TREEREF-BFN: EQUIVARIANCE-FREE DE NOVO MOLECULE GENERATION BASED ON 2D TOPOLOGY AND INTERNAL 3D GEOMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

De novo 3D molecular generation jointly models molecular size, topology, and geometry. Most methods pre-sample molecular size and generate Cartesian coordinates, limiting variable-size conditional tasks such as fragment completion and scaffold decoration while often relying on equivariant architectures. Internal-coordinate methods avoid rigid-body redundancy but typically require a known molecular graph or autoregressive construction, which may accumulate errors. We propose TreeRef, a tree-based molecular representation that assigns molecular topology and topology-dependent local 3D geometry to a naturally variable-size tree. RingRef nodes encode ring closures while preserving the tree structure, while Null nodes allow molecular size to emerge directly from node occupancy. Based on TreeRef, we develop TreeRef-BFN, a Bayesian Flow Network with a standard Transformer backbone that globally couples these locally defined variables and jointly generates discrete molecular variables and continuous local geometry. A single pretrained TreeRef-BFN supports unconditional generation and variable-size structure-conditioned 3D generation through masking alone, without retraining. Empirical studies demonstrate strong chemical validity, molecular stability, and diversity, accurate local geometric distributions, fast sampling, and competitive property-conditioned generation, establishing TreeRef-BFN as an efficient and flexible framework for 3D molecular generation.

1 INTRODUCTION

De novo 3D molecular generation seeks to construct complete molecular structures from scratch. Unlike conformer generation, neither molecular size nor topology is known beforehand, making it a joint modeling problem over molecular size N, 2D topology $\mathcal{G}$, and 3D geometry X.

Most existing methods generate molecular geometry directly in Cartesian coordinates and have achieved strong performance (Hoogeboom et al., 2022; Vignac et al., 2023; Huang et al., 2023). However, two limitations remain. First, many methods determine N before generation and model $p(\mathcal{G}, X \mid N)$. While effective for unconditional generation, this is less natural for downstream tasks such as fragment completion, scaffold decoration, and linker generation, where the final molecular size should adapt to the condition and be generated together with the structure. Fixing molecular size in advance can likewise restrict property-driven inverse design, where a suitable size may depend on the target properties and is not necessarily known beforehand. Second, Cartesian coordinates represent local molecular geometry only indirectly. For methods that do not explicitly generate bonds, the molecular graph must additionally be inferred from the generated coordinates, so geometric errors can directly lead to incorrect connectivity or bond assignments. Moreover, many Cartesian models rely on equivariant architectures to handle global rotations and translations, introducing additional architectural complexity and computational cost.

Internal coordinates provide a more direct description through bond lengths, bond angles, and torsions, while being invariant to global rigid-body transformations. Their main difficulty is that they are *topology-dependent*: an angle or torsion can only be defined after the corresponding local connectivity is known. This is straightforward for conformer generation (Xu et al., 2022; Jing et al., 2022), where the molecular graph is given, but not for *de novo* generation. G-SphereNet (Luo & Ji,

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.

2022) addresses this by constructing atoms sequentially and defining internal coordinates once their references become available. As a result, later geometry depends on earlier construction decisions, making generation order-dependent and allowing errors to propagate and accumulate. We therefore ask whether topology-dependent internal geometry can instead be assigned to a stable structural support before generation, allowing size, topology, and geometry to be generated jointly.

We propose TreeRef, which represents a molecule on a full k-ary tree. A BFS traversal establishes parent–child relations, and each node is described by $z_i = (R_i, E_i, B_i, G_i, Q_i, H_i)$, representing role, element, parent bond, local geometry, formal charge, and handedness. Unused child slots are Null nodes, a prescribed occupancy class rather than padding, so molecular size emerges from which slots are occupied rather than being sampled beforehand. Edges excluded from the spanning tree are represented by virtual RingRef nodes, allowing ring closures to be encoded on the same structural support. TreeRef thus maps molecular size, topology, and local geometry onto a common fixed-width representation. We generate TreeRef with a Bayesian Flow Network (Graves et al., 2023) using a standard Transformer backbone (Vaswani et al., 2017). Since occupancy, topology, and geometry share the same support, the same pretrained model can perform unconditional generation, topology-conditioned 3D generation, and variable-size fragment completion by fixing different subsets of variables. By conditioning generation on target molecular properties, TreeRef-BFN also supports molecular inverse design.

Contributions. Our work makes three contributions. First, TreeRef unifies molecular size, topology, and local 3D geometry in a tree representation, where the number of atoms is determined directly during generation by node occupancy rather than pre-sampled in advance. Second, TreeRef assigns chemical and geometric information locally to individual nodes, while a standard Transformer enables these local representations to interact globally; combined with a Bayesian Flow Network, this supports joint discrete–continuous generation without autoregressive construction or equivariance. Third, the shared TreeRef representation turns molecular conditioning into variable masking, enabling a single pretrained model to perform 2D-to-3D generation and variable-size molecular editing, including scaffold decoration and linker generation, without task-specific retraining.

2 Related Work

2.1 Molecular Graph Generation and Tree-Based Representations

JT-VAE represents molecules as junction trees of chemical substructures (Jin et al., 2018), while STGG/STGG+ decompose molecular graphs into spanning trees and residual edges for autoregressive 2D generation (Ahn et al., 2022; Jolicoeur-Martineau et al., 2025). TreeRef instead uses a tree as a shared index for atomic topology and local 3D geometry, with RingRef nodes encoding ring closures on the same structural support.

2.2 De Novo 3D Molecular Generation

Most *de novo* 3D generators operate in Cartesian space. EDM, GCDM, and GeoLDM generate atom types and 3D coordinates using equivariant diffusion or latent diffusion (Hoogeboom et al., 2022; Morehead & Cheng, 2024; Xu et al., 2023). MiDi, JODO, SemlaFlow, and FlowMol3 further generate 2D topology together with Cartesian geometry through mixed discrete–continuous diffusion or flow matching (Vignac et al., 2023; Huang et al., 2023; Irwin et al., 2025; Dunn & Koes, 2025). GeoBFN applies Bayesian Flow Networks to mixed molecular variables (Song et al., 2024), while VecMol uses continuous vector-field representations (Hua et al., 2026).

Recent methods also reduce reliance on explicitly equivariant backbones. ADiT performs latent diffusion with standard Transformers (Joshi et al., 2025), and Zatom-1 combines multimodal flow matching with a standard Transformer (Morehead et al., 2026). Symphony++ instead generates atoms and coordinates autoregressively, naturally allowing molecular size to emerge during generation (Kim et al., 2026). In contrast, TreeRef represents topology and local 3D geometry directly as internal variables on a shared structural support. A standard Transformer generates these variables; it is not equivariant. Cartesian coordinates are recovered afterwards by decoding, in a frame fixed at the root.

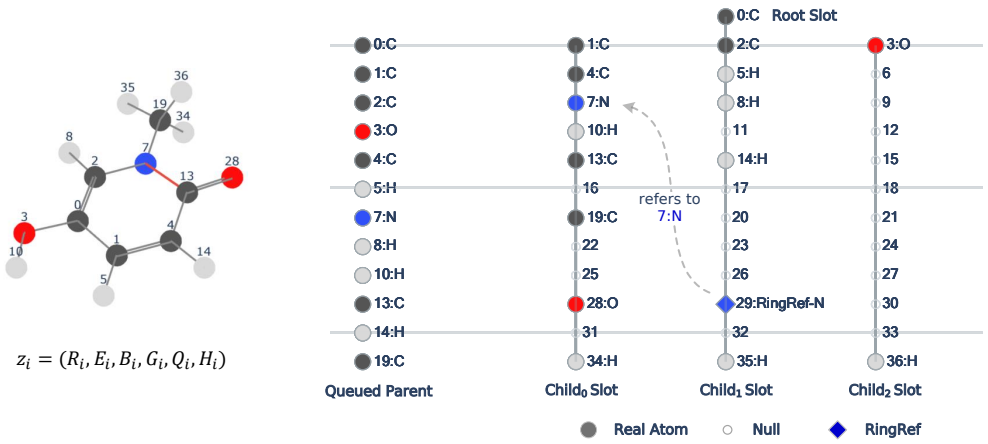


Figure 1: Illustration of the TreeRef representation. The left panel shows the Real atoms of the original 3D molecule. The middle column lists the Real atoms in BFS order, each serving as the parent of the $k = 3$ child slots in the three columns on the right, which are the actual BFN-modeled tokens represented by z_i. Unused child slots are represented as Null nodes, and all slots after the last occupied region are Null. During BFS traversal, edges excluded from the spanning tree (red edge in the molecule) are represented by RingRef nodes, which encode ring-closing relations by referring to the corresponding existing Real atoms (dashed arrow).

2.3 Internal Coordinates and Intrinsic Molecular Geometry

Internal coordinates have been widely used for conformer generation when molecular topology is known. GeoDiff generates conformations conditioned on a molecular graph (Xu et al., 2022), Torsional Diffusion models rotatable-bond torsions (Jing et al., 2022), and GO-Flow models translation, rotation, and conformation on corresponding manifolds (Liu et al., 2026). G-SphereNet extends internal-coordinate generation to the *de novo* setting by constructing atoms sequentially with distances, angles, and torsions (Luo & Ji, 2022). TreeRef instead assigns topology-dependent local geometry to stable tree slots, enabling molecular size, topology, and internal geometry to be generated jointly rather than autoregressively.

3 Method

3.1 Encoding TreeRef Representation

Overview and slot variables. TreeRef represents a molecule as a rooted full k-ary tree whose slots jointly encode molecular occupancy, topology, and local 3D geometry. Let $Z = (z_0, \ldots, z_{L'-1})$ denote the packed representation, where each slot is

$$z_i = (R_i, E_i, B_i, G_i, Q_i, H_i). \tag{1}$$

Here, $R_i \in \{\text{Real}, \text{RingRef}\}$ denotes the role, $E_i \in \mathcal{A} \cup \{\text{Null}\}$ the element, and $B_i \in \{\text{None}\} \cup \mathcal{B}$ the bond to the parent or RingRef host. Q_i is an optional formal-charge channel, G_i stores local geometry, and $H_i \in \{\text{inactive}, +, -\}$ represents local handedness.

Tree construction and packing. We choose a root with degree no larger than k and perform BFS to establish parent–child relations. Each Real atom is assigned k child slots, with unused slots filled by Null.

Null is an occupancy class with a fixed target state, not padding; the model learns which slots receive it. A slot is expanded only when it represents a Real atom with $E_i \neq \text{Null}$; otherwise, its descendants remain unoccupied. Molecular size is therefore determined by the generated occupancy rather than sampled before generation. For Null slots, we set $B = \text{None}$, $H = \text{inactive}$, and the continuous geometry to zero.

The root is stored first, followed by the k child slots of each Real atom in a deterministic BFS order, with the ordering rules specified in Appendix A.1; only Real atoms are further expanded. This yields a packed sequence whose occupied region depends on molecular structure.

Using $N_{\max}$ as the maximum number of physical atoms, the initial shared capacity is bounded by $1 + kN_{\max}$. Since terminal child blocks are entirely Null, trailing all-Null blocks can be removed consistently over the dataset, producing a shared width L'. All molecules are therefore represented by tensors of the same width, while positions beyond the cropped boundary are treated as implicit Null during decoding.

RingRef. BFS retains a spanning-tree edge set $\mathcal{E}_T$, while $\mathcal{E}_R = \mathcal{E} \setminus \mathcal{E}_T$ contains non-tree edges. For each edge in $\mathcal{E}_R$, one endpoint acts as the host and a virtual RingRef node is inserted into an available child slot, referring to the other endpoint as an existing Real atom. RingRef is a leaf and is not further expanded. Its element and optional formal charge follow the referenced atom, while its bond channel records the corresponding ring-closing bond. In this way, ring closures are incorporated into the same tree-structured representation rather than stored separately.

Local geometry and handedness. TreeRef does not directly generate Cartesian coordinates. Instead, each occupied slot stores local internal geometry. The scalar r gives the bond length to the parent, or the host–target distance for a RingRef. Although B specifies the bond type, it does not determine this continuous distance.

For at most $k + 1$ ordered local directions, we represent pairwise bond angles by $\cos\theta$ and include one dihedral angle as $(\cos\phi, \sin\phi)$ to determine the local orientation around the reference bond when the required references are available. After standardizing positive distances, the continuous geometry is

$$G_i = (\widetilde{r}_i, \{\cos\theta_{i;ab}\}, \cos\phi_i, \sin\phi_i)\,, \qquad d_G = 1 + \binom{k+1}{2} + 2. \tag{2}$$

Here, $\widetilde{r}_i = (r_i - \mu_r)/\sigma_r$ for positive physical distances, while the distance entries for root and Null slots remain zero.

Pairwise angles leave a mirror choice, which H_i records. For three ordered unit directions u_i, v_i, w_i, let $s_i = u_i \cdot (v_i \times w_i)$ and set

$$H_i = \begin{cases} +, & s_i > \epsilon_H, \\ -, & s_i < -\epsilon_H, \\ \text{inactive}, & |s_i| \le \epsilon_H \text{ or the reference directions are undefined,} \end{cases} \qquad \epsilon_H = 10^{-3}. \tag{3}$$

Nearly flat neighborhoods, and neighborhoods missing a reference direction, are labeled inactive. Appendix A.1 gives the reference order. A RingRef stores its geometry and handedness around the atom it points to.

3.2 Bayesian Flow Training and Sampling

We adopt Bayesian Flow Networks (BFNs) (Graves et al., 2023) because they naturally accommodate discrete and continuous variables and, when trained in continuous time, allow the sampling step count to be chosen at inference without retraining, making them well suited to joint molecular topology and geometry generation.

TreeRef contains heterogeneous categorical variables and continuous local geometry. We use independent categorical Bayesian flows for $c \in \mathcal{C} = \{R, E, B, H\}$, with Q included when enabled, together with a Gaussian Bayesian flow for G. All channels share the same flow time t, while a single Transformer jointly models dependencies across channels and TreeRef slots.

Bayesian flow. For a categorical channel c with K_c classes and clean state x_i^c, let $e(x_i^c)$ denote its one-hot representation. Using the accumulated accuracy $\beta_c(t)$, the categorical Bayesian state is obtained from

$$z_{i,t}^c \sim \mathcal{N}\big(\beta_c(t)\big(K_c e(x_i^c) - \mathbf{1}\big), K_c\beta_c(t)I\big)\,, \qquad s_{i,t}^c = \mathrm{softmax}(z_{i,t}^c). \tag{4}$$

For continuous geometry, we use minimum variance m and define $\gamma(t) = 1 - m^t$. The corresponding geometry state follows

$$\mu_{i,t}^G \sim \mathcal{N}\big(\gamma(t)G_i,\, \gamma(t)\big(1 - \gamma(t)\big)I\big)\,. \tag{5}$$

The states of all channels are projected to a shared hidden space:

$$h_i^{(0)} = \sum_{c\in\mathcal{C}} P_c(s_{i,t}^c) + P_G(\mu_{i,t}^G) + e_i^{\text{pos}}, \tag{6}$$

Here, e_i^{pos} is the sum of packed-position, child-slot, and parent-group embeddings. A Transformer with AdaLN-Zero conditioning jointly predicts categorical distributions $\widehat{p}_i^c$ and clean local geometry $\widehat{G}_i$. The shared flow time t is encoded by a sinusoidal embedding followed by an MLP, which modulates normalized activations and residual gates in every attention and feed-forward block, together with the final layer normalization. Self-attention operates over all retained packed slots, including Null slots.

We train using the continuous-time BFN objective. With the categorical accuracy schedule $\beta_c(t) = \beta_1 t^2$ and geometry variance schedule m^t, the objective is

$$\mathcal{L}_{\text{BFN}} = \mathbb{E}_{Z,t,\epsilon}\left[\frac{1}{|\mathcal{C}|L'}\sum_{c\in\mathcal{C}}\sum_{i=0}^{L'-1} K_c\beta_1 t\ \|e(x_i^c) - \widehat{p}_i^c\|_2^2 + \frac{-\log m}{2m^t L' d_G}\sum_{i=0}^{L'-1}\left\|G_i - \widehat{G}_i\right\|_2^2\right]. \tag{7}$$

Here, $t \sim \text{Uniform}(0,1)$, ϵ denotes the forward-flow noise, and the predictions are evaluated at the resulting Bayesian states and time t. The categorical term averages equally over the enabled channels and packed slots, while the geometry term averages over slots and geometry dimensions. We use $\beta_1 = 1$ and $m = 10^{-3}$. All packed slots contribute to training, including Null slots with their prescribed categorical targets and zero clean geometry.

Tree-view consistency regularization. The same molecule can yield different BFS serializations: parent relations, RingRef placement, and local frames all change. By default, training uses only the canonical tree. When this regularizer is enabled, several trees produced by random BFS are trained together. They share one flow time and receive independent noise. Let $\overline{h}^{(a)}$ be the pooled final hidden state of view a, and let $\mathcal{P}_b$ be the pairs of distinct views of molecule b. The consistency loss is

$$\mathcal{L}_{\text{inv}} = \mathbb{E}_b\left[\frac{1-t_b}{\max(1,|\mathcal{P}_b|)}\sum_{(a,a')\in\mathcal{P}_b}\left(1-\cos\left(\overline{h}^{(a)},\overline{h}^{(a')}\right)\right)\right], \tag{8}$$

and the training objective is $\mathcal{L} = \mathcal{L}_{\text{BFN}} + \lambda_{\text{inv}}\mathcal{L}_{\text{inv}}$. This term is optional and is primarily used to reduce serialization-specific shortcuts in conditional generation. Pooling, the $(1-t)$ weight, and pair construction are given in Appendix B.3.

Generative sampling. Sampling starts from categorical log-states $z_{i,0}^c = 0$ and a Gaussian geometry prior with mean $\mu_{i,0}^G = 0$ and precision $\rho_0 = 1$. We use S Bayesian updates at times $t_j = j/S$, reserving N for molecular size. For categorical channel c, let $\alpha_j^c = \beta_c(t_j) - \beta_c(t_{j-1})$. At time t_{j-1}, sample a class $\widetilde{x}_{i,j}^c \sim \text{Cat}(\widehat{p}_{i,j-1}^c)$ and update

$$z_{i,j}^c = z_{i,j-1}^c + \alpha_j^c\left(K_c e(\widetilde{x}_{i,j}^c) - \mathbf{1}\right) + \sqrt{K_c\alpha_j^c}\,\epsilon_{i,j}^c, \qquad \epsilon_{i,j}^c \sim \mathcal{N}(0,I). \tag{9}$$

For geometry, let $\rho_j = m^{-t_j}$ and $\alpha_j^G = \rho_j - \rho_{j-1}$. The predicted clean geometry defines a sender observation and the Gaussian update:

$$y_{i,j}^G \sim \mathcal{N}\left(\widehat{G}_{i,j-1}, (\alpha_j^G)^{-1}I\right), \qquad \mu_{i,j}^G = \frac{\rho_{j-1}\mu_{i,j-1}^G + \alpha_j^G y_{i,j}^G}{\rho_j}. \tag{10}$$

All channels share the time grid and interact through every network evaluation. After the S updates, a final evaluation at $t = 1$ provides the categorical distributions to sample and the clean geometry used for decoding. The complete procedure is given in Appendix B.5.

Masked conditional generation. Conditional generation reuses the pretrained BFN with frozen conditioning. Let $M_i^c \in \{0,1\}$ and $M_i^G \in \{0,1\}^{d_G}$ denote categorical and geometry masks, respectively, with one indicating a variable to generate and zero an observed variable. Before each Transformer evaluation at t_j, we form

$$\begin{aligned}\bar{s}_{i,j}^c &= M_i^c s_{i,j}^c + (1-M_i^c)\widetilde{s}_{i,j}^c,\\ \bar{\mu}_{i,j}^G &= M_i^G \odot \mu_{i,j}^G + (\mathbf{1}-M_i^G)\odot\widetilde{\mu}_{i,j}^G.\end{aligned} \tag{11}$$

Frozen conditioning sets $\widetilde{s}^c_{i,j} = e(x^c_{i,\mathrm{obs}})$ and $\widetilde{\mu}^G_{i,j} = G_{i,\mathrm{obs}}$, where $e(\cdot)$ denotes a one-hot state. Unobserved variables start from their priors and evolve through the original categorical and Gaussian Bayesian updates. The schedule follows the forward sampling grid without time jumps and restores the observed values at final readout.

Unconditional generation sets all masks to one. For generation conditioned on the serialized discrete structure, R, E, B, and optional Q are observed, while G and H are regenerated; RingRef targets are determined during decoding. Fragment completion conditions on the supplied discrete variables and selected geometric observations while generating the remaining variables, including occupancy. For structural editing, geometry and handedness at attachment sites are also regenerated. The final molecular size therefore emerges during sampling rather than being specified in advance. All tasks share the same pretrained BFN, with task-specific masks and structural readout detailed in Appendix B.6.

3.3 TreeRef Decoding

Decoding converts the generated TreeRef variables into a molecular graph $\mathcal{G}$ and Cartesian coordinates X through topology construction, geometry construction, and RingRef closure. The variables themselves are produced jointly; the decoder then reads them in tree order.

Topology construction. Starting from a non-Null Real root, we traverse the packed parent–child structure, retaining reachable non-Null Real and RingRef slots while expanding only Real slots. Unreachable slots are set to Null. The retained Real atoms and their parent–child relations define the spanning-tree edges $\mathcal{E}_T$. Their bond types are read from B, with non-None bonds enforced at occupied non-root slots. RingRefs remain virtual leaves until their targets are identified after geometry construction.

Geometry construction. We reconstruct coordinates sequentially from the local distances, angular cosines, dihedrals, and handedness. Each Real node's construction neighborhood contains its parent, when present, followed by its occupied Real and RingRef children. RingRef children are placed as virtual points using the host's local geometry; their own geometric descriptors are centered at the referenced atom.

