriculum stages. Stage 0 precomputes dataset chemistry statistics, trains the variational autoencoder, and pretrains the bond-posterior head. It then freezes both so that subsequent symbolic gradients update only the diffusion model. Stage 1 activates E_{clash}, Stage 2 adds $E_{\text{bond_len}}$, Stage 3 adds E_{angle}, and Stage 4 adds E_{val}. This schedule is not merely heuristic, as it matches when intermediate denoising states become reliable enough for each constraint to provide meaningful supervision. To mitigate optimization instability when adding a new constraint, we assign each constraint an adaptive weight $\lambda_k(e)$ at epoch e. Before activation, $\lambda_k(e) = 0$. After activation at epoch e_k^{start}, the weight is linearly warmed up over R_{ramp} epochs and then further increased based on the current level of rule satisfaction. Formally,

$$\lambda_k(e) = \begin{cases} 0, & e < e_k^{\text{start}}, \\ \dfrac{\lambda_k^{\text{final}}}{2} \min\left(1, \dfrac{e - e_k^{\text{start}}}{R_{\text{ramp}}}\right), & 0 \le e - e_k^{\text{start}} < R_{\text{ramp}}, \\ \dfrac{\lambda_k^{\text{final}}}{2}\Big(1 + g(\bar{E}_k)\Big), & e - e_k^{\text{start}} \ge R_{\text{ramp}}, \end{cases} \tag{9}$$

where e_k^{start} is the activation epoch of constraint k, R_{ramp} is the warmup length, λ_k^{final} is the final weight, and $\bar{E}_k$ is the rolling mean of E_k for smoothing short-term fluctuations. $\bar{E}_k^{(0)}$ is the reference

Table 1: Unconditional 3D molecule generation on QM9 [58] and GEOM-Drugs [2]. **Best** results are marked in bold. We provide a complete comparison of chemical validity and geometric plausibility on QM9 and GEOM-Drugs in Table 8 and Table 9, respectively. All metrics are percentages (%).

Model	QM9				GEOM-Drugs	
	Valid	Atom Sta	Mol Sta	Valid & Unique	Valid	Atom Sta
EDM [30]	87.9	95.5	78.4	96.0	88.8	77.5
EEGSDE [5]	89.8	95.7	80.3	94.3	89.1	79.1
GEOLDM [69]	89.8	94.2	86.2	95.5	95.6	80.7
EquiFM [62]	91.6	94.4	84.3	94.7	95.7	80.3
GeoBFN [63]	90.7	94.9	86.9	93.3	88.5	81.5
GCDM [50]	91.4	94.2	82.5	95.7	91.2	84.5
RADM [14]	90.0	95.9	84.7	92.2	95.7	82.1
SLDM [52]	92.1	95.0	90.9	89.1	95.8	84.8
UAE-3D [49]	94.4	96.2	95.4	94.4	84.4	82.5
UniGEM [20]	92.0	95.2	87.1	93.5	94.7	82.1
SensDiff (**Ours**)	**96.9**	**98.1**	**95.6**	**97.2**	**97.2**	**87.9**

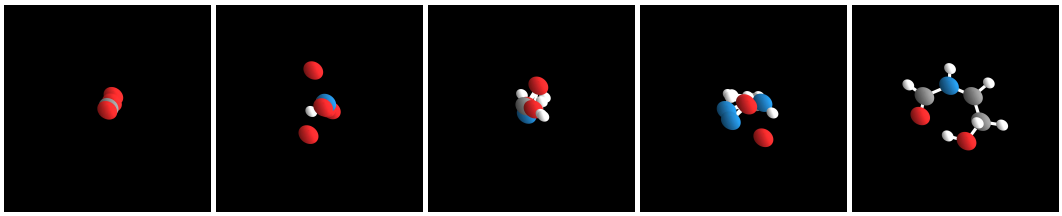


Figure 1: The 3D molecule generation trajectory over intermediate diffusion denoising steps. We also visualize trajectories from multiple baselines for the same molecule in Figures 9, 10, and 11.

energy at the start of the adaptive phase, and τ is the threshold for releasing the full weight. Besides,

$$g(\bar{E}_k) = \min\left(1,\ \frac{\max(0, \bar{E}_k^{(0)} - \bar{E}_k)}{\bar{E}_k^{(0)} - \tau}\right)$$

is the energy-adaptive factor. The linear warmup lets the model adapt to the gradients of a newly introduced rule, while the adaptive phase increases the weight only once the rule becomes reliable at the current noise level. Thus, the curriculum synchronizes symbolic supervision with the stability of intermediate denoising states, rather than applying a fixed sequence of auxiliary losses.

3.4 Diffusion Model Training Objective

SensDiff incorporates the four symbolic constraints into the training objective. Let $\mathcal{L}_{\text{diffusion}}$ denote the native training objective of the diffusion backbone. The symbolic loss is

$$\mathcal{L}_{\text{symbolic}}(e) = \lambda_c(e) \cdot E_{\text{clash}} + \lambda_b(e) \cdot E_{\text{bond_len}} + \lambda_a(e) \cdot E_{\text{angle}} + \lambda_v(e) \cdot E_{\text{val}}, \tag{10}$$

and the final training objective combines data fitting with the enforcement of physicochemical rules, ensuring that the model captures complex molecular patterns from data while simultaneously adhering to prescribed scientific principles, and is defined as

$$\mathcal{L}_{\text{total}}(e) = \mathcal{L}_{\text{diffusion}}(e) + \mathcal{L}_{\text{symbolic}}(e). \tag{11}$$

To make the constraints act on intermediate denoising states rather than only on the final decoded molecule, we evaluate $\mathcal{L}_{\text{symbolic}}(e)$ on a one-step reverse surrogate z_s. In particular, starting from the clean mean code z_{xh}, we sample a random diffusion step $t_{\text{int}} \sim \mathcal{U}\{1, \ldots, T\}$, construct the noisy latent $z_t = \alpha_t z_{xh} + \sigma_t \epsilon$ with $\epsilon \sim \mathcal{N}(0, I)$, and then compute one reverse step

$$z_s \sim \mathcal{N}\left(\frac{z_t}{\alpha_{t|s}} - \frac{\sigma_{t|s}^2}{\alpha_{t|s}\sigma_t}\epsilon_\theta(z_t, t),\ \left(\frac{\sigma_{t|s}\sigma_s}{\sigma_t}\right)^2 I\right). \tag{12}$$

Table 2: Mean absolute error (MAE, ↓) for quantum property conditional generation on QM9 [58].

MODEL	C_v $\left(\frac{\text{cal}}{\text{mol}}\,\text{K}\right)$	μ (D)	α (Bohr^3)	$\Delta\varepsilon$ (meV)	$\varepsilon_{\text{HOMO}}$ (meV)	$\varepsilon_{\text{LUMO}}$ (meV)
U-Bound	6.879	1.613	8.98	1464	645	1457
L-Bound	0.040	0.043	0.09	65	39	36
EDM [30]	1.154	1.164	2.87	676	368	603
EEGSDE [5]	0.986	0.812	2.61	506	314	462
GEOLDM [69]	1.072	1.156	2.46	608	351	541
EquiFM [62]	1.081	1.148	2.52	614	349	548
GeoBFN [63]	0.992	1.041	2.43	596	339	532
GCDM [50]	0.723	0.879	2.06	624	357	495
RADM [14]	0.907	0.851	2.07	474	301	397
SLDM [52]	0.781	0.831	1.53	456	331	362
UAE-3D [49]	0.582	0.638	1.61	327	224	259
UniGEM [20]	0.754	0.801	1.84	441	295	349
SensDiff (Ours)	**0.569**	**0.618**	**1.49**	**314**	**219**	**246**

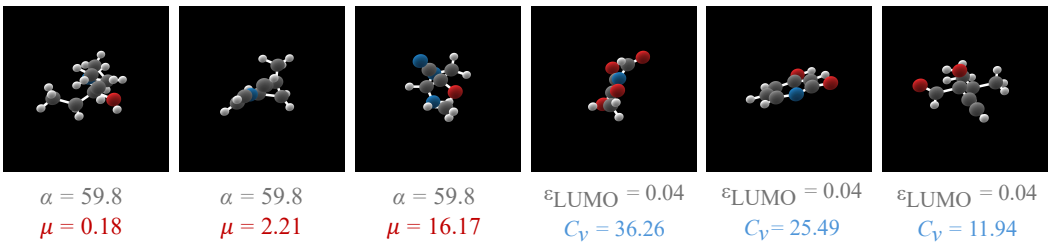


Figure 2: Conditional generation on QM9 [58] targeting two quantum-chemical properties simultaneously. Additional results are provided in Figure 14 and Figure 15.

Based on this, the symbolic state (R, A, B) is constructed from z_s, and the resulting symbolic loss is backpropagated through the reverse denoising step to the diffusion model. Because E_{clash}, $E_{\text{bond_len}}$, E_{angle}, and E_{val} are differentiable in (R, A, B), and (R, A, B) is differentiable in z_s, each rule yields a gradient $\nabla_{z_s} E_k$, which then propagates through the reverse kernel to the denoising network. As a result, the model is encouraged not merely to learn the observed molecular distribution, but to organize its generative dynamics around physicochemical rules.

In summary, SensDiff augments diffusion models with a sequential neuro-symbolic training strategy. It exposes chemically meaningful intermediate states, evaluates physicochemical rules on them, and feeds the resulting constraint signals back into the denoising process. Our design follows the coarse-to-fine behavior of diffusion molecular generation, where global geometry becomes reliable before local bonding and valence. Technically, SensDiff first establishes a differentiable interface from latent states to soft atoms and bonds. It then converts domain knowledge into differentiable energy constraints. These constraints are internalized through a sequential curriculum with epoch-dependent loss weighting. Each constraint is activated only when its required structural signal is stable. Consequently, the model learns from data while being guided by scientific priors, improving chemical validity and rule compliance. We provide implementation details for SensDiff in Appendix D.

4 Experiments

We evaluate SensDiff on two standard molecular benchmarks, QM9 [58] and GEOM-Drugs [2]. QM9 consists of small organic molecules with annotated quantum-chemical properties and is used for both unconditional and conditional generation, while GEOM-Drugs contains larger and more diverse drug-like molecules, providing a more challenging evaluation setting. We compare against a comprehensive set of recent baselines, including EDM [30], EEGSDE [5], GEOLDM [69], EquiFM [62], GeoBFN [63], GCDM [50], RADM [14], SLDM [52], UAE-3D [49], and UniGEM [20]. All methods are evaluated under a unified pipeline using metrics that cover chemical validity, sample diversity, geometric fidelity, constraint compliance, and conditional controllability. Each experiment is repeated ten times, and the mean results are reported. Further details on the datasets, baselines, and evaluation protocols are provided in Appendix C.

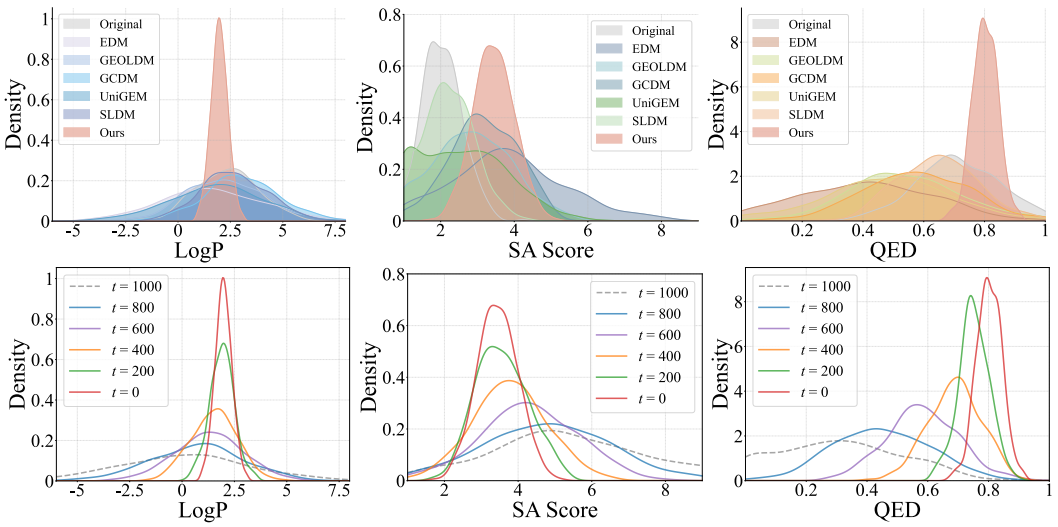


Figure 3: (**Upper**) The generated molecules precisely satisfy the specified oral drug-design constraints. (**Bottom**) The molecules converge to the target values over the denoising steps.

4.1 Unconditional 3D Molecule Generation

Following prior work [69, 52, 14, 20], Table 1 reports results on two benchmarks, with generation trajectories shown in Figure 1. SensDiff achieves the best generation quality across all metrics, with the largest gains in stability. It improves upon the strongest baseline by 0.2% in molecule stability on QM9 and 3.1% in atom stability on GEOM-Drugs. We attribute this to encoding physicochemical rules as symbolic constraints that directly supervise steric, bonding, and valence properties during training, yielding more chemically coherent structures. The coarse-to-fine curriculum further enforces basic geometric validity before enforcing finer chemical consistency, preventing higher-order constraints from being learned on flawed intermediates. More results are provided in Appendix F.

4.2 Conditional 3D Molecule Generation

Targeted Generation with Quantum Properties. QM9 annotates molecules with diverse quantum chemical properties, enabling conditional generation to test whether models can produce molecules that match specified targets. Table 2 reports MAE for property-conditioned generation, where lower is better. SensDiff achieves the lowest MAE across all properties, outperforming the strongest baseline UAE-3D [49]. Since these properties are largely determined by 3D geometry, our explicit geometric constraints provide a stronger foundation for conditional generation. Figure 2 further shows that SensDiff generates structurally diverse molecules that respond coherently to changing conditions, enabling fine-grained control over multiple properties. Additional results are provided in Appendix G.

Multi-Objective Optimization for Oral Drugs. Oral drug design remains challenging because oral bioavailability requires meeting multiple criteria simultaneously, and violating any single criterion can invalidate a candidate [23]. Existing methods often struggle to generate 3D molecules that satisfy all requirements at once because these rules are not explicitly learned. We thus apply SensDiff by freezing the GEOM-Drugs model and enforcing constraints only during sampling to preserve structural validity while steering generation toward target ranges. Specifically, we impose seven constraints aligned with Lipinski's Rule of Five and industry standards [51]: $1 < \text{LogP} < 3$, $\text{MW} \leq 500$, $\text{TPSA} \leq 140$ Å^2, $\text{HBD} \leq 5$, $\text{HBA} \leq 10$, $\text{QED} > 0.7$, and $\text{SA} < 5$. Figure 3 visualizes the constraint distributions and convergence toward the targets across denoising steps. More results on oral drug design and Lipinski's Rule-of-Five–constrained drug design [45, 65] are given in Appendix H.

4.3 Ablation Studies

We conduct a comprehensive set of ablation studies to validate our training strategy. Due to space limitations, we present only the key findings here, with detailed results and analysis in Appendix I.

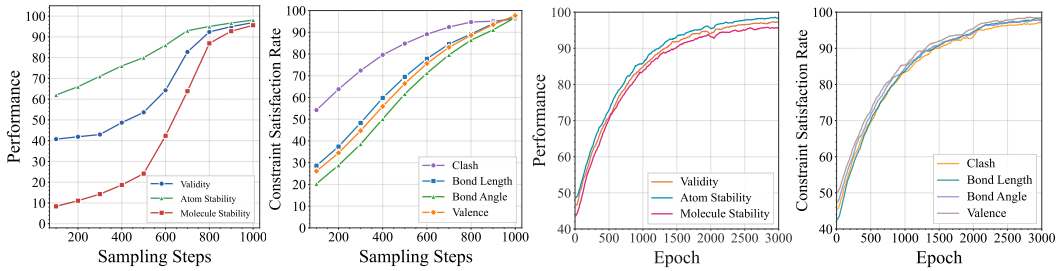


Figure 4: Ablations on sampling steps and training epochs on QM9. We save model checkpoints every 25 epochs (*i.e.,* 120 checkpoints in total). Results on GEOM-Drugs are shown in Figure 21.

Constraint Enforcement Order. As shown in Table 17, we study which enforcement order most improves generation quality by sequentially activating constraint subsets. The order matters substantially, with validity ranging from 74.6% to 90.9% across permutations. The best ordering (our design) applies global geometry early and local valence later, aligning with the denoising trajectory. In contrast, applying fine-grained local rules too early degrades performance because early geometries are too noisy to satisfy bond-level constraints, yielding misleading gradients that hinder refinement.

Non-Sequential Constraint Enforcement. As reported in Table 15 and Table 16, we study how sequential enforcement affects the molecule generation. When all constraints are enforced simultaneously, adding more constraints degrades generation quality on both datasets (*e.g.,* on QM9, molecule stability drops from 88.8% with one constraint to 78.8% with all four), suggesting that naively combining loss terms introduces competing gradients. This motivates our sequential activation strategy, which introduces rules progressively and lets the model consolidate each prior before adding the next.

Weighting Strategies. As shown in Table 18 and Table 20, we investigate the optimal constraint weights for diffusion training and sampling, respectively. Moreover, Table 23 presents how adaptive weighting affects the molecule generation. A hierarchical weighting that prioritizes global geometry over local bonds and valency is essential: flattening or inverting the weights collapses both generation quality and constraint satisfaction, indicating that a dominant structural stabilizer is needed before finer refinements can take effect. Moreover, linear ramping stabilizes training by avoiding abrupt gradients from newly introduced constraints, while energy-adaptive weighting further improves performance by adjusting weights based on constraint satisfaction rather than a fixed schedule.

Training and Sampling Schedules. Figure 4 and Figure 21 visualize how training epochs and sampling steps improve generation quality and constraint compliance. Table 19 and Table 22 compares constraints applied during training versus sampling only, showing that training-time constraints better improve molecule generation by more thoroughly internalizing scientific priors into the model via iterative learning. Table 21 shows that constraints are most effective when applied frequently, since sparse enforcement allows drift to accumulate and reduces both stability and constraint satisfaction.

4.4 Key Empirical Insights

Observation I. The neuro-symbolic method provides an interpretable lens that informs model design: by analyzing generated molecules, we can pinpoint which rules are violated and devise targeted remedies. We find that failures stem from a strict physicochemical hierarchy, where spatial non-overlap must precede meaningful bond geometry, and valid topology emerges only after local geometry stabilizes. Upstream violations thus cascade into downstream defects, motivating sequential training with adaptive weighting that activates each constraint only when its prerequisite becomes reliable. Notably, this analysis reveals that simply imposing more constraints does not necessarily improve generation, since prematurely activated rules can produce misleading or competing gradients. The key is therefore to enforce the right constraints at the stage when their underlying structural signals become sufficiently reliable. We provide a detailed failure case analysis from a domain-science perspective in Appendix M.

Observation II. SensDiff is trained to encode scientific principles into its generative dynamics, and these learned mechanistic regularities generalize to previously unseen data. This highlights the strong generalizability of our training strategy, which endows the model with science-grounded, rule-consistent reasoning beyond black-box neural denoising. In particular, SensDiff remains effective

under cross-dataset transfer and in low-density property regions, where purely data-driven models receive limited distributional support. This suggests that explicit physicochemical rules provide transferable structural inductive biases rather than merely fitting dataset-specific correlations. Details are provided in Appendix K.

Observation III. Our integration of scientific priors enables SensDiff to internalize domain knowledge, improving geometric plausibility, constraint compliance, and fine-grained control (*e.g.,* simultaneously satisfying even conflicting constraints). Importantly, these improvements are complementary rather than mutually exclusive: enforcing scientific rules enhances validity without sacrificing molecular diversity or conditional accuracy. Moreover, the resulting gains extend from directly constrained local structures, such as bond lengths and angles, to higher-order geometry and molecular-level properties. Details are provided in Appendices E, L, and H.

Observation IV. The proposed framework is broadly applicable: we empirically demonstrate that it can be leveraged for sampling-time constraint enforcement and fine-tuning. Moreover, our training strategy generalizes to various diffusion backbones. While sampling-time guidance offers a flexible way to impose new requirements without retraining, training-time integration yields stronger improvements by internalizing the constraints into the model parameters. The consistent gains across backbones further indicate that SensDiff acts as a model-agnostic training principle rather than an architecture-specific modification. Details are provided in Appendices I, N, and J.

5 Related Work

3D Molecule Generation with Diffusion Models. Score-based [61] and denoising diffusion models [28] are dominant methods for 3D molecule generation [35, 53, 47, 52, 12, 49]. They learn to reverse a Gaussian noising process to recover atomic coordinates and features [30, 69], and have been widely applied to molecule generation [68, 71, 20], conformer generation [70, 37, 43], optimization [60, 34, 77], and docking [13, 36, 48]. Existing approaches primarily improve generation through architectural inductive biases, molecular representations, or diffusion trajectory design. These advances effectively enhance geometric modeling and generation quality, but chemical feasibility is still learned largely through statistical correlations in the training data. However, purely data-driven training does not ensure adherence to scientific priors unless they are explicitly specified, so statistically plausible samples may still violate fundamental physicochemical rules. This limitation becomes more pronounced for complex molecular structures and generation beyond the well-represented training distribution. We thus formulate symbolic rules as differentiable energies and incorporate them into the training loss to shape the generative dynamics toward chemically valid structures.

Constrained Molecule Generation. Imposing task-specific constraints on molecular generators is vital for scientific discovery [57, 21, 10]. Large language models (LLMs) represent molecules as string sequences [67, 41] and satisfy user-specified objectives through instruction tuning, chemical feedback loops, or representational manipulation [15, 54, 18, 44]. Flow-based models learn transport maps and use equivariance or conditional vector fields to steer generation toward constraint-satisfying structures [63, 62, 76]. Diffusion models strike a balance between training stability and generation quality, making them particularly well suited for 3D molecule generation [6, 40]. Nevertheless, many constrained generation methods introduce corrective signals only during sampling, which limits the model's ability to internalize the underlying scientific principles. Moreover, applying multiple constraints simultaneously may introduce competing guidance, especially when intermediate molecular structures remain noisy and unreliable. Building on the strengths of diffusion models, we propose a symbolic-rule-constrained training strategy that sequentially integrates physicochemical constraints into the denoising process according to their structural dependencies. Our strategy is compatible with different molecular diffusion backbones and improves physicochemical feasibility while preserving generation quality and conditional controllability.

6 Conclusion

In this work, we propose SensDiff, a diffusion-based 3D molecule generator for physicochemically valid molecule generation. SensDiff internalizes scientific priors via sequential, adaptively weighted symbolic constraints aligned with diffusion's generative dynamics. Experiments show high generation quality and strong constraint compliance, while ablation studies validate our key design choices.

Acknowledgements and Disclosures

This work was supported in part by the Hong Kong Polytechnic University Internal Research Fund (P0057774, Young Innovative Researcher Scheme; P0063303, RIAIoT) and by the Research Grants Council of Hong Kong through the General Research Fund (Ref. No. 15208725). The authors received no other third-party funding or support related to this work during the past 36 months and declare that they have no competing interests.

