SCALEMOL: DEFINING AND LEVERAGING SCALES FOR 3D MOLECULAR GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Small molecules lack a natural definition of scale, making it difficult to organize their generation from coarse structure to fine detail. We address this problem by introducing *resolution-consistent molecular mass grids* (MMGs), which define molecular scales through the spatial resolution of a molecular field. Coarse MMGs capture regional element content, while finer MMGs progressively resolve its spatial distribution, with exact consistency across scales. Based on this scale definition, we introduce ScaleMol, which explicitly predicts and supervises molecular fields from coarse to fine. On QM9 and GEOM-Drugs, ScaleMol improves atom and molecule stability over existing field-based generators, while controlled ablations verify the benefits of both MMGs and coarse-to-fine prediction. The same scale hierarchy also enables conditioning on a reference molecule at different resolutions without additional training. Together, our results provide a simple definition of scale for small molecules and demonstrate its utility for 3D molecular generation.

1 INTRODUCTION

Complex structured objects often exhibit organization at multiple scales, suggesting a natural coarse-to-fine strategy for generation that first establishes global structure and then progressively resolves finer details. Central to this paradigm is a meaningful definition of scale, through which the same object can be represented at different levels of granularity. Images provide a canonical example, where spatial resolution naturally defines scales from broad layouts to fine visual details. Building on this structure, VAR (Tian et al., 2024) formulates image generation as next-scale prediction, progressively predicting finer token maps conditioned on coarser ones. Protein backbones provide another example, where their ordered chain structure supports multiscale representations from overall backbone organization to increasingly local structural detail. PAR (Qu et al., 2026) exploits this hierarchy to generate protein structures progressively across scales.

In parallel, three-dimensional molecular generation has emerged as an important approach to molecular design and drug discovery. A natural question is how scale-based generation can be extended to three-dimensional molecules. Defining meaningful molecular scales, however, is less straightforward. Atoms are not arranged on a regular grid like pixels, and molecular graphs can contain branches and rings rather than following a single ordered chain. It is therefore unclear how atoms should be grouped to form coarser scales without losing the information needed at finer scales. Existing methods such as HierDiff and EBD construct molecular hierarchies using chemically defined fragments (Qiang et al., 2023; Park & Shen, 2024). While chemically meaningful, these fragments replace groups of atoms with different structural units, rather than providing coarse and fine views that can be directly related across scales. How to define such molecular scales and exploit them for coarse-to-fine generation therefore remains underexplored.

We address this challenge by defining molecular scale through spatial resolution rather than molecular decomposition. We first represent a molecule as a continuous field in three-dimensional space. Each atom contributes unit mass to its element channel, giving regional values an additive atom-count interpretation. As illustrated in Figure 1, we observe this field at different resolutions to construct *resolution-consistent **M**olecular **M**ass **G**rid**s*** (**MMGs**). At a coarse resolution, each grid cell describes the elemental content within a larger spatial region. At finer resolutions, the same content is progressively localized to smaller regions, revealing increasingly detailed molecular structure. Im-

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

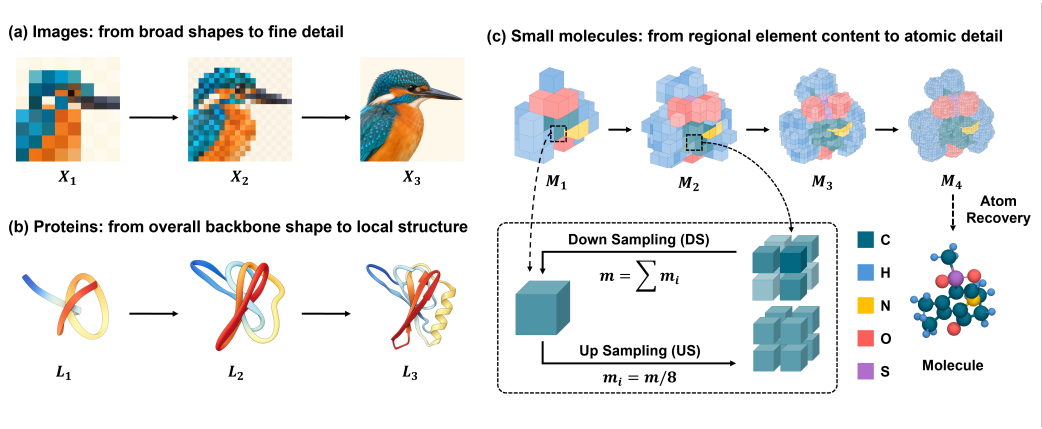

Figure 1: **Scales for structured-data generation.** **(a)** Images progress from broad shapes to fine detail, illustrating the coarse-to-fine principle used by VAR (Tian et al., 2024). **(b)** Protein backbone representations progress from overall shape to local structural detail (adapted from PAR (Qu et al., 2026)). **(c)** MMGs describe coarse atom distributions through regional element content and resolve their spatial allocation at finer scales. The views are rendered from an actual generated molecular field; atoms are recovered from the finest field. Downsampling (DS) sums eight child cells, while upsampling (US) distributes parent mass equally among them. Scale arrows indicate finer representations, not sampling time.

portantly, these representations are directly related across scales, since summing finer cells exactly recovers their corresponding coarser cells.

To leverage these molecular scales for generation, we introduce ScaleMol. ScaleMol uses a shared 3D denoiser to predict the clean molecular field progressively from coarse to fine, with explicit supervision at each scale. Each finer prediction builds on the preceding coarser prediction while remaining free to correct its errors. This coarse-to-fine prediction is performed within each denoising step, while the sampling process evolves only the finest-resolution field.

An additional benefit of this scale hierarchy is that it naturally supports conditioning on a reference molecule without additional training. A reference molecule can be represented at a selected coarse level of the hierarchy and used as a *resolution prefix*, replacing the model prediction at the corresponding scale and guiding subsequent finer predictions. The prefix resolution determines how much spatial detail is specified by the reference, while finer scales remain free to generate unresolved molecular structure. During sampling, protected readouts preserve the supplied regional totals in the predicted field while allowing finer structural variation.

On QM9 and GEOM-Drugs, ScaleMol consistently improves molecular stability over existing field-based methods. The advantage is particularly pronounced on the larger and more structurally diverse GEOM-Drugs dataset, where ScaleMol achieves 88.3% molecule stability without chemical post-processing, compared with 77.4% for VecMol and 57.9% for GeoLDM. Controlled ablations show that both the MMG representation and explicitly organizing prediction across scales contribute to these improvements. Conditioning experiments further show that scale resolution naturally controls the trade-off between reference preservation and structural diversity without additional training.

Our main contributions are as follows:

- We define molecular scales through *resolution-consistent **M**olecular **M**ass **G**rids* (**MMGs**), which represent the same molecular field at different spatial resolutions with exact consistency across scales.
- We introduce ScaleMol, which explicitly predicts and supervises MMGs from coarse to fine, allowing finer predictions to build on coarser structure while correcting earlier errors.
- We show that ScaleMol improves molecular stability over existing field-based generators on QM9 and GEOM-Drugs, with particularly strong gains on the larger and more diverse GEOM-Drugs dataset. Controlled ablations verify the benefits of both MMGs and coarse-

to-fine prediction, while the same scale hierarchy enables reference conditioning at different molecular scales without additional training.

2 Related Work

Multiscale generation and representations. VAR generates image token maps scale by scale (Tian et al., 2024); FlowAR combines continuous multiscale latents with flow matching (Ren et al., 2025). Related hierarchies support protein generation in PAR (Qu et al., 2026), shape generation in Octree Transformer (Ibing et al., 2021), and graph generation in MAG (Belkadi et al., 2025). Cryo-SWAN learns multiscale representations of cryo-EM densities (Li et al., 2026). Our hierarchy instead comprises regional element observations predicted within each denoiser evaluation.

Atom-based molecular generation. G-SchNet generates atoms autoregressively (Gebauer et al., 2019). EDM diffuses coordinates and atom types (Hoogeboom et al., 2022), GeoLDM uses equivariant latents (Xu et al., 2023), and MiDi jointly generates graphs and coordinates (Vignac et al., 2023). HierDiff assembles molecules from generated fragments (Qiang et al., 2023); EBD restores atomic detail from fragment coordinates conditioned on a molecular graph (Park & Shen, 2024). Unlike these fragment-to-atom hierarchies, MMGs retain regional element-content semantics across scales.

Field-based molecular generation. Compared with atom-based methods, field-based approaches model elemental distributions rather than explicit atomic states, recovering atoms without prescribing their number. Early work generates density grids and fits atoms to them (Ragoza et al., 2022). VoxMol denoises voxel grids (Pinheiro et al., 2023); FuncMol uses WJ sampling in neural-field latent space (Kirchmeyer et al., 2024); VecMol diffuses vector-field latent codes (Hua et al., 2026). These methods do not explicitly predict nested regional element content. MMGs instead define scales through regional element content with shared atom-count semantics. ScaleMol uses them as supervised targets and parent conditions.

3 Method

Problem formulation. A conformation is an unordered atom set $\mathcal{M} = \{(r_i, c_i)\}_{i=1}^{N}$, with positions $r_i \in \mathbb{R}^3$, elements $c_i \in \{1, \dots, C\}$, and variable N. Given samples from p_{data}, we learn an unconditional generator q_θ. Bonds are inferred after generating the atom set.

3.1 Resolution-Consistent Molecular Mass Grids (MMGs)

Continuous molecular field. Both VoxMol and ScaleMol represent atoms with Gaussian fields in element-specific channels (Pinheiro et al., 2023). Their interpretation and combination differ: VoxMol uses a probability-like occupancy, combining bounded atomic responses $g_i(x) \in [0, 1]$ as $o_c(x) = 1 - \prod_{i:c_i=c}(1 - g_i(x))$. This union saturates where same-element atoms overlap. We instead assign unit mass to each atom and add its density:

$$\rho_c(x; \sigma_f) = \sum_{i:c_i=c} \mathcal{N}(x; r_i, \sigma_f^2 I). \tag{1}$$

Each atomic density integrates to one, so mass is atom-count-equivalent, not atomic weight: $\int_{\mathbb{R}^3} \rho_c(x)\, dx = N_c$. This additive construction provides unit contributions per atom, an explicit counting interpretation, and nonsaturating accumulation where same-element atoms overlap.

Grid observations. We associate each grid value with the mass of a spatial cell: let $\mathcal{G}_h$ partition a fixed domain Ω into cubic cells of width h, and define

$$M_{h,c}(V) = \int_V \rho_c(x; \sigma_f)\, dx, \qquad V \in \mathcal{G}_h. \tag{2}$$

Each value therefore measures soft, atom-count-equivalent element content within a region rather than pointwise field strength. An atom contributes a fraction of its unit mass to each cell, so this is not a hard count of atom centers. A finite box excludes external Gaussian tails.

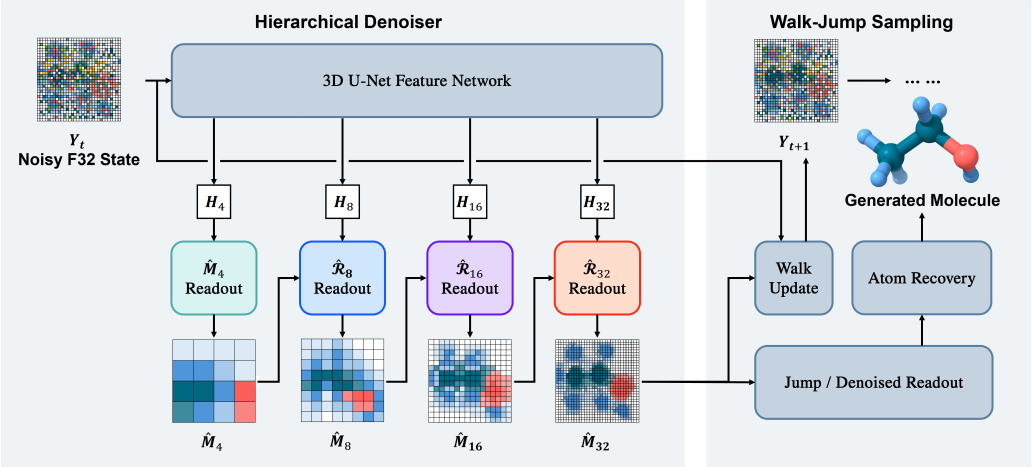

Figure 2: **ScaleMol overview.** Within each denoiser evaluation, shared 3D U-Net features support parent-conditioned MMG predictions at increasing resolutions. The finest prediction defines the Walk drift and the Jump estimate used for atom recovery; Jump outputs are not fed back into Walk. Figure 6 details the architecture and supervision.