For a non-root node i, let e_z point toward its parent. The generated dihedral and a placed reference from the parent's neighborhood determine the first child's radial direction e_x, with $e_y = e_z \times e_x$. Writing $c_j = c_{i;0j}$ and $s_j = \sqrt{1 - c_j^2}$, the first child direction is $u_{i,1} = c_1 e_z + s_1 e_x$. Subsequent directions are constructed as

$$\eta_j = \mathrm{clip}\left(\frac{c_{i;1j} - c_1 c_j}{\max(s_1 s_j, \epsilon)}, -1, 1\right), \qquad u^{\pm}_{i,j} = c_j e_z + s_j \left(\eta_j e_x \pm \sqrt{1 - \eta_j^2}\, e_y\right). \tag{12}$$

The predicted angular cosines need not be mutually realizable. For a non-degenerate local frame with $s_1 s_j \geq \epsilon$, compatibility with the parent-relative angles requires

$$|c_{i;1j} - c_1 c_j| \leq s_1 s_j. \tag{13}$$

Clipping η_j enforces this condition in exact unit arithmetic by relaxing an incompatible first-child cosine while retaining the parent-relative angles. The implementation uses the parent cosine of the normalized first-child direction in place of c_1. When $s_1 s_j < \epsilon$, that exact interval no longer applies; Appendix A.2 gives the regularized formulas. Handedness selects the second child's mirror orientation; if an active sign is required but $H_i = \mathrm{inactive}$, we choose the more probable of $+$ and $-$. Later children select the mirror candidate minimizing the cosine residual to the second child, so additional sibling-angle constraints may remain unsatisfied. Previously placed directions are not revisited.

At the root, the first two child directions define the coordinate frame. If the third child's predicted cosines yield a negative squared out-of-plane component, we set that component to zero and normalize the direction, producing a planar fallback that may relax both angles involving the third child. Numerical degeneracies use deterministic fallbacks detailed in Appendix A.2. Each child is then placed according to

$$x_j = x_i + r_j u_{i,j}, \tag{14}$$

where x_j denotes a Real atom coordinate or a virtual RingRef point. The fast decoder performs no iterative coordinate optimization.

RingRef closure. After constructing all Real atoms and virtual points, we jointly assign each RingRef r to a Real target by minimizing the total virtual-point discrepancy:

$$\pi^\star = \operatorname*{argmin}_{\pi \in \Pi} \sum_{r \in \mathcal{R}} \left\| \widetilde{x}_r - x_{\pi(r)} \right\|_2 . \tag{15}$$

Here, Π contains assignments satisfying element compatibility, BFS consistency, degree limits, and basic bonding constraints. Each assignment adds the corresponding ring edge with bond type B_r, yielding $\mathcal{E} = \mathcal{E}_T \cup \mathcal{E}_R$. If no feasible assignment exists, decoding fails. This step establishes ring connectivity without moving the reconstructed Real atoms.

4 Experiments

4.1 Experimental Setup

We evaluate TreeRef-BFN on QM9 Ramakrishnan et al. (2014) with explicit hydrogens, using the MiDi split and no formal-charge channel, and on GEOM-Drugs Axelrod & Gómez-Bombarelli (2022) with explicit hydrogens, one conformer per molecule, and an $80/10/10$ split. All experiments use $k = 3$. The QM9 Transformer has 16 blocks, hidden size 768, 12 attention heads, and feed-forward width 3072. The GEOM-Drugs Transformer has 16 blocks, hidden size 1024, 16 attention heads, and feed-forward width 4096. Training details are in Appendix B. For property-conditioned generation, a frozen EGNN Satorras et al. (2021) regressor scores each decoded molecule, and we report the mean absolute error between the requested property and this score. Unless noted otherwise, we generate 10,000 samples and use EMA parameters, 100 Bayesian update steps followed by a final prediction at $t = 1$, and the fast decoder. Mask-based structural tasks reuse the unconditional checkpoint, whereas property-conditioned inverse design uses a separately trained model for each target property with CFG scale $w = 2.0$. We report RDKit Landrum et al. (2026) validity, uniqueness, and EDM-style atom and molecule stability, with decoding failures counted as invalid and unstable; uniqueness is computed over valid molecules. Geometric fidelity is evaluated using Wasserstein-1 distances between generated and reference bond-length, bond-angle, and dihedral distributions. Sampling time is measured on one NVIDIA RTX A6000 for generating 10,000 molecules under the corresponding sampling configuration.

Table 1: Unconditional generation results on QM9 with explicit hydrogens. *Results are taken from the corresponding papers because pretrained checkpoints are not publicly available.

			Validity (↑)	Unique (↑)	Stability (↑)		W_1 (↓)		
Model	Steps	Time	RDKit	SMILES	Atom	Mol.	Bond	Angle	Dih.
EDM (Hoogeboom et al., 2022)	1000	104.2	84.08	98.44	98.34	81.53	0.0041	0.762	5.624
GCDM (Morehead & Cheng, 2024)	1000	136	87.03	98.38	98.70	85.79	0.0042	0.843	4.247
GeoLDM (Xu et al., 2023)	1000	101.5	91.36	97.90	98.91	89.90	0.0059	0.745	4.202
ADiT (Joshi et al., 2025)	500	59.6	91.37	96.76	76.43	7.56	0.0438	1.599	5.535
Zatom-1 (Morehead et al., 2026)	100	11.4	91.56	96.69	98.16	84.94	0.0066	0.822	3.590
Symphony++* (Kim et al., 2026)	AR	–	93.8	90.6	93.1	50.5	–	–	–
GeoBFN* (Song et al., 2024)	100	–	93.0	98.4	98.6	87.2	–	–	–
VecMol* (Hua et al., 2026)	1000	–	98.4	98.0	99.8	97.4	0.024	1.01	–
MiDi (Vignac et al., 2023)	1000	162.6	97.83	97.71	98.14	83.57	0.0074	0.701	4.016
JODO (Huang et al., 2023)	1000	151.2	98.90	96.62	99.24	93.46	0.0034	0.699	3.118
SemlaFlow (Irwin et al., 2025)	100	12.9	100.00	95.82	97.13	78.49	0.0113	0.751	3.486
TreeRef-BFN	100	12.3	96.56	96.11	99.19	92.52	0.0040	0.500	1.187
Single-angle geometry	100	12.1	90.12	96.44	97.73	88.27	0.0242	0.745	2.874
Post-hoc ring closure	100	22.4	92.4	97.1	98.2	85.4	0.0038	0.543	1.237
TreeRef-BFN	10	4.5	90.26	88.65	97.83	83.64	0.0132	3.555	7.124
TreeRef-BFN	20	5.2	94.47	94.03	98.75	89.53	0.0081	1.930	3.909
TreeRef-BFN	50	9.3	95.89	95.83	98.99	91.29	0.0055	0.915	2.056
TreeRef-BFN	200	25.5	97.26	96.01	99.26	93.35	0.0033	0.429	1.215
TreeRef-BFN	500	63.1	97.38	96.21	99.30	93.38	0.0031	0.268	0.958
TreeRef-BFN	1000	121.9	97.31	95.84	99.25	93.06	0.0028	0.236	1.058
TreeRef-BFN	2000	239.6	97.53	96.20	99.32	93.67	0.0029	0.226	1.068

4.2 UNCONDITIONAL GENERATION

Molecular Generation on QM9. We evaluate all available baselines under a unified protocol covering RDKit validity, stability, diversity, local geometric fidelity, and sampling efficiency. TreeRef-BFN achieves a strong overall balance across these criteria: its chemical validity and molecular stability are competitive with the strongest baselines, while maintaining high uniqueness and substantially more accurate local geometry. In particular, its bond-, angle-, and dihedral-level distributions closely match the reference data, showing that the generated topology is accompanied by geometrically consistent 3D structures rather than merely chemically valid graphs. The contrast with SemlaFlow is informative: SemlaFlow attains very high RDKit validity, but its lower distance-based molecule stability indicates that the generated topology is more accurate than the corresponding Cartesian geometry. TreeRef-BFN instead maintains strong consistency between chemical structure and local geometry. It also offers a favorable quality–efficiency trade-off: useful samples can already be obtained with very few Bayesian updates, most gains in validity and stability occur within the first 100–200 steps, and larger sampling budgets mainly refine geometric fidelity with diminishing improvements in chemical quality. We therefore use 100 steps as the default setting, which provides competitive generation quality with highly efficient sampling.

Two retrained TreeRef ablations in Table 1 separate these pieces. Single-angle geometry replaces pairwise bond angles with the angle of each child to its parent. Validity falls to 90.12% and molecule stability to 88.27%, so the pairwise angles carry more information and the model learns them more readily. Post-hoc ring closure is a two-stage procedure: the spanning tree is generated first, and a second stage adds the ring edges. It is slower, with sampling time rising from 12.3 to 22.4, and slightly less accurate, with validity 92.4% and molecule stability 85.4%. RingRef keeps ring closure in the same representation, and this unified encoding improves the model's account of rings. Appendix C.2 records both comparisons.

GEOM-Drugs. Table 2 evaluates TreeRef-BFN at 100 steps. The larger and more topologically diverse molecules in GEOM-Drugs reveal a different trade-off from QM9. TreeRef-BFN attains 82.4% RDKit validity, below several baselines, but reaches 80.1% molecule stability and achieves the lowest bond-, angle-, and dihedral-distribution distances. This contrast separates two sources of generation error: producing a fully valid discrete molecular structure becomes more difficult at this scale, whereas the generated connectivity remains strongly aligned with its local three-dimensional geometry. This behavior is consistent with the motivation of TreeRef, which models topology-dependent local geometry explicitly rather than recovering it indirectly from Cartesian coordinates.

Table 2: Unconditional generation on GEOM-Drugs.

		Validity (↑)	Unique (↑)	Stability (↑)		W_1 (↓)		
Model	Steps	RDKit	SMILES	Atom	Mol.	Bond	Angle	Dih.
EDM (Hoogeboom et al., 2022)	1000	41.15	100.00	81.09	0.22	0.0238	5.613	22.418
GCDM (Morehead & Cheng, 2024)	1000	55.51	100.00	87.96	3.79	0.0227	1.479	21.504
GeoLDM (Xu et al., 2023)	1000	45.50	100.00	84.29	0.60	0.0187	3.297	21.222
ADiT (Joshi et al., 2025)	500	58.65	100.00	75.47	0.12	0.0512	2.048	19.374
Zatom-1 (Morehead et al., 2026)	100	79.89	99.90	88.36	5.98	0.0256	1.271	20.227
GeoBFN* (Song et al., 2024)	100	93.05	–	78.89	–	–	–	–
VecMol* (Hua et al., 2026)	1000	89.2	91.2	99.1	80.2	0.024	4.81	–
MiDi (Vignac et al., 2023)	1000	73.07	100.00	81.51	0.37	0.0278	1.495	25.363
JODO (Huang et al., 2023)	1000	90.94	99.98	84.48	1.11	0.0189	0.644	25.150
SemlaFlow (Irwin et al., 2025)	100	100.00	99.99	83.68	1.89	0.0205	1.300	7.488
TreeRef-BFN	100	82.4	99.99	98.1	80.1	0.0104	0.524	3.214

4.3 CONDITIONAL GENERATION

2D-to-3D generation. JODO and MiDi clamp a forward-noised graph while denoising coordinates; SemlaFlow clamps the clean graph at every step. TreeRef-BFN holds the discrete structure fixed (Frozen). Valid is xyz2mol reconstruction followed by RDKit sanitization, over all requested outputs. We report S and M alongside the main model (configurations are in the appendix). Inter-

estingly, Frozen validity decreases from S to M to L, opposite to the trend observed in unconditional generation. Channel-deletion experiments suggest why: during joint generation, the models make strong use of geometry to help determine discrete structure. This is useful when topology and geometry are generated together, but becomes a disadvantage in 2D-to-3D generation, where the topology is fixed and geometry must be generated from scratch. M-inv2 recovers most of this loss; rather than removing the geometry–topology dependence, multi-view consistency makes the model less tied to a particular TreeRef serialization, improving the use of the fixed graph as a condition for geometry generation.

Table 3: Structure-conditioned 3D generation on QM9. Lengths are in Å and angles in degrees.

Method	Conditioning	Steps	Valid ↑	RMSD ↓	Bond W_1 ↓	Angle W_1 ↓	Dihedral W_1 ↓
JODO (Huang et al., 2023)	RePaint-style	1000	88.50	1.223	0.026	0.953	7.823
MiDi (Vignac et al., 2023)	RePaint-style	500	4.40	2.512	1.392	38.801	23.480
SemlaFlow (Irwin et al., 2025)	Frozen	100	63.90	1.849	0.419	11.918	11.806
TreeRef-BFN-S	Frozen	100	93.00	0.902	0.014	0.719	2.966
TreeRef-BFN-M	Frozen	100	78.40	0.804	0.015	1.066	3.553
TreeRef-BFN-M-inv2	Frozen	100	94.55	0.664	0.007	0.383	2.378
TreeRef-BFN	Frozen	100	64.65	1.041	0.062	3.081	5.314

Variable-size fragment completion and editing. Success requires both RDKit validity and preservation of all supplied fragments, while Diff. measures whether the output differs from the source molecule. We remove geometry and handedness at fragment boundaries, so new attachments must be generated rather than copied from the original pose. Performance largely follows how much structure must be rebuilt: linker generation changes only a short bridge and is easiest, whereas scaffold replacement removes the core and is the most challenging. M-inv2 improves Frozen success on all four tasks, most notably from 33.6% to 63.8% for scaffold replacement. This improvement is consistent with multi-view consistency making the model less dependent on a particular TreeRef serialization, which becomes especially important when a retained fragment must be extended into substantially new topology. Full results are reported in Appendix C.6.

Property-conditioned generation. Table 5 evaluates property generation. Because molecular size is determined through node occupancy, the generator can adapt the number of atoms to the target property rather than fixing it in advance. At 100 sampling steps, TreeRef-BFN achieves lower error than NExT-Mol (Liu et al., 2025) and PropMolFlow (Zeng et al., 2026) on the gap, HOMO, and LUMO, while remaining close on the other three properties. At 1,000 steps, it achieves lower error on all six targets. Appendix C.3 reports the full step sweep and results with a TreeRef regressor.

Table 4: Frozen fragment editing at 100 steps. All metrics except Atoms are percentages.

Task	Model	Atoms	V	Success	Diff.
Scaffold decoration	TreeRef-BFN	10.7	79.4	69.7	97.7
	M-inv2		94.0	85.0	97.6
Multi-fragment	TreeRef-BFN	16.2	84.7	62.2	78.9
	M-inv2		92.6	70.5	79.6
Scaffold replacement	TreeRef-BFN	7.5	41.0	33.6	99.5
	M-inv2		71.9	63.8	99.7
Linker generation	TreeRef-BFN	17.0	94.9	92.9	87.8
	M-inv2		97.6	95.3	87.8

Table 5: QM9 property-conditioned generation. MAE (↓).

Method	Steps	α Bohr^3	Gap meV	HOMO meV	LUMO meV	μ D	C_v cal/(mol K)
EDM	1000	2.76	655	356	584	1.111	1.101
GeoLDM	1000	2.37	587	340	522	1.108	1.025
GCDM	1000	1.97	602	344	479	0.844	0.689
GeoBFN	500	2.34	577	328	516	0.998	0.949
JODO	1000	1.42	335	226	256	0.628	0.581
NExT-Mol	100	1.16	297	205	235	0.507	0.512
PropMolFlow	100	1.36	391	246	312	0.632	0.640
TreeRef-BFN	100	1.18	225	142	169	0.531	0.513
TreeRef-BFN	1000	1.14	213	135	162	0.495	0.493

5 Conclusion

We propose TreeRef, a k-ary representation that jointly encodes molecular size, topology, and local geometry, and TreeRef-BFN, a Bayesian-flow generator with a standard Transformer. Occupancy determines size, RingRef encodes non-tree bonds, and masking supports unconditional, 2D-to-3D, and fragment completion; property conditioning enables inverse design. Improving serialization robustness, structural flexibility, and chemical coverage is left for future work.

REFERENCES

Sungsoo Ahn, Binghong Chen, Tianzhe Wang, and Le Song. Spanning tree-based graph generation for molecules. In *International Conference on Learning Representations (ICLR)*, 2022.

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

Ian Dunn and David R. Koes. FlowMol3: Flow matching for 3D de novo small-molecule generation. *arXiv preprint arXiv:2508.12629*, 2025.

Alex Graves, Rupesh Kumar Srivastava, Timothy Atkinson, and Faustino Gomez. Bayesian flow networks. *arXiv preprint arXiv:2308.07037*, 2023.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. URL https://arxiv.org/abs/2207.12598.

Emiel Hoogeboom, Víctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3D. In *International Conference on Machine Learning (ICML)*, 2022.

Yuchen Hua, Xingang Peng, Jianzhu Ma, and Muhan Zhang. VecMol: Vector-field representations for 3D molecule generation. In *International Conference on Machine Learning (ICML)*, 2026.

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.

Ross Irwin, Alessandro Tibo, Jon Paul Janet, and Simon Olsson. SemlaFlow: Efficient 3D molecular generation with latent attention and equivariant flow matching. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2025.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. In *International Conference on Machine Learning (ICML)*, 2018.

Bowen Jing, Gabriele Corso, Jeffrey Chang, Regina Barzilay, and Tommi Jaakkola. Torsional diffusion for molecular conformer generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Alexia Jolicoeur-Martineau, Aristide Baratin, Kisoo Kwon, Boris Knyazev, and Yan Zhang. Any-property-conditional molecule generation with self-criticism using spanning trees. *Transactions on Machine Learning Research*, 2025.

Chaitanya K. Joshi, Xiang Fu, Yi-Lun Liao, Vahe Gharakhanyan, Benjamin Kurt Miller, Anuroop Sriram, and Zachary Ward Ulissi. All-atom diffusion transformers: Unified generative modelling of molecules and materials. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 28393–28417. PMLR, 2025.

Song Kim, Ameya Daigavane, Parthasarathy Suryanarayanan, and Tess Smidt. Autoregressive models enable efficient conditional 3d molecular generation, 2026. URL https://openreview.net/forum?id=FyMNdLs7CO. ICML 2026 GenBio Workshop, Spotlight.

Greg Landrum et al. RDKit: Open-source cheminformatics. https://www.rdkit.org, 2026. version 2026.03.6.

Yunqing Liu, Yi Zhou, and Wenqi Fan. Geometric flow matching for molecular conformation generation via manifold decomposition. *arXiv preprint arXiv:2605.25577*, 2026.

Zhiyuan Liu, Yanchen Luo, Han Huang, Enzhi Zhang, Sihang Li, Junfeng Fang, Yaorui Shi, Xiang Wang, Kenji Kawaguchi, and Tat-Seng Chua. NExT-Mol: 3d diffusion meets 1d language modeling for 3d molecule generation. In *The Thirteenth International Conference on Learning Representations*, 2025.

Youzhi Luo and Shuiwang Ji. An autoregressive flow model for 3D molecular geometry generation from scratch. In *International Conference on Learning Representations (ICLR)*, 2022.

Alex Morehead and Jianlin Cheng. Geometry-complete diffusion for 3D molecule generation and optimization. *Communications Chemistry*, 7:150, 2024. doi: 10.1038/s42004-024-01233-z.

Alex Morehead, Miruna Cretu, Antonia Panescu, Rishabh Anand, Maurice Weiler, Tynan Perez, Samuel Blau, Steven Farrell, Wahid Bhimji, Anubhav Jain, Hrushikesh Sahasrabuddhe, Pietro Lio, Tommi Jaakkola, Rafael Gomez-Bombarelli, Rex Ying, N. Benjamin Erichson, and Michael W. Mahoney. Zatom-1: A multimodal flow foundation model for 3d molecules and materials. *arXiv preprint arXiv:2602.22251*, 2026.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1:140022, 2014.

Víctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In *International Conference on Machine Learning (ICML)*, 2021.

Yuxuan Song, Jingjing Gong, Yanru Qu, Hao Zhou, Mingyue Zheng, Jingjing Liu, and Wei-Ying Ma. Unified generative modeling of 3D molecules via Bayesian Flow Networks. In *International Conference on Learning Representations (ICLR)*, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Clément Vignac, Nagham Osman, Laura Toni, and Pascal Frossard. MiDi: Mixed graph and 3D denoising diffusion for molecule generation. In *Machine Learning and Knowledge Discovery in Databases: Research Track (ECML PKDD)*, 2023.

Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. GeoDiff: A geometric diffusion model for molecular conformation generation. In *International Conference on Learning Representations (ICLR)*, 2022.

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 (ICML)*, 2023.

Cheng Zeng, Jirui Jin, Connor Ambrose, George Karypis, Mark Transtrum, Ellad B. Tadmor, Richard G. Hennig, Adrian Roitberg, Stefano Martiniani, and Mingjie Liu. PropMolFlow: Property-guided molecule generation with geometry-complete flow matching. *Nature Computational Science*, 6:233–242, 2026. doi: 10.1038/s43588-025-00946-y.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing, code implementation, debugging, and document preparation. All AI-assisted code was reviewed, tested, and verified by the authors. The methodological design, experimental setup, analysis of results, and scientific conclusions were determined by the authors. The authors take full responsibility for the content of this paper.

A TREEREF DETAILS

A.1 CONSTRUCTION DETAILS

Construction must give each slot and geometric channel a reproducible meaning: the decoder needs to know which atom, bond, or ordered set of directions a prediction describes. We therefore first fix the traversal and reference conventions, then encode topology and geometry using those conventions.

Canonical ordering. Different roots and neighbor orders can give the same molecule different parent relations and geometric references. To make this choice reproducible, we order atoms by

graph-derived WL ranks and use input indices to resolve remaining ties. Let $\mathcal{G} = (V, \mathcal{E})$ be a connected molecular graph, $N = |V|$, $\mathcal{N}(v)$ its neighbor set, b_{uv} the bond order, and $\iota(v)$ the input atom index. Bond-aware WL refinement is defined by

$$\begin{aligned} \zeta_v^{(0)} &= \left(\mathrm{idx}(E_v), |\mathcal{N}(v)|, \sum_{u \in \mathcal{N}(v)} b_{uv} \right), \\ \zeta_v^{(s+1)} &= \mathrm{rank}_{\mathrm{lex}}\Big(\zeta_v^{(s)}, \mathrm{sort}_{\mathrm{lex}}\big[(b_{uv}, \zeta_u^{(s)}) : u \in \mathcal{N}(v) \big] \Big). \end{aligned} \tag{16}$$

Each refinement step incorporates the neighboring atoms' current labels and their connecting bond orders. The operator $\mathrm{rank}_{\mathrm{lex}}$ assigns consecutive integers to distinct signatures in lexicographic order. The serialization element order is C, N, O, F, H, S, B, Al, Si, P, Cl, As, Br, I, Hg, Bi, independently of the model's element-vocabulary order. Refinement stops at an unchanged label vector or after N iterations; write w_v for the resulting rank.