References

[1] Zeyu An, Wanyu Lin, Feng Tan, and Shujun Wang. MMCP-GEN: A Modality-Extensible Diffusion Language Model for Conditional Protein Sequence Generation. In *Conference on Computer Vision and Pattern Recognition 2026*, 2026. 35

[2] Simon Axelrod and Rafael Gomez-Bombarelli. GEOM, Energy-Annotated Molecular Conformations for Property Prediction and Molecular Generation. *Scientific Data*, 9(1):185, 2022. 6, 7, 30, 36, 37, 42, 49, 52, 53, 54, 55, 56, 58, 59, 60, 61, 62, 63

[3] Richard FW Bader, Ronald J Gillespie, and Preston J MacDougall. A Physical Basis for the VSEPR Model of Molecular Geometry. *Journal of the American Chemical Society*, 110(22):7329–7336, 1988. 58

[4] Dávid Bajusz, Anita Rácz, and Károly Héberger. Why is Tanimoto Index An Appropriate Choice for Fingerprint-Based Similarity Calculations? *Journal of Cheminformatics*, 7(1):20, 2015. 31, 56

[5] Fan Bao, Min Zhao, Zhongkai Hao, Peiyao Li, Chongxuan Li, and Jun Zhu. Equivariant Energy-Guided SDE for Inverse Molecular Design. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 6, 7, 30, 37, 43, 44, 55, 56

[6] Tariq Berrada, Pietro Astolfi, Melissa Hall, Reyhane Askari-Hemmat, Yohann Benchetrit, Marton Havasi, Matthew Muckley, Karteek Alahari, Adriana Romero-Soriano, Jakob Verbeek, et al. On Improved Conditioning Mechanisms and Pre-Training Strategies for Diffusion Models. *Advances in Neural Information Processing Systems*, 37:13321–13348, 2024. 10

[7] G Richard Bickerton, Gaia V Paolini, Jérémy Besnard, Sorel Muresan, and Andrew L Hopkins. Quantifying the Chemical Beauty of Drugs. *Nature Chemistry*, 4(2):90–98, 2012. 30

[8] A van Bondi. Van Der Waals Volumes and Radii. *The Journal of Physical Chemistry*, 68(3):441–451, 1964. 58

[9] Darko Butina. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way to Cluster Small and Large Data Sets. *Journal of Chemical Information and Computer Sciences*, 39(4):747–750, 1999. 56

[10] Ziqi Chen, Bo Peng, Srinivasan Parthasarathy, and Xia Ning. Shape-Conditioned 3D Molecule Generation via Equivariant Diffusion Models. In *NeurIPS 2023 Generative AI and Biology (GenBio) Workshop*, 2023. 10

[11] Beatriz Cordero, Verónica Gómez, Ana E Platero-Prats, Marc Revés, Jorge Echeverría, Eduard Cremades, Flavia Barragán, and Santiago Alvarez. Covalent Radii Revisited. *Dalton Transactions*, pages 2832–2838, 2008. 58

[12] François Cornet, Grigory Bartosh, Mikkel N Schmidt, and Christian A Naesseth. Equivariant Neural Diffusion for Molecule Generation. *Advances in Neural Information Processing Systems*, 37:49429–49460, 2024. 2, 10

[13] Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi S. Jaakkola. DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking. In *The Eleventh International Conference on Learning Representations*, 2023. 10

[14] Yuhui Ding and Thomas Hofmann. Scalable Non-Equivariant 3D Molecule Generation via Rotational Alignment. In *Forty-Second International Conference on Machine Learning*, 2025. 6, 7, 8, 30, 37, 38, 39, 40, 43, 54, 55, 56

[15] Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. Translation Between Molecules and Natural Language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 375–413, 2022. 10

[16] Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-Weighted Linear Units for Neural Network Function Approximation in Reinforcement Learning. *Neural Networks*, 2018. 34

[17] Peter Ertl and Ansgar Schuffenhauer. Estimation of Synthetic Accessibility Score of Drug-Like Molecules Based on Molecular Complexity and Fragment Contributions. *Journal of Cheminformatics*, 1(1):8, 2009. 30

[18] Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. Mol-Instructions: A Large-Scale Biomolecular Instruction Dataset for Large Language Models. In *The Twelfth International Conference on Learning Representations*, 2024. 10

[19] Jonathan Feldstein, Paulius Dilkas, Vaishak Belle, and Efthymia Tsamoura. Mapping the Neuro-Symbolic AI Landscape by Architectures: A Handbook on Augmenting Deep Learning Through Symbolic Reasoning. *arXiv preprint arXiv:2410.22077*, 2024. 3

[20] Shikun Feng, Yuyan Ni, Lu Yan, Zhi-Ming Ma, Wei-Ying Ma, and Yanyan Lan. UniGEM: A Unified Approach to Generation and Property Prediction for Molecules. In *The Thirteenth International Conference on Learning Representations*, 2025. 1, 6, 7, 8, 10, 30, 37, 38, 39, 40, 43, 44, 47, 54, 55, 56

[21] Niklas WA Gebauer, Michael Gastegger, Stefaan SP Hessmann, Klaus-Robert Müller, and Kristof T Schütt. Inverse Design of 3D Molecular Structures with Conditional Generative Neural Networks. *Nature Communications*, 13(1):973, 2022. 10

[22] Dillon T Gentekos, Renee J Sifri, and Brett P Fors. Controlling Polymer Properties Through the Shape of the Molecular-Weight Distribution. *Nature Reviews Materials*, 4(12):761–774, 2019. 30

[23] Isabel Gomez-Orellana. Strategies to Improve Oral Drug Bioavailability. *Expert Opinion on Drug Delivery*, 2(3):419–433, 2005. 8

[24] Xu Han, Caihua Shan, Yifei Shen, Can Xu, Han Yang, Xiang Li, and Dongsheng Li. Training-free multi-objective diffusion model for 3d molecule generation. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[25] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge university press, 2003. 4, 32

[26] Pascal Hitzler, Aaron Eberhart, Monireh Ebrahimi, Md Kamruzzaman Sarker, and Lu Zhou. Neuro-Symbolic Approaches in Artificial Intelligence. *National Science Review*, 9(6):nwac035, 2022. 3

[27] Pascal Hitzler and Md Kamruzzaman Sarker. Neuro-Symbolic Artificial Intelligence: The State of the Art. 2022. 3

[28] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 10

[29] Haokai Hong, Wanyu Lin, Zhang Chusong, and Kay Chen Tan. Geometric Graph Neural Diffusion for Stable Molecular Dynamics Simulations. In *The Fourteenth International Conference on Learning Representations*, 2026. 35

[30] Emiel Hoogeboom, Vıctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant Diffusion for Molecule Generation in 3D. In *International Conference on Machine Learning*, pages 8867–8887. PMLR, 2022. 1, 2, 6, 7, 10, 30, 37, 43, 55, 56

[31] Chenqing Hua, Sitao Luan, Minkai Xu, Zhitao Ying, Jie Fu, Stefano Ermon, and Doina Precup. Mudiff: Unified Diffusion for Complete Molecule Generation. In *Learning on Graphs Conference*, pages 33–1. PMLR, 2024. 2

[32] Haowei Hua and Wanyu Lin. Local-Global Associative Frames for Symmetry-Preserving Crystal Structure Modeling. In *Advances in Neural Information Processing Systems*, 2025. 35

[33] Han Huang, Leilei Sun, Bowen Du, and Weifeng Lv. Learning Joint 2-D and 3-D Graph Diffusion Models for Complete Molecule Generation. *IEEE Transactions on Neural Networks and Learning Systems*, 35(9):11857–11871, 2024. 2

[34] Lei Huang, Tingyang Xu, Yang Yu, Peilin Zhao, Xingjian Chen, Jing Han, Zhi Xie, Hailong Li, Wenge Zhong, Ka-Chun Wong, et al. A Dual Diffusion Model Enables 3D Molecule Generation and Lead Optimization Based on Target Pockets. *Nature Communications*, 15(1):2657, 2024. 10

[35] Lei Huang, Hengtong Zhang, Tingyang Xu, and Ka-Chun Wong. MDM: Molecular Diffusion Model for 3D Molecule Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5105–5112, 2023. 10

[36] Yufei Huang, Odin Zhang, Lirong Wu, Cheng Tan, Haitao Lin, Zhangyang Gao, Siyuan Li, and Stan Z. Li. Re-Dock: Towards Flexible and Realistic Molecular Docking with Diffusion Bridge. In *Forty-First International Conference on Machine Learning*, 2024. 10

[37] Bowen Jing, Gabriele Corso, Jeffrey Chang, Regina Barzilay, and Tommi Jaakkola. Torsional Diffusion for Molecular Conformer Generation. *Advances in Neural Information Processing Systems*, 35:24240–24253, 2022. 10

[38] Hojung Jung, Youngrok Park, Laura Schmid, Jaehyeong Jo, Dongkyu Lee, Bongsang Kim, Se-Young Yun, and Jinwoo Shin. Conditional Synthesis of 3D Molecules with Time Correction Sampler. *Advances in Neural Information Processing Systems*, 37:75914–75941, 2024. 2

[39] Mortimer J Kamlet and RW Taft. The Solvatochromic Comparison Method. I. The. Beta.-Scale of Solvent Hydrogen-Bond Acceptor (HBA) Basicities. *Journal of the American chemical Society*, 98(2):377–383, 1976. 30

[40] Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and Improving the Training Dynamics of Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24174–24184, 2024. 10

[41] Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. Self-Referencing Embedded Strings (SELFIES): A 100% Robust Molecular String Representation. *Machine Learning: Science and Technology*, 1(4):045024, 2020. 10

[42] Tuan Le, Julian Cremer, Frank Noe, Djork-Arné Clevert, and Kristof T Schütt. Navigating the Design Space of Equivariant Diffusion-Based Generative Models for De Novo 3D Molecule Generation. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[43] Danyeong Lee, Dohoon Lee, Dongmin Bang, and Sun Kim. Disco: Diffusion Schrödinger Bridge for Molecular Conformer Optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 13365–13373, 2024. 10

[44] Zhuoran Li, Xu Sun, Chang Wen Chen, and Wanyu Lin. Controllable Molecule Generation via Sparse Representation Editing: An Interpretability-Driven Perspective. In *Forty-third International Conference on Machine Learning*, 2026. 10

[45] Christopher A Lipinski. Lead-and Drug-Like Compounds: the Rule-of-Five Revolution. *Drug Discovery Today: Technologies*, 1(4):337–341, 2004. 8, 47

[46] Christopher A Lipinski, Franco Lombardo, Beryl W Dominy, and Paul J Feeney. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Advanced Drug Delivery Reviews*, 64:4–17, 2012. 30

[47] Zhiyuan Liu, Yanchen Luo, Han Huang, Enzhi Zhang, Sihang Li, Junfeng Fang, Yaorui Shi, Xiang Wang, Kenji Kawaguchi, and Tat-Seng Chua. NExT-Mol: 3D Diffusion Meets 1D Language Modeling for 3D Molecule Generation. In *The Thirteenth International Conference on Learning Representations*, 2025. 1, 10

[48] Wei Lu, Jixian Zhang, Weifeng Huang, Ziqiao Zhang, Xiangyu Jia, Zhenyu Wang, Leilei Shi, Chengtao Li, Peter G Wolynes, and Shuangjia Zheng. DynamicBind: Predicting Ligand-Specific Protein-Ligand Complex Structure with A Deep Equivariant Generative Model. *Nature Communications*, 15(1):1071, 2024. 10

[49] Yanchen Luo, Zhiyuan Liu, Yi Zhao, Sihang Li, Hengxing Cai, Kenji Kawaguchi, Tat-Seng Chua, Yang Zhang, and Xiang Wang. Towards Unified and Lossless Latent Space for 3D Molecular Latent Diffusion Modeling. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems*, 2025. 1, 2, 6, 7, 8, 10, 30, 37, 38, 39, 40, 43, 54, 55, 56, 58, 59

[50] Alex Morehead and Jianlin Cheng. Geometry-Complete Diffusion for 3D Molecule Generation and Optimization. *Communications Chemistry*, 7(1):150, 2024. 1, 2, 6, 7, 30, 37, 43, 54, 55, 56

[51] Manuel A Navia and Pravin R Chaturvedi. Design Principles for Orally Bioavailable Drugs. *Drug Discovery Today*, 1(5):179–189, 1996. 8

[52] Yuyan Ni, Shikun Feng, Haohan Chi, Bowen Zheng, Huan-Ang Gao, Wei-Ying Ma, Zhi-Ming Ma, and Yanyan Lan. Straight-Line Diffusion Model for Efficient 3D Molecular Generation. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems*, 2025. 1, 6, 7, 8, 10, 30, 37, 38, 39, 40, 43, 47, 54, 55, 56, 58, 59

[53] Pedro O O Pinheiro, Joshua Rackers, Joseph Kleinhenz, Michael Maser, Omar Mahmood, Andrew Watkins, Stephen Ra, Vishnu Sresht, and Saeed Saremi. 3D Molecule Generation by Denoising Voxel Grids. *Advances in Neural Information Processing Systems*, 2023. 10

[54] Qizhi Pei, Wei Zhang, Jinhua Zhu, Kehan Wu, Kaiyuan Gao, Lijun Wu, Yingce Xia, and Rui Yan. BioT5: Enriching Cross-Modal Integration in Biology with Chemical Knowledge and Natural Language Associations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1102–1123, 2023. 10

[55] Xingang Peng, Jiaqi Guan, Qiang Liu, and Jianzhu Ma. MolDiff: Addressing the Atom-Bond Inconsistency Problem in 3D Molecule Diffusion Generation. In *International Conference on Machine Learning*, pages 27611–27629. PMLR, 2023. 2

[56] Sivaprakasam Prasanna and Robert J Doerksen. Topological Polar Surface Area: A Useful Descriptor in 2D-QSAR. *Current Medicinal Chemistry*, 16(1):21–41, 2009. 30

[57] Bo Qiang, Yiran Zhou, Yuheng Ding, Ningfeng Liu, Song Song, Liangren Zhang, Bo Huang, and Zhenming Liu. Bridging the Gap Between Chemical Reaction Pretraining and Conditional Molecule Generation with A Unified Model. *Nature Machine Intelligence*, 2023. 10

[58] Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole Von Lilienfeld. Quantum Chemistry Structures and Properties of 134 Kilo Molecules. *Scientific Data*, 1(1):1–7, 2014. 6, 7, 28, 30, 33, 35, 36, 37, 41, 43, 44, 45, 46, 49, 50, 51, 52, 53, 54, 55, 56, 60, 61, 62, 63

[59] Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) Equivariant Graph Neural Networks. In *International Conference on Machine Learning*. PMLR, 2021. 28

[60] Arne Schneuing, Charles Harris, Yuanqi Du, Kieran Didi, Arian Jamasb, Ilia Igashov, Weitao Du, Carla Gomes, Tom L Blundell, Pietro Lio, et al. Structure-Based Drug Design with Equivariant Diffusion Models. *Nature Computational Science*, 4(12):899–909, 2024. 10

[61] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*, 2021. 10

[62] Yuxuan Song, Jingjing Gong, Minkai Xu, Ziyao Cao, Yanyan Lan, Stefano Ermon, Hao Zhou, and Wei-Ying Ma. Equivariant Flow Matching with Hybrid Probability Transport for 3D Molecule Generation. *Advances in Neural Information Processing Systems*, 36:549–568, 2023. 6, 7, 10, 30, 37, 43, 55, 56

[63] Yuxuan Song, Jingjing Gong, Hao Zhou, Mingyue Zheng, Jingjing Liu, and Wei-Ying Ma. Unified Generative Modeling of 3D Molecules with Bayesian Flow Networks. In *The Twelfth International Conference on Learning Representations*, 2024. 6, 7, 10, 30, 37, 43, 55, 56

[64] Clement Vignac, Nagham Osman, Laura Toni, and Pascal Frossard. MiDi: Mixed Graph and 3D Denoising Diffusion for Molecule Generation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 560–576. Springer, 2023. 1, 2

[65] W Patrick Walters. Going Further than Lipinski's Rule in Drug Design. *Expert Opinion on Drug Discovery*, 7(2):99–107, 2012. 8, 47

[66] Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De Novo Design of Protein Structure and Function with RFdiffusion. *Nature*, 620(7976):1089–1100, 2023. 1

[67] David Weininger. 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. 10

[68] Lemeng Wu, Chengyue Gong, Xingchao Liu, Mao Ye, and Qiang Liu. Diffusion-Based Molecule Generation with Informative Prior Bridges. *Advances in Neural Information Processing Systems*, 35:36533–36545, 2022. 10

[69] Minkai Xu, Alexander S Powers, Ron O Dror, Stefano Ermon, and Jure Leskovec. Geometric Latent Diffusion Models for 3D Molecule Generation. In *International Conference on Machine Learning*, pages 38592–38610. PMLR, 2023. 1, 2, 6, 7, 8, 10, 28, 30, 37, 38, 39, 40, 43, 55, 56, 58, 59

[70] Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. GeoDiff: A Geometric Diffusion Model for Molecular Conformation Generation. In *International Conference on Learning Representations*, 2022. 10

[71] Yuning You, Ruida Zhou, Jiwoong Park, Haotian Xu, Chao Tian, Zhangyang Wang, and Yang Shen. Latent 3D Graph Diffusion. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 10

[72] Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Zilong Wang, Aliaksandra Shysheya, Jonathan Crabbé, Shoko Ueda, et al. A Generative Model for Inorganic Materials Design. *Nature*, 639(8055):624–632, 2025. 1

[73] An Zeyu and Wanyu Lin. Discrete Diffusion with Physical Mass Constraints for De Novo Peptide Sequencing. In *Forty-third International Conference on Machine Learning*, 2026. 35

[74] Odin Zhang, Tianyue Wang, Gaoqi Weng, Dejun Jiang, Ning Wang, Xiaorui Wang, Huifeng Zhao, Jialu Wu, Ercheng Wang, Guangyong Chen, et al. Learning on Topological Surface and Geometric Structure for 3D Molecular Generation. *Nature Computational Science*, 2023. 1

[75] Odin Zhang, Jintu Zhang, Jieyu Jin, Xujun Zhang, RenLing Hu, Chao Shen, Hanqun Cao, Hongyan Du, Yu Kang, Yafeng Deng, et al. ResGen Is A Pocket-Aware 3D Molecular Generation Model Based on Parallel Multiscale Modelling. *Nature Machine Intelligence*, 5(9):1020–1030, 2023. 1

[76] Wenyin Zhou, Christopher Iliffe Sprague, Vsevolod Viliuga, Matteo Tadiello, Arne Elofsson, and Hossein Azizpour. Energy-Based Flow Matching for Generating 3D Molecular Structure. In *Forty-Second International Conference on Machine Learning*, 2025. 10

[77] Xiangxin Zhou, Xiwei Cheng, Yuwei Yang, Yu Bao, Liang Wang, and Quanquan Gu. DecompOpt: Controllable and Decomposed Diffusion Models for Structure-based Molecular Optimization. In *The Twelfth International Conference on Learning Representations*, 2024. 10

[78] Wonho Zhung, Hyeongwoo Kim, and Woo Youn Kim. 3D Molecular Generative Framework for Interaction-Guided Drug Design. *Nature Communications*, 15(1):2688, 2024. 1

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Please refer to the contributions in Section 1.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Please refer to the limitation and future work in Section P.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: This work is primarily empirical, with the mathematical formulation presented in Section 3.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Detailed experimental setup and implementation details are provided in Sections C and D, and the official code is publicly available at `https://github.com/WanyuGroup/NeurIPS2026_SensDiff`.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification:We release the code, together with detailed usage instructions, at `https://github.com/WanyuGroup/NeurIPS2026_SensDiff`.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide detailed experimental setup and implementation details in Section C and D, respectively, which are necessary to understand the results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Instead, we run each experiment ten times and report the mean.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Please refer to Section D for compute resources.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have read and followed the Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Please refer to Section Q.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [No]

Justification: This work does not involve data or models that pose a high risk for misuse.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All assets used in this paper are properly credited, and their licenses and terms of use are explicitly stated and respected. Please refer to Section 4 and Section C.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We release the code for implementing the proposed method at `https://github.com/WanyuGroup/NeurIPS2026_SensDiff`.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The LLM is used only for grammar checking, formatting, and writing assistance.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.

Contents

1 Introduction **1**

2 Preliminaries **2**

3 Method **3**

3.1 Differentiable Molecular Interface 3

3.2 Physicochemical Rules as Symbolic Energy Constraints 4

3.3 Coarse-to-Fine Curriculum with Adaptive Loss Weighting 5

3.4 Diffusion Model Training Objective 6

4 Experiments **7**

4.1 Unconditional 3D Molecule Generation 8

4.2 Conditional 3D Molecule Generation 8

4.3 Ablation Studies 8

4.4 Key Empirical Insights 9

5 Related Work **10**

6 Conclusion **10**

A Notations **25**

B Extended Preliminaries **28**

C Experimental Setup **30**

D Implementation Details **31**

E Geometric Plausibility of SensDiff **35**

F Unconditional 3D Molecule Generation **37**

G Conditional 3D Molecule Generation with Quantum-Chemical Properties **43**

H Conditional 3D Molecule Generation for Multi-Objective Drug Design **47**

I Additional Ablation Studies **49**

J Generalizing Our Training Strategy to Multiple Diffusion Backbones **54**

K Out-of-Distribution Generalizability of SensDiff **54**

L Additional Evaluation Metrics for 3D Molecule Generation **56**

M Analysis of Failure Cases **57**

N Extending the Framework to Diffusion Model Fine-Tuning **60**

O Additional Studies on Training Schemes **61**

P Limitation and Future Work **62**

Q Broader Impacts **63**

A Notations

We organize the notation used in our method into four groups: Table 3 summarizes the notation for molecular representation and diffusion models, Table 4 summarizes the notation for the differentiable symbolic state and chemistry statistics, Table 5 summarizes the notation used in the four symbolic constraints, and Table 6 summarizes the notation used in the training curriculum and objective.

Table 3: Notation for molecular representation and diffusion models.

Notation	Meaning
B	Batch size.
N	Padded number of atoms per molecule.
C	Number of atom types.
M	Number of bond classes.
D	Latent dimensionality.
x	Molecular coordinates.
h	Molecular atom features.
h^{cat}	Categorical atom features.
h^{int}	Integer-valued atom features.
m_i	Node mask for atom slot i.
i, j, k	Atom indices; in angle expressions, j is the center atom and i, k are its neighbors.
z	Latent molecular state.
$z^{(x)}$	Coordinate channel of the latent state.
$z^{(h)}$	Feature channel of the latent state.
$z_{x\mu}, z_{x\sigma}$	Encoder mean and standard deviation for the coordinate latent code.
$z_{h\mu}, z_{h\sigma}$	Encoder mean and standard deviation for the feature latent code.
z_{xh}	Mean latent code used as the clean latent state for diffusion.
z_0	Clean latent state.
z_t	Corrupted latent state at diffusion step t.
z_s	Previous-step latent state in the reverse process.
z_T	Terminal Gaussian-noise latent state.
t	Normalized diffusion time.
s	Previous normalized diffusion time.
t_{int}	Integer diffusion step.
T	Total number of diffusion steps.
α_t	Signal coefficient in the forward diffusion process.
σ_t	Noise coefficient in the forward diffusion process.
$\alpha_{t\|s}$	Conditional signal coefficient for the reverse step from t to s.
$\sigma_{t\|s}$	Conditional noise coefficient for the reverse step from t to s.
ϵ	Gaussian noise.
θ	Parameters of the reverse diffusion model.
ϵ_θ	Noise prediction network.
$p_\theta(z_s \mid z_t)$	Reverse transition kernel.
μ	Mean of the Gaussian reverse kernel.
σ	Standard deviation of the Gaussian reverse kernel.

Table 4: Notation for the differentiable symbolic state and chemistry statistics.

Notation	Meaning
R	Decoded coordinates from the soft decoder.
r_i	Coordinate of atom i.
A_i	Soft atom-type distribution for atom i.
A_{ia}	Probability that atom i has atom type a.
h_i^{logit}	Pre-softmax atom-type logits for atom i.
Δ^{C-1}	Probability simplex for valid atom-type distributions.
$B_{ij}^{(m)}$	Bond posterior for atom pair (i, j) and bond class m.
m	Bond-class index.
n	A second bond-class index used in the angle rule.
v	Valence value used in the admissible valence set.
ϕ_{ij}	Pair feature used by the bond posterior head.
ℓ_{ij}	Bond logits for pair (i, j).
$\tilde{\ell}_{ij}$	Symmetrized bond logits for pair (i, j).
a, b	Atom-type indices used in chemistry statistics.
$d_{\min}^{ab}$	Non-bonded distance lower bound for atom types a and b.
$d_{\min,ij}$	Softly interpolated non-bonded distance lower bound for pair (i, j).
$\mu_{abm}^{\mathrm{bond}}$	Bond-length mean conditioned on atom types (a, b) and bond class m.
$\sigma_{abm}^{\mathrm{bond}}$	Bond-length standard deviation conditioned on atom types (a, b) and bond class m.
$\mu_{ijm}^{\mathrm{bond}}$	Interpolated bond-length mean for atom pair (i, j).
$\sigma_{ijm}^{\mathrm{bond}}$	Interpolated bond-length standard deviation for atom pair (i, j).
$\mu_{b,mn}^{\mathrm{cos,fb}}$	Fallback angle-cosine mean conditioned on center atom type b and bond classes (m, n).
$\sigma_{b,mn}^{\mathrm{cos,fb}}$	Fallback angle-cosine standard deviation conditioned on center atom type b and bond classes (m, n).
$\mu_{j,mn}^{\mathrm{cos}}$	Interpolated angle-cosine mean for center atom j.
$\sigma_{j,mn}^{\mathrm{cos}}$	Interpolated angle-cosine standard deviation for center atom j.
$\sigma_{\min}^{\mathrm{bond}}$	Lower bound for interpolated bond-length standard deviation.
$\sigma_{\min}^{\mathrm{angle}}$	Lower bound for interpolated angle standard deviation.
ν_m	Bond-order weight of bond class m.
$\mathcal{V}(a)$	Admissible valence set for atom type a.
$\pi_{a,v}$	Empirical prior for valence v of atom type a.
v_i^{soft}	Soft valence of atom i.

Table 5: Notation used in the four symbolic constraints.

Notation	Meaning
E	Generic differentiable energy.
κ	Soft truth value of a symbolic constraint.
ω	Soft antecedent weight.
Z	Normalization factor in the generic energy definition.
ε	Numerical-stability constant.
$\sigma(\cdot)$	Sigmoid function.
ρ	Pseudo-Huber penalty.
d_{ij}	Euclidean distance between atoms i and j.
$\|\cdot\|_2$	Euclidean norm.
$\langle\cdot,\cdot\rangle$	Inner product.
$\kappa_{ij}^{\text{clash}}$	Soft truth for the steric non-overlap rule.
τ_c	Temperature parameter for the clash rule.
E_{clash}	Steric-clash energy.
Z_c	Normalization factor for the clash energy.
z_{ijm}^{bond}	Standardized bond-length residual.
$\kappa_{ijm}^{\text{bond}}$	Soft truth for the bond-length rule.
$E_{\text{bond_len}}$	Bond-length energy.
Z_b	Normalization factor for the bond-length energy.
c_{ijk}	Cosine of the angle formed by atoms i-j-k.
p_{ji}^{bond}	Total nonzero bond probability between atoms j and i.
K_{nbr}	Number of likely neighbors retained for bond-angle evaluation.
$\mathcal{N}_{K_{\text{nbr}}}(j)$	Top-K_{nbr} likely neighbors of center atom j.
z_{ijkmn}^{angle}	Standardized angle residual.
$\kappa_{ijkmn}^{\text{angle}}$	Soft truth for the bond-angle rule.
E_{angle}	Bond-angle energy.
κ_i^{val}	Soft truth for the valence-validity rule.
σ_v	Bandwidth parameter in the valence kernel.
N_{valid}	Number of valid atoms in the molecule or batch item.
E_{val}	Valence energy.

Table 6: Notation for the training curriculum and objective.

