TOPVAE: REDUCING DARK AREAS IN 3D MOLECULAR LATENT DIFFUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Latent diffusion is a scalable framework for 3D molecular generation, but it requires a latent space that stays valid beyond the posterior samples. A molecular VAE decoder can reconstruct every posterior latent exactly and still return disconnected or chemically invalid graphs a short distance away from it, which is where a diffusion prior draws its samples. We call these regions **dark areas**. They are widest where conditioning or size extrapolation builds a latent code no training molecule supplied. Therefore, we propose **TopVAE**, a topology-optimized VAE whose decoder emits adjacency before atom and bond types and is trained with connectivity and chemical constraints inside the decoding path, so that the region of latents that decode to valid molecules extends well past the posterior. TopVAE matches the strongest baseline at the posterior and keeps returning valid, connected molecules far outside it, where that baseline has collapsed. It continues to return sanitizable products when a scaffold is grown well past its training maximum in atom count. Paired with a standard diffusion transformer it is competitive on de novo generation on QM9 and GEOM-Drugs, with the lowest GEOM-Drugs FCD_{3D} of the methods scored through one pipeline.

1 INTRODUCTION

Latent diffusion models, originally developed for image generation (Rombach et al., 2022; Peebles & Xie, 2023), have also been extended to 3D molecular generation (Xu et al., 2023; Chen et al., 2025). This setting is more challenging because valid molecules must simultaneously satisfy structural, geometric, and chemical constraints (You et al., 2024; Luo et al., 2025). A latent-variable formulation is therefore attractive: if the decoder honors these constraints, diffusion need only model the latent codes.

Molecular latent spaces are typically learned with reconstruction objectives plus regularization, as in VAE or VQ-VAE frameworks (Luo et al., 2025). However, strong reconstruction does not guarantee a latent space suitable for diffusion. Recent work shows that reconstruction quality can poorly predict, and sometimes even conflict with, downstream generation performance (Yao et al., 2025; Skorokhodov et al., 2025; Xu et al., 2026), because a diffusion prior draws latents far from any posterior sample, and the decoder must still return valid outputs there. Latent regularization is meant to enlarge the region of valid latents, but molecules make this difficult: decoding demands strict structural and chemical precision. Even small perturbations around posterior latents can therefore return disconnected structures or graphs that fail chemical sanitization (Fig. 1). We call these failure regions **dark areas**: latent neighborhoods where the decoder breaks.

Dark areas arise mainly from two sources: *topological disconnection*, where the decoded graph fractures into isolated components, and *chemical invalidity*, where predicted bonds violate valence or compatibility rules. Prior work often addresses these issues by enforcing chemical rules during generation (Jin et al., 2018; Liu et al., 2018; Ma et al., 2018; Krenn et al., 2022). Such constraints repair outputs without removing the underlying latent-space fragility. Instead, TopVAE applies these constraints during VAE training, with chemical constraint optimization (ChemCO) in the loop, so that valid molecules occupy a robust latent neighborhood. To this end we propose TopVAE (Topology-OPtimized VAE), which promotes connectivity through BFS-based adjacency refinement, corrects valence and bond-type violations by mirror ascent with adaptive penalty weights during training, and selectively injects ChemCO's corrections into the decoder.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

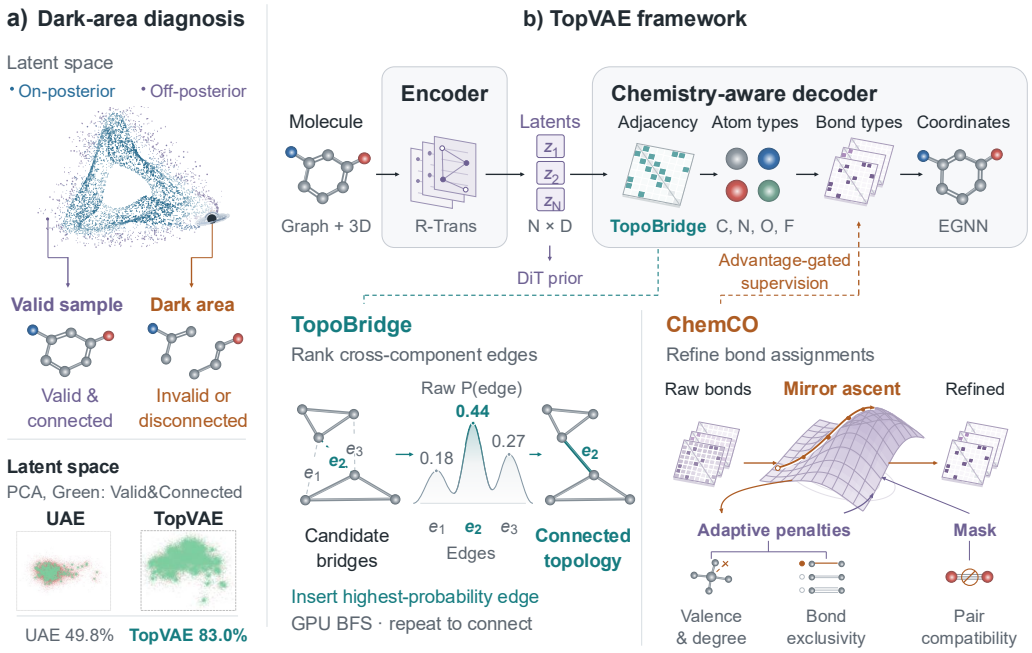


Figure 1: **Dark areas, and the methods to reduce them.** **(a)** Off the posterior a reconstruction-trained decoder returns invalid or disconnected molecules. Pooled over the nonzero σ levels of Appendix D.1, our method gives a latent space with 83.0% valid and connected, against 49.8% for UAE. Each point is one perturbed latent, green where its decode is valid and connected. **(b)** TopVAE decodes topology first, TopoBridge connecting the adjacency and ChemCO refining bond types under valence, degree and atom-pair constraints.

In summary, our contributions are:

1. **Dark areas in molecular latent space.** We identify and formalize dark areas: latent regions reachable by diffusion sampling but decoded as chemically invalid or disconnected molecules.
2. **TopVAE**, a topology-optimized VAE that decodes topology first under chemical and connectivity supervision: a) TopoBridge refines adjacency into a connected candidate topology, b) ChemCO corrects valence and bond-type violations by penalized mirror ascent, c) Advantage-Gated Constraint Learning (AGCL) feeds those corrections into decoder training only when they help.
3. **Generation and out-of-distribution expansion.** Conditioning and size extrapolation build latent codes no training molecule supplied, and that is where the two decoders separate most (Sec. 5.3): grown by 100 heavy atoms on GEOM-Drugs, to products of 106 to 110 heavy atoms against a training maximum of 90, TopVAE decodes 95.2% of its products to sanitizable molecules. Paired with a standard DiT it has the lowest GEOM-Drugs $\text{FCD}_{3\text{D}}$ of the three methods scored through one pipeline (Sec. 5.4).

2 Related Work

3D molecular generation. *Data-space* models generate atom types and coordinates, inferring bonds post hoc (Hoogeboom et al., 2022), while graph-aware variants model topology and geometry jointly (Peng et al., 2023; Huang et al., 2023; Vignac et al., 2023; Le et al., 2023; Reidenbach et al., 2026; Xu et al., 2024). *Latent* models compress molecules and fit a diffusion model to the codes (Xu et al., 2023; You et al., 2024; Luo et al., 2025; Joshi et al., 2025), gaining scalability but requiring high-quality latents, a requirement our dark-area diagnostic makes testable. Within that family, LDM-3DG (You et al., 2024) carries a molecule as a 2D-graph latent and a separate 3D latent and reassembles the graph from a fragment vocabulary, while TopVAE emits topology and coordinates from one atom-level latent sequence in a single decode.

Constraint-aware molecular decoding. Chemical and structural validity has been enforced via grammar-level constraints (Kusner et al., 2017; Krenn et al., 2022), structured graph decoders (Jin et al., 2018; Liu et al., 2018), validity-oriented regularization (Ma et al., 2018), and differentiable constraint-satisfaction layers (Wang et al., 2023; Zeng et al., 2024). TopVAE instead uses ChemCO's corrections as a training signal: the continuous solver supervises the decoder during training and runs no iterations at sampling time (Sec. 4.5, Appendix B).

3 PREREQUISITES

Dark areas in molecular latent space. At inference time, a latent diffusion model samples latents that lie between or outside the posterior codes seen during training. We call a decoded graph *chemically valid* if it (i) passes RDKit sanitization (legal valences, consistent aromaticity, and recognized bond types) and (ii) forms a single connected component. Denoting the sampling-reachable region by $\mathcal{A}_{\mathrm{LDM}}$ and the valid molecule set by $\mathcal{M}_{\mathrm{valid}}$, we define the **dark areas** as

$$\mathcal{D}_{\mathrm{dark}} = \{\boldsymbol{z} \in \mathcal{A}_{\mathrm{LDM}} : D_\theta(\boldsymbol{z}) \notin \mathcal{M}_{\mathrm{valid}}\}. \tag{1}$$

4 TOPVAE: A CHEMISTRY AWARE DECODER FOR OFF-POSTERIOR LATENTS

TopVAE targets dark areas in molecular latent space. It reduces them with a constraint-aware decoder. The decoder emits topology first, and two constraint modules supervise it during training.

4.1 ENCODER

The encoder maps the molecular graph into latent features with a Relational Transformer (Diao & Loynd, 2022). We first embed atoms and atom pairs:

$$\boldsymbol{h}_i^{(0)} = f_{\mathrm{enc-node}}(T_i, \boldsymbol{R}_i), \qquad \boldsymbol{p}_{ij} = f_{\mathrm{enc-edge}}(B_{ij}, \|\boldsymbol{R}_i - \boldsymbol{R}_j\|_2^2). \tag{2}$$

The Relational Transformer updates atom features by edge-aware attention, where pair features $\boldsymbol{p}_{ij}$ are injected into the attention between atoms i and j:

$$\boldsymbol{H} = \mathrm{RTrans}_\phi\left(\{\boldsymbol{h}_i^{(0)}\}_{i=1}^N,\ \{\boldsymbol{p}_{ij}\}_{i,j=1}^N\right). \tag{3}$$

The resulting atom-level features $\boldsymbol{H} = (\boldsymbol{h}_1, \ldots, \boldsymbol{h}_N)$ are used to parameterize the latent distribution in Eq. (15) of Appendix A.

4.2 TOPOLOGY-AWARE MULTI-STAGE DECODER

Using the notation of Appendix A, TopVAE decodes each latent code $\boldsymbol{z}$ in a topology-first order:

$$p_\theta(\boldsymbol{A}, \boldsymbol{T}, \boldsymbol{B}, \boldsymbol{R} \mid \boldsymbol{z}) = p_\theta(\boldsymbol{A} \mid \boldsymbol{z})\, p_\theta(\boldsymbol{T} \mid \boldsymbol{A}, \boldsymbol{z})\, p_\theta(\boldsymbol{B} \mid \boldsymbol{A}, \boldsymbol{T}, \boldsymbol{z})\, p_\theta(\boldsymbol{R} \mid \boldsymbol{A}, \boldsymbol{T}, \boldsymbol{B}, \boldsymbol{z}). \tag{4}$$

Here $\boldsymbol{A}$ is the adjacency, $\boldsymbol{T}$ the atom types, $\boldsymbol{B}$ the bond types and $\boldsymbol{R}$ the coordinates. This factorization makes adjacency the first decoded object. In particular, bond prediction is constrained by

$$A_{ij} = 0 \Rightarrow B_{ij} = 0, \qquad A_{ij} = 1 \Rightarrow B_{ij} \in \{0, \ldots, K\}, \quad K = 3. \tag{5}$$

A node projection head maps each latent token to an initial decoder feature, and an adjacency head scores every atom pair for edge existence. TopoBridge, denoted Π_{TB}, refines the resulting symmetric probability matrix into a connected adjacency $\widetilde{\boldsymbol{A}}$. That adjacency enters every decoder layer as a learned attention bias. An atom-type head and a pair-wise bond head follow, the latter conditioned on both $\widetilde{\boldsymbol{A}}$ and the predicted atom types. Appendix C.2.2 gives each head in full. The full bond distribution is then obtained by adjacency gating:

$$P_{ij,0}^B = 1 - \widetilde{A}_{ij}, \qquad P_{ij,c}^B = \widetilde{A}_{ij}\, \mathrm{softmax}(\boldsymbol{o}_{ij}^B)_c, \quad c = 1, \ldots, K. \tag{6}$$

At decode time the bond type is the argmax over the $K{+}1$ classes, so an edge TopoBridge supplies may still be assigned the no-bond class: adjacency fixes the support on which bonds are scored, and the bond head decides within it.

Coordinates are decoded last. The bond distribution is mapped to a continuous edge feature through learned bond embeddings, and an EGNN refines an initial coordinate prediction under that feature.

4.3 TopoBridge: connected adjacency refinement

TopoBridge is the connectivity counterpart of ChemCO: it supervises the decoder in training and remains in the decoding path at inference (Appendices O and Q).

TopoBridge maps the soft adjacency P^A to a binary connected adjacency $\widetilde{\boldsymbol{A}} = \Pi_{\mathrm{TB}}(P^A)$. It keeps the top-$k$ most confident neighbors per atom, k a per-dataset maximum degree applied after thresholding, symmetrizes, and repairs any disconnection by repeatedly inserting the highest-scoring cross-component pair, with S the reachable set of the largest component,

$$(u^\star, v^\star) = \arg \max_{u \in S,\, v \notin S} P^A_{uv}, \tag{7}$$

and re-running BFS from the largest component. Each insertion grows the reachable set S, so the procedure terminates with $\mathcal{G}(\widetilde{\boldsymbol{A}})$ connected. Because Π_{TB} is discrete, a straight-through surrogate (Bengio et al., 2013) carries gradients to the adjacency head. Appendix C.2 gives the surrogate, the supervision of the adjacency logits, and the connectivity guarantee in full.

The downstream heads condition on the refined adjacency. Edge insertion is rare at inference: on $10{,}000$ GEOM-Drugs samples, inserted edges are 0.078% of all bonds (Appendix Q).

4.4 ChemCO: Chemical Constraint Optimization by Mirror Ascent

TopoBridge fixes the candidate topology. ChemCO assigns the bond *types* on it, constraining them against valence caps and pair compatibility.

Let $k = 1, \ldots, K$ index bond types with bond orders o_k, and $k = 0$ denote no bond. From the decoder's raw logits $U^{(k)}_{ij}$ over all $K+1$ bond classes (including no-bond $k=0$), ChemCO takes the relative neural utility:

$$\bar{U}^{(k)}_{ij} = U^{(k)}_{ij} - U^{(0)}_{ij}, \qquad k = 1, \ldots, K, \tag{8}$$

as a fixed input encoding the decoder's preference for each bond type. ChemCO carries the assignment in a mirror coordinate $\Phi^{(k)}$, initialized as $\Phi^{(k)}_0 = \bar{U}^{(k)}$, and maps it to a continuous assignment score via

$$Y^{(k)} = T_k(\Phi) := \tfrac{1}{2}\big(g(\Phi^{(k)}) + g(\Phi^{(k)})^\top\big) \odot \widetilde{A} \odot \Omega^{(k)}, \quad g(x) = \sigma(x)^2, \tag{9}$$

where $\widetilde{A}$ is the TopoBridge adjacency support restricting optimization to candidate edges, and $\Omega^{(k)}$ is a binary mask that zeros out chemically invalid atom-bond combinations (e.g., a triple bond between two oxygen atoms). During training $\Omega^{(k)}$ is constructed from predicted atom types. From Y, ChemCO derives

$$s_{ij} = \textstyle\sum_k Y^{(k)}_{ij}, \qquad \mathrm{val}_i = \sum_{j \neq i} \sum_k o_k Y^{(k)}_{ij}, \qquad \deg_i = \sum_{j \neq i} s_{ij}, \tag{10}$$

subject to three chemical constraints: $s_{ij} \leq 1$ (each atom pair carries at most one bond type), $\mathrm{val}_i \leq c_i$ (total bond order at each atom does not exceed its valence capacity), and $\deg_i \geq d_{\min}$ (every non-padding atom participates in at least one bond, $d_{\min}=1$).

Objective. ChemCO maximizes expected bond utility while penalizing chemical constraint violations through an adaptive-penalty scheme. Each violation is smoothly approximated by $\psi_\gamma(r) = \gamma^{-1}\log(1 + e^{\gamma r})$, a soft surrogate for $[r]_+$. The per-step score is

$$\begin{aligned}\mathcal{J}_t(\Phi) = \textstyle\sum_{i<j} \sum_{k=1}^{K} \bar{U}^{(k)}_{ij} Y^{(k)}_{ij} - \sum_{i<j} \mu_{ij,t}\psi_\gamma(s_{ij} - 1) \\ - \textstyle\sum_i \lambda_{i,t}\psi_\gamma(\mathrm{val}_i - c_i) - \sum_i \nu_{i,t}\psi_\gamma(d_{\min} - \deg_i),\end{aligned} \tag{11}$$

where $Y = T(\Phi)$. The three penalty terms enforce bond exclusivity, valence limits, and minimum-degree requirements respectively. After each ascent step on Φ, the weights μ, λ, ν are raised on still-violated constraints and never lowered, so pressure accumulates where a violation persists. Full update rules are in Appendix S. We also provide analysis of the computational overhead in Appendix R.

4.5 Advantage-Gated Constraint Learning

Following the positive-advantage principle (Oh et al., 2018; Peng et al., 2019), AGCL injects ChemCO's signal only when it raises the score on the ground-truth bond type, on bonded pairs.

With P_{raw} and P_{chem} the bond distributions before and after ChemCO, and $y^{\star}_{ij}$ the ground-truth labels on the bonded pair set $\mathcal{S}_b$ of molecule b, the per-molecule advantage is

$$\mathcal{E}_b(P) = \frac{1}{|\mathcal{S}_b|} \sum_{(i,j)\in\mathcal{S}_b} -\log P_{ij,y^{\star}_{ij}}, \qquad a_b = [\mathcal{E}_b(P_{\text{raw}}) - \mathcal{E}_b(P_{\text{chem}})]_+ . \tag{12}$$

The advantage, positive only when ChemCO raises the score on the ground-truth bond type, gates an L_2 consistency loss $\mathcal{L}_{\text{AGCL}}$ that pulls P_{raw} toward $\text{sg}(P_{\text{chem}})$ on those pairs, the stop-gradient ensuring it steers only the raw decoder (Appendix C.4). The continuous constraint solver therefore does not run at sampling.

4.6 Training Objective

TopVAE is trained end-to-end with reconstruction losses for each decoded modality, KL regularization, and the AGCL selective teacher:

$$\mathcal{L} = \mathcal{L}_{\text{topo}} + \mathcal{L}_{\text{atom}} + \mathcal{L}_{\text{3D}} + \beta\,\mathcal{L}_{\text{KL}} + \lambda_{\text{AGCL}}\,\mathcal{L}_{\text{AGCL}}. \tag{13}$$

Here $\mathcal{L}_{\text{topo}}$ groups focal binary cross-entropy (Lin et al., 2017) for the adjacency matrix, in which bonded pairs are a small minority of the $O(N^2)$ candidate pairs, and cross-entropy for bond types, $\mathcal{L}_{\text{atom}}$ is cross-entropy over atom types, $\mathcal{L}_{\text{3D}}$ combines coordinate regression and pairwise distance supervision, and $\mathcal{L}_{\text{KL}}$ is the standard VAE KL divergence. All individual loss terms, weights, architectural details, ChemCO hyperparameters, and warmup schedules are reported in Appendix C.

5 Experiments

5.1 Setup

Datasets. We evaluate on QM9 (Ramakrishnan et al., 2014), ∼134k organic molecules with up to 9 heavy atoms (C, N, O, F) at DFT geometries, and GEOM-Drugs (Axelrod & Gomez-Bombarelli, 2022), ∼304k drug-like molecules with up to 90 heavy atoms and multiple conformers each.