The root has only k child slots, whereas every other Real atom can additionally use its parent edge. This gives the capacity conditions $\max_v |\mathcal{N}(v)| \le k+1$ and $V_0 = \{v : |\mathcal{N}(v)| \le k\} \neq \varnothing$. Among admissible roots, minimum eccentricity minimizes the maximum Real-atom depth of the BFS tree. With $\mathrm{dist}_{\mathcal{G}}$ denoting unweighted graph distance, the root and neighbor order are

$$v_\circ = \operatorname*{arg\,min}^{\mathrm{lex}}_{v \in V_0} \left(\max_{u \in V} \mathrm{dist}_{\mathcal{G}}(v,u), w_v, \iota(v) \right), \qquad u \prec v \iff (w_u, \iota(u)) <_{\mathrm{lex}} (w_v, \iota(v)). \tag{17}$$

BFS processes neighbors in order $\prec$, marks each atom on enqueue, and assigns its first discoverer as parent $p(v)$. Children occupy consecutive slots in discovery order. Canonical serialization is therefore deterministic for fixed input indexing, without implying invariance to atom reindexing.

RingRef placement. A spanning tree omits the non-tree bonds, but adding another Real atom for an omitted bond would duplicate a physical atom. We therefore encode each omitted bond with one RingRef leaf pointing to an existing atom. Let $d(v)$ be BFS depth and m_v the number of child slots already occupied, initially the number of Real children. Process $\mathcal{E}_R = \mathcal{E} \setminus \mathcal{E}_T$ by the lexicographic key $(w_u, w_v, \iota(u), \iota(v))$, orienting each input edge so that $\iota(u) < \iota(v)$. For each edge,

$$A_{uv} = \{a \in \{u,v\} : m_a < k\}, \qquad h = \operatorname*{arg\,min}^{\mathrm{lex}}_{a \in A_{uv}} (-d(a), w_a, \iota(a)), \qquad \{t\} = \{u,v\} \setminus \{h\}. \tag{18}$$

The set A_{uv} enforces storage capacity. The key $(-d(a), w_a, \iota(a))$ prefers the deeper endpoint as host, then the WL rank and input index. The RingRef r has host $h(r) = h$, target $t(r) = t$, and child index m_h, after which $m_h \leftarrow m_h + 1$. Denote the occupied RingRef-slot set by $\mathcal{R}$. An empty A_{uv} causes construction failure. Randomized variants replace neighbor order by a random permutation and choose $h \sim \mathrm{Unif}(A_{uv})$; the randomized-root variant additionally uses $v_\circ \sim \mathrm{Unif}(V_0)$. Geometric channels are recomputed for each serialization.

Logical and packed indices. Logical addresses specify parent–child relations, but storing every intervening address would allocate unused tree branches. We instead pack the child blocks of Real atoms consecutively, preserving their order. Identify each physical atom with its Real slot. Let $\lambda(s)$ and $\kappa(s)$ denote a slot's logical address and packed position, respectively, and let $v_0 = v_\circ, v_1, \ldots, v_{N-1}$ be the Real BFS queue. For child index $j \in \{0, \ldots, k-1\}$,

$$\begin{aligned} \lambda(v_\circ) &= \kappa(v_\circ) = 0, \\ \lambda(\mathrm{child}(v_q, j)) &= k\lambda(v_q) + j + 1, \qquad \kappa(\mathrm{child}(v_q, j)) = 1 + kq + j. \end{aligned} \tag{19}$$

The first recurrence preserves the tree relation; the second gives its position in the model tensor. Only occupied Real slots enter the queue; every such slot contributes a full child block, including terminal all-Null blocks. Thus $L_N = 1 + kN$ before cropping. For the collection $\mathcal{D}$ of cached training, validation, and test serializations, including stored training variants, define

$$m_\star = \max_{Z \in \mathcal{D}} \max\{i : E_i(Z) \neq \mathrm{Null}\}, \qquad L' = 1 + k \left\lceil \frac{m_\star}{k} \right\rceil \le 1 + kN_{\max}. \tag{20}$$

The common prefix $[0, L')$ contains every occupied Real and RingRef slot in $\mathcal{D}$. Removing an internal Null would shift the meaning of later child positions, so internal Nulls remain supervised states.

Only the shared trailing suffix is removed, at a complete block boundary; the decoder interprets omitted child blocks as Null.

Ordered local stars. An angular value is meaningful only if its two directions are identified, and a handedness sign additionally depends on their order. We therefore assign geometric channels to explicitly ordered neighborhoods. Let $C(v)$ be the occupied child-slot list ordered by λ, and define $\chi(s) = s$ for a Real slot and $\chi(s) = t(s)$ for a RingRef. With $\|$ denoting list concatenation, the unpadded stars are

$$\begin{aligned} S_v &= \begin{cases} \chi(C(v)), & v = v_\circ, \\ (p(v))\|\chi(C(v)), & v \neq v_\circ, \end{cases} \\ S_r &= (h(r))\|\operatorname{sort}_{(\lambda,w,\iota)}(\mathcal{N}(t(r)) \setminus \{h(r)\}), \qquad r \in \mathcal{R}. \end{aligned} \tag{21}$$

Apply χ entrywise and pad each list to $k+1$ entries with $\perp$. The center is $o_v = v$ for Real slots and $o_r = t(r)$ for RingRefs. Define $\Delta_{i,a} = x_{S_{i,a}} - x_{o_i}$ for nonempty entries. The angular channels, indexed lexicographically by $0 \le a < b \le k$, are

$$c_{i;ab} = \begin{cases} \dfrac{\Delta_{i,a}^\top \Delta_{i,b}}{\|\Delta_{i,a}\|_2 \|\Delta_{i,b}\|_2}, & S_{i,a}, S_{i,b} \neq \perp, \\ 0, & \text{otherwise.} \end{cases} \tag{22}$$

For a Real atom, this star supplies the directions needed to place its construction children, so it can be a strict subset of the physical neighborhood. A RingRef instead carries descriptors centered on the referenced atom; its host already supplies the local directions used to place the virtual leaf. This distinction determines which references are used during decoding.

Distances and torsions. Bond lengths and angles specify a local shape, but a non-root star can still rotate about its parent bond. The dihedral fixes that rotation relative to the parent's neighborhood. Distances satisfy $r_v = \|x_v - x_{p(v)}\|_2$ for $v \neq v_\circ$, $r_r = \|x_{t(r)} - x_{h(r)}\|_2$ for RingRefs, and $r_{v_\circ} = 0$. Let $f(S, a)$ return the first entry of S outside $\{a, \perp\}$. The dihedral quadruple is

$$T_i = \begin{cases} (f(S_{p(v)}, v), p(v), v, f(S_v, p(v))), & i = v \neq v_\circ, \\ (f(S_{h(r)}, t(r)), h(r), t(r), f(S_r, h(r))), & i = r \in \mathcal{R}. \end{cases} \tag{23}$$

The first available references make this rotational convention reproducible. For $T_i = (a, b, c, d)$, let $n = (x_c - x_b)/\|x_c - x_b\|_2$, $v = (I - nn^\top)(x_a - x_b)$, and $w = (I - nn^\top)(x_d - x_c)$. These projections compare the two reference directions in the plane perpendicular to the central bond. The stored pair is

$$q_i = (\cos\phi_i, \sin\phi_i) = \frac{\left(v^\top w, (n \times v)^\top w\right)}{\|v\|_2 \|w\|_2}. \tag{24}$$

Set $q_i = (0, 0)$ at the root, for a missing reference, or if any of $\|x_c - x_b\|_2, \|v\|_2, \|w\|_2$ is below 10^{-8}. The zero pair marks an undefined rotation without assigning an arbitrary torsion. Only positive distances are standardized: $\widetilde{r}_i = (r_i - \mu_r)/\sigma_r$, with μ_r and σ_r fitted to accepted training bond lengths and $\sigma_r \ge 10^{-6}$. Root and Null distances remain zero.

Handedness. Reflection preserves pairwise angular cosines, so these cosines alone cannot distinguish the two mirror orientations of a local star. Handedness supplies this discrete choice. We use the first three star entries, with no substitution of later neighbors. If any is missing or $\|\Delta_{i,a}\|_2 \le 10^{-12}$, set $H_i = \text{inactive}$. Otherwise, normalize these directions to u_0, u_1, u_2, let $s_i = u_0 \cdot (u_1 \times u_2)$, and clip their pairwise dot products to $(a, b, c) \in [-1, 1]^3$, in the order $(01, 02, 12)$. The stable magnitude and label are

$$\tau_i = \sqrt{\max\{0, 1 + 2abc - a^2 - b^2 - c^2\}}, \qquad H_i = \begin{cases} \text{inactive}, & \tau_i \le 10^{-3}, \\ +, & \tau_i > 10^{-3},\ s_i \ge 0, \\ -, & \tau_i > 10^{-3},\ s_i < 0. \end{cases} \tag{25}$$

For exact unit directions, $\tau_i = |s_i|$. Near coplanarity, a small perturbation can change the sign of the triple product, so the threshold assigns such neighborhoods the inactive state. Thus, H_i describes a stable orientation of the prescribed ordered references when one is available; it is not an atom-order-independent stereochemical designation.

A.2 Decoding Details

Predicted channels need not jointly satisfy the structural and geometric relations imposed during encoding. Decoding therefore first recovers a consistent occupied tree, then constructs coordinates under a fixed constraint priority, and finally resolves RingRef targets. The optional chemical stage then adjusts element and bond labels on the recovered connectivity.

Occupancy projection. The role, element, and bond channels can disagree: a reachable occupied slot may still predict no parent bond. We resolve occupancy first because it determines which slots and bonds exist. For single-tree decoding, let $\mathsf{Read}_A(\eta)$ apply the final sampling or argmax rule to logits restricted to classes A. After setting $R_0 = \text{Real}$ and $E_0 = \mathsf{Read}_{\mathcal{A}}(\eta_0^E)$, the Real queue defines the reachable occupied set $\mathcal{O}$, with Null and RingRef slots unexpanded. For the read-out parent bond $\widehat{B}_i$,

$$B_i^{\text{dec}} = \begin{cases} \text{None}, & i = 0 \text{ or } i \notin \mathcal{O}, \\ \mathsf{Read}_{\mathcal{B}}(\eta_i^B), & i \in \mathcal{O} \setminus \{0\},\ \widehat{B}_i = \text{None}, \\ \widehat{B}_i, & \text{otherwise.} \end{cases} \tag{26}$$

Here, $\mathcal{B}$ excludes None. Reachability depends on R, E, so it can retain a slot whose bond still requires correction. For $i \notin \mathcal{O}$, set $(E_i, G_i, H_i, Q_i) = (\text{Null}, 0, \text{inactive}, 0)$, omitting Q when disabled.

Inverse packing and traversal order. Packed neighbors in the tensor are not necessarily neighbors in the molecular graph. To recover the intended relations, each queued Real slot consumes exactly one child block. This inverts the packed BFS map and determines the reachable set $\mathcal{O}$. Let $\Lambda(i)$ be the logical address of packed position i, $P(i)$ its packed parent, and $d(i)$ its depth. Initialize $\mathcal{Q}^{(0)} = (0)$, $\Lambda(0) = 0$, and $d(0) = 0$. At step $q < |\mathcal{Q}^{(q)}|$, the q-th queued Real slot $p_q = \mathcal{Q}^{(q)}[q]$ owns the next complete block. For $j = 0, \ldots, k-1$,

$$\begin{aligned} i_{qj} &= 1 + kq + j, \qquad P(i_{qj}) = p_q, \\ \Lambda(i_{qj}) &= k\Lambda(p_q) + j + 1, \qquad d(i_{qj}) = d(p_q) + 1, \\ \mathcal{Q}^{(q+1)} &= \mathcal{Q}^{(q)} \| \big(i_{qj} : j = 0, \ldots, k-1,\ R_{i_{qj}} = \text{Real},\ E_{i_{qj}} \neq \text{Null}\big). \end{aligned} \tag{27}$$

In this recurrence, q selects the parent block and j selects a child within that block. The appended entries retain increasing j. Every block consumes k positions, including Nulls; occupied RingRefs acquire a host $P(i)$ but never enter the queue. If $1 + kq = L'$, all remaining child blocks are implicit Null. A partially available block is invalid. When the Real queue terminates, any unconsumed suffix is unreachable. The resulting Λ is the inverse of the construction-side packed map on visited packed slots. Thus, root and sibling positions are inherited from the serialization; WL ranks are required only for encoding. Real atoms are processed in the recovered BFS order, and each parent's children are placed in increasing logical-address order before the next Real atom is processed.

Geometric domains. Coordinate construction requires bounded angular cosines and usable distance and torsion values. We therefore first apply scalar domain corrections. For occupied non-root distances and available angular channels, use

$$\begin{aligned} r_i^{\text{dec}} &= \text{clip}(\sigma_r \widehat{\widetilde{r}}_i + \mu_r, 0.5, 3.0)\ \text{Å}, \qquad c_{i;ab} = \text{clip}(\widehat{c}_{i;ab}, -1, 1), \\ q_i^{\text{dec}} &= \widehat{q}_i / \|\widehat{q}_i\|_2 \quad \text{if } \|\widehat{q}_i\|_2 \geq \epsilon, \qquad \epsilon = 10^{-8}. \end{aligned} \tag{28}$$

A smaller dihedral norm provides no azimuthal constraint. However, valid individual values do not ensure a valid joint geometry: for example, three directions cannot all be pairwise opposite, even though each requested cosine -1 lies in the valid interval. For m_i available directions, form the symmetric target cosine matrix C_i with unit diagonal. Exact realization by three-dimensional directions requires

$$C_i = U_i^\top U_i \text{ for some } U_i \in \mathbb{R}^{3 \times m_i} \quad \Longleftrightarrow \quad C_i \succeq 0,\ \text{diag}(C_i) = \mathbf{1},\ \text{rank}(C_i) \leq 3. \tag{29}$$

Positive semidefiniteness expresses consistency of the dot products, and the rank bound restricts their realization to three dimensions. Because scalar clipping does not enforce either condition, conflicting angle predictions still require a decision about which constraints to retain. We use a fixed priority so that coordinates can be constructed in one pass, with scalar clipping at non-root nodes and a planar fallback at the root. The following formulas specify these choices and the constraints they may relax. Write $\mathsf{n}(v) = v/(\|v\|_2 + \delta)$ with $\delta = 10^{-12}$.

Decoder neighborhood indices. Recovering the parent–child tree does not by itself identify the angular channel for each pair of directions. Because channel lookup groups children by role while placement follows logical addresses, the decoder also needs an explicit index map. Let $C_{\mathrm{Real}}(v)$ and $C_{\mathrm{RingRef}}(v)$ list the occupied children of each role in logical-address order. The lookup star and placement order are

$$\widehat{S}_v = (p(v))\|C_{\mathrm{Real}}(v)\|C_{\mathrm{RingRef}}(v), \qquad C(v) = \mathrm{sort}_\lambda(C_{\mathrm{Real}}(v) \cup C_{\mathrm{RingRef}}(v)), \tag{30}$$

omitting the parent at the root. If $\alpha_v(s)$ is the index of slot s in $\widehat{S}_v$, define

$$\gamma_v(s,t) = c_{v;ab}, \qquad (a,b) = \big(\min\{\alpha_v(s),\alpha_v(t)\}, \max\{\alpha_v(s),\alpha_v(t)\}\big). \tag{31}$$

RingRef entries denote virtual slots. Encoding gives identical child orders because Real children precede RingRefs. Generated roles can change their relative order; α_v retrieves the corresponding angular channel without reordering the packed sequence. The first already placed reference in $\widehat{S}_{p(v)} \setminus \{v\}$ anchors the non-root dihedral frame, whereas handedness reads channels $(01, 02, 12)$ in the lookup-star order. At the root, no parent entry is reserved. After matching, a RingRef's target-centered references are recovered as $\widehat{S}_r = (h(r))\|\,\mathrm{sort}_\lambda(\mathcal{N}_{\mathrm{dec}}(t(r)) \setminus \{h(r)\})$, where $\mathcal{N}_{\mathrm{dec}}$ includes tree and recovered ring neighbors. Logical addresses are unique, so this sorting agrees with the construction-side (λ, w, ι) order.

Construction signs. The predicted geometry and handedness can disagree about whether a mirror choice is needed. We first decide from the available references and predicted cosines whether the star is non-degenerate, then use the handedness prediction to choose a sign. Let $a_i = \mathbf{1}[\text{three prescribed references exist}]\,\mathbf{1}[\widehat{\tau}_i > 10^{-3}]$, where $\widehat{\tau}_i$ is computed from clipped predicted cosines by Eq. equation 25. For Real slots, the sign is

$$s_i^H = \begin{cases} +1, & a_i = 0, \\ \mathrm{sgn}(H_i), & a_i = 1,\ H_i \in \{+,-\}, \\ \underset{s\in\{+1,-1\}}{\arg\max}\ \eta_i^H(s), & a_i = 1,\ H_i = \text{inactive}. \end{cases} \tag{32}$$

Here, $\mathrm{sgn}(+) = +1$ and $\mathrm{sgn}(-) = -1$. In the last case, an inactive label cannot resolve the required mirror choice, so the decoder selects the more probable active class. Ties favor $+1$; if the required logits are unavailable, decoding fails. Degenerate stars use the fixed positive sign. This generated-sample policy selects one construction without branch-mirror enumeration.

Root construction. Internal geometry leaves the molecule's global position and orientation unspecified. We fix this freedom by placing the root at the origin, its first child along the x-axis, and its second child in the xy-plane. Set $x_{v_\circ} = 0$. Number its m occupied children $s^{(1)}, \ldots, s^{(m)}$ in placement order and write $g_{ab} = \gamma_{v_\circ}(s^{(a)}, s^{(b)})$. The first three directions are

$$\begin{aligned} u_1 &= (1,0,0), \qquad u_2 = (g_{12}, \sqrt{\max(1-g_{12}^2, 0)}, 0), \\ y &= \frac{g_{23} - g_{12}g_{13}}{\max\{\sqrt{\max(1-g_{12}^2,0)}, \epsilon\}}, \\ u_3 &= \mathsf{n}\left(g_{13}, y, s_{v_\circ}^H \sqrt{\max(1-g_{13}^2-y^2, 0)}\right). \end{aligned} \tag{33}$$

Use only directions corresponding to existing children. For $\sqrt{1-g_{12}^2} \geq \epsilon$, the untruncated squared height is

$$\xi_3 = 1 - g_{13}^2 - y^2 = \frac{1 + 2g_{12}g_{13}g_{23} - g_{12}^2 - g_{13}^2 - g_{23}^2}{1 - g_{12}^2}. \tag{34}$$

The numerator is the determinant of the first three children's target Gram matrix. A negative ξ_3 would require a negative squared height, so no third direction can satisfy all three requested cosines while retaining the first two directions. We therefore set the height to zero and normalize: the third direction becomes $\mathsf{n}(g_{13}, y, 0)$, while u_1, u_2 remain fixed. With $a = \sqrt{g_{13}^2 + y^2}$, its realized angular cosines are $g_{13}^{\mathrm{rec}} = g_{13}/a$ and $g_{23}^{\mathrm{rec}} = g_{23}/a$; both predictions can therefore change. This is a deterministic planar fallback, without a minimum-error guarantee over all three angles. If the

first two directions are nearly collinear, the denominator bound ϵ in the construction formula applies instead. For $j = 4, \ldots, m$ when $k > 3$,

$$\alpha_j = \frac{2\pi(j-1)}{m}, \qquad u_j = \mathsf{n}\Big(g_{1j}, \sqrt{1-g_{1j}^2}\cos\alpha_j, \sqrt{1-g_{1j}^2}\sin\alpha_j\Big). \tag{35}$$

Non-root frame. At a non-root atom, the already placed parent supplies an axis, but an axis alone leaves the child's azimuth undetermined. A reference from the parent's neighborhood and the generated dihedral supply that azimuth. At Real atom i with parent p, let $e_z = \mathsf{n}(x_p - x_i)$ and $n = -e_z$. Choose the first already placed reference a in $\widehat{S}_p$ other than i, using virtual coordinates for RingRef entries. Define $w = (I - nn^\top)(x_a - x_p)$ and

$$e_x = \begin{cases} \mathsf{n}\big(q_{i,1}^{\mathrm{dec}} v + q_{i,2}^{\mathrm{dec}}(n \times v)\big), & v = w/\|w\|_2, \quad a \text{ exists, } \|w\|_2 \geq \epsilon, \ q_i^{\mathrm{dec}} \text{ available,} \\ \mathsf{n}(e_z \times a_0), & \text{otherwise,} \end{cases} \tag{36}$$

If the reference or torsion is unusable, the Cartesian fallback still provides a perpendicular direction. We choose $a_0 = (0,0,1)$ if $|e_z^\top(0,0,1)| \leq 0.9$ and $a_0 = (0,1,0)$ otherwise, avoiding a nearly parallel cross-product reference. Set $e_y = e_z \times e_x$.

Child placement. Fixing a child's angle to the parent confines its direction to a cone around the parent axis. Its angle to the first child then restricts the remaining azimuth to two mirror candidates. This reduction lets us construct a new direction while keeping earlier directions fixed. For children $s^{(j)}$ in placement order, write $c_j = \gamma_i(p, s^{(j)})$, $s_j = \sqrt{\max(1-c_j^2, 0)}$, and $g_{ab} = \gamma_i(s^{(a)}, s^{(b)})$. The first direction is $u_1 = \mathsf{n}(c_1 e_z + s_1 e_x)$. To retain the numerical normalization used in the implementation, define $\bar{c}_1 = u_1^\top e_z$ and $\bar{s}_1 = \sqrt{\max(1-\bar{c}_1^2, 0)}$. For $j \geq 2$ with $s_j > \epsilon$,

$$\eta_j = \mathrm{clip}\left(\frac{g_{1j} - \bar{c}_1 c_j}{\max(\bar{s}_1 s_j, \epsilon)}, -1, 1\right), \qquad v_j^\sigma = c_j e_z + s_j\left(\eta_j e_x + \sigma\sqrt{1-\eta_j^2}\, e_y\right). \tag{37}$$

The selected directions are

$$u_j = \begin{cases} \mathsf{n}(c_j e_z + s_j e_x), & s_j \leq \epsilon, \\ \mathsf{n}(v_2^{s_i^H}), & j = 2, \ s_j > \epsilon, \\ \mathsf{n}(v_j^{\sigma_j}), \quad \sigma_j = \underset{\sigma\in\{+1,-1\}}{\arg\min} \left|(v_j^\sigma)^\top u_2 - g_{2j}\right|, & j \geq 3, \ s_j > \epsilon. \end{cases} \tag{38}$$

Ties favor $+1$. The construction prioritizes the parent-relative cosine, the first-child cosine, and then the cosine to child 2. For every Real or virtual child slot s of i, assign $x_s = x_i + r_s^{\mathrm{dec}} u_s$ once. Finally, apply $x_s \leftarrow x_s - N^{-1}\sum_{v\in V} x_v$ to all Real and virtual positions.

