VEDA: Generation of 3D Molecules via Variance-Exploding Diffusion with Annealing

Peining Zhang[1], **Jinbo Bi**[1], **Minghu Song**[2]

[1]University of Connecticut, Storrs, Connecticut 06269, USA
[1]Institute of Health and Medicine, Hefei Comprehensive National Science Center, Hefei 230601, China
peining.zhang@uconn.edu, jinbo.bi@uconn.edu, minghu.song@ihm.ac.cn

Abstract

Diffusion models show promise for 3D molecular generation, but face a fundamental trade-off between sampling efficiency and conformational accuracy. While flow-based models are fast, they often produce geometrically inaccurate structures, as they have difficulty capturing the multimodal distributions of molecular conformations. In contrast, denoising diffusion models are more accurate but suffer from slow sampling, a limitation attributed to sub-optimal integration between diffusion dynamics and SE(3)-equivariant architectures. To address this, we propose **VEDA**, a unified SE(3)-equivariant framework that combines variance-exploding diffusion with annealing to efficiently generate conformationally accurate 3D molecular structures. Specifically, our key technical contributions include: (1) a VE schedule that enables noise injection functionally analogous to simulated annealing, improving 3D accuracy and reducing relaxation energy; (2) a novel preconditioning scheme that reconciles the coordinate-predicting nature of SE(3)-equivariant networks with a residual-based diffusion objective, and (3) a new arcsin-based scheduler that concentrates sampling in critical intervals of the logarithmic signal-to-noise ratio. On the QM9 and GEOM-DRUGS datasets, VEDA matches the sampling efficiency of flow-based models, achieving state-of-the-art valency stability and validity with only 100 sampling steps. More importantly, VEDA's generated structures are remarkably stable, as measured by their relaxation energy (ΔE_{relax}) during GFN2-xTB optimization. The median energy change is only 1.72 kcal/mol, significantly lower than the 32.3 kcal/mol from its architectural baseline, SemlaFlow. Our framework demonstrates that principled integration of VE diffusion with SE(3)-equivariant architectures can achieve both high chemical accuracy and computational efficiency.

Code — https://github.com/peiningzhang/VEDA
Extended version — https://arxiv.org/abs/2511.09568

Introduction

Deep generative models like AlphaFold3 (Abramson et al. 2024) and RFdiffusion (Watson et al. 2023) are revolutionizing computational drug discovery. Diffusion models have emerged as a dominant tool for generating 3D molecular conformations, as their ability to model continuous generative processes—transforming random noise into structured data—aligns naturally with atomic coordinates. By combining these processes with equivariant architectures (e.g., EGNNs (Satorras, Hoogeboom, and Welling 2021)), these models generate geometrically realistic and chemically valid structures (Hoogeboom et al. 2022).

In recent years, numerous diffusion models have been developed for 3D molecular generation or 3D molecular conformation generation. While most research focuses on architecture (e.g., equivariant transformers (Liao and Smidt 2023)) or domain-specific constraints (e.g., torsion angles (Jing et al. 2022)), the principled design of the diffusion process remains under-explored (Zhang et al. 2025). A key challenge is aligning the learning objective with the inductive biases of SE(3)-equivariant networks, which tend to learn identity mappings due to their message-passing mechanism. Preconditioning (Karras et al. 2022) provides a framework to correct this bias and improve stability, yet it remains largely overlooked in molecular generation.

To address this gap, we propose VEDA, a principled framework for 3D molecular diffusion where:

- To the best of our knowledge, VEDA is the first framework to apply the Variance-Exploding (VE) diffusion paradigm to the hybrid discrete-continuous domain of 3D molecules, unifying atomic types (discrete) and coordinates (continuous) within a single diffusion process that is functionally analogous to simulated annealing.
- We propose a theoretically grounded preconditioning scheme to correct the inductive bias of coordinate-predicting SE(3) equivariant networks. As a complementary innovation in diffusion dynamics, we introduce a noise schedule based on the arcsine function, which achieves a better balance between early-stage exploration and late-stage refinement in molecular generation.
- VEDA substantially reduces relaxation energy of generated molecules, by 90% compared to those generated by SemlaFlowa (Irwin et al. 2025), while achieving state-of-the-art performance on QM9 and GEOM-DRUGS, and matching the efficiency of strong flow-based models with significantly fewer sampling steps.