Baselines. UDM-3D (Luo et al., 2025) pairs the unified autoencoder UAE with UDM, a diffusion prior over its latents, and we write that pair UAE+UDM. Ours is TopVAE+DiT. UAE's encoder is of the same Relational Transformer family as TopVAE's. The decoder and its training are what this paper changes, at the matched latent width of the QM9 pair (Appendix C.2). JODO and LDM-3DG release their generated samples, so those two and TopVAE+DiT are also scored through one code path (Appendix G).

Metrics. An atom is stable when its valence equals the reference value for its element, and a molecule is stable when all of its atoms are (Hoogeboom et al., 2022). We report atom stability, molecular stability (MolStab), validity (RDKit sanitization) and connectivity, and their conjunctions **V&C** (validity ∧ connectivity), **V&U** (validity ∧ uniqueness) and **S∧C** (MolStab ∧ connectivity). V&C is the chemical validity of Sec. 3. **VCB** is the fraction of decoded graphs that are valid, connected and fully bonded. **FCD** (Preuer et al., 2018) is taken on the predicted bond graph and **FCD$_{\text{3D}}$** on the graph re-inferred from the generated coordinates alone. **iFID** (Xu et al., 2026) is its latent-space adaptation, taken between the validation molecules and the set decoded from interpolated latents. Appendix C.7 defines each metric and states its bond source.

5.2 Dark-Area Diagnosis

A reconstruction score is taken at the posterior alone, and a diffusion prior does not draw posterior codes. What decides whether its samples decode is how far off the posterior a decoder still returns molecules. For each validation molecule M, we encode its posterior mean $\boldsymbol{z} = \boldsymbol{\mu}_\phi(M)$ and decode the perturbed latent

$$\tilde{\boldsymbol{z}} = \boldsymbol{z} + \sigma \cdot s_{\text{emp}} \cdot \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \tag{14}$$

Table 1: **VAE dark-area diagnosis** on QM9 (validation split, seed-42 subsample, n=10,000, and n=2,000 for the ADiT-VAE cells, Appendix E). $\bar{s}_{\text{emp}}$ averages s_{emp}, the spread that scales σ in Eq. (14), over the d dimensions. UAE and TopVAE share the QM9 latent width d=16.

	Post. MolStab ↑	Post. S∧C ↑	$\bar{s}_{\text{emp}}$ (d)	σ=0.5 s_{emp} VCB ↑	iFID ↓
QM9					
UAE	**1.000**	**1.000**	18.03 (16)	0.014	2.88
ADiT-VAE	0.962	0.962	0.69 (8)	0.000	6.58
TopVAE	**1.000**	**1.000**	6.18 (16)	**0.460**	**0.395**

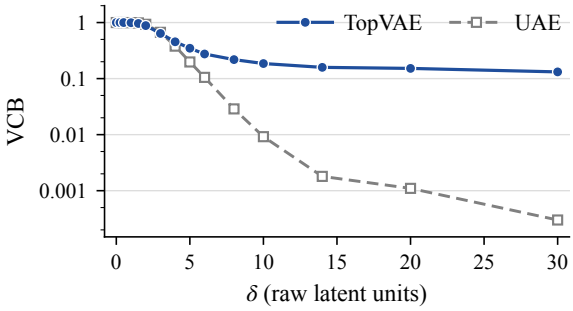


Figure 2: **Cross-model dark-area diagnosis at matched absolute displacement** (QM9, n=10,000, on the checkpoints and subset of Table 1). VCB (log scale) against δ, the per-coordinate standard deviation of isotropic Gaussian noise added to the latent code, in raw latent units and the same for both models. δ=0 is the reconstruction anchor.

where $s_{\text{emp}} = \text{std}(\{\boldsymbol{\mu}_\phi(M)\}_{M\in\mathcal{D}})$ is the per-dimension standard deviation of the encoded latents. UAE's posterior means spread 18.03 per dimension against TopVAE's 6.18, so a displacement fixed in this convention is not fixed in raw units, and Figure 2 repeats the comparison at a matched absolute displacement δ. That ratio, 2.92, is the conversion: σ=0.5 is 9.01 raw units for UAE against 3.09 for TopVAE. The prior is trained on whitened latents (Appendix C.6), so a displacement of a given size in the space the prior samples is that displacement in units of σ, the convention of Eq. (14).

Table 1 reads the three autoencoders at σ=0 and 0.5. Over interpolated latents TopVAE returns the lowest iFID of the three, 0.395 against 2.88 and 6.58. UAE and TopVAE both reconstruct the split. At σ=0.5 the strongest baseline returns 0.014 of perturbed latents as valid, connected and fully bonded graphs, against 0.460 for TopVAE. TopVAE leads at every nonzero σ of Appendix D.3 and, at matched absolute displacement, from δ=5 to δ=30 (Figure 2). We retrained the baseline autoencoder five times under its published recipe, changing only the training seed, and swept all five on the same grid. From δ=5 to δ=30 TopVAE is above the highest of the five, and its margin over their mean runs from 3.6 to 354 standard deviations: at δ=10, 0.186 against a mean of 0.014. A second TopVAE training is above all five from δ=4. All seven trainings reconstruct the split to within 0.0005 of one another, so none of this separation is visible at the posterior (Appendix I).

This robustness extends to the QM9 diffusion trajectory: at the pure-noise end it decodes 10.3% of the latents the sampler visits to valid, connected molecules, where UAE+UDM returns 0, over 1,024 trajectories each (Appendix D).

5.3 Dark Areas Out of Distribution

Scaffold expansion drives the decoder out of distribution on demand, and we use it to measure the decoding path rather than to benchmark scaffold completion. Given a scaffold with k atoms, we encode it into k latent tokens and fill the remaining $n-k$ slots (heavy atoms) via reverse diffusion. Only those k tokens are posterior codes. The rest come from the prior, so the share of the sequence the decoder was never supervised on grows with the expansion. Growing benzene by 100 heavy atoms leaves 6 posterior tokens in 106. On QM9, where the displacement grid is calibrated, conditioning moves the latents outward: the median displacement of the completions is 2.1 times its unconditional value at a one-atom expansion and 2.4 to 2.5 times at a three-atom expansion, monotone in the expansion (Appendix D.3).

Table 2 sweeps the expansion band on GEOM-Drugs over four aromatic scaffolds, and Figure 3 plots the same sweep against the baseline: between +20 and +80 heavy atoms the baseline's sanitization

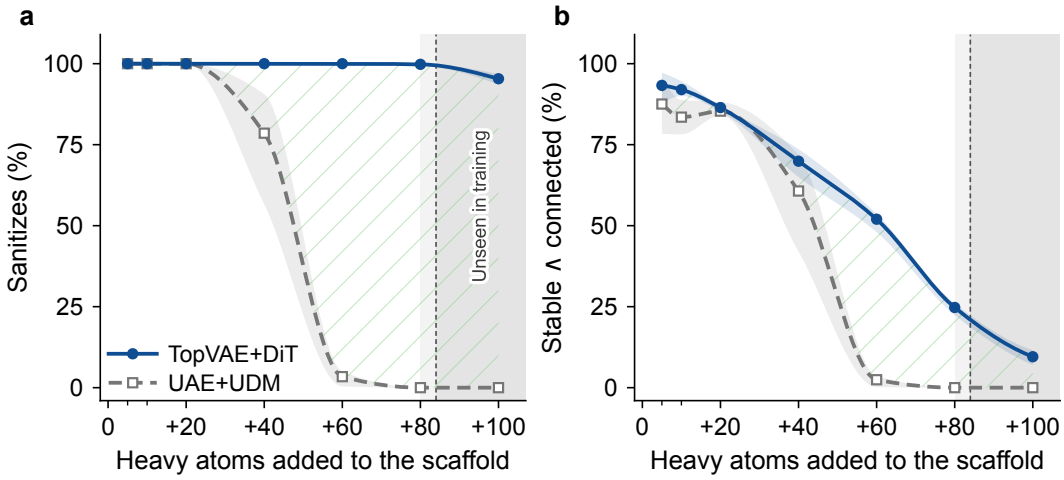


Figure 3: **Decoding past the training maximum.** Mean over the four aromatic scaffolds, 500 completions per band. Each band spans the four scaffolds from lowest to highest, the hatch is the margin between the two means, and the curves interpolate between the measured markers. **(a)** The fraction RDKit sanitizes. **(b)** The fraction that is stable and a single connected component. Shading on the right marks products at or past the 90-heavy-atom training maximum, from +80 for the largest scaffold and from +84 for all four. The two arms are TopVAE+DiT and UAE+UDM, each autoencoder with its own diffusion prior, the baseline trained by us on this split.

Table 2: **Zero-shot scaffold expansion on GEOM-Drugs.** Mean over benzene, pyridine, naphthalene and indole, 500 completions per band, percentages. *Joint* is the scaffold-completion criterion: the product preserves the conditioning scaffold, is valence-legal and forms a single connected component. UAE is UAE+UDM, a latent diffusion prior over a reconstruction-trained autoencoder trained by us on this split, and TopVAE is TopVAE+DiT. * marks a band whose scaffold-clustered bootstrap 95% confidence interval excludes zero. S∧C is the fraction of TopVAE's products that are stable and a single connected component. *Sanitizes* is the fraction RDKit sanitizes. The training maximum is 90 heavy atoms. Per-scaffold cells in Appendix Tables 26 and 27.

	Joint ↑			TopVAE	
+n heavy atoms	UAE	TopVAE	Δ	**S∧C ↑**	**Sanitizes ↑**
+5	21.3	**23.1**	+1.8	93.3	100.0
+10	5.1	**15.4**	+10.3*	92.0	100.0
+20	2.6	**10.4**	+7.8*	86.4	100.0
+40	5.6	**15.0**	+9.4*	69.9	100.0
+60	0.2	**9.5**	+9.3*	52.0	100.0
+80	0.0	**3.8**	+3.8*	24.9	99.8
+100	0.0	**0.5**	+0.5	9.6	95.2
Mean	5.0	**11.1**	+6.1*	61.2	99.3

rate falls from 100% to 0, while TopVAE holds 100% through +60, returns 99.8% at +80 and 95.2% at +100.

In the +100 band, products of 106 to 110 heavy atoms, the reported checkpoint decodes 9.6% of products to stable and connected molecules and 95.2% to sanitizable ones. A latent diffusion prior over a reconstruction-trained autoencoder returns 0 sanitizable molecules of 4,000 across the +80 and +100 bands, against 97.5% for TopVAE (Appendix N). Conditioning drives the decoder out of distribution the same way: steering a frozen prior toward property targets with a small regressor fitted in the frozen latent space reaches a mean absolute error of 4.18 Å^2 for topological polar surface area and 0.28 for the octanol-water partition coefficient, at validity 0.971 and 0.981 (Appendix O).

Table 3: **QM9 de novo 3D molecule generation** (n=10,000). Baseline cells are published values (sources per row in Appendix F). A dash means that method's own paper does not report the quantity, except in the training-set row, where an FCD would be the reference measured against itself. Every unmarked FCD cell is computed against its own source's reference set. †Our own evaluation of that method's released samples, all three through one code path against one reference rebuilt from the test split, at sampling seed 42 (Appendix G). Every other FCD cell is a published value. Best in bold, second best underlined, except the FCD columns, left unranked because their cells do not share a reference set. TopVAE+DiT is our own evaluation (τ=1.0). Each unmarked TopVAE+DiT cell is the mean over three sampling seeds.

2D metrics, on the predicted bond graph

Method	FCD↓	Atom Stab.↑	Mol Stab.↑	V&C↑	V&U↑
Training set	–	0.999	0.988	0.989	0.989
CDGS	0.798	0.997	0.951	0.951	0.936
LDM-3DG	2.440†	0.9932†	0.8623†	**0.9999**†	**0.9756**†
MiDi	0.187	0.998	0.976	0.980	0.954
JODO	0.136†	<u>0.999</u>	<u>0.988</u>	0.990	0.960
EQGAT-diff	2.088	<u>0.999</u>	0.971	0.965	0.950
SemlaFlow	0.863	0.995	0.949	0.857	0.821
UDM-3D	0.130	<u>0.999</u>	<u>0.988</u>	0.983	<u>0.973</u>
TopVAE+DiT	1.056†	**1.000**	**1.000**	<u>0.998</u>	0.966

3D metrics

Method	FCD_{3D}↓	Atom Stab.↑	Mol Stab.↑
Training set	–	0.994	0.953
EDM	1.285	0.986	0.817
MDM	4.861	0.992	0.896
GeoLDM	1.030	0.989	0.897
LDM-3DG	2.157†	0.9462†	0.6913†
MolFLAE	–	**0.994**	0.920
GFMDiff	–	0.989	0.877
ADiT	2.884	0.211	–
MiDi	1.100	0.983	0.842
JODO	0.137†	0.992	<u>0.934</u>
EQGAT-diff	1.520	0.988	0.888
SemlaFlow	1.127	0.971	0.787
UDM-3D	0.881	<u>0.993</u>	**0.935**
TopVAE+DiT	0.340†	0.991	0.928

5.4 De Novo Generation with a Standard DiT

Tables 3 and 4 carry published values, each taken under its source's own protocol. The three methods that release their samples are scored instead through one code path against a single reference rebuilt from the test split, marked † in both tables (Appendix G).

QM9. TopVAE+DiT is the only method at 1.000 2D atom and 2D molecular stability, and its V&C of 0.998 is second to LDM-3DG's 0.9999 (Table 3). Its 3D stabilities sit within 0.003 and 0.007 of the highest listed, 0.991 against 0.994 and 0.928 against 0.935. JODO has the lower QM9 FCD_{3D}, 0.137 against 0.340, and the lower 2D FCD, 0.136 against 1.056.

GEOM-Drugs. Table 4 gives GEOM-Drugs. Through the one pipeline, TopVAE+DiT has the lowest FCD_{3D}, 1.501 against JODO's 1.956 and LDM-3DG's 6.284, and the highest V&U. LDM-3DG has the highest V&C and both molecular stabilities, and JODO the lower 2D FCD.

Summary. We provide additional analysis of chemical properties, energy relaxation, stereochemistry and diversity (Appendices J to M). Sampled molecules are shown in Figure 6.

5.5 Ablation

Component ablation. Retraining one change at a time, removing ChemCO and removing AGCL each cost VCB at δ=0.5, 4 and 5, more at δ=4 than at δ=0.5 (Table 5). All four settings reconstruct the split at δ=0 at 0.9999, so the spread is entirely off the posterior. At δ=0.5 the three ablated experiments return 0.6745 to 0.7454 against A0's 0.9999, and at δ=5 they return 0.0611 to 0.0867 against 0.3758, so the ratio to A0 rises from roughly 1.4 to between 4.3 and 6.2 as the displacement grows. Removing the gate is worst at δ=0.5, at 0.6745, and a teacher without chemistry is worst at δ=4 and δ=5, at 0.0923 and 0.0611.

Table 4: **GEOM-Drugs de novo 3D molecule generation** (n=10,000). Every baseline cell is a published value except the cells marked † (Appendix F). Conventions and seeds follow Table 3. TopVAE+DiT is sampled at τ=0.95. †Our own evaluation of that method's released samples, all three through one code path against one reference rebuilt from the test split, at sampling seed 42 (Appendix G). Every other FCD cell is a published value.

2D metrics, on the predicted bond graph

Method	FCD↓	Atom Stab.↑	Mol Stab.↑	V&C↑	V&U↑
Training set	–	1.000	1.000	1.000	1.000
CDGS	22.051	0.991	0.706	0.285	0.285
LDM-3DG	4.448†	0.9997†	**0.991**†	**0.9998**†	0.9997†
MiDi	7.054	0.968	0.822	0.633	0.654
JODO	2.565†	**1.000**	0.981	0.874	0.905
EQGAT-diff	5.898	**1.000**	0.989	0.845	0.863
SemlaFlow	–	0.998	0.973	–	0.939
UDM-3D	0.692	**1.000**	0.925	0.879	0.913
TopVAE+DiT	3.573†	0.9995	0.979	0.950	**0.9999**

3D metrics

Method	FCD_{3D}↓	Atom Stab.↑	Mol Stab.↑
Training set	–	0.861	0.028
EDM	31.290	0.831	0.002
GeoLDM	30.68	0.843	0.008
LDM-3DG	6.284†	0.791†	**0.0331**†
MiDi	23.14	0.750	0.003
JODO	1.956†	0.845	0.010
EQGAT-diff	26.33	0.825	0.007
UDM-3D	17.36	**0.852**	0.014
TopVAE+DiT	1.501†	0.833	0.0205

Table 5: **Controlled ablations of AGCL** (QM9 validation split, n=10,000, `sample_seed` 42). Each configuration retrains the reported recipe with one change. A0 makes none. Entries are VCB. All four reconstruct the split at δ=0 at 0.9999, so the arms are read from a common reconstruction anchor. Best in each column in bold.

	Matched absolute displacement		
Configuration	δ=0.5	δ=4	δ=5
A0 reported recipe	**0.9999**	**0.4883**	**0.3758**
A1 without AGCL	0.6745	0.1101	0.0728
A2 chemistry-free teacher	0.7040	0.0923	0.0611
A3 without ChemCO	0.7454	0.1191	0.0867

6 LIMITATIONS

TopVAE decodes element identity, bond order and coordinates, and formal charge is not among the decoded channels. All SMILES in this paper are written with `isomericSmiles=False`, so validity, uniqueness, novelty and FCD are computed on the constitutional graph, and Appendix L assigns stereochemistry from the generated coordinates for TopVAE+DiT and JODO.

ChemCO's constraints are graph-level, namely valence caps, degree bounds and pair compatibility, and geometry is not among them, so the coordinate head is trained by the reconstruction loss alone. Any differentiable constraint with a primal-dual form can be carried the same way, which makes 3D physical priors such as van der Waals clash penalties and force-field energy bounds the natural next case. The conditional experiment steers a frozen prior toward two descriptors computed in closed form from the graph. Targets that need a learned property model, such as quantum properties or pocket-conditioned binding, are outside what we test here.

7 CONCLUSION

We gave a perturbation protocol that measures *dark areas*, the latent regions where a VAE decoder returns chemically invalid or disconnected structures while reconstructing the posterior exactly. Reconstruction at the posterior does not certify a latent space for diffusion: the two decoders separate where conditioning and size extrapolation build latent codes no training molecule supplied. TopVAE reduces dark areas by carrying connectivity and chemistry inside the decoding path, and with a standard DiT it reaches the lowest GEOM-Drugs FCD_{3D} of the three methods scored through one pipeline.

ACKNOWLEDGMENTS

Omitted to preserve anonymity.

ETHICS STATEMENT

This work studies unconditional and scaffold-conditioned generation of small drug-like molecules on two public benchmarks, QM9 and GEOM-Drugs, and involves no human subjects, no personal data and no new data collection. Generative models for molecules are dual-use: the same capability that proposes candidate therapeutics can propose harmful compounds. Two properties of this setting bound that concern in practice. The benchmarks contain no toxicity, potency or synthesis-route information, so a model trained on them has no pathway-specific knowledge to transfer, and every molecule it produces is a hypothesis that still requires synthesis, assay and regulatory review before it can affect anyone. We consider the marginal risk of releasing this work small relative to its contribution to reliable molecular design, and we make no claim that the generated molecules are safe, synthesizable or biologically active.

REPRODUCIBILITY STATEMENT

Every number we report comes from a checkpoint and a command we will publish. Upon acceptance we will release the complete implementation, the training and evaluation configurations, and every checkpoint used for every table, under a permissive license. Appendix C.7 defines every metric and states its bond source. Appendix C gives the training protocol, architecture and hyperparameters. Sampling temperature and reference set are stated beside each table.

AI USE STATEMENT

We used generative AI tools as writing assistants, to improve the grammar, clarity and concision of text drafted by the authors. We reviewed all AI-assisted material and take full responsibility for the content of this work.

REFERENCES

Brandon Amos and J Zico Kolter. Optnet: Differentiable optimization as a layer in neural networks. In *International conference on machine learning*, pp. 136–145. PMLR, 2017.

Simon Axelrod and Rafael Gomez-Bombarelli. Geom, energy-annotated molecular conformations for property prediction and molecular generation. *Scientific data*, 9(1):185, 2022.