Non-root Gram conflicts. The two prescribed parent-relative angles limit how close or far apart two child directions can be. A predicted sibling angle outside that range cannot be satisfied unless a higher-priority angle is changed. The preceding clipping rule retains the parent-relative angles and adjusts the sibling cosine to its feasible range. To derive that range, in exact unit arithmetic consider the parent and child directions $1, j$, whose target Gram matrix is

$$C_{1j} = \begin{pmatrix} 1 & c_1 & c_j \\ c_1 & 1 & g_{1j} \\ c_j & g_{1j} & 1 \end{pmatrix}, \qquad \det C_{1j} = s_1^2 s_j^2 - (g_{1j} - c_1 c_j)^2. \tag{39}$$

For fixed parent-relative cosines $c_1, c_j \in [-1, 1]$, feasibility is equivalent to $g_{1j} \in \mathcal{I}_j$, where

$$\mathcal{I}_j = [c_1 c_j - s_1 s_j, \ c_1 c_j + s_1 s_j], \qquad g_{1j}^{\mathrm{proj}} = \underset{g\in\mathcal{I}_j}{\arg\min}\, |g - g_{1j}|^2 = \mathrm{clip}(g_{1j}, \mathcal{I}_j). \tag{40}$$

A cosine already inside $\mathcal{I}_j$ is retained; one outside is moved to the nearest endpoint. When $s_1 s_j \geq \epsilon$, clipping η_j implements this scalar projection in the unit-vector idealization. Thus, the change is confined to the incompatible sibling constraint while the prescribed parent-relative angles are retained. The preceding direction formulas specify the δ-regularized implementation. For $s_1 s_j < \epsilon$, the regularized denominator is used and the result need not equal this exact interval projection; a child with $s_j \leq \epsilon$ takes the deterministic radial fallback already specified.

For $j \geq 3$ with $s_j > \epsilon$, the remaining cosine to child 2 is tested only over the two available candidates. Its minimum residual is

$$\rho_{2j} = \min_{\sigma \in \{+1,-1\}} \left| (v_j^{\sigma})^{\top} u_2 - g_{2j} \right|. \tag{41}$$

The two candidates may both disagree with the second-child cosine. We then keep the one with smaller residual, accepting a positive ρ_{2j} so that earlier directions and higher-priority choices remain fixed. Other angles between later children do not enter this selection. Thus, the priority is parent-relative cosine, first-child cosine, then the residual to child 2, with handedness selecting child 2's mirror branch. For normalized reconstructed directions $\bar{U}_i$, the realized matrix $C_i^{\mathrm{rec}} = \bar{U}_i^{\top} \bar{U}_i$ satisfies the Gram constraints but need not equal C_i. This correction is induced by sequential coordinate construction; it does not minimize a global matrix or angular reconstruction objective.

RingRef candidates. A virtual RingRef point predicts where its target should lie, but does not directly identify that Real atom. Distance alone could select an atom incompatible with the encoded tree, so candidate targets are first filtered by element, connectivity, degree, and BFS constraints. Let $\mathcal{T}_m(a, b) = 1$ indicate that element b occurs under a in the implementation's order-m bond table. Define the physical degree limit $D_v = k + \mathbf{1}[v \neq v_\circ]$. For RingRef r with host $h = h(r)$, the target set is

$$\begin{aligned} \mathcal{C}_r = \{ t \in V : E_t = E_r,\ t \neq h,\ \{h, t\} \notin \mathcal{E}_T,\ \mathcal{T}_1(E_h, E_t) = 1, \\ \deg_T(t) < D_t, \quad |d(t) - d(h)| \leq 1, \\ d(t) = d(h) - 1 \Rightarrow \lambda(t) \geq \lambda(p(h)), \\ d(t) = d(h) + 1 \Rightarrow \lambda(h) \geq \lambda(p(t)) \}. \end{aligned} \tag{42}$$

The depth conditions follow from BFS: an edge can connect only the same or adjacent levels. The address conditions exclude a target that would have discovered the deeper atom before its recorded parent. These filters use no virtual-to-target distance threshold and impose no equality between RingRef and target charges. A virtual point's direction is constructed from its host's geometry; its own distance gives its displacement from the host.

Joint ring closure. Candidate choices share physical degree budgets, so individually nearest targets can form an invalid combination. We therefore minimize the total virtual-point discrepancy over jointly admissible assignments, as in the main-text objective. For a target mapping π, define $e_r(\pi) = \{h(r), \pi(r)\}$ and $\mathcal{E}_R(\pi) = \{e_r(\pi) : r \in \mathcal{R}\}$. Every feasible assignment $\pi \in \Pi$ in the main-text distance objective satisfies

$$\pi(r) \in \mathcal{C}_r \quad (\forall r \in \mathcal{R}), \qquad \deg_{\mathcal{E}_T \cup \mathcal{E}_R(\pi)}(v) \leq D_v \quad (\forall v \in V). \tag{43}$$

A Real atom can participate in several distinct ring edges, so these constraints do not require π to be injective. The decoder processes RingRefs by $\lambda(r)$ and candidates by $(\|\widetilde{x}_r - x_t\|_2, \lambda(t))$, pruning a partial assignment whenever $C_{\mathrm{partial}} \geq C_{\mathrm{best}}$. Under the fast decoder, an empty Π causes structural decoding failure.

Ring closure recovers missing bonds using the constructed coordinates. With coordinate fusion disabled, closure leaves X fixed, so a residual virtual-to-target discrepancy is retained as geometric error. In the default single-orientation decoder, a RingRef's target-centered angles, dihedral, and handedness do not modify π or X. After matching, the active-sign requirement is checked from its predicted cosines and the resolved target's available neighbors, applying the same inactive-to-active policy when needed. No constraint equates the predicted RingRef sign with the sign recomputed from X.

Isolated RingRef promotion. The optional `rules` decoder applies one structural repair before the joint assignment above. For each RingRef r, let $\widetilde{x}_r$ be its virtual point and E_r its element. Using the order-1 bond length $L_1(E_r, E_r)$ from `dataset/bond_analyze.py`, converted from picometres to angstroms, declare r isolated when no Real atom of element E_r lies inside the window

$$\|\widetilde{x}_r - x_t\|_2 \leq L_1(E_r, E_r)/100 + 0.10. \tag{44}$$

The 0.10 Å margin is the table's `margin1` of 10 pm. If E_r occurs in the table but not paired with itself, the window uses the longest order-1 length tabulated for E_r, still plus 0.10 Å. If E_r is absent from the table, r is not treated as isolated and remains in the matching set. An isolated RingRef is removed from $\mathcal{R}$. Atom t other than the host collides with the virtual point when

$$\|\widetilde{x}_r - x_t\|_2 \leq \tfrac{1}{2} L_1(E_r, E_t)/100. \tag{45}$$

A collision means that Real atom is the closure target predicted with the wrong role and the wrong element, so the decoder sets $E_t = E_r$ and bonds the host to t with type B_r. No new atom is written. If $L_1(E_r, E_t)$ is missing, the test uses $L_1(E_t, E_r)$, then $L_1(E_t, E_t)$. Otherwise the RingRef is materialized as a Real atom at $\widetilde{x}_r$, bonded to its host $h(r)$ with bond type B_r. The host is left out of the collision test, because $\widetilde{x}_r$ is already a bond length away from it. Coordinates are not moved. The chemical stage below then relabels elements and bond orders on this connectivity. RingRefs that still have a same-element atom inside the window stay in the matching set, including those later rejected for BFS order, degree, or valence. The fast decoder does not perform this promotion. If the remaining set has no feasible assignment, decoding still fails.

Chemical scores. Structural decoding establishes atoms and connectivity, but the predicted element and bond labels may still violate allowed valences. After the fast match, or after the promotion above when `rules` is used, this chemical stage addresses those label inconsistencies while preserving the geometric construction. It fixes $(V, \mathcal{E}, X, Q)$, where $\mathcal{E} = \mathcal{E}_T \cup \mathcal{E}_R$ already includes any edge created by promoting an isolated RingRef. Let η^E, η^B be final logits. Under the single-RingRef representation used in this work, each ring edge e has one associated RingRef r_e. Define its bond-bearing slot by

$$s(e) = \begin{cases} v, & e = \{p(v), v\} \in \mathcal{E}_T, \\ r_e, & e \in \mathcal{E}_R. \end{cases} \tag{46}$$

The element and bond scores are

$$\begin{aligned} \ell_v^E(a) &= \eta^E_{\kappa(v),a} - \log \sum_{a' \in \mathcal{A}} \exp \eta^E_{\kappa(v),a'}, \\ \ell_e^B(b) &= \eta^B_{\kappa(s(e)),b} - \log \sum_{b' \in \mathcal{B}} \exp \eta^B_{\kappa(s(e)),b'}. \end{aligned} \tag{47}$$

Because occupancy and connectivity are already fixed, Null and None are excluded from these normalizations. Each physical edge contributes the score of the slot that encodes its bond, allowing feasible label assignments to be ranked by the model predictions.

Chemical feasibility. The objective can favor a high-probability assignment that violates a valence rule, so the search must restrict its feasible set explicitly. Index physical atoms by decoded order $0, \ldots, N-1$. Let $\mathcal{V}(a, q)$ be the allowed bond-order sums for element a and charge q, and define

$$\begin{aligned} \mathcal{B}_{\mathrm{pair}}(a, b) &= \{1\} \cup \{m \in \{2, 3\} : \mathcal{T}_m(a, b) = 1\}, \\ \mathcal{F}(\mathcal{E}, Q) &= \left\{ (E, B) : \begin{array}{ll} E_v \in \mathcal{A}, & \displaystyle\sum_{e \ni v} B_e \in \mathcal{V}(E_v, Q_v) \quad (\forall v), \\ & B_{\{u,v\}} \in \mathcal{B}_{\mathrm{pair}}(E_v, E_u) \quad (\forall \{u, v\} \in \mathcal{E},\ u < v) \end{array} \right\}. \end{aligned} \tag{48}$$

The first condition requires each atom's total bond order to belong to its allowed valence set; the second limits which bond orders its endpoint elements can form. Because hydrogens are explicit, unrepresented hydrogens cannot be used to fill a deficient bond-order sum. The argument order $\mathcal{B}_{\mathrm{pair}}(E_v, E_u)$ follows the implementation: the later processed atom is the first table index. With explicit hydrogens, $\mathcal{V}(\mathrm{C}, 0) = \{4\}$, $\mathcal{V}(\mathrm{N}, 0) = \{3\}$, $\mathcal{V}(\mathrm{O}, 0) = \{2\}$, and $\mathcal{V}(\mathrm{F}, 0) = \mathcal{V}(\mathrm{H}, 0) = \{1\}$. Remaining charge-dependent sets are specified in `codec/valence.py`, and $\mathcal{T}_m$ in `dataset/bond_analyze.py`; unsupported element–charge combinations have $\mathcal{V}(a, q) = \varnothing$. If Q is disabled, set $Q_v = 0$. This chemical stage optimizes (E, B) over $\mathcal{F}(\mathcal{E}, Q)$ without moving atoms or changing the connectivity fixed by structural decoding.

Search and pruning. The unrestricted label space grows as $|\mathcal{A}|^N 3^{|\mathcal{E}|}$, so we discard impossible partial assignments before enumerating their completions. When processing atom v, let $\mathcal{N}_v^- = \{u < v : \{u, v\} \in \mathcal{E}\}$ and $m_v = |\{u > v : \{u, v\} \in \mathcal{E}\}|$. After proposing E_v and bonds to $\mathcal{N}_v^-$, let ν_v be its assigned bond-order sum, ν'_u the updated sum at a predecessor, and $\mathcal{V}_v = \mathcal{V}(E_v, Q_v)$. Retain the partial assignment only if

$$\begin{aligned} &\mathcal{V}_v \neq \varnothing, \qquad \nu_v + m_v \leq \max \mathcal{V}_v, \qquad \nu_v + 3m_v \geq \min \mathcal{V}_v, \\ &m_v = 0 \Rightarrow \nu_v \in \mathcal{V}_v, \qquad \nu'_u \leq \max \mathcal{V}_u \quad (u \in \mathcal{N}_v^-). \end{aligned} \tag{49}$$

Each remaining edge contributes between one and three bond-order units, which gives the lower and upper completion bounds above. An updated predecessor must also stay below its maximum allowed

valence. These bounds are necessary but not sufficient, so complete assignments additionally require $\nu_v \in \mathcal{V}_v$ for every atom. Initialize $\mathcal{S}_{-1} = \{\varnothing\}$ with score zero. If Extend_v enumerates the retained extensions and Top_W keeps the W largest partial log-probability scores, beam search is

$$\mathcal{S}_v = \mathrm{Top}_W\left(\bigcup_{s\in\mathcal{S}_{v-1}} \mathrm{Extend}_v(s)\right), \qquad (W, K_E, K_B) = (128, 4, 2). \tag{50}$$

Extensions use the top K_E element candidates and, for each edge, the top K_B orders within its compatible set. The QM9 profile starts with $W = K_E = K_B = \infty$; when its cumulative state count exceeds $200{,}000$, search restarts with the above beam parameters. GEOM-Drugs profiles use beam search directly. Exhaustive search is exact if completed within the limit; beam search is approximate. The highest-scoring feasible completion is returned, or a chemical decoding failure if none is found.

B Additional Implementation Details

B.1 Data Preprocessing

QM9 preprocessing. We use the DeepChem OpenBabel-converted QM9 SDF and retain explicit hydrogens. The dataset split follows the MiDi procedure: the complete metadata table is shuffled with random state 42, the first $100{,}000$ entries are assigned to training, the final $\lfloor 0.1 N_{\mathrm{raw}} \rfloor$ entries to testing, and the remaining entries to validation, where N_{raw} is the number of records in the complete metadata table. After split assignment, we remove the $3{,}054$ molecules listed as uncharacterized in QM9 because they failed the dataset's geometry–connectivity consistency check, and exclude unreadable SDF records. RDKit reads the SDF with hydrogen removal and sanitization disabled at loading. Atom identities, coordinates, and bonds are taken directly from the molecular records, and aromatic bonds are Kekulized into the bond-order vocabulary $\{1, 2, 3\}$. The element vocabulary is $\{\mathrm{C}, \mathrm{N}, \mathrm{O}, \mathrm{F}, \mathrm{H}\}$, and the formal-charge channel is disabled. We construct TreeRef representations with branching factor $k = 3$ and additionally exclude 146 molecules with disconnected molecular graphs: 105 from training, 31 from validation, and 10 from testing.

The TreeRef scan accepts $97{,}629$ of $97{,}734$ training records, $20{,}011$ of $20{,}042$ validation records, and $13{,}045$ of $13{,}055$ test records. All rejected records in these three splits correspond to disconnected molecular graphs, giving 105, 31, and 10 rejected records, respectively.

GEOM-Drugs preprocessing. We use the GEOM-Drugs RDKit release and retain the lowest-energy available conformer of each molecule with explicit hydrogens. Molecules are assigned deterministically to training, validation, and test sets by hashing their source SMILES with SHA-256 and a split seed of 1, giving nominal proportions of $80/10/10$ before filtering. Each conformer must pass RDKit sanitization, Kekulization, and a single-fragment check; bonded distances must differ from the sums of the corresponding covalent radii by no more than 40%. Atom identities, connectivity, and formal charges are taken from RDKit, with bond orders in $\{1, 2, 3\}$ and charges in $\{-1, 0, +1\}$. We use a fixed element vocabulary that retains uncommon elements such as B, Si, P, I, and Bi, while excluding 58 molecules (0.019%) containing unsupported elements such as Se or Na. We construct single-RingRef representations with $k = 3$, which supports atom degrees up to four. Only nine molecules (0.0030%), all in training, contain degree-five atoms; we exclude these rather than increase k to four and enlarge the representation for the entire dataset. An additional 459 molecules (0.15%) are excluded for disconnected structures. All percentages are relative to the 304,348 scanned molecules. Overall, preprocessing removes approximately 0.17% of the data, leaving 242,992 training, 30,518 validation, and 30,312 test molecules.

The final accepted sets contain $242{,}992$ training molecules, $30{,}518$ validation molecules, and $30{,}312$ test molecules. The reported counts include all filtering steps described above.

Representation preparation. Each accepted molecule is encoded using the canonical serialization described above. Distance normalization statistics are fitted only to accepted training molecules and reused for validation and testing. The other geometric channels are not standardized. The retained packed widths are 88 for QM9 and 526 for GEOM-Drugs in the main model configurations. The QM9 cache contains up to eight stored serializations per training molecule, but the main training configuration uses only the canonical variant. Validation and test representations are canonical.

B.2 Model Details

We adopt Bayesian Flow Networks (BFNs) (Graves et al., 2023) because they provide a common framework for discrete molecular variables and continuous geometry, while continuous-time training allows sampling with any chosen positive number of steps without retraining.

Transformer architecture. We use a Transformer (Vaswani et al., 2017) with AdaLN-Zero conditioning (Peebles & Xie, 2023). The BFN time t is multiplied by 1000, encoded sinusoidally into a d-dimensional vector, and passed through a two-layer MLP with dimensions $d \to 4d \to d$ and a SiLU activation. Each block contains full multi-head self-attention and a GELU feed-forward network. The conditioning vector determines the shifts, scales, and residual gates of the attention and feed-forward branches. The modulation layers are initialized to zero, so each block initially acts as an identity mapping. A final adaptive layer normalization precedes the output heads.

We use QM9-L as the main QM9 model and additionally evaluate QM9-S and QM9-M to examine the effect of model size. These models contain 10.50M, 60.74M, and 179.04M trainable parameters, respectively. QM9-M therefore uses only 33.9% of the parameters of QM9-L, providing an intermediate configuration for evaluating the trade-off between model size and generation quality. The configurations and parameter counts are listed in Table 6. The property-conditioned QM9 models use the QM9-L backbone with an additional scalar-property embedding network and a learned null-condition vector, bringing the parameter count to 179.64M. GEOM-Drugs uses a wider XL backbone, with hidden dimension 1024, 16 blocks, 16 attention heads, and feed-forward dimension 4096, for 319.83M parameters. All configurations use dropout 0.1.

Table 6: Transformer configurations and total trainable parameter counts. M denotes one million parameters. The property-conditioned QM9 configuration includes the scalar-property embedding network and learned null-condition vector. GEOM-Drugs-XL uses $k = 3$, explicit hydrogens, and a formal-charge channel.

Model	Blocks	Hidden dimension	Attention heads	FFN dimension	Parameters (M)
QM9-S	8	256	8	1024	10.50
QM9-M	12	512	8	2048	60.74
QM9-L	16	768	12	3072	179.04
QM9-L (property-conditioned)	16	768	12	3072	179.64
GEOM-Drugs-XL	16	1024	16	4096	319.83

Input representation and embeddings. The network receives the current categorical probability vectors for R, E, B, H, and the optional Q, together with the current continuous geometry mean. Each categorical channel is independently projected to the hidden dimension through a bias-free linear map, while geometry uses a separate linear projection. These projections are summed with three learned positional embeddings:

$$e_i^{\mathrm{pos}} = e^{\mathrm{abs}}(i) + e^{\mathrm{slot}}(\mathrm{slot}(i)) + e^{\mathrm{group}}(\mathrm{group}(i)). \tag{51}$$

The absolute-position embedding identifies the packed position $i \in \{0, \dots, L' - 1\}$. The child-slot embedding identifies a position within its parent's ordered block of k children. The parent-group embedding identifies the child block itself, so all k positions belonging to the same parent share this embedding. For a non-root packed position,

$$\mathrm{slot}(i) = 1 + ((i-1) \bmod k), \qquad \mathrm{group}(i) = 1 + \left\lfloor \frac{i-1}{k} \right\rfloor, \qquad i > 0. \tag{52}$$

The root uses $\mathrm{slot}(0) = \mathrm{group}(0) = 0$. Thus, group $q + 1$ identifies the child block of the q-th Real atom in the packed BFS queue, with the root indexed by $q = 0$; it is not the packed position of that parent. For example, when $k = 3$, positions $1, 2, 3$ have child-slot identifiers $1, 2, 3$ and share group identifier 1, while positions $4, 5, 6$ share group identifier 2. These embeddings expose absolute position, sibling order, and shared-parent membership to the network. They describe the packed layout rather than logical heap addresses. All retained positions, including Null slots, participate in attention; no padding mask excludes Null states.

Output heads. Separate two-layer MLP heads produce logits for each categorical channel and a prediction of the clean geometry vector. Each head has a hidden layer of width d with a GELU

activation, and its final linear layer is initialized to zero. For $k = 3$, the geometry dimension is $1 + \binom{4}{2} + 2 = 9$. Including the categorical probability-vector dimensions, the complete per-slot input has dimension 24 for QM9 and 38 for GEOM-Drugs. The unconditional models use time-only conditioning. Property-conditioned models retain the same input channels and output heads, with the target property supplied through the additional conditioning pathway described below.