Notation	Meaning
k	Index of a symbolic constraint in the curriculum.
e	Training epoch.
$\lambda_k(e)$	Epoch-dependent weight of constraint k.
e_k^{start}	Activation epoch of constraint k.
R_{ramp}	Warmup length for a newly activated constraint.
λ_k^{final}	Final weight of constraint k.
$\bar{E}_k$	Rolling mean of the energy of constraint k.
$\bar{E}_k^{(0)}$	Reference rolling energy at the start of the adaptive phase.
τ	Threshold used to release the full constraint weight.
$g(\bar{E}_k)$	Energy-adaptive factor used in the weighting schedule.
$\lambda_c(e)$	Weight of the clash energy.
$\lambda_b(e)$	Weight of the bond-length energy.
$\lambda_a(e)$	Weight of the bond-angle energy.
$\lambda_v(e)$	Weight of the valence energy.
$\mathcal{L}_{\text{diffusion}}$	Native diffusion loss of the backbone.
$\mathcal{L}_{\text{symbolic}}(e)$	Weighted symbolic loss at epoch e.
$\mathcal{L}_{\text{total}}(e)$	Final training loss at epoch e.
$\nabla_{z_s} E$	Gradient of an energy with respect to the intermediate latent state z_s.
$\nabla_{z_s} E_k$	Gradient of constraint energy E_k with respect to z_s.

B Extended Preliminaries

Overview. We consider a batch of molecules with coordinates $x \in \mathbb{R}^{B\times N\times 3}$, atom features $h = \{h^{\text{cat}}, h^{\text{int}}\}$ (*e.g.,* cat and int denote categorical and integer, respectively), and node masks $m_i \in \{0, 1\}$, where B is the batch size, N is the padded number of atoms, C is the number of atom types, M is the number of bond classes, and D is the latent dimensionality. Here, N denotes the maximum number of atoms after padding. For example, if molecules in a batch contain different numbers of atoms, they are padded to the same size N. The quantity C denotes the number of possible atom types. For example, in QM9 [58] we typically have $C = 5$, corresponding to H, C, N, O, and F. The quantity M denotes the number of bond classes. For example, one may use $M = 5$ to represent no bond, single, double, triple, and aromatic bonds.

The latent molecular state is $z = [z^{(x)} \| z^{(h)}] \in \mathbb{R}^{B\times N\times D}$, where $z^{(x)}$ and $z^{(h)}$ denote the coordinate and feature channels. At diffusion step $t = t_{\text{int}}/T$, where T is the total number of diffusion steps, we denote the corrupted latent by z_t and the previous-step latent by z_s, with $s = (t_{\text{int}} - 1)/T$. We use α_t and σ_t for the signal and noise coefficients, $\epsilon_\theta(z_t, t)$ for the noise predictor, and $z_{xh} = [z_{x\mu} \| z_{h\mu}]$ for the mean latent code produced by the encoder. Soft decoding yields coordinates $R = \{r_i\}_{i=1}^N$, atom-type distributions A_i, and bond posteriors $B_{ij}^{(m)}$, where $m = 0$ denotes the no-bond class. These variables define the symbolic state used for energy functions: $E_{\text{clash}}, E_{\text{bond_len}}, E_{\text{angle}}$, and E_{val}.

Diffusion Model and Molecular Representation. As discussed above, we use GEOLDM [69] as our molecular diffusion backbone. The molecular state (x, h) is encoded into a latent representation z, and the clean mean code $z_{xh} = [z_{x\mu} \| z_{h\mu}]$ is used as the latent state modeled by diffusion. The forward process is defined as

$$z_t = \alpha_t z_0 + \sigma_t \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I),$$

so z_t gradually interpolates between clean molecular structure and Gaussian noise. The corresponding reverse transition $p_\theta(z_s \mid z_t)$ is parameterized by $\epsilon_\theta(z_t, t)$. For the reverse step used here, the Gaussian kernel is expressed as

$$\mu = \frac{z_t}{\alpha_{t|s}} - \frac{\sigma_{t|s}^2}{\alpha_{t|s}\sigma_t}\epsilon_\theta(z_t, t), \qquad \sigma = \frac{\sigma_{t|s}\sigma_s}{\sigma_t},$$

which yields $z_s \sim \mathcal{N}(\mu, \sigma^2 I)$. This latent-space formulation is important because it exposes a continuous denoising state z_s on which symbolic constraints can be applied.

3D Molecular Variational Autoencoder. Since GEOLDM operates in latent space, we require a molecular variational autoencoder to encode molecular inputs into a latent 3D representation. A 3D molecular variational autoencoder provides a continuous latent representation for molecular geometry and atom features. Specifically, it aims to map the structured molecular input (x, h) into a latent state z that can be modeled more smoothly by diffusion models, and to reconstruct molecular variables from latent states during molecule generation. In our setting, the encoder outputs channel-wise latent statistics $(z_{x\mu}, z_{x\sigma}, z_{h\mu}, z_{h\sigma})$, and the mean code $z_{xh} = [z_{x\mu} \| z_{h\mu}]$ is used as the clean latent representation passed to the diffusion model. Because the underlying object is a 3D molecule, the encoder and decoder are designed to preserve geometric structure while jointly modeling node features [59]. The decoder thus provides the latent-to-molecule mapping needed for both 3D molecule generation and evaluating symbolic constraint satisfaction.

Latent Diffusion Model. Latent diffusion applies the diffusion process in the learned latent space z rather than directly on molecular state (x, h). This separation is useful because the variational autoencoder first compresses molecular geometry and node attributes into a continuous representation, after which diffusion only models the distribution of latent states. In practice, one starts from either an encoded clean latent z_0 during training or a noisy latent $z_T \sim \mathcal{N}(0, I)$ during sampling, applies the reverse dynamics $z_t \to z_s \to \cdots \to z_0$, and finally decodes the recovered latent into a molecule. The key to SensDiff is that latent diffusion exposes a continuous state z_s on which one can define gradients of symbolic energy constraints. Since R, A, and B are differentiable functions of z_s, a constraint energy $E(R, A, B)$ induces gradients $\nabla_{z_s} E$, making latent space the natural level at which neural denoising and symbolic reasoning can be effectively integrated.

Differentiable Molecular Interface. The method requires a differentiable map from latent states to chemically meaningful variables. A standard hard decoding path is unsuitable because discretization destroys the gradients needed for training-time evaluation of symbolic constraint satisfaction.

Therefore, the soft decoding path maps z_s to

$$(R, A) = \texttt{decode_soft}(z_s), \qquad A_i = \texttt{softmax}(h_i^{\text{logit}}) \cdot m_i,$$

so that, for valid nodes, $A_i \in \Delta^{C-1}$ remains a continuous probability vector rather than collapsing to a one-hot atom type. This is crucial because the symbolic branch must evaluate constraints on intermediate denoising states, where atomic identity is still uncertain.

The decoded geometry and soft atom identities are then lifted to a relational molecular representation through a bond posterior head. For each pair (i, j), the pair feature is

$$\phi_{ij} = \left[|r_i - r_j|_2 \,,\, A_i,\, A_j\right]$$

which is mapped to bond logits ℓ_{ij}, symmetrized as $\tilde{\ell}_{ij} = \frac{1}{2}(\ell_{ij} + \ell_{ji})$, and normalized by

$$B_{ij}^{(m)} = \frac{\exp(\tilde{\ell}_{ij}^{(m)})}{\sum_{m'=0}^{M-1} \exp(\tilde{\ell}_{ij}^{(m')})}.$$

The resulting $B_{ij}^{(m)}$ is symmetric in i and j. This molecular interface is essential because bond-length, bond-angle, and valence constraints are not defined on 3D coordinates alone. They require a probabilistic and differentiable relational representation derived from the latent state.

Neuro-Symbolic Constraints as Differentiable Energies. Our neuro-symbolic formulation evaluates symbolic physicochemical rules on the symbolic state (R, A, B) and maps each rule to a differentiable energy, which can be expressed as

$$E = -\frac{1}{Z} \sum \omega \log(\kappa + \varepsilon),$$

where $\kappa \in (0, 1)$ is a soft truth, ω is a soft antecedent weight, Z is a normalization factor, and $\varepsilon = 10^{-8}$. Because atom identities are represented by soft distributions, discrete chemistry statistics must be interpolated continuously. For example, the non-bonded distance lower bound is

$$d_{\min,ij} = \sum_{a=1}^{C} \sum_{b=1}^{C} A_{ia} A_{jb} d_{\min}^{ab}.$$

The bond-length statistics are

$$\mu_{ijm}^{\text{bond}} = \sum_{a,b} A_{ia} A_{jb} \mu_{abm}^{\text{bond}}, \qquad \sigma_{ijm}^{\text{bond}} = \max(\sigma_{\min}^{\text{bond}}, \sum_{a,b} A_{ia} A_{jb} \sigma_{abm}^{\text{bond}}),$$

and the angle statistics are

$$\mu_{j,mn}^{\cos} = \sum_{b=1}^{C} A_{jb} \mu_{b,mn}^{\cos,\text{fb}}, \qquad \sigma_{j,mn}^{\cos} = \max(\sigma_{\min}^{\text{angle}}, \sum_{b=1}^{C} A_{jb} \sigma_{b,mn}^{\cos,\text{fb}}),$$

where $\sigma_{\min}^{\text{bond}}$ and $\sigma_{\min}^{\text{angle}}$ are lower bounds used to stabilize the interpolated standard deviations. Likewise, the soft valence of atom i is defined as

$$v_i^{\text{soft}} = \sum_{j \neq i} \sum_{m=0}^{M-1} \nu_m B_{ij}^{(m)},$$

where ν_m denotes the bond order weight of bond class m. This quantity is the expected total incident bond order of atom i under the soft bond posterior $B_{ij}^{(m)}$, and therefore provides a differentiable relaxation of discrete chemical valence. It is then evaluated against the admissible valence set $\mathcal{V}(a)$ of atom type a, weighted by the corresponding empirical prior $\pi_{a,v}$, so that the valence constraint favors values of v_i^{soft} that are chemically valid for plausible atom identities. These quantities define the four energy constraints E_{clash}, $E_{\text{bond_len}}$, E_{angle}, and E_{val}, corresponding to steric non-overlap, bond-length consistency, bond-angle consistency, and valence validity, respectively.

C Experimental Setup

Datasets. We evaluate SensDiff on QM9 [58] and GEOM-Drugs [2]. QM9 contains 134k small organic molecules with annotated quantum-chemical properties and is a standard benchmark for assessing whether a model can generate chemically plausible 3D molecules while controlling molecular properties. In our setting, the atom vocabulary is $\{\mathrm{H}, \mathrm{C}, \mathrm{N}, \mathrm{O}, \mathrm{F}\}$, and each molecule has at most nine heavy atoms. We use QM9 for both unconditional and conditional molecule generation. GEOM-Drugs is a larger and more challenging benchmark of drug-like molecules with diverse 3D conformers. Compared with QM9, it features larger structures, richer topologies, and broader geometric variability, offering a more demanding test of scalability and robustness. We use GEOM-Drugs to evaluate whether the proposed constraints remain effective beyond the small-molecule regime. We split each dataset into 80%, 10%, and 10% for training, validation, and testing, respectively.

Baselines. We compare SensDiff against a broad set of recent 3D molecular generators chosen to cover the main modeling paradigms used in the literature. The first group consists of strong geometry-aware generative backbones. EDM [30] is the canonical equivariant diffusion baseline for 3D molecule generation and serves as the standard reference for direct diffusion in Euclidean space. GCDM [50] is another diffusion-based model that aims to model molecular geometry more completely, making it a useful comparator for testing whether better geometric realism can be achieved by geometry modeling alone, without explicit symbolic constraints. SLDM [52] emphasizes a simplified straight-line diffusion trajectory and represents methods that improve generation through trajectory design and efficiency rather than through domain constraints. RADM [14] provides a complementary comparison because it handles rotations through explicit alignment rather than built-in equivariance, allowing us to assess whether SensDiff remains competitive against strong models with a different inductive bias. The second group captures methods that inject extra information during generation. EEGSDE [5] is especially relevant because it uses energy-guided sampling. This makes it a natural comparator for SensDiff: both methods leverage scientific signals, but EEGSDE applies guidance primarily at inference time, whereas SensDiff integrates symbolic constraints into the training objective itself. The third group focuses on latent-space and representation-centric generation. GEOLDM [69] is the most direct backbone baseline because SensDiff is built on top of it. This comparison isolates the contribution of the proposed neuro-symbolic component from the underlying latent diffusion architecture. UAE-3D [49] is another strong latent-space generator that emphasizes a more unified and information-preserving 3D representation. Finally, we compare against alternative generative paradigms and property-aware models. EquiFM [62] represents equivariant flow matching, which learns a continuous probability transport process instead of a denoising diffusion chain. GeoBFN [63] represents Bayesian flow networks and provides a non-diffusion comparison point while targeting the same 3D molecule generation problem. UniGEM [20] jointly models molecule generation and property prediction, making it particularly relevant for conditional generation experiments. In summary, these baselines cover direct diffusion, guided diffusion, latent diffusion, flow matching, Bayesian flow networks, and unified generation-property modeling. All metrics are computed by us for all baselines under a unified evaluation pipeline, and the values may differ slightly from those reported in the original papers.

Evaluation Metrics. For generation quality, we report validity, validity & uniqueness, and novelty. Validity is the fraction of generated samples that are chemically valid under the standard benchmark pipeline. Validity & uniqueness is the fraction of valid molecules that are non-duplicated, preventing inflated performance from repeated valid samples. Novelty is the fraction of generated molecules absent from the training set. To assess 3D plausibility, we report atom and molecule stability, along with bond-length, bond-angle, and dihedral-angle statistics. Atom and molecule stability measure whether local environments and whole molecules pass standard stability checks in prior 3D benchmarks, while the internal-coordinate statistics quantify deviations from the empirical geometry distribution, where smaller deviation is better. For QM9 quantum properties, we evaluate six target properties, namely α, $\Delta\varepsilon$, $\varepsilon_{\mathrm{HOMO}}$, $\varepsilon_{\mathrm{LUMO}}$, μ, and C_v, with units Bohr^3, meV, meV, meV, D, and $\frac{\mathrm{cal}}{\mathrm{mol}}$ K, respectively. For all six properties, we report mean absolute error (MAE) between the requested target value and the achieved value of the generated molecule. MAE provides a direct measure of controllability: lower MAE means the model follows the desired scientific target more accurately. For constrained generation in drug-like molecules, we report the achieved values of LogP [46], Molecular Weight (MW) [22], Hydrogen-Bond Acceptors (HBA) [39], Hydrogen-Bond Donors (HBD) [39], Topological Polar Surface Area (TPSA) [56], Quantitative Estimate of Drug-likeness (QED) [7], and Synthetic Accessibility (SA) Score [17]. These descriptors cover hydrophobicity, molecular size,

hydrogen-bonding capacity, polarity, overall drug-likeness, and synthetic accessibility. We report the raw descriptor values because the goal is not to optimize every descriptor in the same direction, but to place generated molecules inside a user-specified target range of chemical property space. To measure whether all requested conditions are satisfied, we report the constraint satisfaction rate (CSR), defined as the fraction of generated molecules that satisfy the specified constraints. CSR is important in multi-objective settings because property-wise averages can be misleading. A molecule may be close to one target yet conflict with several others. When similarity preservation is part of the task, we report Tanimoto similarity, a standard fingerprint-based overlap score in $[0, 1]$ where larger values indicate greater structural similarity to the reference molecule or scaffold [4]. We repeat each experiment ten times and report the mean results.

D Implementation Details

Overview. Our implementation follows the pipeline described in the main text, and we provide additional details here. We use QM9 with the standard GEOLDM preprocessing pipeline, keep integer charges in the node representation, and retain the default normalization factors $[1, 4, 10]$. Molecules are represented by coordinates $x \in \mathbb{R}^{B\times N\times 3}$, atom features $h = \{h^{\text{cat}}, h^{\text{int}}\}$, and node masks m_i, with $N = 29$ after padding on QM9, $C = 5$ atom types, and $M = 5$ bond classes. The latent representation has $\text{latent_nf} = 1$, so the latent dimensionality is $D = 4$. The encoder, decoder, and diffusion denoiser all inherit the EGNN configuration with 9 layers, hidden size 256, attention, and tanh gating. Diffusion model uses $T = 1000$ steps, the polynomial-2 noise schedule, noise precision 10^{-5}, and the standard ℓ_2 denoising objective. The role of SensDiff is to expose a differentiable symbolic state around the reverse-time latent variable z_s, so that the same denoising trajectory can be supervised by the four physicochemical constraints defined in Section 3, rather than to replace the GEOLDM generative backbone itself.

Preparation of the symbolic branch. Before training the constrained diffusion model, we first make the symbolic branch fully operational. The first step is an offline pass over the QM9 training split, where RDKit is used to reconstruct bond graphs from the reference 3D geometries. From this pass, we accumulate the atom-pair non-bonded distance lower bounds $d^{ab}_{\min}$, the bond-length statistics $(\mu^{\text{bond}}_{abm}, \sigma^{\text{bond}}_{abm})$, the center-conditioned fallback angle-cosine statistics $(\mu^{\cos,\text{fb}}_{b,mn}, \sigma^{\cos,\text{fb}}_{b,mn})$, and the admissible valence sets $\mathcal{V}(a)$ together with their empirical priors $\pi_{a,v}$. These tables are not used as hard lookups during training. Instead, they are converted into differentiable quantities through the soft interpolation rules in Equation 1, Equation 2, and Equation 3. In practice, although the full triplet-conditioned angle table is precomputed, the implementation uniformly uses the center-conditioned fallback table in Equation 3 for the vectorized angle kernel, because many fully conditioned triplets are sparse on QM9 and the fallback version is both more stable and computationally easier. We then train the GEOLDM VAE for 300 epochs with batch size 64, learning rate 10^{-4}, KL weight 0.01, and EMA decay 0.9999, and we separately pretrain the bond posterior head with RDKit-derived bond orders as five-class supervision. Once Stage 0 is complete, both the VAE and the bond posterior head are frozen so that all subsequent symbolic gradients reshape only the diffusion dynamics.

Differentiable symbolic interface. The implementation uses the same soft molecular interface described in Section 3, but we make the engineering role of each component explicit here. First, the decoder exposes a dedicated `decode_soft` path, because a hard $\arg\max$ over atom-type logits would destroy the gradients needed to optimize the symbolic energies. Given a reverse-time latent z_s, `decode_soft` returns recentered coordinates R and simplex-valued atom identities A_i, where padding positions are suppressed by the node mask. Second, a `BondPosteriorHead` maps each pair feature $\phi_{ij} = \left[|r_i - r_j|_2, A_i, A_j\right]$ to logits over the five bond classes and then converts them into $B^{(m)}_{ij}$. This head uses two hidden layers of width 128 with SiLU activations, directly symmetrizes the logits so that $B^{(m)}_{ij} = B^{(m)}_{ji}$, and suppresses self-loops and padded pairs by adding large negative penalties before softmax, which ensures that the no-bond class dominates on the diagonal and

at masked positions. The resulting state (R, A, B) is the implementation-level realization of the symbolic state used in Section 3. It is also the critical bridge that makes the general energy definition in Equation 4 differentiable with respect to the latent variable.

Constraint evaluation. The four symbolic energies are implemented exactly according to the manuscript: steric non-overlap in Equation 5, bond-length consistency in Equation 6, bond-angle consistency in Equation 7, and valence validity in Equation 8. The implementation directly follows the soft-truth construction in Equation 4, so each rule is evaluated as a normalized negative log soft satisfaction score. For E_{clash}, the temperature is fixed to $\tau_c = 0.05$. For $E_{\text{bond_len}}$ and E_{angle}, standardized residuals are regularized with the Pseudo-Huber penalty $\rho(u) = \sqrt{1+u^2} - 1$ [25], which avoids the severe underflow and vanishing-gradient behavior that would arise from a pure Gaussian penalty on highly corrupted denoising states. For numerical stability, we impose lower bounds of 0.03 on the interpolated bond length and angle standard deviations, use $\varepsilon = 10^{-8}$ in all logarithmic and normalization terms, clamp cosine values to a safe interval before evaluating the angle term, and restrict the angle computation to the top-K_{nbr} neighbors with $K_{\text{nbr}} = 4$, where neighbors are ranked by total bonding probability. For valence, the implementation uses the bond-order weights $\nu_0 = 0$, $\nu_1 = 1$, $\nu_2 = 2$, $\nu_3 = 3$, and $\nu_{\text{aroma}} = 1.5$, and the RBF bandwidth is fixed to $\sigma_v = 0.15$. All energy computations remain fully vectorized and mask-aware, so padded atoms do not contribute either to the energy values or to their gradients.

Training objective and one-step reverse surrogate. The training objective follows the exact decomposition already stated in Equation 10 and Equation 11. In every iteration, we first load the batch with non-blocking GPU transfers, recenter the coordinates with `remove_mean_with_mask`, optionally apply Gaussian data augmentation, and compute the native GEOLDM objective $\mathcal{L}_{\text{diffusion}}$. We then run the frozen encoder to obtain the clean mean latent code $z_{xh} = [z_{x\mu} \| z_{h\mu}]$, immediately detach this tensor so that symbolic gradients do not flow back into the encoder, sample a random timestep $t_{\text{int}} \sim \mathcal{U}\{1, \ldots, T\}$, and corrupt the latent using the same forward diffusion process as the backbone. Rather than evaluating the symbolic branch only on the final decoded molecule, we use the one-step reverse surrogate in Equation 12. Concretely, we form z_t, compute one reverse transition $z_t \to z_s$ through the GEOLDM reverse kernel, decode z_s into (R, A), infer B, and then evaluate whichever of the four energies are active at the current training stage, while the inactive terms are set to zero by the stage scheduler. This choice is a key engineering compromise as it keeps the symbolic supervision attached to intermediate denoising states, exactly as required by the method, while avoiding the prohibitive cost of unrolling a full T-step reverse chain inside each training iteration. Once the energies are computed, we assemble $\mathcal{L}_{\text{symbolic}}(e)$ from Equation 10, add it to the diffusion objective as in Equation 11, and backpropagate through the one-step reverse kernel so that only the diffusion network is updated. No additional connectivity penalty, loss term, or extra KL term is introduced beyond the native GEOLDM objective and the four symbolic energies.

Curriculum and adaptive weighting. The activation schedule matches the coarse-to-fine nature of diffusion molecular generation. Stage 1 covers epochs 1–500 and activates only E_{clash}, Stage 2 covers epochs 501–1400 and adds $E_{\text{bond_len}}$, Stage 3 covers epochs 1401–2200 and adds E_{angle}, and Stage 4 covers epochs 2201–3000 and adds E_{val}. The final weights are $\lambda_c = 4.0$, $\lambda_b = 3.0$, $\lambda_a = 2.0$, and $\lambda_v = 1.0$. Their epoch-dependent values are controlled by Equation 9. Each newly activated term is first linearly ramped for 100 epochs until it reaches half of its target weight, after which the remaining half is released according to the rolling energy decrease. The implementation maintains a 20-epoch rolling mean for each constraint, records the ramp-end snapshot $\bar{E}_k^{(0)}$, uses the target threshold $\tau = 0.03$, and applies the same adaptive factor that appears in Equation 9. To prevent abrupt destabilization when a newly introduced rule yields unusually large gradients, the scheduler includes a safeguard: if the instantaneous energy exceeds twice its rolling mean, the constraint is frozen at half strength for 50 epochs before adaptive release resumes. This mechanism is not a separate modeling component. It implements the curriculum and is crucial for stabilizing the stage-wise introduction of the four constraints during training.

Optimization and validation protocol. Training uses learning rate 10^{-4}, EMA decay 0.9999, and adaptive queue-based gradient clipping rule inherited from the implementation notes. We enable TF32 matrix multiplication on GPU. Mixed precision is handled by using BF16 autocast for the forward and backward passes. Otherwise, the code falls back to FP16 with GradScaler. Gradients are unscaled before clipping, after which the optimizer step and EMA update are performed. A dedicated logger records the epoch, iteration, stage, individual symbolic energies, diffusion loss, symbolic loss, and total loss throughout training. Each epoch ends with a stability check, and every 20 epochs we

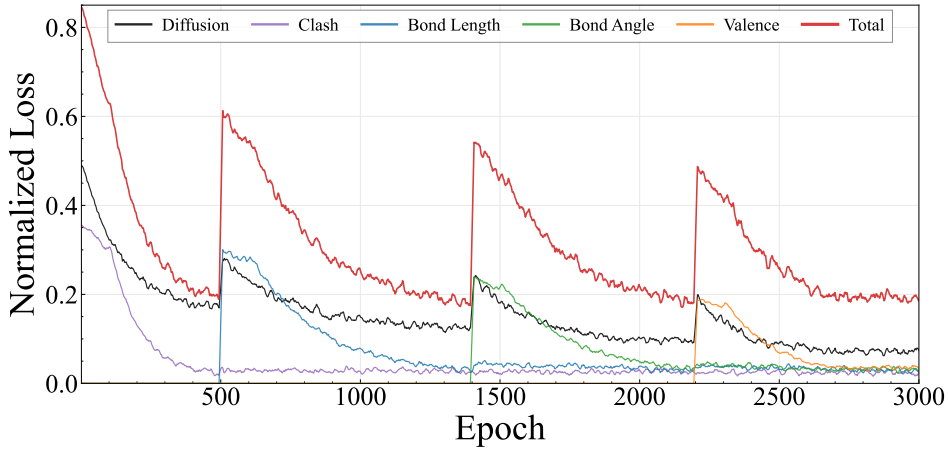


Figure 5: **Training loss on QM9** [58]. Symbolic energy constraints are activated sequentially in a coarse-to-fine curriculum. Each triggers a transient spike in total loss that is quickly absorbed as adaptive weighting ramps up the newly introduced constraint.

run validation, compute NLL-style statistics, generate 1000 molecules with EMA parameters, and select the best checkpoint. These details matter because the symbolic branch increases computational cost and gradient variance. Mixed precision, adaptive clipping, EMA, and stage-wise logging are therefore key to making the training recipe reproducible rather than merely well defined.

