AUTOREGRESSIVE LATENT DIFFUSION FOR 3D MOLECULE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Three-dimensional (3D) molecule generation has been dominated by diffusion models, which achieve strong generation quality but typically require molecular size to be specified (or predicted) separately before generation. This can be limiting for fragment-based molecule generation, central to drug discovery, where the size of the generated structure is itself part of the design problem. Autoregressive models determine size during generation and naturally support partial-structure conditioning, but balancing unconditional and fragment-conditioned generation remains challenging. We introduce KRONOS, a latent autoregressive diffusion framework that generates molecules in the latent space of a Unified AutoEncoder (UAE), jointly modeling molecular graph topology and geometry, while retaining the flexibility of autoregressive generation. We further introduce a mixed training strategy inspired by the Fill-in-the-Middle (FIM) paradigm, enabling a single left-to-right autoregressive model to support both unconditional and fragment-conditioned generation. Experiments on QM9 and GEOM-Drugs demonstrate strong unconditional generation performance and competitive fragment-conditioned generation.

1 INTRODUCTION

Generative modeling of three-dimensional (3D) molecules is a central problem in machine learning for drug discovery and molecular design. Recent progress has largely been driven by diffusion (Hoogeboom et al., 2022b; Huang et al., 2024; Luo et al., 2025; Joshi et al., 2025) and flow matching (Irwin et al., 2024), which achieve excellent generation quality but typically require molecular size to be specified or predicted separately before generation. This is particularly relevant for molecule generation from partial structures, where the size of the generated structure is itself part of the design problem.

Autoregressive models provide a complementary paradigm by generating molecules sequentially (Gebauer et al., 2019; Luo & Ji, 2022; Daigavane et al., 2024), determining molecular size during generation while naturally supporting conditioning on partial molecular structures. This flexibility is particularly valuable in medicinal chemistry, where new molecules are often designed by elaborating known scaffolds or functional fragments (Lim et al., 2020; Erlanson et al., 2016). Recent autoregressive models (Cheng et al., 2025; Rose et al., 2025) have narrowed the performance gap to diffusion-based approaches, but balancing unconditional generation with partial-structure conditioning remains challenging.

Unified latent representations (Joshi et al., 2025; Xu et al., 2023; Park & Walsh, 2025) provide a natural interface for generative modeling by jointly encoding discrete structural information and continuous geometry. Existing latent molecular generators, however, have primarily relied on diffusion-based generation.

In this work, we introduce **KRONOS**, a latent autoregressive diffusion framework that operates directly in the latent space of a Unified AutoEncoder (UAE) (Luo et al., 2025). KRONOS models latent molecular tokens autoregressively, while parameterizing each conditional distribution with a diffusion model (Li et al., 2024). An explicit stop-prediction mechanism enables variable-length generation. To support fragment-conditioned generation, we further introduce a mixed training strategy inspired by the Fill-in-the-Middle (FIM) paradigm (Bavarian et al., 2022) enabling both unconditional and fragment-conditioned generation within a single left-to-right architecture.

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.

Experiments on QM9 and GEOM-Drugs datasets demonstrate that KRONOS achieves leading unconditional performance among autoregressive methods, while remaining competitive with diffusion-based approaches. Furthermore, the results show that enabling fragment-conditioned generation has negligible impact on unconditional generation quality. A comprehensive discussion of related work is provided in Appendix A.1. Our contributions are summarized below:

- We introduce KRONOS, a latent autoregressive diffusion framework for 3D molecule generation.
- We introduce a fragment-conditioned training strategy that supports both unconditional and fragment-conditioned generation within a single autoregressive model.
- We demonstrate strong unconditional generation performance, together with competitive fragment-conditioned generation.

2 PRELIMINARIES

3D molecule representation A 3D molecule consists of discrete chemical attributes (atom types and bond connectivity) together with continuous geometry (3D coordinates). This hybrid nature makes joint generative modeling challenging, as it requires simultaneously handling discrete and continuous variables while respecting permutation and SE(3) symmetries. Following previous work (Luo et al., 2025; Huang et al., 2024), we represent a 3D molecule as a tuple $\mathbf{G} = (\mathbf{H}, \mathbf{A}, \mathbf{X})$, where $\mathbf{H} \in \mathbb{R}^{|\mathcal{V}| \times d_1}$ encodes atomic features, $\mathbf{A} \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}| \times d_2}$ is the bond adjacency matrix, and $\mathbf{X} \in \mathbb{R}^{|\mathcal{V}| \times 3}$ specifies 3D atomic coordinates. $\mathcal{V}$ represents the set of atoms of $\mathbf{G}$, while d_1 and d_2 represent the dimension of atomic and bond features, respectively.

Unified Autoencoder (UAE). To obtain a unified representation, we adopt the UAE framework (Luo et al., 2025), which encodes each atom into a continuous latent vector that jointly captures its chemical identity, bonding environment, and spatial position. The UAE encoder $\mathcal{E}_{\text{UAE}}$ employs a graph relational transformer (Diao & Loynd, 2023), which maps a molecule $\mathbf{G}$ into a latent sequence $\mathcal{E}_{\text{UAE}}(\mathbf{G}) = \mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$, where $\mathbf{z}_i \in \mathbb{R}^d$. A transformer decoder $\mathcal{D}_{\text{UAE}}$ decodes latents back into the reconstructed molecule $\hat{\mathbf{G}} = \mathcal{D}_{\text{UAE}}(\mathbf{Z})$. Notably, both the encoder $\mathcal{E}_{\text{UAE}}$ and decoder $\mathcal{D}_{\text{UAE}}$ are permutation equivariant with respect to atom ordering. Furthermore, the UAE is trained with random SE(3) transformations of the input coordinates, encouraging the encoder–decoder pair to learn SE(3)-equivariant reconstruction.

Autoregressive Diffusion for Continuous Tokens. To model conditional distributions over continuous tokens, we adopt the Diffusion Loss introduced by Li et al. (2024), which provides a unified framework for continuous-valued generation in both autoregressive (e.g., Vaswani et al. (2017)) and masked modeling settings (e.g., Devlin et al. (2019); Chang et al. (2022)). This formulation replaces discrete token prediction with a denoising objective, enabling flexible modeling of continuous tokens without discretization. In particular, each conditional distribution $p(\mathbf{z} \mid \mathbf{c})$ is parameterized via a diffusion-based denoising process:

$$\mathcal{L}(\mathbf{c}, \mathbf{z}) = \mathbb{E}_{t,\varepsilon}\left[\left\|\varepsilon - \varepsilon_\theta(\mathbf{z}_t \mid t, \mathbf{c})\right\|^2\right], \tag{1}$$

where $\mathbf{z}_t$ is a noise-perturbed version of $\mathbf{z}$ and $\mathbf{c}$ represents the conditioning context produced by the autoregressive model. This formulation has been successfully applied to 3D molecule generation for modeling continuous 3D geometry (Cheng et al., 2025).

3 METHODS

Prior autoregressive approaches to 3D molecule generation operate in data space (Cheng et al., 2025; Rose et al., 2025). This requires the model to learn over heterogeneous discrete and continuous generation targets. We instead perform autoregressive generation directly in the latent space of a UAE (Luo et al., 2025), where each latent token jointly encodes atom identity, bond topology, and 3D geometry. KRONOS combines three key components: (i) autoregressive modeling over latent 3D molecular tokens (Section 3.1), (ii) a mixed autoregressive training strategy supporting

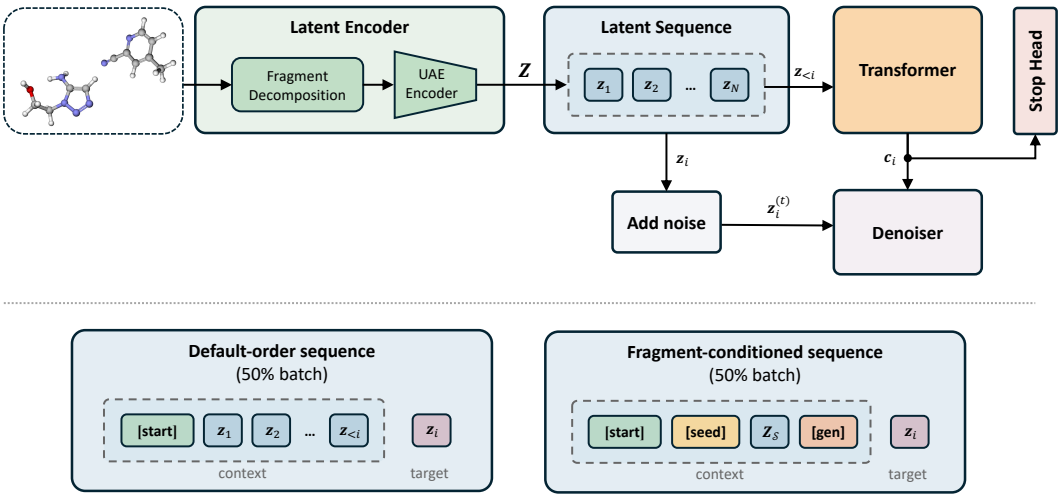

Figure 1: Visual overview of KRONOS. Molecules are decomposed into fragments and encoded into latent molecular sequences by the UAE encoder. During training, each example follows either the default-order or fragment-conditioned sequence construction. A causal transformer encodes the resulting autoregressive context, which conditions both a latent diffusion denoiser for next-token prediction and a stop head for sequence termination. This unified training strategy supports both unconditional and fragment-conditioned generation without architectural modifications.

both unconditional and fragment-conditioned generation (Section 3.2) and (iii) a Diffusion Loss formulation (Li et al., 2024) with a stop classifier for sequence termination (Section 3.3). We provide a visual overview of KRONOS in Figure 1.

3.1 Autoregressive Latent Generation

Our goal is to learn a generative model over latent molecular sequences $p_\theta(\mathbf{z}_{1:n})$, where each latent token corresponds to an atom. Samples from this distribution are decoded into molecular structures using the UAE decoder $\mathcal{D}_{\text{UAE}}$. We model the joint distribution over latent tokens $p_\theta(\mathbf{z}_{1:n})$ using a left-to-right autoregressive factorisation:

$$p_\theta(\mathbf{z}_{1:n}) = \prod_{i=1}^{n} p_\theta(\mathbf{z}_i \mid \mathbf{z}_{<i}), \tag{2}$$

where generation is initialized with a special [START] token. At each step i, a causal transformer encodes the preceding tokens into a context representation

$$\mathbf{c}_i = \text{Transformer}([\text{START}], \mathbf{z}_1, \ldots, \mathbf{z}_{i-1}), \tag{3}$$

which conditions a diffusion model used to parameterize $p_\theta(\mathbf{z}_i \mid \mathbf{z}_{<i})$ (described in Section 3.3). As the latent space does not admit a natural end-of-sequence token, we model sequence termination separately via a stop classifier $p_\theta([\text{STOP}] \mid \mathbf{c}_i)$ (described in Section 3.3). During training, the autoregressive factorization in Eq. 2 corresponds to the *default-order* sequence, which follows the token ordering inherited from the input molecules, and is used for unconditional (*de novo*) molecule generation. This ordering likely traces back to how the 3D structures were initialized from SMILES (Cheng et al., 2025).

3.2 Fragment-Conditioned Generation

Previous autoregressive models for 3D molecule generation employ a fixed atom ordering (Cheng et al., 2025). While effective for unconditional generation, training with a fixed atom ordering limits generalization to arbitrary subsets of molecular atoms. To address this, we introduce a mixed autoregressive training strategy inspired by the Fill-in-the-Middle (FIM) paradigm (Bavarian et al., 2022), originally proposed to unlock document-infilling capabilities within left-to-right transformer

models and later extended to protein language models (Lee & Yu, 2023). During training, each molecule is transformed into either a default-order sequence (Section 3.1) or a *fragment-conditioned* sequence with probabilities $1-p_f$ and p_f, respectively, where p_f denotes the fragment-conditioning probability. We use $p_f = 0.5$ throughout the main experiments and study the effect of varying p_f in Appendix A.11. The fragment-conditioned sequence construction can be viewed as defining autoregressive factorizations at the fragment level rather than at the atom level. We introduce this perspective below.