Property conditioning and classifier-free guidance. For inverse design, we train a separate conditional model for each target property using the QM9-L backbone. Let y be the scalar target value, and let μ_y and a_y be its mean and mean absolute deviation computed on the generator-training subset. We normalize the target and combine its embedding with the time embedding:

$$\widetilde{y} = \frac{y - \mu_y}{a_y}, \qquad \xi(t, y) = e_t(t) + e_y(\widetilde{y}). \tag{53}$$

Here, $e_t(t)$ is the time embedding defined above, and e_y is a two-layer MLP with dimensions $1 \to 768 \to 768$ and a SiLU activation. Its final linear layer is initialized to zero. The combined condition $\xi(t, y)$ is shared across packed slots and injected through AdaLN-Zero: each Transformer block maps it to shifts, scales, and residual gates for both the attention and feed-forward branches. It also determines the shift and scale of the final layer normalization before the output heads. Thus, the target property modulates the network throughout its depth and influences both categorical and geometric predictions.

To enable classifier-free guidance (CFG) (Ho & Salimans, 2022), we replace the property embedding with a learned null-condition vector $e_\varnothing$ with probability 0.1 per molecule during training; this vector is initialized to zero. Conditional and null-conditioned predictions therefore share the same model parameters. At sampling time, both branches receive the same Bayesian states and time, using conditions $e_t(t) + e_y(\widetilde{y})$ and $e_t(t) + e_\varnothing$, respectively. Their predictions are combined as

$$\eta^c_{i,\mathrm{cfg}} = \eta^c_{i,\varnothing} + w\big(\eta^c_{i,y} - \eta^c_{i,\varnothing}\big), \qquad \widehat{G}_{i,\mathrm{cfg}} = \widehat{G}_{i,\varnothing} + w\big(\widehat{G}_{i,y} - \widehat{G}_{i,\varnothing}\big). \tag{54}$$

Here, η^c_i denotes the logits of categorical channel c, and w is the guidance scale. Softmax is applied after combining the categorical logits, while the guided clean geometry prediction enters the Gaussian Bayesian update. We use $w = 2.0$ for property-conditioned evaluation, applying the same scale to all categorical channels and geometry at every sampling time and at the final readout. Under this convention, $w = 1$ recovers the ordinary conditional prediction.

B.3 Training Details

Notation and forward Bayesian flows. We use the same notation as the main text: x^c_i is the clean class of channel $c \in \mathcal{C}$ at packed slot i, $e(x^c_i)$ is its one-hot vector, and $\widehat{p}^c_{i,t}$ is the predicted categorical distribution. Here, $\mathcal{C} = \{R, E, B, H\}$, with Q included when enabled, and K_c is the corresponding class count. The symbol $z^c_{i,t}$ denotes a categorical log-state, rather than a clean class. For each molecule, we sample one time $t \sim \mathrm{Uniform}(0, 1)$ shared by all slots and channels. With $\beta_c(t) = \beta_1 t^2$ and $\beta_1 = 1$, the categorical states are

$$z^c_{i,t} = \beta_c(t)\big(K_c e(x^c_i) - \mathbf{1}\big) + \sqrt{K_c \beta_c(t)}\,\epsilon^c_i, \qquad s^c_{i,t} = \mathrm{softmax}(z^c_{i,t}), \qquad \epsilon^c_i \sim \mathcal{N}(0, I). \tag{55}$$

For the standardized clean geometry $G_i \in \mathbb{R}^{d_G}$, let $m = 10^{-3}$ and $\gamma(t) = 1 - m^t$, as in the main text. The Gaussian mean state is

$$\mu^G_{i,t} = \gamma(t) G_i + \sqrt{\gamma(t)(1 - \gamma(t))}\,\epsilon^G_i, \qquad \epsilon^G_i \sim \mathcal{N}(0, I). \tag{56}$$

Noise is independent across slots and channels. At $t = 0$, the categorical states are uniform and the geometry means are zero. Sampling these states directly at a randomly selected time avoids simulating a complete sampling trajectory during training. The network jointly processes the states and time to predict $\widehat{p}^c_{i,t}$ and the clean geometry $\widehat{G}_{i,t}$.

Training objectives. The categorical and geometry losses use the same weights and normalization as the main-text objective:

$$\ell_c(t) = \frac{K_c \beta_1 t}{L'} \sum_{i=0}^{L'-1} \big\| e(x^c_i) - \widehat{p}^c_{i,t} \big\|_2^2, \qquad \ell_G(t) = \frac{-\log m}{2 m^t L' d_G} \sum_{i=0}^{L'-1} \Big\| G_i - \widehat{G}_{i,t} \Big\|_2^2. \tag{57}$$

The geometry weight equals one half of the derivative of the precision $\rho(t) = m^{-t}$, so errors at more informative states receive greater weight. Categorical errors are summed over classes and averaged over slots; geometry errors are averaged over both slots and dimensions. The full objective is

$$\mathcal{L}_{\mathrm{BFN}} = \mathbb{E}_{Z,t,\epsilon}\left[\frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} \ell_c(t) + \ell_G(t)\right]. \tag{58}$$

Every packed slot contributes. Null slots therefore remain supervised occupancy targets with their prescribed categorical labels and zero clean geometry. Their noisy geometry states are generally nonzero at $t > 0$; a zero target does not remove the forward-flow noise or its loss contribution.

Tree-view augmentation and consistency. Multiple cached BFS serializations can represent the same molecule with different parent relations, RingRef placements, and local geometric references. In experiments using multiple views per molecule, the views share a sampled time and receive independent forward-flow noise. Let V_b be the number of sampled views for molecule b, and let $h_i^{\mathrm{out},(a)}$ be the final normalized Transformer hidden state of view a. We pool all packed positions, including Null slots, to obtain

$$\overline{h}^{(a)} = \frac{1}{L'}\sum_{i=0}^{L'-1} h_i^{\mathrm{out},(a)}. \tag{59}$$

Let $\mathcal{P}_b$ contain the unordered pairs of sampled views with different serialization identifiers. Repeated copies of the same cached variant are excluded. We encourage agreement between these molecule-level summaries through

$$\mathcal{L}_{\mathrm{inv}} = \mathbb{E}_b\left[\frac{1-t_b}{\max(1,|\mathcal{P}_b|)}\sum_{(a,a')\in\mathcal{P}_b}\left(1-\cos\left(\overline{h}^{(a)},\overline{h}^{(a')}\right)\right)\right], \qquad \mathcal{L} = \mathcal{L}_{\mathrm{BFN}} + \lambda_{\mathrm{inv}}\mathcal{L}_{\mathrm{inv}}. \tag{60}$$

This is Equation 8. The BFN term is averaged over views, and $\lambda_{\mathrm{inv}} = 1$ when consistency training is enabled. Cosine similarity is computed after dividing each pooled vector by $\max(\|\overline{h}\|_2, 10^{-8})$. The factor $1 - t_b$ decreases the consistency weight toward the clean-data endpoint. A molecule with no distinct-view pair contributes zero consistency loss. Canonical-only runs use one view and omit this term; storing several variants in the cache alone does not enable consistency training.

Property-conditioned training. The accepted QM9 training molecules are partitioned into two disjoint subsets,

$$\mathcal{D}_{\mathrm{train}} = \mathcal{D}_{\mathrm{gen}} \,\dot\cup\, \mathcal{D}_{\mathrm{reg}},$$

with an approximately equal split, random seed 1, and all serialization variants of one molecule assigned to the same subset. The conditional generator is trained on $\mathcal{D}_{\mathrm{gen}}$. For each target property, we independently train an EGNN regressor and a TreeRef-based regressor on $\mathcal{D}_{\mathrm{reg}}$. Both regressors are frozen during evaluation. Their mean absolute errors with respect to the requested target value are reported separately. Property-conditioned generation uses the same 10,000-sample protocol as the other experiments and applies CFG with $w = 2.0$.

Optimization. We use AdamW with moment coefficients $(0.9, 0.999)$, numerical constant 10^{-8}, zero weight decay, and global gradient-norm clipping at 1. Global batch sizes are given in Appendix B.4. From a fresh initialization, the learning rate increases linearly from 10^{-6} to 10^{-4} over the first 10 epochs and then follows cosine decay toward 10^{-6}, with the scheduler advanced after each optimization step. The configured training horizons are 3,000 epochs for QM9 and 600 epochs for GEOM-Drugs. For an extended run, the optimizer state is restored and cosine decay continues from the resumed learning rate over the remaining steps, without restarting warmup. We maintain an exponential moving average (EMA) of model parameters, updated after every optimization step with decay 0.9999. Validation evaluates the same BFN loss with EMA parameters and a fixed random seed for the sampled times and noise; property-condition dropout is disabled during validation. The checkpoint with the lowest validation loss is saved alongside the latest checkpoint. Generation-metric checkpoints, when used for evaluation, are identified separately from the validation-loss checkpoint.

B.4 HARDWARE

The QM9 models were trained together on NVIDIA RTX A6000 GPUs. QM9-S uses a global batch size of 64. GEOM-Drugs-XL was trained on three NVIDIA H100 GPUs with a global batch size of 16.

B.5 UNCONDITIONAL SAMPLING

Initialization and schedules. Let S denote the number of Bayesian updates, reserving N for molecular size. Sampling begins with uniform categorical probabilities and a Gaussian geometry prior with zero mean and unit precision:

$$s^c_{i,0} = \frac{\mathbf{1}}{K_c}, \qquad \mu^G_{i,0} = \mathbf{0}, \qquad \rho_0 = 1. \tag{61}$$

The prior is represented by its parameters; we do not initialize $\mu^G_{i,0}$ with a sampled Gaussian vector. We use the uniform grid $t_j = j/S$, $j = 0, \dots, S$, and the same schedules $\beta_c(t) = \beta_1 t^2$ and $\rho(t) = m^{-t}$ as in training. Their increments are

$$\alpha^c_j = \beta_c(t_j) - \beta_c(t_{j-1}) = \beta_1 \frac{2j-1}{S^2}, \qquad \alpha^G_j = m^{-j/S} - m^{-(j-1)/S}. \tag{62}$$

Changing S changes the inference discretization without changing the trained model. Unless specified otherwise, the main evaluation protocol uses $S = 100$.

Categorical Bayesian updates. At step j, the network receives the current states at t_{j-1} and returns $\widehat{p}^c_{i,j-1}$ and $\widehat{G}_{i,j-1}$. A sampled class determines the mean of a noisy sender message:

$$\widetilde{x}^c_{i,j} \sim \mathrm{Cat}(\widehat{p}^c_{i,j-1}), \qquad y^c_{i,j} \sim \mathcal{N}\big(\alpha^c_j\big(K_c e(\widetilde{x}^c_{i,j}) - \mathbf{1}\big), K_c \alpha^c_j I\big)\,. \tag{63}$$

The message updates the current categorical belief through

$$s^c_{i,j} = \mathrm{softmax}\big(\log s^c_{i,j-1} + y^c_{i,j}\big)\,. \tag{64}$$

This is equivalent to the additive log-state update in the main text. The implementation stores normalized log probabilities and subtracts their log-sum-exp after each update for numerical stability; adding or subtracting a common scalar does not change the resulting probabilities.

Geometry Bayesian updates. The predicted clean geometry supplies the center of a Gaussian sender observation, which is combined with the current belief according to their precisions:

$$y^G_{i,j} \sim \mathcal{N}\Big(\widehat{G}_{i,j-1}, (\alpha^G_j)^{-1} I\Big)\,, \qquad \rho_j = \rho_{j-1} + \alpha^G_j = m^{-t_j}, \qquad \mu^G_{i,j} = \frac{\rho_{j-1}\mu^G_{i,j-1} + \alpha^G_j y^G_{i,j}}{\rho_j}. \tag{65}$$

All slots and channels use the same time grid, with independent sender noise. Slots that ultimately become Null are updated throughout sampling because their final occupancy is not known beforehand.

Final readout and decoding. After the S Bayesian updates, the network is evaluated once more at $t_S = 1$. Thus, unconditional sampling uses $S + 1$ network evaluations. The default readout samples each categorical class at temperature 1 and takes the clean geometry prediction:

$$x^{c,\mathrm{raw}}_i \sim \mathrm{Cat}(\widehat{p}^c_{i,S}), \qquad G^{\mathrm{raw}}_i = \widehat{G}_{i,S}. \tag{66}$$

The final geometry is therefore the predicted clean vector, not the accumulated Gaussian mean $\mu^G_{i,S}$. Occupancy projection enforces a non-Null Real root, removes unreachable slots, and supplies a non-None parent bond to each occupied non-root slot. Null geometry is set to zero and Null handedness to inactive. The projected representation is passed to the topology, geometry, and RingRef reconstruction procedures described above. The main protocol uses the fast decoder and leaves the reconstructed Real coordinates fixed during RingRef closure; chemically constrained decoding is reported separately.

B.6 CONDITIONAL SAMPLING

Masks and initial beliefs. Mask-based structural generation reuses an unconditional checkpoint without parameter updates. We distinguish two conditioning schemes: *frozen*, which supplies clean observations to the network, and *RePaint-style*, which supplies freshly sampled forward-flow states of the observations at each sampling time. Both use the same task masks and Bayesian updates for unobserved variables.

We retain the main-text convention that $M_i^c = 1$ denotes a categorical variable to generate. For geometry, the sampler uses one binary decision g_i for the whole vector, corresponding to $M_i^G = g_i \mathbf{1}_{d_G}$. Boundary-geometry resampling therefore releases the entire G_i. Given observations $x_{i,\mathrm{obs}}^c$ and $G_{i,\mathrm{obs}}$, the stored states are initialized as

$$\begin{aligned} s_{i,0}^c &= M_i^c \frac{\mathbf{1}}{K_c} + (1 - M_i^c)\overline{e}(x_{i,\mathrm{obs}}^c), \\ \mu_{i,0}^G &= (1 - g_i)G_{i,\mathrm{obs}}, \qquad \rho_{i,0} = 1. \end{aligned} \tag{67}$$

Here, $\overline{e}(x)$ lower-bounds the entries of the one-hot vector $e(x)$ by 10^{-30} and normalizes, keeping categorical log states finite. Unknown categorical variables start uniformly, while unknown geometry starts with zero mean and unit precision. Clearing source entries to Null during fragment construction is only preprocessing: regenerated variables are subsequently reset to these priors.

Frozen and RePaint-style conditioning. Before each Transformer evaluation at t_j, the network inputs are

$$\begin{aligned} \bar{s}_{i,j}^c &= M_i^c s_{i,j}^c + (1 - M_i^c)\widetilde{s}_{i,j}^c, \\ \bar{\mu}_{i,j}^G &= g_i \mu_{i,j}^G + (1 - g_i)\widetilde{\mu}_{i,j}^G. \end{aligned} \tag{68}$$

Frozen conditioning uses $\widetilde{s}_{i,j}^c = \overline{e}(x_{i,\mathrm{obs}}^c)$ and $\widetilde{\mu}_{i,j}^G = G_{i,\mathrm{obs}}$. RePaint-style conditioning instead draws

$$\begin{aligned} \widetilde{s}_{i,j}^c &= \mathrm{softmax}\Big(\beta(t_j)\big[K_c e(x_{i,\mathrm{obs}}^c) - \mathbf{1}\big] + \sqrt{K_c \beta(t_j)}\,\epsilon_{i,j}^c\Big), \\ \widetilde{\mu}_{i,j}^G &= \big[1 - v(t_j)\big]G_{i,\mathrm{obs}} + \sqrt{v(t_j)\big[1 - v(t_j)\big]}\,\epsilon_{i,j}^G, \end{aligned} \tag{69}$$

where $\beta(t) = \beta_1 t^2$, $v(t)$ is the geometry variance schedule, and the noise vectors are independently sampled standard Gaussians at every evaluation. For the schedule $v(t) = m^t$, these are the corresponding categorical and Gaussian BFN forward distributions. RePaint-style conditioning replaces only the observed network inputs; it does not overwrite the running states of unobserved variables. The same rule applies at the final evaluation at $t = 1$.

Masked updates and readout. The network receives the entire packed sequence under either conditioning scheme. Let $s_{i,j+1}^{c,*}$ and $\mu_{i,j+1}^{G,*}$ denote candidate Bayesian updates obtained from the stored states and the predictions at t_j. Only unobserved entries are updated:

$$\begin{aligned} s_{i,j+1}^c &= M_i^c s_{i,j+1}^{c,*} + (1 - M_i^c)\overline{e}(x_{i,\mathrm{obs}}^c), \\ \mu_{i,j+1}^G &= g_i \mu_{i,j+1}^{G,*} + (1 - g_i)G_{i,\mathrm{obs}}, \\ \rho_{i,j+1} &= \rho_{i,j} + g_i \alpha_{j+1}^G. \end{aligned} \tag{70}$$

Thus, regenerated geometry follows the original precision schedule, while observed geometry retains bookkeeping precision 1, including under RePaint-style conditioning. At structural readout, supplied labels and selected geometric observations are restored, subject to the root and Null conventions. Both schemes start at $t = 0$ unless specified otherwise and follow a single forward traversal of the sampling grid, without time jumps or repeated backward–forward cycles.

Retained slots for structural editing. For scaffold decoration, multi-fragment completion, scaffold replacement, and linker completion, let $A \subseteq V$ be the retained physical atom set and $\kappa(a)$ its Real slot in the source serialization. With $h(r)$ and $t(r)$ denoting a source RingRef's host and target, retain

$$\begin{aligned} \mathcal{F}(A) = \{&\kappa(a) : a \in A\} \\ &\cup \{\kappa(r) : r \in \mathcal{R},\ h(r) \in A,\ t(r) \in A\}, \qquad b_i = \mathbf{1}[i \notin \mathcal{F}(A)]. \end{aligned} \tag{71}$$

Retained observations remain at their original packed positions. Slots outside $\mathcal{F}(A)$, including source Null positions, are regenerated. For the structural categorical channels $\mathcal{C}_{\text{str}} = \mathcal{C} \setminus \{H\}$, set $M_i^c = b_i$. These masks specify which variables are observed; frozen or RePaint-style conditioning determines how those observations are presented to the network.

Geometry at attachment sites. A retained atom can lose a neighbor when the source molecule is cut. Its original local descriptors may reveal the removed region or constrain the replacement geometry. Under the junction policy, define

$$J(A) = \{a \in A : \exists u \in V \setminus A,\ \{a, u\} \in \mathcal{E}\}, \qquad \mathcal{J} = \mathcal{F}(J(A)). \tag{72}$$

The geometry and handedness masks are

$$g_i = M_i^H = b_i \vee \mathbf{1}[i \in \mathcal{J}], \qquad M_i^G = g_i \mathbf{1}_{d_G}. \tag{73}$$

Structural labels remain observed at retained slots. At attachment slots, the complete geometry vector and handedness are reset to their priors and regenerated. Other retained slots use their source geometry and handedness as observations, supplied cleanly under frozen conditioning or through forward-flow samples under RePaint-style conditioning. The set $\mathcal{J}$ includes Real slots of attachment atoms and RingRefs whose host and target both belong to $J(A)$. This local rule does not perform a component-wise dependency analysis of all geometric references.

Generation conditioned on serialized structure. For the standard 2D-to-3D task, R, E, B and optional Q are observed, while G, H are regenerated at every position: $M_i^c = 0$ for $c \in \mathcal{C}_{\text{str}}$ and $g_i = M_i^H = 1$. Frozen conditioning supplies clean structural states; RePaint-style conditioning supplies their time-dependent forward-flow samples. Both restore the structural labels at final readout and apply zero geometry and inactive handedness to Null slots. Independent sampling noise yields different geometries from the same serialized condition. RingRef targets are inferred from geometry during decoding, so observing these categorical channels alone does not explicitly fix the final ring-edge assignments.

Fragment growth. Let $a_1, \ldots, a_N$ be the source Real atoms in packed BFS order, with a_1 the root, and retain $A_{\text{grow}} = \{a_1, \ldots, a_{n_{\text{keep}}}\}$. The retained size counts explicit atoms, including hydrogens. This prefix is connected through its construction-tree edges. The growth sampler also retains occupied RingRef children hosted by retained Real atoms:

$$\begin{aligned} \mathcal{F}_{\text{grow}} = &\{\kappa(a) : a \in A_{\text{grow}}\} \\ &\cup \{\kappa(r) : r \in \mathcal{R},\ h(r) \in A_{\text{grow}}\}. \end{aligned} \tag{74}$$

All supplied channels on $\mathcal{F}_{\text{grow}}$ are observed, and the complement is regenerated from the prior. Unlike $\mathcal{F}(A)$ for structural editing, this host-based rule does not require the source RingRef target to be retained; target assignments are resolved during decoding. Prefix growth observes the source geometry and handedness on retained slots, whereas the junction policy applies to the four editing tasks. Either frozen or RePaint-style conditioning can be used with this growth mask; fragment growth is a task, distinct from the choice of conditioning scheme.

Scaffold decoration and replacement. We extract the RDKit Murcko scaffold and include hydrogens directly attached to its atoms, obtaining A_{sc}. Let $D_1, \ldots, D_J$ be the connected components of the induced graph on $V \setminus A_{\text{sc}}$. Decoration retains $A = A_{\text{sc}}$ and regenerates its complement. Scaffold replacement retains $A = \bigcup_j D_j$ and regenerates the scaffold region. Both require a nonempty scaffold and at least one side-chain component. The replacement mask permits, but does not require, a changed scaffold.

Eligibility of two retained fragments. Multi-fragment and linker completion retain two source subgraphs A_1, A_2. Each must be connected and contain at least three heavy atoms. Their atom sets are disjoint, their union induces exactly two connected components, and the source shortest path between them includes an atom outside the retained union. These conditions ensure a nonempty removed region between the fragments. All atoms in $V \setminus (A_1 \cup A_2)$ and unused packed slots are regenerated. Sources without an eligible pair are excluded.

Multi-fragment completion. We first remove A_{sc} and search the remaining components for two eligible fragments, ordering candidates by decreasing heavy-atom count and then explicit atom count. If this fails, we enumerate cuts containing one heavy atom or two bonded heavy atoms, together with their attached hydrogens. Among valid pairs of connected remnants, we maximize

$$(\min\{|A_1|_{\text{heavy}}, |A_2|_{\text{heavy}}\}, |A_1|_{\text{heavy}} + |A_2|_{\text{heavy}}, -|V \setminus (A_1 \cup A_2)|_{\text{heavy}}) \tag{75}$$

lexicographically. Here, $|A|_{\text{heavy}}$ counts non-hydrogen atoms. The retained set is $A = A_1 \cup A_2$; neither fragment is required to contain a ring.