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Zitao Chen, Yinjun Jia, Zitong Tian, Wei-Ying Ma, and Yanyan Lan. Manipulating 3d molecules in a fixed-dimensional E(3)-equivariant latent space. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2506.00771.

Hanjun Dai, Yingtao Tian, Bo Dai, Steven Skiena, and Le Song. Syntax-directed variational autoencoder for structured data. *arXiv preprint arXiv:1802.08786*, 2018.

Cameron Diao and Ricky Loynd. Relational attention: Generalizing transformers for graph-structured tasks. *arXiv preprint arXiv:2210.05062*, 2022.

Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276, 2018.

Matthew D Hoffman, Carlos Riquelme, and Matthew J Johnson. The β-vae's implicit prior. In *Workshop on Bayesian Deep Learning, NIPS*, pp. 1–5, 2017.

Emiel Hoogeboom, Vıctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In *International conference on machine learning*, pp. 8867–8887. PMLR, 2022.

Han Huang, Leilei Sun, Bowen Du, and Weifeng Lv. Learning joint 2d & 3d diffusion models for complete molecule generation. *arXiv preprint arXiv:2305.12347*, 2023.

Lei Huang, Hengtong Zhang, Tingyang Xu, and Ka-Chun Wong. Mdm: Molecular diffusion model for 3d molecule generation. *arXiv preprint arXiv:2209.05710*, 2022.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. In *International conference on machine learning*, pp. 2323–2332. PMLR, 2018.

Chaitanya K Joshi, Xiang Fu, Yi-Lun Liao, Vahe Gharakhanyan, Benjamin Kurt Miller, Anuroop Sriram, and Zachary W Ulissi. All-atom diffusion transformers: Unified generative modelling of molecules and materials. *arXiv preprint arXiv:2503.03965*, 2025.

Mario Krenn, Qianxiang Ai, Senja Barthel, Nessa Carson, Angelo Frei, Nathan C Frey, Pascal Friederich, Théophile Gaudin, Alberto Alexander Gayle, Kevin Maik Jablonka, et al. Selfies and the future of molecular string representations. *Patterns*, 3(10), 2022.

Matt J Kusner, Brooks Paige, and José Miguel Hernández-Lobato. Grammar variational autoencoder. In *International conference on machine learning*, pp. 1945–1954. PMLR, 2017.

Tuan Le, Julian Cremer, Frank Noe, Djork-Arné Clevert, and Kristof Schütt. Navigating the design space of equivariant diffusion-based generative models for de novo 3d molecule generation. *arXiv preprint arXiv:2309.17296*, 2023.

Hyomin Lee, Minseon Kim, Sangwon Jang, Jongheon Jeong, and Sung Ju Hwang. Enhancing variational autoencoders with smooth robust latent encoding. *arXiv preprint arXiv:2504.17219*, 2025.

Ruizhe Li, Xutan Peng, Chenghua Lin, Wenge Rong, and Zhigang Chen. On the low-density latent regions of vae-based language models. In *NeurIPS 2020 Workshop on Pre-registration in Machine Learning*, pp. 343–357. PMLR, 2021.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Qi Liu, Miltiadis Allamanis, Marc Brockschmidt, and Alexander Gaunt. Constrained graph variational autoencoders for molecule design. *Advances in neural information processing systems*, 31, 2018.

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. *arXiv preprint arXiv:2503.15567*, 2025.

Youzhi Luo, Keqiang Yan, and Shuiwang Ji. Graphdf: A discrete flow model for molecular graph generation. In *International conference on machine learning*, pp. 7192–7203. PMLR, 2021.

Tengfei Ma, Jie Chen, and Cao Xiao. Constrained generation of semantically valid graphs via regularizing variational autoencoders. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/1458e7509aa5f47ecfb92536e7dd1dc7-Paper.pdf.

Manuel Madeira, Clement Vignac, Dorina Thanou, and Pascal Frossard. Generative modelling of structurally constrained graphs. *Advances in Neural Information Processing Systems*, 37: 137218–137262, 2024.

Arkadi Nemirovski and David Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1983.

Toshiki Ochiai, Tensei Inukai, Manato Akiyama, Kairi Furui, Masahito Ohue, Nobuaki Matsumori, Shinsuke Inuki, Motonari Uesugi, Toshiaki Sunazuka, Kazuya Kikuchi, et al. Variational autoencoder-based chemical latent space for large molecular structures with 3d complexity. *Communications Chemistry*, 6(1):249, 2023.

Junhyuk Oh, Yijie Guo, Satinder Singh, and Honglak Lee. Self-imitation learning. In *International conference on machine learning*, pp. 3878–3887. PMLR, 2018.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Xingang Peng, Jiaqi Guan, Qiang Liu, and Jianzhu Ma. Moldiff: Addressing the atom-bond inconsistency problem in 3d molecule diffusion generation. *arXiv preprint arXiv:2305.07508*, 2023.

Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*, 2019.

Kristina Preuer, Philipp Renz, Thomas Unterthiner, Sepp Hochreiter, and Gunter Klambauer. Fréchet chemnet distance: a metric for generative models for molecules in drug discovery. *Journal of chemical information and modeling*, 58(9):1736–1741, 2018.

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.

Danny Reidenbach, Filipp Nikitin, Olexandr Isayev, and Saee Gopal Paliwal. Applications of modular co-design for de novo 3d molecule generation. *Digital Discovery*, 5(2):754–768, 2026.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Chence Shi, Minkai Xu, Zhaocheng Zhu, Weinan Zhang, Ming Zhang, and Jian Tang. Graphaf: a flow-based autoregressive model for molecular graph generation. *arXiv preprint arXiv:2001.09382*, 2020.

Ivan Skorokhodov, Sharath Girish, Benran Hu, Willi Menapace, Yanyu Li, Rameen Abdal, Sergey Tulyakov, and Aliaksandr Siarohin. Improving the diffusability of autoencoders. *arXiv preprint arXiv:2502.14831*, 2025.

Clement Vignac, Igor Krawczuk, Antoine Siraudin, Bohan Wang, Volkan Cevher, and Pascal Frossard. Digress: Discrete denoising diffusion for graph generation. *arXiv preprint arXiv:2209.14734*, 2022.

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*, pp. 560–576. Springer, 2023.

Runzhong Wang, Yunhao Zhang, Ziao Guo, Tianyi Chen, Xiaokang Yang, and Junchi Yan. Linsatnet: the positive linear satisfiability neural networks. In *International Conference on Machine Learning*, pp. 36605–36625. PMLR, 2023.

Can Xu, Haosen Wang, Weigang Wang, Pengfei Zheng, and Hongyang Chen. Geometric-facilitated denoising diffusion model for 3d molecule generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 338–346, 2024.

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*, pp. 38592–38610. PMLR, 2023.

Tongda Xu, Mingwei He, Shady Abu-Hussein, Jose Miguel Hernandez-Lobato, Haotian Zhang, Kai Zhao, Chao Zhou, Ya-Qin Zhang, and Yan Wang. Making reconstruction fid predictive of diffusion generation fid. *arXiv preprint arXiv:2603.05630*, 2026.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15703–15712, 2025.

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.

Chengxi Zang and Fei Wang. Moflow: an invertible flow model for generating molecular graphs. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 617–626, 2020.

Hongtai Zeng, Chao Yang, Yanzhen Zhou, Cheng Yang, and Qinglai Guo. Glinsat: The general linear satisfiability neural network layer by accelerated gradient descent. *Advances in Neural Information Processing Systems*, 37:122584–122615, 2024.

APPENDIX CONTENTS

A Background **15**

B Extended Related Work **15**

C Training Protocol **16**

D Extended Dark-Area Analysis **22**

E Reproduction Ledger **24**

F Provenance of the Baseline Numbers in Tables 3 and 4 **25**

G Three Methods Scored From Their Released Samples **25**

H VAE Posterior Reconstruction Quality **28**

I Sampling Variability of the Headline Runs **28**

J Molecular Property Distributions **29**

K Energy Relaxation Evaluation **30**

L Stereochemistry of the Generated Coordinates **31**

M Diversity and Novelty **32**

N Scaffold Inpainting: Per-Scaffold Results **32**

O Conditional Generation: Protocol and Extended Results **32**

P Controlled Ablations of AGCL **36**

Q TopoBridge Chemical Plausibility Analysis **38**

R Computational Cost Analysis **38**

S Details of ChemCO **39**

A Background

The following definitions are standard and were moved here from Section 3 for length.

3D molecular VAE. A 3D molecule can be represented as $M = (\boldsymbol{A}, \boldsymbol{T}, \boldsymbol{B}, \boldsymbol{R})$, where $\boldsymbol{A} \in \{0,1\}^{N\times N}$ is the binary adjacency matrix, $\boldsymbol{T}$ denotes atom types, $\boldsymbol{B}$ denotes bond types, and $\boldsymbol{R} \in \mathbb{R}^{N\times 3}$ stores atomic coordinates. A molecular VAE encodes M into a latent variable $\boldsymbol{z}$ and reconstructs the molecule through a decoder D_θ:

$$q_\phi(\boldsymbol{z} \mid M) = \mathcal{N}\big(\boldsymbol{\mu}_\phi(M), \mathrm{diag}(\boldsymbol{\sigma}^2_\phi(M))\big), \qquad \hat{M} = D_\theta(\boldsymbol{z}), \quad \boldsymbol{z} = \boldsymbol{\mu}_\phi(M) + \boldsymbol{\sigma}_\phi(M) \odot \boldsymbol{\epsilon}. \tag{15}$$

Here $q_\phi(\boldsymbol{z} \mid M)$ is the encoder posterior, parameterized by encoder parameters ϕ, which maps an input molecule to the mean and variance of a Gaussian latent distribution. It is trained by balancing reconstruction against the KL from $q_\phi(\boldsymbol{z} \mid M)$ to the prior $p(\boldsymbol{z})$, at weight β.

The prior is trained over this latent space, and the decoder must therefore work on the terminal latents that prior produces, which need not lie in the support of the encoder's posterior.

B Extended Related Work

This appendix expands on the related work discussion in Sec. 2, providing detailed comparisons with prior methods and clarifying where TopVAE fits relative to VAE latent-space analyses and constraint-aware decoding methods.

Latent-space quality in VAEs. A VAE decoder is trained on samples from the encoder posterior, but downstream latent generation may query regions closer to the aggregated posterior, interpolation paths between encoded molecules, or trajectories produced by a learned latent prior. This creates a gap between reconstruction-time decoding and generation-time decoding. Prior analyses relate this gap to aggregate-posterior mismatch and to the phenomenon of low-density latent "holes" (Hoffman et al., 2017; Li et al., 2021). For molecular generation, such holes are especially consequential: an off-posterior latent may decode not merely to a low-quality sample, but to a disconnected graph or a molecule that violates chemical valence and sanitization constraints.

Earlier molecular VAEs already showed that continuous molecular latent spaces are useful for optimization, while also exposing the importance of validity and representation bias in molecular decoding: grammar and syntax constraints, substructure-level generation and robust string representations were all introduced to raise the validity of decoded molecules (Gómez-Bombarelli et al., 2018; Kusner et al., 2017; Dai et al., 2018; Jin et al., 2018; Krenn et al., 2022; Ochiai et al., 2023). Beyond molecules, SRL-VAE shows that adversarial smoothing can improve latent robustness in image VAEs (Lee et al., 2025).

These methods motivate the importance of smooth and valid latent decoding, but they do not directly test a 3D molecular VAE under the latent distribution induced by modern latent diffusion. TopVAE makes this test explicit through dark-area analysis: we evaluate whether the decoder remains valid under posterior perturbations, interpolation, prior sampling, and diffusion sampling, and we identify topological disconnection and chemical invalidity as two major failure modes.

Constraint-aware graph and molecule decoding. Validity-aware molecular generation has been addressed at multiple levels. At the representation level, grammar-based methods and SELFIES restrict the output language so that decoded strings better respect syntactic or semantic molecular validity (Kusner et al., 2017; Krenn et al., 2022). At the substructure or graph-decoder level, JT-VAE and constrained graph VAEs impose chemical validity through the generative procedure itself: JT-VAE builds molecules from chemical substructure trees (Jin et al., 2018), while CGVAE uses constrained graph extension to guide molecular graph generation (Liu et al., 2018). At the graph-regularization level, regularized graph VAEs penalize invalid matrix-valued outputs to encourage properties such as connectivity and valence consistency (Ma et al., 2018).

Another line of work introduces differentiable optimization or satisfiability layers inside neural networks. OptNet embeds a differentiable quadratic programming solver as a neural network layer (Amos & Kolter, 2017), while LinSATNet and GLinSAT project neural outputs toward feasible

sets defined by linear satisfiability constraints (Wang et al., 2023; Zeng et al., 2024). Those layers are differentiated through, so the constraint set enters the decoder's gradient. ChemCO runs under `no_grad`, so no gradient reaches the decoder along it. It acts on the decoder in two places: the advantage-gated teacher loss of Sec. 4.5, and the decoding path itself, where the projected bond distribution is what the coordinate stage conditions on. Its update is mirror ascent (Nemirovski & Yudin, 1983; Beck & Teboulle, 2003) on a penalized bond-assignment objective, with penalty weights raised on persistent violations.

Constrained diffusion methods impose hard structural constraints along the generative trajectory. For example, ConStruct maintains graph constraints such as planarity or acyclicity by combining edge-absorbing noise with projection operators (Madeira et al., 2024). These structural graph constraints are related to molecular validity, but they are not identical to chemical sanitization constraints. In particular, molecular decoding must handle both global graph structure and local chemistry-specific rules.

Discrete graph diffusion, and valence enforced at generation time. Three further families of graph generator enforce constraints that overlap with the ones TopVAE trains against, each at a different point in the pipeline. Discrete graph diffusion puts them in the noise process: DiGress denoises a categorical graph under transition matrices whose limit distribution is the training marginal over node and edge types, and MiDi couples that discrete process to a continuous one over 3D coordinates (Vignac et al., 2022; 2023). Both act in data space, with no latent code and no decoder, so the decodability of a perturbed latent that Sec. 5.2 measures has no counterpart in them. Autoregressive and flow-based molecular generators instead enforce valence on the sample. GraphAF and GraphDF mask the action space of a sequential sampler so that no step can add a bond exceeding an atom's remaining valence (Shi et al., 2020; Luo et al., 2021), and MoFlow, which produces the graph in one shot, applies a post-hoc validity correction to the decoded molecule (Zang & Wang, 2020). Masking presupposes a generation process with steps at which to act, and TopVAE's decoder (Eq. (4)) scores every atom pair of each stage in one pass. A post-hoc correction is applicable to such a decoder, and the difference is where the correction is spent: routed through AGCL, ChemCO's corrections alter the decoder's parameters during training rather than the molecules a sampler returns (Sec. 4.5).

C Training Protocol

This section provides complete training details for reproducing TopVAE.

C.1 Data Preprocessing

Kekulization. All molecules are kekulized via `Chem.Kekulize(mol, clearAromaticFlags=True)`, converting aromatic bonds to alternating single/double bonds and reducing the bond vocabulary from {single, double, triple, aromatic} to $K = 3$ kekulized types {single, double, triple}. Molecules that fail kekulization are discarded, identically for every model we train and evaluate.

Atom and bond vocabularies. QM9 (heavy-only): 4 atom types {C, N, O, F}, max 9 heavy atoms. GEOM-Drugs (heavy-only): 12 atom types {C, N, O, F, S, Cl, Br, P, I, B, Si, Bi}, max 90 heavy atoms.

Input features. The encoder receives a fully-connected edge graph ($N \times N$). Bond edges carry one-hot bond type, non-bond edges carry zeros, and self-loops receive a dedicated flag. Node features concatenate atom-type one-hot with 39 GeoMol-style features (atomic number, aromaticity, degree, hybridization, implicit valence, formal charge, ring membership), plus 3D coordinates.

Augmentation. All coordinates are mean-centered. GEOM-Drugs training applies random SO(3) rotation and Gaussian translation (scale 0.1). QM9 does not.

C.2 Model Architectures

C.2.1 TopoBridge straight-through surrogate

Π_{TB} is discrete, so the decoder stages downstream of it receive

$$\boldsymbol{A}^{\mathrm{ST}} = P^A + \mathrm{stopgrad}\Big(\widetilde{\boldsymbol{A}} - P^A\Big). \tag{16}$$

The forward value is the binary connected adjacency $\widetilde{\boldsymbol{A}}$, and gradients flow through the soft adjacency P^A. The adjacency head is supervised from its own logits against the ground-truth adjacency, so the head is trained by both paths.

TopoBridge guarantees a connected candidate adjacency $\widetilde{\boldsymbol{A}}$, and bond assignment is confined to that support. Connectivity of the molecule finally returned is measured: the rate is reported in Tables 3 and 4 for de novo samples and in Tables 16 and 17 at the posterior.

C.2.2 Decoder heads

This subsection states in full the decoder stages summarised in Sec. 4.2.

A node projection head f_{node} maps each latent token $\boldsymbol{z}_i$ to an initial decoder feature $\boldsymbol{g}_i^{(0)}$, and the adjacency head f_A scores the pair feature $[\boldsymbol{g}_i^{(0)}, \boldsymbol{g}_j^{(0)}]$ for edge existence:

$$\boldsymbol{g}_i^{(0)} = f_{\mathrm{node}}(\boldsymbol{z}_i), \qquad P_{ij}^A = \sigma\Big(f_A([\boldsymbol{g}_i^{(0)}, \boldsymbol{g}_j^{(0)}])\Big). \tag{17}$$

Because molecular bonds are undirected and self-bonds are invalid, P^A is symmetrized and its diagonal set to zero. TopoBridge then refines it into the connected adjacency $\widetilde{\boldsymbol{A}} = \Pi_{\mathrm{TB}}(P^A)$. The refined adjacency enters TopoFormer as an additive bias $\boldsymbol{e}_A(\widetilde{A}_{ij})$, learned per head, on the scaled dot-product attention logits between atoms i and j.

Given the atom-type distribution $P_i^T = \mathrm{softmax}(f_T(\boldsymbol{g}_i))$, an atom-conditioned feature $\boldsymbol{a}_i^T$ is formed from P_i^T and used in the bond head, which predicts logits over real bond types for each pair:

$$\boldsymbol{o}_{ij}^B = f_B(\boldsymbol{g}_i, \boldsymbol{g}_j, \boldsymbol{a}_i^T, \boldsymbol{a}_j^T, \widetilde{A}_{ij}), \qquad \boldsymbol{o}_{ij}^B \in \mathbb{R}^K. \tag{18}$$

For coordinate reconstruction the coordinate head predicts an initial coordinate matrix $\boldsymbol{R}^{(0)} = f_R(\boldsymbol{g})$ from the final decoder features, the bond distribution is mapped to a continuous edge feature $\boldsymbol{e}_{ij}^B = \sum_{c=0}^K P_{ij,c}^B \boldsymbol{w}_c$ through learned bond embeddings $\{\boldsymbol{w}_c\}$, and L EGNN layers then update features and coordinates jointly under $\boldsymbol{e}^B$, returning $\widehat{\boldsymbol{R}} = \boldsymbol{R}^{(L)}$.

Encoder. UAE and TopVAE use encoders of the same Relational Transformer family, with Gaussian Basis Function distance embedding and message-passing attention. The encoder projects per-atom hidden states to $\boldsymbol{\mu}$ and $\log \boldsymbol{\sigma}^2$ via linear heads, with posterior log-variance clamped to $[-10, 4]$ on GEOM-Drugs.

TopoFormer blocks. The topology-conditioned transformer (Sec. 4.2) uses Graphormer-style additive pair bias from the binary adjacency: bias $=$ `nn.Embedding`$(2, n_{\mathrm{heads}})[\widetilde{A}]$. Each block applies pre-LayerNorm, multi-head self-attention with additive bias, and a GELU FFN ($4\times$ expansion), with residual connections.