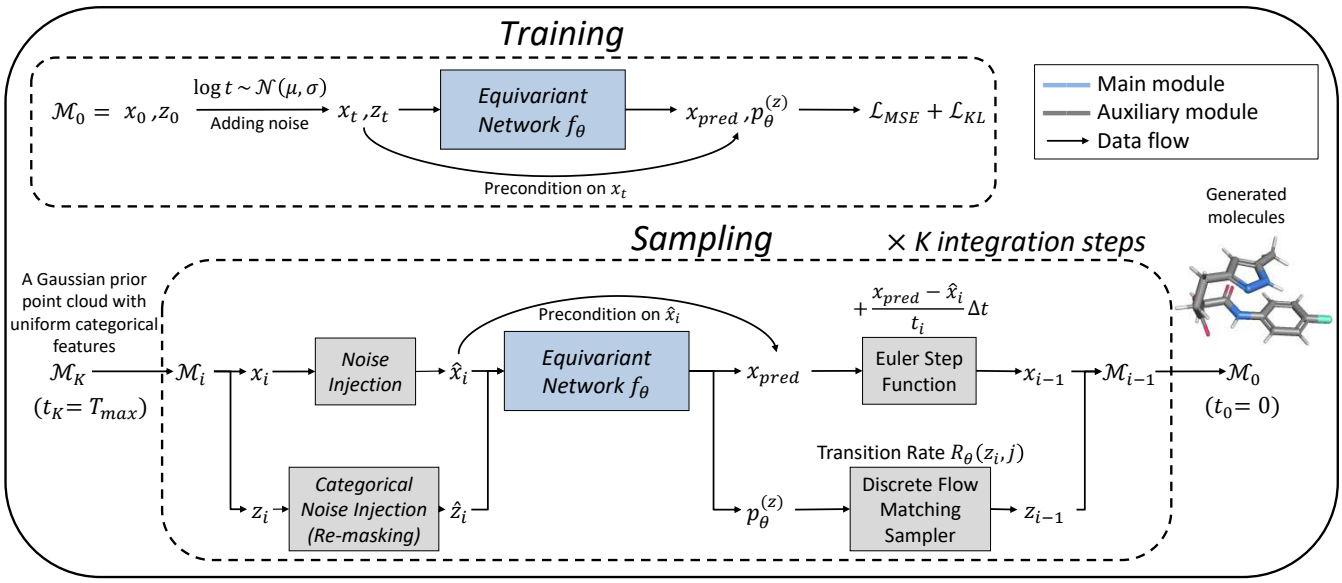

Figure 1: An overview of the VEDA framework, detailing its training and sampling processes. During **Training** (top), a clean molecule $\mathcal{M}_0$ is perturbed via Gaussian noise for coordinates ($\mathbf{x}_t$) and categorical masking for features ($\mathbf{z}_t$), defined by Eq. 1 and Eq. 3. The equivariant network f_θ is then trained to predict the original molecule by minimizing the combined Mean Squared Error (MSE) and Cross-Entropy (CE) loss in Eq. 12. During **Sampling** (bottom), the process starts from a pure noise distribution (a Gaussian point cloud with uniform categorical features) and iteratively refines the sample over K steps. Each integration step i involves: (1) a noise injection from $(\mathbf{x}_i, \mathbf{z}_i)$ to $(\hat{\mathbf{x}}_i, \hat{\mathbf{z}}_i)$; (2) the network f_θ is applied to both coordinates and features, with preconditioning affecting only the coordinate predictions; it outputs x_{pred} and the category probabilities $p_\theta^{(z)}$ (Eq. 6); and (3) an update combining a continuous Euler step and a Discrete Flow Matching sampler to obtain $\mathcal{M}_{i-1}$, formally given in Eq. 13 and Eq. 14. In the diagram, blue boxes represent the main network module, gray boxes are auxiliary operations, and arrows indicate the data flow.