Resolution and interscale operators. We define observation resolution by cell width: hold σ_f fixed and use nested grids indexed from coarsest $k = 0$ to finest $k = L$, with $h_{k+1} = h_k/2$ and $M_k = M_{h_k} = \mathcal{O}_k(\mathcal{M})$. For the eight children $\mathrm{ch}(V)$ of a coarse cell, aggregation and equal-mass lifting are

$$(D_k M_{k+1})_c(V) = \sum_{v \in \mathrm{ch}(V)} M_{k+1,c}(v) = M_{k,c}(V), \tag{3}$$

$$(U_k m)_c(v) = m_c(V)/8, \qquad v \in \mathrm{ch}(V). \tag{4}$$

Thus $D_k U_k = I$: lifting preserves regional mass but not unresolved geometry. Exact aggregation follows from integration over nested regions (Appendix A).

In contrast, pooling VoxMol's center-sampled occupancies does not generally reproduce occupancy at coarse-grid locations. On QM9, mean pooling from 32^3 to 16^3 and 8^3 yields relative L_1 discrepancies of 12.7% and 52.2%, respectively, whereas summed MMGs agree with direct coarse integration to numerical precision (Appendix H).

3.2 Hierarchical Field Prediction and Supervision

Figure 2 summarizes the denoiser and sampling process; Figure 6 in Appendix B expands the architecture and supervision.

Shared features. Normalize the finest grid as $X = M_L/a \in \mathbb{R}^d$, where $d = C|\mathcal{G}_{h_L}|$ and a is the centered single-atom cell mass (Appendix B.3). Only this grid is corrupted:

$$Y = X + \sigma_n \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I). \tag{5}$$

A 3D U-Net-style backbone (Çiçek et al., 2016) extracts decoder features $(H_0, \ldots, H_L) = E_\theta(Y)$ aligned to the grids. As an implementation refinement, our full model uses gated fusion of pooled finer-resolution encoder features at each downsampling transition (Appendix B.1). Predicted fields condition readouts, without feeding back into the backbone within a forward pass.

Hierarchical readout. Writing $\widehat{M}_k$ for predicted observations, the heads satisfy

$$\widehat{M}_0 = R_{\theta,0}(H_0), \qquad \widehat{M}_k = R_{\theta,k}(H_k, \widehat{M}_{k-1}), \quad k = 1, \ldots, L. \tag{6}$$

Zero-sum detail can redistribute a parent's mass but cannot correct an erroneous atom count. We therefore predict both fine-scale detail $\widetilde{r}_k$ and a parent-mass correction Δ_{k-1} from backbone features

and the parent field (Figure 6 in Appendix B.2):

$$\begin{aligned} r_k &= (I - U_{k-1}D_{k-1})\widetilde{r}_k, \\ \widehat{M}_k &= U_{k-1}(\widehat{M}_{k-1} + \Delta_{k-1}) + r_k. \end{aligned} \tag{7}$$

Detail redistributes mass within parent cells, while the correction revises their totals. Exact aggregation defines the target hierarchy, not a hard constraint between predicted heads.

Multiresolution supervision. Let $D_{k\leftarrow L} = D_k \circ \cdots \circ D_{L-1}$, with $D_{L\leftarrow L} = I$, and let ℓ_k be mean squared error in finest-cell-equivalent units (Appendix B.3). Intermediate losses supervise individual heads; projection losses supervise the final prediction's regional mass:

$$L_k = \ell_k(\widehat{M}_k, M_k), \qquad J_k = \ell_k(D_{k\leftarrow L}\widehat{M}_L, M_k), \quad k < L. \tag{8}$$

With $L_L = \ell_L(\widehat{M}_L, M_L)$, the objective is

$$\mathcal{L}(\theta) = \mathbb{E}_{M_L,\epsilon}\left[L_L + \sum_{k<L} \lambda_k L_k + \sum_{k<L} \mu_k J_k\right]. \tag{9}$$

3.3 Walk–Jump Sampling

Walk–jump sampling explores a Gaussian-smoothed distribution and denoises sampled states (Saremi & Hyvärinen, 2019; Pinheiro et al., 2023). Let p_{σ_n} denote the noisy-field distribution in Equation 5. Using only the normalized finest prediction $\widehat{X}_\theta(Y) = \widehat{M}_L(Y)/a$, we estimate its score as

$$\nabla_Y \log p_{\sigma_n}(Y) = \frac{\mathbb{E}[X \mid Y] - Y}{\sigma_n^2} \approx s_\theta(Y) := \frac{\widehat{X}_\theta(Y) - Y}{\sigma_n^2}. \tag{10}$$

Walk updates the single noisy state Y and its velocity by underdamped Langevin dynamics at fixed σ_n. After K steps, Jump returns

$$\widehat{M}_L = a\big(Y_K + \sigma_n^2 s_\theta(Y_K)\big) = a\widehat{X}_\theta(Y_K). \tag{11}$$

The clean output is used for atom recovery and never fed back into Walk, allowing different molecules to be read out at different steps along the same trajectory (Figure 7). Appendix F specifies initialization, dynamics, discrete updates, and branching.

3.4 Atom Recovery

MMG recovery combines plateau-aware peak detection with local log-quadratic interpolation. Above-threshold local maxima in a bounded transform of $\widehat{M}_L$ define atoms, with connected plateaus merged into single candidates. Channels determine element types, and peak counts determine atom counts. Local log-quadratic interpolation gives subgrid coordinates without further optimization. Algorithm 1 in Appendix B.4 specifies the readout; Appendix E separately evaluates fixed-count/type L-BFGS refinement.

3.5 Prefix-Conditioned Generation

For reference-guided generation, we supply an MMG observation $m_k = \mathcal{O}_k(\mathcal{M}^{\mathrm{ref}})$ at level $0 \le k < L$ in a specified coordinate frame and sample $\widetilde{\mathcal{M}} \sim q_\theta(\cdot\,; k, m_k)$. Aggregating m_k supplies all coarser observations $m_j = D_{j\leftarrow L}M_L^{\mathrm{ref}}$ for $j \le k$. At each denoiser evaluation, these observations replace predicted parents in both readout branches (Figure 10):

$$\widehat{M}_{j+1} = R_{\theta,j+1}(H_{j+1}, c_j), \qquad c_j = \begin{cases} m_j, & j \le k, \\ \widehat{M}_j, & j > k, \end{cases} \quad 0 \le j < L. \tag{12}$$

Reference fields replace parents in the same units without retraining; the backbone still processes only Y. To prevent later mass corrections from overwriting the prefix, we set known heads to m_j and project subsequent corrections onto the null space of aggregation to level k:

$$\Delta_j^{(k)} = (I - U_{j\leftarrow k}D_{k\leftarrow j})\Delta_j, \qquad j \ge k, \tag{13}$$

where $U_{j\leftarrow k}$ repeatedly applies equal-mass lifting, with identity operators for $j = k$. Using $\Delta_j^{(k)}$ in Equation 7 gives $D_{k\leftarrow L}\widehat{M}_L = m_k$. The immediate parent correction is zero; finer corrections can still redistribute mass within known cells. Without a prefix, all corrections remain unchanged.

We use this protected readout in both Walk drift and Jump, without projecting the noisy state or feeding Jump outputs back into Walk. It is a sampling intervention, not an exact conditional score model. Preservation holds for the predicted grid, not necessarily for atoms recovered from it; we therefore evaluate molecular quality and reference agreement separately (Appendix J).

4 EXPERIMENTS

We evaluate whether MMG-defined molecular scales improve generation, capture progressively finer molecular information, and benefit from explicit coarse-to-fine prediction.

4.1 EXPERIMENTAL SETUP

Model training. We evaluate ScaleMol on QM9 (Ramakrishnan et al., 2014) and GEOM-Drugs (Axelrod & Gómez-Bombarelli, 2022), following the QM9 preprocessing and split of VoxMol (Pinheiro et al., 2023) and the GEOM-Drugs split of MiDi (Vignac et al., 2023), with explicit hydrogens retained. ScaleMol uses the same four-level 3D U-Net design on both datasets, with dataset-specific grid sizes and element channels. Full architecture, preprocessing, and training details are provided in Appendices B and C.

Baselines. *Atom-based methods* generate atom types and coordinates directly (EDM; Hoogeboom et al., 2022) or through equivariant latents (GeoLDM; Xu et al., 2023). *Field-based methods* recover atoms from generated fields: VoxMol samples occupancy grids (Pinheiro et al., 2023), while FuncMol and VecMol generate occupancy- and vector-field latent codes, respectively (Kirchmeyer et al., 2024; Hua et al., 2026). Published baseline scores come from Hua et al. (2026); controlled VoxMol variants are trained for the ablation study.

Evaluation protocol. We follow MiDi (Vignac et al., 2023) and evaluate atom and molecule stability, validity, uniqueness, atom- and bond-type total variation, and Wasserstein distances (W_1) for valency, bond lengths, and bond angles. For unconditional generation, we report mean and standard deviation over three independent sampling seeds with 10,000 attempted outputs per seed, using the 350-epoch checkpoint on QM9 and the 24-epoch checkpoint on GEOM-Drugs, with $J = 2000$ unless otherwise specified. For comparison with VecMol, $\text{ScaleMol}_{\text{ref}}$ applies its public postprocessing procedure to the same generated samples. Metrics and protocols specific to reference control and ablations are described in their respective sections and Appendix G.

4.2 UNCONDITIONAL MOLECULAR GENERATION

QM9 results. On QM9, ScaleMol improves molecule stability over the field-based VoxMol and FuncMol baselines by 6.0 and 6.1 percentage points, respectively, alongside better atom stability and valency matching (Table 1, Mol. stab., Atom stab., and Valency). Unlike existing field-based methods that generate molecular fields at a single scale, ScaleMol organizes prediction from coarse regional element content to finer atomic detail, providing coarse structural context for resolving local atomic arrangements. VecMol applies bond repair and hydrogen supplementation, which can affect stability (Hua et al., 2026). Applying the same public procedure, $\text{ScaleMol}_{\text{ref}}$ exceeds VecMol's molecule stability by 0.5 percentage points and matches its reported atom stability (Table 1). Without postprocessing, ScaleMol also exceeds VoxMol and FuncMol in uniqueness, while atom-based methods remain stronger in molecule stability and uniqueness on QM9.

GEOM-Drugs results. The advantage of atom-based methods on QM9 does not carry over to GEOM-Drugs. As molecules become larger and more structurally diverse, the molecule stability of EDM and GeoLDM drops substantially to 40.3% and 57.9%, whereas ScaleMol maintains 88.3% molecule stability and 99.5% atom stability (Table 2). This contrast is consistent with the benefit of organizing generation across scales. Rather than resolving the entire molecular structure directly

Table 1: Unconditional generation on QM9. Mean ± SD over three independent sampling seeds from one 350-epoch checkpoint (10,000 outputs each, J2000). Ref applies VecMol-style repair ($\rho = 1.5$) to the same samples.

	Validity & diversity				**Distribution matching**				
Model	Mol. stab. (%) ↑	Atom stab. (%) ↑	Valid (%) ↑	Unique (%) ↑	Valency W_1 ↓	Atom TV ↓	Bond TV ↓	Length W_1 ↓	Angle W_1 ↓
Data	98.7	99.8	98.9	99.9	0.001	0.003	0.000	0.000	0.120
Atom-based methods									
EDM	97.9	99.8	99.0	98.5	0.011	0.021	0.002	0.001	0.440
GeoLDM	97.5	99.9	100.0	98.0	0.005	0.017	0.003	0.007	0.435
Field-based methods									
VoxMol	89.3	99.2	98.7	92.1	0.023	0.029	0.009	0.003	1.960
FuncMol	89.2	99.0	100.0	92.8	0.021	0.012	0.006	0.005	1.560
VecMol	97.4	99.8	98.4	98.0	0.038	0.011	0.006	0.024	1.010
ScaleMol	95.3 (±0.2)	99.7 (±0.0)	98.5 (±0.0)	93.1 (±0.3)	0.011 (±0.000)	0.033 (±0.002)	0.008 (±0.000)	0.003 (±0.000)	2.659 (±0.116)
$\text{ScaleMol}_{\text{ref}}$	97.9 (±0.1)	99.8 (±0.0)	97.9 (±0.1)	91.9 (±0.2)	0.014 (±0.000)	0.029 (±0.002)	0.009 (±0.000)	0.003 (±0.000)	2.805 (±0.119)

Table 2: Unconditional generation on GEOM-Drugs. Mean ± SD over three independent sampling seeds from one 24-epoch checkpoint (10,000 outputs each, J2000). Ref applies VecMol-style repair ($\rho = 1.5$) to the same samples.