Inference-time latent refinement. Sampling keeps the standard GEOLDM reverse chain but inserts symbolic constraints after every reverse step. We first sample z_T with center-of-gravity-zero Gaussian noise for the coordinate channels and standard Gaussian noise for the feature channels, and then run the reverse kernel from $t = T$ to $t = 1$, using the same reverse transition as Equation 12. At inference, the stage is fixed to 4, so all four constraints are active regardless of which training stage was used to produce the checkpoint. After each reverse step, we perform $K = 3$ refinement steps on the current latent state. In each refinement step, a detached float32 copy of the latent is marked as requiring gradients, decoded to (R, A), converted to B, and evaluated with the full weighted symbolic energy corresponding to the final form of Equation 10. We then compute the gradient with respect to the latent variable only, replace non-finite values with zero if necessary, mask out padded positions, and normalize the gradient independently for each molecule in the batch. This per-sample normalization matters because otherwise a single molecule with large energy could dominate the batch update direction. We apply the normalized gradient with step size $\eta_t = \alpha_t / K$, where $\alpha_t = \alpha_{\max}(1 - t/T)^2$ and $\alpha_{\max} = 1.0$. This quadratic schedule makes refinement negligible at the noise-dominated start of the reverse chain and strongest near the end, when structure is sufficiently formed for constraints to be informative. After each step, we recenter the coordinate channels and remask the feature channels. If the symbolic energy is NaN or Inf, we skip refinement and keep the original reverse sample.

Hardware and efficiency. Figure 5 visualizes SensDiff's training loss on QM9. We adopt a sequential training strategy that progressively introduces constraints on top of the diffusion loss, as reflected by the periodic spikes in the total loss curve whenever a new term is activated. Despite these transient increases, all individual loss components converge rapidly and stabilize, demonstrating that our curriculum-based weight scheduling enables the model to effectively internalize each symbolic constraint without disrupting previously learned objectives. Training is run on an NVIDIA H200 GPU, while sampling is run on an NVIDIA RTX 4090 GPU. We use QM9 as an example and report diffusion hyperparameters in Table 7. On average, training takes 62.358 minutes on QM9 and 74.263 minutes on GEOM-Drugs. Sampling takes 420 milliseconds per molecule on average, close to our backbone GEOLDM (417 milliseconds). SensDiff is faster than many baselines, such as EDM (528 milliseconds), EEGSDE (731 milliseconds), GCDM (571 milliseconds), and UniGEM (587 milliseconds). With symbolic constraints enabled at sampling time, the cost increases slightly to 472 milliseconds per molecule. In summary, our training and sampling efficiency are comparable to those of popular molecular diffusion backbones.

Table 7: Hyperparameters used in the QM9 implementation of SensDiff.

Hyperparameter	**Value**
Data	
atom types C	5 (H, C, N, O, F)
normalize_factors	[1, 4, 10]
include_charges	True
3D Molecular Variational Autoencoder	
latent_nf	1
kl_weight	0.01
n_epochs	300
batch_size	64
learning rate	10^{-4}
ema_decay	0.9999
Diffusion Model	
diffusion_steps T	1000
noise_schedule	`polynomial_2`
noise_precision	10^{-5}
loss_type	ℓ_2
n_epochs	3000
batch_size	64
learning rate	10^{-4}
ema_decay	0.9999
ode_regularization	10^{-3}
gradient clipping	adaptive queue
test_epochs	20
n_stability_samples	1000
EGNN Backbone	
n_layers	9
nf (hidden)	256
attention	True
tanh gating	True
Bond Posterior Head	
hidden_nf	128
n_bond_classes M	5
activation	SiLU [16]
Symbolic Energies	
τ_c	0.05 Å
δ (Pseudo-Huber)	1.0
$\sigma_{\min}^{\text{bond}}$	0.03 Å
$\sigma_{\min}^{\text{angle}}$	0.03
σ_v	0.15
K_{nbr}	4
ε	10^{-8}
Training Curriculum	
stage_epochs	(500, 1400, 2200)
ramp_epochs	100
λ_c	4.0
λ_b	3.0
λ_a	2.0
λ_v	1.0
freeze spike ratio	$2.0\times$ rolling mean
freeze duration	50 epochs
Diffusion Sampling	
$\alpha_{\max}$	1.0
K	3

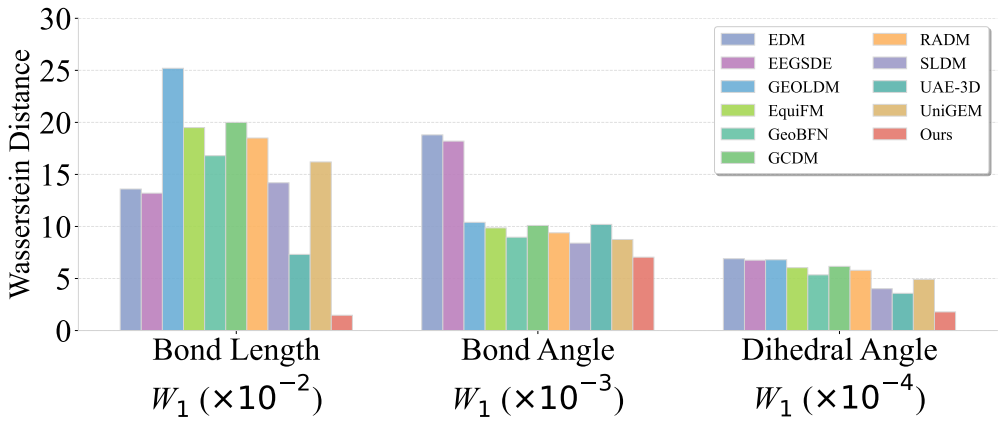


Figure 6: Distributional geometric compliance measured by Wasserstein distance on QM9 [58]. Lower values indicate a better match between the generated and reference geometry distributions.

E Geometric Plausibility of SensDiff

Empirical Bond-Length Distribution. The geometric structure of a molecule is important because it largely determines its physicochemical properties and functional behavior [32, 1, 73, 29]. As presented in Figure 7 and Figure 8, this study evaluates geometric compliance at the level of bond-specific empirical distributions. This analysis is motivated by the role of $E_{\text{bond_len}}$ in our method. Since $E_{\text{bond_len}}$ explicitly constrains 1-hop geometry by matching interatomic distances to chemically plausible bond-length statistics, a natural question is whether the generated molecules recover the same bond-length distributions as the original data, rather than merely achieving higher validity or stability scores. Therefore, we compare the empirical bond-length distributions of common bond types between the reference molecules and the molecules generated by SensDiff on both QM9 and GEOM-Drugs. This study directly tests whether the symbolic bond-length constraint is reflected in the local geometry of the generated samples. The results show that generated bond-length distributions closely track those of the original data across common bond types. On QM9, the agreement is strong in both peak location and spread, indicating that SensDiff preserves the dominant local bond geometry of the training distribution. On GEOM-Drugs, where the bond environments are more diverse and several bond types exhibit broader or multi-modal structure, the generated distributions still follow the main modes of the reference data well. This supports the central claim of the paper that the symbolic bond-length constraint improves geometric compliance at the local structural level, rather than only improving assessment metrics.

Distributional Geometric Compliance Across Methods. Figure 6 compares different methods through a unified distributional metric. While the empirical bond-length distribution provide a fine-grained view for individual bond types, they do not directly summarize how well a method matches the overall geometry distribution relative to competing baselines. Consequently, we measure geometric compliance using the Wasserstein distance between the generated and reference distributions. Importantly, we report this discrepancy not only for bond length, but also for bond angle and dihedral angle. This gives a comprehensive evaluation across 1-hop, 2-hop, and higher-order local geometry, and therefore tests whether improvements in bond-level structure propagate to richer geometric structure. The Wasserstein analysis further supports our main claim. SensDiff achieves the lowest distributional discrepancy among all compared methods on bond length, bond angle, and dihedral angle. The strongest gain appears in bond length, which is consistent with our design: bond length is the first chemically meaningful local signal enforced by the coarse-to-fine curriculum. At the same time, the advantage extends to bond angle and dihedral angle, showing that better 1-hop geometric consistency also improves higher-order local structure. Accordingly, these results indicate that SensDiff improves geometric compliance not only at the level of individual bond statistics, but also at the level of the overall geometry distribution.

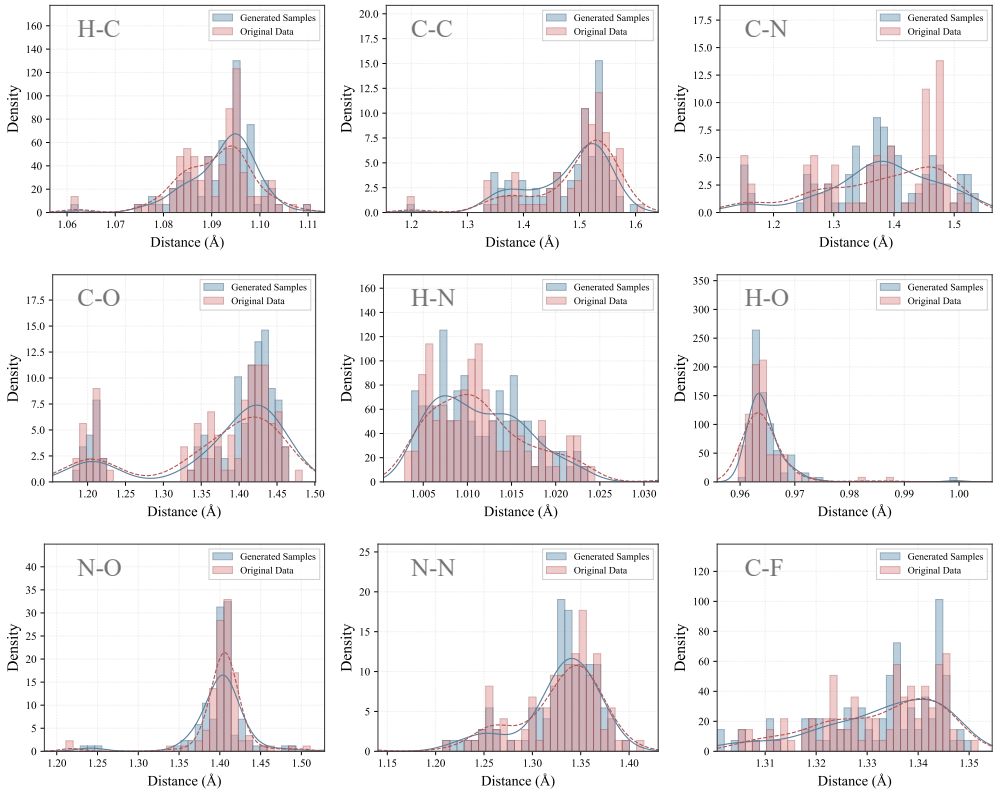


Figure 7: Common bond-length distribution in QM9 [58].

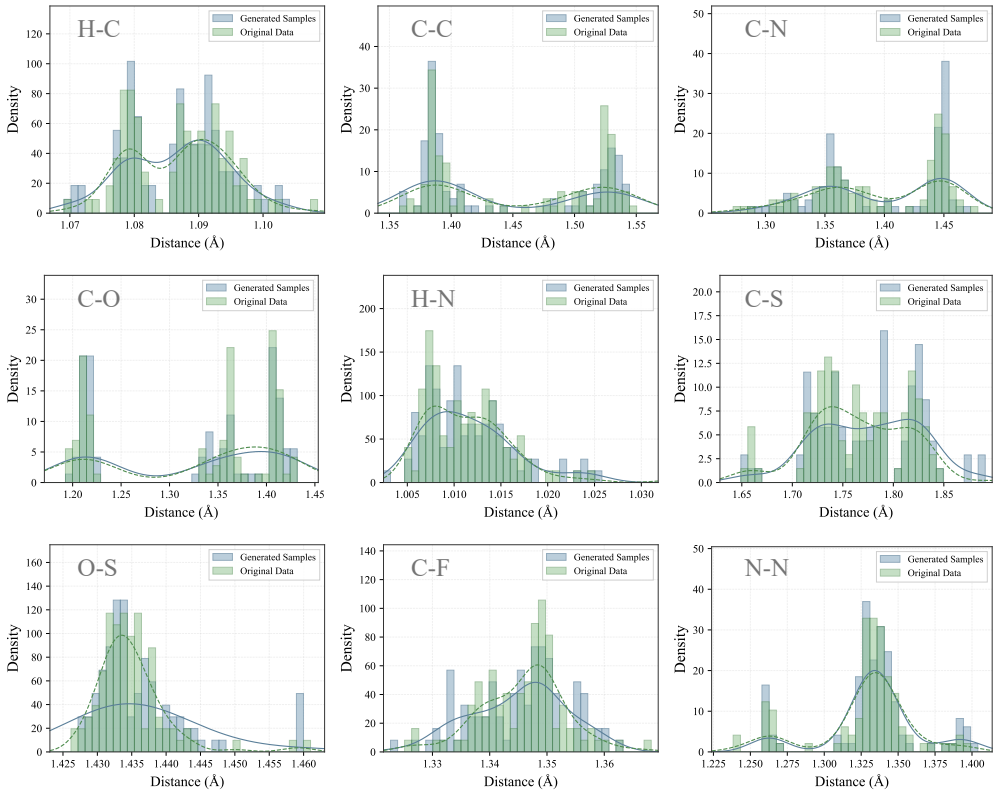


Figure 8: Common bond-length distribution in GEOM-Drugs [2].

F Unconditional 3D Molecule Generation

As presented in Table 8 and Table 9, SensDiff achieves the best performance across nearly all metrics on both QM9 and GEOM-Drugs. A particularly noteworthy observation is that existing methods tend to exhibit a trade-off between chemical validity and geometric accuracy: UAE-3D [49] attains low bond-length and bond-angle errors but suffers a significant drop in validity on GEOM-Drugs, whereas high-validity methods such as SLDM [52] and GEOLDM [69] show considerably larger geometric deviations. SensDiff resolves this tension by simultaneously achieving 97.2% validity and geometric errors comparable to or lower than UAE-3D [49], which we attribute to the coarse-to-fine curriculum that first establishes correct spatial geometry through steric constraints before enforcing finer chemical rules. The advantage of explicit constraint supervision becomes more pronounced on GEOM-Drugs, where molecular complexity is substantially higher. SensDiff surpasses the strongest baseline by 12.6% in molecule stability, suggesting that purely data-driven methods struggle to implicitly capture all relevant physicochemical regularities as molecular size grows. Meanwhile, SensDiff maintains the highest validity & uniqueness and competitive novelty, indicating that the symbolic constraints do not overly restrict generative diversity but rather guide the model to explore chemically plausible regions of the molecular space more effectively. These results collectively demonstrate that structured domain knowledge and data-driven generation are complementary. The former provides reliable inductive biases for fundamental physicochemical properties, while the latter captures complex distributional patterns that are difficult to encode symbolically.

As shown in Figure 9, Figure 10, and Figure 11, SensDiff exhibits a more structured denoising trajectory: atoms gradually coalesce into chemically valid arrangements, with bonds separating from early steps onward. In contrast, several baselines remain cluttered or exhibit overlapping atoms at intermediate steps, followed by abrupt rearrangements near the final step. We also show 30 randomly picked molecules from QM9 and GEOM-Drugs generations in Figure 12 and Figure 13, respectively.

Table 8: Unconditional 3D molecule generation results on QM9 [58]. Validity, uniqueness, novelty, and stability are reported as percentages (%). Geometric metrics (*e.g.,* bond length, bond angle, and dihedral angle) are reported as errors. **Best** results are in bold.

Model	Generation Quality (↑)					Geometric Fidelity (↓)		
	Valid	Valid & Unique	Novelty	Atom Sta	Mol Sta	Bond Length	Bond Angle	Dihedral Angle
EDM [30]	87.9	96.0	63.0	95.5	78.4	0.13600	0.01880	0.00069
EEGSDE [5]	89.8	94.3	61.3	95.7	80.3	0.13200	0.01820	0.00068
GEOLDM [69]	89.8	95.5	54.4	94.2	86.2	0.25200	0.01040	0.00068
EquiFM [62]	91.6	94.7	58.8	94.4	84.3	0.19500	0.00986	0.00060
GeoBFN [63]	90.7	93.3	63.6	94.9	86.9	0.16800	0.00897	0.00054
GCDM [50]	91.4	95.7	56.8	94.2	82.5	0.20000	0.01010	0.00062
RADM [14]	90.0	92.2	59.9	95.9	84.7	0.18500	0.00939	0.00058
SLDM [52]	92.1	89.1	71.3	95.0	90.9	0.14200	0.00840	0.00040
UAE-3D [49]	94.4	94.4	92.5	96.2	95.4	0.07320	0.01020	0.00036
UniGEM [20]	92.0	93.5	67.0	95.2	87.1	0.16200	0.00875	0.00049
SensDiff (Ours)	**96.9**	**97.2**	**92.6**	**98.1**	**95.6**	**0.01470**	**0.00705**	**0.00018**

Table 9: Unconditional 3D molecule generation results on GEOM-Drugs [2]. Validity, uniqueness, novelty, and stability are reported as percentages (%). Geometric metrics (*e.g.,* bond length, bond angle, and dihedral angle) are reported as errors. **Best** results are in bold.

Model	Generation Quality (↑)					Geometric Fidelity (↓)		
	Valid	Valid & Unique	Novelty	Atom Sta	Mol Sta	Bond Length	Bond Angle	Dihedral Angle
EDM [30]	88.8	96.6	89.0	77.5	41.5	0.44700	0.51200	0.01520
EEGSDE [5]	89.1	94.0	87.2	79.1	46.5	0.42700	0.49300	0.01440
GEOLDM [69]	95.6	95.6	90.5	80.7	49.3	0.40700	0.43500	0.01750
EquiFM [62]	95.7	94.9	88.5	80.3	47.5	0.41700	0.45600	0.01640
GeoBFN [63]	88.5	96.4	88.1	81.5	52.2	0.38300	0.41300	0.01370
GCDM [50]	91.2	96.3	91.7	84.5	58.8	0.30000	0.29200	0.00884
RADM [14]	95.7	94.9	89.6	82.1	53.2	0.36600	0.40000	0.01310
SLDM [52]	95.8	95.3	89.7	84.8	62.0	0.24900	0.23500	0.00678
UAE-3D [49]	84.4	94.4	92.8	82.5	66.8	**0.01040**	0.00529	**0.00018**
UniGEM [20]	94.7	95.1	90.4	82.1	54.2	0.35300	0.39000	0.01180
SensDiff (Ours)	**97.2**	**98.8**	**91.8**	**87.9**	**79.4**	0.01470	**0.00505**	**0.00018**

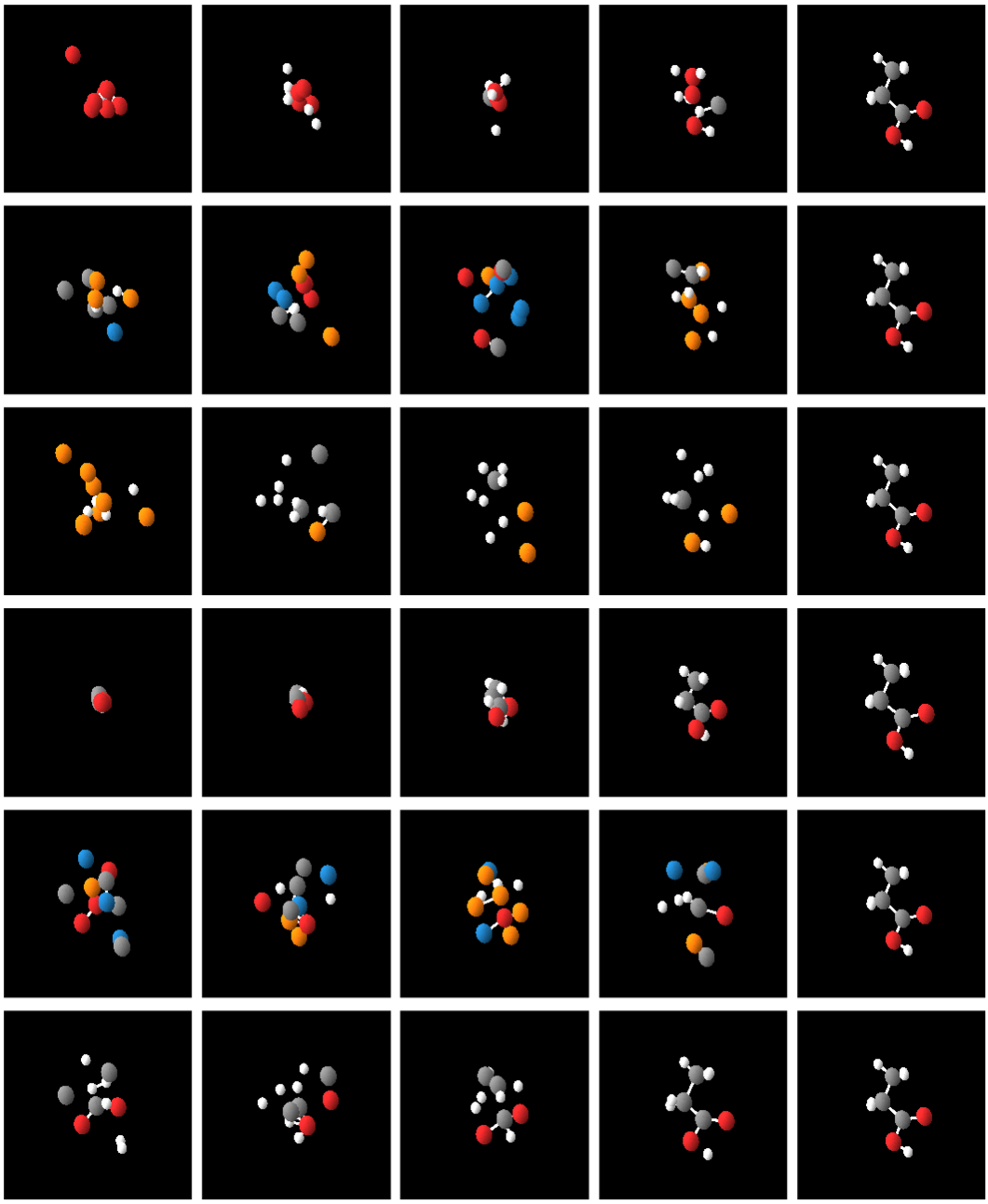


Figure 9: Comparison of molecule generation trajectories for the same sample (ID: 004) across methods. From top to bottom, we show samples generated by SensDiff (ours), GEOLDM [69], RADM [14], SLDM [52], UAE-3D [49], and UniGEM [20].

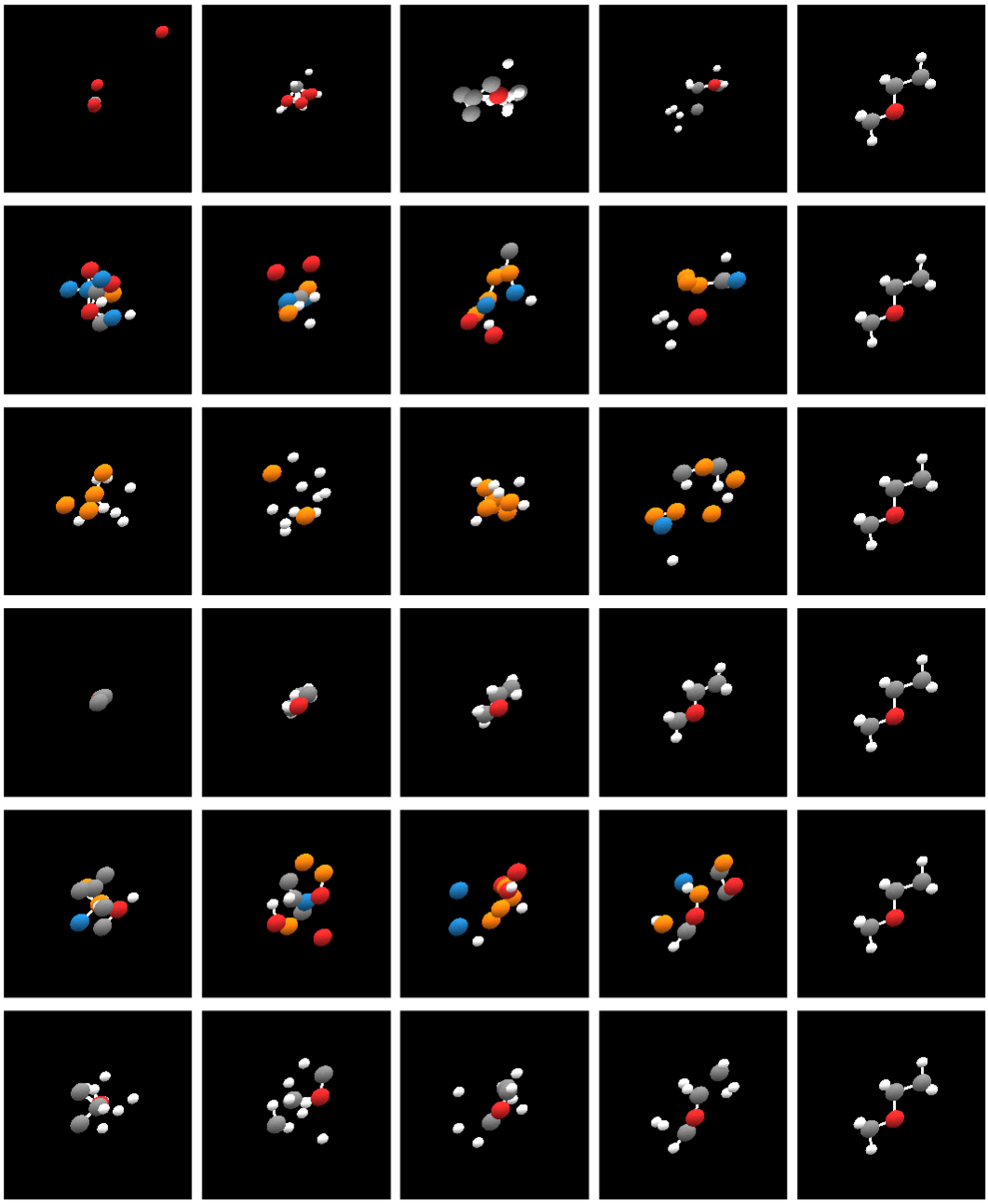


Figure 10: Comparison of molecule generation trajectories for the same sample (ID: 013) across methods. From top to bottom, we show samples generated by SensDiff (ours), GEOLDM [69], RADM [14], SLDM [52], UAE-3D [49], and UniGEM [20].