Permutation space By the chain rule, a joint distribution over n latent variables can be factorized according to any permutation $\pi \in S_n$:

$$p_\theta(\mathbf{z}_{1:n}) = \prod_{i=1}^{n} p_\theta\big(\mathbf{z}_{\pi(i)} \mid \mathbf{z}_{\pi(<i)}\big)\,. \tag{4}$$

The permutation equivariance of the UAE encoder–decoder pair ensures that different orderings of the same molecule correspond to consistently permuted latent sequences, which decode to the same molecular structure. This allows us to construct alternative autoregressive factorizations by choosing different molecular token orderings. Order-agnostic autoregressive (OA-AR) models (Uria et al., 2016; Yang et al., 2019) treat π as a latent variable and maximise the expected log-likelihood over S_n. Because the full expectation is intractable, they optimise the variational lower bound

$$\log p_\theta(\mathbf{z}) \;\geq\; \mathbb{E}_{\pi\sim S_n}\left[\sum_{i=1}^{n} \log p_\theta\big(\mathbf{z}_{\pi(i)} \mid \mathbf{z}_{\pi(<i)}\big)\right], \tag{5}$$

which can be estimated with Monte-Carlo sampling of permutations and positions, as originally proposed by Uria et al. (2016) and later reformulated to enable efficient, single-step training (Hoogeboom et al., 2022a). Rather than marginalizing over S_n as in OA-AR, we instead define autoregressive factorisations based on *molecular fragments*. Specifically, each molecule is decomposed into chemically meaningful fragments. A random subset of these fragments is then selected as the conditioning context, and a single deterministic ordering is assigned to the resulting conditioning and generation sequences. We examine the importance of preserving structured atom orderings in Appendix A.6.

Fragment decomposition To construct fragment-conditioned training sequences, each molecule is partitioned into K fragments $\{\mathcal{F}_0, \ldots, \mathcal{F}_{K-1}\}$ using a BRICS-based decomposition (Degen et al., 2008), a common fragmentation strategy employed in molecular machine learning (Zhang et al., 2021; Wang et al., 2022; Lee et al., 2026). Hydrogens bonded to heteroatoms are retained within their parent fragment, as they can contribute to the fragment's chemical identity. Carbon-bound structural hydrogens are instead assigned to a dedicated pseudo-fragment $\mathcal{F}_H \notin \{\mathcal{F}_0, \ldots, \mathcal{F}_{K-1}\}$, as they follow from the heavy-atom skeleton. This yields a collection of contiguous fragment blocks used to construct fragment-based orderings. Full details are provided in Appendix A.2.

Fragment-conditioned sequence Given the fragment decomposition above, for molecules with $K \geq 2$ we sample a seed size $s \sim \mathrm{Uniform}\{1, \ldots, K-1\}$ and sample a fragment subset $\mathcal{S} \subset \{\mathcal{F}_0, \ldots, \mathcal{F}_{K-1}\}$ uniformly among all subsets of size s. For molecules with $K = 1$, we instead use the default autoregressive sequence. The selected subset $\mathcal{S}$ serves as the *seed*, while the complement $\mathcal{S}^c$ defines the generation target. For a sampled subset $\mathcal{S}$, we construct $\mathbf{Z}_\mathcal{S}$ and $\mathbf{Z}_{\mathcal{S}^c}$ as the latent token sequences associated with fragments in $\mathcal{S}$ and its complement $\mathcal{S}^c$, respectively. Fragments and their constituent atoms are ordered deterministically within each subset $\mathcal{S}$ and $\mathcal{S}^c$ (see Appendix A.2). Importantly, the latent tokens associated with structural hydrogens $\mathbf{Z}_{\mathcal{F}_H}$ are excluded from the fragment sampling procedure. Instead, they are appended after $\mathbf{Z}_{\mathcal{S}^c}$ in every training sequence and generated autoregressively. The resulting training sequence is

$$[\mathrm{START}] \circ [\mathrm{SEED}] \circ \mathbf{Z}_\mathcal{S} \circ [\mathrm{GEN}] \circ \mathbf{Z}_{\mathcal{S}^c} \circ \mathbf{Z}_{\mathcal{F}_H} \circ [\mathrm{STOP}], \tag{6}$$

where sentinel tokens mark the conditioning and generation regions, $\circ$ denotes sequence concatenation, and [STOP] delimits the end of the training sequence. Sentinel tokens are structural markers

rather than generation targets: the diffusion objective is applied only at atom positions. Conditioned on the seed sequence $\mathbf{Z}_{\mathcal{S}}$, the remaining tokens are generated autoregressively:

$$p_\theta(\tilde{\mathbf{Z}}_{\mathcal{S}^c} \mid \mathbf{Z}_{\mathcal{S}}) = \prod_{j=1}^{m} p_\theta(\mathbf{z}_{r_j} \mid \mathbf{Z}_{\mathcal{S}}, \mathbf{z}_{r_{<j}}), \tag{7}$$

where $\tilde{\mathbf{Z}}_{\mathcal{S}^c} = \mathbf{Z}_{\mathcal{S}^c} \circ \mathbf{Z}_{\mathcal{F}_H} = (\mathbf{z}_{r_1}, \ldots, \mathbf{z}_{r_m})$ denotes the concatenated target sequence. Unlike conventional FIM, our formulation does not perform two-sided infilling between a prefix and suffix; instead, the FIM-style transformation constructs seed-conditioned prefix-completion tasks, where a sampled subset of fragments provides the context, and the remaining fragments are generated autoregressively. In contrast to order-agnostic autoregressive training (Uria et al., 2016), which marginalizes over token orderings, we use a single deterministic factorization for each sampled fragment partition, fixing the fragment and within-fragment atom orderings. Diversity across training factorizations therefore arises from sampling different seed subsets.

3.3 Training objective

Diffusion loss To model autoregressive conditionals $p_\theta(\mathbf{z}_i \mid \mathbf{z}_{<i})$, we adopt a Diffusion Loss formulation (Li et al., 2024) and model each using a diffusion process in the continuous latent space. Let $\mathbf{c}_i$ denote the causal transformer context at step i. To model the distribution of the next latent token $\mathbf{z}_i$, we adopt a Variance-Preserving (VP) forward process (Song et al., 2021b) that perturbs $\mathbf{z}_i$ with Gaussian noise:

$$\mathbf{z}_i^{(t)} = \alpha_t \mathbf{z}_i + \sigma_t \varepsilon, \quad t \sim \mathcal{U}[0,1],\ \varepsilon \sim \mathcal{N}(0, I)\,, \tag{8}$$

where α_t and σ_t denote the signal and noise coefficients of the VP noise schedule. The denoising network $\varepsilon_\theta(\cdot)$ is implemented as a residual Multi-Layer Perceptron (MLP) with Adaptive Layer Normalization (AdaLN) conditioning (Li et al., 2024). The network predicts the injected noise ε conditioned on the noisy latent, the autoregressive context, and the diffusion timestep. Following standard diffusion training (Ho et al., 2020), we optimize the noise prediction objective

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{(\mathbf{z}_i, \mathbf{c}_i),\, t,\, \varepsilon}\left[\left\| \varepsilon_\theta(\mathbf{z}_i^{(t)}, \mathbf{c}_i, t) - \varepsilon \right\|^2\right], \tag{9}$$

where $\mathbf{c}_i$ denotes the autoregressive transformer context (Eq. 3), and $\mathbf{z}_i$ and $\mathbf{z}_i^{(t)}$ denote the clean and noisy latent targets, respectively.

Stop loss Unlike autoregressive models operating directly in data space (Cheng et al., 2025; Rose et al., 2025), our latent representations do not contain a dedicated end-of-sequence token. We therefore train a lightweight stop classifier MLP_{stop} on the transformer hidden states $\mathbf{c}_i$. At each step, it predicts the probability that generation should terminate conditioned on the current context:

$$p_\theta([\text{STOP}] \mid \mathbf{c}_i) = \sigma(\text{MLP}_{\text{stop}}(\mathbf{c}_i))\,, \tag{10}$$

where $\sigma(\cdot)$ denotes the sigmoid function. We train the classifier jointly with the diffusion model using binary cross-entropy

$$\mathcal{L}_{\text{stop}} = \mathbb{E}_{(\mathbf{c}_i, y_i^{\text{stop}})}\left[\text{BCE}\big(p_\theta([\text{STOP}] \mid \mathbf{c}_i), y_i^{\text{stop}}\big)\right], \tag{11}$$

where $y_i^{\text{stop}} \in \{0, 1\}$ is one when $\mathbf{c}_i$ corresponds to the context after the final latent token, and zero otherwise. At inference, generation stops when $\sigma(\text{MLP}_{\text{stop}}(\mathbf{c}_i)) > \tau$, with $\tau = 0.5$. Additional details are provided in Appendix A.7.

Global objective All components, the transformer, the MLP denoiser, and the stop classifier, are trained jointly, while the UAE encoder is kept frozen. The full objective is:

$$\mathcal{L} = \mathcal{L}_{\text{diff}} + \lambda_{\text{stop}} \mathcal{L}_{\text{stop}}\,, \tag{12}$$

where $\mathcal{L}_{\text{diff}}$, $\mathcal{L}_{\text{stop}}$ are defined in Eq. 9, Eq. 11, respectively, and λ_{stop} is set to 1.0.

Table 1: 3D metrics on the QM9 dataset for the unconditional generation task. **Best** and second-best entries are highlighted in bold and underlined, respectively. Models marked with (*) were evaluated using released generated molecules. Models marked with (**) were evaluated using official pre-trained weights and source code. Models marked with (†) were re-trained utilizing official source code. We report metrics evaluated on the test set of QM9 dataset in gray. For (**) and (†) we report mean ± std over three independent inference runs.

Model	3D Validity (↑) (%)	3D AtomStable (↑) (%)	3D MolStable (↑) (%)	Bond length (↓) ($\times 10^{-2}$)	Bond angle (↓) ($\times 10^{-2}$)	Dihedral angle (↓) ($\times 10^{-3}$)
QM9	98.4	99.53	96.3	0.003	1.06	0.05
Diffusion / Flow						
ADiT*	79.3	76.43	7.6	111.5	14.90	3.60
JODO*	96.3	**99.25**	**93.6**	38.1	2.23	1.37
EDM†	$95.3_{\pm 0.1}$	$97.56_{\pm 0.19}$	$87.6_{\pm 0.1}$	$23.7_{\pm 2.3}$	$2.21_{\pm 0.06}$	$0.50_{\pm 0.10}$
GeoLDM**	$93.3_{\pm 0.2}$	$98.88_{\pm 0.03}$	$89.5_{\pm 0.1}$	$42.8_{\pm 0.6}$	$2.24_{\pm 0.06}$	$0.56_{\pm 0.04}$
SemlaFlow**	$85.8_{\pm 0.1}$	$97.07_{\pm 0.03}$	$78.6_{\pm 0.2}$	$75.2_{\pm 0.8}$	$1.94_{\pm 0.02}$	$1.55_{\pm 0.21}$
MiDi**	$95.3_{\pm 0.3}$	$98.21_{\pm 0.05}$	$83.5_{\pm 0.2}$	$110.5_{\pm 1.7}$	$3.00_{\pm 0.10}$	$0.80_{\pm 0.10}$
Autoregressive						
Symphony*	68.1	90.75	43.9	19.4	11.00	2.06
G-SchNet*	80.1	93.71	63.7	41.5	9.25	6.49
G-SphereNet*	16.4	67.82	14.0	27.8	37.60	14.30
QUETZAL**	$94.8_{\pm 0.1}$	$98.28_{\pm 0.04}$	$87.4_{\pm 0.2}$	$40.1_{\pm 1.1}$	$2.20_{\pm 0.10}$	$\mathbf{0.41}_{\pm 0.03}$
NEAT**	$93.4_{\pm 0.1}$	$98.09_{\pm 0.02}$	$84.0_{\pm 0.3}$	$33.1_{\pm 0.1}$	$1.90_{\pm 0.10}$	$4.30_{\pm 0.20}$
KRONOS (ours)	$\mathbf{97.3}_{\pm 0.1}$	$99.07_{\pm 0.02}$	$92.6_{\pm 0.1}$	$\mathbf{4.8}_{\pm 0.6}$	$\mathbf{1.33}_{\pm 0.03}$	$1.00_{\pm 0.10}$

Table 2: 3D metrics on the GEOM-Drugs dataset for the unconditional generation task. **Best** and second-best entries are highlighted in bold and underlined, respectively. We report metrics evaluated on the test set of GEOM-Drugs dataset in gray. Models marked with (*) were evaluated using released generated molecules. Models marked with (**) were evaluated using official pre-trained weights and source code. Models marked with (†) were re-trained utilizing the official source code. For (**) and (†) we report mean ± std over three independent inference runs.