	Validity & diversity				**Distribution matching**				
Model	Mol. stab. (%) ↑	Atom stab. (%) ↑	Valid (%) ↑	Unique (%) ↑	Valency W_1 ↓	Atom TV ↓	Bond TV ↓	Length W_1 ↓	Angle W_1 ↓
Data	99.9	99.9	99.8	100.0	0.001	0.001	0.025	0.000	0.050
Atom-based methods									
EDM	40.3	97.8	87.8	99.9	0.285	0.212	0.048	0.002	6.420
GeoLDM	57.9	98.7	100.0	100.0	0.197	0.099	0.024	0.009	2.960
Field-based methods									
VoxMol	75.0	98.1	93.4	99.6	0.254	0.033	0.024	0.002	0.640
FuncMol	69.7	98.8	100.0	95.3	0.245	0.109	0.052	0.003	2.490
VecMol	77.4	99.0	82.3	81.4	0.102	0.034	0.269	0.104	8.980
ScaleMol	88.3 (±0.4)	99.5 (±0.0)	95.3 (±0.5)	97.7 (±0.4)	0.252 (±0.001)	0.111 (±0.005)	0.045 (±0.001)	0.003 (±0.001)	1.470 (±0.053)
$\text{ScaleMol}_{\text{ref}}$	90.4 (±0.3)	99.7 (±0.0)	92.8 (±0.7)	97.6 (±0.3)	0.264 (±0.001)	0.122 (±0.005)	0.043 (±0.001)	0.004 (±0.000)	2.458 (±0.021)

at the atomic level, ScaleMol first establishes regional element content at coarse scales and progressively resolves finer atomic arrangements. Without chemical postprocessing, ScaleMol exceeds EDM, GeoLDM, and VecMol in molecule stability by 48.0, 30.4, and 10.9 percentage points, respectively.

4.3 Controlled Generation at Different Molecular Scales

We next examine how control at different molecular scales affects the generated structures. We condition ScaleMol on QM9 reference MMGs at 4^3, 8^3, and 16^3 resolution through the resolution-prefix mechanism in Section 3.5, without additional training. We use 32 references with 32 sampling chains per reference and 2,500 total Walk steps. As shown in Figure 3(a,b), increasing the prefix resolution raises reference similarity from 0.087 to 0.894 and 0.986, while reducing graph diversity from 31.16 to 4.94 and 1.31. This reflects progressively stronger control across scales: coarse prefixes constrain regional element content while leaving finer arrangements flexible, whereas finer prefixes increasingly specify its spatial distribution. Figure 3(c,d) further shows that molecule stability increases from 57.03% to 91.89% and 99.22%, while the absolute atom-count deviation decreases

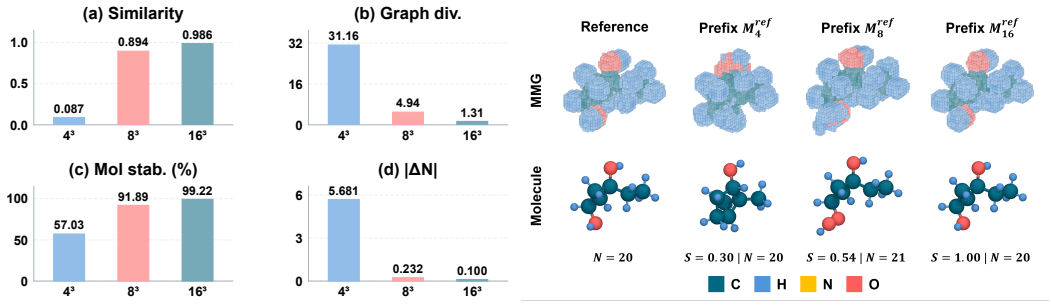

Figure 3: **Reference control across molecular scales at 2,500 total Walk steps.** All conditions share 32 references, 32 randomly initialized chains per reference, and the 250-epoch checkpoint. Left (a–d): reference retention and diversity; Graph div. counts distinct valid molecular graphs per 32 outputs, averaged over references. Right: finest MMGs and recovered molecules conditioned on reference fields M_s^{ref} at $s = 4, 8, 16$ cells per axis; S is Tanimoto similarity and N is atom count including hydrogens.

from 5.681 to 0.232 and 0.100. Chemical-property deviations follow the same trend (Appendix J), showing that finer-scale control retains more information from the reference molecule. Although the high diversity at 4^3 is accompanied by lower stability and larger atom-count deviations, the overall progression across scales provides a direct way to trade off reference preservation and structural variation within the same model, without additional training.

Table 3: **MMGs and architecture ablations on QM9.** Full ScaleMol is the shared reference for both groups. Means over three independent sampling-seed runs per model from the same 250-epoch checkpoint, each with 1,024 outputs and nonoverlapping roots; $W = J = 1000$.

Model	Mol. stab. (%) ↑	Atom stab. (%) ↑	Valid (%) ↑	Unique (%) ↑	Valency W_1 ↓	Atom TV ↓	Bond TV ↓	Length W_1 ↓	Angle W_1 ↓
Incremental additions: representation, supervision, and architecture									
VoxMol(Occupancy)	85.74	98.50	98.37	98.35	0.039	0.032	0.009	**0.002**	**1.950**
VoxMol(MMGs)	89.19	99.24	98.76	**98.78**	0.025	0.024	0.011	0.004	3.292
VoxMol+J(MMGs)	91.89	99.45	**98.93**	97.00	0.018	0.034	0.015	0.004	2.960
VoxMol+Split(MMGs)	93.13	99.54	98.27	94.97	0.015	0.044	0.011	0.004	3.289
Single-component removals from full ScaleMol									
ScaleMol w/o L	92.97	99.48	98.40	95.70	0.016	0.039	0.009	0.003	2.544
ScaleMol w/o Cond	93.49	99.55	98.37	96.16	0.015	0.036	0.010	0.003	3.299
ScaleMol w/o Corr	93.10	99.47	98.70	98.25	0.015	0.047	0.013	0.003	2.738
ScaleMol	**94.43**	**99.60**	98.63	94.45	**0.013**	**0.008**	**0.006**	**0.002**	2.694

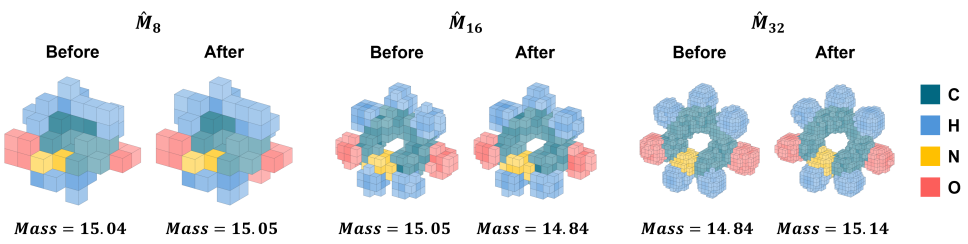

Figure 4: **Layerwise MMG predictions and mass correction.** Predictions $\widehat{M}_s$ at $s = 8, 16, 32$ cells per axis before and after mass correction in a single denoiser forward pass during unconditional QM9 generation.

4.4 Ablation Studies

MMGs and Architecture Ablations We progressively modify VoxMol to separate representation and prediction effects (Table 3). Replacing occupancy (VoxMol(Occupancy)) with MMGs (VoxMol(MMGs)) raises molecule stability from 85.74% to 89.19% and improves valency matching from 0.039 to 0.025. Supervising coarse projections of the final field with J_k (VoxMol+J(MMGs)) further raises stability to 91.89%. Adding parent-conditioned hierarchical readouts and intermediate supervision L_k (VoxMol+Split(MMGs)) reaches 93.13%, supporting molecular scales as explicit prediction targets beyond projection supervision alone. Adding gated encoder fusion (ScaleMol) reaches 94.43% and improves several distribution metrics. Single-component removals further test this design: removing L_k while retaining J_k (ScaleMol w/o L) reduces stability to 92.97%; zeroing parent-field condition inputs while retaining parent upsampling (ScaleMol w/o Cond) reduces it to 93.49%; removing mass correction (ScaleMol w/o Corr) reduces it to 93.10%. These results support both coarse context for finer predictions and the ability to correct inherited errors. Figure 4 illustrates how mass correction adjusts local field content without discarding the inherited coarse structure. Beyond the representation change, successive additions increase stability but reduce uniqueness, indicating a trade-off between structural consistency and diversity.

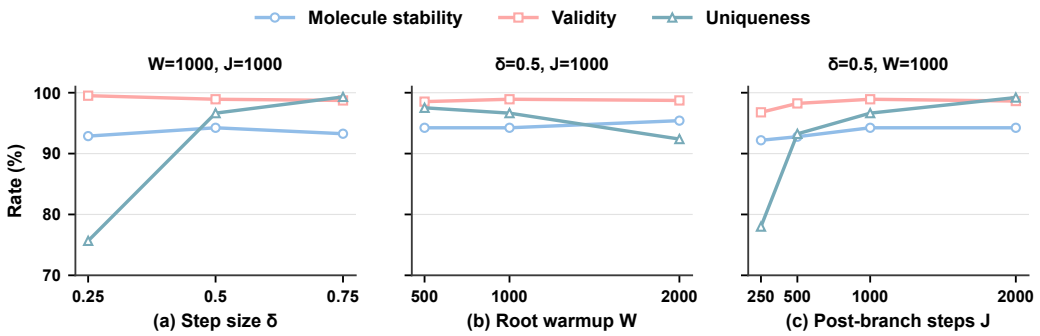

Figure 5: **Sampling hyperparameter ablations on QM9.** Each panel varies one parameter with the others fixed, using one 1,024-output run per setting with matched root seeds.

Sampling Hyperparameter Ablations. We examine the sensitivity of Walk–Jump sampling to step size δ, root warmup W, and post-branch length J (Figure 5). Longer post-branch sampling substantially increases uniqueness while preserving molecule stability, whereas longer root warmup reduces uniqueness. Increasing δ from 0.5 to 0.75 also improves uniqueness but lowers stability. These results reveal a stability–diversity trade-off in the sampling configuration. We use $\delta = 0.5$, $W = 1000$, and $J = 2000$ for the main experiments, with full results reported in Appendix K.

5 Conclusion

We introduced a definition of scale for small molecules based on the spatial resolution of molecular fields, and instantiated it through resolution-consistent Molecular Mass Grids (MMGs), which progressively resolve regional element content from coarse to fine. Building on these molecular scales, we developed ScaleMol to explicitly organize molecular generation from coarse structure to fine detail. Experiments on QM9 and GEOM-Drugs demonstrate improved molecular stability over existing field-based generators, with ScaleMol achieving 88.3% molecule stability on the larger and more diverse GEOM-Drugs dataset without chemical postprocessing. Controlled ablations further verify that both the MMG representation and coarse-to-fine prediction contribute to these improvements. Beyond unconditional generation, the same scale hierarchy enables flexible reference control at different resolutions without additional training. We hope that this notion of molecular scale can provide a useful basis for future work on controllable and multiscale molecular generation.

AI USE STATEMENT

We used generative AI tools to assist with software implementation, research methodology and conceptual refinement, reviewing and analyzing prior work, identifying relevant literature, and interpreting experimental results. We did not use these tools to formulate mathematical claims, drafting the manuscript or supply key steps in mathematical proofs. Synthetic dataset generation, dataset cleaning and reformatting, and qualitative or thematic data analysis were not applicable to their use in this work. We reviewed all AI-assisted work, including manually proofreading generated text and checking source documents to identify fabricated references and potential plagiarism. We take responsibility for the final content of this work, including all AI-assisted text, claims, and artifacts.