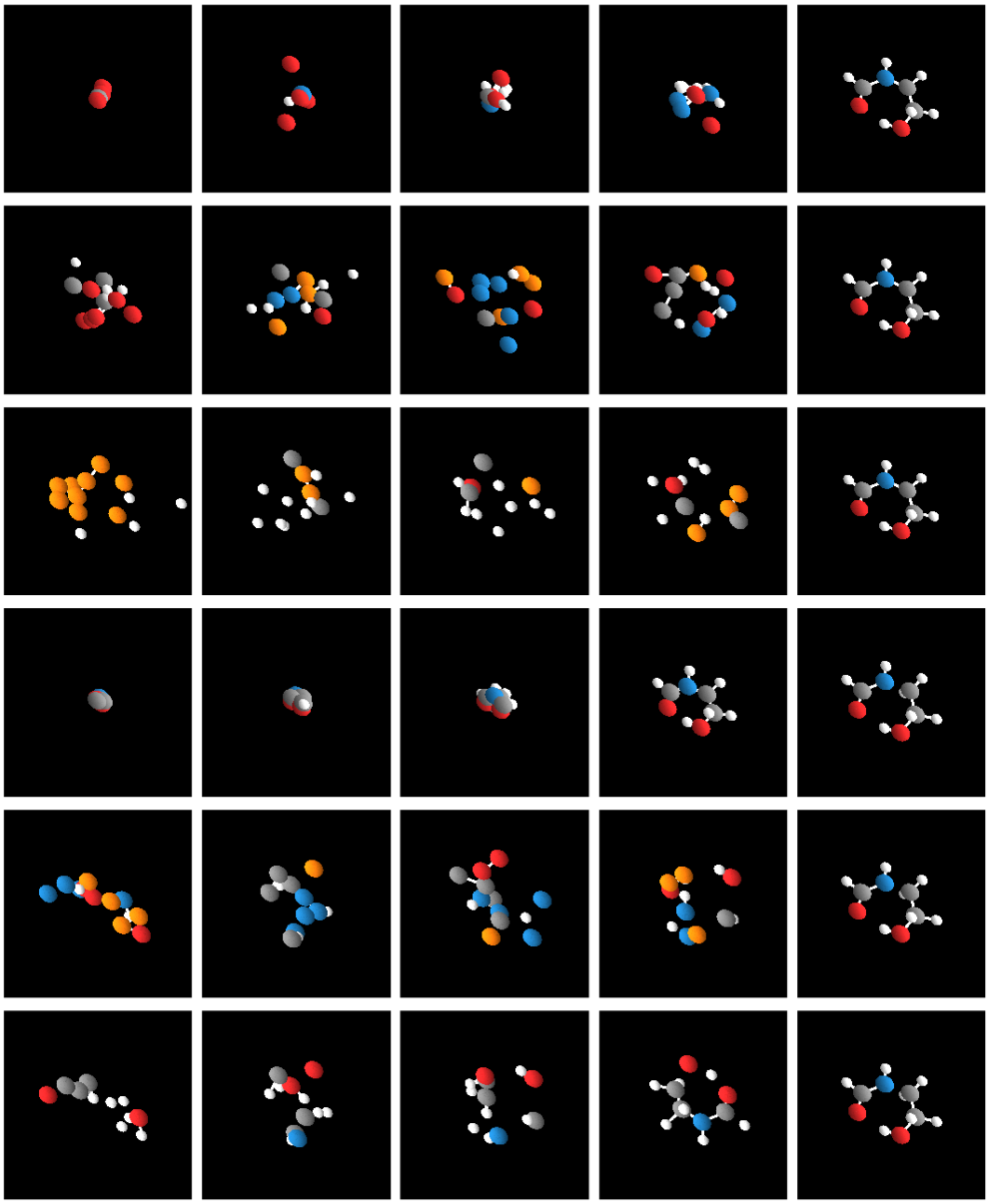


Figure 11: Comparison of molecule generation trajectories for the same sample (ID: 026) across methods. From top to bottom, we show samples generated by SensDiff (ours), GEOLDM [69], RADM [14], SLDM [52], UAE-3D [49], and UniGEM [20].

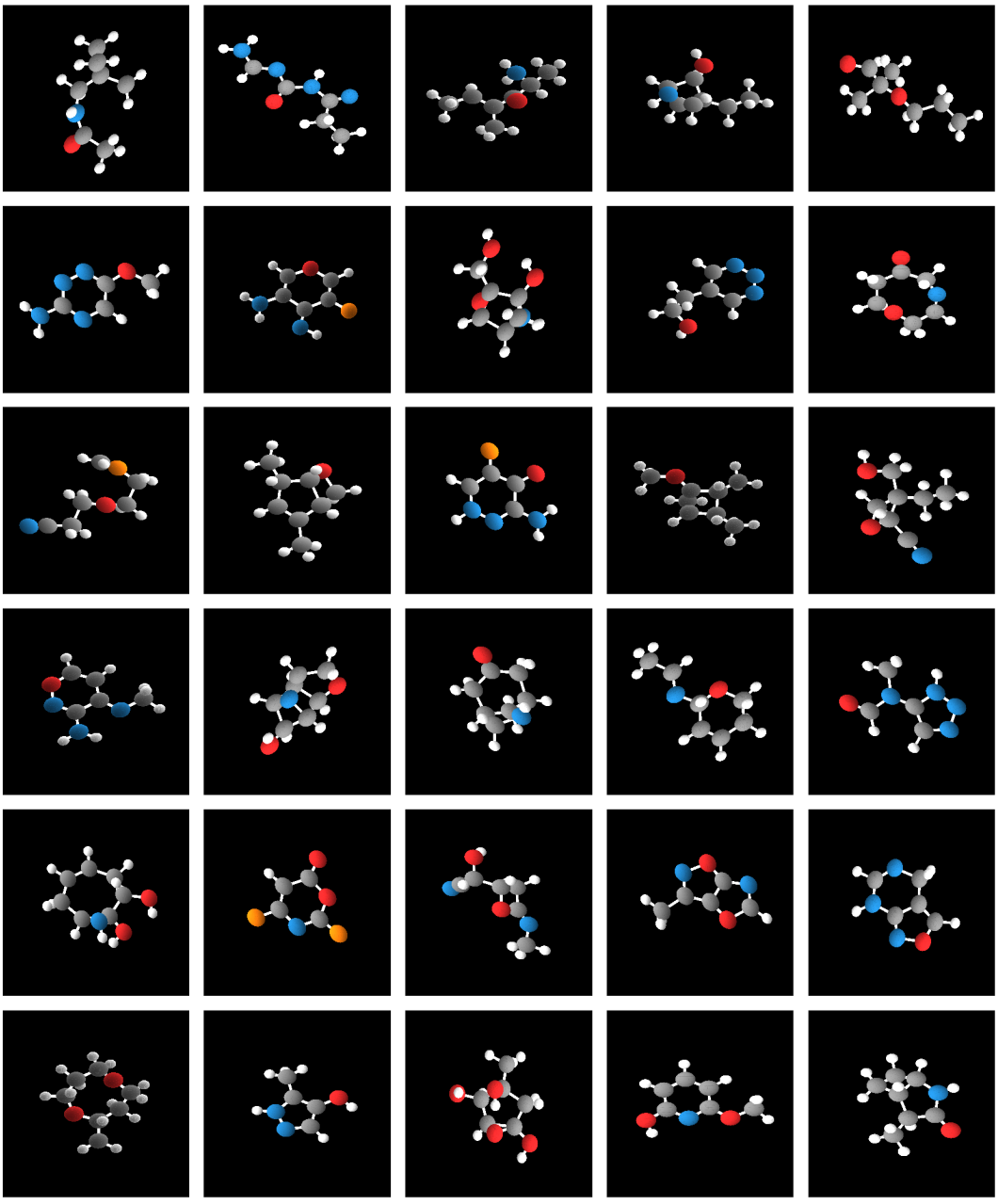


Figure 12: Random samples from QM9 [58] generations.

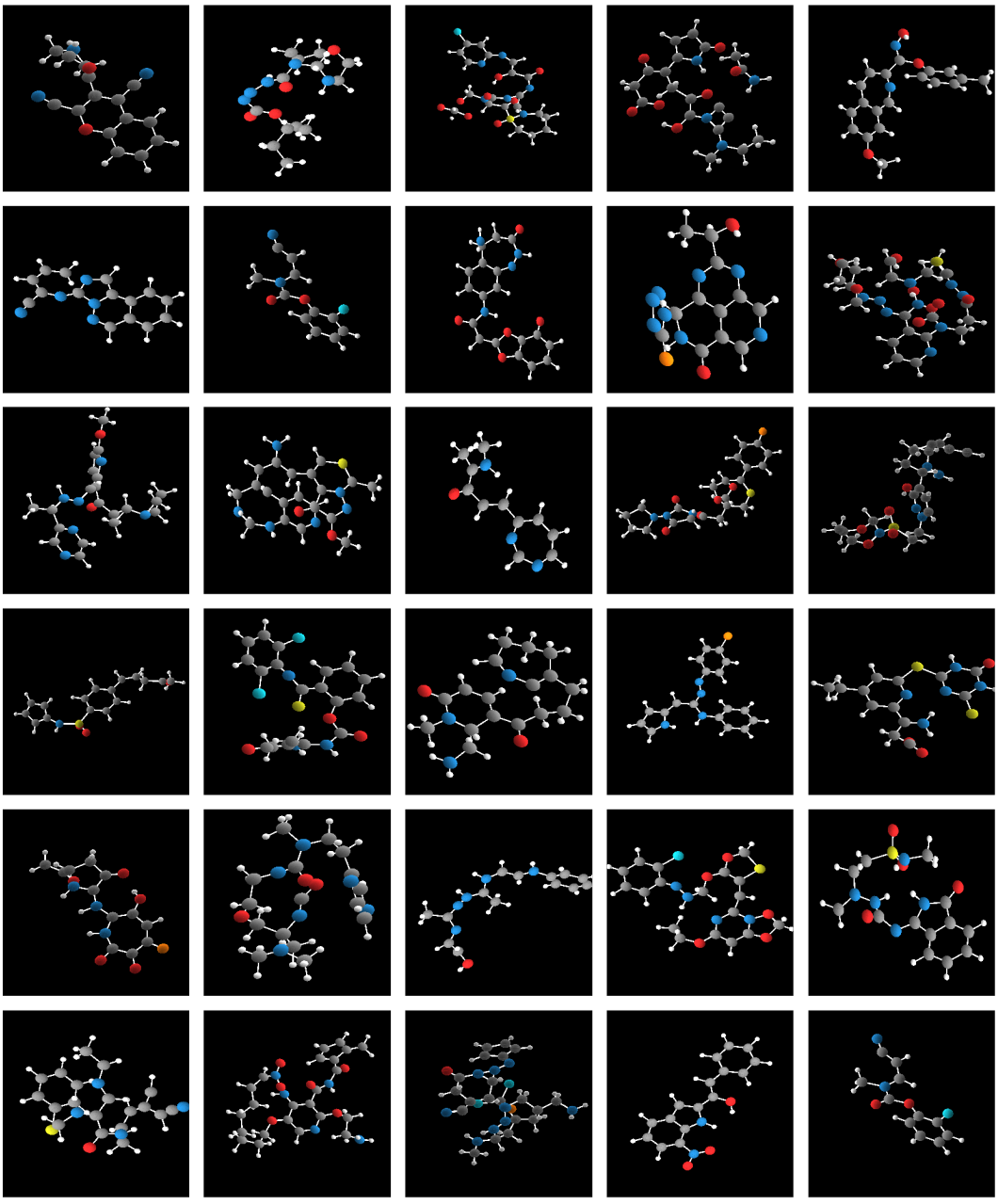


Figure 13: Random samples from GEOM-Drugs [2] generations.

G Conditional 3D Molecule Generation with Quantum-Chemical Properties

As reported in Table 10, SensDiff achieves the highest molecule stability and validity across all six conditional generation tasks. The gains are especially significant in molecule stability, where SensDiff surpasses the strongest baseline SLDM [52] by up to 2.1% on μ and 1.9% on $\Delta\varepsilon$, suggesting that the explicit physicochemical constraints facilitate the model to avoid generating chemically implausible structures. These results confirm that enforcing geometric rules during training not only improves property-level accuracy but also enhances the overall structural quality of the generated molecules.

As shown in Table 11–14, SensDiff achieves the lowest MAE on every property pair while simultaneously obtaining the highest molecule stability and validity, confirming that physicochemical plausibility and property accuracy are complementary rather than competing objectives. The improvements are most pronounced for geometry-sensitive electronic properties such as $\varepsilon_{\mathrm{LUMO}}$ and $\Delta\varepsilon$, where MAE drops by over 30% relative to the strongest baseline (Table 12), suggesting that explicit bond-length and valence constraints yield more physically realistic geometries from which accurate electronic structure naturally follows. Under dual-property conditioning, a more challenging setting that tempts the model to distort geometry to satisfy two targets simultaneously, SensDiff still maintains validity above 0.925 across all pairs (Table 14), whereas baselines frequently fall below 0.90 (Table 11 and 13), indicating that symbolic constraints effectively regularize against invalid structural deformation. The consistent gains over EEGSDE [5] (Table 13), which also leverages constraints during sampling, further demonstrate that embedding domain knowledge into training is more effective than purely post-hoc gradient guidance, as the model internalizes geometric priors rather than correcting an already imperfect score function. Finally, the uniform improvements across thermodynamic (C_v), electronic ($\varepsilon_{\mathrm{HOMO}}$, $\varepsilon_{\mathrm{LUMO}}$), and electrostatic ($\mu$) properties indicate that the benefit is not property-specific but stems from a fundamentally improved molecular representation. We also visualize molecules generated under quantum-property control, with single-property results in Figure 14 and double-property results in Figure 15.

Table 10: Validity and molecule stability (%, ↑) for quantum-property conditional generation on QM9 [58]. **Best** results are marked in bold.

Model	C_v ($\frac{\text{cal}}{\text{mol}}$ K)		μ (D)		α (Bohr^3)		$\Delta\varepsilon$ (meV)		$\varepsilon_{\mathrm{HOMO}}$ (meV)		$\varepsilon_{\mathrm{LUMO}}$ (meV)	
	Mol Sta	Valid	Mol Sta	Valid	Mol Sta	Valid	Mol Sta	Valid	Mol Sta	Valid	Mol Sta	Valid
EDM [30]	79.6	86.4	83.4	88.9	80.3	87.1	82.7	89.2	77.8	84.8	81.9	88.3
EEGSDE [5]	81.1	87.8	84.7	90.1	82.4	88.4	84.1	89.6	79.9	86.9	83.6	89.2
GEOLDM [69]	82.5	88.1	85.9	91.0	83.7	89.4	85.3	90.2	80.9	87.4	84.4	89.8
EquiFM [62]	82.2	87.9	85.6	90.7	83.4	89.1	84.9	89.9	80.6	87.2	84.1	89.6
GeoBFN [63]	81.8	87.6	85.1	90.2	82.9	88.8	84.6	89.6	80.2	87.1	83.9	89.1
GCDM [50]	80.9	87.3	84.2	89.3	81.9	87.9	83.7	89.1	79.6	86.8	82.9	88.6
RADM [14]	81.9	87.7	84.9	89.9	82.8	88.7	84.6	89.9	80.4	87.2	83.7	89.4
SLDM [52]	83.4	88.9	86.7	91.2	84.6	89.8	86.2	90.9	81.6	87.9	**85.4**	90.3
UAE-3D [49]	82.8	**90.2**	87.4	91.8	86.2	88.2	86.5	89.4	83.0	86.7	84.9	**90.6**
UniGEM [20]	82.9	88.3	86.2	90.8	84.1	89.2	85.6	90.1	81.2	87.6	84.8	89.7
SensDiff (Ours)	**83.6**	89.8	**88.8**	**92.1**	**86.5**	**89.7**	**88.1**	**91.8**	**83.0**	**88.4**	85.3	90.5

Table 11: Double-property conditional generation on QM9 [58] with EDM [30].

Property 1				Property 2			
Metric	MAE ↓	Mol Sta ↑	Valid ↑	Metric	MAE ↓	Mol Sta ↑	Valid ↑
C_v ($\frac{\text{cal}}{\text{mol}}$ K)	1.079	0.768	0.898	μ (D)	1.156	0.768	0.898
α (Bohr^3)	2.760	0.772	0.899	μ (D)	1.158	0.772	0.899
$\Delta\varepsilon$ (meV)	683	0.759	0.896	μ (D)	1.130	0.759	0.896
$\varepsilon_{\mathrm{HOMO}}$ (meV)	371	0.760	0.895	μ (D)	1.041	0.760	0.895
$\varepsilon_{\mathrm{HOMO}}$ (meV)	364	0.748	0.892	$\varepsilon_{\mathrm{LUMO}}$ (meV)	612	0.748	0.892
$\varepsilon_{\mathrm{HOMO}}$ (meV)	382	0.745	0.888	C_v ($\frac{\text{cal}}{\text{mol}}$ K)	1.118	0.745	0.888
$\varepsilon_{\mathrm{HOMO}}$ (meV)	376	0.749	0.890	α (Bohr^3)	2.880	0.749	0.890
$\Delta\varepsilon$ (meV)	702	0.754	0.894	$\varepsilon_{\mathrm{LUMO}}$ (meV)	505	0.754	0.894

Table 12: Double-property conditional generation on QM9 [58] with UniGEM [20].

Property 1				Property 2			
Metric	MAE ↓	Mol Sta ↑	Valid ↑	Metric	MAE ↓	Mol Sta ↑	Valid ↑
C_v $\left(\frac{\text{cal}}{\text{mol}}\,\text{K}\right)$	0.952	0.811	0.907	μ (D)	0.876	0.811	0.907
α (Bohr^3)	2.430	0.818	0.911	μ (D)	0.838	0.818	0.911
$\Delta\varepsilon$ (meV)	544	0.796	0.903	μ (D)	0.856	0.796	0.903
$\varepsilon_{\text{HOMO}}$ (meV)	246	0.829	0.909	μ (D)	0.848	0.829	0.909
$\varepsilon_{\text{HOMO}}$ (meV)	252	0.822	0.905	$\varepsilon_{\text{LUMO}}$ (meV)	556	0.822	0.905
$\varepsilon_{\text{HOMO}}$ (meV)	264	0.808	0.898	C_v $\left(\frac{\text{cal}}{\text{mol}}\,\text{K}\right)$	0.926	0.808	0.898
$\varepsilon_{\text{HOMO}}$ (meV)	259	0.814	0.901	α (Bohr^3)	2.490	0.814	0.901
$\Delta\varepsilon$ (meV)	548	0.801	0.900	$\varepsilon_{\text{LUMO}}$ (meV)	482	0.801	0.900

Table 13: Double-property conditional generation on QM9 [58] with EEGSDE [5].

Property 1				Property 2			
Metric	MAE ↓	Mol Sta ↑	Valid ↑	Metric	MAE ↓	Mol Sta ↑	Valid ↑
C_v $\left(\frac{\text{cal}}{\text{mol}}\,\text{K}\right)$	0.981	0.754	0.892	μ (D)	0.912	0.754	0.892
α (Bohr^3)	2.610	0.770	0.897	μ (D)	0.855	0.770	0.897
$\Delta\varepsilon$ (meV)	563	0.727	0.890	μ (D)	0.866	0.727	0.890
$\varepsilon_{\text{HOMO}}$ (meV)	328	0.752	0.889	μ (D)	0.894	0.752	0.889
$\varepsilon_{\text{HOMO}}$ (meV)	319	0.741	0.887	$\varepsilon_{\text{LUMO}}$ (meV)	536	0.741	0.887
$\varepsilon_{\text{HOMO}}$ (meV)	341	0.739	0.882	C_v $\left(\frac{\text{cal}}{\text{mol}}\,\text{K}\right)$	1.006	0.739	0.882
$\varepsilon_{\text{HOMO}}$ (meV)	333	0.745	0.885	α (Bohr^3)	2.670	0.745	0.885
$\Delta\varepsilon$ (meV)	587	0.736	0.888	$\varepsilon_{\text{LUMO}}$ (meV)	458	0.736	0.888

Table 14: Double-property conditional generation on QM9 [58] with SensDiff (Ours).

Property 1				Property 2			
Metric	MAE ↓	Mol Sta ↑	Valid ↑	Metric	MAE ↓	Mol Sta ↑	Valid ↑
C_v $\left(\frac{\text{cal}}{\text{mol}}\,\text{K}\right)$	0.672	0.846	0.934	μ (D)	0.721	0.846	0.934
α (Bohr^3)	1.920	0.855	0.938	μ (D)	0.748	0.855	0.938
$\Delta\varepsilon$ (meV)	449	0.832	0.929	μ (D)	0.757	0.832	0.929
$\varepsilon_{\text{HOMO}}$ (meV)	219	0.864	0.936	μ (D)	0.782	0.864	0.936
$\varepsilon_{\text{HOMO}}$ (meV)	226	0.857	0.932	$\varepsilon_{\text{LUMO}}$ (meV)	323	0.857	0.932
$\varepsilon_{\text{HOMO}}$ (meV)	238	0.841	0.925	C_v $\left(\frac{\text{cal}}{\text{mol}}\,\text{K}\right)$	0.752	0.841	0.925
$\varepsilon_{\text{HOMO}}$ (meV)	234	0.849	0.929	α (Bohr^3)	1.930	0.849	0.929
$\Delta\varepsilon$ (meV)	437	0.836	0.927	$\varepsilon_{\text{LUMO}}$ (meV)	368	0.836	0.927

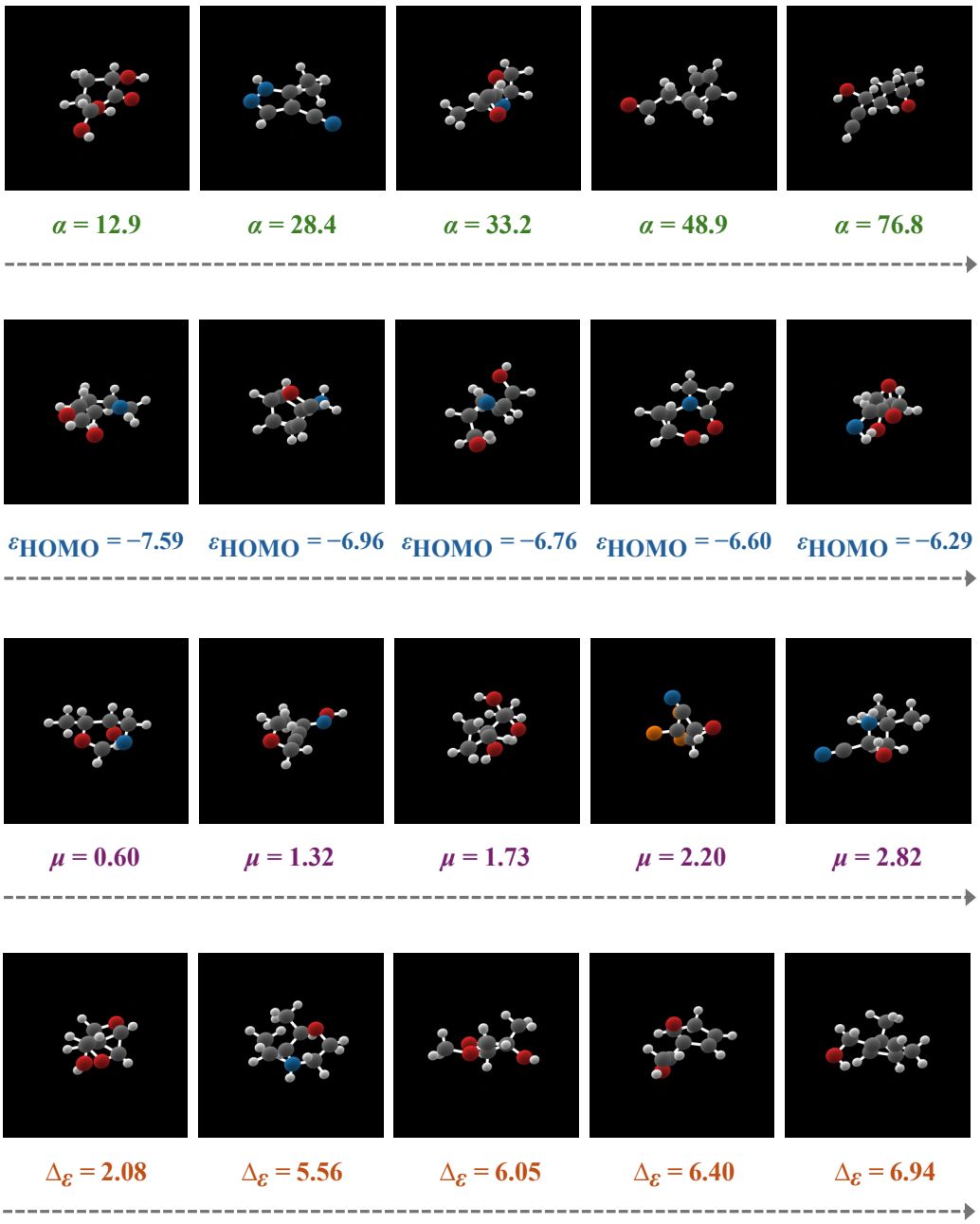


Figure 14: Conditional generation on QM9 [58] targeting one quantum property.