Linker completion. Linker completion requires two ring-containing retained components and does not fall back to generic multi-fragment selection. We enumerate RDKit ring atom sets with their attached hydrogens in decreasing explicit atom count and inspect disjoint pairs. For each pair, initialize a candidate linker with the internal atoms of a shortest path between the rings. Expand it by adding attached hydrogens and non-anchor degree-two atoms whose neighbors all lie in the linker or anchor sets, until no further atom is added. After deleting this set, retain the two distinct connected components containing the respective rings, provided they satisfy the eligibility conditions. We use the first eligible pair. The retained anchors therefore include surviving substituents, and the regenerated complement has no prescribed size.

Structural readout. Frozen and RePaint-style conditioning use the same root-based structural readout. After occupancy projection, observed structural labels and the selected observed geometry and handedness are restored. Reachability from the root is then recomputed, and unreachable regenerated slots are cleared before decoding. Structural editing retains source descriptors at their original packed positions without re-encoding the selected fragments. Because generated Real occupancy can change child-block ownership, preserving these entries does not by itself guarantee preservation of the retained fragment graph or its Cartesian geometry. Fragment retention, connectivity, and geometric consistency are therefore evaluated on the decoded molecules.

Forest traversal is an optional alternative: successive packed trees are decoded separately and merged, without adding edges or matching RingRefs across trees. Under either traversal, generated Real occupancy can change child-block ownership. Preserving packed entries therefore does not guarantee preservation of the source fragment graph or a common Cartesian placement of retained fragments. Fragment retention, connectivity, and geometry are evaluated on the decoded output.

Property-conditioned sampling. Inverse design uses the separately trained conditional models and property injection described in Appendix C.3. All molecular variables start from their unconditional priors. At each sampling time, including final readout, CFG predictions with $w = 2.0$ replace the ordinary categorical logits and clean geometry predictions in the same Bayesian update and readout rules. The target property remains fixed throughout sampling. This property conditioning is separate from the frozen and RePaint-style treatment of partial structural observations.

B.7 Evaluation Details

Chemical validity, uniqueness, and novelty. RDKit validity (RDKit V) requires successful TreeRef decoding, a single connected molecular graph, and successful RDKit sanitization, with decoding failures counted as invalid. Uniqueness is the fraction of distinct canonical SMILES among RDKit-valid molecules. Novelty is computed after deduplication as the fraction of generated reference-compatible SMILES that are absent from the training reference set. Generated SMILES are obtained by sanitizing the largest molecular fragment and applying RDKit's default SMILES conversion. The QM9 reference set is constructed from the training SDF molecules with `isomericSmiles=False`, whereas the GEOM-Drugs reference set uses the source SMILES of accepted training molecules.

Atom and molecule stability. Stability follows the implemented MiDi allowed-valence rules using the decoded elements, formal charges, and sums of incident bond orders. Explicit hydrogens participate in the calculation. Atom stability is the fraction of stable atoms among all decoded atoms. A molecule is stable only if every atom satisfies the rule. We distinguish molecule stability conditional on successful decoding from overall molecule stability, for which decoding failures count as unstable. These denominators are reported explicitly. The stricter exact-valence diagnostic implemented in the repository is recorded separately and is not substituted for the MiDi stability metric.

Geometric accuracy and ring closure. Geometric metrics are measured from reconstructed Cartesian coordinates rather than directly from predicted TreeRef channels. Reference distributions are obtained from the validation set. Bond lengths are grouped by endpoint elements and bond order, bond angles by the element triplet, and dihedrals by the element quadruplet and central bond order, identifying reversed descriptions of the same category. We compute a Wasserstein-1 distance within each category represented by at least 20 observations in both generated and reference data,

and aggregate category distances using reference observation counts as weights. Bond-length distances are reported in Å, and angular distances in degrees. For dihedrals, both distributions are wrapped to a common interval centered at the reference circular mean before applying the one-dimensional Wasserstein calculation. Successfully decoded structures contribute even when they subsequently fail chemical validity. Ring closure is assessed by the Euclidean distance between each virtual RingRef coordinate and its matched Real target; with coordinate fusion disabled, these discrepancies are retained in the output geometry.

Conditional generation metrics. For 2D-to-3D generation, we report equality of the fixed $R/E/B/Q$ channels, atom-count agreement, and decoded SMILES agreement with the source. Cartesian agreement is measured by Kabsch-aligned RMSD when the generated and reference atom counts agree, using their decoded atom ordering without an additional atom-permutation search. For fragment growth and editing, retained-slot agreement checks the role and element values at the supplied positions. Growth records whether the decoded output contains more atoms than the decoded condition. We additionally assess retention of the supplied element multiset, substructure matches to the retained fragments, and whether the output SMILES reproduces the source molecule. The substructure score is the fraction of supplied fragment queries found in each generated molecule, averaged over evaluable conditions; hydrogens are removed from the fragment queries before matching. Retained-slot agreement, chemical validity, and decoded fragment preservation are reported separately.

Evaluation protocol. Each experiment specifies its checkpoint, sampling seed, sampling steps, number of source conditions, and number of generated samples or conformers per condition. Fragment experiments report the eligible-condition count and exclusions caused by the fragment-selection rules. Comparisons across conditions use the same source subset when required by the experimental design. Main evaluations use EMA parameters, the `fast` decoder; constrained chemical decoding is reported only in its designated ablation. Unless otherwise stated, every experiment uses $10{,}000$ generated samples, EMA parameters, 100 Bayesian update steps followed by a final prediction at $t = 1$, and the fast decoder.

C Additional Experimental Results

C.1 Model Size and Sampling Steps on QM9

Evaluation protocol. We evaluate the S, M, and L variants of TreeRef-BFN on unconditional QM9 generation with explicit hydrogens, using 10, 20, 50, 100, 200, 500, 1,000, and 2,000 Bayesian update steps. Each S, M, and L configuration contains 10,000 generated samples. All configurations use the fast decoder without RingRef coordinate fusion. RDKit validity and molecule stability use all generated samples as the denominator, including decoding failures. Atom and molecule stability follow the MiDi allowed-valence rules; atom stability is measured over decoded atoms. Uniqueness is computed among RDKit-valid molecules, and novelty is the fraction of distinct generated MiDi-compatible SMILES absent from the training reference set. Geometric quality is measured by category-weighted Wasserstein-1 distances between decoded and validation-reference bond-length, bond-angle, and dihedral distributions. These distances measure distributional agreement rather than reconstruction error against individual reference conformers. Each configuration is evaluated in one sampling run; variation across independent runs is not estimated.

Effect of model size. Table 7 and Figure 2 show that larger models consistently improve RDKit validity and molecule stability across the sampling budgets considered. At 100 steps, validity increases from 85.98% for S to 95.15% for M and 96.56% for L, while molecule stability increases from 84.60% to 94.52% and 96.19%, respectively. Larger models also reduce bond-length distribution errors. This ordering is not uniform across all geometric metrics: at 100 steps, M has a lower bond-angle distance than L, whereas L has a lower dihedral distance. Uniqueness remains above 96% for all three models at this budget, although novelty decreases with model size.

Invariance-trained M. M-inv2 is the same M model with the consistency term in Equation 8 turned on, so two tree serializations of one molecule are asked to share a pooled summary. Figure 2 and Table 7 include this run at every budget in the sweep. RDKit validity and molecule stability stay next to M and short of L. At 100 steps they are 94.70% and 94.00%, against 95.15% and 94.52% for M. At 2,000 steps they are 96.10% and 96.10%, against 96.04% and 95.57% for M and 97.53%

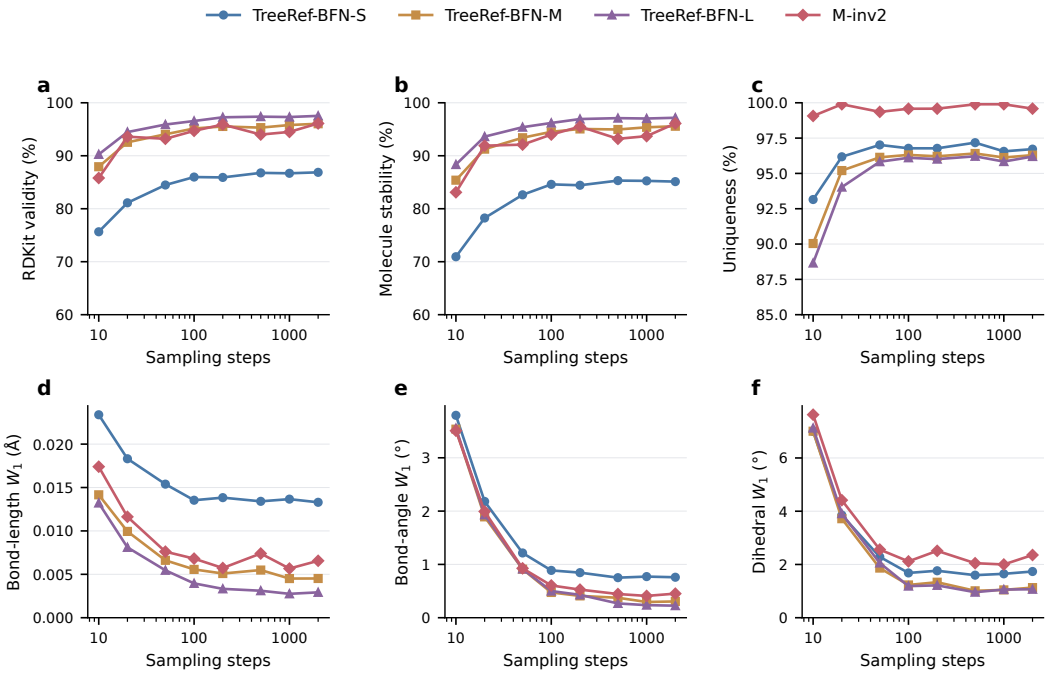


Figure 2: Unconditional QM9 generation across model sizes and sampling budgets. Panels show (a) RDKit validity, (b) molecule stability, (c) uniqueness, and Wasserstein-1 distances for (d) bond lengths, (e) bond angles, and (f) dihedrals. Higher values are better in (a–c), and lower values are better in (d–f). S, M, and L markers summarize 10,000 samples. The horizontal axis is logarithmic, and lines connect evaluated configurations. No variability across independent runs is reported.

and 97.17% for L. The curve is not monotone: validity is 94.00% at 500 steps. Bond, angle, and dihedral distances remain larger than M at every budget. At 100 steps they are 0.0068 Å, 0.605°, and 2.12°, against 0.0056 Å, 0.474°, and 1.23° for M. Novelty at 100 steps is 47.9%, against 43.2% for M. The extra term does not buy the jump from M to L when the model draws a whole molecule from the prior. It keeps an M-sized model in the same range, which is the setting where the fragment edits later pick up the gain. One reading is that unconditional sampling never shows the model a fragment whose traversal it could mistake for the molecule. The sample is one tree drawn from the prior, so a summary that has been made insensitive to the traversal has nothing extra to condition on, and validity stays with M. The same term still pools away the local frames that carry bond lengths and angles. That is a plausible reason the geometric distances stay larger than M at every budget, while the width of the network remains that of M and so stays short of L.

Effect of sampling steps. Increasing the sampling budget yields the largest quality improvements between 10 and 100 steps, followed by smaller and sometimes non-monotonic changes. For L, increasing the budget from 100 to 2,000 steps improves validity from 96.56% to 97.53% and molecule stability from 96.19% to 97.17%. Bond-length and bond-angle distribution errors continue to decrease overall, while dihedral agreement does not improve monotonically. Thus, 100 steps provides a useful quality–step-budget trade-off in this sweep, whereas larger budgets offer additional gains in selected metrics. This comparison concerns sampling steps and does not establish a wall-clock speed advantage over other architectures.

C.2 Ablations on QM9

Table 1 includes two QM9-L models retrained at the same size and evaluated at 100 steps with 10,000 samples. Each removes one piece of the representation used by TreeRef-BFN.

Single-angle geometry. The full local vector G_i stores a pairwise cosine for every pair among the parent direction and the child directions, together with one dihedral. This variant keeps the

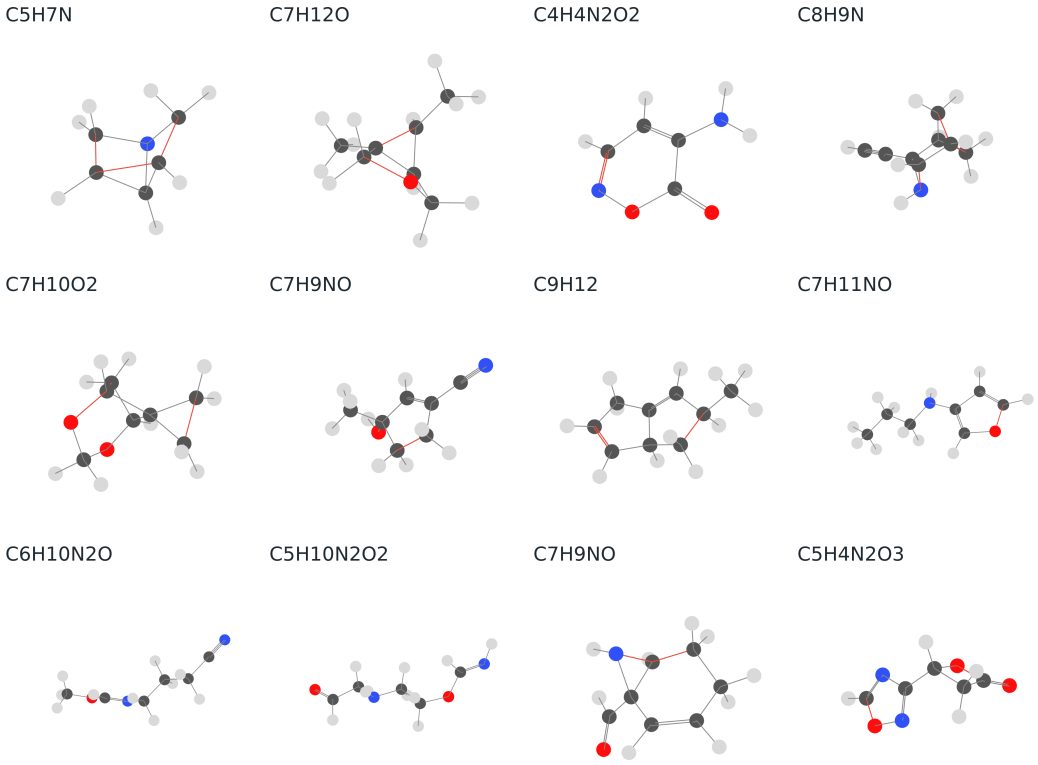


Figure 3: Unconditional QM9 molecules at 100 steps, drawn at random from valid samples. Red edges indicate RingRef closures.

Table 7: Unconditional QM9 results for TreeRef-BFN-S, M, L, and M-inv2. Valid denotes RDKit validity. MiDi stability uses the generated bond graph and MiDi allowed-valence rules; EDM stability uses bonds inferred from Cartesian distances and EDM valence rules. Atom and molecule stability are reported separately, with decoding failures counted as unstable molecules. Validity, stability, uniqueness (Unique), and novelty (Novel) are percentages. Bond-length W_1 is in Å; bond-angle and dihedral W_1 are in degrees.

Model	Steps	Valid ↑	MiDi stability ↑		EDM stability ↑		Unique ↑	Novel	W_1 ↓		
			Atom	Mol.	Atom	Mol.			Bond	Angle	Dih.
S	10	75.65	96.63	70.94	93.25	58.50	93.15	67.82	0.0234	3.797	6.998
S	20	81.12	97.71	78.25	94.62	64.33	96.18	58.79	0.0183	2.181	3.853
S	50	84.48	98.27	82.63	95.30	67.69	97.02	54.80	0.0154	1.215	2.286
S	100	85.98	98.48	84.60	95.59	69.25	96.78	53.19	0.0135	0.888	1.682
S	200	85.91	98.47	84.43	95.52	68.69	96.78	50.64	0.0138	0.846	1.764
S	500	86.77	98.58	85.31	95.58	69.07	97.18	50.95	0.0134	0.751	1.598
S	1000	86.69	98.57	85.26	95.59	69.26	96.56	50.40	0.0137	0.771	1.649
S	2000	86.87	98.56	85.12	95.53	68.92	96.72	51.83	0.0133	0.759	1.736
M	10	87.93	98.43	85.39	96.06	74.94	90.04	59.11	0.0142	3.537	7.005
M	20	92.54	99.14	91.25	97.34	81.49	95.20	50.47	0.0099	1.892	3.719
M	50	94.05	99.42	93.40	97.86	84.50	96.14	45.79	0.0066	0.910	1.865
M	100	95.15	99.52	94.52	97.98	85.25	96.31	43.18	0.0056	0.474	1.230
M	200	95.53	99.56	95.06	98.02	85.68	96.21	41.80	0.0051	0.409	1.333
M	500	95.31	99.55	94.96	98.03	85.63	96.41	41.35	0.0055	0.374	1.011
M	1000	95.80	99.58	95.37	98.18	86.69	96.13	41.07	0.0045	0.295	1.045
M	2000	96.04	99.61	95.57	98.16	86.61	96.29	40.62	0.0045	0.305	1.131
M-inv2	10	85.80	97.92	83.10	95.17	72.80	99.07	63.18	0.0174	3.508	7.626
M-inv2	20	93.60	99.29	91.90	97.32	80.60	99.89	54.01	0.0116	1.993	4.415
M-inv2	50	93.20	99.22	92.10	97.31	81.90	99.36	52.16	0.0076	0.926	2.560
M-inv2	100	94.70	99.45	94.00	97.59	83.20	99.58	47.93	0.0068	0.605	2.116
M-inv2	200	95.90	99.60	95.50	97.68	83.40	99.58	42.41	0.0057	0.527	2.512
M-inv2	500	94.00	99.43	93.20	97.80	84.30	99.89	46.86	0.0074	0.445	2.048
M-inv2	1000	94.50	99.40	93.70	97.86	84.90	99.89	43.11	0.0057	0.409	2.000
M-inv2	2000	96.10	99.55	96.10	97.54	83.10	99.58	43.36	0.0066	0.452	2.355
L	10	90.26	98.74	88.34	96.60	78.89	88.65	55.39	0.0132	3.555	7.124
L	20	94.47	99.39	93.61	97.89	85.55	94.03	46.52	0.0081	1.930	3.909
L	50	95.89	99.59	95.39	98.25	87.63	95.83	42.19	0.0055	0.915	2.056
L	100	96.56	99.68	96.19	98.46	89.08	96.11	40.00	0.0040	0.500	1.187
L	200	97.26	99.73	96.93	98.48	89.20	96.01	38.08	0.0033	0.429	1.215
L	500	97.38	99.75	97.10	98.56	89.49	96.21	38.05	0.0031	0.268	0.958
L	1000	97.31	99.74	97.03	98.62	90.05	95.84	37.57	0.0028	0.236	1.058
L	2000	97.53	99.76	97.17	98.55	89.57	96.20	37.10	0.0029	0.226	1.068

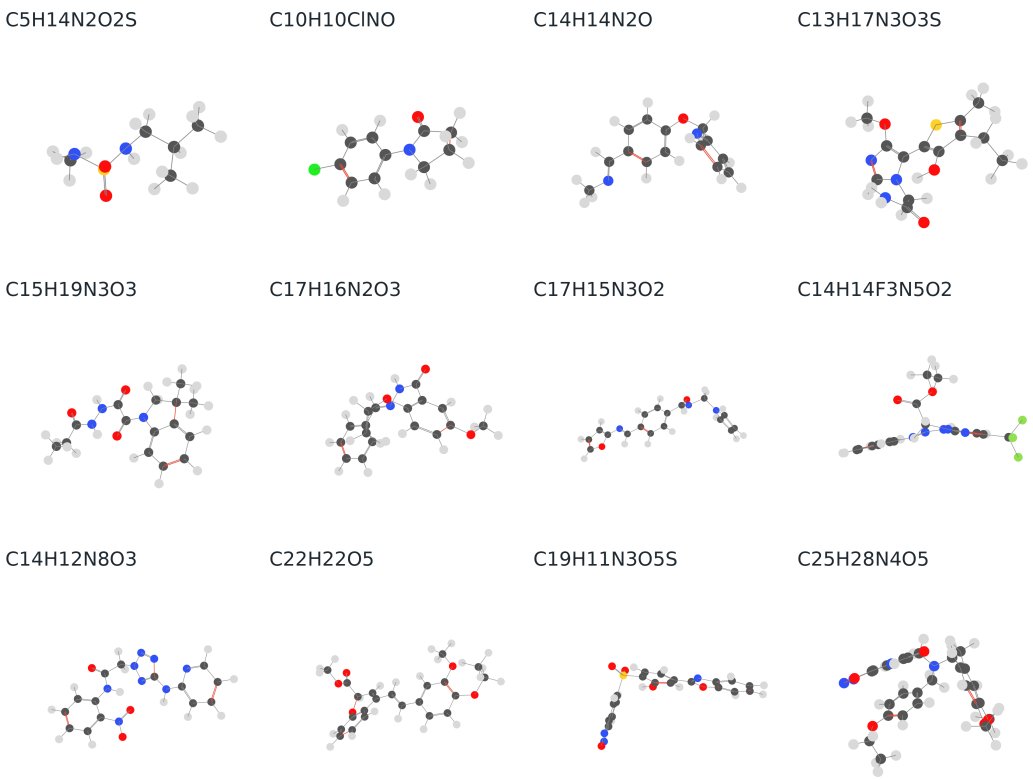


Figure 4: Unconditional GEOM-Drugs samples from TreeRef-BFN at 100 steps. Twelve molecules, four per row. The label above each molecule is its heavy-atom formula. Red edges are RingRef closures.

parent distance, the cosine of each child against the parent, and the dihedral, and drops the cosines between children. Validity falls from 96.56% to 90.12%, and molecule stability from 92.52% to 88.27%. Bond-length W_1 rises from 0.0040 Å to 0.0242 Å, and the dihedral distance from 1.187° to 2.874°. Sampling time is 12.1, against 12.3 for the full model. Those omitted cosines are what pin the children relative to one another. Without them the decoder still places every atom, but the bond-length and torsion distributions leave the reference.