Related Works

Diffusion-Based Generative Models Denoising Diffusion Probabilistic Models (DDPMs) (Ho, Jain, and Abbeel 2020) have become a leading class of generative models which learn to reverse a forward process that incrementally corrupts data with Gaussian noise. DDPMs are extended to continuous-time settings using score-based stochastic differential equations (SDEs) (Song et al. 2021), incorporating either variance-preserving or variance-exploding dynamics. Recent advances, particularly those utilizing Variance-Exploding (VE) SDEs with preconditioning techniques (Karras et al. 2022), have substantially enhanced generation quality and sampling efficiency. An alternative approach, Flow Matching (Lipman et al. 2023), directly learns the velocity field of the data distribution's probability flow, offering a different yet highly competitive approach to generative modeling. Consistent with recent literature (Gao et al. 2024), we consider flow matching a specialized formulation within the broader diffusion framework.

Discrete Diffusion Models The principles of diffusion have also been extended to discrete data generation, drawing inspiration from masked language modeling (Devlin et al. 2019). These non-autoregressive methods are a natural fit for the unordered structure of molecules, enabling de novo generation without imposing artificial sequential or ordering assumptions. D3PM (Austin et al. 2021) established the theoretical foundations for discrete diffusion using continuous-time Markov chains. Subsequent developments of variants of Discrete Flow Matching (Campbell et al. 2024; Gat et al. 2024) have further advanced the field with improved sampling strategies and scalability.

***De Novo* 3D Molecular Generation** Modern 3D molecular generation heavily relies on E(3)-equivariant neural networks, from pioneering architectures like EGNN (Satorras, Hoogeboom, and Welling 2021) to subsequent transformer-based models (Zhang, Chen, and Chu 2025). From a probabilistic modeling perspective, the field is dominated by a fundamental trade-off. Denoising diffusion-based methods like EDM (Hoogeboom et al. 2022) and GeoLDM (Xu et al. 2023) achieve high conformational accuracy but suffer from slow sampling. Conversely, flow-based models such as EquiFM (Song et al. 2023a), GeoBFN (Song et al. 2023c), and SemlaFlow (Irwin et al. 2025) are efficient but often struggle with geometric precision. This challenge is underscored by recent benchmark reports, which highlight the limitations of flow-based approaches in generating chemically accurate structures (Nikitin et al. 2025). Balancing high fidelity with computational efficiency remains one of the most pressing challenges in 3D molecular generation.

Methodology

Preliminaries

We represent molecules as point clouds of chemical elements in 3D space, denoted by $\mathcal{M} = (\mathbf{x}, \mathbf{z})$, where $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_N) \in \mathbb{R}^{N\times 3}$ are the atomic coordinates of N atoms, and $\mathbf{z}$ denotes non-geometric molecular attributes. To keep notation unified, we decompose $\mathbf{z} = (\mathbf{h}, \mathbf{b})$, where $\mathbf{h}$ includes atom-level properties (e.g., atom types, charges) and $\mathbf{b}$ denotes pairwise bond information.

In most graph-based models, a molecule is further represented by a graph $G = (V, E)$, where nodes V represent atoms and edges E represent either inferred or explicitly defined interactions between atoms. These interactions can be constructed based on distance thresholds or chemical bond annotations. Some models treat bond information as part of the output (explicit bond modeling), while others infer it post hoc based on generated atomic coordinates (implicit bond inference). This distinction influences the architecture and training strategy of generative models, as detailed later.