REFERENCES

Simon Axelrod and Rafael Gómez-Bombarelli. GEOM, energy-annotated molecular conformations for property prediction and molecular generation. *Scientific Data*, 2022. doi: 10.1038/s41597-022-01288-4.

Samuel Belkadi, Steve Hong, Marian Chen, Miruna Cretu, Charles Harris, and Pietro Liò. Diffusion-free graph generation with next-scale prediction. *arXiv preprint arXiv:2503.23612*, 2025. URL https://arxiv.org/abs/2503.23612.

Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, pp. 424–432, 2016. doi: 10.1007/978-3-319-46723-8_49.

Peter Ertl, Bernhard Rohde, and Paul Selzer. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *Journal of Medicinal Chemistry*, 43(20):3714–3717, 2000. doi: 10.1021/jm000942e.

Niklas W. A. Gebauer, Michael Gastegger, and Kristof T. Schütt. Symmetry-adapted generation of 3D point sets for the targeted discovery of molecules. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://papers.neurips.cc/paper_files/paper/2019/hash/a4d8e2a7e0d0c102339f97716d2fdfb6-Abstract.html.

Emiel Hoogeboom, Víctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3D. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 8867–8887, 2022. URL https://proceedings.mlr.press/v162/hoogeboom22a.html.

Yuchen Hua, Xingang Peng, Jianzhu Ma, and Muhan Zhang. VecMol: Vector-field representations for 3D molecule generation. *arXiv preprint arXiv:2603.12734*, 2026. URL https://arxiv.org/abs/2603.12734.

Moritz Ibing, Gregor Kobsik, and Leif Kobbelt. Octree transformer: Autoregressive 3D shape generation on hierarchically structured sequences. *arXiv preprint arXiv:2111.12480*, 2021. URL https://arxiv.org/abs/2111.12480.

Matthieu Kirchmeyer, Pedro O. Pinheiro, and Saeed Saremi. Score-based 3D molecule generation with neural fields. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/1486bbd863ba396398ff4227c5b3dccd-Abstract-Conference.html.

Rui Li, Artsemi Yushkevich, Mikhail Kudryashev, and Artur Yakimovich. Cryo-SWAN: the multiscale wavelet-decomposition-inspired autoencoder network for molecular density representation of molecular volumes. *arXiv preprint arXiv:2603.03342*, 2026. URL https://arxiv.org/abs/2603.03342.

Dong C. Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45:503–528, 1989. doi: 10.1007/BF01589116.

Noel M. O'Boyle, Michael Banck, Craig A. James, Chris Morley, Tim Vandermeersch, and Geoffrey R. Hutchison. Open babel: An open chemical toolbox. *Journal of Cheminformatics*, 3:33, 2011. doi: 10.1186/1758-2946-3-33.

Jiwoong Park and Yang Shen. Equivariant blurring diffusion for hierarchical molecular conformer generation. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/edcbd38b2509b57954d38de6cd9c05b3-Abstract-Conference.html.

Pedro O. Pinheiro, Joshua Rackers, Joseph Kleinhenz, Michael Maser, Omar Mahmood, Andrew Martin Watkins, Stephen Ra, Vishnu Sresht, and Saeed Saremi. 3D molecule generation by denoising voxel grids. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/da1131a86ac3c70e0b7cae89c3d4df22-Abstract-Conference.html.

Bo Qiang, Yuxuan Song, Minkai Xu, Jingjing Gong, Bowen Gao, Hao Zhou, Wei-Ying Ma, and Yanyan Lan. Coarse-to-fine: a hierarchical diffusion model for molecule generation in 3D. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 28277–28299. PMLR, 2023. URL https://proceedings.mlr.press/v202/qiang23a.html.

Yanru Qu, Cheng-Yen Hsieh, Zaixiang Zheng, Ge Liu, and Quanquan Gu. Protein autoregressive modeling via multiscale structure generation. *arXiv preprint arXiv:2602.04883*, 2026. URL https://arxiv.org/abs/2602.04883.

Matthew Ragoza, Tomohide Masuda, and David Ryan Koes. Generating 3D molecules conditional on receptor binding sites with deep generative models. *Chemical Science*, 13:2701–2713, 2022. doi: 10.1039/D1SC05976A. URL https://pubs.rsc.org/en/content/articlelanding/2022/sc/d1sc05976a.

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. doi: 10.1038/sdata.2014.22.

RDKit contributors. RDKit: Open-source cheminformatics, n.d. URL https://www.rdkit.org.

Sucheng Ren, Qihang Yu, Ju He, Xiaohui Shen, Alan Yuille, and Liang-Chieh Chen. FlowAR: Scale-wise autoregressive image generation meets flow matching. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 51489–51502. PMLR, 2025. URL https://proceedings.mlr.press/v267/ren25e.html.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. doi: 10.1021/ci100050t.

Saeed Saremi and Aapo Hyvärinen. Neural empirical bayes. *Journal of Machine Learning Research*, 20(181):1–23, 2019. URL https://jmlr.org/papers/v20/19-216.html.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/9a24e284b187f662681440ba15c416fb-Abstract-Conference.html.

Clément Vignac, Nagham Osman, Laura Toni, and Pascal Frossard. MiDi: Mixed graph and 3D denoising diffusion for molecule generation. *arXiv preprint arXiv:2302.09048*, 2023. URL https://arxiv.org/abs/2302.09048.

Scott A. Wildman and Gordon M. Crippen. Prediction of physicochemical parameters by atomic contributions. *Journal of Chemical Information and Computer Sciences*, 39(5):868–873, 1999. doi: 10.1021/ci9903071.

Minkai Xu, Alexander S. Powers, Ron O. Dror, Stefano Ermon, and Jure Leskovec. Geometric latent diffusion models for 3D molecule generation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 38592–38610. PMLR, 2023. URL https://proceedings.mlr.press/v202/xu23n.html.

A PROPERTIES OF MMGS

A.1 CELL INTEGRATION AND FINITE DOMAINS

For a cubic cell $V = \prod_{j=1}^{3}[\ell_j, u_j]$, separability of the isotropic Gaussian gives

$$M_c(V) = \sum_{i:c_i=c} \prod_{j=1}^{3} \left[\Phi\left(\frac{u_j - r_{ij}}{\sigma_f}\right) - \Phi\left(\frac{\ell_j - r_{ij}}{\sigma_f}\right) \right], \tag{14}$$

where Φ is the standard normal cumulative distribution function.

For cells partitioning a finite box Ω,

$$\sum_{V \subset \Omega} M_c(V) = \int_{\Omega} \rho_c(x)\,\mathrm{d}x = N_c - \int_{\mathbb{R}^3 \setminus \Omega} \rho_c(x)\,\mathrm{d}x. \tag{15}$$

Gaussian tails can extend beyond the box even when all atom centers lie inside it. We do not renormalize truncated atomic contributions, preserving their interpretation as integrals of a common continuous field. The difference between in-box mass and the full atom count depends on truncation, not grid resolution. Appendix E examines the associated atom-recovery errors.

A.2 AGGREGATION, LIFTING, AND SPATIAL DETAIL

For adjacent nested grids, let D sum each group of eight child cells and let U distribute a parent value equally among those children, as in Equations 3–4. Additivity of integration yields $DM_{k+1} = M_k$. Successive aggregation also agrees with direct aggregation to the target grid. These identities require aligned grids covering the same domain with the same σ_f and hold up to floating-point error in implementation.

Flattening a single channel and using the Euclidean inner product gives

$$DD^\top = 8I, \qquad U = \tfrac{1}{8}D^\top, \qquad DU = I. \tag{16}$$

Consequently, $P = UD$ satisfies $P^\top = P$ and $P^2 = P$: it is the orthogonal projection onto fields constant within each parent block. With $Q = I - P$, any fine-grid field z decomposes as

$$z = UDz + Qz, \qquad DQz = 0, \qquad \langle UDz, Qz\rangle = 0. \tag{17}$$

In these unweighted grid coordinates, the two components satisfy $\|z\|_2^2 = \|UDz\|_2^2 + \|Qz\|_2^2$.

Every field satisfying $Dz = m$ can be written as

$$z = Um + r, \qquad r \in \ker D. \tag{18}$$

The detail r may be signed even though an MMG of a molecule is nonnegative. This is a decomposition of field values, not of the network's hidden features.

A.3 OBSERVATION RESOLUTION AND ATOMIC BANDWIDTH

Increasing cell width at fixed σ_f aggregates the same field over larger regions; increasing σ_f instead changes the atomic density. Our hierarchy varies cell width while keeping σ_f fixed.

A.4 NOISE STATISTICS OF SHARED MULTIRESOLUTION VIEWS

For explicit aggregation of the noisy field in Equation 5, write

$$Y_L = X_L + \sigma_n \epsilon, \qquad X_L = M_L/a, \qquad \epsilon \sim \mathcal{N}(0, I). \tag{19}$$

Let $A_k = D_{k \leftarrow L}$ aggregate the finest grid to level k, with n_k finest cells per coarse cell. Then

$$Y_k = A_k Y_L = X_k + \sigma_n A_k \epsilon, \qquad X_k = A_k X_L, \qquad A_k A_k^\top = n_k I. \tag{20}$$

Within any one coarse view, the noise remains isotropic Gaussian, but its variance is $n_k\sigma_n^2$, not σ_n^2. Aggregation from 32^3 to 16^3, 8^3, and 4^3 multiplies the variance by 8, 64, and 512, respectively. These factors apply before any additional feature standardization.

Views at different resolutions are correlated:

$$\mathrm{Cov}(Y_k - X_k, Y_j - X_j) = \sigma_n^2 A_k A_j^\top. \tag{21}$$

B Model Implementation

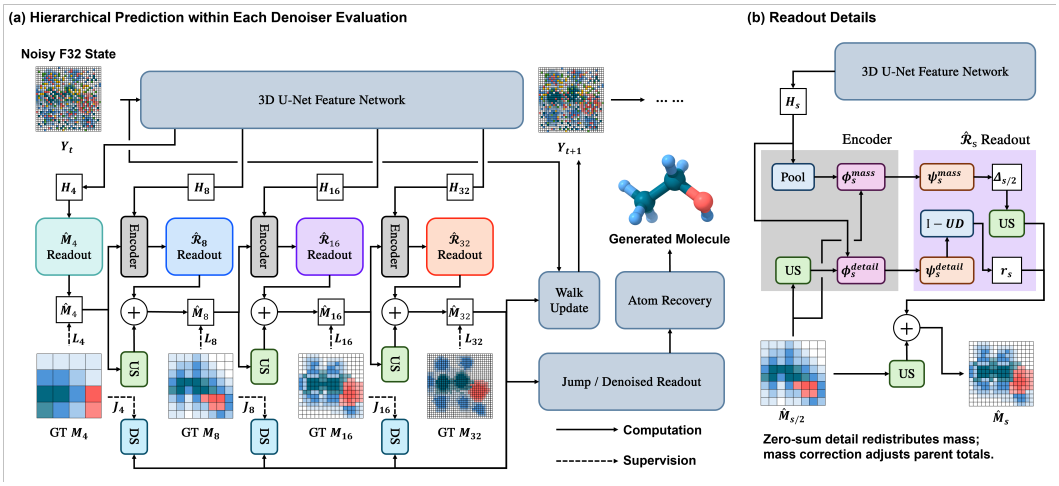

Figure 6: **Detailed architecture and hierarchical readout.** (a) Shared features support hierarchical predictions within each denoiser evaluation. L_s supervises each prediction, while J_s supervises aggregations of the finest prediction. (b) Parent-conditioned branches predict zero-sum detail and a parent-mass correction, combined with the lifted parent as in Equation 7. US and DS denote mass-preserving lifting and sum aggregation; s denotes cells per axis.

B.1 Encoder Feature Fusion

The encoder supplements each native downsampling operation with

$$c^+ = c + W_{\text{out}}\left[\text{sigmoid}(G([c, \bar{f}])) \odot P(\bar{f})\right], \qquad \bar{f} = \text{AvgPool}_2(f), \tag{22}$$

where f and c are features immediately before and after downsampling. P is a 1^3 convolution followed by SiLU; G and W_{out} are 1^3 convolutions. The gate acts per location and channel, and W_{out} is initialized to zero. This branch fuses hidden features, not predicted molecular fields.