EGNN coordinate head. The E(n)-equivariant coordinate prediction uses normalized directions with tanh-bounded magnitude:

$$\Delta \boldsymbol{r}_i = \sum_j \frac{\boldsymbol{r}_i - \boldsymbol{r}_j}{\|\boldsymbol{r}_i - \boldsymbol{r}_j\| + \epsilon} \tanh\big(\phi_x(m_{ij})\big), \qquad \boldsymbol{r}_i \leftarrow \boldsymbol{r}_i + \Delta \boldsymbol{r}_i, \tag{19}$$

where $\epsilon = 10^{-6}$ and ϕ_x is a learned scalar MLP. Coordinates are re-centered after each EGNN layer.

Table 6: Architecture configurations.

Parameter	QM9	GEOM-Drugs
Latent dim d	16	32
Encoder hidden / heads / blocks	64 / 4 / 3	128 / 8 / 4
Decoder (Topo / Geo layers)	3 / 3	4 / 4
Decoder hidden / heads	64 / 4	128 / 8
Edge dim	64	128
Dropout	0.1	0.1
RBF distance features	No	Yes (16 centers)

Table 7: Training hyperparameters for TopVAE (Stage 1: VAE).

Parameter	QM9	GEOM-Drugs
Optimizer	AdamW	AdamW
Learning rate	10^{-3}	3×10^{-4}
Weight decay	10^{-5}	10^{-5}
LR scheduler	Constant	Cosine ($T_{\max} = 1200$)
LR warmup	–	Linear, 50 epochs
Max epochs	2000	1200
Batch size $\times$ GPUs	1024×4	128×4
Precision	fp32	fp32
Gradient clipping	–	1.0
Loss weights (Eq. 13)		
w_{adj}	1.0	10.0
w_{atom}, w_{bond}, λ_R	1, 1, 1	1, 1, 1
λ_D (all-pairs + bonded-pair distance)	0	1.0
λ_{AGCL}	0.1	0.1
β (KL weight)	10^{-6}	5×10^{-6}
TopoBridge		
Threshold τ	0.5	0.3
Max degree cap	4	6 (Considering charges)
Adjacency loss	BCE	Focal ($\alpha = 0.75$, $\gamma = 1.0$)
ChemCO (Sec. 4.4, Appendix S)		
T_{CO} (training)	20	40
Primal base rate α / decay γ_α	0.05 / 0.99	0.05 / 0.99
Penalty-weight base rates η_μ, η_+, η_-	0.1, 0.2, 0.2	0.1, 0.2, 0.2
Active-set sharpness γ	10.0	10.0
Chemical mask Ω	Yes	Yes

Bond type head. For each pair (i, j) on the support $\widetilde{A}$, the bond head concatenates node pair features, learned atom-type embeddings, adjacency embedding, and optionally RBF distance features (16 Gaussian centers spanning $[0.8, 3.0]$ Å on GEOM-Drugs).

Adjacency head. Edge-existence logits are produced by a 2-layer MLP on concatenated node pairs, symmetrized and zero-diagonal before sigmoid.

Architecture summary. Table 6 collects the encoder and decoder configurations used for each dataset: latent width, hidden widths, head counts, block counts, edge dimension, dropout, and whether the bond head receives RBF distance features.

C.3 Training Hyperparameters

Table 7 gives the complete Stage-1 (VAE) configuration for both datasets: optimizer, learning-rate schedule and warmup, epoch budget, batch size and device count, precision and gradient clipping, the loss weights of Eq. (13), the TopoBridge threshold, degree cap and adjacency loss, and the ChemCO step counts, rates and mask setting. Stage-2 (DiT) settings are in Appendix C.6.

C.4 Loss Function Details

The full training objective (Eq. 13) consists of the terms defined below. Let $\boldsymbol{M}_{\text{node}} \in \{0,1\}^{B\times N}$ denote the valid-atom mask and $\boldsymbol{M}_{\text{pair}} = \boldsymbol{M}_{\text{node}}^{(i)} \wedge \boldsymbol{M}_{\text{node}}^{(j)} \wedge \mathbb{1}[i<j]$ the upper-triangular valid-pair mask.

(i) Adjacency loss $\mathcal{L}_{\text{adj}}$. On GEOM-Drugs we use focal binary cross-entropy:

$$\mathcal{L}_{\text{adj}} = \frac{1}{|\boldsymbol{M}_{\text{pair}}|} \sum_{(i,j)\in\boldsymbol{M}_{\text{pair}}} \alpha_t \,(1-p_t)^{\gamma}\,\text{BCE}(s_{ij}, t_{ij}), \tag{20}$$

with $\alpha = 0.75$ (up-weight bond edges) and $\gamma = 1.0$. QM9 uses standard BCE ($\gamma = 0$).

(ii) Atom type loss $\mathcal{L}_{\text{atom}}$. Cross-entropy over predicted atom types, averaged over the valid atoms $\boldsymbol{M}_{\text{node}}$.

(iii) Bond type loss $\mathcal{L}_{\text{bond}}$. Conditional cross-entropy computed only on ground-truth bond edges within the TopoBridge support $\widetilde{A}$:

$$\mathcal{L}_{\text{bond}} = \frac{1}{|\boldsymbol{M}_{\text{bond}}|} \sum_{(i,j)\in\boldsymbol{M}_{\text{bond}}} \text{CE}\big(q_{ij},\, t_{ij}^{\text{bond}} - 1\big), \tag{21}$$

where $\boldsymbol{M}_{\text{bond}} = \boldsymbol{M}_{\text{pair}} \wedge (t_{ij}^{\text{bond}} > 0) \wedge \widetilde{A}$, and $q_{ij} \in \mathbb{R}^K$ are the bond-type logits. The target is shifted by -1 because the bond head predicts over $\{1,\dots,K\}$. Bond absence is handled by the adjacency gate.

(iv) Coordinate loss $\mathcal{L}_{\text{coord}}$. MSE between predicted and ground-truth coordinates, both zero-centered per molecule, averaged over the valid atoms and the three axes.

(v) Distance loss $\mathcal{L}_{\text{dist}}$. This term provides rotation-invariant geometric supervision and is the sum of an all-pairs distance MSE and a bonded-pair distance MSE:

$$\mathcal{L}_{\text{dist}} = \frac{1}{|\boldsymbol{M}_{\text{full}}|} \sum_{(i,j)\in\boldsymbol{M}_{\text{full}}} (\hat{D}_{ij} - D_{ij}^{\text{gt}})^2 \;+\; \frac{1}{|\boldsymbol{M}_{\text{bonded}}|} \sum_{(i,j)\in\boldsymbol{M}_{\text{bonded}}} (\hat{D}_{ij} - D_{ij}^{\text{gt}})^2, \tag{22}$$

where $D_{ij} = \|\boldsymbol{r}_i - \boldsymbol{r}_j\|_2$. The bonded-pair term focuses geometric learning on bond lengths, where distance errors have the largest chemical impact. Active on GEOM-Drugs ($\lambda_D = 1$), inactive on QM9 ($\lambda_D = 0$).

(vi) KL divergence $\mathcal{L}_{\text{KL}}$. The encoder produces per-atom posteriors $q(z_i \mid M) = \mathcal{N}(\mu_i, \text{diag}(\exp(\ell_i)))$ with $\ell_i = \text{clamp}(\log\sigma_i^2, -10, 4)$. $\mathcal{L}_{\text{KL}}$ is the closed-form KL from that posterior to $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$, averaged over atoms and the batch. The near-zero weight ($\beta = 5\times10^{-6}$ on GEOM, 10^{-6} on QM9) yields a near-deterministic autoencoder whose latent space remains structured enough for downstream diffusion.

(vii) AGCL loss $\mathcal{L}_{\text{AGCL}}$. This is the advantage-gated selective teacher loss whose advantage a_b is defined in Sec. 4.5, Eq. (12). Let P_{raw} and P_{chem} denote the raw decoder and ChemCO-projected bond distributions, respectively. $\mathcal{E}_b$ is a bonded-pair score rather than a normalized negative log-likelihood: P_{raw} is normalized over the $K+1$ classes, while ChemCO returns $\sum_{k\geq1} P_{\text{chem},ij}^{(k)} = s_{ij}$ (Eq. 33), the pair mass its exclusivity penalty drives toward at most one. For each molecule b in the batch, the advantage $a_b = [\mathcal{E}_b(P_{\text{raw}}) - \mathcal{E}_b(P_{\text{chem}})]_+$ gates an L_2 consistency loss that steers P_{raw} toward $\text{sg}(P_{\text{chem}})$,

$$\mathcal{L}_{\text{AGCL}} = \frac{\sum_b a_b \sum_{(i,j)\in\mathcal{S}_b} \|P_{\text{raw},ij} - \text{sg}(P_{\text{chem},ij})\|_2^2}{(K+1)\sum_b a_b|\mathcal{S}_b| + \epsilon}, \tag{23}$$

where $\mathcal{S}_b$ is the bonded pair set of molecule b and $K+1$ the number of bond classes. The stop-gradient detaches the ChemCO output so that only the raw decoder is updated. When $\sum_b a_b < 10^{-8}$ (i.e., ChemCO provides no advantage), the loss returns zero. Weight: $\lambda_{\text{AGCL}} = 0.1$ on both datasets.

Table 8: DiT configurations for Stage 2 latent diffusion.

Parameter	QM9	GEOM DiT-B
Hidden dim	512	768
Heads	8	12
Layers	8	12
MLP ratio	$4\times$	$4\times$
Optimizer	AdamW, lr $= 10^{-4}$, weight decay $= 0.05$	
LR schedule	Linear warmup (1k steps) + cosine decay	
Max epochs	10000	10000
Batch size $\times$ GPUs	1024×4	768×4
Precision	fp16-mixed	

Raw bond distribution. P_{raw} is the adjacency gating of Eq. (6) with $\widetilde{A}$ replaced by the straight-through adjacency $\widetilde{A}^{\text{ste}}$ of Eq. (16). Pairs outside the support are forced to $[1, 0, \ldots, 0]$ (no bond).

C.5 GT-Support Warmup Schedule

Motivation. Early in training, the adjacency predictor produces near-random graphs. If the bond-type and coordinate heads must condition on these noisy adjacencies, they receive inconsistent inputs that slow convergence. GT-support warmup decouples this by providing ground-truth topology initially, then transitioning to predicted TopoBridge supports.

Three-phase schedule. Let W_{gt} and W_{tr} be the warmup and transition durations (in epochs). The probability of using ground-truth support is:

$$p_{\text{gt}}(\text{epoch}) = \begin{cases} 1.0 & \text{epoch} < W_{\text{gt}} \quad \text{(Phase A: pure GT)}, \\ 1 - \frac{\text{epoch} - W_{\text{gt}}}{W_{\text{tr}}} & W_{\text{gt}} \le \text{epoch} < W_{\text{gt}} + W_{\text{tr}} \quad \text{(Phase B: linear transition)}, \\ 0.0 & \text{epoch} \ge W_{\text{gt}} + W_{\text{tr}} \quad \text{(Phase C: fully predicted)}. \end{cases} \tag{24}$$

During Phase B, gating is a per-batch Bernoulli coin flip: with probability p_{gt}, the entire batch receives ground-truth adjacency and bond types. Otherwise the batch uses predicted TopoBridge supports. Settings: $W_{\text{gt}} = 100$, $W_{\text{tr}} = 100$ on GEOM-Drugs. QM9 uses $W_{\text{gt}} = 0$.

C.6 Latent Diffusion Model (Stage 2)

The diffusion prior is trained on frozen TopVAE latents. We describe the backbone, noise schedule, and sampling procedure.

DiT backbone. A standard TransformerEncoder with sinusoidal timestep embeddings (base 10000). Timestep embedding is projected through a 2-layer MLP and added to input token embeddings.

Architecture configurations. Table 8 gives the Stage-2 DiT configuration on each dataset.

Noise schedule. VP-SDE with the cosine schedule: $\bar{\alpha}(t) = \cos\left(\frac{t+s}{1+s} \cdot \frac{\pi}{2}\right)^2$ with offset $s = 0.008$.

Latent whitening. Before diffusion training, latent codes from the frozen VAE encoder are whitened: $z_{\text{norm}} = (z - \mu_z)/\sigma_z$, where μ_z, σ_z are computed over all training latents. The inverse transform is applied before decoding at generation time.

Sampling. We use ancestral DDPM reverse sampling with $T = 100$ steps on $t_{\text{array}} = \text{linspace}(1-\epsilon, \epsilon, T)$, $\epsilon = 10^{-3}$. Noise temperature $\tau = 1.0$ on QM9 and $\tau = 0.95$ on GEOM-Drugs, as stated beside each table. No exponential moving average is used.

VAE freezing. The VAE is fully frozen during Stage 2: all parameters are detached, and online encoding runs under `torch.no_grad()`.

C.7 METRICS

Which bond graph each metric is computed on. Every metric in Tables 3 and 4 that we compute is computed on heavy atoms after hydrogen removal, and the two blocks differ only in where the bonds come from. Cells quoted from other papers follow those papers' own conventions. Appendix G states the convention for the published JODO QM9 cells. The *2D* block uses the bond graph the model itself predicts, so it is defined only for methods that produce one. EDM, MDM, GeoLDM,MolFLAE, GFMDiff and ADiT do not, and they are listed in the 3D table only. The *3D* block discards any predicted bonds and re-infers connectivity from the generated coordinates alone, using the distance-lookup rule of EDM (Hoogeboom et al., 2022) as adopted by JODO (Huang et al., 2023): a bond is placed between two atoms when their distance falls within the tabulated range for that element pair and bond order. FCD_{3D} is therefore the Fréchet ChemNet distance between the reference set and the SMILES obtained from that coordinate-derived graph, and it is undefined for a generated molecule whose coordinate-derived graph does not sanitize. Such molecules are excluded from the FCD population. Atom and molecular stability in each block use the same valence table, applied to whichever graph that block specifies.

Stability, connectivity, and the conjunctions V&C and S∧C. An atom is stable when its summed bond order equals an entry in the reference valence set for its element, and a molecule is stable when every one of its atoms is. The check is run after `Chem.AddHs(addCoords=True)`, so hydrogens saturate every heavy atom before valences are summed, and a heavy atom left without a heavy-atom neighbor is scored as its hydride. Connectivity is evaluated on the heavy-atom graph before hydrogens are added and requires exactly one connected component. No largest-fragment extraction is applied anywhere in this paper.

Estimators and populations. *San.* is RDKit sanitization of the decoded graph alone. *Conn.* is the single-connected-component rate. *Fully bonded* is the condition that every atom of the decoded graph carries at least one bond. *AtomStab* and *MolStab* are atom and molecular stability against the reference valence set, summed after `Chem.AddHs(addCoords=True)`. V&C is *San.* ∧ *Conn.*, V&U is *San.* ∧ uniqueness, and S∧C is *MolStab* ∧ *Conn. VCB* is *San.* ∧ *Conn.* ∧ fully bonded, the quantity the perturbation sweeps describe as valid, connected and fully bonded. *Joint* is the scaffold-completion criterion of Table 25: the conditioning scaffold preserved, the product valence-legal, and the product a single connected component. *Recall* and *Prec.* are micro bond recall and precision over heavy-heavy atom pairs. *Eligible* is *San.* ∧ *Conn.* ∧ finite 3D coordinates.

iFID Computation Details. Our iFID metric adapts the interpolated FID from pixel space to the latent space of molecular VAEs (Xu et al., 2026). The pipeline proceeds in four stages: encoding, interpolation, decoding, and scoring.

Encoding. Every molecule x in the training set $\mathcal{D}_{\text{train}}$ or validation set $\mathcal{D}_{\text{val}}$ is mapped to its deterministic posterior mean $\mu_\phi(x)$ by a frozen encoder with all augmentations disabled (no rotation, translation, or noise).

Interpolation. Each dense latent $z \in \mathbb{R}^{N_{\max} \times d}$ is mean-pooled along the node dimension to $\bar{z}$. For each validation molecule x_v, we retrieve the top-K=10 neighbors by cosine similarity from training molecules with the *same* atom count $|x|=|x_v|$, avoiding size-mismatch artefacts. One neighbor z_t is drawn categorically with weights

$$\pi_k = \text{softmax}\big(-\|z_v - z_t^{(k)}\|_2^2\big).$$

The midpoint latent is then constructed via spherical linear interpolation:

$$z_\alpha = \text{SLERP}(z_v,\, z_t,\, \alpha), \qquad \alpha = 0.5,$$

computed on the node-level flattened vectors, falling back to LERP when $\sin\theta < 10^{-6}$.

Decoding. The interpolated latent z_α is decoded by a frozen decoder following each model's standard protocol, assembled into RDKit molecules, and converted to canonical SMILES (`RemoveAllHs`, `MolToSmiles(isomericSmiles=False)`).

Scoring and invalid-molecule handling. FID features are 512-d ChemNet activations (`fcd_torch`), from which we estimate means and covariances for the reference and the interpolated sets and

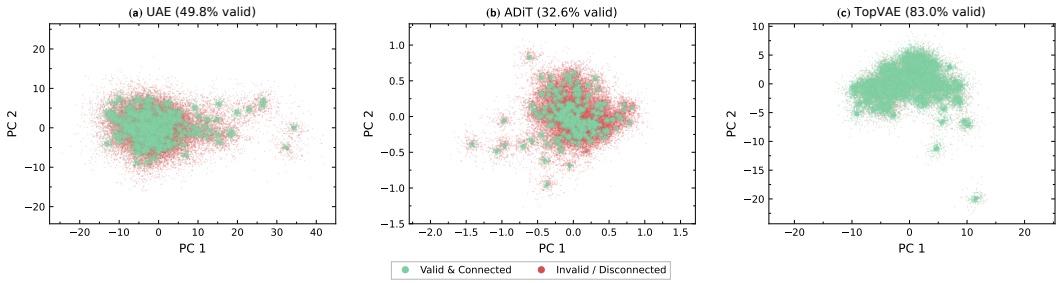


Figure 4: **VAE dark-area embeddings.** 2-D PCA of the pooled perturbed latents of Appendix D.1. Green points decode to valid and connected molecules, red points do not. Axis ranges are set per panel and are not shared across panels.

compute the Fréchet distance:

$$\mathrm{FID} = \|\mu_1 - \mu_2\|_2^2 + \mathrm{Tr}\big(\Sigma_1 + \Sigma_2 - 2(\Sigma_1\Sigma_2)^{1/2}\big).$$

iFID is this distance evaluated on one pair: set 1 is the reference set, the ground-truth validation molecules, and set 2 is the decoded interpolated set. Decoding failures yield empty strings. A subsequent `MolFromSmiles(sanitize=True)` filter removes unsanitizable or overly long (>350 character) SMILES. Only molecules that survive this filter enter ChemNet. This is the convention `fcd_torch` enforces, and it is applied identically to every row of Table 1.

D Extended Dark-Area Analysis

D.1 VAE Latent Space Geometry

We randomly sample 200 QM9 validation molecules and encode each through three VAEs (UAE, TopVAE, ADiT). For each latent code $z = \mu_\phi(M)$ we add Gaussian perturbations $z' = z + \sigma \cdot s_{\mathrm{emp}} \cdot \varepsilon$ at $\sigma \in \{0, 0.05, 0.1, 0.2, 0.3, 0.5\}$ (50 draws per nonzero σ), decode each z', and score RDKit sanitization and single-fragment connectivity. Pooled over every nonzero level of this grid, 83.0% of TopVAE's perturbed latents decode to valid and connected molecules, against 49.8% for UAE and 32.6% for ADiT. Figure 4 projects those pooled latents.

D.2 Latent Scale of Each Autoencoder

Table 9 gives s_{emp} for the reported checkpoints of both datasets. The QM9 ADiT-VAE entry is measured on the checkpoint that produced that row: $\bar{s}_{\mathrm{emp}} = 0.6905$ over its eight dimensions, on a 2,000-molecule subset of the same split. Its per-dimension range is not reported. The GEOM-Drugs TopVAE entry is measured on the released GEOM-Drugs checkpoint.