Model Architecture

To show our VEDA framework is both general and scalable, we apply it to two architectures at different levels of complexity. This dual implementation proves VEDA's effectiveness across different modeling paradigms, covering both implicit and explicit bond generation.

VEDA-E: Implicit Bond Modeling The VEDA-E variant is built upon the EGNN architecture (Satorras, Hoogeboom, and Welling 2021) from EDM (Hoogeboom et al. 2022). In this setup, the model generates only node-centric features $\mathbf{h} = (\mathbf{h}_1, \ldots, \mathbf{h}_N) \in \mathbb{R}^{N\times d_h}$, which include atom types and charges. Diffusion is performed by adding Gaussian noise to all continuous features. The chemical bond structure $\mathbf{b}$ is not explicitly modeled during generation; instead, it is inferred post hoc based on interatomic distances and chemical valence rules, a standard practice in EDM-based molecule generation.

VEDA-S: Explicit Bond Modeling In contrast, VEDA-S adopts the Semla architecture from SemlaFlow (Irwin et al. 2025) to explicitly model bonds. This variant directly generates a complete molecular graph representation $\mathbf{z} = (\mathbf{h}, \mathbf{b})$, where $\mathbf{h}$ represents node features and $\mathbf{b} \in \mathbb{R}^{N\times N\times d_b}$ denotes the bond types between atoms. This approach is compatible with discrete data types by using a mask-based diffusion process and a classification objective. A diagram illustrating the main concept of VEDA-S is provided in Figure 1.

Diffusion Dynamics

Our generative model is based on Score Diffusion (Song et al. 2021). It includes a forward diffusion process gradually perturbing the clean data into noise, and a learned denoising process that recover the noise into data. The denoising process is governed by the denoiser D_θ, which takes as input the noisy sample $\mathbf{x}_t$ and the noise level t. We define $\mathbf{x}_0$ as the clean data point (e.g., molecular coordinates) and $\mathbf{x}_\infty$ as the fully noised sample.

Forward Process We corrupt the continuous coordinates $\mathbf{x}$ by adding Gaussian noise:

$$\mathbf{x}_t = \mathbf{x}_0 + t\,\epsilon, \quad \epsilon\sim\mathcal{N}(0, I), \tag{1}$$

The noise level t is sampled from a log-normal distribution: $t \sim \text{LogNormal}\left(\ln\sqrt{T_{\min}T_{\max}}, [\frac{1}{8}\ln(T_{\max}/T_{\min})]^2\right)$, so $t \in [T_{\min}, T_{\max}])$ in most case.

For discrete features $\mathbf{z}$, we consider two variants.

In **VEDA-E**, Gaussian noise is applied:

$$\mathbf{z}_t = \mathbf{z}_0 + t\,\eta, \quad \eta \sim \mathcal{N}(0, I). \tag{2}$$

In **VEDA-S**, we use time-dependent categorical masking:

$$\mathbf{z}_t \sim \text{Cat}\left[(1 - m(t))\,\delta(\mathbf{z}_t = \mathbf{z}_0) + \frac{m(t)}{S}\right], \tag{3}$$

where S is the number of categories, and the masking rate function $m(t) = (\ln t - \ln T_{\min})/(\ln T_{\max} - \ln T_{\min})$ aligns the discrete corruption schedule with the continuous noise schedule, reaching uniform distribution at $t = T_{\max}$.

Our approach uses this principled VE scheduler during training. This variance-exploding (VE) formulation is particularly suitable for modeling continuous data such as 3D molecular structures, as it enables flexible noise schedules and accurate noise level control in log-space, injecting a massive amount of noise to reach lower energy. We believe infinite separation at $t \to \infty$ between atoms is an ideal noisy state for molecule generation, superior to a standard Gaussian prior, as it corresponds to no chemical interactions between atoms.