B.2 Hierarchical Readout Implementation

Each readout branch encodes the normalized parent field and concatenates its encoding with backbone features (Figure 6b):

$$\begin{aligned} C_k^{\text{detail}} &= \phi_k^{\text{detail}}\left(H_k, U_{k-1}\widehat{M}_{k-1}\right), \\ C_k^{\text{mass}} &= \phi_k^{\text{mass}}\left(\text{AvgPool}_2(H_k), \widehat{M}_{k-1}\right). \end{aligned} \tag{23}$$

The heads predict $\widetilde{r}_k = \psi_k^{\text{detail}}(C_k^{\text{detail}})$ and $\Delta_{k-1} = \psi_k^{\text{mass}}(C_k^{\text{mass}})$, with fixed rescaling to mass units. Parameters are independent across branches and levels; outputs are combined by Equation 7.

Figure 4 illustrates the local contribution of the mass branch while holding the parent prediction and detail branch fixed. All levels are predicted without a reference prefix.

B.3 Field Normalization and Supervision Weights

We normalize the finest field by the mass of an isolated atom inside a cell centered on that atom:

$$a = \left[2\Phi\left(\frac{h_L}{2\sigma_f}\right) - 1\right]^3. \tag{24}$$

The centered single-atom response is thus one. This scalar is shared across elements and resolutions and commutes with aggregation and lifting; it does not normalize individual molecules or their total mass.

Separately, the readout uses frozen training-set means and standard deviations for each scale and element channel of M_k/a. The detail branch standardizes the lifted parent at the child scale, while the mass branch standardizes it at the parent scale. Their outputs are rescaled by the corresponding standard deviations; the coarsest head also restores its mean. These internal operations are included in the readout functions in Equation 23 and do not change the physical mass semantics or aggregation rules.

Let $n_k = (h_k/h_L)^3$ count the finest cells in a level-k cell. The discrepancy used in Section 3.2 is

$$\ell_k(A, B) = \frac{1}{C|\mathcal{G}_{h_k}|} \sum_{c=1}^{C} \sum_{V \in \mathcal{G}_{h_k}} w_{k,c} \left(\frac{A_c(V) - B_c(V)}{a n_k} \right)^2 . \tag{25}$$

Division by n_k converts mass errors to finest-cell-equivalent units. Fixed $w_{k,c}$ weight elements; λ_k and μ_k weight intermediate and final-projection losses.

For optimization, the implementation multiplies the combined mean loss by the number of finest-grid scalar entries in the batch (`clean.numel()`). Distributed reduction and accumulation are compensated to recover the summed objective over the global batch of 128. This common scaling preserves the relative loss weights but must be retained to reproduce the optimization settings.

B.4 Atom Recovery Implementation

Algorithm 1 recovers atoms from the finest predicted MMG without a prescribed composition. We use $\tau = 0.20$, cell width $h = 0.25$ Å, and output caps $N_{\max} = 32$ for QM9 and 512 for GEOM-Drugs. The cell-center map $x(i)$ specifies the grid's coordinate frame. Peak detection and interpolation use the same bounded field; archived fields are converted to FP32 before decoding. No collision deletion, coordinate optimization, or chemical repair is included in this readout.

Algorithm 1 MMG recovery: plateau-aware detection and local log-quadratic interpolation

Require: Finest MMG $\widehat{M}_L$, normalization a, cell centers $x(i)$, cell width h, threshold τ, output cap $N_{\max}$
Ensure: Atom set $\widehat{\mathcal{M}}$ with inferred positions and element types

$b \leftarrow 1 - \exp[-(\log 2)\max(\widehat{M}_L, 0)/a]$; $\mathcal{P} \leftarrow \varnothing$
for each element channel c **do**
  $\mathcal{C}_c \leftarrow \{i : b_c(i) \geq \tau,\ b_c(i) = \max_{j \in \mathcal{N}_{27}(i)} b_c(j)\}$ ▷ Outside-grid values are $-\infty$
  for each 26-connected component S of $\mathcal{C}_c$ **do**
    $\bar{i} \leftarrow |S|^{-1} \sum_{i \in S} i$
    $i^\star \leftarrow \arg\min_{i \in S} \|i - \bar{i}\|_2^2$ ▷ First grid-order index breaks ties
    Append $(c, i^\star, b_c(i^\star))$ to $\mathcal{P}$
  end for
end for
Keep at most $N_{\max}$ candidates in $\mathcal{P}$, ranked by decreasing peak value
$\widehat{\mathcal{M}} \leftarrow \varnothing$
for each retained candidate (c, i, s) **do**
  $r \leftarrow x(i)$
  for axis $d \in \{1, 2, 3\}$ with both neighbors $i - e_d$ and $i + e_d$ inside the grid **do**
    $\ell_t \leftarrow \log \max(b_c(i + te_d), \epsilon_{\text{tiny}})$ for $t \in \{-1, 0, 1\}$
    $q \leftarrow \ell_{-1} - 2\ell_0 + \ell_1$
    if $q < -10^{-8}$ **then**
      $r_d \leftarrow r_d + h\ \text{clip}\left(\frac{\ell_{-1} - \ell_1}{2q}, -\frac{1}{2}, \frac{1}{2}\right)$
    end if
  end for
  Append (r, c) to $\widehat{\mathcal{M}}$
end for
return $\widehat{\mathcal{M}}$ ▷ No candidates returns an empty atom set

Here ϵ_{tiny} is the smallest positive normal value of the decoding dtype. Missing axis neighbors or nonconcave/near-flat log curvature leave that coordinate at the cell center. Bond inference and the optional chemical-repair evaluation are separate procedures (Appendix G).

C EXPERIMENTAL CONFIGURATION

Table 4: QM9 main-experiment configuration. Parameter count excludes the EMA copy.

Setting	Value
Training set / exposure epoch	97,734 molecules
Domain / finest cell width	8 Å cube / 0.25 Å
Readout grids / Gaussian bandwidth	$4^3, 8^3, 16^3, 32^3$ / $\sigma_f = 0.35$ Å
U-Net base channels / channel multipliers	32 / $(1, 2, 2, 4)$
Blocks per encoder level / attention	2 / $8^3, 4^3$
Parameters	112,361,919
Optimizer / learning rate	AdamW / constant 2×10^{-5}
Adam coefficients / weight decay	$(0.99, 0.999)$ / 0.01
Global batch / training budget	128 / 350 sample-exposure epochs
Noise standard deviation / EMA decay	0.9 / 0.999
Loss weights	$w_{k,c} = 1$; $\lambda_k = \mu_k = (0.025, 0.05, 0.10)$, coarse to fine
Augmentation	Random rotations; translations up to ± 0.5 Å per axis
Training / sampling precision	FP32 / FP32 EMA; TF32 off
Walk: δ / warmup / post-branch steps	0.5 / 1,000 / 2,000

Table 5: GEOM-Drugs main-experiment configuration. Parameter count excludes the EMA copy.

Setting	Value
Training set / exposure epoch	1,172,577 conformers from 243,480 molecules
Domain / finest cell width	16 Å cube / 0.25 Å
Readout grids / Gaussian bandwidth	$8^3, 16^3, 32^3, 64^3$ / $\sigma_f = 0.35$ Å
U-Net base channels / channel multipliers	32 / $(1, 2, 2, 4)$
Blocks per encoder level / attention	2 / $16^3, 8^3$
Parameters	112,507,223
Optimizer / learning rate	AdamW / constant 2×10^{-5}
Adam coefficients / weight decay	$(0.99, 0.999)$ / 0.01
Global batch / training budget	128 / 24 sample-exposure epochs
Noise standard deviation / EMA decay	0.9 / 0.999
Loss weights	$w_{k,c} = 1$; $\lambda_k = \mu_k = (0.025, 0.05, 0.10)$, coarse to fine
Augmentation	Random rotations; translations up to ± 0.5 Å per axis
Training / sampling precision	BF16 mixed / FP32 EMA; TF32 off
Walk: δ / warmup / post-branch steps	0.5 / 1,000 / 2,000

Table 6: QM9 ablation ladder and parameter counts. J_k and L_k denote additional projection and intermediate-head losses. MMG recovery follows Algorithm 1. Parameter counts exclude EMA copies and buffers.

Model	MMGs	J_k	L_k	Parent conditioning	Mass correction	Encoder fusion	Atom Recovery	Parameters
VoxMol(Occupancy)	—	—	—	—	—	—	L-BFGS	111,515,237
VoxMol(MMGs)	✓	—	—	—	—	—	MMG recovery	111,515,237
VoxMol+J(MMGs)	✓	✓	—	—	—	—	MMG recovery	111,515,237
VoxMol+Split(MMGs)	✓	✓	✓	✓	✓	—	MMG recovery	112,275,231
ScaleMol w/o L	✓	✓	—	✓	✓	✓	MMG recovery	112,361,919
ScaleMol w/o Cond	✓	✓	✓	—	✓	✓	MMG recovery	112,361,919
ScaleMol w/o Corr	✓	✓	✓	✓	—	✓	MMG recovery	111,988,498
ScaleMol	✓	✓	✓	✓	✓	✓	MMG recovery	112,361,919

D COMPUTATIONAL COST

Tables 7 and 8 summarize training and sampling costs. On QM9, ScaleMol incurs an 11–13% training-time overhead and an 11–12% sampling-time overhead relative to the VoxMol occupancy baseline under matched base U-Net configurations, batch sizes, and hardware settings, with similar training peak memory and modest additional sampling peak memory. Both methods share the cost of dense three-dimensional field processing. Generating compact latent representations, as in FuncMol (Kirchmeyer et al., 2024), is therefore a promising efficiency direction for both; extending this approach to ScaleMol would require retaining MMG scale semantics. For atom recovery, the non-iterative MMG readout is approximately four times faster than occupancy recovery with L-BFGS on matched ground-truth molecules (Table 9).

Table 7: QM9 training cost on L40 GPUs, global batch 128. Means (standard deviations) over three runs of 200 updates after 30 warmup updates. Peak memory is the maximum allocated CUDA memory reported by PyTorch per GPU. Overhead is additional update time relative to VoxMol at the same GPU count. Validation and checkpoint I/O are excluded.

	Samples/s ↑		Seconds/update ↓		Peak memory (GiB) ↓		
GPUs	VoxMol (Occupancy)	ScaleMol	VoxMol (Occupancy)	ScaleMol	VoxMol (Occupancy)	ScaleMol	Overhead
2	110.28 (0.48)	99.02 (0.29)	1.161 (0.005)	1.293 (0.004)	21.05	20.63	+11.37%
4	237.00 (0.64)	209.65 (1.10)	0.540 (0.001)	0.611 (0.003)	11.81	11.60	+13.05%

Table 8: QM9 sampling cost on one L40S, with 1,000 root-warmup steps and the indicated post-branch Walk steps. Times are means (standard deviations) over three runs, including the final Jump but excluding atom recovery and evaluation. Peak memory is the maximum allocated CUDA memory reported by PyTorch per GPU. Overhead is ScaleMol's additional time per output relative to VoxMol.

		Seconds/output ↓		Peak memory (GiB) ↓		
Children/ root	Walk steps	VoxMol (Occupancy)	ScaleMol	VoxMol (Occupancy)	ScaleMol	Overhead
32	500	3.051 (0.020)	3.391 (0.011)	1.64	1.67	+11.17%
32	1,000	5.220 (0.032)	5.816 (0.013)	1.64	1.67	+11.42%
32	2,000	9.554 (0.022)	10.661 (0.014)	1.64	1.67	+11.59%
64	500	2.527 (0.007)	2.832 (0.005)	2.84	2.90	+12.07%
64	1,000	4.614 (0.011)	5.176 (0.004)	2.84	2.90	+12.18%
64	2,000	8.781 (0.011)	9.864 (0.008)	2.84	2.90	+12.34%

Table 9: QM9 GT-field atom-recovery cost on one L40S. Both pipelines use the same 256 unfiltered test molecules, batch 16, four CPU threads, and FP32. Parentheses give standard deviations over three repeats. Timing includes internal CPU/GPU transfers but excludes field construction, input upload, file I/O, and chemical evaluation. Peak memory is PyTorch's maximum allocated CUDA memory per GPU.