Table 9: **Latent scale of each autoencoder.** Per-dimension standard deviation s_{emp} of the posterior means, and d the latent dimension. QM9 n=10,000 (ADiT-VAE n=2,000). GEOM-Drugs n=1,997.

			s_{emp}			
Dataset	Model	d	mean	min	median	max
QM9	UAE	16	18.03	5.80	17.44	31.72
QM9	ADiT-VAE	8	0.69	–	–	–
QM9	TopVAE	16	6.18	1.34	6.55	12.24
GEOM-Drugs	UAE	16	9.77	7.60	8.98	14.77
GEOM-Drugs	TopVAE	32	1.38	0.37	0.90	5.78

D.3 THE NOISE-SCALED DIAGNOSIS ON QM9

Table 11 gives the noise-scaled diagnosis at the posterior and at σ=0.3, 0.5 and 1.0, on the capacity-matched QM9 pair. The two decoders are within 0.1 points of each other at σ=0, and TopVAE leads at every nonzero level. Table 10 gives the matched-absolute cells that Figure 2 plots.

Table 10: **Cross-model dark-area diagnosis at matched absolute displacement, cell by cell** (QM9, n=10,000, on the checkpoints and subset of Table 1, one run per arm). Entries are VCB. δ is as in Figure 2, which plots these cells. Higher in bold. δ=0 is the reconstruction anchor.

δ	0	0.25	0.5	1	1.5	2	3	4	5	6	8	10	14	20	30
UAE	0.9999	**0.9999**	**0.9998**	**0.9980**	**0.9851**	**0.9306**	**0.6692**	0.3793	0.1980	0.1052	0.0288	0.0092	0.0018	0.0011	0.0003
TopVAE	0.9994	0.9995	0.9992	0.9925	0.9627	0.8823	0.6391	**0.4536**	**0.3480**	**0.2748**	**0.2182**	**0.1856**	**0.1587**	**0.1520**	**0.1317**

Conditioning moves the latents outward. A scaffold is supplied as posterior codes and the remaining heavy-atom slots are filled by reverse diffusion, on the same QM9 checkpoint and axis as the grid above. Each decoded molecule is re-encoded and its displacement taken as $\delta = \|\boldsymbol{z} - \boldsymbol{\mu}_\phi(\hat{M})\|_F/\sqrt{16\,n}$, the per-coordinate unit the grid uses. The round trip is exact to 1.4×10^{-6}. Entries below are relative to unconditional sampling on the same checkpoint.

	product size	median δ ($\times$ unconditional)	p_{95} ($\times$ unconditional)
unconditional	—	1.00	1.00
benzene +1	7	2.11	1.49
benzene +3	9	2.39	1.94
pyridine +3	9	2.47	1.97

Each row is 500 completions at full coverage. The shift is monotone in the expansion, and it is the displacement of the conditioned tokens rather than of the filled slots that grows.

Table 11: **The noise-scaled diagnosis on QM9** (Eq. (14), n=10,000, on the checkpoint and subset of Table 1, one run per arm). Entries are the VCB estimator of Appendix C.7, the one the σ=0.5 column of Table 1 reports. The σ=0 column is the reconstruction anchor and is not marked. Higher value in bold in the others.

	σ=0	σ=0.3	σ=0.5	σ=1.0
QM9				
UAE	0.9999	0.1260	0.0142	0.0009
TopVAE	0.9994	**0.7571**	**0.4599**	**0.2780**

D.4 DIFFUSION DENOISING TRAJECTORY

Table 12 reports the decoded validity of both models at 1,024 trajectories each on the sampler's own step grid. The baseline stays at 0.000 over the noisy half of the trajectory and reaches 0.262 at t=0.102. TopVAE+DiT is at 0.103 at t=0.999, where the latent is pure noise, and 0.386 at t=0.253.

1. **Latent spread.** UDM-3D's intermediate latents span a PC 1 range of $\sim$300 units early in denoising, reflecting high variance in the unconstrained latent space. TopVAE+DiT operates within a $\sim$60-unit range, and ADiT within $\sim$1 unit.
2. **Validity transition.** UDM-3D's validity ratio stays near zero over the noisy half of the trajectory and rises only as denoising completes. The axis is the sampled progress grid rather than a linear one, and no point is sampled between 0.50 and 0.75. TopVAE+DiT decodes valid molecules from the noisy latent throughout, and its intermediate molecules are already chemically plausible at early denoising stages.

Figure 5 shows two of those trajectories decoded at five points.

Table 12: **Decoded validity along the denoising trajectory** (QM9, 1,024 trajectories per model, seed 42, on the sampler's own grid, with t=1 pure noise and t=0 the finished sample). A decode counts as valid if it sanitizes and is a single connected component, the criterion of Sec. 3. The two converge only once denoising is essentially complete.

t	0.999	0.747	0.495	0.253	0.102	0.051	0.011	0.001
UAE+UDM	0.000	0.000	0.003	0.007	0.262	0.828	1.000	0.999
TopVAE+DiT	**0.103**	**0.114**	**0.129**	**0.386**	**0.925**	**0.983**	0.990	0.991

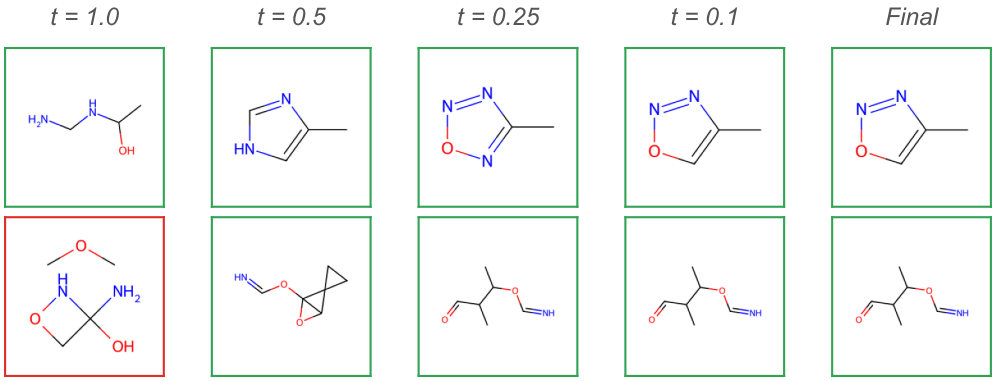


Figure 5: **Two QM9 sampling trajectories decoded at five points** (TopVAE+DiT, seed 42). Each row is one trajectory and each column one point on the sampler's grid, from pure noise at t=1.0 to the finished sample. A green border marks a decode that sanitizes and is a single connected component and a red border one that does not. The upper trajectory meets that criterion at every point and the lower one from t=0.5 onward. Table 12 gives the rate over 1,024 trajectories.

In a batch of 16 decoded denoising trajectories per model, thirteen of the 16 UDM-3D trajectories decode at t=1.0 to valences so far out of range that RDKit cannot produce a 2D depiction. These count as invalid. No TopVAE+DiT trajectory is undepictable at any timestep.

E Reproduction Ledger

UDM-3D's authors release source code but no trained checkpoint, for either dataset. The request for one on their repository is open and unanswered.[1] Every UDM-3D number we produce is therefore a re-training from that repository at its default configuration, and we use the paper's own published values wherever they exist (Table 13).

Published values and our own evaluations. Tables 3 and 4 use *published* values for every baseline except the cells marked †, which are our own evaluation of released samples. Table 14 names the source of every row. Our own re-training of UDM-3D is used only where no published value exists for the quantity in question.

Provenance of the ADiT-VAE row of Table 1. That row is our own evaluation. ADiT's paper reports no posterior or perturbed-latent decodability for its autoencoder, so there is no published value to set beside it. The autoencoder is our own training of the official ADiT implementation, run at that repository's `qm9_only` autoencoder configuration and through its own QM9 data pipeline rather than the preprocessing of Appendix C.1. It is not a checkpoint we obtained already trained. The posterior, $\bar{s}_{\text{emp}}$ and σ entries of this row come from one run, measured on a 2,000-molecule subset rather than the 10,000-molecule subset the other QM9 rows use, as the captions of Table 1 and Table 9 state. The ADiT-VAE row of Table 16 is a separate evaluation of the same checkpoint at n=10,000, on the one subset the UAE and TopVAE rows of that table also use.

[1] `https://github.com/lyc0930/UAE-3D`

Table 13: **Reproduction ledger.** Every entry we produced ourselves for a method that is not ours, set against the value its own paper reports. A dash means the quantity is not reported by that paper. Every entry here is a re-training from the official source.

Method	Dataset	Quantity	Published	Ours	Status
UAE (VAE)	QM9	posterior V&C	–	1.000	reproduces
UAE (VAE)	GEOM	posterior V&C	–	0.942	reconstructs
UDM-3D	QM9	FCD (2D)	0.130	0.207	re-trained, $+0.077$ against published
UDM-3D	GEOM	FCD (2D)	0.692	25.73	does not reproduce
UDM-3D	GEOM	valid molecules	–	1,823 / 10,000	does not reproduce

F Provenance of the Baseline Numbers in Tables 3 and 4

Two papers re-score this benchmark for the field, and most published 3D molecular generators do not report every metric in it. Table 14 therefore records, per row, where each cell comes from. Four conventions govern the table. First, a method that produces no predicted bond graph has no native 2D-block quantity, and appears in the 3D table only. Six rows are of this kind: EDM, MDM, GeoLDM, MolFLAE, GFMDiff and ADiT. Each was classified from the method's own paper and from the output head of its official implementation, not from its description elsewhere. Second, FCD is reported by neither EDM, GeoLDM, MiDi, EQGAT-diff nor SemlaFlow in any version. Every FCD cell for those methods originates with JODO or UDM-3D. Third, V&U and V&U&N are adjacent columns in both JODO's and UDM-3D's tables, and V&U is the one reported here. Fourth, each cell is attributed to a single named source rather than merged across papers.

F.1 De Novo Sample Sets Used in Table 22

Table 15 records the sample set for each row of the stereochemistry comparison. Its LDM-3DG GEOM-Drugs row is also the sample set used for that method's row in Table 4. JODO's rows are its released `rdkit_mols/*_ancestral_ckpt_*.pkl` samples, verified md5-identical to that release, and LDM-3DG's are the de novo samples in its Zenodo release. `RemoveAllHs` is applied identically to every sample set before every comparison.

LDM-3DG (You et al., 2024) is not in Table 22. Its release runs MMFF94 optimization on every conformer before writing it. Re-running MMFF94 from the released QM9 coordinates moves them by a median maximum per-atom displacement of 0.00000 Å, with 98.3% of molecules moving by less than 0.01 Å, against 0.65 Å for TopVAE+DiT and 0.30 Å for JODO. A geometry statistic computed on coordinates that already sit at an MMFF94 minimum measures the force field's minimum rather than the generator's output.

G Three Methods Scored From Their Released Samples

TopVAE+DiT, JODO (Huang et al., 2023) and LDM-3DG (You et al., 2024) are scored here through one pipeline, applied identically to every method, from the samples each release ships. `RemoveAllHs` is applied to every sample set first, so all three are compared on heavy atoms, and the metrics are the EDM/JODO definitions used throughout this paper, with the full released sample count as the denominator and every failure counted as a failure. Hydrogens are then completed by RDKit for every method before the 3D metrics are computed, the convention those metrics assume. A bond-order normalization is applied to all three before scoring. It modified 45 of JODO's QM9 molecules and 28 of its GEOM-Drugs molecules, and none of TopVAE+DiT's. TopVAE+DiT is the seed-42 run, whereas the TopVAE+DiT cells of Tables 3 and 4 are three-seed means. Both TopVAE+DiT runs are sampled at the noise temperature reported beside Tables 3 and 4. The FCD reference is built from the test split under the same pipeline as the generated side. On GEOM-Drugs it is the $14{,}120$ unique molecules that the $30{,}429$-molecule JODO-pipeline test split of Table 15 yields under the coordinate-derived bond construction. On QM9 it is the heavy-only test split rebuilt the same way. Table 15 records the sample sets.

LDM-3DG's release uses one GEOM-trained autoencoder for both benchmarks, and its QM9 samples are off-support: 58.5% of them exceed QM9's maximum of nine heavy atoms, and it produces twelve

Table 14: **Per-row provenance for Tables 3 and 4.** *own* denotes the method's own paper. *JODO* and *UDM-3D* denote a value those papers report for that method, in several cases a reproduction they ran with official code.

Method	Dataset	Source	Note
Training set	QM9	UDM-3D	Reference row. Stability and joint-criterion columns from UDM-3D Tab. 2, both blocks. JODO Tab. 1 reports the identical values. Its FCD cells are left empty.
EDM	QM9	JODO (3D)	No predicted bond graph and no FCD in any version. 2D cells removed. The same column also carries 0.943 for the implicit-hydrogen row.
MDM (Huang et al., 2022)	QM9	UDM-3D	UDM-3D Tab. 2, 3D block. Reports no 2D-block quantity.
CDGS	QM9	JODO	2D-only model, so the 3D block is not applicable. V&C and V&U corrected here from a one-column shift against UDM-3D's row, whose V&U and V&U&N were taken as V&C and V&U.
GeoLDM	QM9	UDM-3D (3D)	No predicted bond graph and no FCD. 2D cells removed.
LDM-3DG (You et al., 2024)	QM9	ours	Every cell is our own evaluation of all 10,000 released samples and is marked †. The method itself decodes a bond graph, so it is listed with the bond-graph methods. The release uses one GEOM-trained autoencoder for both benchmarks, and its QM9 samples are off-support: 58.5% of them exceed QM9's maximum of nine heavy atoms and it produces twelve element types against QM9's four, so this row is measured on a different population from the others (Appendix G).
MolFLAE (Chen et al., 2025)	QM9	own	Configuration `MolFLAE 100`. Stabilities are coordinate-derived.
GFMDiff	QM9	own	Stabilities are coordinate-derived.
MiDi	QM9	UDM-3D	Reproduced by UDM-3D with official code. MiDi reports no FCD.
JODO	QM9	own	The only external baseline whose full row is self-reported. Its two FCD cells are marked † and are our own evaluation of its released samples.
EQGAT-diff	QM9	UDM-3D	Reproduced with official code. EQGAT-diff reports no FCD.
ADiT (Joshi et al., 2025)	QM9	UDM-3D	UDM-3D Tab. 2, 3D block. Its molecular-stability cell is not reported there.
SemlaFlow	QM9	UDM-3D	Every cell in this row is UDM-3D's. SemlaFlow reports no FCD in any version. Its AISTATS camera-ready gives AtomStab 99.9, MolStab 99.7, Validity 99.4, Uniqueness 95.4 (97.5 in the preprint versions) on MiDi and EQGAT-diff's explicit-hydrogen splits. That Validity is "not capable of flagging disconnected molecules", not a V&C.
UDM-3D	QM9	own	
Training set	GEOM	UDM-3D	Reference row. Stability and joint-criterion columns, both blocks, from UDM-3D Tab. 1. JODO Tab. 2 reports the identical values. Its FCD cells are left empty.
EDM	GEOM	JODO (3D)	No predicted bond graph. Listed in the 3D block only.
CDGS	GEOM	JODO	
GeoLDM	GEOM	UDM-3D (3D)	2D cells removed, same search. GeoLDM reports no DRUG molecule-stability value.
LDM-3DG (You et al., 2024)	GEOM	ours	The one measured row in Table 4: our own evaluation of the 9,995 de novo samples in that method's Zenodo release, through the code path of Appendix G. Its two FCD cells are marked †: they are our own evaluation on the rebuilt reference, not that paper's published values. Its released coordinates are MMFF94-relaxed.
MiDi	GEOM	UDM-3D	Reproduced with official code.
JODO	GEOM	own	V&U corrected here from the adjacent V&U&N column. Its two FCD cells are marked † and are our own evaluation of its released samples.
EQGAT-diff	GEOM	UDM-3D	Reproduced with official code.
SemlaFlow	GEOM	own	Stabilities are its AISTATS camera-ready Tab. 2, 99.8 and 97.3, on the bond graph the model generates (97.7 and 95.2 in the preprint versions). No FCD and no coordinate-derived metric on this dataset. Its Validity carries no connectivity clause, so it is not a V&C. V&U is that Validity, 93.9, left unchanged by Uniqueness 100.0.
UDM-3D	GEOM	own	V&U corrected here from the adjacent V&U&N column.

element types against QM9's $\{\mathrm{C}, \mathrm{N}, \mathrm{O}, \mathrm{F}\}$. Only 35.0% of its $10{,}000$ QM9 samples are within nine heavy atoms and drawn from those four elements, and the heavy-atom count runs to 72 against a mean of 10.42. Its QM9 cells marked † in Table 3 are our own evaluation of all $10{,}000$ released samples, through the same code path as the other † cells, so they are measured on that population rather than on the one the other rows use. Its GEOM-Drugs samples are on support: mean heavy-atom count 25.88 against the test set's 24.90, and six atoms outside the GEOM-Drugs vocabulary in $259{,}000$. Its released coordinates are MMFF94-relaxed.

Figure 6 is a gallery of unconditional TopVAE+DiT samples.

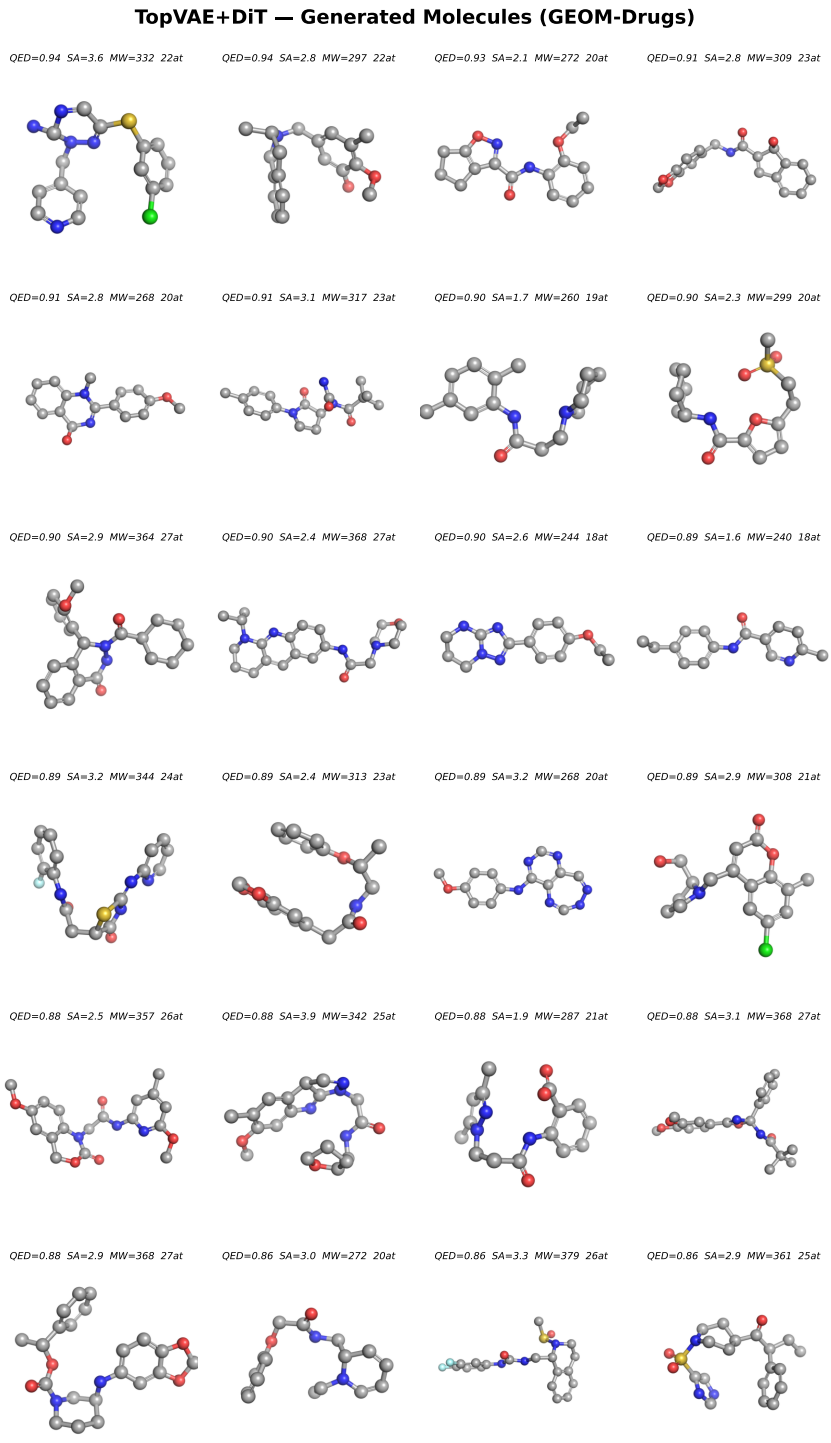


Figure 6: **Unconditional TopVAE+DiT sample gallery on GEOM-Drugs.** Twenty-four decoded conformers from the run of Table 4, with drug-likeness, synthetic accessibility, molecular weight and heavy-atom count. Selected by descending drug-likeness under a scaffold-diversity filter, so the panel is the top of that ordering rather than a random draw.