Model	3D Validity (↑) (%)	3D AtomStable (↑) (%)	3D MolStable (↑) (%)	Bond length (↓)	Bond angle (↓)	Dihedral angle (↓)
GEOM	96.6	86.17	2.80	0.253	0.2695	0.0153
Diffusion / Flow						
ADiT*	97.3	75.46	0.12	1.323	1.0474	0.0548
JODO*	95.6	84.49	1.11	0.089	0.0099	**0.0005**
EDM†	$90.5_{\pm 0.6}$	$83.35_{\pm 0.05}$	$1.01_{\pm 0.04}$	$0.865_{\pm 0.005}$	$0.8368_{\pm 0.0060}$	$0.0567_{\pm 0.0004}$
GeoLDM**	$\mathbf{99.4}_{\pm 0.1}$	$84.30_{\pm 0.04}$	$0.59_{\pm 0.02}$	$0.916_{\pm 0.006}$	$0.7885_{\pm 0.0024}$	$0.0474_{\pm 0.0003}$
SemlaFlow**	$92.4_{\pm 0.2}$	$84.10_{\pm 0.09}$	$2.17_{\pm 0.24}$	$0.569_{\pm 0.001}$	$0.1241_{\pm 0.0015}$	$0.0007_{\pm 0.0000}$
MiDi**	$92.0_{\pm 0.4}$	$75.48_{\pm 0.16}$	$0.30_{\pm 0.04}$	$0.113_{\pm 0.001}$	$0.0915_{\pm 0.0029}$	$0.0046_{\pm 0.0003}$
Autoregressive						
QUETZAL†	$93.8_{\pm 0.2}$	$86.27_{\pm 0.03}$	$2.04_{\pm 0.06}$	$0.882_{\pm 0.008}$	$0.9370_{\pm 0.0068}$	$0.0614_{\pm 0.0015}$
NEAT**	$98.1_{\pm 0.1}$	$81.16_{\pm 0.04}$	$0.50_{\pm 0.01}$	$1.194_{\pm 0.002}$	$1.3939_{\pm 0.0342}$	$0.1543_{\pm 0.0012}$
KRONOS (ours)	$94.7_{\pm 0.2}$	$\mathbf{86.45}_{\pm 0.03}$	$\mathbf{2.70}_{\pm 0.10}$	$\mathbf{0.048}_{\pm 0.003}$	$\mathbf{0.0094}_{\pm 0.0004}$	$\mathbf{0.0005}_{\pm 0.0002}$

4 EXPERIMENTS

We evaluate KRONOS on three complementary tasks: unconditional 3D molecule generation on the QM9 and GEOM-Drugs datasets (Tables 1 and 2), fragment-conditioned generation from partial molecular context (Table 3), and conformational stability assessed through semi-empirical GFN2-xTB optimization (Figure 2). The UAE is trained from scratch with inter-fragment message passing disabled in the encoder to ensure fragment-local latent representations (see Appendix A.3). Additional experiments on 2D molecule generation are provided in Appendix A.9.

4.1 EXPERIMENTAL SETUP

Datasets We evaluate our framework on QM9 (Ramakrishnan et al., 2014) and GEOM-Drugs (Axelrod & Gómez-Bombarelli, 2022) datasets. For QM9, we use the same train/validation/test split established in existing work (Luo et al., 2025; Huang et al., 2024), resulting in 97, 609/14, 405/12, 405 molecules. Similarly, for GEOM-Drugs we follow the pre-processing and train/validation/test splits

utilized in existing literature (Vignac et al., 2023), retaining the five lowest-energy conformers per molecule. This results in $1,172,532/146,405/146,793$ conformers for training, validation, and testing, respectively.

Baselines We compare the proposed method with several existing autoregressive and diffusion baselines for 3D molecule generation. Autoregressive baselines include G-SchNet (Gebauer et al., 2019), G-SphereNet (Luo & Ji, 2022), Symphony (Daigavane et al., 2024), QUETZAL (Cheng et al., 2025), NEAT (Rose et al., 2025). Diffusion models include EDM (Hoogeboom et al., 2022b), MiDi (Vignac et al., 2023), ADiT (Joshi et al., 2025), GeoLDM (Xu et al., 2023) and JODO (Huang et al., 2024). SemlaFlow (Irwin et al., 2024) is included as a flow matching baseline. We do not report performance of UDM-3D (Luo et al., 2025), because we are unable to reproduce reasonable results utilizing the source code, and pre-trained checkpoints or pre-generated molecules are unavailable at the time of writing.

Metrics We evaluate generated molecules using standard metrics for 3D molecule generation. Specifically, we report validity and stability of reconstructed molecular graphs together with *Maximum Mean Discrepancy* (MMD) (Gretton et al., 2012) between generated and reference bond length, bond angle, and dihedral angle distributions. Complete metric definitions are provided in Appendix A.5.

4.2 Results

Unconditional molecule generation Results on QM9 and GEOM-Drugs are reported in Tables 1 and 2, respectively. On QM9, KRONOS achieves the highest validity and the strongest overall performance among autoregressive methods, while remaining competitive with diffusion-based models. As shown in Appendix A.11, reducing the fragment-conditioning probability p_f further improves unconditional generation performance, indicating that the small performance trade-off primarily arises from the mixed training strategy. On GEOM-Drugs, KRONOS demonstrates strong unconditional generation performance across the evaluated metrics. Compared with recent autoregressive baselines such as QUETZAL and NEAT, these results demonstrate that latent autoregressive generation can improve both molecular stability and geometric fidelity.

Assessment of conformational stability While the molecular stability criterion (**3D MolStable**) introduced by Hoogeboom et al. (2022b) is a standard evaluation metric for molecule generation, it relies on heuristic bond-assignment procedures that can be sensitive to aromatic systems and distance thresholds (Rose et al., 2025; Daigavane et al., 2024; Irwin et al., 2024). To complement this graph-based evaluation, we perform semi-empirical quantum-chemical refinement using GFN2-xTB (Bannwarth et al., 2019). For each generated molecule, we compute its single-point energy before optimization and report (i) the Root Mean Square Deviation (RMSD) between the generated and optimized conformations after optimal rigid-body alignment using the Kabsch algorithm (Kabsch, 1976), and (ii) the absolute energy difference $|\Delta E|$. Lower values indicate closer agreement with the optimized structure. We evaluate $10{,}000$ generated molecules per model. Figure 2 summarizes the results. On QM9, KRONOS attains the third-lowest median energy difference (within 0.01 eV of the best model), together with a competitive median RMSD. On GEOM-Drugs, it attains the second-lowest median energy difference, while geometric agreement with the optimized structures is more variable. Overall, the results indicate consistent energetic agreement with the xTB-optimized structures, whereas geometric agreement is more variable across datasets.

Fragment-conditioned molecule generation We evaluate whether the proposed framework can generate chemically coherent 3D molecules from partial molecular context. This task follows the spirit of *prefix completion* introduced by Rose et al. (2025). However, rather than conditioning on a fixed set of manually designed scaffolds, we construct conditioning prefixes by our fragment decomposition scheme (Appendix A.2) applied to molecules from the GEOM-Drugs test set, and treat each resulting fragment as conditioning context. The same fragment-derived prefixes are used to evaluate KRONOS, NEAT, and QUETZAL. This setting represents a practical, conditional-design scenario in which a fragment from a known three-dimensional structure is retained while alternative molecular structures are generated around it, as arises in scaffold-based design (Lim et al., 2020); conditioning on a fragment resolved by X-ray crystallography and elaborated from its bound pose is likewise

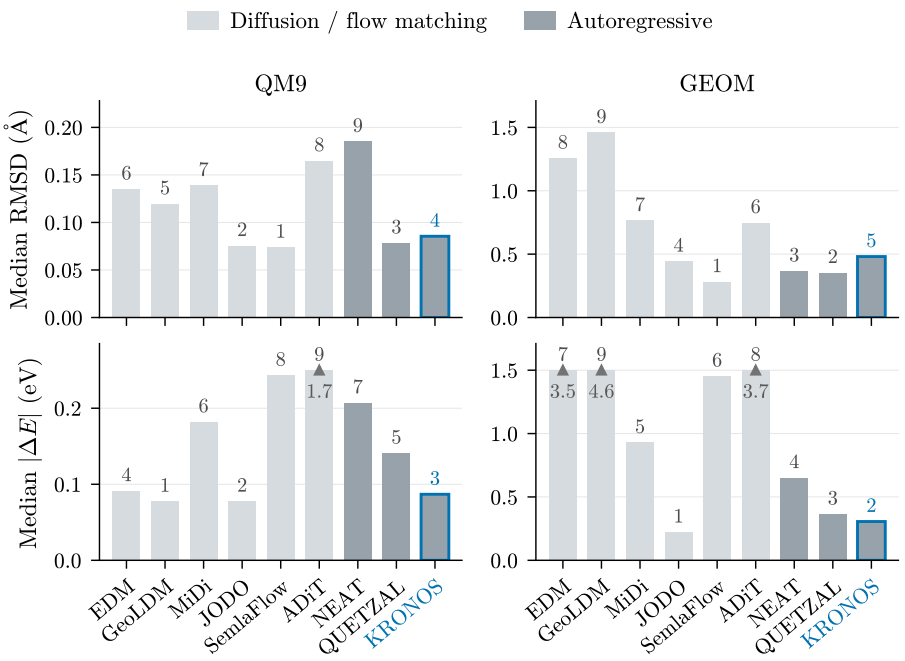

Figure 2: Physical plausibility of generated conformers. We report the median RMSD between generated and xTB-relaxed geometries (top) and the median absolute energy difference between them (bottom). Numbers above the bars indicate rank within each panel (1 = best); bars marked with a caret exceed the axis limit and show their true value.

Table 3: Fragment-conditioned molecule generation performance over $\sim$ 15k molecule-fragment pairs. Models marked with (**) were evaluated using official pre-trained weights and source code. Models marked with (†) were re-trained utilizing the official source code. **Best** and second-best entries are highlighted in bold and underlined, respectively. Results are aggregated across molecule–fragment pairs and reported as mean $\pm$ 95% confidence interval.