Recovery pipeline	ms/molecule ↓	Molecules/s ↑	Peak memory (MiB) ↓
Occupancy + L-BFGS	69.917 (0.050)	14.30 (0.01)	190.5
MMG recovery (Algorithm 1)	17.598 (0.015)	56.82 (0.05)	74.6

E GROUND-TRUTH FIELD RECONSTRUCTION FIDELITY

We isolate atom-readout errors by reconstructing molecules from fields rendered from ground-truth coordinates, without a learned generator.

Table 10: **QM9 paired GT-field recovery.** Each boundary condition uses the same 10,000 test molecules. Stability and distribution metrics follow MiDi; distributions use the same molecules' GT coordinates and OB-inferred graphs as reference. Position uses common count-correct subsets of 10,000/10,000/9,999 molecules for interior/near-boundary/outside placements.

Field	Refine	Count (%) ↑	Position (Å) ↓	Atom stab. (%) ↑	Mol. stab. (%) ↑	Bond TV (10^{-5}) ↓	Length W_1 (10^{-4} Å) ↓	Angle W_1 (°) ↓
Interior								
GT	–	100.00	0	99.9	98.8	0.000	0.000	0.04013
Union	No	100.00	3.16×10^{-5}	99.9	98.8	0.000	0.070	0.04082
Union	Yes	100.00	8.78×10^{-7}	99.9	98.8	0.000	0.027	0.04014
MMGs	No	100.00	2.25×10^{-3}	99.9	98.8	0.763	0.969	0.04917
MMGs	Yes	100.00	5.07×10^{-9}	99.9	98.8	0.509	0.038	0.04013
Near boundary								
GT	–	100.00	0	99.9	98.8	0.000	0.000	0.04013
Union	No	100.00	2.21×10^{-3}	99.9	98.4	5.344	3.018	0.06602
Union	Yes	100.00	8.71×10^{-7}	99.9	98.8	0.127	0.030	0.04014
MMGs	No	100.00	4.36×10^{-3}	99.9	98.4	5.344	3.198	0.07253
MMGs	Yes	100.00	5.89×10^{-9}	99.9	98.8	0.382	0.047	0.04013
Outside								
GT	–	100.00	0	99.9	98.8	0.000	0.000	0.04013
Union	No	100.00	3.16×10^{-2}	99.9	97.9	35.756	22.642	0.72480
Union	Yes	100.00	9.55×10^{-7}	99.9	98.8	0.000	0.030	0.04014
MMGs	No	99.99	3.36×10^{-2}	99.8	97.9	35.616	23.014	0.72991
MMGs	Yes	99.99	1.00×10^{-7}	99.9	98.7	0.243	0.041	0.04034

GT's nonzero angle distance reflects differences between MiDi's reference-statistics and evaluation implementations; it is not a reconstruction error.

Setup. We select 10,000 distinct QM9 test molecules and reuse them across three translated boundary conditions, retaining the stored orientation and explicit hydrogens. Eligibility requires that the axis-aligned molecular extent fit inside the box with at least 0.5 Å clearance on every face; 11,862 of 13,055 test molecules satisfy this condition. We sample without replacement from this eligible set.

Interior poses are centered by their bounding boxes. For each molecule, a uniformly sampled box face determines the translation direction for both remaining poses; the minimum signed atom-to-box distance is sampled uniformly from $[0.125, 0.375]$ Å for near-boundary poses and $[-0.375, -0.125]$ Å for outside poses. These paired stress tests isolate placement effects conditional on geometric eligibility; they are not estimates for natural boundary groups or the full dataset. We use an 8 Å box, a 32^3 grid, FP32 rendering, VoxMol occupancy radius 0.5 Å, and MMG bandwidth $\sigma_f = 0.35$ Å. These are representation-specific settings, not a matched-bandwidth comparison.

Within each representation, the paired readouts share peak detection, element assignments, and subgrid initialization. For MMGs, detection and local interpolation use the bounded field

$$b(m) = 1 - \exp\left[-(\log 2)\frac{\max(m, 0)}{a}\right], \tag{26}$$

where a is the centered single-atom cell mass defined in Equation 24. We use a detection threshold of 0.20 on $b(m)$. This transform is used for localization, not as the refinement target.

Refinement fits each representation with its own renderer: Gaussian-union occupancy for union fields and cell-integrated Gaussians for MMGs. For detected types $\mathbf{c}$ and subgrid coordinates $\mathbf{R}^{(0)}$, the mass-field objective is

$$\mathbf{R}^\star = \arg\min_{\mathbf{R}\in\mathcal{B}_{\delta_r}(\mathbf{R}^{(0)})} \left\| \mathcal{R}_L(\mathbf{R}, \mathbf{c}) - \widehat{M}_L \right\|_2^2, \tag{27}$$

where $\mathcal{R}_L$ renders cell-integrated Gaussians and $\mathcal{B}_{\delta_r}$ bounds each coordinate's displacement from initialization. Optimization keeps atom types and counts fixed and uses no reference coordinates,

bond constraints, or force-field energies. For these recovery experiments, we use two L-BFGS (Liu & Nocedal, 1989) runs of at most 200 iterations, restarting the optimizer between runs, with $\delta_r = 0.5$ Å per axis.

We use the MiDi evaluation pipeline (Vignac et al., 2023), with Open Babel (O'Boyle et al., 2011) inferring bonds for both recovered and GT molecules instead of using GT bonds. Position error is evaluated only on molecules with correct counts of every element across all readouts; we report this common subset's size. All other metrics include the full sample set, including failures.

F Walk–Jump Sampling Implementation

Algorithm 2 specifies the unconditional sampler using the normalized finest-grid state Y and score $s_\theta(Y) = (\widehat{X}_\theta(Y) - Y)/\sigma_n^2$ from Section 3.3. Each score evaluation uses the complete hierarchical predictor, but only its finest-resolution output enters the update.

Continuous-time interpretation. We evolve the noisy state and auxiliary velocity using underdamped Langevin dynamics:

$$\begin{aligned} \mathrm{d}Y_t &= V_t \,\mathrm{d}t, \\ \mathrm{d}V_t &= s_\theta(Y_t)\,\mathrm{d}t - \gamma V_t \,\mathrm{d}t + \sqrt{2\gamma}\,\mathrm{d}B_t, \end{aligned} \tag{28}$$

where $\gamma > 0$ is the continuous-time friction rate and B_t is standard Brownian motion. The noise level stays fixed; t indexes Langevin evolution rather than noise annealing.

Discrete walk. We use the VoxMol-style update implemented in our sampler. With $\alpha = \exp(-1)$, one step is

$$\begin{aligned} Y_{1/2} &= Y + \tfrac{\delta}{2} V, \\ g &= s_\theta(Y_{1/2}), \\ V^+ &= \alpha\big(V + \tfrac{\delta}{2} g\big) + \tfrac{\delta}{2} g + \sqrt{1-\alpha^2}\,\xi, \qquad \xi \sim \mathcal{N}(0, I), \\ Y^+ &= Y_{1/2} + \tfrac{\delta}{2} V^+. \end{aligned} \tag{29}$$

Each step reuses one score evaluation, with no Metropolis correction. We fix the per-step damping factor at $\alpha = e^{-1}$, independently of δ. Under the Ornstein–Uhlenbeck substep interpretation, $\alpha = e^{-\gamma\delta}$, so this corresponds to $\gamma\delta = 1$ rather than a fixed continuous-time friction rate ($\gamma = 2$ at the default $\delta = 0.5$).

Algorithm 2 Unconditional walk–jump sampling with shared warmup

Require: EMA predictor $\widehat{X}_\theta$, normalization a, noise level σ_n;
number of roots R, children per root B, warmup W,
observation steps $0 < K_1 < \cdots < K_m$, walk parameter δ
Ensure: Field sets $\mathcal{S}_1, \ldots, \mathcal{S}_m$, each containing RB predictions

$\mathcal{S}_j \leftarrow \varnothing$ for $j = 1, \ldots, m$
for $r = 1, \ldots, R$ **do**
  Draw $Q \sim \mathcal{U}([0,1]^d)$ and $\epsilon \sim \mathcal{N}(0, I)$ independently
  $Y \leftarrow Q + \sigma_n \epsilon$; $V \leftarrow 0$
  for $t = 1, \ldots, W$ **do**
    $(Y, V) \leftarrow$ WalkStep(Y, V) using Equation 29
  end for
  $(Y_b, V_b) \leftarrow (\mathrm{copy}(Y), \mathrm{copy}(V))$ for $b = 1, \ldots, B$
  for $t = 1, \ldots, K_m$ **do**
    Advance every (Y_b, V_b) by WalkStep, with independent ξ_b
    if $t = K_j$ for some j **then**
      $\widehat{M}_{j,r,b} \leftarrow a\widehat{X}_\theta(Y_b)$ for $b = 1, \ldots, B$ ▷ Jump
      Append all $\widehat{M}_{j,r,b}$ to $\mathcal{S}_j$; leave (Y_b, V_b) unchanged
    end if
  end for
end for
return $\mathcal{S}_1, \ldots, \mathcal{S}_m$

WJ can produce different molecules by applying Jump at different Walk steps along the same ongoing trajectory, without restarting the chain or feeding the readouts back into Walk. Figure 7 illustrates this behavior for an unconditional QM9 chain.

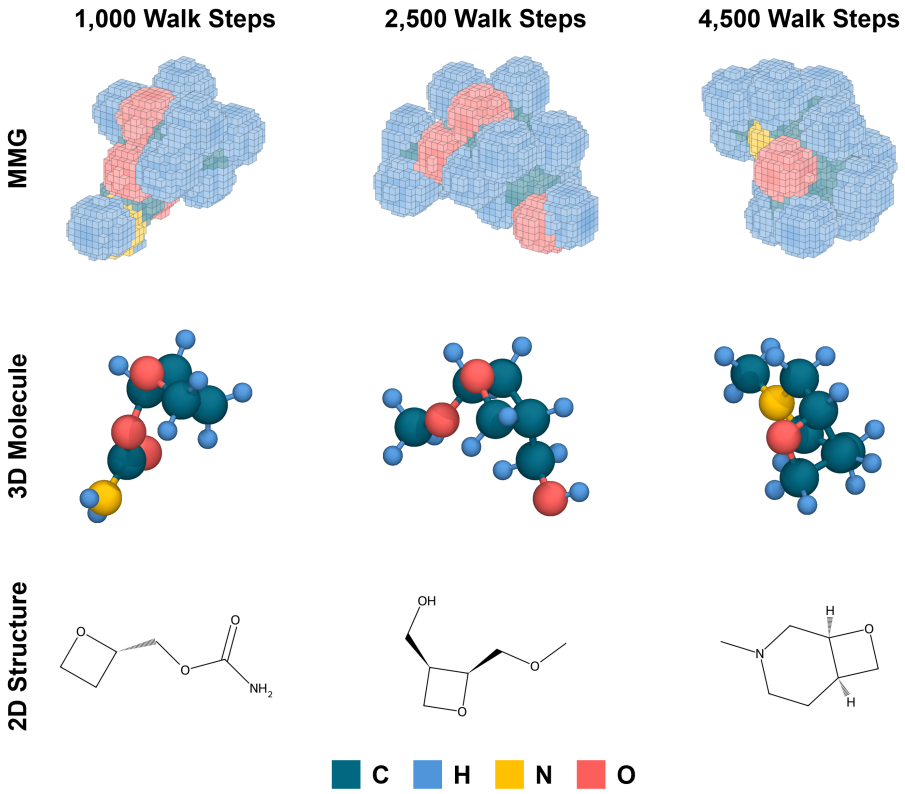

Figure 7: **Qualitative Walk trajectory in WJ sampling.** A single unconditional QM9 chain is read out after 1,000, 2,500, and 4,500 total Walk steps.

G Evaluation Metrics

We use MiDi's unconditional generation metrics (Vignac et al., 2023). Open Babel (O'Boyle et al., 2011) infers bonds and formal charges from coordinate outputs, retaining explicit hydrogens. For our unconditional generation tables, distribution metrics use native test-set graphs as references. The paired GT-field recovery experiment instead uses the same molecules' GT coordinates and OB-inferred graphs to isolate reconstruction-induced differences (Appendix E).