Table 15: **Sample sets used in Table 22.** *eligible* is the subset that is sanitizable, connected and carries 3D coordinates, which is the population every comparison in that table is restricted to. Source hydrogens records what the released file contains. `RemoveAllHs` is applied to every sample set before every comparison, so all are compared on heavy atoms.

Dataset	Sample set	n	eligible	Source hydrogens	Mean heavy atoms
QM9	TopVAE+DiT	10,000	9,978	heavy-only as generated	8.79
	JODO	10,000	9,912	explicit, stripped	8.85
	LDM-3DG	10,000	9,999	explicit, stripped	10.42
	test set	13,083	12,892	explicit, stripped	8.80
GEOM-Drugs	TopVAE+DiT	10,000	9,454	heavy-only as generated	24.78
	JODO	10,000	8,789	explicit, stripped	25.08
	LDM-3DG	9,995	9,984	explicit, stripped	25.88
	test set	30,429	30,381	explicit, stripped	24.90

H VAE Posterior Reconstruction Quality

Tables 16 and 17 report VAE posterior reconstruction quality on QM9 and GEOM-Drugs respectively. These results complement the dark-area diagnosis in Section 5.2 by establishing what both autoencoders do at the posterior. On GEOM-Drugs the two are indistinguishable there: every one of the 1,997 reconstructions is a single connected component in both, at molecular stability 0.907 and 0.906, and the decoders recover the reference bond set at recall 1.0000 and precision 1.0000. On QM9 they are indistinguishable as well, at 1.000 against 1.000. The QM9 separation reported in Sec. 5.2 is therefore not a reconstruction gap. It appears only once the latent is displaced from the posterior. Each row of both tables is scored on what its own decoder produces.

Table 16: **VAE posterior reconstruction on QM9** (n=10k validation set, all rows are our own evaluation). Each row is scored on what its own decoder produces.

Method	FCD↓	Atom Stab.↑	Mol Stab.↑	V&C↑
UAE	0.087	1.000	1.000	1.000
ADiT-VAE	0.090	1.000	0.960	0.921
TopVAE	0.087	1.000	1.000	1.000

Table 17: **VAE posterior reconstruction on GEOM-Drugs** (n=1997 validation set, all rows are our own evaluation). Each row is scored on what its own decoder produces. They agree to 0.001 on every column, and both return every molecule as a single connected component.

Method	FCD↓	Atom Stab.↑	Mol Stab.↑	V&C↑
UAE	2.633	0.966	0.907	0.942
TopVAE	2.633	0.966	0.906	0.942

I Sampling Variability of the Headline Runs

Tables 3 and 4 report the mean over three sampling seeds. This appendix gives the standard deviation alongside it (Table 18), and then the spread over five independent *training* runs of the baseline autoencoder.

Training-seed replication of the baseline. The spread above is over sampling. To separate it from the variation introduced by training, we trained the baseline autoencoder five times from scratch under its published recipe, changing only the training seed, and swept each of the five on the grid of Figure 2 under the identical protocol (n=10,000, the same molecule subset and the same perturbation stream). Table 19 reports that spread beside TopVAE. The same table carries a second TopVAE

Table 18: **Three-seed sampling variability** of the runs reported in Tables 3 and 4 (seeds 42, 43 and 44, at n=10,000 each). Mean ± standard deviation.

	Atom Stab.↑	Mol Stab.↑	V&C↑	V&U↑	Atom Stab.↑ (3D)	Mol Stab.↑ (3D)
QM9	1.0000 ± 0.0000	1.0000 ± 0.0000	0.9984 ± 0.0004	0.9662 ± 0.0011	0.9907 ± 0.0004	0.9276 ± 0.0026
GEOM-Drugs	0.9995 ± 0.0001	0.9788 ± 0.0017	0.9504 ± 0.0030	0.9999 ± 0.0002	0.8325 ± 0.0002	0.0205 ± 0.0012

training under the same recipe, differing in seed and in how its batches were divided across devices, above the first TopVAE training at every $\delta \geq 3$. The table's first row is the same measurement at δ=0, the reconstruction anchor, where the seven trainings span five decoded graphs in ten thousand, while at δ=8 they span 0.0266 to 0.2559.

Table 19: **Training-seed spread of the baseline autoencoder** on the matched-absolute grid (five independent trainings on QM9, seeds 42, 777, 1234, 2024 and 3407, at n=10,000 each, under the estimator of Figure 2). The two TopVAE columns are independent trainings under one recipe. From δ=5 to δ=30 the weaker TopVAE training exceeds the baseline's five-seed maximum, by 3.6 to 354 times the baseline's training-seed standard deviation.

δ	Baseline mean ± sd	[min, max]	TopVAE	TopVAE 2
0	0.9999 ± 0.0000	$[0.9999, 0.9999]$	0.9994	0.9999
0.5	0.9999 ± 0.0000	$[0.9999, 0.9999]$	0.9992	0.9996
1	0.9990 ± 0.0009	$[0.9977, 0.9998]$	0.9925	0.9917
1.5	0.9826 ± 0.0096	$[0.9703, 0.9936]$	0.9627	0.9503
2	0.9262 ± 0.0289	$[0.8942, 0.9632]$	0.8823	0.8646
3	0.6666 ± 0.0667	$[0.6013, 0.7557]$	0.6391	0.6498
4	0.3894 ± 0.0640	$[0.3310, 0.4843]$	**0.4536**	**0.4860**
5	0.2029 ± 0.0403	$[0.1651, 0.2599]$	**0.3480**	**0.3758**
6	0.1070 ± 0.0274	$[0.0817, 0.1479]$	**0.2748**	**0.3109**
8	0.0347 ± 0.0103	$[0.0266, 0.0487]$	**0.2182**	**0.2559**
10	0.0144 ± 0.0048	$[0.0099, 0.0215]$	**0.1856**	**0.2335**
14	0.0044 ± 0.0015	$[0.0028, 0.0068]$	**0.1587**	**0.1957**
20	0.0012 ± 0.0007	$[0.0004, 0.0021]$	**0.1520**	**0.1850**
30	0.0007 ± 0.0004	$[0.0003, 0.0013]$	**0.1317**	**0.1718**

Every stability and connectivity figure here is stable to within 0.003.

J Molecular Property Distributions

To assess whether TopVAE+DiT generates molecules that faithfully reproduce the training distribution beyond aggregate FCD scores, we compute 8 physicochemical descriptors on the generated set (n=10,000) and compare them to the reference. No public checkpoint has been released for UDM-3D. Tables 3 and 4 therefore quote UDM-3D's published numbers. *Every other* UDM-3D number in this paper, in this appendix and in Appendices D, K, Q and M, and the sentences in the main text that cite them, is produced by our own re-training with the official repository and its default configuration.

Analysis. On GEOM-Drugs, TopVAE+DiT matches the reference distribution more closely than UDM-3D on seven of eight properties. The improvement is particularly pronounced for molecular weight (1.1% deviation vs. 9.6%), heavy-atom count (exact match vs. 10.8% undercount), and ring count (9% deviation vs. 26%): the re-trained UDM-3D run evaluated here returns structurally simpler molecules on this larger-molecule benchmark. TopVAE+DiT also achieves 100% validity (10,000/10,000), compared to 1,823/10,000 for UDM-3D.

On QM9 both models sit at the reference values for molecular weight, QED and heavy-atom count. TopVAE+DiT achieves perfect validity (10,000 vs. 9,986). Figures 7 and 8 give the full distributions

Table 20: **Property statistics on GEOM-Drugs** (n=10,000). Reference values computed on the GEOM-Drugs test set. Bold marks TopVAE+DiT values closer to the reference than UDM-3D in the property rows, and the higher validity yield in the last row.

Property	Dataset Ref.	UDM-3D	TopVAE+DiT
MW (Da)	355.5	321.5	**351.5**
logP	2.86	2.41	**2.91**
QED	0.646	0.530	**0.668**
HeavyAtoms	24.9	22.2	**24.9**
Rings	3.00	2.22	**3.27**
Fsp3	0.305	0.537	**0.334**
TPSA (Å^2)	73.9	67.2	**68.2**
RotBonds	5.05	4.88	4.75
Valid molecules	—	1,823 / 10,000	**10,000** / 10,000

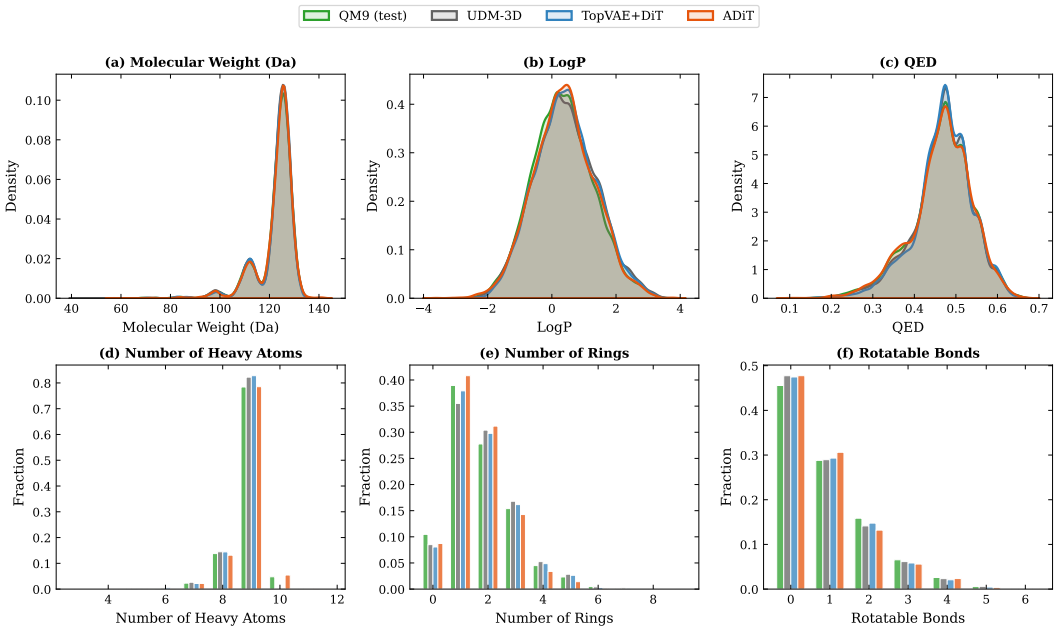


Figure 7: **Property distributions on QM9** (n=10,000). Green is the test-set reference, blue TopVAE+DiT, gray UDM-3D as re-trained by us, and orange ADiT. All three track the reference on molecular weight, QED and heavy-atom count.

summarised here, on QM9 and on GEOM-Drugs respectively. Table 20 gives the GEOM-Drugs statistics numerically.

K Energy Relaxation Evaluation

To measure how close the generated conformers sit to a force-field minimum, we conduct an MMFF94 energy-relaxation study on GEOM-Drugs. For each generated molecule, we compute the MMFF94 single-point energy of the generated conformer, run a local geometry optimization to the nearest force-field minimum, and report (i) the relaxation energy $\Delta E = E_{\text{generated}} - E_{\text{relaxed}}$ (lower indicates the generated geometry is closer to equilibrium) and (ii) the heavy-atom RMSD between the generated and relaxed conformers.

Results. Table 21 reports MMFF94 relaxation statistics for TopVAE+DiT, UDM-3D, and reference molecules from the GEOM-Drugs test set (n=1,994 per source). TopVAE+DiT generates 3D

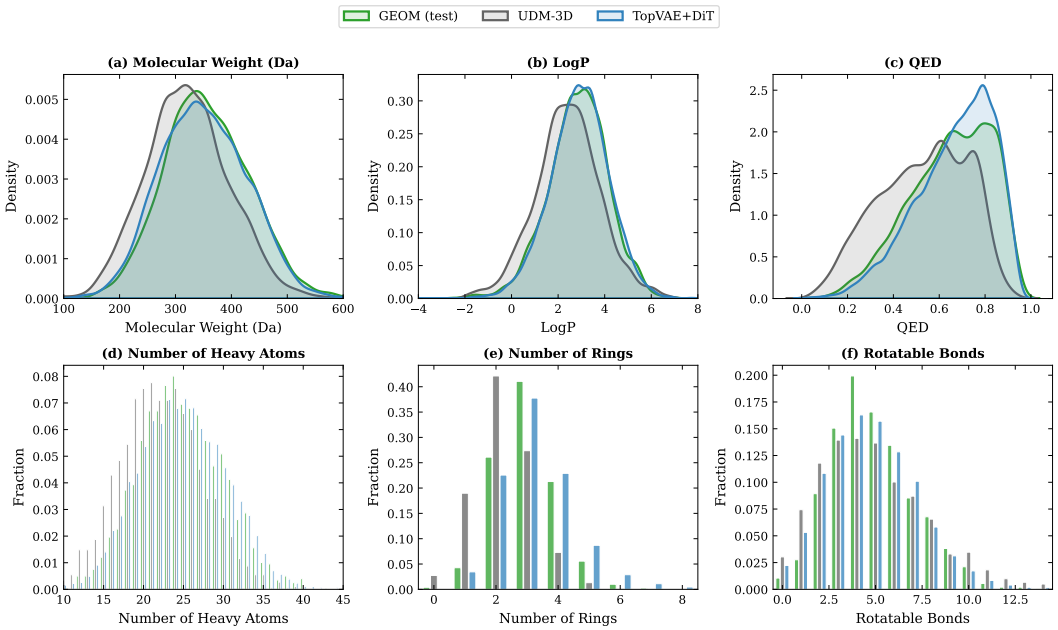


Figure 8: **Property distributions on GEOM-Drugs** (n=10,000). Green is the test-set reference, blue TopVAE+DiT, and gray UDM-3D as re-trained by us. TopVAE+DiT closely tracks the reference across the six properties shown. Table 20 gives all eight. UDM-3D exhibits a systematic leftward shift in molecular weight and heavy-atom count, underrepresentation of three- and four-membered ring systems, and a bimodal QED distribution, consistent with this re-trained run's tendency to decode structurally simpler molecules.

Table 21: MMFF94 energy relaxation on GEOM-Drugs (n=1,994 per source). ΔE is the energy difference between the generated and locally optimized conformer (lower is better). RMSD is the heavy-atom root-mean-square deviation between the two conformers. The UDM-3D column is our own re-training (no public checkpoint, see Appendix J).

Metric	Reference	TopVAE+DiT	UDM-3D
ΔE mean (kcal/mol) ↓	15.9	**40.6**	104.4
ΔE median (kcal/mol) ↓	13.6	**19.8**	76.8
RMSD mean (Å) ↓	0.92	**1.14**	1.24
RMSD median (Å) ↓	0.75	**0.99**	1.12

geometries substantially closer to force-field minima than our own re-training of UDM-3D: the median relaxation energy is 3.9× lower (19.8 vs. 76.8 kcal/mol), and the median RMSD is 12% lower (0.99 vs. 1.12 Å). The Reference column reports the same statistics for experimental GEOM-Drugs conformers under an identical protocol (median $\Delta E = 13.6$ kcal/mol), and fixes the scale of the gap.

L Stereochemistry of the Generated Coordinates

Stereochemistry is measured from the generated coordinates. No release in the 3D-generative literature writes wedge or parity information, so the coordinates are the only source. Table 22 assigns it with RDKit's `AssignStereochemistryFrom3D` and reports the fraction of molecules carrying at least one specified stereo element, counting a tetrahedral stereocenter and a double-bond configuration alike.

The reference sets check the assignment itself, since their stereochemistry is already annotated in the deposited files: as shipped and assigned from 3D agree, 72.99% against 74.60% on QM9 and

41.67% against 41.59% on GEOM-Drugs, so the procedure reproduces the stored annotation rather than inventing stereocenters.

Table 22: **Fraction of molecules carrying at least one specified stereo element**, a tetrahedral stereocenter or a double-bond configuration, assigned from the 3D coordinates with `AssignStereochemistryFrom3D`. Heavy atoms throughout. Sample sets as in Table 15.

	QM9	GEOM-Drugs
TopVAE+DiT	78.95%	54.86%
Test set (reference)	74.60%	41.59%
JODO	70.70%	46.24%

M Diversity and Novelty

Tables 24 and 23 report diversity and novelty metrics on both datasets. Internal diversity is computed as $1 - \overline{\text{Tanimoto}}$ over Morgan fingerprints (radius 2, 2048 bits, up to 500k sampled pairs). Novelty is the fraction of unique SMILES absent from the training set. Unique scaffolds are Murcko scaffolds of unique molecules.

QM9 analysis. On QM9, internal diversity is effectively tied (0.914 vs. 0.916) and so is uniqueness (96.7% vs. 96.6%). TopVAE+DiT produces 3,477 unique Murcko scaffolds versus 2,864 for this re-training of UDM-3D, a 21% increase, and achieves higher novelty (45.1% vs. 42.3%).

GEOM-Drugs analysis. TopVAE+DiT achieves high internal diversity (0.874), near-perfect novelty (99.95%), and 88.7% unique Murcko scaffolds across all 10,000 generated molecules. UDM-3D's internal diversity (0.898) and novelty (100%) are metrics computed on only 1,823 valid molecules. The remaining 81.8% fail RDKit sanitization.

N Scaffold Inpainting: Per-Scaffold Results

Table 26 provides the per-scaffold breakdown of the GEOM-Drugs inpainting results summarized in Table 25, and Table 27 the per-scaffold preservation and S∧C cells. All experiments use differential inpainting with noise scale=0.3, 500 samples per cell. The noise scale multiplies the sampler's own per-step noise level on the conditioned tokens: 0 holds them fixed through the trajectory and 1 noises them exactly as the new tokens are noised. On the four aromatic scaffolds TopVAE preserves the conditioning scaffold 1.7× to 3.3× as often as the baseline, with the largest margin on benzene.

Table 26 reports the joint rate for every scaffold and every band, and the sanitization rate on the same cells in its last two columns.

Figures 9 and 10 show expansion outputs for all four scaffolds.

O Conditional Generation: Protocol and Extended Results

Every result in this appendix runs on the frozen prior and frozen decoder evaluated in Tables 3 and 4. No model is retrained, and no training-time hyperparameter is changed.