	Mean@3			Best@3			Unique@3
Model	**3D Validity**	**3D AtomStable**	**3D MolStable**	**3D Validity**	**3D AtomStable**	**3D MolStable**	
QUETZAL†	$48.7_{\pm 0.6}$	$67.6_{\pm 0.2}$	$0.2_{\pm 0.1}$	$70.6_{\pm 0.7}$	$74.2_{\pm 0.2}$	$0.5_{\pm 0.1}$	$41.0_{\pm 0.5}$
NEAT**	$\mathbf{93.7}_{\pm 0.3}$	$\mathbf{83.4}_{\pm 0.1}$	$\underline{0.8}_{\pm 0.1}$	$\mathbf{97.4}_{\pm 0.2}$	$\mathbf{87.9}_{\pm 0.1}$	$\mathbf{2.1}_{\pm 0.2}$	$\mathbf{93.7}_{\pm 0.3}$
KRONOS (ours)	$\underline{79.7}_{\pm 0.4}$	$\underline{82.1}_{\pm 0.1}$	$\mathbf{0.9}_{\pm 0.1}$	$\underline{96.2}_{\pm 0.3}$	$\underline{86.5}_{\pm 0.1}$	$\underline{1.8}_{\pm 0.2}$	$\underline{74.2}_{\pm 0.4}$

standard in fragment-based drug discovery (Erlanson et al., 2016). We collect $\sim$ 15k molecule-fragment pairs and sample 3 completions per pair. We report Best@3, Mean@3 and Unique@3 in Table 3; Mean@3 averages each metric across the three completions, Best@3 selects the best value, and Unique@3 measures the fraction of valid completions that are distinct based on their inferred 2D molecular graphs. Figure 3 (left) illustrates the task.

The results show that fragment-conditioned generation can be learned without a fully permutation-invariant model. NEAT achieves higher overall metrics on this task, consistent with its set-based, permutation-invariant formulation and training objective, in which every example is a partial-structure completion task. KRONOS nevertheless achieves competitive fragment-conditioned generation while retaining strong unconditional generation performance (Tables 1, 2) and efficient sampling (Appendix A.8). It substantially improves over QUETZAL, which shares its ordered autoregressive formulation but is trained only on fixed left-to-right sequences.

To assess how the amount of conditioning structure affects molecular exploration, we measure the similarity of generated completions to the parent molecule from which each conditioning fragment was extracted. Similarity is computed on 2D graphs obtained by perceiving bond orders from the generated coordinates using `rdDetermineBonds` from `RDKit` (Landrum); the distance-based

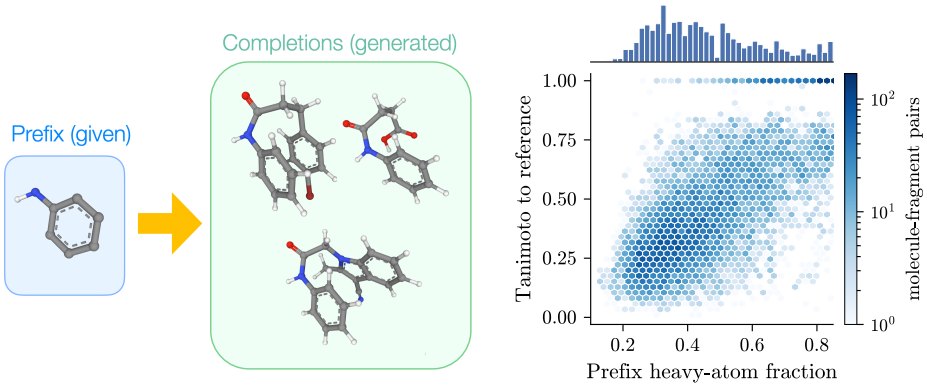

Figure 3: Fragment-conditioned molecule generation. *Left*) a conditioning fragment ("prefix") extracted from a molecule in the GEOM-Drugs test set, and three independent completions generated by KRONOS, that extend the fragment into diverse, chemically plausible 3D molecules. *Right*) prefix heavy-atom fraction versus Tanimoto similarity to the *reference* (the parent molecule the prefix was taken from).

inference used for 3D metrics assign bond orders, but for GEOM-Drugs these are reduced to single bonds when constructing the molecular graph, following established conventions Hoogeboom et al. (2022b); Rose et al. (2025); the resulting SMILES therefore cannot distinguish, e.g., aromatic from saturated rings. For each molecule-fragment pair, we retain the completion with highest Tanimoto similarity to the reference. When at most half of the reference's heavy atoms are supplied (63% of pairs), the median similarity is 0.32 and the retained completion is identical to the reference molecule in only 0.4% of pairs. Similarity increases with the size of the conditioning fragment, with reconstruction becoming common only once more than 60% of the reference heavy atoms are supplied (Figure 3 (right)). The size of the conditioning fragment therefore acts as a control on exploration: small fragments yield diverse elaborations around a retained core, while larger fragments increasingly determine the resulting molecule. Appendix A.15 (Figure 12) shows representative completions.

5 Conclusion

We introduced KRONOS, a latent autoregressive diffusion framework for 3D molecule generation. By operating in the unified latent space of a UAE, KRONOS generates molecular topology and geometry through a single, continuous token representation while its autoregressive formulation naturally supports variable-length generation. A FIM-inspired, fragment-conditioned training strategy further enables unconditional and fragment-conditioned generation within a single architecture. Experiments on QM9 and GEOM-Drugs demonstrate strong unconditional generation performance and competitive fragment-conditioned generation, while semi-empirical geometry optimization supports the physical plausibility of generated conformations. Overall, our results establish latent autoregressive generation as a flexible pathway for 3D molecule design. We discuss limitations and directions for future work in Appendix A.13.

AI Disclosure

In this work, we used generative AI tools to assist with research execution, including adapting existing baseline implementations and accelerating experimental workflows. AI tools were also used for literature search and to polish and improve the readability of the manuscript. We did not use generative AI tools to design the core methodology or experiments. All AI-assisted work was reviewed and verified by the authors. In particular, AI-assisted code and analyses were checked for correctness, while AI-assisted text was reviewed and edited for accuracy, consistency, and readability. The authors take full responsibility for the final content of this work, including all claims, analyses, code, and other artifacts.

Code Availability Statement

We provide an anonymous repository containing the source code and instructions for reproducing our experiments: https://anonymous.4open.science/r/kronos-AF75/. Processed data and pre-trained checkpoints will be released upon publication.

References

Simon Axelrod and Rafael Gómez-Bombarelli. GEOM, energy-annotated molecular conformations for property prediction and molecular generation. *Scientific data*, 9(1):185, April 2022. ISSN 2052-4463. doi: 10.1038/s41597-022-01288-4. URL https://europepmc.org/articles/PMC9023519.

Christoph Bannwarth, Sebastian Ehlert, and Stefan Grimme. GFN2-xTB—an accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *Journal of Chemical Theory and Computation*, 15(3):1652–1671, 2019. doi: 10.1021/acs.jctc.8b01176. URL https://doi.org/10.1021/acs.jctc.8b01176. PMID: 30741547.

Mohammad Bavarian, Heewoo Jun, Nikolas Tezak, John Schulman, Christine McLeavey, Jerry Tworek, and Mark Chen. Efficient training of language models to Fill in the Middle, 2022. URL https://arxiv.org/abs/2207.14255.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. MaskGIT: Masked generative Image Transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11315–11325, 2022.

Austin H. Cheng, Chong Sun, and Alán Aspuru-Guzik. Scalable Autoregressive 3D Molecule Generation, 2025. URL https://arxiv.org/abs/2505.13791.

Ameya Daigavane, Song Eun Kim, Mario Geiger, and Tess Smidt. Symphony: Symmetry-equivariant point-centered spherical harmonics for 3D molecule generation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=MIEnYtlGyv.

Jörg Degen, Christof Wegscheid-Gerlach, Andrea Zaliani, and Matthias Rarey. On the art of compiling and using 'drug-like' chemical fragment spaces. *ChemMedChem*, 3(10):1503–1507, 2008. doi: https://doi.org/10.1002/cmdc.200800178. URL https://chemistry-europe.onlinelibrary.wiley.com/doi/abs/10.1002/cmdc.200800178.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Cameron Diao and Ricky Loynd. Relational attention: Generalizing transformers for graph-structured tasks. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=cFuMmbWiN6.

Daniel A Erlanson, Stephen W Fesik, Roderick E Hubbard, Wolfgang Jahnke, and Harren Jhoti. Twenty years on: the impact of fragments on drug discovery. *Nature reviews Drug discovery*, 15 (9):605–619, 2016.

Niklas Gebauer, Michael Gastegger, and Kristof Schütt. Symmetry-adapted generation of 3d point sets for the targeted discovery of molecules. *Advances in neural information processing systems*, 32, 2019.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL http://jmlr.org/papers/v13/gretton12a.html.

Adi Haviv, Ori Ram, Ofir Press, Peter Izsak, and Omer Levy. Transformer language models without positional encodings still learn positional information. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 1382–1390, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.99. URL https://aclanthology.org/2022.findings-emnlp.99/.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Emiel Hoogeboom, Alexey A. Gritsenko, Jasmijn Bastings, Ben Poole, Rianne van den Berg, and Tim Salimans. Autoregressive diffusion models. In *International Conference on Learning Representations (ICLR)*, 2022a.

Emiel Hoogeboom, Victor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3D. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 8867–8887. PMLR, 17–23 Jul 2022b. URL https://proceedings.mlr.press/v162/hoogeboom22a.html.

Han Huang, Leilei Sun, Bowen Du, and Weifeng Lv. Learning joint 2-D and 3-D graph diffusion models for complete molecule generation. *IEEE Transactions on Neural Networks and Learning Systems*, 35(9):11857–11871, 2024. doi: 10.1109/TNNLS.2024.3416328.

Ilia Igashov, Hannes Stärk, Clément Vignac, Arne Schneuing, Victor Garcia Satorras, Pascal Frossard, Max Welling, Michael Bronstein, and Bruno Correia. Equivariant 3D-conditional diffusion model for molecular linker design. *Nature Machine Intelligence*, 6(4):417–427, 2024.

Ross Irwin, Alessandro Tibo, Jon Paul Janet, and Simon Olsson. Efficient 3D molecular generation with flow matching and scale optimal transport. In *ICML 2024 AI for Science Workshop*, 2024. URL https://openreview.net/forum?id=CxAjGjdkqu.

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 *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=89QPmZjIhv.

Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 32(5):922–923, 1976.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=k7FuTOWMOc7.

Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Gregory Landrum. RDKit: Open-source cheminformatics. URL https://www.rdkit.org.

Joongwon Lee, Seonghwan Kim, Seokhyun Moon, Hyunwoo Kim, and Woo Youn Kim. FragFM: Hierarchical framework for efficient molecule generation via fragment-level discrete flow matching, 2026. URL https://arxiv.org/abs/2502.15805.

Youhan Lee and Hasun Yu. ProtFIM: Fill-in-Middle protein sequence design via protein language models. *arXiv preprint arXiv:2303.16452*, 2023.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. In *Advances in Neural Information Processing Systems*, 2024.

Jaechang Lim, Sang-Yeon Hwang, Seokhyun Moon, Seungsu Kim, and Woo Youn Kim. Scaffold-based molecular design with a graph generative model. *Chemical Science*, 11(4):1153–1164, 01 2020. ISSN 2041-6520. doi: 10.1039/c9sc04503a. URL https://doi.org/10.1039/c9sc04503a.

Gang Liu, Jie Chen, Yihan Zhu, Michael Sun, Tengfei Luo, Nitesh V Chawla, and Meng Jiang. Graph diffusion transformers are in-context molecular designers. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=lJ87GN5zJc.