Preconditioning the Denoiser In 3D molecular generation with diffusion models, directly predicting coordinates (x-prediction) is the prevailing approach, as equivariant network architectures like GNNs are not well-suited for predicting noise or velocity (Irwin et al. 2025). We adopt the preconditioning framework from Karras et al. (2022), where the denoiser D_θ is defined as a combination of the input $\mathbf{x}_t$ and a neural network output F_θ:

$$D_\theta(\mathbf{x}_t; t) = c_{\text{skip}}\mathbf{x}_t + c_{\text{out}}\left[F_\theta(c_{\text{in}}\mathbf{x}_t; c_{\text{noise}})\right] \tag{4}$$

$$F_{\text{target}} = \frac{1}{c_{\text{out}}}(\mathbf{x}_0 - c_{\text{skip}}(\mathbf{x}_0 + t\,\epsilon)) \tag{5}$$

Optimizing the preconditioning framework with a coordinate-space MSE loss requires the network (F_θ) to output residuals that are uncorrelated with the input. However, our backbone (GNN or Graph Transformer) outputs coordinates via a residual connection design (He et al. 2016), which causes F_θ to produce outputs highly correlated with $\mathbf{x}_t$, conflicting with the residual learning objective.

To address this mismatch, we subtract a scaled identity component from the network output. Instead of forcing the network to suppress its inherent bias through the loss function, we explicitly subtract the undesired identity component from its output. Our modified denoiser is:

$$D_\theta(\mathbf{x}_t; t) = c_{\text{skip}}\mathbf{x}_t + c_{\text{out}}\left(F_\theta(c_{\text{in}}\mathbf{x}_t; c_{\text{noise}}) - \alpha_t\, c_{\text{in}}\,\mathbf{x}_t\right) \tag{6}$$

The coefficient α_t is our main contribution in this section. It serves as the optimal linear predictor of the ground-truth noise $\frac{\sigma_d \boldsymbol{\epsilon}}{\sqrt{\sigma_d^2+t^2}}$ from F_{target}, where σ_d is the standard deviation of the data distribution (Karras et al. 2022). This corresponds to the Linear Minimum Mean Squared Error (LMMSE) solution, which yields the following closed-form coefficient:

$$\alpha_t = \arg\min_{\alpha} \mathbb{E}[\| \frac{\sigma_d \boldsymbol{\epsilon}}{\sqrt{\sigma_d^2 + t^2}} - \alpha_t \, c_{\text{in}} \, \mathbf{x}_t \|^2] \tag{7}$$

$$= \arg\min_{\alpha} \mathbb{E}[\| \sigma_d \boldsymbol{\epsilon} - \alpha_t \mathbf{x}_t \|^2] \tag{8}$$

$$= \frac{\text{Cov}(\sigma_d \boldsymbol{\epsilon}, \mathbf{x}_t)}{\text{Var}(\mathbf{x}_t)} = \frac{\sigma_d \, t}{\sigma_d^2 + t^2} \tag{9}$$

Using this optimal α_t reduces the residual correlation with the input and better aligns the training objective with the network's inductive bias, i.e., its preference for modeling absolute coordinates.We use the standard definitions of c_{skip}, c_{out}, c_{in}, c_{noise}, and λ_σ from Karras et al. (2022).

Training Objectives

For VEDA-E, we apply a mean squared error (MSE) loss to both continuous coordinates and categorical features:

$$\mathcal{L}_{\text{MSE}} = \mathbb{E}_{\mathbf{x}_0, \mathbf{z}_0, \epsilon, t} \left[\| D_\theta(\mathbf{x}_t, \mathbf{z}_t; t) - (\mathbf{x}_0, \mathbf{z}_0) \|^2 \right], \tag{10}$$

where $(\mathbf{x}_0, \mathbf{z}_0)$ represents the concatenated true coordinates and categorical features. Following the implementation of EDM (Hoogeboom et al. 2022), VEDA-E treats categorical features $\mathbf{z}_0$ as continuous during denoising and applies MSE loss, despite their inherently discrete nature..