How the prior is steered. A DeepSets regressor is fitted on the frozen latent space, one per property: 25,000 QM9 training molecules are encoded to their posterior means through the frozen autoencoder, and the regressor predicts the RDKit property of the source molecule from that latent. Held out, it reaches R^2=0.993 at a mean absolute error of 1.23 Å^2 on topological polar surface area and R^2=0.985 at 0.085 on the octanol-water partition coefficient. At each of the 100 reverse steps the prior's denoised estimate is displaced along the gradient of the regressor's squared error to the requested target before the step's posterior mean is formed. The gradient is taken through the regressor alone and never through the diffusion network, is masked to the occupied tokens and normalized per sample, and is applied with step size $2.0\,\sigma_t$, so the displacement anneals with the

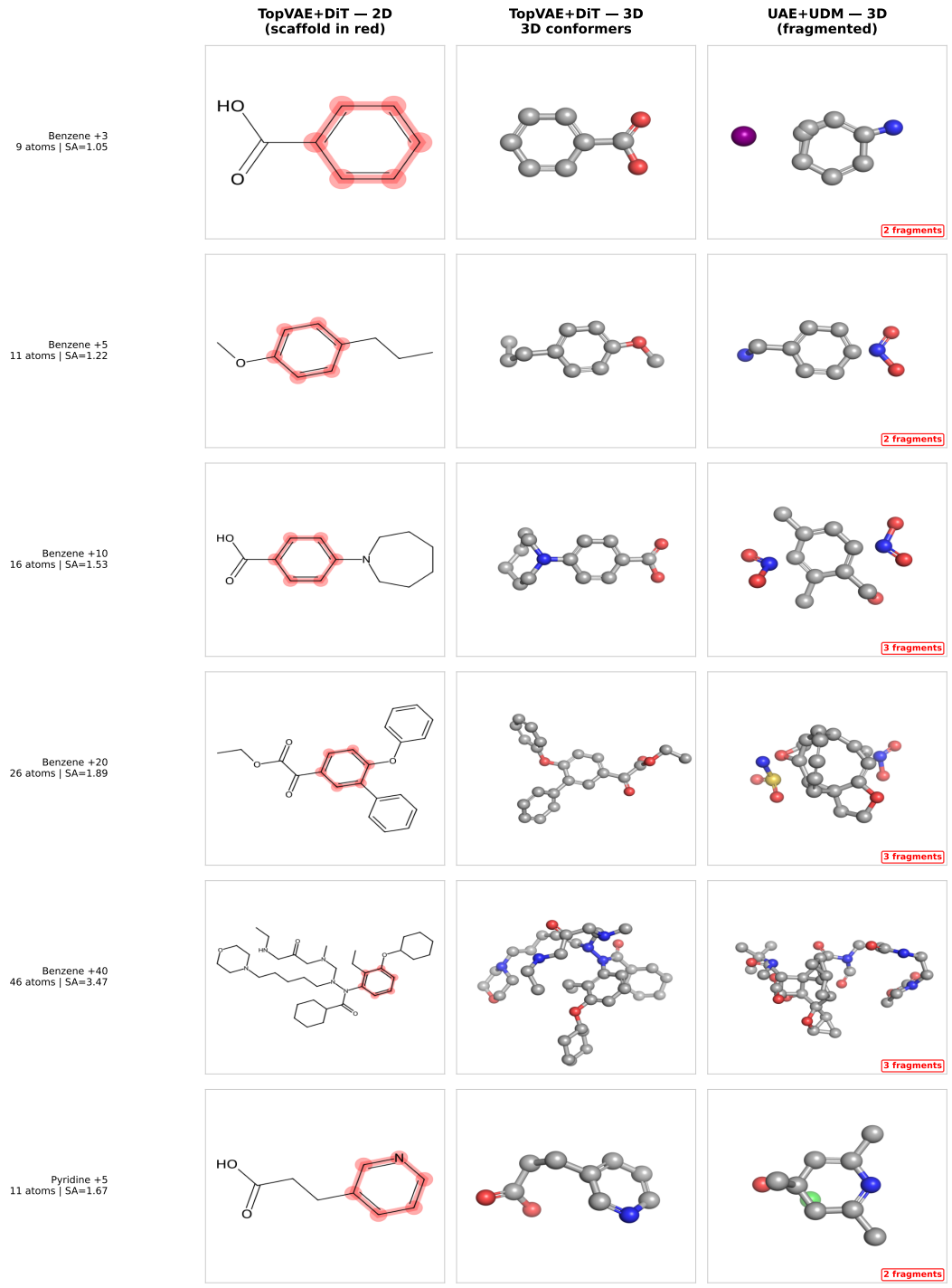


Figure 9: **Scaffold expansion, part 1.** Benzene expanded by 3 to 40 heavy atoms and pyridine by 5. Each row gives the decoded 2D graph with the conditioning scaffold in red, the TopVAE+DiT 3D conformer of the same molecule, and the UAE+UDM output on the same condition. Panels are chosen by a fixed rule rather than at random: the TopVAE panel is the most synthetically accessible sample that sanitizes, stays connected and contains the scaffold, and the UAE+UDM panel is the sample with the most disconnected components. Rates over 500 completions per cell are in Table 26.

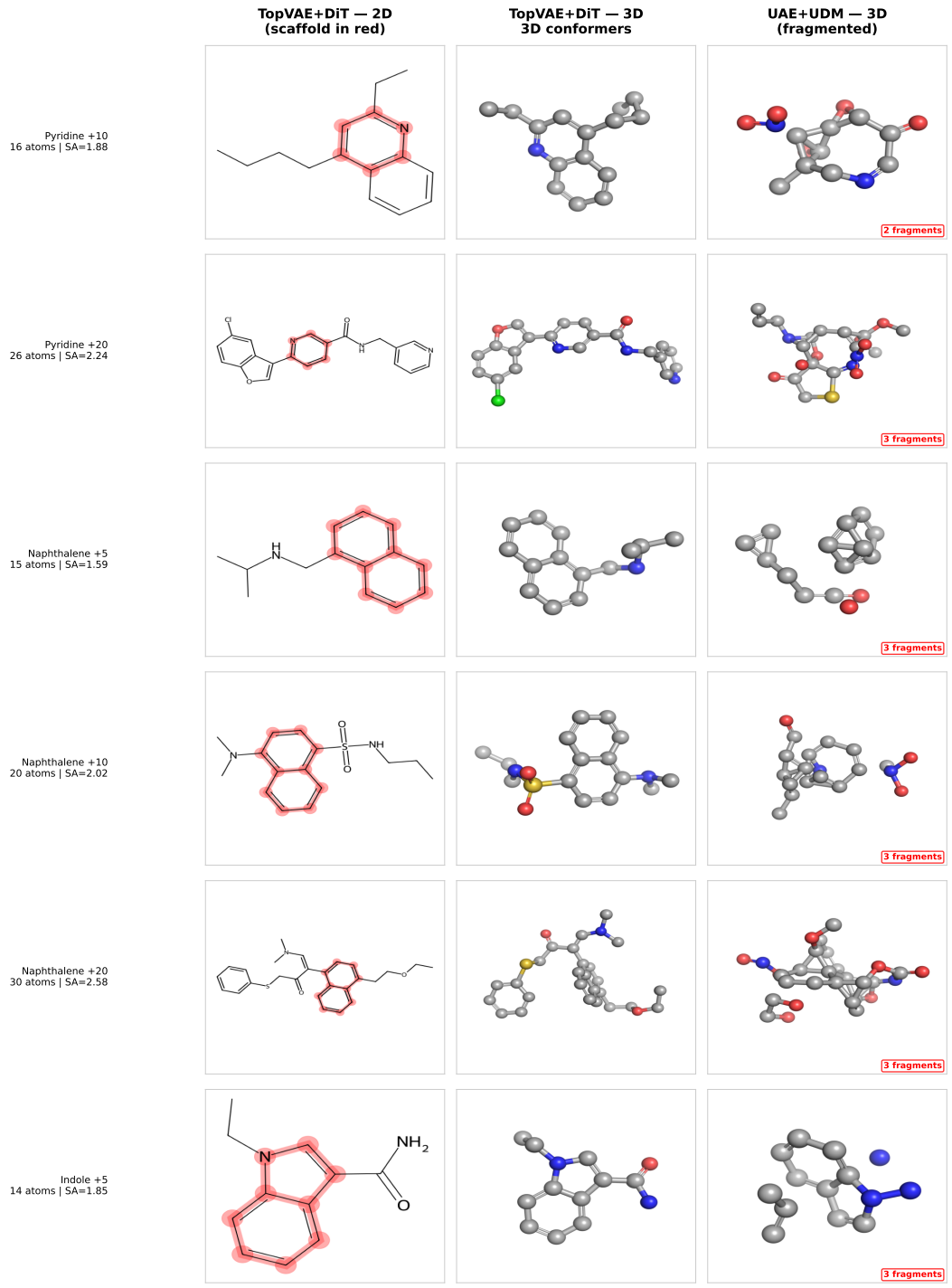


Figure 10: **Scaffold expansion, part 2.** Pyridine expanded by 10 and 20 heavy atoms, naphthalene by 5 to 20 and indole by 5. Each row gives the decoded 2D graph with the conditioning scaffold in red, the TopVAE+DiT 3D conformer of the same molecule, and the UAE+UDM output on the same condition. Panels are chosen by a fixed rule rather than at random: the TopVAE panel is the most synthetically accessible sample that sanitizes, stays connected and contains the scaffold, and the UAE+UDM panel is the sample with the most disconnected components. Rates over 500 completions per cell are in Table 26.

Table 23: **Diversity and novelty on QM9** (n=10,000). The UDM-3D column is our own re-training (no public checkpoint, see Appendix J).

Metric	TopVAE+DiT	UDM-3D
Valid molecules	**10,000** (100%)	9,986 (99.9%)
Internal diversity ↑	0.914	0.916
Novelty ↑	**45.1%**	42.3%
Unique Murcko scaffolds ↑	**35.9%**	29.6%
Uniqueness ↑	96.7%	96.6%
Unique SMILES	9,672	9,660
Unique scaffold count	**3,477**	2,864

Table 24: **Diversity and novelty on GEOM-Drugs** (n=10,000 sampled). The UDM-3D column is our own re-training with the official repository and its default configuration (no public checkpoint, see Appendix J). UDM-3D produces only 1,823 valid molecules (18.2%). Its diversity and novelty metrics are computed on this smaller set.

Metric	TopVAE+DiT	UDM-3D
Valid molecules	**10,000** (100%)	1,823 (18.2%)
Internal diversity ↑	0.874	0.898*
Novelty ↑	99.95%	100%*
Unique Murcko scaffolds ↑	**88.7%**	85.4%*
Uniqueness ↑	99.99%	100%*

*Computed on 1,823 valid molecules.

Table 25: **Zero-shot scaffold inpainting on GEOM-Drugs** (mean over the four aromatic scaffolds, 500 samples per cell, percentages). Cells report the joint task-success rate: the decoded product preserves the conditioning scaffold, is valence-legal, and forms a single connected component. Δ is the paired difference in percentage points. * marks a band whose scaffold-clustered bootstrap 95% confidence interval excludes zero. The per-scaffold joint rates these means are computed from are in Table 26. The column labeled UAE is UAE+UDM, a latent diffusion prior over a reconstruction-trained autoencoder. In the bands where the baseline returns products, they are connected because they are over-bonded, at 1.24 to 1.46 bonds per atom against 1.08 in the GEOM-Drugs data. Across the +80 and +100 bands the UAE column returns 0 sanitizable molecules of 4,000, against 97.5% for TopVAE. Per-scaffold breakdown in Appendix Table 27.

$+n$	UAE	TopVAE	Δ	$+n$	UAE	TopVAE	Δ
5	21.3	**23.1**	+1.8	60	0.2	**9.5**	+9.3*
10	5.1	**15.4**	+10.3*	80	0.0	**3.8**	+3.8*
20	2.6	**10.4**	+7.8*	100	0.0	**0.5**	+0.5
40	5.6	**15.0**	+9.4*				
				Mean	5.0	**11.1**	+6.1*

noise level and vanishes as the trajectory ends. Step size 0 recovers unconditional sampling exactly and is the unconditional column of Table 28. The step size was chosen at the hardest target, the 90th percentile of surface area, over $\{0, 1, 2, 4, 8\}$ at n=300: error falls from 31.5 at 0 to 5.5 at 1 and 5.3 at 2 and does not improve at 4.

O.1 Scaffold-conditioned completion

The scaffold is supplied as a subgraph together with its 3D coordinates and the model fills the remaining heavy-atom slots by reverse diffusion. On the QM9 completions where TopVAE and the baseline both preserve the conditioning scaffold (n=3,535), TopVAE's products pass all five headline

Table 26: **Joint task-success rate by scaffold and expansion band** (GEOM-Drugs, 500 samples per cell). A completion counts only if it preserves the conditioning scaffold, is valence-legal, and forms a single connected component. *San.* is the sanitization rate on the same cells.

	Joint success (%)			San. (%)	
Scaffold / band	UAE	TopVAE	Δ	UAE	TopVAE
Benzene					
+5	28.8	40.6	+11.8	100.0	100.0
+10	7.8	24.4	+16.6	100.0	100.0
+20	8.2	21.2	+13.0	100.0	100.0
+40	20.4	46.6	+26.2	81.4	100.0
+60	0.6	25.0	+24.4	3.6	100.0
+80	0.0	11.0	+11.0	0.0	99.8
+100	0.0	1.4	+1.4	0.0	96.4
Pyridine					
+5	31.6	30.8	−0.8	100.0	100.0
+10	9.8	18.0	+8.2	100.0	100.0
+20	2.0	11.2	+9.2	100.0	100.0
+40	2.0	6.8	+4.8	90.6	100.0
+60	0.0	4.4	+4.4	5.0	100.0
+80	0.0	1.2	+1.2	0.0	99.8
+100	0.0	0.6	+0.6	0.0	95.4
Naphthalene					
+5	6.4	5.4	−1.0	100.0	100.0
+10	0.4	6.2	+5.8	100.0	100.0
+20	0.0	3.8	+3.8	100.0	100.0
+40	0.0	2.8	+2.8	56.6	100.0
+60	0.0	2.0	+2.0	0.4	100.0
+80	0.0	0.4	+0.4	0.0	100.0
+100	0.0	0.0	+0.0	0.0	95.0
Indole					
+5	18.4	15.6	−2.8	100.0	100.0
+10	2.6	13.0	+10.4	100.0	100.0
+20	0.0	5.4	+5.4	100.0	100.0
+40	0.0	3.6	+3.6	85.6	100.0
+60	0.0	6.4	+6.4	4.6	100.0
+80	0.0	2.6	+2.6	0.0	99.6
+100	0.0	0.0	+0.0	0.0	93.8

tests of the official PoseBusters `mol` configuration, which are bond lengths, bond angles, internal steric clash, internal energy and flatness, at 0.9615 against 0.8931.

P Controlled Ablations of AGCL

Under a matched absolute perturbation on QM9 the full recipe is the most robust of the four configurations. Removing AGCL costs 32.5 points of VCB, replacing its teacher with an identically gated but chemistry-free one costs 29.6, and removing ChemCO costs 25.4, all at δ=0.5 (Table 29). At δ=4 the same three costs are 37.8, 39.6 and 36.9 points, and at δ=5 they are 30.3, 31.5 and 28.9.

AGCL acts during training and is never applied at inference. We retrain the reported QM9 configuration from a fixed seed, changing one thing at a time, and evaluate all four under the posterior-noise protocol of Sec. 5.2.

A0 is the reported recipe with the seed fixed. **A1** is without AGCL, with the bond supervision matched, so it differs from A0 only in where that supervision comes from. **A2** replaces the distillation target with one that carries no chemistry and changes nothing else. **A3** is without ChemCO.

Table 27: **GEOM-Drugs scaffold inpainting: per-scaffold breakdown.** 500 samples per cell, differential inpainting (noise scale=0.3). Values in %. Bold = better per cell.

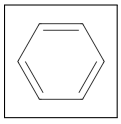
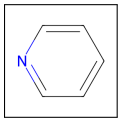
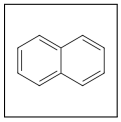
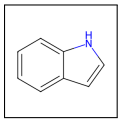
Scaffold	$+n$	N	Scaff. Pres. ↑ UAE	Scaff. Pres. ↑ TopVAE	S∧C ↑ UAE	S∧C ↑ TopVAE
	5	11	31.0	**41.8**	95.6	**96.2**
	10	16	10.2	**25.6**	88.0	**92.2**
	20	26	9.6	**23.4**	**88.2**	86.8
	40	46	25.8	**66.2**	62.0	**69.8**
	60	66	0.8	**48.2**	2.8	**53.4**
	80	86	0.0	**35.0**	0.0	**22.4**
	100	106	0.0	**12.8**	0.0	**6.8**
	5	11	**35.0**	31.4	93.2	**97.2**
	10	16	12.2	**18.4**	88.8	**94.6**
	20	26	2.2	**11.8**	84.2	**86.8**
	40	46	2.4	**9.4**	**71.8**	71.2
	60	66	0.0	**8.8**	3.2	**53.0**
	80	86	0.0	**5.2**	0.0	**25.8**
	100	106	0.0	**2.2**	0.0	**11.8**
	5	15	**9.0**	5.4	83.0	**87.6**
	10	20	1.0	**6.4**	79.0	**89.8**
	20	30	0.0	**4.0**	84.0	**86.6**
	40	50	0.0	**3.8**	42.0	**73.6**
	60	70	0.0	**4.0**	0.2	**48.0**
	80	90	0.0	**2.0**	0.0	**26.2**
	100	110	0.0	**0.8**	0.0	**11.6**
	5	14	**24.8**	15.8	78.4	**92.2**
	10	19	5.2	**13.8**	78.2	**91.2**
	20	29	0.0	**5.8**	85.0	**85.2**
	40	49	0.0	**5.6**	**67.0**	65.0
	60	69	0.0	**10.6**	3.6	**53.6**
	80	89	0.0	**8.4**	0.0	**25.0**
	100	109	0.0	**3.0**	0.0	**8.0**

Table 28: **Property-conditioned generation on a frozen prior** (QM9, targets at the data 10th, 50th and 90th percentile, at n=1,000 per target). MAE is to the requested target. Both columns come from one run on one set of targets: the unconditional column is the same sampler with the guidance step size set to zero. Every draw is kept, so both columns are one molecule per request.

Target	Unconditional MAE	Guided MAE	Guided validity ↑	Guided connectivity ↑
TPSA (Å^2)	23.43	**4.18**	0.971	≈1.000
MolLogP	1.143	**0.28**	0.981	≈0.999

Table 29: **Controlled ablations of AGCL**, on the QM9 validation split (n=10,000, `sample_seed` 42). Every entry is the VCB estimator of Appendix C.7, evaluated at matched absolute displacements of 0.5, 4 and 5, the same value for all four configurations at each displacement. Solver off and solver on are whether ChemCO's continuous solver runs at decode time, and the body reports the solver-off column, the setting every reported checkpoint is sampled at. Best value in each column in bold. The δ=0 column is the reconstruction anchor, 0.9999 for all four, so the arms are read from a common reconstruction anchor.

Configuration	δ=0 anchor	Matched absolute displacement δ=0.5 Solver off	δ=0.5 Solver on	δ=4 Solver off	δ=4 Solver on	δ=5 Solver off	δ=5 Solver on
A0 reported recipe	0.9999	**0.9999**	**0.9999**	**0.4883**	**0.5202**	**0.3758**	**0.4238**
A1 without AGCL	0.9999	0.6745	0.7045	0.1101	0.1355	0.0728	0.0891
A2 chemistry-free teacher	0.9999	0.7040	0.7417	0.0923	0.1186	0.0611	0.0795
A3 without ChemCO	0.9999	0.7454	0.7454	0.1191	0.1191	0.0867	0.0867

Table 30: BFS-inserted edge audit on GEOM-Drugs (n=10,000 generated molecules). A BFS-inserted edge is an edge absent after initial adjacency thresholding that was explicitly added by BFS connectivity repair. For comparison, the corresponding statistics for all graph-theoretic bridge bonds (cut edges) are shown in the right column. Repair inserts 211 of the 96,791 bridge bonds and the decoder predicts the other 99.8%.

Statistic	**BFS-inserted edges**	**All bridge bonds**
Molecules affected	2.1% (210/10,000)	99.9% (9,989/10,000)
Total edges	211	96,791
Mean per molecule	0.02	9.68
Mean fraction of all edges	0.078%	35.8%
Chemically allowed atom pairs	98.1%	99.97%
Bond type: Single / Double / Triple	98.1% / 1.5% / <1%	81% / 18% / <1%

Table 31: Degree, ring, and graph-theoretic bridge-bond statistics on GEOM-Drugs (n=10,000). Lower EMD ↓ indicates closer match to the reference distribution. The UDM-3D column is our own re-training (no public checkpoint, see Appendix E).