Yanchen Luo, Zhiyuan Liu, Yi Zhao, Sihang Li, Hengxing Cai, Kenji Kawaguchi, Tat-Seng Chua, Yang Zhang, and Xiang Wang. Towards unified and lossless latent space for 3d molecular latent diffusion modeling. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=g2XE40zTrj.

Youzhi Luo and Shuiwang Ji. An Autoregressive Flow Model for 3D Molecular Geometry Generation from Scratch. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=C03Ajc-NS5W.

Hyunsoo Park and Aron Walsh. Guiding generative models to uncover diverse and novel crystals via reinforcement learning. *arXiv preprint arXiv:2511.07158*, 2025.

Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1(1):140022, 2014. doi: 10.1038/sdata.2014.22. URL https://doi.org/10.1038/sdata.2014.22.

Daniel Rose, Roxane Axel Jacob, Johannes Kirchmair, and Thierry Langer. NEAT: Neighborhood-Guided, Efficient, Autoregressive Set Transformer for 3D Molecular Generation. *arXiv preprint arXiv:2512.05844*, 2025.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=St1giarCHLP.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=PxTIG12RRHS.

Benigno Uria, Marc-Alexandre Côté, Karol Gregor, Iain Murray, and Hugo Larochelle. Neural autoregressive distribution estimation. *J. Mach. Learn. Res.*, 17(1):7184–7220, January 2016. ISSN 1532-4435.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, 2017.

Clement Vignac, Nagham Osman, Laura Toni, and Pascal Frossard. MiDi: Mixed Graph and 3D Denoising Diffusion for Molecule Generation. In *ICLR 2023 - Machine Learning for Drug Discovery workshop*, 2023. URL https://openreview.net/forum?id=M6Ifac3G4HK.

Yuyang Wang, Rishikesh Magar, Chen Liang, and Amir Barati Farimani. Improving molecular contrastive learning via faulty negative mitigation and decomposed fragment contrast. *Journal of Chemical Information and Modeling*, 62(11):2713–2725, 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*, pp. 38592–38610. PMLR, 2023.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf.

Zaixi Zhang, Qi Liu, Hao Wang, Chengqiang Lu, and Chee-Kong Lee. Motif-based graph self-supervised learning for molecular property prediction. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=gwGYN1fQY8H.

A Appendix

A.1 Related work

Diffusion-based models for 3D molecule generation Diffusion models have become the dominant paradigm for 3D molecule generation, achieving state-of-the-art performance across standard benchmarks. Early approaches such as EDM (Hoogeboom et al., 2022b) introduced E(3)-equivariant diffusion models that jointly denoise atom coordinates and atom types to generate 3D molecular structures. MiDi (Vignac et al., 2023) extends this framework by jointly generating molecular graph topology and 3D geometry. JODO (Huang et al., 2024) further develops unified joint 2D/3D diffusion through a dedicated graph transformer that explicitly models interactions between molecular topology and geometry. More recently, ADiT (Joshi et al., 2025) explored latent diffusion over unified all-atom representations, enabling a single transformer-based model to generate both molecules and materials within a shared latent space. Beyond unconditional generation, diffusion has also been adapted to conditional design tasks such as linker generation (Igashov et al., 2024), where a connector is generated between two given fragments, and its size must be supplied or produced by a separately trained predictor. Parallel to diffusion, flow-matching approaches such as SemlaFlow (Irwin et al., 2024) learned continuous transport maps from noise to molecular distributions, achieving competitive generation quality while requiring significantly fewer sampling steps. However, in both unconditional and conditional settings the molecular size is fixed or predicted before generation; in the conditional case this limits the ability to adapt the generated structure's size to the given context.

Autoregressive models for 3D molecule generation Early autoregressive approaches generated molecules sequentially by predicting one atom at a time together with its 3D placement (Gebauer et al., 2019; Luo & Ji, 2022). Symphony (Daigavane et al., 2024) subsequently demonstrated that incorporating higher-order E(3)-equivariant representations substantially improves autoregressive molecule generation, narrowing the gap to diffusion-based methods. More recent autoregressive approaches, including QUETZAL (Cheng et al., 2025) and NEAT (Rose et al., 2025), combine autoregressive generation with continuous generative models for molecular geometry, using diffusion and flow-matching objectives, respectively. These methods substantially narrow the performance gap to diffusion models while adaptively determining molecular size during generation. However, existing autoregressive models still face a performance trade-off between unconditional generation and flexible conditioning on partial molecular structures. Our work addresses this trade-off through a unified autoregressive training framework that jointly supports both objectives.

A.2 FRAGMENT-BASED MOLECULAR DECOMPOSITION

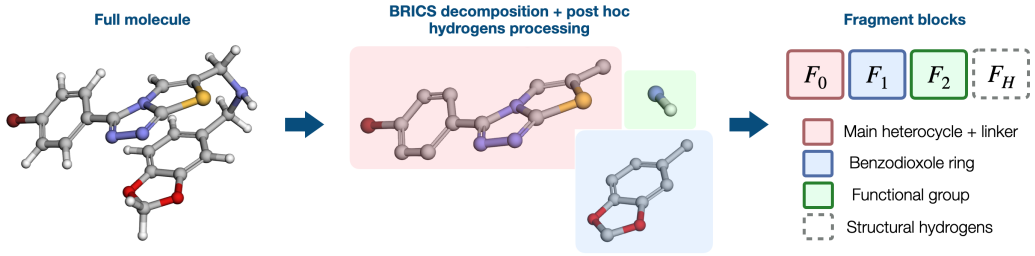

Figure 4: Visual example of the adopted molecular decomposition strategy. The initial molecule is first decomposed utilizing the BRICS algorithm. Then, hydrogens bonded to heteroatoms are retained by their parent fragment (preserving common functional groups as coherent units); structural hydrogens (carbon-bonded) are collected into a separate, pseudo-fragment that is always appended at the end of the obtained sequence.

Rather than considering arbitrary atom permutations, we decompose each molecule into chemically meaningful fragments and assign a deterministic ordering within each fragment. During training, random fragment partitions define the conditioning and generation regions, yielding diverse fragment-conditioned autoregressive generation tasks while avoiding arbitrary atom-level permutations. Fragment boundaries are obtained from a BRICS-based decomposition (Degen et al., 2008), which cleaves bonds at retrosynthetically meaningful sites to produce chemically coherent building blocks, and has been explored in molecular generative modeling (Lee et al., 2026; Liu et al., 2026). Removing these bonds partitions the molecular graph into connected components. Hydrogens bonded to heteroatoms are assigned to the same fragment as their parent atom, preserving functional groups such as $-\mathrm{OH}$, $-\mathrm{NH_2}$, and $-\mathrm{SH}$ as coherent units. In contrast, carbon-bound structural hydrogens are collected into a separate pseudo-fragment $\mathcal{F}_H$, which is always appended after all fragment blocks. Within each fragment $\mathcal{F}_k$, atoms follow a fixed ordering obtained through a deterministic depth-first search (DFS) traversal of the fragment subgraph. During training, the fragment set is randomly partitioned into conditioning and generation fragments. Within each set, fragments follow a fixed ordering, while atoms within each fragment follow the deterministic DFS ordering defined above. Structural hydrogens are always appended last. Consequently, the model is exposed to many fragment-conditioned generation tasks while avoiding arbitrary atom-level permutations. We provide an illustrative example of the adopted decomposition in Figure 4.

A.3 FRAGMENT-LOCAL LATENT ENCODING

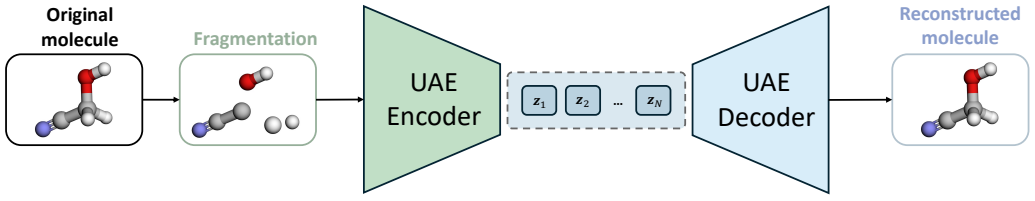

Figure 5: To enable fragment-conditioned generation, molecules are decomposed into disconnected fragment components before UAE encoding by removing inter-fragment bonds. Latent representations are therefore computed using only intra-fragment information. The UAE decoder subsequently reconstructs the full molecular graph, including bonds between fragments.

In the original UAE architecture (Luo et al., 2025), latent representations are computed through message passing across the full molecular graph, allowing each latent token to aggregate information from progressively larger molecular neighborhoods. Consequently, the latent representation of each fragment depends not only on its internal structure but also on atoms belonging to neighboring fragments. However, our fragment-conditioned generation setting requires latent representations that can be computed for fragments in isolation at inference time. To achieve this, we modify the

encoder connectivity during UAE training. Given a fragment decomposition $\{\mathcal{F}_0, \ldots, \mathcal{F}_{K-1}\}$ of a molecular graph $\mathbf{G}$, we remove all inter-fragment edges prior to encoder message passing, restricting information propagation to atoms within the same fragment (Figure 5). Importantly, this modification affects only the graph used by the encoder to compute latent representations. The decoder continues to reconstruct the complete molecule graph, including inter-fragment connectivity and global 3D geometry. Consequently, latent tokens are encouraged to capture fragment-local information, while the decoder integrates information across fragments and recovers global structure. Empirically, UAE reconstruction performance remains unchanged even when all inter-atomic edges are removed during encoding, suggesting that the decoder can recover global relationships from the collection of local latent representations. We attribute this flexibility to the decoder's joint attention over latent tokens and to the atom embeddings already encoding local identity and spatial information.

A.4 Fragment statistics

Figure 6 shows the distribution of BRICS fragment counts K across molecules in QM9 and GEOM; the structural-hydrogen pseudo-fragment $\mathcal{F}_H$ is not included. QM9 molecules contain an average of 2.46 fragments (median = 2), while GEOM molecules exhibit greater structural diversity, averaging 3.40 fragments (median = 3).

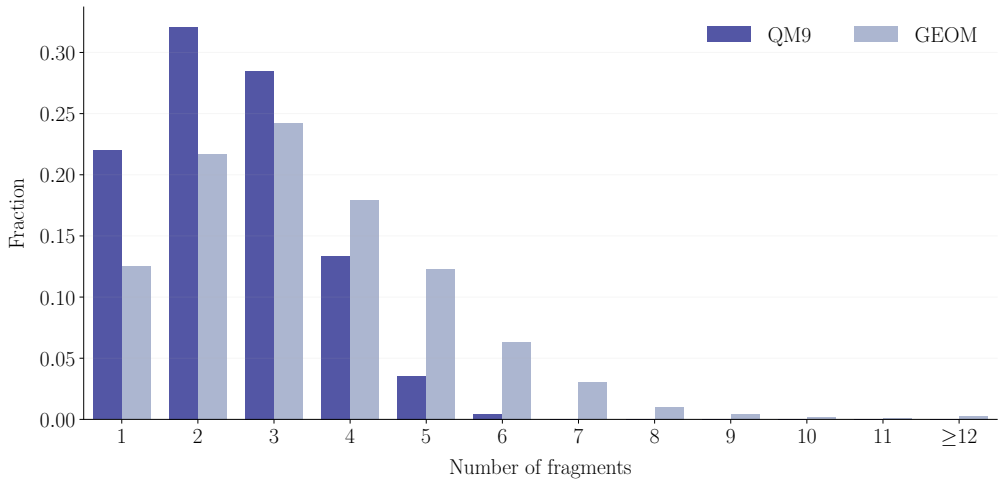

Figure 6: Distribution of molecular fragment counts across molecules in QM9 and GEOM.

A.5 Evaluation Protocol

All metrics are computed following the evaluation protocols used in prior 3D molecule generation work (Luo et al., 2025).