For VEDA-S, the model predicts the original discrete features $\mathbf{z}_0$ via a categorical distribution $\hat{p}_\theta(\mathbf{z}_0 \mid \tilde{\mathbf{z}}_t, t)$. This is optimized using a KL divergence loss:

$$\mathcal{L}_{\text{KL}} = \mathbb{E}_{\mathbf{z}_0, \tilde{\mathbf{z}}_\mathbf{t}, t} \Big[\text{KL}(\delta(\mathbf{z}_0) \,\|\, \hat{p}_\theta(\mathbf{z}_0 \mid \tilde{\mathbf{z}}_\mathbf{t}, t)) \Big]. \tag{11}$$

The overall loss combines continuous and discrete terms:

$$\mathcal{L}_{\text{VEDA-S}} = \lambda_{\text{cont}} \, \mathcal{L}_{\text{MSE}} \; + \; \lambda_{\text{disc}} \, \mathcal{L}_{\text{KL}} \tag{12}$$

where λ_{cont} and λ_{disc} balance the two objectives.

Sampling

We adopt the standard denoising framework where the generative process iteratively refines noisy inputs into data samples, following a predefined SDE formulation (Karras et al. 2022). Conceptually, this process is analogous to *simulated annealing*. The noise level t acts as a temperature parameter that is gradually lowered, enabling broad exploration of the energy landscape at high noise and convergence to low-energy, stable conformations as noise decreases. Our model operates under the x-prediction parameterization to directly estimate the clean sample from its noisy counterpart. During sampling, VEDA-E applies continuous noise injection to all features, while VEDA-S uses continuous noise injection for coordinates and discrete token refinement for categorical features. We now describe three interlocking sampling components: (1) continuous denoising with amplified noise injection, (2) discrete masked-token refinement, and (3) the proposed arcsin noise scheduler.

Stochastic Annealing for Continuous Coordinates Our sampling process employs an amplified noise injection strategy. This approach is governed by a hyperparameter, $\gamma > 0$, which controls the degree of noise amplification. This process is equivalent in expectation to Gaussian smoothing on the molecular potential energy surface with a Gaussian kernel bandwidth of $\sqrt{\gamma^2 + 2\gamma} \cdot t_i$. This pronounced smoothing suppresses local minima and surface roughness, making it easier for the sample trajectory to find the global low-energy basin and thereby improving the final energy metric (Miao, Feher, and McCammon 2015).

Sampling proceeds in two substeps for each iteration i:

1. **Perturbation**: We first set the value of $\hat{t}_i = (1+\gamma)t_i$ inject amplified noise into the current sample $\mathbf{x}_i$ to reach an intermediate state $\hat{\mathbf{x}}_i = \mathbf{x}_i + \sqrt{\hat{t}_i^2 - t_i^2} \cdot \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$. The noise added is substantially greater than in standard DDPM (Ho, Jain, and Abbeel 2020) or Flow-Matching (Lipman et al. 2023) schedules.
2. **Denoising & Update**: The denoiser $D_\theta(\hat{\mathbf{x}}_i; \hat{t}_i)$ predicts the clean structure, from which we extrapolate the next sample $\mathbf{x}_{i+1}$ via a Euler step:

$$\mathbf{x}_{i+1} = \hat{\mathbf{x}}_i + (t_{i+1} - \hat{t}_i)(\hat{\mathbf{x}}_i - D_\theta(\hat{\mathbf{x}}_i; \hat{t}_i))/\hat{t}_i. \tag{13}$$

In VEDA-E, this two-step procedure is applied uniformly to both coordinates and categorical features. In VEDA-S, only the coordinates undergo amplified noise injection, while categorical features are handled separately via discrete token refinement.

Masked Token Refinement for Discrete Variables For discrete features in VEDA-S, we use masked token refinement process. This process is governed by the time-dependent mask rate function $m(t)$, which is designed to align the discrete corruption with the continuous noise schedule. Specifically, when injecting noise from noise level t_1 to t_2, each token is randomly re-masked with probability $\frac{m(t_2)-m(t_1)}{1-m(t_1)}$. Additionally, we found that a simple uniform random re-masking strategy consistently outperformed more complex confidence-based strategies proposed in prior work (Nie et al. 2025). We included our comparison with the strategies of low prediction confidence (Chang et al. 2022) or small probability margins (Kim et al. 2025) .