Published scores can reflect method-specific postprocessing. For ScaleMol$_{\text{ref}}$, we use the official VecMol repository's bond-repair and hydrogen-supplementation functions, with bond-distance ratio $\rho = 1.5$ and at most ten repair iterations (Hua et al., 2026). These operations can substantially affect molecule stability.

Atom and molecule stability. Atom stability is the percentage of generated atoms whose graph valence satisfies MiDi's allowed-valence rules for their element and formal charge. Molecule stability is the percentage of outputs in which every atom is stable. Both are computed on the full graph and measure valence consistency rather than energetic stability.

Validity. Validity is the percentage of generated outputs whose largest connected fragment passes RDKit (RDKit contributors, n.d.) sanitization. A valid output therefore need not be a single connected molecule.

Uniqueness. Uniqueness is the number of distinct canonical SMILES divided by the number of valid outputs, using the same fragments as for validity.

Atom- and bond-type TV. These metrics compare normalized type histograms over generated and test samples using MiDi's sum of absolute frequency differences. Atom-type histograms count all atoms; bond-type histograms count unordered atom pairs, including the no-bond category.

Valency W_1. This metric compares valency distributions separately for each element and averages their Wasserstein distances using test-set element frequencies.

Bond-length W_1. This metric compares bond-length distributions separately for each bond type and weights their Wasserstein distances by the corresponding test-set bond-type frequencies. Distances are measured in Å using the original generated coordinates.

Bond-angle W_1. This metric compares distributions of angles between pairs of bonds sharing an atom, grouped by the central element. MiDi weights the per-element Wasserstein distances by test-set element frequencies and the fraction of atoms with valency at least two, then normalizes the weights. Angles are measured in degrees.

The following metrics evaluate prefix-conditioned molecular generation.

Field agreement. We render recovered atoms as MMGs in the reference coordinate frame and measure their L_1 distance to the reference field at the supplied prefix resolution, divided by the reference's total grid mass.

Reference similarity. Reference similarity is the Tanimoto similarity of Morgan fingerprints (Rogers & Hahn, 2010) (radius two, 2,048 bits, without chirality).

Chemical-property changes. We report mean absolute generated-minus-reference differences in RDKit-computed Crippen logP (Wildman & Crippen, 1999), topological polar surface area (TPSA) (Ertl et al., 2000), and hydrogen-bond donor and acceptor counts.

Atom-count change. We report the mean absolute difference $|\Delta N|$ between generated and reference total atom counts, including explicit hydrogens and averaging over all attempted outputs.

Within-reference diversity. We report the number of distinct valid canonical SMILES among 32 attempts per reference, averaged over references (Graph div.).

Reference similarity and property changes use valid largest fragments from the same bond-inference pipeline for reference and generated molecules. We average these metrics over valid pairs and graph diversity over references, reporting each Jump separately and retaining validity as coverage.

H Cross-Scale Aggregation Consistency

We compare directly rendered coarse fields with sum- or mean-pooled fine fields on 1,024 interior GT molecules per dataset. QM9 uses an 8 Å box and GEOM-Drugs a 16 Å box, with VoxMol occupancy radius 0.5 Å and MMG bandwidth $\sigma_f = 0.35$ Å. We evaluate analytic Gaussian cell integrals and VoxMol center-sampled occupancies in FP64 without kernel cutoffs. For fine and coarse fields F_h, F_H, Table 11 reports the mean relative error $\|A(F_h) - F_H\|_1 / \|F_H\|_1$, where A is the indicated pooling operator.

Summed MMGs match direct coarse integration, whereas pooled occupancies differ from coarse-grid point queries. The discrepancy also occurs for an isolated atom (13.5% under $32 \to 16$ mean pooling), independently of atomic overlap. This tests aggregation consistency, not generation quality.

Table 11: Mean relative L_1 aggregation error on 1,024 GT molecules per dataset. Grid sizes denote cells per axis.

Dataset	Grid	MMGs: sum Rel. $L_1 \downarrow$	Occupancy: mean Rel. $L_1 \downarrow$	Occupancy: sum Rel. $L_1 \downarrow$
QM9	$32 \to 16$	1.07×10^{-16}	0.127	6.999
	$32 \to 8$	1.30×10^{-16}	0.522	65.411
	$32 \to 4$	1.99×10^{-16}	2.373	1229.828
GEOM-Drugs	$64 \to 32$	1.08×10^{-16}	0.127	7.000
	$64 \to 16$	1.32×10^{-16}	0.517	64.235
	$64 \to 8$	2.02×10^{-16}	1.701	812.536

I Training Dynamics

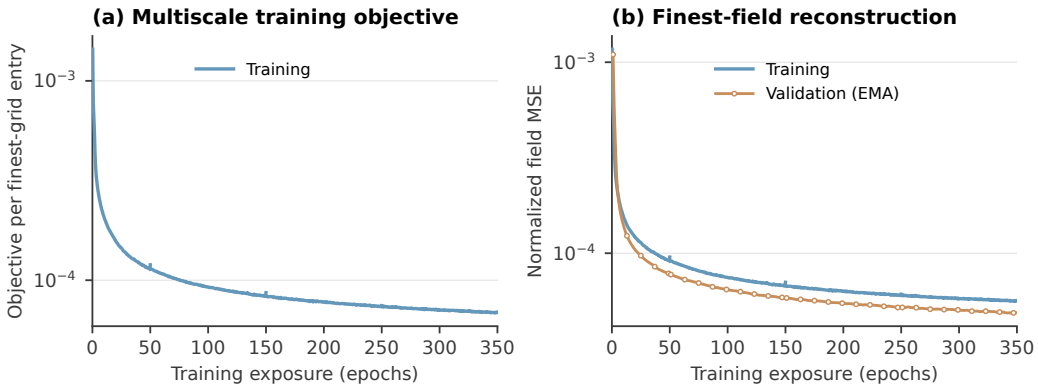

Figure 8: QM9 training dynamics: (a) multiscale training objective and (b) finest-field reconstruction error.

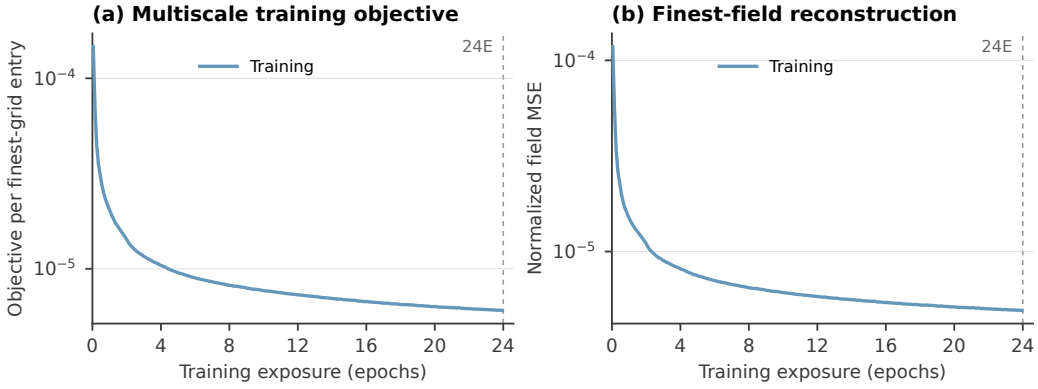

Figure 9: GEOM-Drugs training dynamics: (a) multiscale training objective and (b) finest-field reconstruction error.

J Prefix-Controlled Generation

Protocol. We uniformly select 32 distinct molecules from the held-out QM9 test set with a fixed seed, without screening by reconstruction quality. Reference coordinates are centered by their all-atom mean and retain their stored orientation. Each reference initializes 32 independent chains as

$Y_0 \sim \mathcal{U}(0,1) + \sigma_n \mathcal{N}(0, I)$, with zero velocity. Random streams are paired across prefix resolutions. We use the frozen 250-epoch checkpoint, $\delta = 0.5$, $\sigma_n = 0.9$, and per-step damping $\alpha = e^{-1}$, with EMA weights in FP32. The prefix is active throughout Walk and Jump. Each reported time contains 1,024 attempted outputs; total Walk times include the initial 500-step warmup. Successive Jump observations reuse the same chains and are not independent repeats. Atom recovery is unchanged, without VecMol chemical repair.

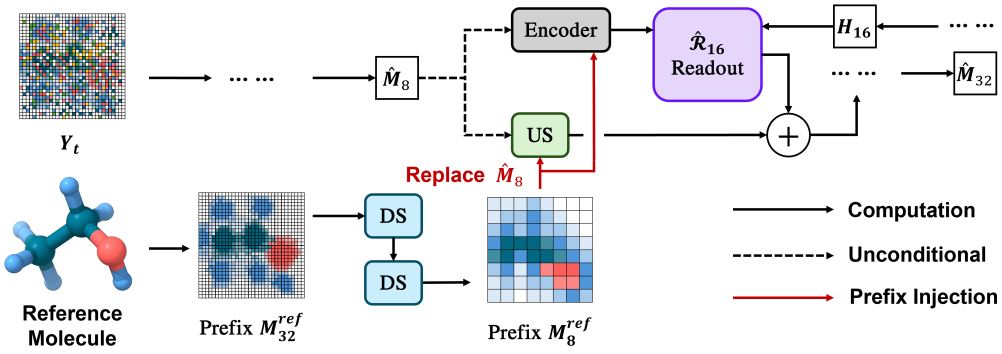

Figure 10: **Prefix construction and injection.** A reference molecule defines a finest-resolution MMG; repeated sum-downsampling (DS) constructs coarser MMGs. The selected prefix replaces the predicted parent at both the parent-field encoder and upsampling (US) inputs, illustrated here at resolution 8.

Metrics and temporal behavior. Tables 14 and 15 report eight Walk times using the metrics in Appendix G; stability and validity include all attempts. Similarity and diversity retain their ordering across scales, while stability does not improve monotonically with Walk time.

Field preservation versus atom recovery. Exact grid-level preservation in Equation 13 does not guarantee preservation after atom recovery. Table 12 quantifies this gap at the supplied resolution.

Protecting the supplied observations. We compare unprotected and protected prefix sampling using the same 250-epoch checkpoint, eight fixed references, and eight paired chains per reference at 2,500 Walk steps. Both variants inject the same reference parents; only the latter projects subsequent corrections as in Equation 13. Without this protection, finer readout heads can overwrite the supplied observations (Table 12). Protection substantially improves reference agreement, atom-count retention, and stability at 8^3 and 16^3. At 4^3, field agreement improves but stability declines: preserving coarse mass alone does not ensure a recoverable atomic structure.

Table 12: Paired prefix-protection diagnostic with 64 outputs per condition. Raw and recovered errors are relative L_1 at the supplied scale; $|\Delta N|$ is the mean absolute atom-count error.

Prefix	Prefix protection	Raw error	Recovered error	$\lvert\Delta N\rvert$	Mol. stab. (%)	Similarity
4^3	Unprotected	1.115	1.093	4.828	90.63	0.066
	Protected	$< 10^{-6}$	0.810	6.078	50.00	0.089
8^3	Unprotected	1.304	1.195	7.375	65.63	0.072
	Protected	$< 10^{-6}$	0.196	0.250	96.88	0.879
16^3	Unprotected	0.687	0.491	0.922	73.44	0.515
	Protected	$< 10^{-6}$	0.064	0.125	100.00	0.967

Sensitivity to reference orientation. Because ScaleMol uses a Cartesian-grid backbone rather than an explicitly rotation-equivariant architecture, we test whether prefix-conditioned generation depends strongly on the global orientation of the reference. Using 32 boundary-safe references and

a protected 8^3 prefix, we compare the stored orientation with four random global rotations per reference.[1] Across both 1,000 and 2,500 Walk steps, molecule stability, reference similarity, diversity, atom-count error, and recovered-field error remain similar, with no systematic degradation under rotation (Table 13). This suggests substantial robustness to global reference orientation, consistent with training using random-rotation augmentation, although ScaleMol does not enforce exact rotational equivariance.

Table 13: Rotation sensitivity with a protected 8^3 prefix on boundary-safe references. Original and rotated values are means over 32 references. Δ is rotated minus original, with percentile 95% CIs from 2,000 bootstrap resamples of reference-level paired differences; stability differences are in percentage points. Field error is relative L_1 after atom recovery. For the rotated arm, Graph div. pools 8 outputs from each of four random orientations per reference.