Post-hoc ring closure. This variant generates the spanning tree only. Ring-closing edges are assigned after Cartesian coordinates have been built, rather than being drawn as RingRef nodes together with the rest of the molecule. Bond-length W_1 stays at 0.0038 Å, next to 0.0040 Å for the full model, and the angle and dihedral distances remain close. Validity falls to 92.4% and molecule stability to 85.4%, the larger of the two chemical drops. The added closure step raises the sampling time from 12.3 to 22.4. Local tree geometry does not require RingRef. The stability that disappears is the ring connectivity that was no longer part of the generated representation.

C.3 Property-Conditioned Generation on QM9

Each target is a separate QM9-L model trained on one half of the QM9 training set. The other half trains the regressors. The main table uses a frozen EGNN. We also train a TreeRef regressor on the same molecules and the same targets. It sees only the canonical tree. Multi-tree augmentation and the hidden-state alignment in Equation 8 are off. Table 8 is their mean absolute error on the QM9 validation split. The EGNN is the more accurate of the two on every target, which is why the main comparison uses it. Turning the same augmentation on for the TreeRef regressor may close part of that gap. Orbital errors are converted from Hartree by $27.2114\,\mathrm{eV}$.

Table 8: Validation mean absolute error of the two property regressors. Both are trained on the scorer half of the QM9 training set and evaluated on the QM9 validation split.

Property	EGNN	TreeRef
α (Bohr^3)	0.132	0.318
$\Delta\epsilon$ (meV)	99.8	137
ϵ_{HOMO} (meV)	69.1	95.6
ϵ_{LUMO} (meV)	53.0	93.5
μ (D)	0.057	0.434
C_v	0.055	0.101

Sampling uses classifier-free guidance (Ho & Salimans, 2022) with $w = 2$, 10,000 molecules, and the fast decoder. Table 9 is the error at 100 steps under both regressors. Figure 5 is the same error at every sampling budget. The TreeRef regressor reads the generated molecules as closer to the requested target than the EGNN does, on every property and every budget. That gap is the difference between the two evaluators, not a second generator. RDKit validity at 100 steps is 94.7% (α), 97.2% ($\Delta\epsilon$), 97.1% (HOMO), 96.9% (LUMO), 96.6% (μ), and 93.8% (C_v).

Table 9: Property mean absolute error at 100 steps. The generator is the same in both columns. NExT-Mol is the number in Table 5.

Property	EGNN	TreeRef	NExT-Mol
α (Bohr^3)	1.18	1.14	1.16
$\Delta\epsilon$ (meV)	225	174	297
ϵ_{HOMO} (meV)	142	136	205
ϵ_{LUMO} (meV)	169	136	235
μ (D)	0.531	0.495	0.507
C_v	0.513	0.445	0.512

C.4 Sampling Error Diagnosis

Two different failures. Figure 6 summarizes unconditional QM9 samples from TreeRef-BFN-S, M, L, and M-inv2. S, M, and L use one run of 10,000 molecules. All are decoded with the fast decoder.

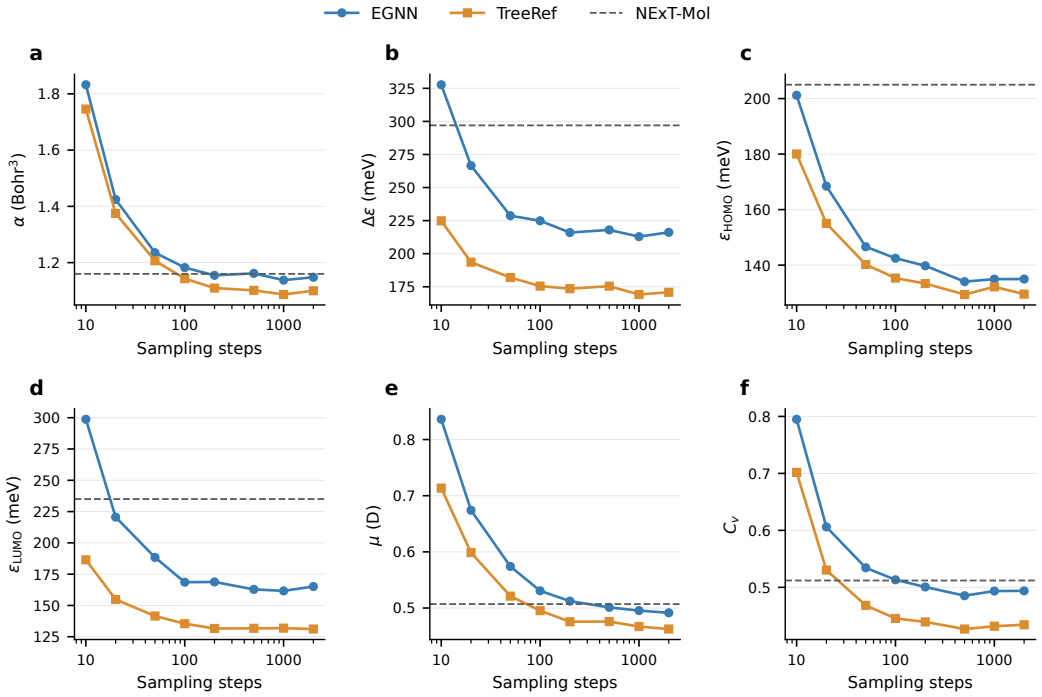


Figure 5: Property error across sampling budgets. Solid lines are the frozen EGNN and TreeRef regressors on the same generated molecules. The dashed line is NExT-Mol from Table 5. Orbital errors are converted from Hartree to meV. Each point is 10,000 samples. The horizontal axis is logarithmic.

A decode failure is almost always an unmatched RingRef. At 100 steps these are 1.48% (S), 0.62% (M), 0.34% (L), and 0.20% (M-inv2). The larger share is chemical: among decoded molecules, exact valence fails for 20.5% (S), 9.3% (M), 7.2% (L), and 10.3% (M-inv2). Both rates drop from 10 to 100 steps and then flatten. The rings in Figure 6c–f are the mix of those failures at 100 steps. M-inv2 has the same mix as M: under-valent carbon is the largest slice, and the combined failure share is 10.5%, next to 9.9% for M. The main unstable atom is an under-valent carbon, usually bond-order sum 3 where 4 is allowed, then an over-valent hydrogen. That failure is a label error on a graph whose coordinates are already placed: the element, the bond orders, or both do not match. RDKit rejects only the over-valent cases, so invalidity is lower than instability. An under-valent atom still sanitizes.

What rules can change. Rules does not draw a new sample, and it does not move coordinates. It repairs in two steps. First the RingRef. If its virtual point has no atom of the same element inside the single-bond window (`bonds1` plus 10 pm), and it does not lie on another atom, it is kept as a real atom at that point and bonded to its host. If it lies within half an order-1 bond length of some other atom, that atom is a RingRef predicted as a Real with the wrong element: rules changes the atom to the RingRef element and bonds the host to it, without adding a second atom. The second step relabels elements and bond orders on that connectivity. An isolated RingRef can still be under-valent after the first step. On the stored L samples at 200 steps, the RingRef step recovers 5 of 18 failures. The other 13 still have a same-element atom nearby and fail for BFS order or valence; rules leaves them unmatched. After the element and bond relabeling, 4 of those 5 also pass exact valence and RDKit. On the stored 10,000-sample runs at 100 steps the same two steps recover 51/148 (S), 25/62 (M), and 9/34 (L) RingRef failures. Of the decoded molecules that are unstable or RDKit-invalid, the relabeling restores both checks for 1988/2015 (S), 918/924 (M), and 707/714 (L). Element edits are more common than bond-order edits: 3836 against 754 on S, 1393 against 490 on M, and 957 against 431 on L. It does not move the atoms, so a long bond stays long; only the labels change.

Figure 7 shows both steps on QM9-L. In the first two rows the unmatched RingRef is the diamond on the left. Sample 3777 places a nitrogen 0.09 Å from an oxygen. Rules changes that oxygen to

nitrogen and bonds the RingRef hosts to it. The site then has bond-order sum 3, so the second step leaves it. Sample 5472 promotes the nitrogen in open space, 1.31 Å from its carbon host. That nitrogen and a nearby oxygen are still under-valent, so the second step changes the nitrogen to fluorine and the carbon–oxygen single bond to a double bond. The last row is an over-valent nitrogen. Rules changes two nitrogens into carbon and oxygen. The coordinates are the fast decoder's coordinates.

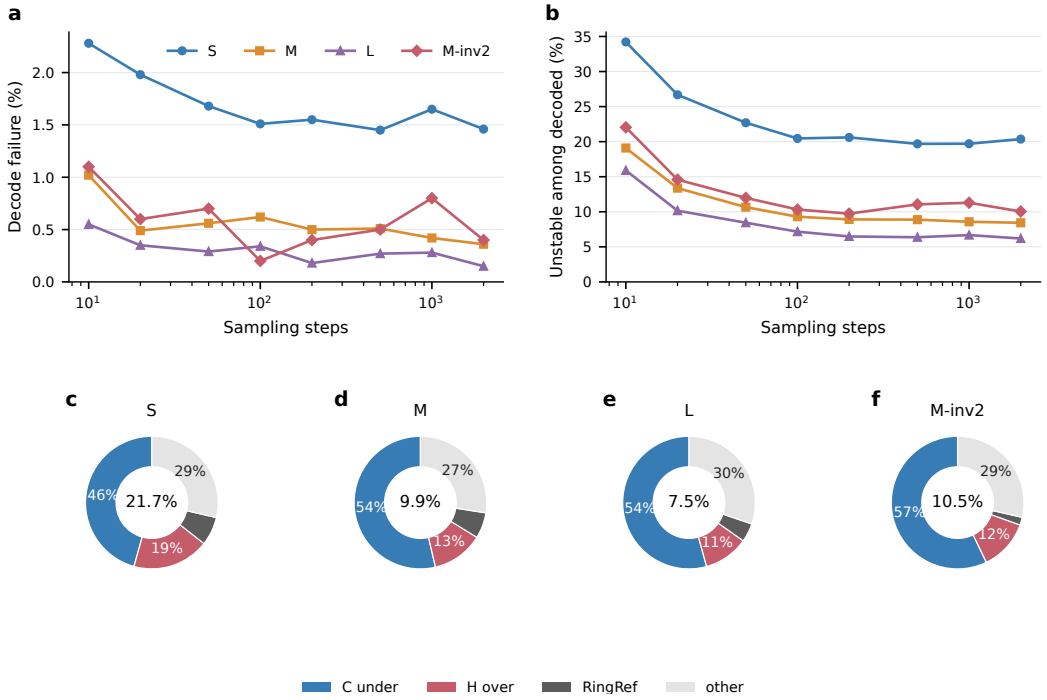


Figure 6: Sampling errors for unconditional QM9. (a) Decode-failure rate and (b) exact-valence instability among decoded molecules. S, M, and L use 10,000 fast-decoded samples. (c–f) Composition of the failures at 100 steps: unmatched RingRefs, under-valent carbon, over-valent hydrogen, and the remaining valence errors. The number in the ring is their combined share of all samples. The horizontal axes in (a) and (b) are logarithmic.

C.5 Conditioning Strategies and Model Scaling for 2D-to-3D Generation

What the sweep shows. Frozen conditioning gives the model the real graph from the start and asks it only for geometry and handedness. RePaint-style conditioning keeps that graph noisy during sampling and puts the clean tokens back only at the end. With the graph held fixed, S is the most often valid of S, M, and L once bonds are read back from the coordinates, and its bond and angle distributions stay closest to the reference, while M lands nearest the reference conformer (Figure 8). L is behind both, so the usual gain from a larger unconditional model does not show up here. M-inv2 breaks that size ordering. It is the M model trained with two tree serializations and the hidden-state alignment of Equation 8. With the graph held fixed, xyz2mol validity stays near 94.5% from 10 to 2,000 steps, above S, and the mean aligned RMSD stays near 0.66 Å, below M. Bond, angle, and dihedral distances are the smallest of the four. Extra steps do not create this gap. RePaint-style runs, including M-inv2, sit near 5% xyz2mol validity, and extra steps do not help. The aligned summary can use a clean graph. It does not help while that graph is still noised.

Deleting a channel. We take validation molecules, noise every channel together, then replace one channel by its prior and ask for a single denoising step (Figure 9a). At $t = 0.1$, dropping the element or the bond barely moves the distance error, by less than 0.004 Å on S, M, L, and M-inv2. Dropping geometry is different: the distance error rises from about 0.09 to 0.16 Å, and element accuracy falls from about 0.67 to 0.21–0.26. At $t = 0.5$ the same split is sharper. With geometry left in place the distance error is about 0.001 Å; with it removed, element accuracy drops below 0.04. We also run ordinary unconditional sampling, 32 molecules and 20 steps, but blank one channel's input before

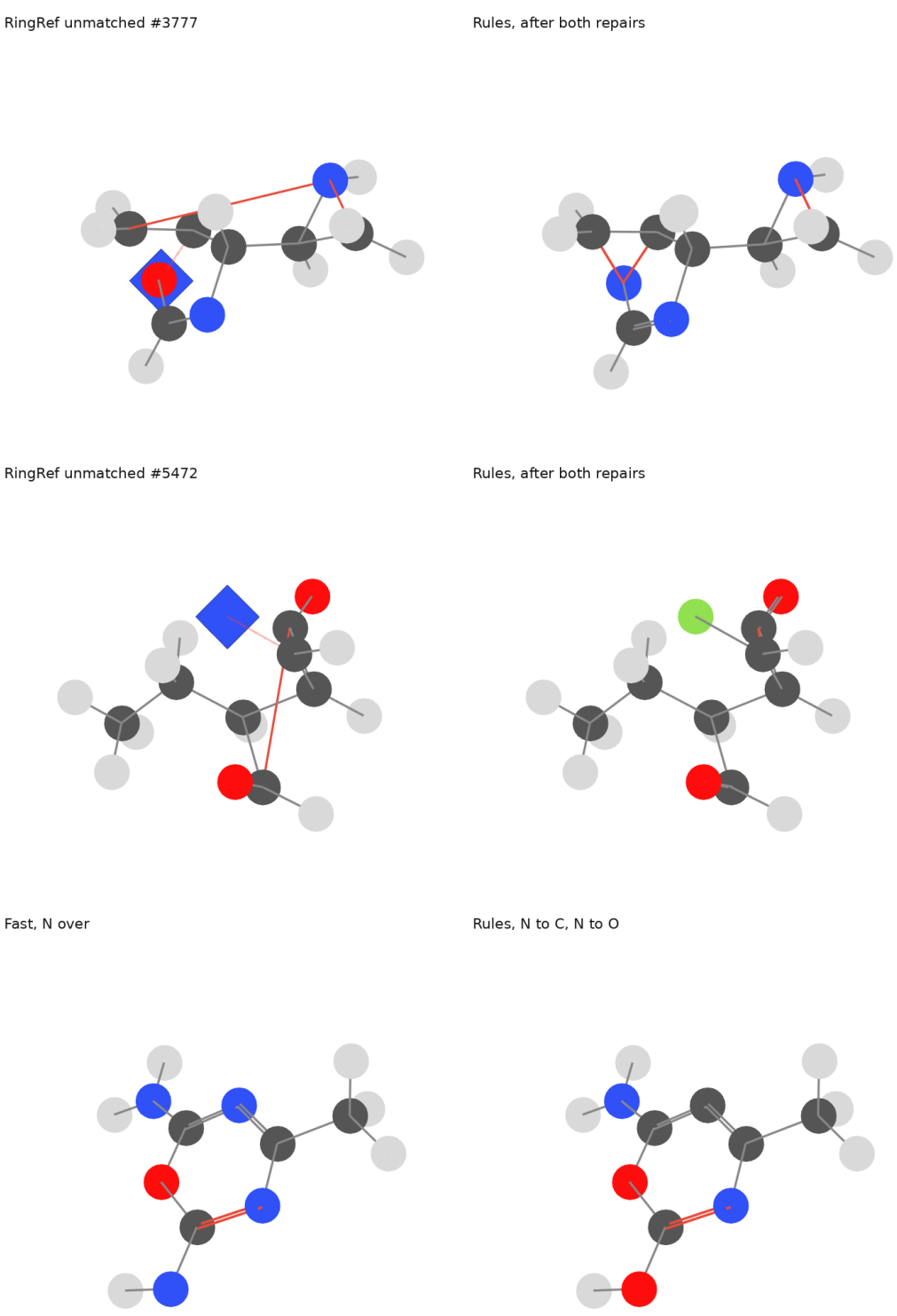


Figure 7: QM9-L repairs. Left is the fast result, right is rules. The first two rows are RingRef samples 3777 (200 steps) and 5472 (100 steps). The diamond is the unmatched RingRef; the dotted segment joins it to its host. In 3777 rules changes the oxygen under the diamond to nitrogen. In 5472 the promoted nitrogen is still under-valent, so the second step writes it as fluorine and doubles the carbon–oxygen bond. The last row is a 100-step molecule whose nitrogens are over-valent. Rules replaces two of them by carbon and oxygen. Coordinates are not moved.

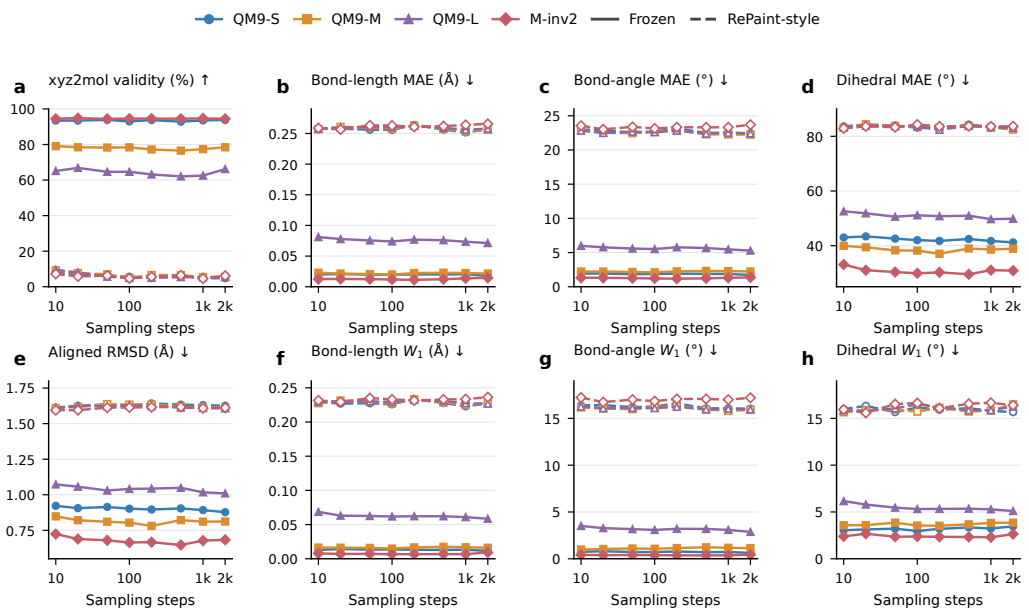


Figure 8: Frozen versus RePaint-style 2D-to-3D generation on QM9. Colors indicate QM9-S, QM9-M, QM9-L, and M-inv2; solid and dashed lines denote Frozen and RePaint-style conditioning. Panel (a) is xyz2mol validity: bonds are inferred from the generated coordinates, then RDKit sanitization and a single component are required. Panels (b–d) are mean absolute errors of bond lengths, bond angles, and dihedral angles against the reference connectivity. Panels (e–h) are mean aligned RMSD and Wasserstein-1 distances of the three local geometric distributions. The sampling-step axis is logarithmic. Each point uses 1,000 conditioning molecules, with two conformers for Frozen and one for RePaint-style. Errors are means over eligible conformers, not a best-of-two score.

every denoiser call (Figure 9b). Blanking geometry hurts M, L, and M-inv2. Blanking the element or the bond does not. The network is reading the coordinates to decide the graph, not the other way around.

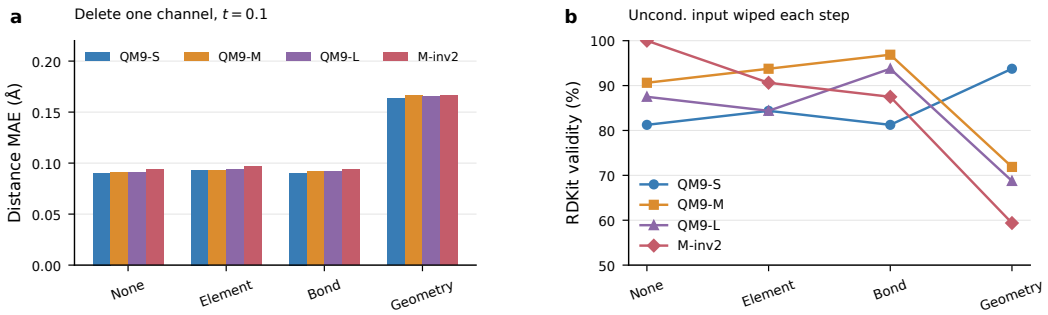


Figure 9: Channel deletion. (a) One-step distance error at $t = 0.1$ after replacing the named input channel by its prior; other channels stay at the training forward state. Sixty-four validation molecules, one noise seed. (b) Unconditional sampling with that channel's input reset to the prior before every denoiser call (32 molecules, 20 steps). Curves are QM9-S, QM9-M, QM9-L, and M-inv2. Validity in (b) uses the generated bond table.

Why the smaller models win here. Frozen 2D-to-3D hands the model a finished graph and a geometry that still starts from the prior. That is the reverse of how these models were trained, where the coordinates become informative first and the discrete labels follow them. L leans on that coupling, so an exact graph paired with an uninformed geometry pulls its distances off and the rings stay open. S leans on it less, and can still place local geometry from the graph it is given. M-inv2 has the same width as M, but the multi-tree augmentation and the hidden-state alignment stop it from treating one traversal as the molecule, so the clean graph is enough to place the geometry. RePaint-style sampling never gets a clean graph while the coordinates are being formed, so putting the right bonds back at the end cannot fix them, and the alignment has nothing clean to read.