3D Structural Metrics. We first evaluate generated molecules directly from their predicted 3D coordinates. 2D molecular graphs are reconstructed from atomic coordinates via distance-based bond inference (Hoogeboom et al., 2022a), after which chemical validity and stability are evaluated on the reconstructed graphs.

- **3D Validity**: fraction of generated conformations that can be converted into chemically-valid 2D molecular graphs.
- **3D AtomStable**: fraction of atoms satisfying valence constraints under the reconstructed 2D bond order.
- **3D MolStable**: fraction of molecules whose reconstructed 2D molecular graph satisfies chemical stability constraints.

We also compare local geometric statistics between generated and reference conformations. Specifically, we measure the discrepancy of bond length, bond angle, and dihedral angle distributions using *Maximum Mean Discrepancy* (MMD). Lower values indicate closer agreement with the reference molecular geometry distribution and therefore more realistic 3D conformations.

- **Bond Length**: MMD between generated and reference bond length distributions for eight frequent bond types (e.g., C–C, C–N, C–O). Computed using Gaussian kernel with bandwidth tuned via the median heuristics.
- **Bond Angle**: MMD between bond angle distributions of generated and reference molecules. Bond angles are defined over triplets of bonded atoms (e.g. C–C–C).
- **Dihedral Angle**: MMD between dihedral (torsional) angle distributions of generated and reference molecules. Dihedral angles are defined over bonded four-atom sequences (e.g. C–C–C–C).

2D Graph Metrics We additionally evaluate the quality of the explicitly predicted 2D molecular graph topology obtained directly from the decoder outputs, without relying on coordinate-based bond reconstruction. In this setting, validity and stability are evaluated directly on the predicted 2D molecular graphs.

- **2D Validity**: fraction of predicted molecular graphs that satisfy chemical validity constraints.
- **2D AtomStable**: fraction of atoms satisfying valence constraints in the predicted 2D graph topology.
- **2D MolStable**: fraction of predicted 2D molecular graphs satisfying molecule-level chemical stability constraints.
- **Validity & Completeness (V&C)**: percentage of generated molecules that correspond to connected molecular graphs without disconnected fragments, and that can be converted into valid SMILES representations.
- **Validity & Uniqueness (V&U)**: percentage of chemically valid generated molecules that are unique under canonical SMILES comparison.
- **Validity & Uniqueness & Novelty (V&U&N)**: percentage of generated molecules that are simultaneously chemically valid, unique, and absent from the training set under canonical SMILES comparison.

V&C, **V&U**, **V&U&N** are evaluated only on directly predicted 2D molecular graphs, since uniqueness and novelty are defined with respect to 2D molecular identity (atom and bond topology). Different 3D conformations of the same 2D molecular graph are therefore not considered distinct or novel molecules.

A.6 Effect of Atom Ordering

Autoregressive molecular generators are sensitive to the order in which atoms are generated, as the chosen factorization determines the sequential dependencies learned by the model (Cheng et al., 2025; Haviv et al., 2022; Kazemnejad et al., 2023). Our fragment-based construction introduces variation at the level of molecular fragments while retaining a deterministic ordering of atoms within each fragment. To assess the importance of this structured ordering, we compare against training with fully random atom permutations. As shown in Table 4, random atom ordering substantially degrades unconditional generation: 3D Validity and 3D MolStable decrease from 98.0% and 94.9% under the default ordering to 74.2% and 35.2%, respectively. In contrast, the mixed fragment-conditioned training strategy ($p_f = 0.5$) retains performance close to the default ordering, achieving 97.3% 3D Validity and 92.6% 3D MolStable. These results suggest that introducing diversity through structured fragment-level factorizations preserves useful sequential regularities, whereas arbitrary atom-level permutations substantially disrupt autoregressive generation.

Table 4: Effect of atom ordering on QM9 dataset. We report Mean $\pm$ std over three independent inference runs.

Atom ordering	3D Validity↑ (%)	3D AtomStable↑ (%)	3D MolStable↑ (%)
Fragment-based (p_f=0.5)	$97.3_{\pm 0.1}$	$99.07_{\pm 0.02}$	$92.6_{\pm 0.1}$
Default (p_f=0.0)	$98.0_{\pm 0.1}$	$99.34_{\pm 0.01}$	$94.9_{\pm 0.0}$
Random permutation	$74.2_{\pm 0.4}$	$88.40_{\pm 0.09}$	$35.2_{\pm 0.2}$

A.7 KRONOS Sampling Procedure

Algorithm 1 describes sampling for both unconditional and fragment-conditioned generation modes; the two differ only in the prompt P. For unconditional generation $P = [\text{START}]$ and $M = N_{\max}$, with $N_{\max}$ denoting the maximum number of allowed atoms. For fragment-conditioned generation $P = [\text{START}] \circ [\text{SEED}] \circ \mathbf{Z}_{\mathcal{S}} \circ [\text{GEN}]$, where $\mathbf{Z}_{\mathcal{S}}$ are the seed fragment latents, $\circ$ denotes sequence concatenation, and $M = N_{\max} - |\mathbf{Z}_{\mathcal{S}}|$; in the unconditional case $\mathbf{Z}_{\mathcal{S}} = ()$ and the returned molecule is decoded from the generated tokens alone.

Algorithm 1 KRONOS sampling

Require: transformer, denoiser ε_θ, stop head MLP_{stop}, decoder $\mathcal{D}_{\text{UAE}}$; prompt P, threshold τ, DDIM steps S, maximum number of atoms M

$\mathbf{Z} \leftarrow ()$
for $i = 1, \ldots, M$ **do**
  $\mathbf{c}_i \leftarrow \text{Transformer}(P \circ \mathbf{Z})$ ▷ context at the final position
  if $\sigma(\text{MLP}_{\text{stop}}(\mathbf{c}_i)) > \tau$ **then**
    break ▷ terminate with $|\mathbf{Z}|$ generated atoms
  end if
  $\mathbf{z}_i \leftarrow \text{DDIM}(\varepsilon_\theta, \mathbf{c}_i, S)$ ▷ S denoiser evaluations
  $\mathbf{Z} \leftarrow \mathbf{Z} \circ \mathbf{z}_i$
end for
return $\mathcal{D}_{\text{UAE}}(\mathbf{Z}_{\mathcal{S}} \circ \mathbf{Z})$

A.8 SAMPLING EFFICIENCY

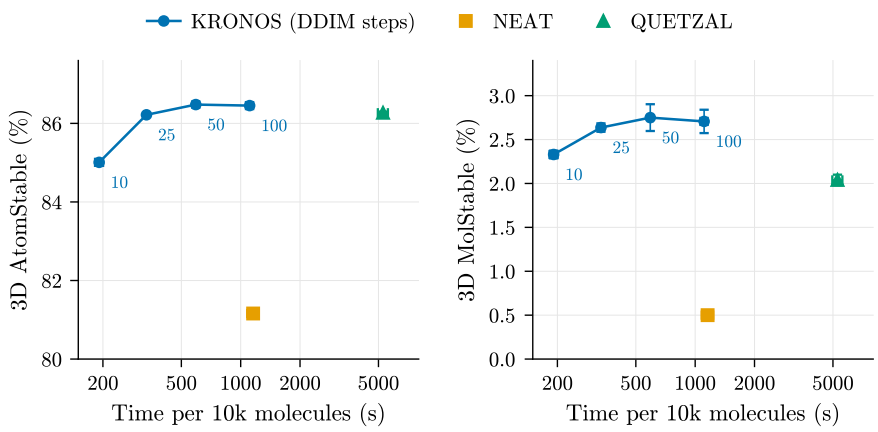

Figure 7: Quality versus sampling cost for 10k molecules generated by each model trained on GEOM-Drugs dataset (mean $\pm$ std over three runs). KRONOS retains the highest 3D MolStable down to 25 DDIM steps, at under a third of NEAT's sampling time.

We measure wall-clock time of KRONOS compared to two main autoregressive baselines (QUETZAL and NEAT). We sample 10k GEOM-Drugs molecules on a single NVIDIA Quadro RTX 6000 (24 GB), averaged over three runs, each method using the largest batch size fitting in GPU memory (KRONOS: 512; QUETZAL: $2,048$; NEAT: $10,000$). Absolute times are not comparable to those reported by Rose et al. (2025) on an RTX 4090, where NEAT and QUETZAL are respectively $5.6\times$ and $4.8\times$ faster; the ordering between methods is unchanged.

Sampling cost splits between a transformer backbone, run once per atom, and a denoiser, run for a number of steps per atom. As sampler order differs, we report denoiser network function evaluations (NFE) alongside the number of steps each method is configured with: KRONOS uses first-order DDIM (Song et al., 2021a) with 100 steps, NEAT employs first-order Euler–Maruyama with 60 steps (NFE = steps), whereas QUETZAL uses a second-order Heun sampler (Karras et al., 2022) with 120 steps (NFE = $2 \cdot$ steps $- 1$), giving $4,400$, $2,640$ and $10,516$ evaluations for an average 44-atom molecule in GEOM-Drugs. The three denoisers share width and depth (1,536 channels, 6 residual blocks), so per-evaluation cost is comparable.

Backbone cost differs structurally. KRONOS attends causally, so an atom's representation is unaffected by atoms placed after it and is retained in a key–value (KV) cache: one backbone pass per atom. NEAT attends bidirectionally over the atom set, so cached state cannot be reused and the partial molecule is re-encoded at every step. For a 44-atom molecule this amounts to 44 backbone passes for KRONOS against 990 for NEAT, offsetting NEAT's lower number of denoiser evaluations. QUETZAL is causal and could in principle cache, but its released implementation re-encodes the sequence at each step; even with caching, its second-order sampler at 120 steps requires 239 denoiser evaluations per atom against KRONOS's 100.

At its default setting of 100 DDIM steps, KRONOS samples in 1110 ± 47 s, comparable to NEAT (1156 ± 77 s) and $4.8\times$ faster than QUETZAL (5273 ± 309 s). Figure 7 reports quality against sampling cost for all three. Fewer steps scale denoiser work down proportionally while backbone cost is fixed by the atom count: at 25 steps KRONOS takes 333 ± 6 s, $3.5\times$ faster than NEAT and $15.9\times$ faster than QUETZAL, with stability essentially unchanged (3D AtomStable 86.2 vs 86.5 (%); 3D MolStable 2.64 vs 2.71 (%)).

A.9 2D MOLECULE GENERATION

Because KRONOS operates in the unified latent space of the UAE, generated latent tokens can be decoded into both 3D molecular structures and explicit 2D molecular graphs (utilizing predicted bond topology). This enables evaluation of the corresponding 2D molecules and direct comparison with prior methods operating on mixed 2D/3D molecular representations (Vignac et al., 2023; Huang et al., 2024; Irwin et al., 2024). We evaluate 10k generated molecules per model and report the resulting 2D metrics in Table 5. Overall, KRONOS achieves competitive 2D generation performance across both benchmarks. While SemlaFlow attains the strongest overall 2D metrics, KRONOS consistently generates highly valid and stable molecular graphs.

Table 5: 2D molecule generation metrics. We report metrics on the test sets of the corresponding datasets in gray. Models marked with (*) were evaluated using released generated molecules. Models marked with (**) were evaluated using official pre-trained weights and source code. **Best** and second-best entries are highlighted in bold and underlined, respectively. We report the mean ± std over three independent runs.