Metric	Original	Rotated	Δ (rot $-$ orig)	95% CI for Δ
1,000 Walk steps				
Mol. stability (%)	87.40	88.48	$+1.07$	$[-2.15,\ 4.49]$
Similarity	0.822	0.845	$+0.023$	$[-0.016,\ 0.067]$
Graph div.	7.31	7.22	-0.09	$[-1.50,\ 1.13]$
$\lvert\Delta N\rvert$	0.279	0.229	-0.051	$[-0.139,\ 0.021]$
Field error	0.262	0.251	-0.012	$[-0.038,\ 0.010]$
2,500 Walk steps				
Mol. stability (%)	87.70	87.21	-0.49	$[-4.10,\ 3.42]$
Similarity	0.817	0.825	$+0.008$	$[-0.030,\ 0.048]$
Graph div.	7.69	7.72	$+0.03$	$[-1.28,\ 1.25]$
$\lvert\Delta N\rvert$	0.279	0.250	-0.029	$[-0.082,\ 0.023]$
Field error	0.265	0.256	-0.009	$[-0.034,\ 0.010]$

[1]References are selected by geometry alone to retain at least $1.05\,\text{Å}$ ($3\sigma_f$) of atom-to-boundary clearance under any rotation. The control uses 32 chains per reference; both arms contain 1,024 attempts per time, with paired chain seeds and unchanged sampling and readout.

Table 14: Protected-prefix quality and reference agreement at every measured Walk time.

Walk	Prefix	AS (%)	Mol. stab. (%)	Valid (%)	Similarity	Graph div.
1000	4^3	94.85	59.77	97.27	0.085	31.00
1000	8^3	99.29	92.48	99.90	0.892	5.38
1000	16^3	99.96	99.61	100.00	0.993	1.22
1500	4^3	94.63	58.89	97.85	0.087	31.19
1500	8^3	99.06	90.43	99.80	0.878	5.84
1500	16^3	99.96	99.41	100.00	0.988	1.22
2000	4^3	94.83	59.28	97.85	0.088	31.22
2000	8^3	99.25	92.19	99.71	0.898	5.00
2000	16^3	99.97	99.51	100.00	0.992	1.16
2500	4^3	94.32	57.03	97.46	0.087	31.16
2500	8^3	99.23	91.89	99.41	0.894	4.94
2500	16^3	99.91	99.22	100.00	0.986	1.31
3000	4^3	94.91	60.45	96.88	0.086	30.94
3000	8^3	99.26	92.58	99.41	0.899	5.16
3000	16^3	99.92	99.22	100.00	0.991	1.25
3500	4^3	95.02	60.64	98.14	0.084	31.34
3500	8^3	99.21	91.31	99.32	0.897	4.94
3500	16^3	99.91	99.02	100.00	0.981	1.34
4000	4^3	94.48	57.81	98.24	0.086	31.31
4000	8^3	99.02	91.60	99.71	0.887	4.94
4000	16^3	99.87	98.63	100.00	0.978	1.34
4500	4^3	94.68	58.69	98.05	0.085	31.31
4500	8^3	99.25	92.29	99.71	0.891	4.84
4500	16^3	99.91	99.02	100.00	0.985	1.38

Table 15: Protected-prefix atom-count and chemical-property errors at every measured Walk time. All changes are mean absolute differences; TPSA is in Å^2.

Walk	Prefix	$\lvert\Delta N\rvert$	$\lvert\Delta\mathrm{logP}\rvert$	$\lvert\Delta\mathrm{TPSA}\rvert$	$\lvert\Delta\mathrm{HBD}\rvert$	$\lvert\Delta\mathrm{HBA}\rvert$
1000	4^3	5.629	0.861	20.033	0.795	1.059
1000	8^3	0.230	0.104	1.622	0.106	0.096
1000	16^3	0.095	0.006	0.054	0.007	0.003
1500	4^3	5.758	0.860	19.803	0.809	1.028
1500	8^3	0.243	0.117	1.973	0.110	0.100
1500	16^3	0.097	0.010	0.054	0.012	0.002
2000	4^3	5.693	0.872	20.033	0.789	1.018
2000	8^3	0.220	0.090	1.361	0.083	0.078
2000	16^3	0.094	0.006	0.019	0.010	0.000
2500	4^3	5.681	0.890	19.566	0.811	0.987
2500	8^3	0.232	0.106	1.565	0.095	0.087
2500	16^3	0.100	0.012	0.133	0.016	0.003
3000	4^3	5.616	0.874	19.939	0.824	1.003
3000	8^3	0.237	0.093	1.535	0.078	0.083
3000	16^3	0.097	0.009	0.085	0.009	0.004
3500	4^3	5.643	0.877	19.914	0.793	1.065
3500	8^3	0.219	0.099	1.724	0.108	0.096
3500	16^3	0.097	0.015	0.087	0.018	0.004
4000	4^3	5.730	0.883	20.846	0.795	1.054
4000	8^3	0.237	0.104	1.751	0.098	0.099
4000	16^3	0.103	0.019	0.212	0.023	0.010
4500	4^3	5.776	0.872	19.870	0.825	1.042
4500	8^3	0.203	0.102	1.566	0.094	0.092
4500	16^3	0.100	0.011	0.093	0.009	0.005

K SAMPLING HYPERPARAMETER SENSITIVITY

Table 16 reports the full sweep of δ, W, and J with the checkpoint, atom recovery, and evaluation protocol fixed. The step-size sweep holds per-step damping $\alpha = e^{-1}$ fixed, not the continuous-time friction rate (Appendix F). Each run produces 1,024 outputs per Jump from 16 roots with 64 children each.

Table 16: Full QM9 sampling sweep at 250 exposure epochs. Each setting uses one run of 1,024 outputs (16 roots with 64 children each), with matched root seeds across settings and the same trajectories read out at each Jump.

δ	W	J	Mol. stab. (%) ↑	Atom stab. (%) ↑	Valid (%) ↑	Unique (%) ↑	Valency W_1 ↓	Atom TV ↓	Bond TV ↓	Length W_1 ↓	Angle W_1 ↓
0.25	1000	250	86.9	99.2	96.6	42.7	0.021	0.089	0.023	0.003	3.217
0.25	1000	500	91.9	99.5	98.4	57.8	0.015	0.089	0.024	0.004	3.016
0.25	1000	1000	92.9	99.6	99.5	75.7	0.013	0.080	0.024	0.004	2.979
0.25	1000	2000	94.8	99.7	99.0	88.1	0.010	0.066	0.023	0.002	2.977
0.5	1000	250	92.2	99.4	96.8	78.0	0.017	0.044	0.007	0.001	3.124
0.5	1000	500	92.8	99.5	98.2	93.2	0.016	0.023	0.006	0.002	2.938
0.5	1000	1000	94.2	99.6	98.9	96.6	0.013	0.003	0.005	0.002	2.578
0.5	1000	2000	94.2	99.6	98.6	99.2	0.011	0.014	0.010	0.002	2.756
0.75	1000	250	93.7	99.5	98.7	91.9	0.014	0.031	0.010	0.002	1.942
0.75	1000	500	93.8	99.5	98.5	95.7	0.014	0.037	0.011	0.002	2.031
0.75	1000	1000	93.3	99.5	98.7	99.3	0.016	0.039	0.011	0.002	1.973
0.75	1000	2000	92.4	99.4	98.6	99.4	0.018	0.028	0.008	0.002	1.709
0.5	500	250	93.4	99.5	99.3	77.1	0.013	0.017	0.006	0.003	3.347
0.5	500	500	91.8	99.4	98.7	91.8	0.017	0.014	0.007	0.003	2.993
0.5	500	1000	94.2	99.6	98.5	97.5	0.012	0.015	0.010	0.003	2.862
0.5	500	2000	94.2	99.6	99.1	98.5	0.013	0.023	0.012	0.003	2.608
0.5	2000	250	94.3	99.6	98.4	69.1	0.013	0.046	0.010	0.004	3.179
0.5	2000	500	95.0	99.6	98.4	85.0	0.012	0.033	0.008	0.003	2.933
0.5	2000	1000	95.4	99.7	98.7	92.4	0.010	0.021	0.009	0.003	2.885
0.5	2000	2000	96.0	99.7	98.9	97.4	0.010	0.013	0.011	0.003	2.902

L GENERATED MOLECULES

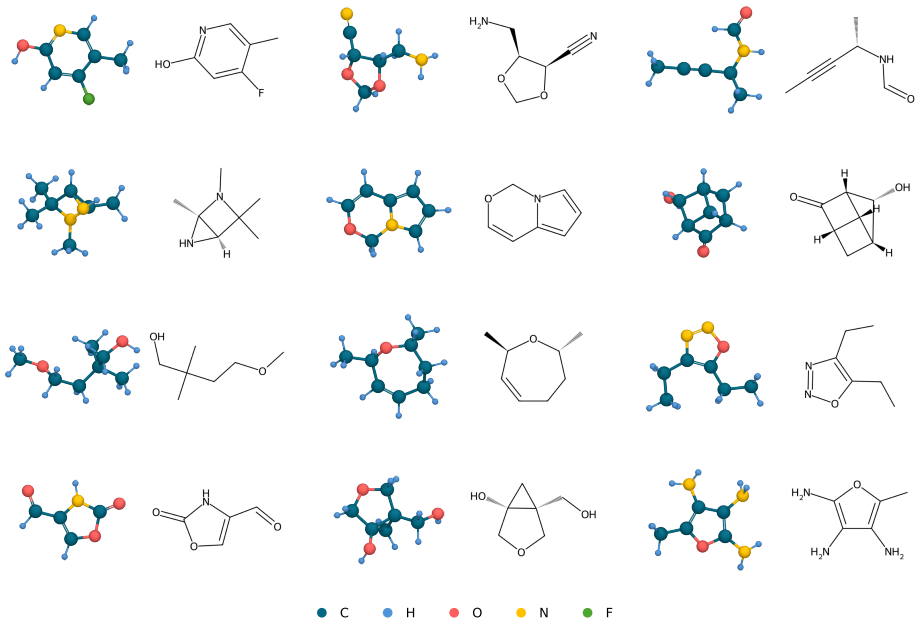

Figure 11: **Molecules generated by ScaleMol trained on QM9.**

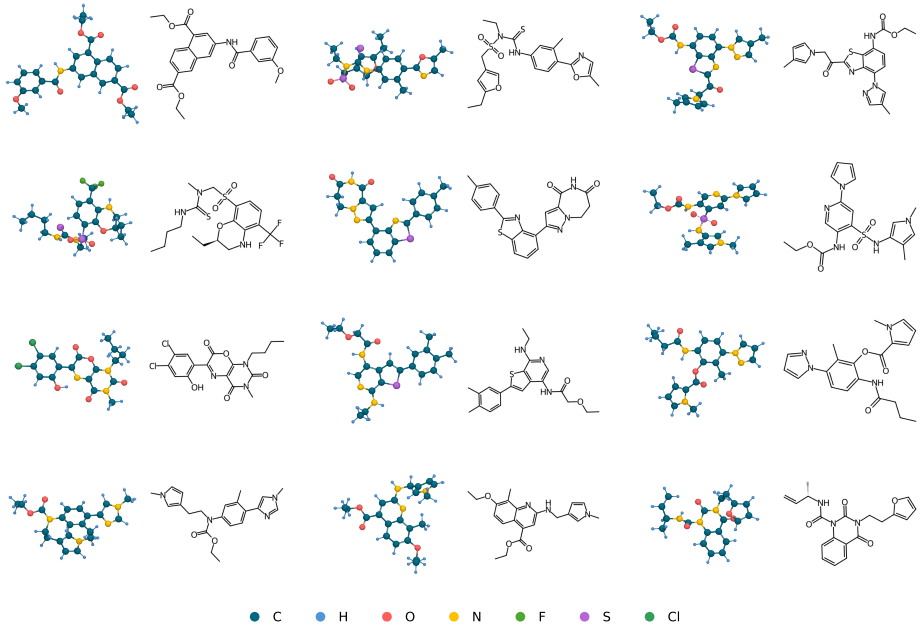

Figure 12: **Molecules generated by ScaleMol trained on GEOM-Drugs.**