C.6 Fragment Completion and Editing

The four edits leave a piece of the source molecule where it already sits in the tree, and ask the model to grow the rest. Geometry and handedness on the cut are thrown away, so the new attachment is not a copy of the old pose. What changes from one task to the next is how much is left to invent. A linker is the easy case: most of the molecule stays, and only a short bridge is new. Decoration keeps one connected scaffold and grows substituents onto it. Multi-fragment completion keeps more atoms than decoration, but they arrive as two separate pieces that have to meet again, and the model more often just hands the source molecule back. Scaffold replacement keeps the least and has to rebuild the core. We call an edit a success when the fragment we supplied is still there: one fragment after decoration, both fragments on the other three tasks. A success is already a valid molecule. Frozen conditioning shows those atoms as clean tokens. RePaint-style conditioning noises them along with everything else and puts the clean tokens back only when the trajectory is finished. The fragment is not moved around the tree to find a better slot.

Sampling longer does not change which task is hard, and it does not close the gap between the two ways of showing the fragment (Figure 11). When the fragment is shown cleanly, M beats S and L on every task. The miss is largest on scaffold replacement, where L has to invent a core from a small piece and falls well behind. Once the fragment is noised, that order turns around: L is a little ahead on decoration, multi-fragment completion, and linkers. Scaffold replacement stays poor for all three, around 10–15% success. The clean fragment was doing the work. Without it, a larger model is only slightly better at finishing the molecule.

M-inv2 is this M model with two cached tree serializations per molecule and the consistency term in Equation 8. The two views differ in parents, RingRef placement, and local frames. They are noised

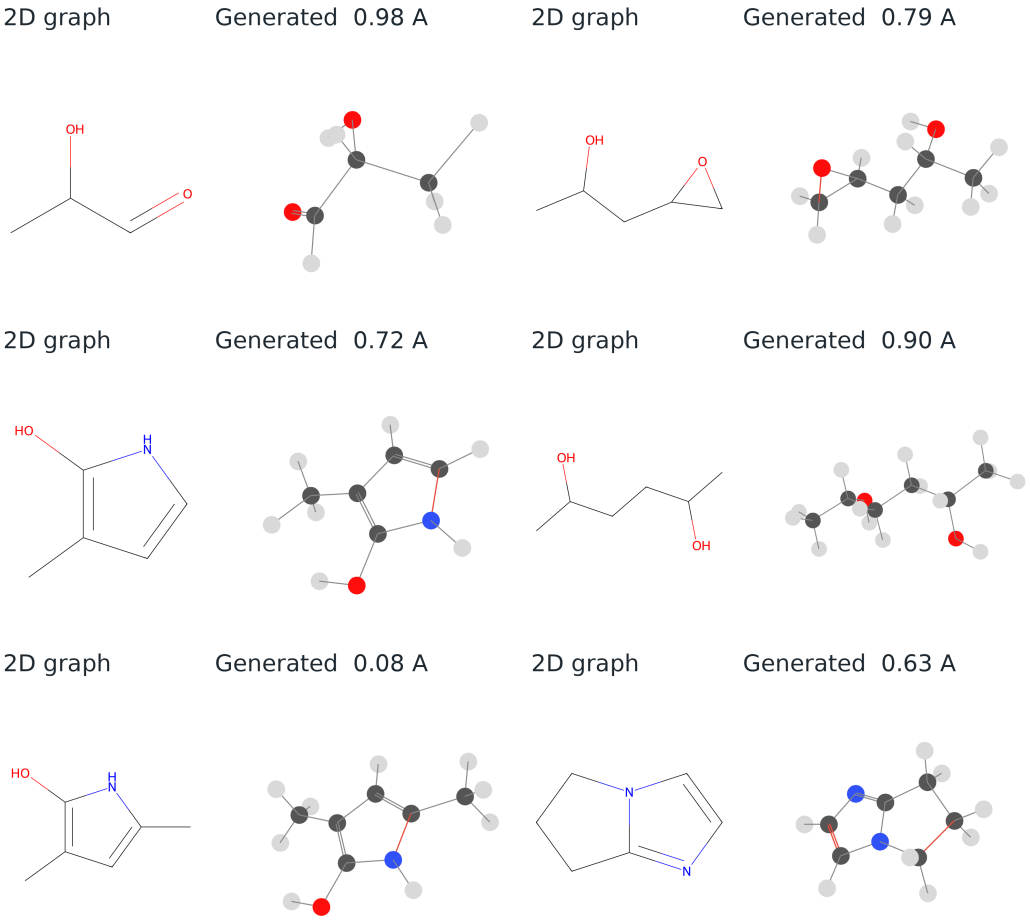


Figure 10: QM9-L 2D-to-3D generation at 100 steps. Each pair places the input graph on the left and one generated conformer on the right. The left drawing is the RDKit 2D depiction with hydrogens removed. The label on the 3D panel is the heavy-atom formula and the RMSD to the reference conformer, in angstroms.

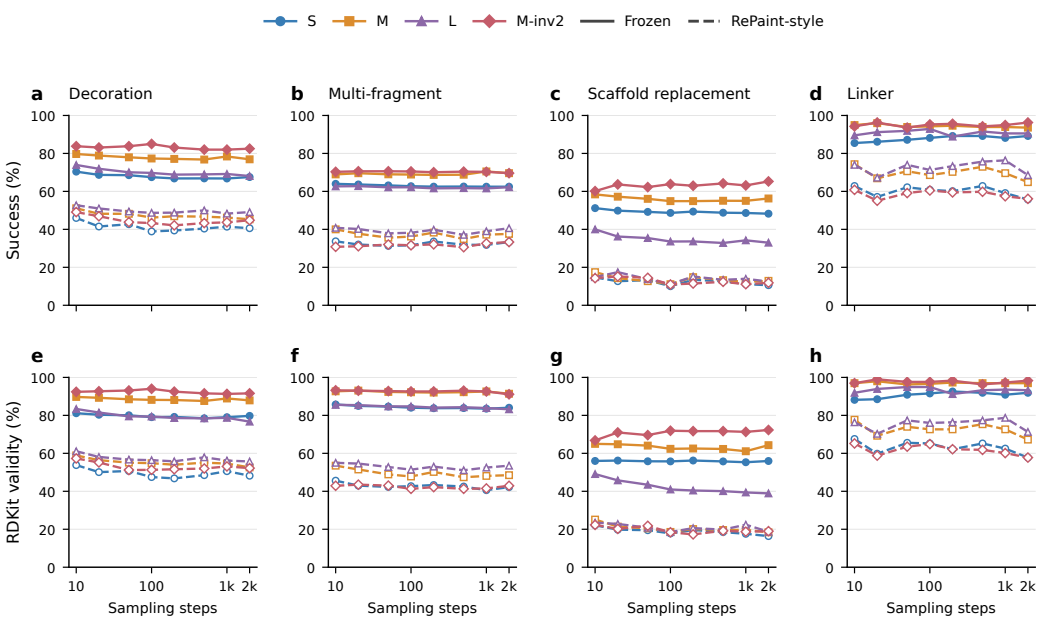


Figure 11: Fragment editing on QM9 across sampling budgets. Top row, success. Bottom row, RDKit validity. Colors are S, M, L, and M-inv2. Solid lines are Frozen conditioning and dashed lines are RePaint-style. Columns are decoration, multi-fragment completion, scaffold replacement, and linker generation. The step axis is logarithmic. S, M, and L Frozen runs use 10,000 conditions. RePaint-style runs use 1,000. Linker generation has 296 eligible conditions.

separately at one shared time, and their pooled hidden states are pulled together. S, M, and L see only one traversal.

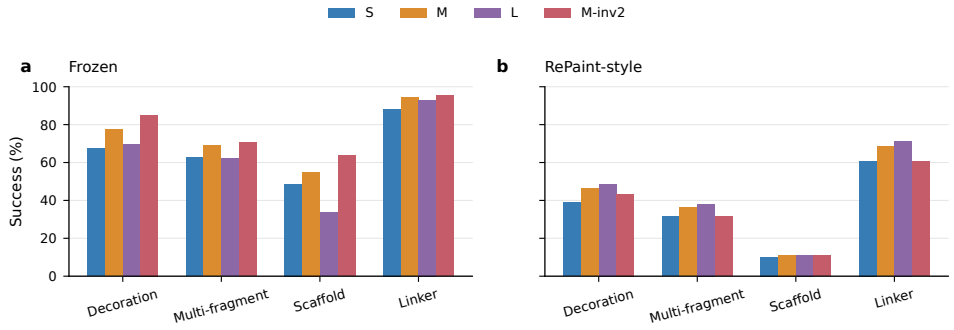


Figure 12: Fragment-edit success at 100 steps for S, M, L, and M-inv2. (a) Frozen conditioning. (b) RePaint-style conditioning. S, M, and L Frozen use 10,000 conditions. RePaint-style runs use 1,000.

These edits were never in that loss. The kept atoms stay in the slots the source tree already assigned, and we do not search for a more convenient serialization. A one-view model can treat that particular traversal as part of the molecule. The extra views and the alignment remove that shortcut: the summary has to match after the traversal changes. Figure 11 shows that the resulting gain is not a single operating point. From 10 to 2,000 steps, Frozen M-inv2 stays above L on every task and above M on decoration and scaffold replacement. At 100 steps scaffold replacement is 63.8%, against 54.9% for M and 33.6% for L. Decoration, multi-fragment completion, and linkers stay with M or a little above it. Extra sampling steps do not create the gap and do not remove it. The

clean fragment is being read as the molecule, rather than as one way of writing it down. That is the generalization from multi-tree augmentation and hidden-state alignment.

RePaint-style sampling does not get this. The fragment is noised with the atoms being drawn and restored only at the end, so the aligned summary is not what the model sees while it samples. M-inv2 then sits a little below M (Figure 12b).

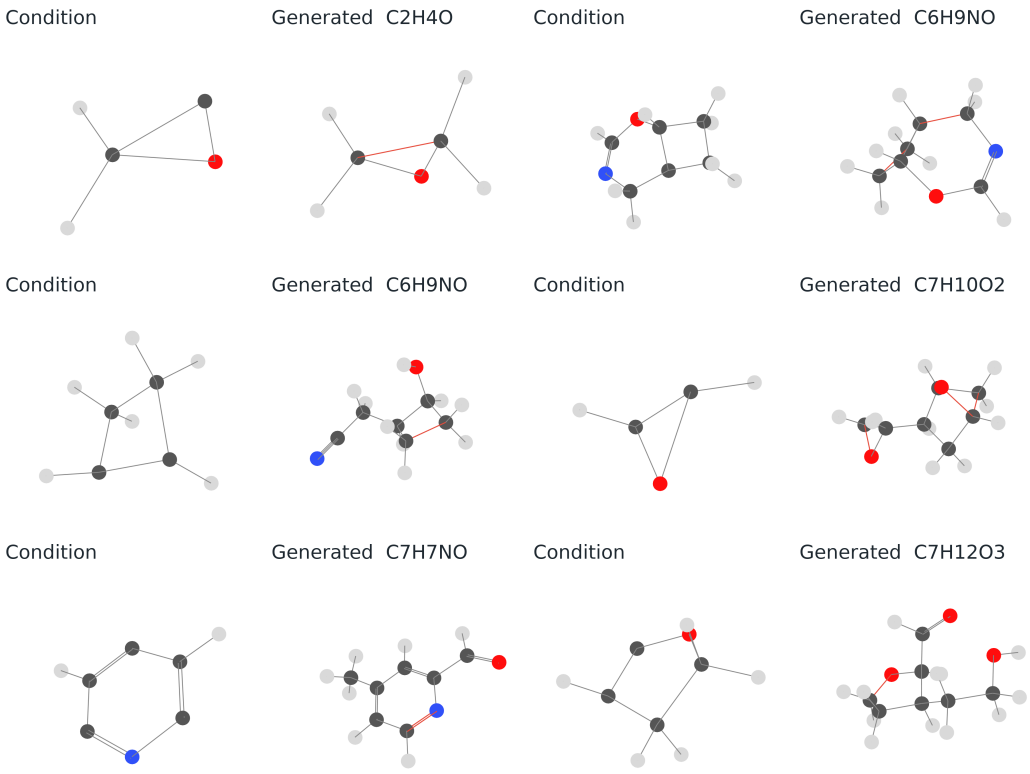


Figure 13: QM9-L scaffold decoration at 100 steps. The left panel is a 3D embedding of the conditioning scaffold. The right panel is the generated molecule.

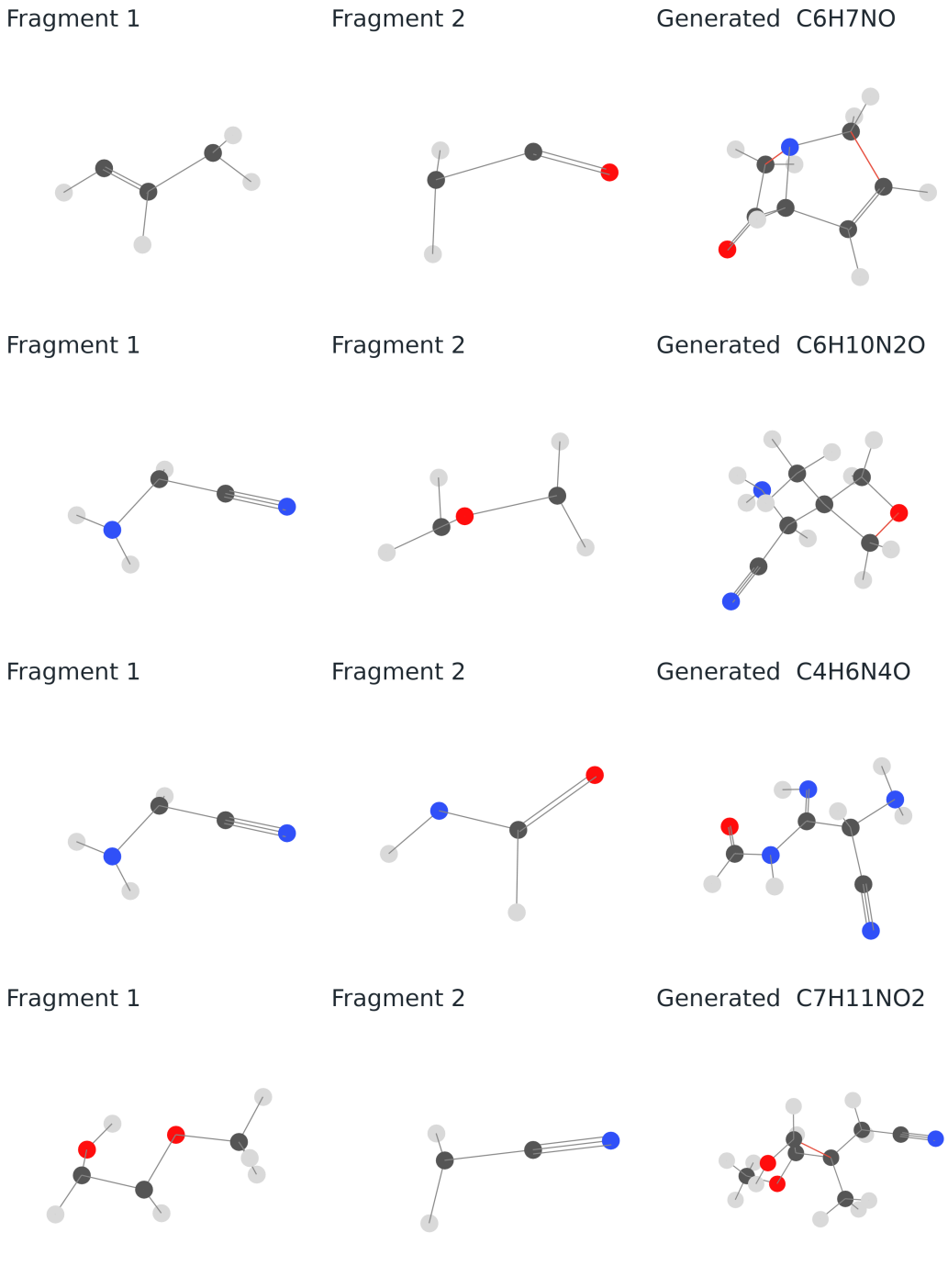


Figure 14: QM9-L multi-fragment completion at 100 steps. Each case shows the two conditioning fragments and the generated molecule. Fragments are embedded from SMILES without adding hydrogens at the cut.

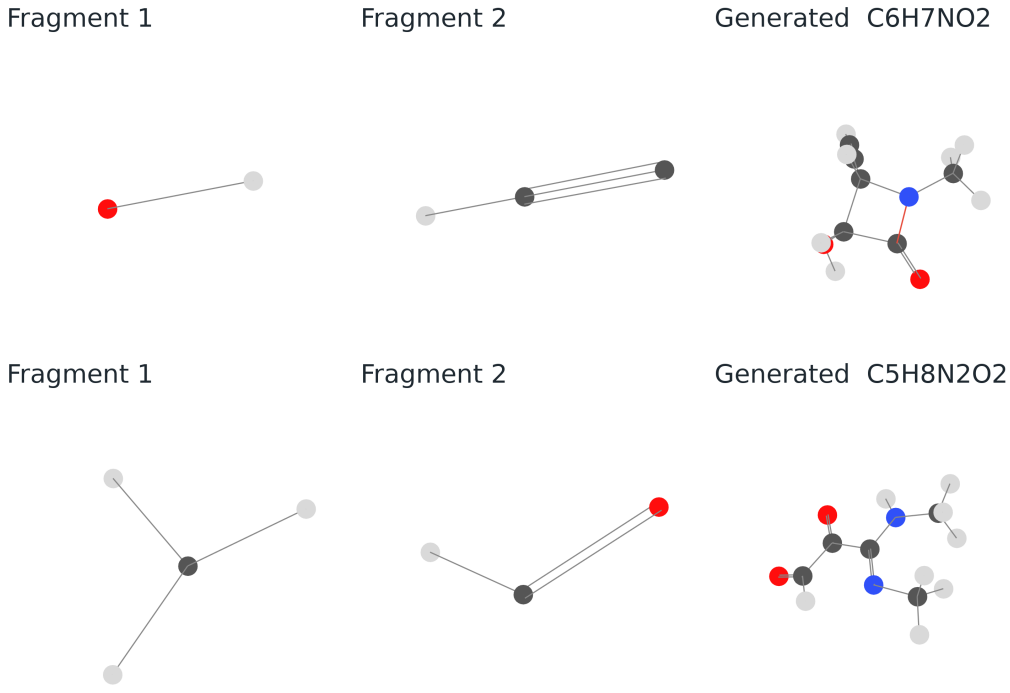


Figure 15: QM9-L scaffold replacement at 100 steps. Only cases with at least two side chains are shown. The left panels are those side chains; the right panel is the generated molecule.

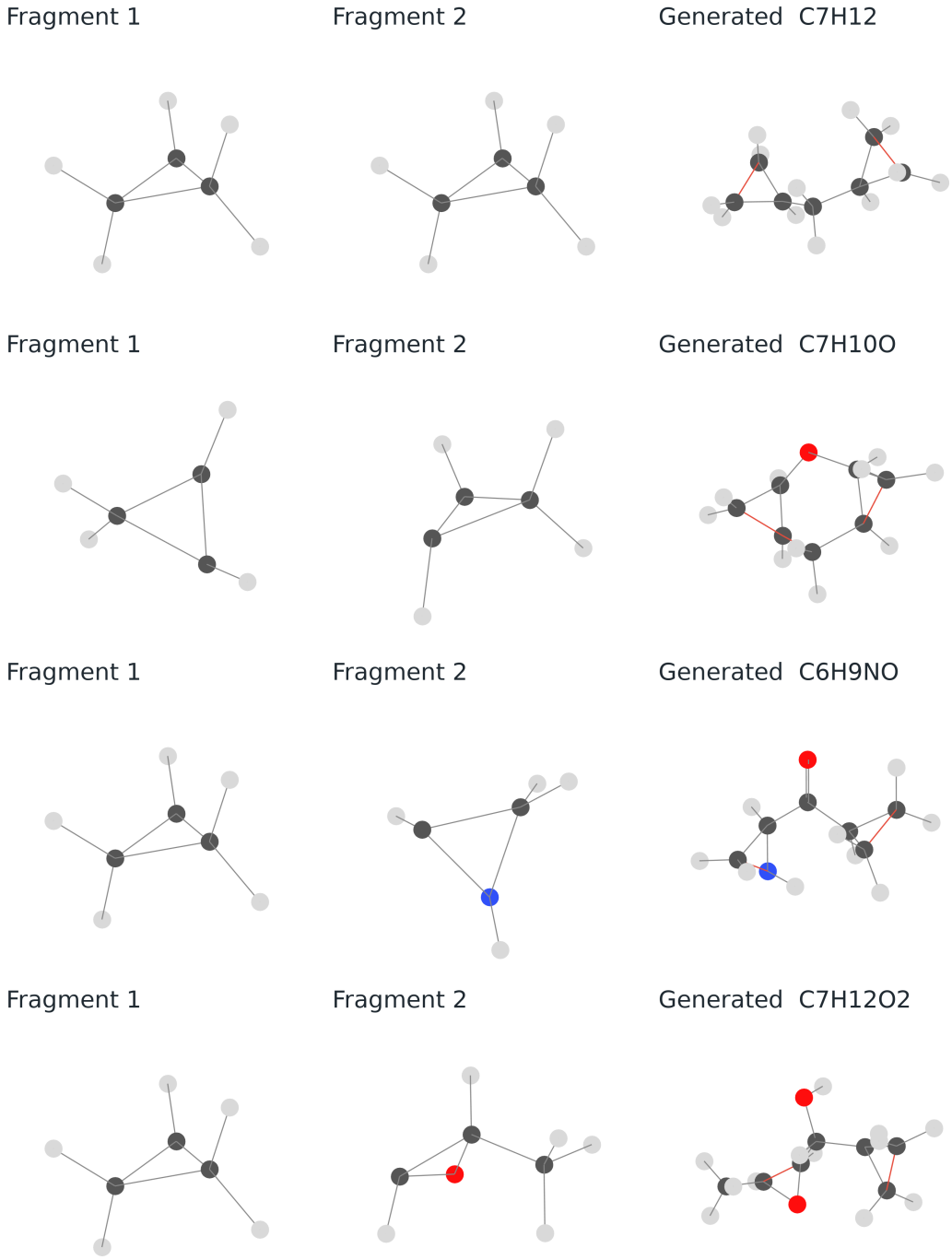


Figure 16: QM9-L linker generation at 100 steps. The left panels are the two anchors that are kept. The right panel is the generated molecule that joins them.