Model	2D Validity (↑) (%)	2D AtomStable (↑) (%)	2D MolStable (↑) (%)	V&C (↑) (%)	V&U (↑) (%)	V&U&N (↑) (%)
QM9	100.0	100.00	100.00	100.0	99.9	99.9
Diffusion / Flow						
JODO*	98.1	99.83	97.83	98.1	**95.1**	**95.1**
MiDi**	$96.8_{\pm 0.1}$	$99.74_{\pm 0.01}$	$96.33_{\pm 0.13}$	$96.7_{\pm 0.1}$	$94.4_{\pm 0.1}$	$94.3_{\pm 0.1}$
SemlaFlow**	$\mathbf{99.4}_{\pm 0.1}$	$\mathbf{100.00}_{\pm 0.00}$	$\mathbf{99.63}_{\pm 0.06}$	$\mathbf{99.3}_{\pm 0.1}$	$\underline{94.8}_{\pm 0.1}$	$\underline{94.8}_{\pm 0.1}$
Autoregressive						
KRONOS (ours)	$\underline{99.2}_{\pm 0.2}$	$\underline{99.92}_{\pm 0.02}$	$\underline{99.10}_{\pm 0.02}$	$\underline{99.1}_{\pm 0.2}$	$92.5_{\pm 0.6}$	$92.5_{\pm 0.6}$
GEOM	89.4	99.44	86.27	89.4	18.5	18.5
Diffusion / Flow						
JODO*	79.1	<u>99.26</u>	80.39	75.6	78.1	78.1
MiDi**	$59.9_{\pm 0.2}$	$96.43_{\pm 0.11}$	$72.85_{\pm 0.32}$	$57.9_{\pm 0.2}$	$59.9_{\pm 0.2}$	$59.9_{\pm 0.2}$
SemlaFlow**	$\mathbf{95.1}_{\pm 0.1}$	$\mathbf{100.00}_{\pm 0.00}$	$\mathbf{98.63}_{\pm 0.11}$	$\mathbf{92.4}_{\pm 0.2}$	$\mathbf{95.1}_{\pm 0.2}$	$\mathbf{95.1}_{\pm 0.2}$
Autoregressive						
KRONOS (ours)	$\underline{84.6}_{\pm 0.4}$	$99.16_{\pm 0.02}$	$\underline{83.45}_{\pm 0.29}$	$\underline{84.4}_{\pm 0.4}$	$\underline{84.5}_{\pm 0.5}$	$\underline{84.5}_{\pm 0.5}$

A.10 SCALING ABLATIONS

Tables 6 and 7 evaluate the effect of scaling the transformer backbone while keeping the diffusion head fixed at dimension 1536. We vary the number of layers, attention heads, and hidden dimension. Across both QM9 and GEOM-Drugs, increasing model capacity consistently improves or maintains generation performance.

Table 6: Model ablations results on the QM9 dataset. We report mean ± std across three independent runs. Best in **bold**, second-best underlined. The best configuration is highlighted in blue.

Transformer heads	Transformer layers	Transformer dimension	Model parameters ($\times 10^6$)	3D Validity	3D AtomStable	3D MolStable
8	8	512	105.27	$95.17_{\pm 0.04}$	$98.55_{\pm 0.01}$	$88.7_{\pm 0.2}$
10	10	640	129.61	$\underline{95.90}_{\pm 0.11}$	$98.67_{\pm 0.04}$	$89.5_{\pm 0.3}$
12	12	768	165.78	$95.87_{\pm 0.08}$	$\underline{98.73}_{\pm 0.01}$	$\underline{90.1}_{\pm 0.1}$
16	16	1024	283.05	$\mathbf{97.29}_{\pm 0.13}$	$\mathbf{99.07}_{\pm 0.02}$	$\mathbf{92.6}_{\pm 0.1}$

Table 7: Model scaling ablations results on the GEOM-Drugs dataset. We report mean ± std across three independent runs. Best in **bold**, second-best underlined. The best configuration is highlighted in blue.

Transformer heads	Transformer layers	Transformer dimension	Model parameters ($\times 10^6$)	3D Validity	3D AtomStable	3D MolStable
8	8	512	105.35	$93.2_{\pm 0.1}$	$86.42_{\pm 0.11}$	$\mathbf{2.73}_{\pm 0.08}$
10	10	640	129.71	$93.1_{\pm 0.2}$	$86.23_{\pm 0.05}$	$2.69_{\pm 0.05}$
12	12	768	165.90	$\underline{93.4}_{\pm 0.2}$	$\underline{86.43}_{\pm 0.06}$	$2.64_{\pm 0.11}$
16	16	1024	283.21	$\mathbf{94.7}_{\pm 0.1}$	$\mathbf{86.45}_{\pm 0.03}$	$\underline{2.71}_{\pm 0.13}$

A.11 EFFECT OF FRAGMENT-CONDITIONING PROBABILITY p_f

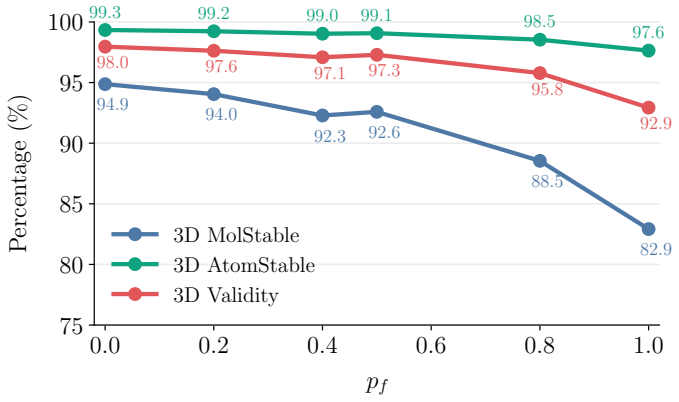

Figure 8: Effect of the fragment-conditioning probability p_f on unconditional molecule generation performance (QM9). Increasing p_f gradually trades unconditional generation quality for fragment-conditioning capability.

During training, each molecule is transformed to a fragment-conditioned sequence with probability p_f. Figure 8 reports unconditional generation performance on QM9 as p_f varies. We observe that model's performance on unconditional generation remains stable for moderate values of p_f, with performance degrading only as training becomes dominated by fragment-conditioned examples. Based on these observations, we select $p_f = 0.5$ throughout the main experiments as a reasonable trade-off between unconditional and fragment-conditioned generation.

A.12 GENERATED MOLECULE SIZE DISTRIBUTION

Figure 9 compares the distribution of the number of atoms in 10k generated molecules against 10k molecules sampled from the corresponding test sets. On QM9, the generated and test distributions exhibit identical median molecule sizes (18 atoms), while on GEOM-Drugs the generated molecules have a median of 42 atoms compared to 43 atoms in the test set. Overall, the learned stopping criterion produces molecule-size distributions that closely match those observed in the data, indicating that KRONOS effectively learns when to terminate generation.

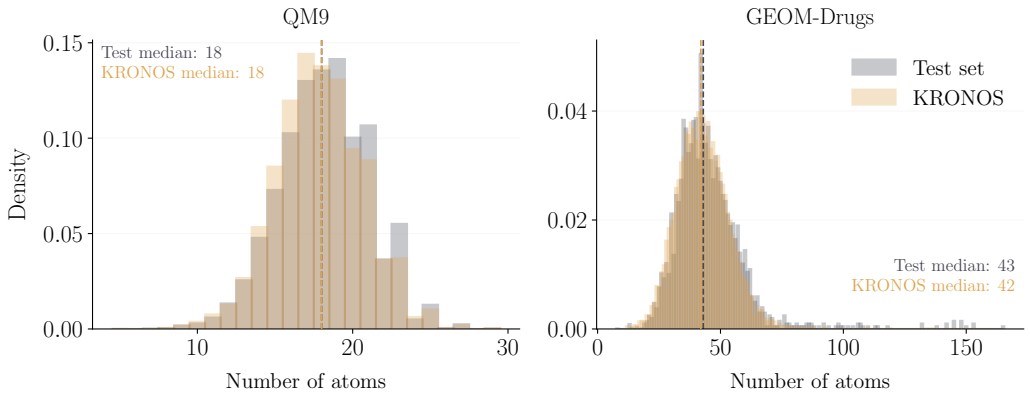

Figure 9: Distribution of the number of atoms in molecules generated by KRONOS and in the corresponding test sets for QM9 and GEOM-Drugs. KRONOS closely reproduces the molecule size distributions of both datasets.

A.13 LIMITATIONS AND FUTURE WORK

KRONOS inherits limitations from the underlying latent autoregressive formulation. First, generation quality is bounded by the reconstruction quality of a separately trained autoencoder (Luo et al., 2025). Improvements in latent representation learning may therefore translate directly into improved generative performance.

Second, while fragment-conditioned training substantially improves completion from partial molecular context, KRONOS is not permutation-invariant. Rose et al. (2025) obtain this property by sum pooling atom representations into a single vector, which they note may cause oversquashing. Future work may find better trade-offs between expressivity and permutation invariance.

Third, conditioning fragments are supplied with the atomic coordinates they occupy in a reference conformer, and KRONOS is sensitive to this placement: fragments positioned outside the distribution seen during training yield shorter and less reliable completions. We found that discarding absolute positions in favour of relative geometry alone is not a straightforward remedy: in our early experiments, encoding distance-based features instead of absolute coordinates degraded UAE reconstruction considerably. A potential direction, which we leave to future work, is to supply the missing placement rather than remove the dependence on it. A fragment specified without coordinates can be embedded with a standard conformer generator, and positioned by sampling a placement observed for a training fragment of comparable size. This would keep the encoder input in-distribution and enable conditioning on a fragment vocabulary without a reference molecule.

Fourth, our fragment decomposition relies on a fixed BRICS-derived partitioning scheme. While this provides chemically meaningful conditioning units, alternative decomposition strategies may expose different structural regularities, and learning the decomposition jointly with the generative model remains an open direction.

Finally, while we focus on unconditional and fragment-conditioned molecule generation, the proposed framework could be naturally extended to property-guided molecular design. In particular, future work could condition generation on a molecular fragment while simultaneously optimizing for desired molecular properties.

A.14 HYPERPARAMETERS

Tables 8 and 9 report the hyperparameters used to train KRONOS on QM9 and GEOM-Drugs, respectively.

Table 8: Hyperparameters of KRONOS trained on the QM9 dataset.

Parameter	**Value**
batch size	2048
bias	false
dropout	0.0
epochs	10,000
diffusion head hidden dimension	1,536
diffusion head layers	6
diffusion noise schedule	cosine
number of diffusion steps	100
sampler	DDIM
learning rate, weight decay	1.0×10^{-4}, 10^{-6}
learning rate decay number of epochs (cosine annealing scheduling)	10,000
learning rate minimum (cosine annealing scheduling)	10%
learning rate warm-up number of epochs (cosine annealing scheduling)	50
time step resampling	4
transformer heads	16
stop classifier threshold τ	0.5
transformer hidden dimension	1024
transformer layers	16
stop classifier weight λ_{stop}	1.0
fragment conditioning probability p_f	0.5

Table 9: Hyperparameters of KRONOS trained on the GEOM-Drugs dataset.