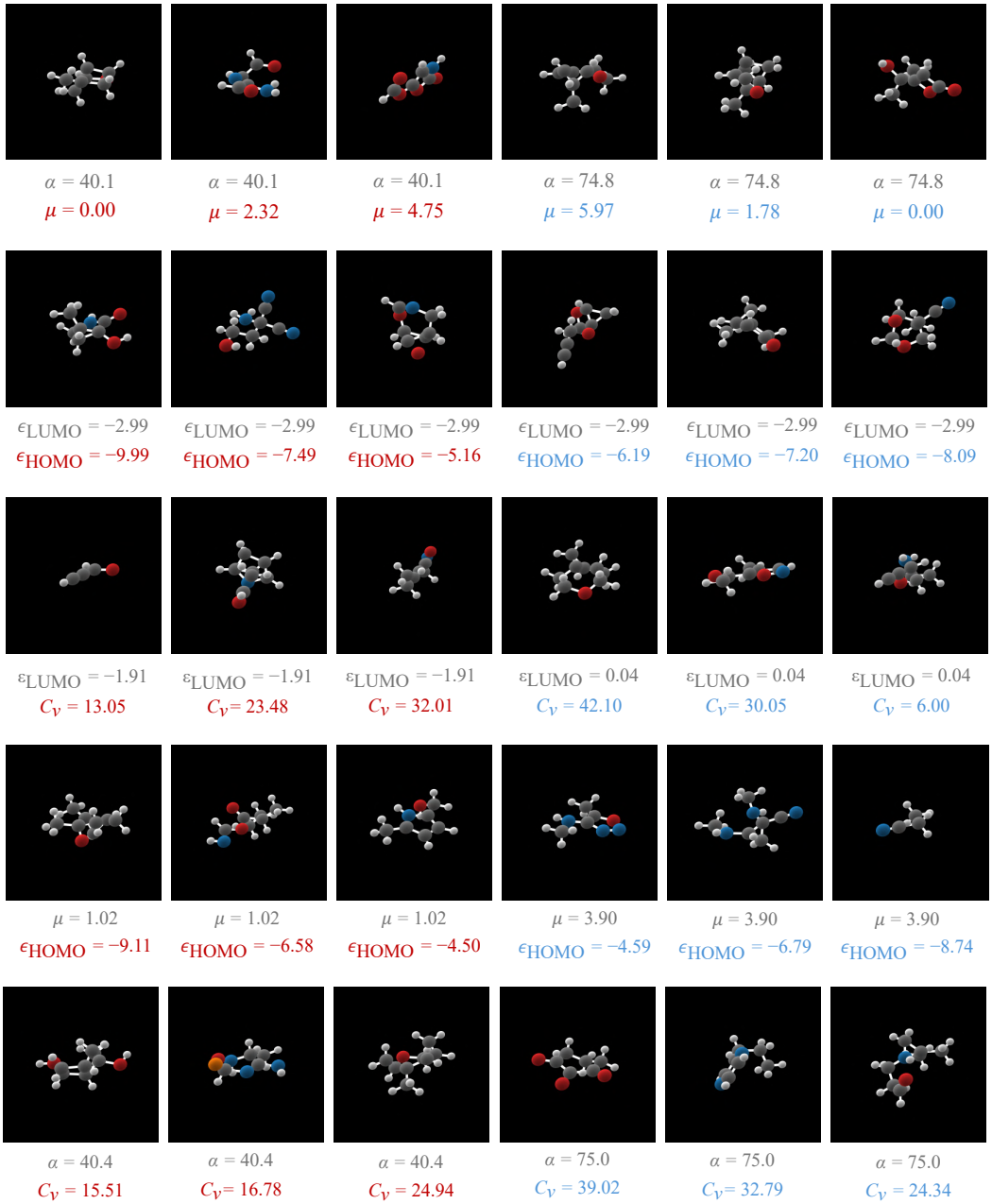


Figure 15: Conditional generation on QM9 [58] targeting two quantum properties simultaneously (increasing or decreasing each).

H Conditional 3D Molecule Generation for Multi-Objective Drug Design

In this task, we consider two settings: (1) oral drug design (detailed in Section 4) and (2) designing drug under Lipinski's Rule-of-Five constraints, requiring HBA ≤ 10, HBD ≤ 5, MW ≤ 500 Da, LogP ≤ 5, SA Score < 5, and QED > 0.7 [45, 65]. Figure 16 shows the distributions and denoising trajectory of MW, HBA, HBD, and TPSA for oral drug design. In this study, we impose these constraints only during sampling by formulating them as symbolic energy constraints and applying them to the pretrained model at sampling time. Figure 17 and Figure 18 show the distributions and denoising trajectory of LogP/SA Score/QED and HBA/HBD/MW, respectively, for Lipinski-constrained drug design. The property distributions of molecules generated by SensDiff exhibit the sharpest peaks and the closest alignment with the target distributions across all descriptors, whereas baselines such as UniGEM [20] and SLDM [52] produce noticeably broader and shifted distributions. The bottom rows of each figure further reveal that the generated property distributions progressively sharpen from an uniform spread at t=1000 to a concentrated mode at t=0, suggesting that the denoising process smoothly steers molecules toward the desired property targets. These results collectively showcase that SensDiff achieves fine-grained property control under both tasks. We also visualize representative generated molecules for both settings in Figures 19 and 20, respectively.

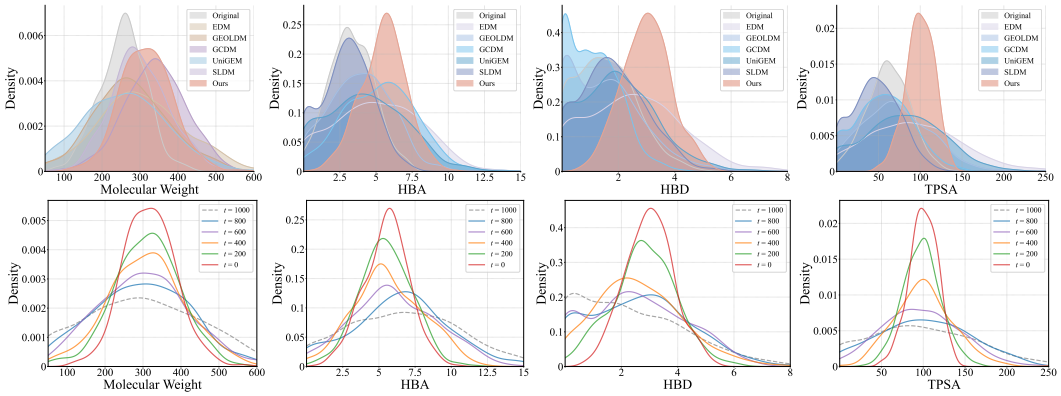


Figure 16: (**Upper**) The generated molecules precisely satisfy the specified oral drug design constraints (*e.g.,* MW, HBA, HBD, and TPSA). (**Bottom**) Molecules converge toward the target values across denoising steps.

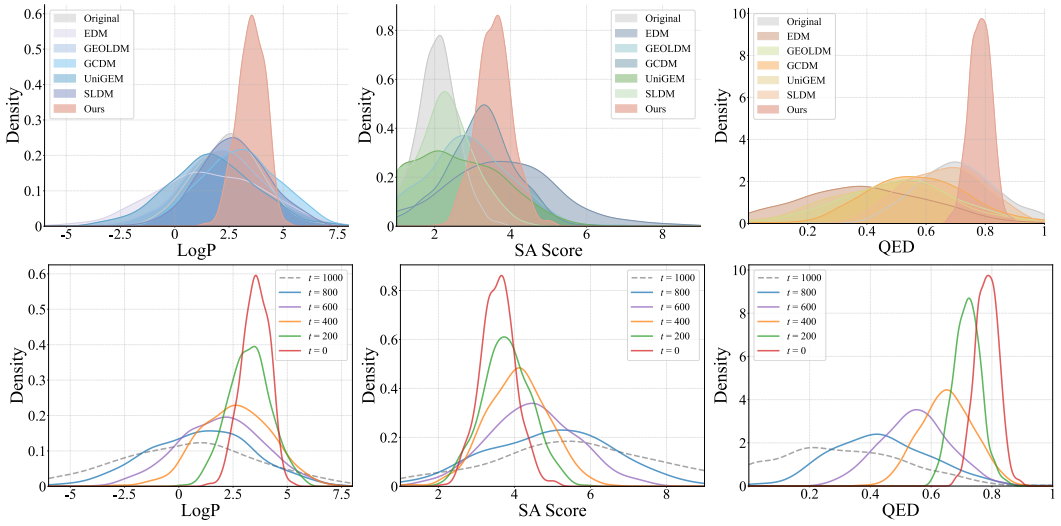


Figure 17: (**Upper**) The generated molecules precisely satisfy the specified Lipinski's Rule-of-Five constraints (*e.g.,* LogP, SA Score, and QED). (**Bottom**) Molecules converge toward the target values across denoising steps.

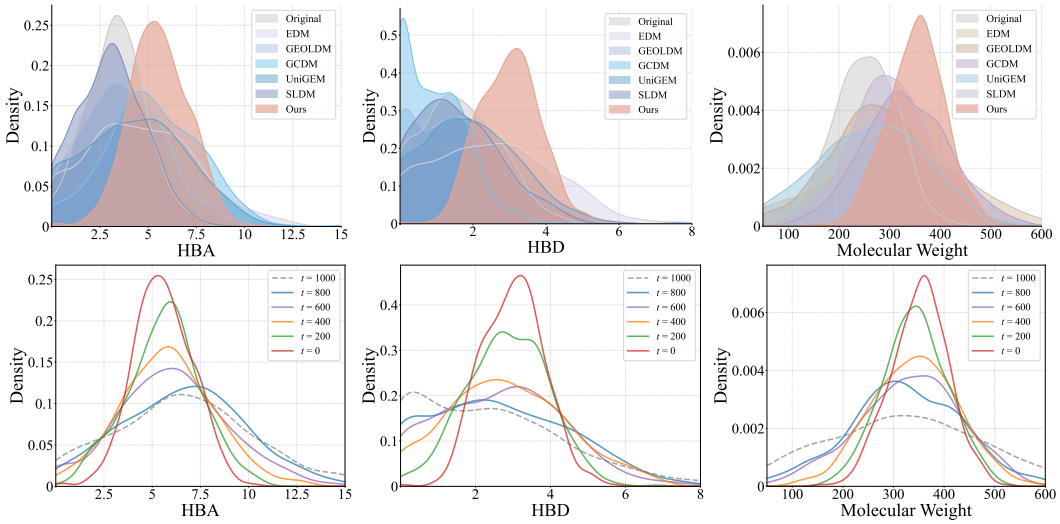


Figure 18: (**Upper**) The generated molecules precisely satisfy the specified Lipinski's Rule-of-Five constraint (*e.g.,* HBA, HBD, and MW). (**Bottom**) Molecules converge toward the target values across denoising steps.

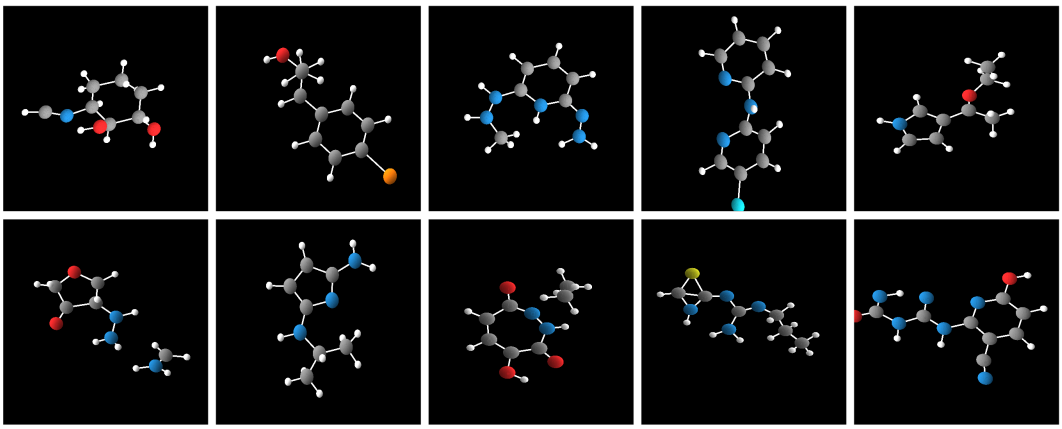


Figure 19: Random samples of molecules satisfying the oral drug design constraints.

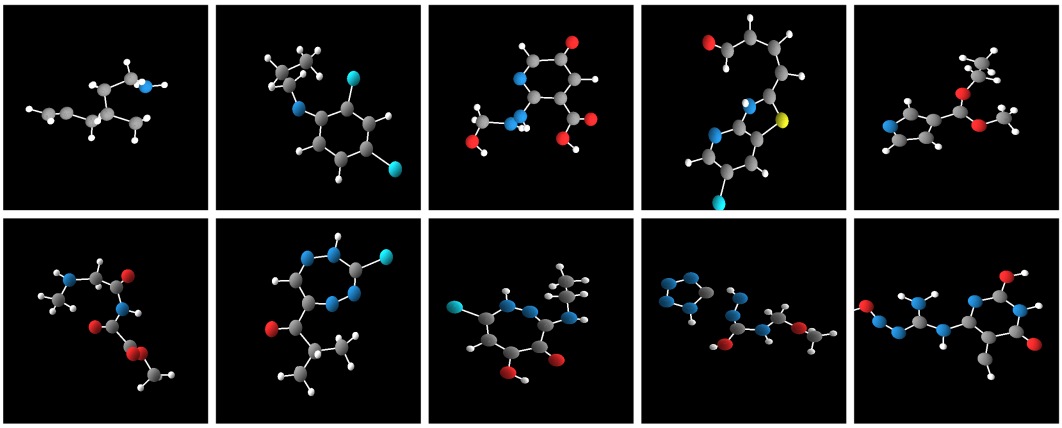


Figure 20: Random samples of molecules satisfying Lipinski's Rule-of-Five constraints.

I Additional Ablation Studies

In this section, we systematically ablate each component of SensDiff. Notably, "/" indicates the enforcement order (*e.g.,* E_3/E_1 enforces E_3 followed by E_1). Due to space limitations, we denote the four constraints E_{clash} (*i.e.,* steric non-overlap), $E_{\text{bond_len}}$ (*i.e.,* bond length), E_{angle} (*i.e.,* bond angle), and E_{val} (*i.e.,* valence validity) as E_1, E_2, E_3, and E_4, respectively, following their enforcement order during training. The four columns, E_1, E_2, E_3, and E_4, report the constraint satisfaction rates of the four constraints. Generation quality and constraint-compliance metrics are reported as percentages, whereas MAE is reported as a numeric value.

Table 15: Ablation on constraint sets without sequential enforcement (*i.e.,* all constraints are activated simultaneously from the start of training). The diffusion models are trained on QM9 [58].

Constraint	Generation Quality				Constraint Satisfaction			
	Valid	Valid & Unique	Atom Stab	Mol Stab	E_1	E_2	E_3	E_4
E_1	90.9	98.4	94.8	87.0	94.8	75.4	77.2	80.5
E_2	92.9	98.6	94.9	86.2	78.4	96.6	81.2	84.1
E_3	93.2	98.5	94.0	87.1	79.9	83.5	95.3	85.3
E_4	93.5	98.6	95.3	88.8	81.7	84.5	86.2	96.9
E_1, E_3	90.1	97.7	94.9	84.4	92.2	76.8	94.5	81.5
E_1, E_4	91.3	97.6	92.7	83.7	92.8	78.2	82.7	95.9
E_1, E_2	91.7	97.9	94.6	84.3	92.3	93.6	79.2	82.6
E_2, E_3	92.4	98.0	94.4	85.0	79.6	93.6	94.1	84.5
E_2, E_4	92.7	98.1	94.8	85.9	81.1	94.0	83.5	95.6
E_3, E_4	92.5	97.9	94.5	85.7	80.5	83.9	94.2	95.8
E_1, E_2, E_3	89.3	97.0	92.2	79.5	90.2	92.4	92.8	83.2
E_1, E_2, E_4	89.6	97.1	92.5	80.2	90.5	92.5	84.1	93.8
E_2, E_3, E_4	90.6	97.3	93.0	81.8	81.8	92.7	92.9	94.0
E_1, E_3, E_4	88.9	96.9	91.8	79.2	90.6	75.2	93.3	94.7
E_1, E_2, E_3, E_4	89.4	96.5	92.4	78.8	89.6	93.2	92.4	94.1

Table 16: Ablation on constraint sets without sequential enforcement (*i.e.,* all constraints are activated simultaneously from the start of training). The diffusion models are trained on GEOM-Drugs [2].

Constraint	Generation Quality				Constraint Satisfaction			
	Valid	Valid & Unique	Atom Stab	Mol Stab	E_1	E_2	E_3	E_4
E_1	91.3	98.5	83.2	68.5	91.5	70.7	72.1	74.3
E_2	92.6	98.6	83.6	69.2	72.9	92.9	74.6	76.8
E_3	92.8	98.5	83.0	70.0	74.1	76.8	91.7	78.2
E_4	93.3	98.6	84.3	72.0	75.2	77.7	79.5	93.3
E_1, E_3	90.2	97.8	82.8	66.0	89.3	72.2	90.5	75.3
E_1, E_4	90.7	97.8	82.1	65.4	89.9	73.2	76.2	91.8
E_1, E_2	91.1	97.9	83.1	65.7	88.8	90.1	73.6	76.0
E_2, E_3	91.8	98.0	83.2	67.2	74.3	90.2	90.3	77.6
E_2, E_4	92.3	98.1	83.8	68.8	75.5	90.7	77.2	92.0
E_3, E_4	92.1	98.0	83.5	68.5	74.9	77.8	90.4	92.2
E_1, E_2, E_3	88.8	97.1	81.1	61.7	86.9	88.9	89.2	76.8
E_1, E_2, E_4	89.2	97.2	81.7	63.0	87.2	89.2	77.9	90.2
E_2, E_3, E_4	90.1	97.3	82.2	64.7	76.2	89.4	89.4	90.4
E_1, E_3, E_4	88.5	97.1	80.7	61.3	87.2	70.6	89.4	90.7
E_1, E_2, E_3, E_4	89.0	96.9	81.3	60.8	86.3	89.3	88.5	90.0

Table 17: Ablation on constraint enforcement order (sequential constraint activation). The diffusion models are trained on QM9 [58].

Constraint	Generation Quality				Constraint Satisfaction			
	Valid	Valid & Unique	Atom Stab	Mol Stab	E_1	E_2	E_3	E_4
E_3/E_1	86.5	97.9	86.2	75.2	88.1	78.4	92.1	82.6
E_4/E_1	82.7	97.3	84.6	69.6	89.8	77.2	80.4	92.7
E_3/E_2	90.2	98.8	89.8	81.3	82.9	95.5	91.9	84.9
E_4/E_2	87.3	98.2	87.2	77.4	82.5	95.2	83.8	93.9
E_1/E_2	84.2	97.5	85.4	71.2	90.2	90.3	79.2	82.0
E_1/E_3	85.9	97.7	86.0	73.6	89.5	77.8	91.5	82.2
E_1/E_4	83.4	97.4	85.1	70.3	90.6	76.6	80.2	91.9
E_2/E_3	89.0	98.6	88.9	79.7	81.7	94.8	91.4	84.2
E_2/E_4	86.3	98.1	86.7	76.2	81.6	94.4	82.7	93.1
E_3/E_4	87.8	98.2	88.1	78.0	82.1	83.3	91.7	92.7
E_4/E_3	89.3	98.7	89.3	80.4	82.9	84.4	91.6	94.4
E_2/E_1	84.7	97.7	85.7	72.5	88.5	90.6	79.0	81.8
$E_3/E_2/E_1$	88.5	98.5	88.7	78.4	88.1	92.3	91.4	83.8
$E_4/E_2/E_1$	87.2	98.3	88.0	76.6	88.2	92.3	82.6	93.2
$E_4/E_3/E_1$	88.0	98.3	88.5	77.3	88.5	83.5	91.4	93.2
$E_4/E_3/E_2$	90.2	98.8	90.3	81.9	83.7	92.1	92.3	94.5
$E_3/E_4/E_2$	89.5	98.7	89.6	80.7	83.0	91.8	91.7	93.6
$E_3/E_1/E_4$	80.3	96.6	84.3	66.8	86.4	81.3	90.3	88.7
$E_3/E_1/E_2$	82.6	97.0	85.1	70.5	86.3	88.0	90.6	83.3
$E_4/E_1/E_3$	84.3	97.5	85.9	72.4	87.3	82.5	89.8	92.0
$E_4/E_1/E_2$	83.6	97.3	85.5	71.1	87.2	87.9	81.7	91.9
$E_4/E_3/E_1/E_2$	89.9	99.0	91.0	83.0	89.5	91.4	92.4	94.6
$E_3/E_2/E_1/E_4$	84.7	97.6	87.0	72.8	86.9	90.1	91.1	88.5
$E_2/E_4/E_1/E_3$	83.1	97.3	86.3	70.9	85.2	90.3	87.6	90.6
$E_2/E_4/E_3/E_1$	87.5	98.2	88.8	77.4	86.9	91.3	90.4	92.1
$E_2/E_3/E_1/E_4$	83.9	97.4	86.7	71.9	85.6	90.4	89.9	87.7
$E_2/E_3/E_4/E_1$	88.9	98.4	89.6	79.8	86.5	91.3	91.3	92.5
$E_2/E_1/E_4/E_3$	78.2	96.3	84.2	64.5	84.5	87.0	85.5	87.8
$E_1/E_3/E_4/E_2$	79.2	96.4	84.7	65.7	85.3	84.7	88.9	88.9
$E_3/E_4/E_2/E_1$	90.6	99.0	90.9	82.9	87.7	91.7	92.0	94.0
$E_3/E_4/E_1/E_2$	87.3	98.1	89.0	77.9	87.5	89.4	91.3	92.0
$E_1/E_3/E_2/E_4$	74.9	95.9	83.8	62.0	84.9	84.3	88.9	86.6
$E_1/E_2/E_4/E_3$	76.3	96.0	84.1	63.1	85.2	86.7	85.1	87.9
$E_4/E_1/E_2/E_3$	85.7	97.8	87.3	74.8	87.6	89.3	88.4	92.8
$E_1/E_2/E_3/E_4$	96.9	97.2	98.1	95.6	96.1	97.2	96.8	97.7
$E_4/E_2/E_1/E_3$	88.1	98.4	89.1	78.5	87.5	91.7	89.2	93.1
$E_4/E_2/E_3/E_1$	89.7	98.7	90.0	81.2	88.5	92.0	91.5	93.7
$E_4/E_1/E_3/E_2$	87.7	98.8	88.9	79.5	87.5	89.2	91.0	93.7
$E_3/E_2/E_4/E_1$	90.0	99.1	90.7	81.9	89.3	91.9	92.2	94.0
$E_3/E_1/E_2/E_4$	74.6	95.8	83.6	61.7	85.3	84.1	89.1	86.5

Table 18: Ablation on constraint weights for training diffusion models on QM9 [58]. The weight column follows the sequence $\lambda_c(e)$, $\lambda_b(e)$, $\lambda_a(e)$, and $\lambda_v(e)$. Values in parentheses give the corresponding loss weights for E_1, E_2, E_3, and E_4 (*e.g.,* $(4, 3, 2, 1)$ indicates $\lambda_c(e) = 4$, $\lambda_b(e) = 3$, $\lambda_a(e) = 2$, and $\lambda_v(e) = 1$).