Metric	Reference	TopVAE+DiT	UDM-3D
Mean atom degree	2.089	2.185	2.146
Graph-theoretic bridge-bond fraction	0.643	0.356	0.384
Ring-size EMD vs. ref. ↓	—	**0.561**	0.706

Q TopoBridge Chemical Plausibility Analysis

TopoBridge makes the candidate graph connected by adding repair edges between disconnected components through greedy BFS refinement (Eq. 7 and 16). To isolate the effect of BFS connectivity repair from the decoder's own predictions, we instrument the inference pipeline to track *BFS-inserted edges*: edges that did not exist after initial thresholding/top-k of the adjacency logits and were explicitly added by the BFS repair procedure (isolated-node fix or component-merging) to ensure connectivity. Concretely, we snapshot the adjacency matrix $\mathbf{A}_{\text{init}}$ immediately after thresholding and before any repair, then identify inserted edges via $\mathbf{M}_{\text{ins}} = \text{clamp}(\mathbf{A}_{\text{final}} - \mathbf{A}_{\text{init}}, \min=0)$. This differs from the broader set of graph-theoretic bridge bonds (cut edges found by Tarjan's algorithm), most of which originate from the decoder's adjacency head and are not artifacts of repair.

BFS-inserted edge chemistry. Table 30 summarizes the chemical properties of BFS-inserted edges in TopVAE+DiT generations on GEOM-Drugs, against the same statistics for all graph-theoretic bridge bonds.

The repair draws from the highest-probability edge predicted by the adjacency head, so even forced connections involve plausible atom pairs.

Degree and ring distributions. Table 31 compares atom-degree, ring-size and bridge-bond distributions against the reference test set and against our own re-training of UDM-3D, which uses no TopoBridge module.

Summary. BFS insertion is rare at inference. The connectivity reported in Table 4 comes from the adjacency head under Π_{TB}'s thresholding and top-k, with insertion supplying the rest.

R Computational Cost Analysis

The constraint machinery is overwhelmingly a training-time cost: ChemCO's continuous constraint solver is not run at inference. The remainder of this section profiles the training-time budget.

We profile TopVAE's training cost on an NVIDIA H100 NVL GPU with batch size 64 on GEOM-Drugs. Component timings are forward-only and averaged over 50 forward passes after 10 warmup

Table 32: Per-component forward-pass cost of TopVAE on GEOM-Drugs (H100 NVL, batch size 64). TopoBridge timing uses the BFS-only cost from a converged model (α=1.0). *Total forward* is measured end to end rather than as the sum of the rows above it. *Other* is the remainder, and holds the costs not timed separately, among them the adjacency head, Tarjan connectivity and host-device transfer.

Component	**Time (ms)**	**% of Forward**
Encoder	36.8	36.7%
TopoBridge BFS-STE (converged, BFS only)	2.8	2.8%
TopoFormer	2.8	2.8%
AtomHead	0.2	0.2%
BondTypeHead	1.1	1.1%
Bond Distribution Assembly	0.5	0.5%
ChemCO (40 Steps)	**24.1**	**24.0%**
CoordHead + EGNN	7.8	7.8%
Loss computation	2.7	2.7%
Other (Tarjan, data transfer, etc.)	21.5	21.4%
Total forward	**100.3**	**100%**

iterations. Training-step timings are averaged over 50 forward+backward passes after the same warmup.

Component-level breakdown. Table 32 reports the per-component cost of a single TopVAE forward pass. The encoder accounts for 36.7% of forward time. ChemCO contributes 24.0%. TopoBridge's BFS-STE adds only 2.8 ms per batch (2.8% of forward time) when measured at convergence (α=1.0). Note that Tarjan bridge detection is a standard $O(V+E)$ graph algorithm independent of TopoBridge and is folded into the "Other" category.

S Details of ChemCO

This appendix provides the complete optimization details for ChemCO. We use the same notation as Sec. 4.4: $\widetilde{A}$ is the TopoBridge adjacency support, $\Omega^{(k)}$ is the chemical mask for bond type k, $\bar{U}$ is the relative neural utility (Eq. 8), Φ denotes the free primal logits, and $Y^{(k)} = T_k(\Phi)$ is the induced continuous bond-type assignment (Eq. 9).

S.1 Initialization

ChemCO optimizes only real bond types $k = 1, \ldots, K$. The no-bond class $k=0$ is represented by the residual mass. The primal logits are initialized from the relative neural utility, $\Phi^{(k)}_{0,ij} = \bar{U}^{(k)}_{ij}$ for $k = 1, \ldots, K$, which seeds the optimizer at the decoder's own belief, so that the subsequent iteration only needs to *correct* constraint violations rather than discover the bond structure from scratch.

Continuous statistics. At iteration t, ChemCO computes the soft bond-type assignment $Y^{(k)}_t = T_k(\Phi_t)$ using the transform of Eq. (9). T_k includes masking by the TopoBridge support $\widetilde{A}$ and the chemical mask $\Omega^{(k)}$, so $Y^{(k)}_t$ is nonzero only on candidate edges with chemically allowed bond types. The three continuous statistics of Eq. (10) are evaluated at iteration t as $s_{ij,t}$, $\mathrm{val}_{i,t}$ and $\deg_{i,t}$, and the residuals formed from them are the h of Eq. (26). A positive residual always indicates a violated constraint.

S.2 Chemical Mask $\Omega^{(k)}$ Construction

The chemical mask $\Omega^{(k)} \in \{0,1\}^{N\times N}$ encodes which bond types are chemically allowed for each atom pair. It is *static*: each dataset's allowed-pair table is fixed before training, admits only bond types that standard valence-bond chemistry allows for that pair, and is never learned or updated from

Table 33: Allowed bond types per atom-pair combination in $\Omega^{(k)}$, by dataset. S = single, D = double, T = triple. A dash marks a pair the mask disallows, including pairs outside that dataset's element vocabulary.

Atom Pair	**QM9**	**GEOM-Drugs**
C–C	S, D, T	S, D, T
C–N	S, D, T	S, D, T
C–O	S, D	S, D
C–F	S only	S only
C–Cl / C–Br / C–I	—	S only
C–S	—	S, D
N–N	S, D, T	S, D, T
N–O	S, D	S, D
N–F	S only	—
O–O	S only	S only
O–F	S only	—
F–F	S only	—
S–S	—	S only
S–N	—	S, D
S–O	—	S, D
P–N / P–O / P–S	—	S, D
P–C	—	S only
B–C / B–O	—	S only
Si–C / Si–O	—	S only
Bi–C / Bi–O	—	S only

Table 34: Per-element valence capacities c_i. A dash marks an element absent from that dataset's vocabulary.

	C	N	O	F	S	Cl	Br	P	I	B	Si	Bi
QM9	4	3	2	1	—	—	—	—	—	—	—	—
GEOM-Drugs	4	4	2	1	6	1	1	5	1	3	4	3

data. During training, $\Omega^{(k)}$ is constructed from the predicted atom-type $\arg\max$. The solver as a whole runs under `no_grad` (Appendix S.7), so this discrete operation carries no gradient. The mask is constructed using the following rules:

For atom-pair combinations not listed above, no bond of any order is allowed ($\Omega_{ij}^{(k)} = 0$ for every k). Over the 78 unordered element pairs of the 12-element GEOM-Drugs vocabulary, 24 carry a bond and 54 carry none.

The mask is active on both datasets, with a vocabulary-specific column in Table 33. Over {C, N, O, F} on QM9 all ten unordered element pairs carry a bond, and the mask restricts the bond order alone.

The valence capacity c_i of Eq. (10) is a per-element constant taken from a fixed table (Table 34). It is set once per dataset and is not learned. The bond orders are $o_k = (1, 2, 3)$ for single, double and triple on both datasets, the data being kekulized, and c_{N} is the one entry that differs between them.

S.3 Adaptive penalty objective and smoothed active-set direction

We introduce nonnegative penalty weights $\mu_{ij} \geq 0$ for pair exclusivity, $\lambda_i^+ \geq 0$ for valence upper bounds, and $\lambda_i^- \geq 0$ for minimum-degree constraints. These correspond to the main-text weights μ, λ, and ν, respectively. The split notation λ^+/λ^- makes the sign convention explicit.

At iteration t, ChemCO uses an adaptive penalty score at fixed weights

$$\begin{aligned}\mathcal{J}_t(\Phi) = &\sum_{\substack{i<j\\ i,j\in\mathcal{V}}}\sum_{k=1}^{K}\bar{U}_{ij}^{(k)}Y_{ij}^{(k)} - \rho_{\mathrm{pair}}\sum_{\substack{i<j\\ i,j\in\mathcal{V}}}\mu_{ij,t}\psi_\gamma(h_{ij,t}^{\mathrm{pair}})\\ &-\sum_{i\in\mathcal{V}}\lambda_{i,t}^{+}\psi_\gamma(h_{i,t}^{\mathrm{val}}) - \sum_{i\in\mathcal{V}}\lambda_{i,t}^{-}\psi_\gamma(h_{i,t}^{\mathrm{deg}}),\end{aligned} \tag{25}$$

where

$$h_{ij,t}^{\mathrm{pair}} = s_{ij,t} - 1, \qquad h_{i,t}^{\mathrm{val}} = \mathrm{val}_{i,t} - c_i, \qquad h_{i,t}^{\mathrm{deg}} = d_{\min} - \deg_{i,t}. \tag{26}$$

Here $\psi_\gamma(r) = \gamma^{-1}\log(1+\exp(\gamma r))$ is a smooth approximation to $[r]_+$, with $\psi_\gamma'(r) = \sigma(\gamma r)$, so the smoothed active-set gates are

$$\omega_{ij,t}^{\mathrm{pair}} = \sigma(\gamma h_{ij,t}^{\mathrm{pair}}), \qquad \omega_{i,t}^{\mathrm{val}} = \sigma(\gamma h_{i,t}^{\mathrm{val}}), \qquad \omega_{i,t}^{\mathrm{deg}} = \sigma(\gamma h_{i,t}^{\mathrm{deg}}). \tag{27}$$

When γ is large, these gates approach binary indicators of violated constraints. For finite γ, they provide a smooth active-set approximation.

The induced smoothed direction with respect to the continuous assignment mass $Y_{ij}^{(k)}$ is

$$\begin{aligned}\widetilde{G}_{ij,t}^{(k)} = &\bar{U}_{ij}^{(k)} - \rho_{\mathrm{pair}}\omega_{ij,t}^{\mathrm{pair}}\mu_{ij,t}\\ &- o_k\big(\omega_{i,t}^{\mathrm{val}}\lambda_{i,t}^{+} + \omega_{j,t}^{\mathrm{val}}\lambda_{j,t}^{+}\big) + \omega_{i,t}^{\mathrm{deg}}\lambda_{i,t}^{-} + \omega_{j,t}^{\mathrm{deg}}\lambda_{j,t}^{-}.\end{aligned} \tag{28}$$

The implementation uses $\rho_{\mathrm{pair}} = 2$, matching the row-wise pair penalty followed by explicit symmetrization. Equivalently, this constant can be absorbed into the pair weight scale.

S.4 Row-centric direction and direct logit-space update

For efficient batched implementation, ChemCO constructs the direction in two stages. First, it forms a row-centric direction that contains atom i's node-level penalty contributions:

$$\widehat{G}_{ij,t}^{(k)} = \tfrac{1}{2}\bar{U}_{ij}^{(k)} - \tfrac{\rho_{\mathrm{pair}}}{2}\omega_{ij,t}^{\mathrm{pair}}\mu_{ij,t} - o_k\omega_{i,t}^{\mathrm{val}}\lambda_{i,t}^{+} + \omega_{i,t}^{\mathrm{deg}}\lambda_{i,t}^{-}. \tag{29}$$

Symmetrizing, $\widetilde{G}_{ij,t}^{(k)} = \widehat{G}_{ij,t}^{(k)} + \widehat{G}_{ji,t}^{(k)}$ recovers Eq. (28): $\bar{U}$ and μ are symmetric, and the two endpoints contribute the valence and degree terms.

Mirror-ascent update. ChemCO applies this direction directly to Φ:

$$\Phi_{t+1}^{(k)} = \Phi_t^{(k)} + \eta_\Phi^{(t)}\widetilde{G}_t^{(k)} \odot \Omega^{(k)}, \qquad k = 1,\dots,K, \tag{30}$$

where $\eta_\Phi^{(t)} = \alpha\gamma_\alpha^t$ is exponentially decayed with base rate α and decay factor $\gamma_\alpha \in (0,1]$. The TopoBridge support $\widetilde{A}$ is enforced by the transform $T_k(\Phi)$, and the chemical mask $\Omega^{(k)}$ zeros out forbidden atom-bond combinations.

Eq. (30) is a mirror-ascent step on Eq. (25) in the geometry of the separable potential $H(Y) = \sum_{i<j}\sum_k h\big(Y_{ij}^{(k)}\big)$ with $h' = g^{-1}$ (Nemirovski & Yudin, 1983; Beck & Teboulle, 2003): Algorithm 1 symmetrizes both U and $\widetilde{G}_t$, so on the support $Y = g(\Phi)$ and $\Phi = \nabla H(Y)$. The chain-rule gradient in Φ would instead carry the factor $g'(\Phi) = 2\sigma(\Phi)^2(1-\sigma(\Phi))$, which vanishes as $\sigma(\Phi) \to 0$ and pins currently absent bonds at $\Phi \ll 0$. The mirror step removes that factor, and lets the correction activate bonds the decoder scores as absent.

S.5 Penalty-weight updates

Penalty-weight warm-start. Rather than initializing all penalty weights at zero, ChemCO warm-starts each at its initial positive residual, $[s_{ij,0}-1]_+$, $[\mathrm{val}_{i,0}-c_i]_+$ and $[d_{\min}-\deg_{i,0}]_+$, so the solver has a nonzero correction signal from the first iteration.

After the logit update, ChemCO recomputes $Y_{t+1}^{(k)} = T_k(\Phi_{t+1})$ and refreshes s_{t+1}, val_{t+1}, and $\deg_{t+1}$. The penalty weights are then updated by accumulating positive residuals:

$$\mu_{ij,t+1} = \mu_{ij,t} + \eta_\mu^{(t)}\big[s_{ij,t+1}-1\big]_+, \qquad \mu_{t+1} \leftarrow \tfrac{1}{2}(\mu_{t+1} + \mu_{t+1}^{\top}), \tag{31}$$

Algorithm 1 ChemCO: implemented mirror-ascent logit-space correction

Require: Raw bond logits $U \in \mathbb{R}^{N\times N\times(1+K)}$; node mask m; TopoBridge support $\widetilde{A}$; chemical masks $\{\Omega^{(k)}\}_{k=1}^{K}$; valence caps $\{c_i\}$; minimum degree $d_{\min}$; iterations T_{CO}.
Ensure: Projected bond distribution $P \in [0,1]^{N\times N\times(1+K)}$.

Build valid-pair mask M from m and effective support $M_e \leftarrow M \odot \widetilde{A}$.
Symmetrize logits: $U \leftarrow \frac{1}{2}(U + U^\top)$.
Compute relative utility: $\bar{U}^{(k)} \leftarrow U^{(k)} - U^{(0)}$, $k = 1, \dots, K$.
Initialize logits: $\Phi_0^{(k)} \leftarrow \bar{U}^{(k)}$.
Set $a_0^{(k)} \leftarrow \sigma(\Phi_0^{(k)}) \odot M_e \odot \Omega^{(k)}$.
Compute $Y_0^{(k)} \leftarrow \widehat{T}_k(a_0)$, then s_0, val_0, and $\deg_0$.
Warm-start penalty weights: $\mu_0 \leftarrow [s_0 - 1]_+$, $\lambda_0^+ \leftarrow [\mathrm{val}_0 - c]_+$, and $\lambda_0^- \leftarrow [d_{\min} - \deg_0]_+$.
for $t = 0, \dots, T_{\mathrm{CO}} - 1$ **do**
  Compute gates $\omega_t^{\mathrm{pair}}, \omega_t^{\mathrm{val}}, \omega_t^{\mathrm{deg}}$ using Eq. (27).
  Compute row-centric direction $\widehat{G}_t^{(k)}$ using Eq. (29).
  Symmetrize: $\widetilde{G}_t^{(k)} \leftarrow \widehat{G}_t^{(k)} + (\widehat{G}_t^{(k)})^\top$.
  Update logits: $\Phi_{t+1}^{(k)} \leftarrow \Phi_t^{(k)} + \alpha\gamma_\alpha^t \widetilde{G}_t^{(k)} \odot \Omega^{(k)}$.
  Recompute $a_{t+1}^{(k)} \leftarrow \sigma(\Phi_{t+1}^{(k)}) \odot M_e \odot \Omega^{(k)}$, $Y_{t+1}^{(k)} \leftarrow \widehat{T}_k(a_{t+1})$, and refresh s_{t+1}, val_{t+1}, $\deg_{t+1}$.
  Update μ_{t+1} using Eq. (31).
  Update λ_{t+1}^+ and λ_{t+1}^- using Eq. (32).
end for
Assemble soft distribution: $P^{(k)} \leftarrow Y_{T_{\mathrm{CO}}}^{(k)}$ for $k = 1, \dots, K$.
Set no-bond mass: $P^{(0)} \leftarrow \mathrm{clamp}(1 - \sum_{k=1}^{K} P^{(k)}, 0, 1)$.
For invalid pairs, set $P_{ij} = [1, 0, \dots, 0]$.
return P.

$$\lambda_{i,t+1}^+ = \lambda_{i,t}^+ + \eta_+^{(t)}\big[\mathrm{val}_{i,t+1} - c_i\big]_+, \qquad \lambda_{i,t+1}^- = \lambda_{i,t}^- + \eta_-^{(t)}\big[d_{\min} - \deg_{i,t+1}\big]_+. \tag{32}$$

Here $\eta_\mu^{(t)} = \eta_\mu \gamma_\mu^t$, $\eta_+^{(t)} = \eta_+ \gamma_+^t$, and $\eta_-^{(t)} = \eta_- \gamma_-^t$ are exponentially decayed weight step sizes.

S.6 Complete algorithm

Algorithm 1 summarizes the implemented ChemCO core. It contains the mirror-ascent iteration in Φ and the positive-residual weight accumulation.

Algorithm 1 carries the post-sigmoid quantity $a^{(k)} = \sigma(\Phi^{(k)}) \odot M_e \odot \Omega^{(k)}$ rather than the logits, where M_e is the valid-pair mask restricted to $\widetilde{A}$. For that argument we write $\widehat{T}_k(a) := \frac{1}{2}\big(a \odot a + (a \odot a)^\top\big) \odot \widetilde{A}$, so that the transform T_k of Eq. (9) is $T_k(\Phi) = \widehat{T}_k\big(\sigma(\Phi) \odot M_e \odot \Omega^{(k)}\big)$ and the sigmoid is applied exactly once.

S.7 Full bond distribution and hard decoding

After T_{CO} iterations, ChemCO returns the optimized assignment $\widehat{Y}^{(k)} = T_k(\widehat{\Phi})$. The full soft bond distribution over $\{0, 1, \dots, K\}$ assigns no-bond mass as the clamped residual:

$$P_{ij}^{(0)} = \mathrm{clamp}\Big(1 - \sum_{k=1}^{K} \widehat{Y}_{ij}^{(k)},\ 0,\ 1\Big), \qquad P_{ij}^{(k)} = \widehat{Y}_{ij}^{(k)}, \quad k = 1, \dots, K. \tag{33}$$

The clamp keeps every entry in $[0, 1]$ and assigns the residual mass to the no-bond channel. The real-bond channels carry $\sum_{k\geq 1} P_{ij}^{(k)} = s_{ij}$, so P_{ij} sums to one exactly when $s_{ij} \leq 1$.

The discrete bond type is $\widehat{B}_{ij} = \arg\max_{k\in\{0,1,\dots,K\}} P_{ij}^{(k)}$.

The constraint solver is never differentiated through. In every model reported in this paper it is executed under `no_grad`, so no gradient reaches the decoder along $\Phi_0 \mapsto \Phi_{T_{\mathrm{CO}}} \mapsto \widehat{Y}$, and the

advantage-gated teacher loss of Sec. 4.5 additionally detaches its target $\widehat{Y}$. The decoder is trained toward a black-box solver's corrections. Memory is therefore independent of the iteration count T_{CO}, because no intermediate iterate is retained. Time is linear in T_{CO}, and Appendix R measures the solver at 24.0% of the training forward pass. Hard decoding is used only for discrete molecule construction.