Parameter	**Value**
batch size	512
bias	false
dropout	0.0
epochs	1,000
diffusion head hidden dimension	1,536
diffusion head layers	6
diffusion noise schedule	cosine
number of diffusion steps	100
sampler	DDIM
learning rate, weight decay	1.0×10^{-4}, 10^{-6}
learning rate decay number of epochs (cosine annealing scheduling)	1,000
learning rate minimum (cosine annealing scheduling)	10%
learning rate warm-up number of epochs (cosine annealing scheduling)	10
time step resampling	4
transformer heads	16
transformer hidden dimension	1024
transformer layers	16
stop classifier threshold τ	0.5
stop classifier weight λ_{stop}	1.0
fragment conditioning probability p_f	0.5

A.15 GENERATED MOLECULES

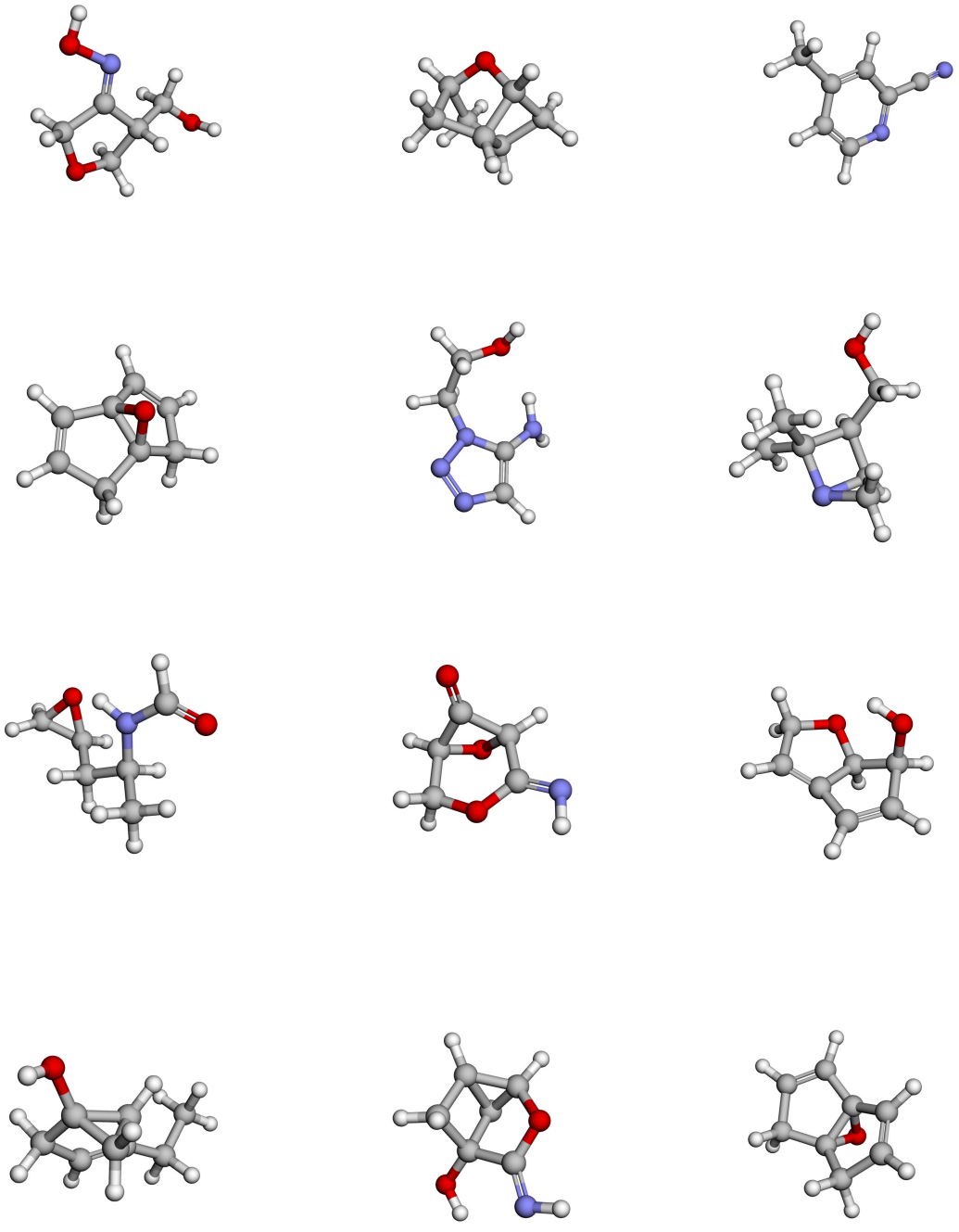

Figure 10: Examples of 3D molecules generated by KRONOS trained on QM9 dataset.

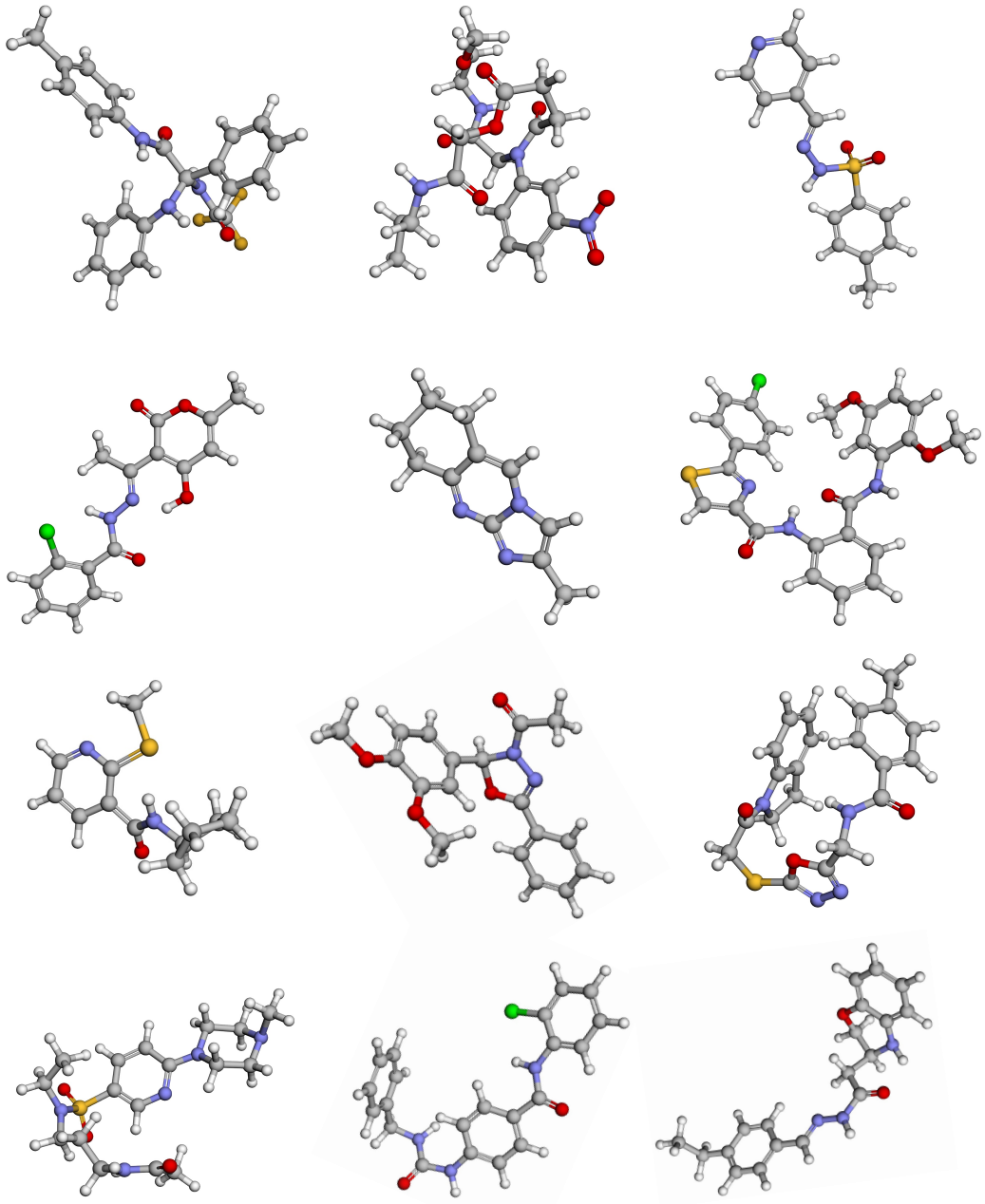

Figure 11: Examples of 3D molecules generated by KRONOS trained on GEOM-Drugs dataset.

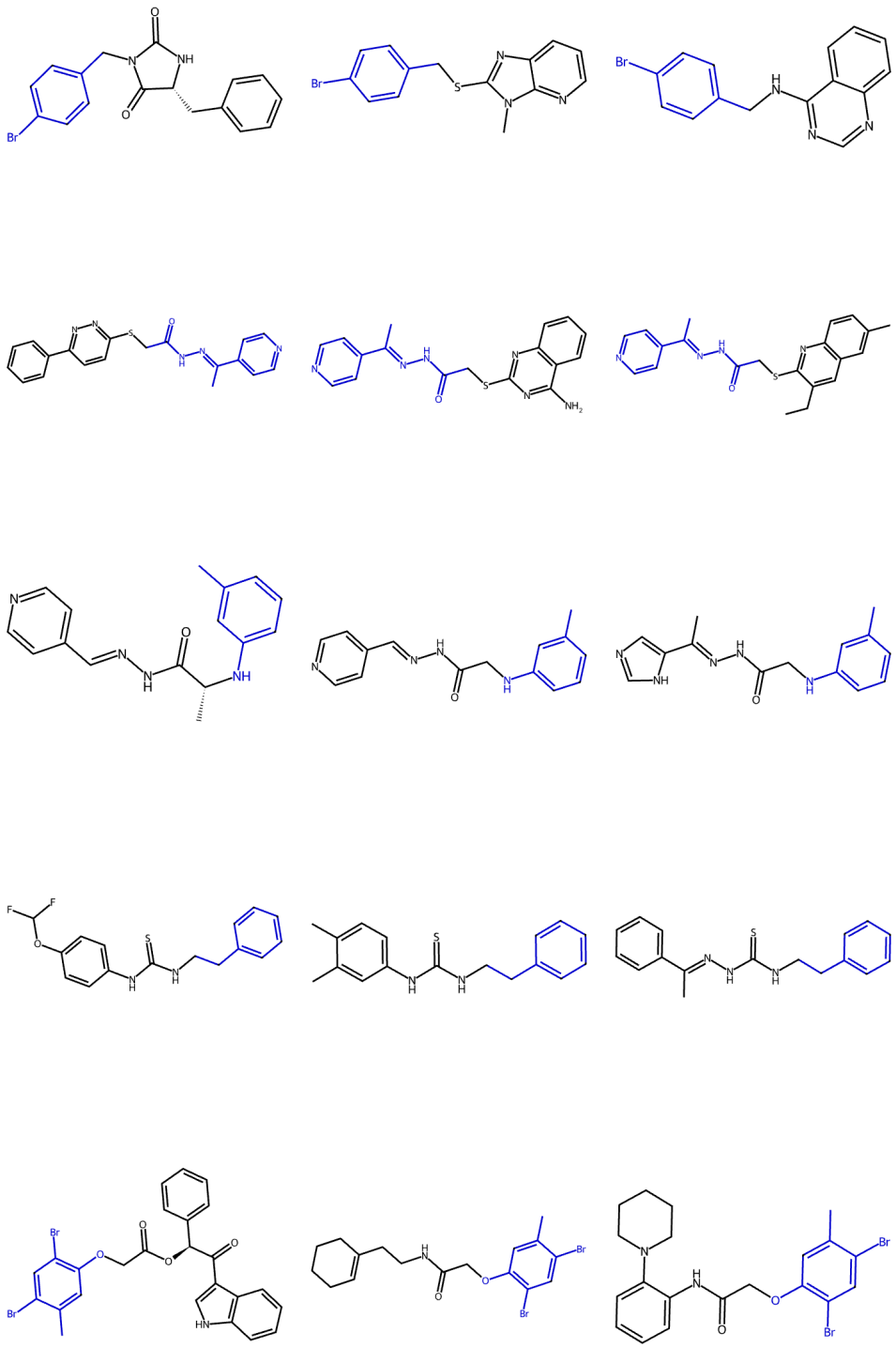

Figure 12: Fragment-conditioned completions generated by KRONOS on GEOM-Drugs test molecules. Each row shows three independent completions of the same conditioning fragment; the prefix fragment is shown in blue.