Weight	Generation Quality				Constraint Satisfaction			
	Valid	Valid & Unique	Atom Stab	Mol Stab	E_1	E_2	E_3	E_4
(3, 3, 2, 2)	93.2	99.3	97.1	89.1	92.6	94.7	93.2	93.2
(4, 2.5, 2.5, 1)	92.4	99.1	96.6	89.3	94.6	92.8	94.4	89.9
(3, 2.5, 2, 1.5)	92.9	99.2	96.9	86.9	93.0	93.5	92.3	92.1
(5, 3, 1.5, 0.5)	91.9	98.6	96.1	84.7	96.4	95.2	89.2	86.2
(5, 2.5, 1.5, 1)	91.6	98.7	95.9	84.0	96.5	94.2	88.9	88.1
(4.5, 3.5, 1.5, 0.5)	91.2	98.5	95.8	82.5	95.9	96.1	88.4	86.0
(5, 3, 1, 1)	90.9	98.6	95.4	80.9	96.5	95.1	86.9	88.0
(3, 3.5, 2.5, 1)	86.7	97.8	93.4	74.9	88.9	96.3	95.5	88.1
(3, 3, 3, 1)	52.5	93.2	78.6	49.4	70.2	84.2	85.0	68.9
(3.5, 3, 2.5, 1)	82.2	97.5	91.6	70.4	91.4	95.2	94.6	86.8
(3, 2.5, 3, 1.5)	47.9	92.7	74.9	44.2	66.8	79.2	83.5	79.6
(4, 2.5, 2, 1.5)	88.9	98.2	94.4	76.6	93.4	92.8	90.9	93.8
(3.5, 2.5, 2, 2)	49.4	93.0	76.2	45.2	72.5	76.3	78.1	85.4
(4, 2, 2, 2)	44.8	92.6	72.5	40.8	74.2	70.4	76.4	84.8
(3.5, 3, 1.5, 2)	56.3	94.2	81.3	52.1	78.6	88.5	74.3	85.0
(4.5, 2.5, 2, 1)	94.2	99.4	97.2	90.0	96.1	93.4	93.7	93.6
(4.5, 2, 2.5, 1)	89.3	98.3	94.5	77.1	94.9	89.2	94.8	88.3
(4, 3.5, 1.5, 1)	93.0	99.1	96.7	88.0	94.6	96.4	89.0	89.7
(3.5, 3.5, 2, 1)	93.6	99.3	96.9	88.7	92.9	96.6	93.5	91.3
(3.5, 2, 2.5, 2)	46.2	92.5	73.9	42.0	71.7	70.7	80.1	86.4
(4, 2, 3, 1)	50.7	93.3	77.0	46.8	72.4	73.5	85.3	70.7
(4, 3, 2, 1)	96.9	97.2	98.1	95.6	96.1	97.2	96.8	97.7
(2.5, 3.5, 2.5, 1.5)	80.4	97.1	90.3	68.6	84.3	96.2	95.0	87.5
(2.5, 3, 3, 1.5)	53.6	93.8	79.7	50.3	71.3	85.2	86.1	79.3
(5, 2, 2, 1)	88.3	98.1	94.1	75.8	96.5	89.3	90.6	89.9
(4.5, 3, 1, 1.5)	88.7	98.2	94.4	78.1	95.6	95.5	85.2	90.4
(3, 4, 2, 1)	89.5	98.3	94.6	78.9	87.5	96.8	93.3	89.0
(2, 4, 2.5, 1.5)	73.5	96.0	88.3	61.9	79.1	96.5	94.4	85.6
(2.5, 2.5, 3, 2)	49.3	93.0	75.8	44.9	70.2	77.6	84.1	84.3
(1.5, 4, 3, 1.5)	45.2	92.4	73.0	41.1	64.1	91.3	84.4	76.8
(1, 4.5, 3, 1.5)	42.9	92.0	70.9	39.2	60.2	93.5	84.3	75.5

Table 19: MAE (↓) under different neuro-symbolic usage (training only vs. sampling only) for quantum-property-conditioned generation. The diffusion models are trained on QM9 [58].

Use During	C_v $\left(\frac{\text{cal}}{\text{mol}}\,\text{K}\right)$	μ (D)	α (Bohr^3)	$\Delta\varepsilon$ (meV)	$\varepsilon_{\text{HOMO}}$ (meV)	$\varepsilon_{\text{LUMO}}$ (meV)
Training	0.612	0.676	1.62	348	232	260
Sampling	0.658	0.724	1.76	382	248	278
Both	0.569	0.618	1.49	314	219	246

Table 20: Ablation on constraint weights during sampling on QM9 [58]. The weight column follows the sequence $\lambda_c(e)$, $\lambda_b(e)$, $\lambda_a(e)$, and $\lambda_v(e)$. Values in parentheses give the corresponding loss weights for E_1, E_2, E_3, and E_4 (*e.g.*, $(4, 3, 2, 1)$ indicates $\lambda_c(e) = 4$, $\lambda_b(e) = 3$, $\lambda_a(e) = 2$, and $\lambda_v(e) = 1$).

Weight	Generation Quality				Constraint Satisfaction			
	Valid	Valid & Unique	Atom Stab	Mol Stab	E_1	E_2	E_3	E_4
(4, 3, 2, 1)	93.8	97.6	97.9	89.4	96.2	95.5	91.6	85.8
(3, 3, 2, 2)	93.7	97.6	97.9	87.2	94.8	95.0	92.0	90.7
(4, 2.5, 2.5, 1)	93.6	97.6	97.8	88.1	96.4	93.4	94.2	85.4
(3, 2.5, 2, 1.5)	93.8	97.7	98.0	88.6	94.4	93.1	91.3	89.5
(5, 3, 1.5, 0.5)	92.9	97.5	97.5	87.5	97.3	95.2	84.4	82.4
(5, 2.5, 1.5, 1)	92.9	97.5	97.5	87.3	97.4	93.6	84.2	85.2
(4.5, 3.5, 1.5, 0.5)	92.7	97.5	97.4	86.2	97.1	96.4	83.8	81.9
(5, 3, 1, 1)	92.7	97.5	97.4	87.0	97.3	95.2	82.7	85.9
(3, 3.5, 2.5, 1)	93.1	97.6	97.7	86.8	90.1	96.6	95.4	84.1
(3, 3, 3, 1)	93.0	97.5	97.6	88.5	89.7	95.8	96.4	83.8
(3.5, 3, 2.5, 1)	93.3	97.6	97.7	87.9	93.8	95.1	94.9	84.3
(3, 2.5, 3, 1.5)	92.6	97.4	97.2	85.8	88.9	93.0	96.3	92.1
(4, 2.5, 2, 1.5)	93.1	97.5	97.6	85.7	95.7	92.8	90.9	92.8
(3.5, 2.5, 2, 2)	92.4	97.4	97.1	87.4	87.8	91.9	89.8	95.8
(4, 2, 2, 2)	92.1	97.3	96.9	86.1	95.1	84.2	90.2	96.0
(3.5, 3, 1.5, 2)	92.5	97.4	97.2	84.6	89.2	95.5	86.9	95.4

Table 21: Ablation on constraint enforcement schedules across sampling steps on QM9 [58].

Enforcement	Generation Quality				Constraint Satisfaction			
	Valid	Valid & Unique	Atom Stab	Mol Stab	E_1	E_2	E_3	E_4
Every 2 steps	95.5	96.7	97.9	90.3	94.5	95.8	94.6	93.8
Every 10 steps	93.7	96.6	96.1	84.1	91.8	93.7	92.4	91.3
Every 100 steps	90.1	98.5	94.3	82.0	83.3	87.5	85.6	84.8
First 50 steps	92.0	97.3	96.3	80.0	79.1	84.3	82.6	81.5
First 250 steps	93.1	97.6	97.4	83.2	87.6	91.2	89.5	88.4
First 500 steps	93.9	98.7	98.0	84.7	90.5	93.2	91.8	90.9
Last 50 steps	93.2	96.6	97.3	82.6	86.4	90.3	88.7	88.0
Last 250 steps	94.1	96.7	98.1	85.0	93.1	95.5	94.2	93.4
Last 500 steps	94.3	97.7	98.3	85.5	93.9	96.1	94.9	94.3

Table 22: Comparison of symbolic constraints usage (training only vs. sampling only) on two datasets.

Dataset	Use During	Generation Quality				Constraint Satisfaction			
		Valid	Valid & Unique	Atom Stab	Mol Stab	E_1	E_2	E_3	E_4
QM9 [58]	Training	94.5	95.9	96.7	91.9	94.1	95.4	93.5	95.0
	Sampling	91.5	96.5	93.9	88.3	90.9	92.8	91.2	92.9
	Both	96.9	97.2	98.1	95.6	96.1	97.2	96.8	97.7
GEOM-Drugs [2]	Training	95.9	97.5	86.5	75.8	87.8	91.7	90.1	91.6
	Sampling	94.9	95.1	85.1	72.5	86.9	90.8	88.3	91.7
	Both	97.2	98.8	87.9	79.4	92.6	95.8	93.6	96.2

Table 23: Ablation study of the weighting strategy on two datasets. Our adaptive weighting has two phases: Phase 1 (*i.e.,* linear ramping) linearly ramps up loss weights, and Phase 2 (*i.e.,* energy adaptive) increases weights based on the degree of constraint satisfaction (see Section 3 for details).

Dataset	Weighting	Generation Quality				Constraint Satisfaction			
		Valid	Valid & Unique	Atom Stab	Mol Stab	E_1	E_2	E_3	E_4
QM9 [58]	Neither (Non-adaptive)	88.5	96.8	89.2	84.7	75.3	81.1	78.5	85.1
	Linear Ramping Only	91.9	97.0	93.5	87.6	81.3	86.5	86.2	90.3
	Energy Adaptive Only	94.3	97.1	94.7	90.8	86.4	91.7	89.2	93.6
	Both	96.9	97.2	98.1	95.6	96.1	97.2	96.8	97.7
GEOM-Drugs [2]	Neither (Non-adaptive)	85.2	96.3	78.6	62.4	75.2	81.2	76.5	77.9
	Linear Ramping Only	88.7	96.4	81.3	66.2	78.5	86.8	80.6	83.7
	Energy Adaptive Only	91.5	96.7	85.3	72.6	82.3	92.6	88.3	91.2
	Both	97.2	98.8	87.9	79.4	92.6	95.8	93.6	96.2

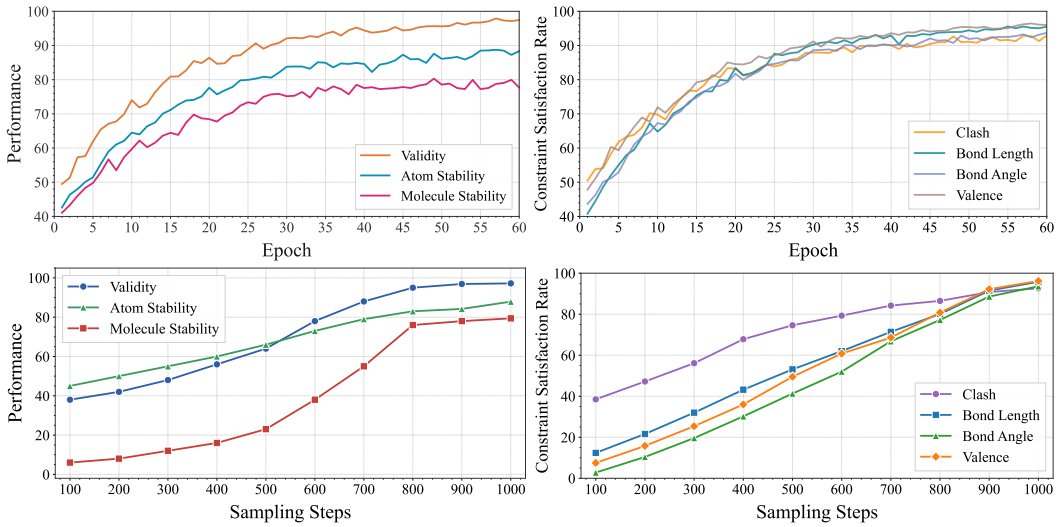


Figure 21: Ablation on sampling steps and training epochs on GEOM-Drugs [2]. We save diffusion model checkpoints every epoch. Increasing either sampling steps or training epochs consistently improves generation quality and constraint compliance.

In summary, our ablation studies include:

1. sequential versus non-sequential enforcement (*i.e.*, activating all constraints simultaneously from the start of training) of symbolic constraint subsets (Table 15 and Table 16);
2. training curricula and schedules (*e.g.*, constraint activation order in Table 17 and epoch schedules in Figure 21);
3. constraint weighting schemes (Table 18, Table 20, and Table 23);
4. applying neuro-symbolic constraints during training versus sampling only (Table 19 and Table 22);
5. sampling-time strategies (*e.g.*, when and how constraints are enforced during denoising in Table 21, and the denoising-step schedule in Figure 21).

Altogether, the extensive ablation studies validate that the proposed training strategies, including sequential constraint activation and adaptive weighting, effectively improve generation quality, constraint compliance, and physicochemical plausibility in diffusion-based 3D molecule generation. We also show that our framework supports diffusion model fine-tuning under the same symbolic constraints, with an ablation over fine-tuning epochs. Further details are provided in Section N.

J Generalizing Our Training Strategy to Multiple Diffusion Backbones

As shown in Table 24, the proposed training strategy yields consistently high generation quality and constraint satisfaction across all five backbones on both datasets, confirming its generalizability. On QM9 [58], all backbones exceed 95% validity and 92% on every constraint metric. On the more challenging Geom-Drugs [2], constraint satisfaction remains above 87% despite a drop in molecule stability, suggesting that the symbolic constraints provide a reliable quality floor regardless of molecular complexity. Among the backbones, UAE-3D [49] and SLDM [52] benefit most, indicating that stronger base architectures better leverage the introduced physicochemical priors.

Table 24: Generalization of our training strategy to different diffusion backbones on two datasets.

Dataset	Method	Generation Quality				Constraint Satisfaction			
		Valid	Valid & Unique	Atom Stab	Mol Stab	E_1	E_2	E_3	E_4
QM9 [58]	GCDM [50]	95.8	97.6	96.6	95.1	92.7	95.9	94.4	95.7
	UniGEM [20]	96.4	96.7	97.9	96.9	94.4	97.2	96.8	97.0
	UAE-3D [49]	97.1	96.8	98.3	98.8	95.7	98.1	97.5	97.9
	RADM [14]	96.1	96.3	97.1	95.7	93.5	96.1	95.1	96.2
	SLDM [52]	96.7	96.6	98.1	96.3	94.9	97.6	96.9	97.4
Geom-Drugs [2]	GCDM [50]	94.7	95.4	90.6	70.4	87.1	90.8	89.5	91.3
	UniGEM [20]	96.5	94.6	91.3	72.3	90.0	92.3	93.1	95.7
	UAE-3D [49]	97.3	95.8	92.2	78.0	93.9	93.7	94.4	96.8
	RADM [14]	95.2	95.4	89.9	74.3	88.5	93.3	92.6	92.5
	SLDM [52]	96.8	96.7	90.6	78.7	92.0	93.1	93.6	96.1

K Out-of-Distribution Generalizability of SensDiff

To assess out-of-distribution generalizability, we conduct cross-dataset evaluations by training on one dataset and evaluating on another (QM9 [58] $\leftrightarrow$ GEOM-Drugs [2]), where the two datasets differ significantly in molecular property and complexity. As shown in Tables 25 and 26, SensDiff achieves the best performance across all metrics in both transfer directions, obtaining 87.9% and 92.9% validity with 69.2% and 82.0% molecule stability, respectively, surpassing the strongest baselines by a substantial margin. The constraint compliance metrics (E_1 to E_4) exhibit a similarly dominant trend, confirming the structural soundness of generated molecules even under severe distribution shift. We attribute this strong transferability to the neuro-symbolic nature of our approach. Rather than merely memorizing data-specific patterns, the symbolic constraints encode fundamental physical and chemical principles, such as bond length rules, valence rules, and steric constraints, that are universally valid regardless of the molecular domain. By internalizing these scientific priors during training and enforcing them during sampling, SensDiff acquires an effective understanding of molecular geometry that naturally generalizes to unseen datasets, whereas purely data-driven baselines degrade more severely when the test distribution departs from the training one.

To evaluate the ability to generate rare (out-of-distribution) molecules, we select molecules with extreme LogP or QED values from the tail of the training distribution and task each model to generate molecules targeting these underrepresented property regimes. As shown in Table 27, SensDiff consistently achieves the best generation quality, reaching 84.2% validity, 88.1% atom stability, and 68.6% molecule stability, outperforming the strongest baseline, UAE-3D [49], by 2.6%, 2.9%, and 2.2%, respectively. The constraint compliance metrics (E_1 to E_4) further suggest this merit, with SensDiff surpassing all baselines by a clear margin. These improvements stem from the inherent difficulty of generating out-of-distribution molecules. As the target drifts toward low-density regions of the data distribution, purely data-driven methods lack sufficient training signal and tend to produce implausible structures. Our neuro-symbolic-guided sampling addresses this by enforcing explicit structural constraints during generation, helping preserve molecular validity even when the model extrapolates beyond well-represented property regions.

Table 25: Out-of-distribution performance comparison. We train the model on GEOM-Drugs [2] and evaluate on QM9 [58] to assess its cross-data generalizability.

Model	Generation Quality				Constraint Satisfaction			
	Valid	Valid & Unique	Atom Stab	Mol Stab	E_1	E_2	E_3	E_4
EDM [30]	58.4	98.6	72.2	41.5	57.6	64.8	62.1	67.5
EEGSDE [5]	61.9	98.8	74.6	44.9	60.2	67.4	64.3	69.8
GEOLDM [69]	66.5	98.9	78.2	50.7	64.7	71.9	68.9	74.3
EquiFM [62]	71.2	99.0	81.6	56.6	69.2	76.2	73.5	78.6
GeoBFN [63]	74.9	99.1	84.2	61.3	72.8	79.5	76.9	81.9
GCDM [50]	77.5	99.2	85.9	64.9	75.5	82.1	79.6	84.4
RADM [14]	79.7	99.2	87.3	67.2	77.3	84.3	81.7	86.1
SLDM [52]	82.7	99.3	89.5	71.8	81.3	87.9	85.6	89.7
UAE-3D [49]	84.2	99.3	90.3	73.6	82.9	89.2	87.1	90.9
UniGEM [20]	83.4	99.3	89.9	72.8	82.2	88.7	86.5	90.1
SensDiff (Ours)	**92.9**	**99.4**	**93.3**	**82.0**	**89.6**	**93.3**	**92.2**	**95.7**

Table 26: Out-of-distribution performance comparison. We train the model on QM9 [58] and evaluate on GEOM-Drugs [2] to assess its cross-data generalizability.

Model	Generation Quality				Constraint Satisfaction			
	Valid	Valid & Unique	Atom Stab	Mol Stab	E_1	E_2	E_3	E_4
EDM [30]	46.3	98.8	61.8	18.5	48.6	56.2	53.8	58.3
EEGSDE [5]	49.7	98.9	64.2	20.9	51.2	58.9	56.1	60.7
GEOLDM [69]	54.6	99.0	68.0	25.8	55.9	63.2	60.5	65.1
EquiFM [62]	59.9	99.1	70.8	31.5	60.7	68.0	65.2	69.8
GeoBFN [63]	64.1	99.2	71.3	36.9	65.0	72.2	69.4	73.9
GCDM [50]	67.5	99.2	73.9	41.3	68.2	75.4	72.7	76.9
RADM [14]	70.4	99.3	76.0	45.2	71.1	78.1	75.9	79.7
SLDM [52]	75.3	99.3	79.2	52.7	76.1	82.7	80.4	84.3
UAE-3D [49]	77.2	99.4	80.6	55.8	77.9	84.3	82.1	85.9
UniGEM [20]	76.2	99.3	80.0	54.4	77.1	83.6	81.3	85.1
SensDiff (Ours)	**87.9**	**99.4**	**85.5**	**69.2**	**89.6**	**90.2**	**91.1**	**94.5**

Table 27: Performance comparison on generating rare molecular samples with extreme (high) LogP and QED values from the tail of the training distribution.

Model	Descriptor		Generation Quality			Constraint Satisfaction			
	LogP	QED	Valid	Atom Stab	Mol Stab	E_1	E_2	E_3	E_4
GEOLDM [69]	3.86	0.708	64.1	69.2	43.8	65.2	74.2	71.8	77.4
GCDM [50]	3.94	0.731	68.4	73.6	51.2	69.4	79.2	77.1	82.6
RADM [14]	4.08	0.748	72.1	76.4	54.8	73.2	82.1	79.6	84.2
SLDM [52]	4.19	0.764	74.8	78.6	59.2	75.8	84.6	82.1	86.6
UniGEM [20]	4.22	0.771	75.6	79.4	60.4	76.4	85.2	82.8	87.4
UAE-3D [49]	4.48	0.792	81.6	85.2	66.4	82.2	90.4	88.8	92.6
SensDiff (Ours)	4.33	0.779	**84.2**	**88.1**	**68.6**	**84.8**	**93.6**	**91.2**	**96.4**

L Additional Evaluation Metrics for 3D Molecule Generation

In this section, we consider two additional evaluation metrics: Tanimoto similarity and constraint satisfaction rate (CSR). Tanimoto similarity measures structural similarity by comparing molecular fingerprints of generated samples to those from the prior distribution [4, 9]. The constraint satisfaction rate is the fraction of generated molecules that satisfy the prescribed rules.

As shown in Table 28, SensDiff achieves the highest Tanimoto similarity on both datasets, outperforming the strongest baseline by 10.7% and 13.7%, respectively, indicating that SensDiff generates molecules whose fingerprints more closely match the training distribution. The consistent gains across both datasets suggest that the constraints not only enforce structural validity but also steer generation toward more chemically realistic substructures. As shown in Figure 22, SensDiff consistently achieves the highest constraint satisfaction rate across all four symbolic constraints on both QM9 and Geom-Drugs, surpassing the strongest baseline in every category. Notably, the improvement is most pronounced on the clash and bond angle constraints, which are typically harder to satisfy, indicating that our training strategy particularly excels at capturing complex geometric dependencies.

Table 28: Tanimoto similarity between generated samples and the training distribution on two datasets.

QM9 [58]		GEOM-DRUGS [2]	
MODEL	TANIMOTO SIMILARITY (%, ↑)	MODEL	TANIMOTO SIMILARITY (%, ↑)
EDM [30]	56.2	EDM [30]	24.2
EEGSDE [5]	60.8	EEGSDE [5]	27.3
GEOLDM [69]	63.6	GEOLDM [69]	32.7
EquiFM [62]	65.8	EquiFM [62]	36.1
GeoBFN [63]	68.1	GeoBFN [63]	39.6
GCDM [50]	69.8	GCDM [50]	42.2
RADM [14]	71.2	RADM [14]	45.2
SLDM [52]	73.4	SLDM [52]	49.7
UAE-3D [49]	74.1	UAE-3D [49]	51.4
UniGEM [20]	73.7	UniGEM [20]	50.3
SensDiff (Ours)	**84.8**	SensDiff (Ours)	**65.1**

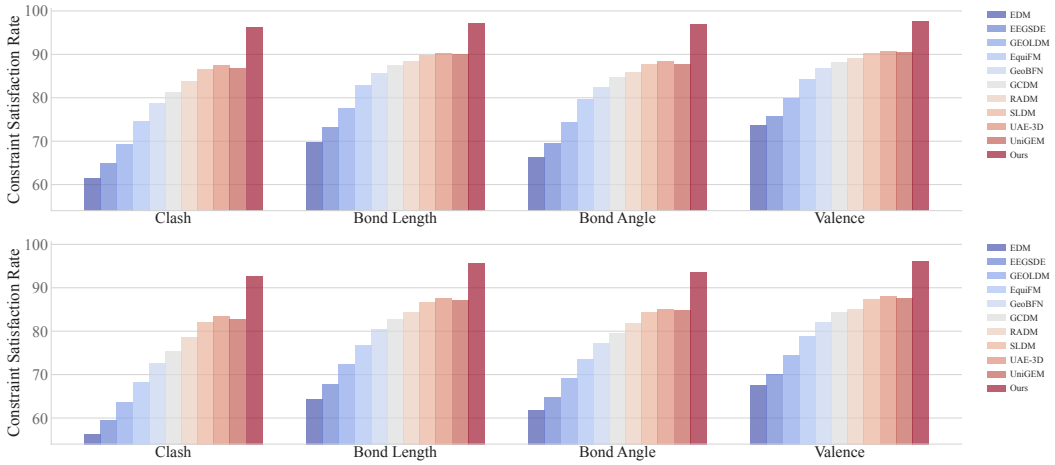


Figure 22: Constraint satisfaction rate on QM9 [58] (**upper**) and GEOM-Drugs [2] (**bottom**).

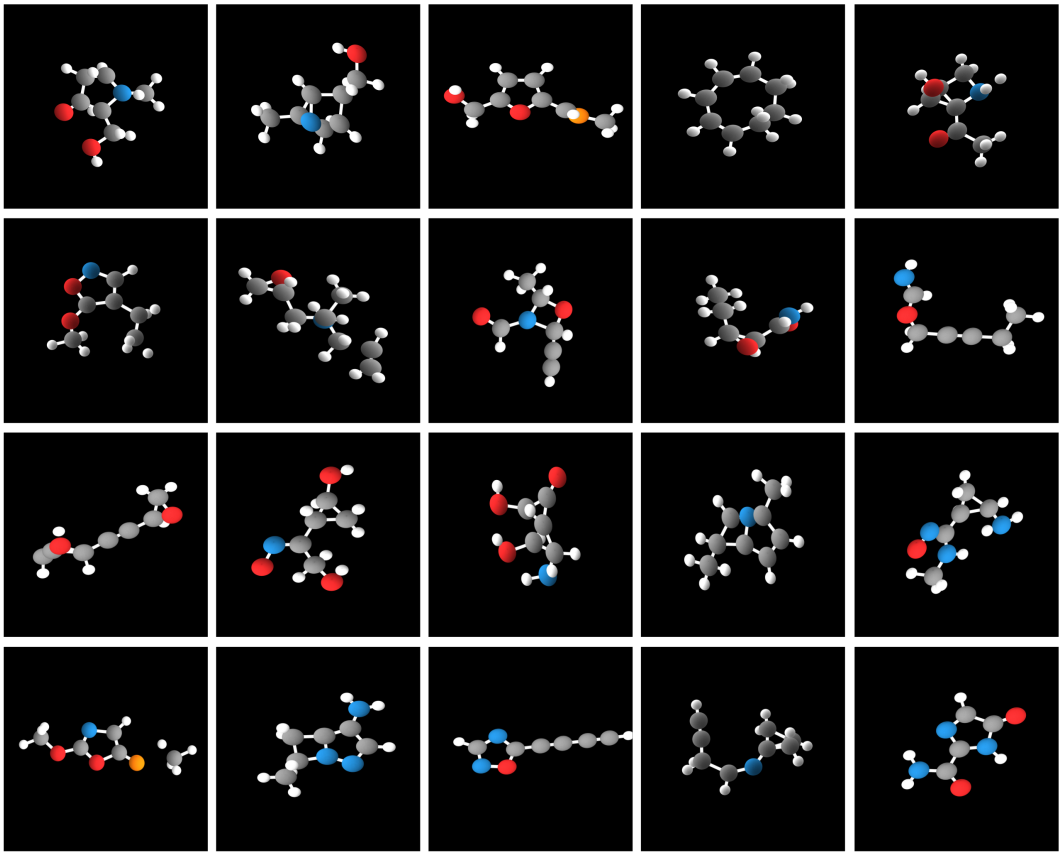


Figure 23: Randomly selected failure cases. In each row, the five molecules violate E_{clash}, $E_{\text{bond_len}}$, E_{angle}, and E_{val}, respectively.

M Analysis of Failure Cases

A key advantage of symbolic constrained molecule generation is that it makes model failures interpretable. In purely implicit molecule generators, failures typically appear as low validity or unrealistic 3D structures, but it is difficult to pinpoint which physicochemical regularities were not learned. In SensDiff, the symbolic energies E_{clash} (*i.e.*, steric non-overlap), $E_{\text{bond_len}}$ (*i.e.*, bond length), E_{angle} (*i.e.*, bond angle), and E_{val} (*i.e.*, valence validity) provide human-understandable diagnostics via Equation 5–8. This helps us view failures not merely as invalid molecules, but as evidence that the model has not yet internalized a specific physicochemical rules. Accordingly, we conduct a qualitative diagnostic study of representative failure cases from earlier versions of our model and weaker training strategies. We use these findings to inform the final constraint design, the overall objective in Equation 11, and the sequential curriculum in Equation 9. We inspected representative generated molecules together with their recovered local geometry and decoded topology, and grouped them by the violated symbolic rules. We reuse the same rule vocabulary that defines $\mathcal{L}_{\text{symbolic}}$ in Equation 10 as the analysis interface, forming a closed loop between interpretability and model design. This enables fine-grained diagnosis by distinguishing failures due to non-bonded repulsion, local bonded geometry, topology restoration, or interactions among constraints, rather than treating all invalid generations as a single error class. We visualize randomly selected failure cases for each of the four constraints in Figure 23.

The most consequential failure mode is the isolated violation of E_{clash}. In these samples, the generated structure exhibits clear steric collapse. Non-bonded atoms become too close, local clusters are compressed, and some coordinates even overlap. Specifically, this suggests that the model has not yet learned the basic excluded-volume regularity induced by non-bonded repulsion and van der

Table 29: Subset-wise evaluation on strained and atypical motifs in GEOM-Drugs [2]. Sanitization is the fraction of generated molecules that pass RDKit sanitization. Ring BL Err is the mean absolute bond-length error computed only on bonds inside the target motif. Ring BA Err is the mean absolute bond-angle error computed only on angles inside the target motif.

Motif	Metric	SensDiff	GEOLDM [69]	SLDM [52]	UAE-3D [49]
Cyclopropane	Sanitization (↑)	0.95	0.89	0.88	0.91
	Mol Stab (↑)	0.90	0.81	0.80	0.83
	Ring BL Err (↓)	0.024	0.038	0.042	0.035
	Ring BA Err (deg) (↓)	2.8	5.2	5.8	4.5
Cyclobutane	Sanitization (↑)	0.92	0.84	0.83	0.87
	Mol Stab (↑)	0.87	0.76	0.74	0.79
	Ring BL Err (↓)	0.028	0.045	0.050	0.041
	Ring BA Err (deg) (↓)	3.5	6.8	7.2	5.5
β-Lactam	Sanitization (↑)	0.90	0.80	0.78	0.83
	Mol Stab (↑)	0.82	0.71	0.68	0.74
	Ring BL Err (↓)	0.030	0.052	0.058	0.048
	Ring BA Err (deg) (↓)	4.2	7.5	8.0	6.2
Bridged ring	Sanitization (↑)	0.91	0.82	0.80	0.85
	Mol Stab (↑)	0.84	0.73	0.71	0.76
	Ring BL Err (↓)	0.025	0.042	0.048	0.038
	Ring BA Err (deg) (↓)	3.2	6.0	6.5	5.0

Waals separation [8]. Such failures typically arise in the early stage, high-noise portion of the reverse process, before the latent state enters a molecule-like geometric manifold. This failure is upstream, since once atomic placement is implausible, later constraints no longer operate on a chemically meaningful configuration. Based on this, our analysis indicates that E_{clash} should serve as the first stabilizer of the reverse process rather than one constraint among many, motivating why we activate E_{clash} first and assign it the largest weight in Equation 10.

This upstream role becomes clearer in coupled failure modes involving E_{clash}. When E_{clash} and $E_{\text{bond_len}}$ fail together, the typical issue is not that the bond-length prior is inherently weak. Rather, steric collapse has already corrupted the local configuration, pushing bonded distances outside their empirical regime. Therefore, violations of $E_{\text{bond_len}}$ are often secondary to violations of E_{clash}. A similar dependency appears when E_{clash} and E_{angle} fail together, where abnormal bond angles usually reflect unstable atomic placement rather than a purely angular defect. This is chemically consistent with the fact that realistic bond lengths and angular preferences are well defined only after atoms occupy a physically plausible local neighborhood [11, 3]. The same logic applies to co-failures of E_{clash} and E_{val}: once the geometry collapses, bond inference becomes unreliable and valence assignments naturally break. In summary, many seemingly distinct failures can be traced back to the same upstream geometric defect, motivating our curriculum that establishes spatial non-overlap before introducing higher-order physicochemical rules.

Isolated violations of E_{angle}, as well as joint violations of $E_{\text{bond_len}}$ and E_{angle}, point to a different issue. Bond angles capture chemically meaningful coordination preferences, but they are intrinsically weaker and more ambiguous than steric exclusion or bond length. An angle only constrains the directional relation among three atoms, does not uniquely determine a 3D structure, and the same angle can correspond to multiple 3D geometries, which is especially problematic in noisy diffusion states. Moreover, E_{angle} relies on the bond posterior to decide which atom triplets are relevant, so uncertainty in inferred connectivity can propagate into the angular term itself. As a result, its gradients become informative only after local connectivity is sufficiently reliable. These failure cases showcase that activating angular rules too early or too strongly yields weak guidance or conflicts with more fundamental geometric objectives. This explains why we place E_{angle} after $E_{\text{bond_len}}$, and keep its weight below those of steric exclusion and bond length in Equation 10. Physicochemically, angle realism is a vital refinement signal, but it is not an effective first-order guidance for generation.

Table 30: Generation-frequency fidelity for strained and atypical motifs on GEOM-Drugs [2]. GEOM-Drugs Reference is the empirical frequency of each motif in the dataset. The model columns report the corresponding frequency among generated molecules. Valid is the overall validity of generated molecules under the standard validity metric used in the main text.

Motif	GEOM-Drugs Reference [2]	SensDiff	GEOLDM [69]	SLDM [52]	UAE-3D [49]
Cyclopropane	4.8%	4.5%	3.8%	4.1%	4.3%
Cyclobutane	2.5%	2.3%	1.9%	2.1%	2.2%
β-Lactam	0.75%	0.68%	0.42%	0.55%	0.60%
Bridged ring	3.4%	3.1%	2.5%	2.8%	2.9%
Valid	–	93%	84%	83%	87%

Table 31: Inspection of motif-specific geometric statistics in the symbolic rule tables. Here μ and σ denote the mean and standard deviation of the corresponding bond-length or bond-angle statistic used by the rule table. The comparison shows that cyclopropane is assigned a distinct geometric prior instead of being matched to a generic acyclic alkane.

Feature	Cyclopropane	Acyclic alkane	Difference
Mean of C–C bond length (μ)	1.510 Å	1.540 Å	−0.030 Å
Standard deviation of C–C bond length (σ)	0.035 Å	0.028 Å	–
Mean of C–C–C bond angle (μ)	60.0°	109.5°	−49.5°
Standard deviation of C–C–C bond angle (σ)	3.2°	4.5°	–

Isolated violations of E_{val} are also informative. In these failure cases, the continuous geometry can appear partially plausible, but the molecule becomes chemically invalid after mapping back to a discrete graph. Some atoms exhibit invalid valence, recovered bond orders are inconsistent, or the resulting topology is illegal. This pattern suggests a failure of geometry–topology alignment rather than pure geometric collapse. Therefore, the model has not yet learned a stable correspondence between local structure and admissible bonding patterns, which is precisely the regime targeted by valence-based reasoning and bond–valence consistency. Crucially, this does not suggest enforcing valence first, since valence validity is downstream of geometry and bond inference. Nevertheless, it does indicate that legality cannot be postponed to a purely terminal repair step. Once the reverse process drifts too far from a topologically admissible regime, post-hoc restoration becomes infeasible. This is why we place E_{val} last in the curriculum, yet still incorporate it into both training via Equation 10 and Equation 11, rather than treating it as an external validator.

Moreover, the most revealing compound failure is the simultaneous violation of E_{clash}, E_{angle}, and E_{val}. We observed this pattern most often in harder settings, especially when the model was required to satisfy multiple structural objectives at the same. These samples suggest that the issue is not merely weak constraints but cross-constraint interference. Multiple rules activate before the structural hierarchy stabilizes, pulling the sample in competing directions and preventing effective satisfaction of any single rule. This was a key negative result in our preliminary study. It demonstrates that enabling all constraints from the start, or increasing all weights in lockstep, does not improve generation even when each constraint is chemically motivated. The challenge is not only which rules to impose, but when they become physicochemically meaningful during denoising. This observation directly supports the adaptive stage-wise weighting schedule in Equation 9.

Taken together, all these failure cases reveal a clear hierarchy of physicochemical dependencies. E_{clash} governs whether 3D molecule generation first enters a physically admissible spatial regime. $E_{\text{bond_len}}$ and E_{angle} then refine local bonded geometry once that regime has stabilized. E_{val} finally enforces consistency between continuous structure and chemically legal topology. This is the core logic implemented in our sequential training curriculum. Motivated by this, the order $E_{\text{clash}} \rightarrow E_{\text{bond_len}} \rightarrow E_{\text{angle}} \rightarrow E_{\text{val}}$ should be viewed not as a heuristic training trick, but as the outcome of the diagnostic loop. The same idea also holds for the descending weight pattern in Equation 10. In particular, the most upstream and physically indispensable rule is emphasized most strongly, while more indirect and uncertainty-sensitive rules are introduced later and more gently. This analysis also clarifies why SensDiff leverages exactly these four constraints. Across failure cases,

we repeatedly observed four missing pieces of internalized scientific structure: excluded-volume control, bond-length consistency, bond-angle consistency, and valence validity. Each deficit maps naturally to a differentiable symbolic energy that can be incorporated into $\mathcal{L}_{\text{symbolic}}$ and ultimately $\mathcal{L}_{\text{total}}$ in Equation 11. Meanwhile, this study indicates that progress does not come from adding constraints indiscriminately. Overly aggressive or poorly sequential constraint enforcement often increases interference rather than validity. The resulting design principle is thus not more rules, but the right rules, activated when their underlying structural signal becomes reliable.

Empirically, a natural concern is whether symbolic constraints may unintentionally penalize chemically valid but geometrically unusual motifs, especially strained rings and bridged structures whose bond lengths and bond angles differ from standard acyclic patterns. To examine this issue, we perform a failure mode analysis on GEOM-Drugs [2] using three studies. First, we conduct a subset-wise evaluation on molecules containing cyclopropane, cyclobutane, β-lactam, and bridged-ring motifs in Table 29. We report Sanitization, the fraction of generated molecules that pass RDKit sanitization, Mol Stab, the fraction that satisfies the molecular stability criterion used in the main text, Ring BL Err, the mean absolute bond-length error on bonds inside the target ring motif, and Ring BA Err, the mean absolute bond-angle error on angles inside the target ring motif in degrees. Second, we evaluate generation-frequency fidelity in Table 30, where the GEOM-Drugs Reference column gives the empirical frequency of each motif in the dataset and the model columns report the corresponding frequency among generated molecules. We also report Valid, the overall molecular validity under the same metric used throughout the paper. Third, we inspect the statistical rule tables in Table 31, where μ and σ denote the mean and standard deviation of the corresponding bond-length or bond-angle statistic, to verify that unusual motifs are assigned motif-specific geometric priors rather than generic acyclic ones. The results show that SensDiff does not suppress unusual but chemically valid motifs. In Table 29, SensDiff achieves higher sanitization and molecular stability than GEOLDM, SLDM, and UAE-3D across all four motif categories, while also obtaining lower ring-specific bond-length and bond-angle errors. Moreover, Table 30 shows that the generation frequency of these motifs remains close to the GEOM-Drugs reference distribution. As shown in Table 31, this behavior is consistent with the context-dependent symbolic rules, which assign strained motifs distinct geometric priors instead of generic acyclic ones. Overall, the results indicate that the symbolic rules improve common geometric and valence-related failure modes while remaining compatible with strained and atypical molecular motifs.

In summary, this preliminary study highlights a concrete methodological advantage of neuro-symbolic generative modeling. The same symbolic formulation that improves the model also enables human understandable debugging. Rather than reporting only lower validity, we can diagnose whether a model fails to suppress early steric collapse, applies angular supervision before local bonding is reliable, or defers chemical legality until geometry and topology have already drifted apart. Such fine-grained diagnosis is difficult to obtain from purely implicit generative models. In this sense, the failure analysis is not merely an empirical add-on, but evidence that rule-level interpretability can directly drive model improvements. Based on this, the final SensDiff formulation in Equation 10 and Equation 11, and Equation 9 can be viewed as the outcome of this analytical process.

N Extending the Framework to Diffusion Model Fine-Tuning

In this section, we further extend our training strategy to fine-tune the model using the same symbolic constraints. Figure 24 and Figure 25 report performance across fine-tuning epochs on QM9 [58] and GEOM-Drugs [2]. On both datasets, all metrics improve steadily with fine-tuning, with the largest gains appearing in the early epochs (*e.g.,* on GEOM-Drugs, validity increases from 82.5% to 89.6% within only 8 epochs). Notably, constraint compliance improves in tandem with generation quality, demonstrating that symbolic constrained fine-tuning improves structural validity and physicochemical plausibility simultaneously rather than trading one for the other. The rapid convergence further highlights the efficiency of our approach. Only a few fine-tuning epochs are needed to internalize symbolic constraints into a pretrained diffusion model, making it practical for large-scale 3D molecule generation. We attribute this efficiency to the dense supervision provided by symbolic constraints, which complements the data-driven objective by supplying explicit gradient feedback at each denoising step. Moreover, this improvement suggests that these constraints are aligned

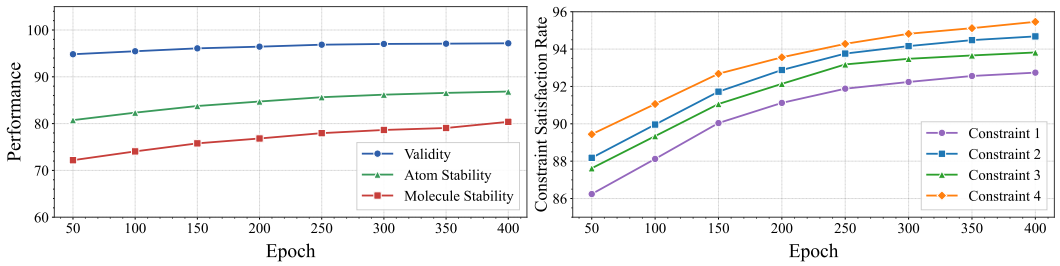


Figure 24: Fine-tuning the diffusion model with symbolic constraints on QM9 [58]. (**Left**) Generation quality over fine-tuning epochs. (**Right**) Constraint satisfaction rate over fine-tuning epochs.

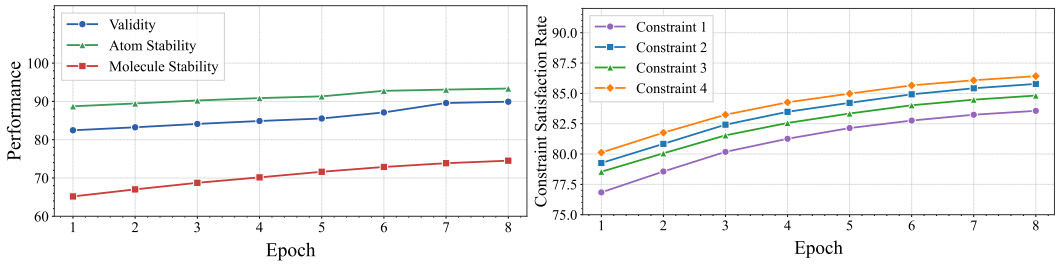


Figure 25: Fine-tuning the diffusion model with symbolic constraints on GEOM-Drugs [2]. (**Left**) Generation quality over fine-tuning epochs. (**Right**) Constraint satisfaction rate over fine-tuning epochs.

with the data distribution, acting as a regularizer instead of an adversarial objective. Essentially, neuro-symbolic fine-tuning distills scientific priors into the model parameters, yielding lasting gains with more efficient computation than large-scale pretraining.

O Additional Studies on Training Schemes

In this section, we conduct a systematic experimental study of training schemes for improving diffusion models with symbolic supervision. Specifically, we compare epoch-wise and timestep-wise supervision strategies, analyze the reliability of intermediate denoising states across timesteps, and perform dataset-specific ablations on curriculum activation epochs and ramp length to evaluate their effects on generation quality, validity, stability, and constraint satisfaction.

Epoch-wise versus timestep-wise supervision. An important question is whether symbolic supervision should be scheduled only across epochs or also across diffusion timesteps. Since reverse states at high noise levels are less reliable, timestep-aware supervision might weaken noisy symbolic signals early in denoising and strengthen them later when molecular structure is more formed. We therefore study whether explicit timestep-aware scheduling brings additional benefits beyond the epoch-wise curriculum used in SensDiff. Specifically, we first analyze the reliability of intermediate denoising states across timesteps in Table 32, using bond top-1 confidence, bond entropy, per-constraint activation strength, and the fraction of samples with near-zero effective symbolic weight. We then compare four supervision strategies on QM9 in Table 33: the default epoch-only curriculum, a timestep-only strategy, and two joint epoch+timestep strategies with either a shared or a per-constraint timestep threshold. In all cases, only the supervision schedule is changed. The results show that the current formulation already has clear implicit timestep adaptivity: early noisy states receive near-zero effective symbolic weights, while later denoising steps activate the constraints much more strongly. As a result, explicit timestep-aware scheduling brings only limited changes and does not improve the overall stability–quality trade-off over the default epoch-wise curriculum.

Table 32: Timestep-wise reliability of intermediate denoising states on QM9 [58]. As denoising proceeds, bond confidence increases, bond entropy decreases, and all symbolic constraints become progressively more active. The last column reports the fraction of samples whose effective symbolic weight is approximately zero, showing that very noisy steps are already naturally suppressed under our formulation.

t	Bond Top-1 Confidence (↑)	Bond Entropy (↓)	E_1	E_2	E_3	E_4	Fraction of $\lambda \approx 0$ (%, ↓)
900	0.32	3.62	0.25	0.09	0.04	0.02	0.92
700	0.52	2.68	0.52	0.30	0.19	0.10	0.72
500	0.73	1.88	0.77	0.58	0.43	0.28	0.42
300	0.87	1.28	0.92	0.82	0.68	0.53	0.16
100	0.95	0.85	0.98	0.96	0.91	0.82	0.02
50	0.97	0.72	0.99	0.98	0.95	0.90	0.01

Table 33: Ablation of supervision schedules on QM9 [58]. V_0 is the default epoch-only curriculum used in SensDiff. V_1 applies timestep-only weighting. V_2 and V_3 combine epoch-wise and timestep-wise scheduling, using either a shared threshold or per-constraint thresholds. The epoch-only curriculum remains the most effective and stable overall, while explicit timestep-aware strategies provide at most marginal gains over the default formulation.

Variant	Valid	Atom Stab	Mol Stab	V&U	E_1	E_2	E_3	E_4	Loss fluctuation (↓)	Convergence epoch (↓)
V_0 (SensDiff)	96.9	98.1	95.6	97.2	0.965	0.958	0.949	0.957	±12%	2400
V_1 (timestep-only)	93.2	94.5	91.0	94.1	0.946	0.938	0.928	0.935	±32%	2800
V_2 (joint shared $\tau = T/2$)	96.4	97.6	95.2	96.8	0.961	0.954	0.945	0.953	±16%	2500
V_3 (joint per-constraint τ)	96.7	97.9	95.4	97.0	0.963	0.957	0.947	0.956	±14%	2450

Hyperparameter selection. Another practical question is how to choose the activation epochs for the staged symbolic curriculum, as well as the ramp length R_{ramp}, which denotes the number of training epochs over which the weight of a newly introduced symbolic constraint is linearly increased from zero to its target value after that stage is activated. This design is important because activating constraints too early or too abruptly may inject unstable supervision when the denoising model is still inaccurate, whereas activating them too late may underuse symbolic guidance and limit the final improvement. Moreover, the appropriate schedule is expected to depend on the dataset, since the reliability of intermediate molecular structures and the convergence speed of training differ between QM9 [58] and GEOM-Drugs [2]. We therefore perform dataset-specific ablations to understand how the curriculum hyperparameters affect generation quality, validity, stability, and symbolic satisfaction. Specifically, we first vary the activation epochs of the three curriculum stages on QM9 [58] and GEOM-Drugs [2] in Tables 34 and 36. This experiment isolates when each group of symbolic constraints should be introduced during training. We then study the joint effect of activation epochs and R_{ramp} in Tables 35 and 37. This allows us to evaluate not only when new constraints should be activated, but also how smoothly their influence should increase afterward. In all cases, the model architecture, optimizer, and symbolic energies are kept fixed, and only the curriculum hyperparameters are varied. The results indicate that the selected schedules provide a favorable balance between training stability and generation quality. They also show that the optimal activation epochs are dataset-dependent rather than universal.

P Limitation and Future Work

While SensDiff performs strongly across benchmarks, several limitations remain. The symbolic constraints in this work are manually specified from established physicochemical knowledge, and automatically learning such rules from data or scientific literature is an important direction for future work. Moreover, extending the framework to more diverse molecular distributions and larger chemical spaces would further test its robustness and scalability. In addition, although the sequential activation curriculum is chemically intuitive, its optimal ordering and schedule may vary across domains, calling for more adaptive curriculum design. We hope SensDiff provides a useful foundation for integrating domain knowledge into generative modeling.

Table 34: Ablation of stage activation epochs on QM9 [58].

Stage 1	Stage 2	Stage 3	Valid	V&U	Atom Stab	Mol Stab	E_1	E_2	E_3	E_4
400	1120	1760	91.5	92.8	92.7	89.2	93.6	92.9	89.2	92.8
400	1680	2640	92.4	94.7	94.6	91.1	94.1	93.4	91.5	90.3
600	1120	2640	92.2	93.5	91.4	90.9	95.0	91.3	92.4	94.2
600	1680	1760	93.6	93.9	93.9	92.3	96.2	92.5	91.6	93.4

Table 35: Joint ablation of stage activation epochs and ramp length R_{ramp} on QM9 [58].

Stage 1	Stage 2	Stage 3	R_{ramp}	Valid	V&U	Atom Stab	Mol Stab	E_1	E_2	E_3	E_4
375	1050	1650	50	90.7	88.0	91.9	91.4	91.6	91.9	90.2	90.8
375	1050	2750	200	89.9	87.2	92.1	90.6	91.2	92.5	89.6	90.4
375	1750	1650	100	92.6	89.9	93.8	93.3	92.6	91.9	92.0	90.8
375	1750	2750	50	93.8	89.1	93.0	92.5	92.2	91.5	93.6	91.1
625	1050	1650	200	88.9	87.2	93.1	89.6	92.8	88.9	89.6	90.4
625	1750	1650	100	92.1	90.5	95.4	91.8	93.9	94.2	91.8	92.1
625	1750	2750	100	90.1	90.4	92.3	94.8	92.3	92.4	93.0	92.6
625	1750	2750	50	91.3	89.6	93.5	93.1	94.5	91.8	87.9	91.7

Table 36: Ablation of stage activation epochs on GEOM-Drugs [2].

Stage 1	Stage 2	Stage 3	Valid	V&U	Atom Stab	Mol Stab	E_1	E_2	E_3	E_4
8	22	34	90.2	89.8	79.9	71.4	89.6	86.3	83.6	85.5
8	32	50	89.7	91.3	80.4	72.9	90.8	89.2	89.8	88.4
12	22	50	92.3	91.0	82.0	72.6	89.1	85.2	86.3	86.6
12	32	34	93.1	91.7	80.8	73.2	92.3	90.3	86.2	87.7

Table 37: Joint ablation of stage activation epochs and ramp length R_{ramp} on GEOM-Drugs [2].

Stage 1	Stage 2	Stage 3	R_{ramp}	Valid	V&U	Atom Stab	Mol Stab	E_1	E_2	E_3	E_4
8	20	32	5	85.9	84.5	78.6	68.1	83.7	87.8	80.0	84.3
8	20	53	20	84.6	83.2	77.2	65.7	85.4	82.7	81.1	82.0
8	34	32	10	89.1	87.6	80.7	67.3	87.7	84.1	83.0	84.2
8	34	53	5	86.7	86.3	75.4	69.9	84.9	83.5	81.3	83.6
13	20	32	20	84.6	83.1	76.3	64.7	88.9	82.0	83.6	85.7
13	20	53	10	90.0	88.6	81.6	70.2	86.7	84.6	85.1	85.7
13	34	32	10	91.0	88.5	80.7	69.0	86.0	84.3	84.5	85.1
13	34	53	5	87.6	87.2	79.3	67.9	85.6	84.1	83.5	84.4

Q Broader Impacts

Integrating symbolic domain knowledge into diffusion models can improve the scientific plausibility and structural consistency of generated molecules, which may be valuable for early-stage drug discovery and materials design. Because SensDiff operates at sampling time without retraining, it also lowers the barrier to incorporating updated chemical knowledge into existing pipelines and makes the framework easier to adapt across evolving application settings. However, depending on how rules and thresholds are specified, the constraints may bias generation toward over-regularized, undesirable, or potentially toxic molecules. We hope this work encourages further exploration of neuro-symbolic methods for scientific discovery.