For discrete sampling, our approach is a step-wise masked sampling algorithm based on Continuous-Time Markov Chains (CTMC) within the Discrete Flow Matching framework. We implement and compare two complementary sampling strategies based on this foundation.

The full transition rate combines base interpolation dynamics with detailed balance corrections:

$$R_\theta(z_t, j) = \omega(t) \cdot p_\theta(z_0 = j | z_t) + \eta_t \cdot p_\theta(z_0 = z_t | z_t) \tag{14}$$

where $\omega(t) = \frac{\eta_t S(1-m(t)) + \eta_t m(t) + m'(t)}{m(t)}$ is the time-dependent scaling factor, and η_t is based on the categorical noise hyperparameter η. To determine the optimal configuration, we performed a comprehensive grid search. The results confirmed that the variable setting from Campbell et al.

(2024), where $\eta_t = \frac{\eta}{m'(t)}$, achieves the best and most robust performance, outperforming both the fixed setting ($\eta_t = \eta$) and the simplified DFM formulation from Gat et al. (2024). The transition rates are thus derived to optimally combine the model's confidence in clean data predictions with the stability of maintaining current token assignments

Proposed Noise Scheduler We propose an **arcsin-based noise scheduler** that further improves the sampling process. Through empirical analysis, we observe that sampling steps corresponding to near-zero log signal-to-noise ratio (log-SNR), i.e. $\log(\text{signal variance/noise variance})$ are especially critical for final molecular structure formation. It aligns well with the distribution of $t \sim \text{LogNormal}\left(\ln\sqrt{T_{\min}T_{\max}}, [\frac{1}{8}\ln(T_{\max}/T_{\min})]^2\right)$ used during training. To leverage this observation, let $u = i/N$ is the normalized step index for a total of N steps. We design an arcsin-shaped scheduler parameterized by a tunable scalar ρ. Our scheduler is defined as:

$$w(u) = (1-\rho)\,u \;+\; \rho\,\frac{2}{\pi}\arcsin(\sqrt{u}) \tag{15}$$

$$t^{(i)} = T_{\min}\left(\frac{T_{\max}}{T_{\min}}\right)^{w(u)} \tag{16}$$

Where $\rho \in [0, \frac{\pi}{\pi-2}]$ modulates the concentration around log-SNR ≈ 0. When $\rho = 0$, the scheduler reduces to a log-uniform schedule commonly used in prior work; as ρ increases, more steps cluster in the mid-range. This targeted allocation of sampling steps improves structural fidelity and overall sample quality, especially in chemically sensitive configurations. As shown in Figure 2, the arcsin scheduler (ρ=2) closely matches the log-normal training distribution, further supporting its design rationale.

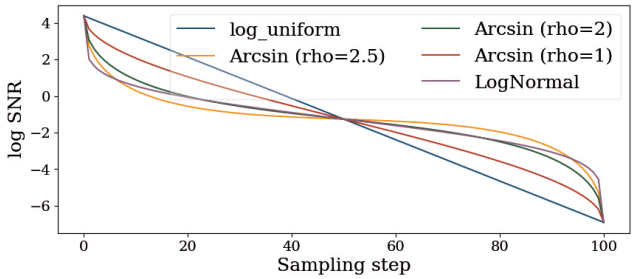

Figure 2: The arcsin sampling scheduler is proposed to focus on the middle part of sampling where log(SNR) is close to 0

Experiments

Experimental Setup

Model Variants To evaluate our VE-diffusion paradigm, we implement two variants based on distinct backbones. **VEDA-E** is built upon the EDM framework (Hoogeboom et al. 2022), using an EGNN backbone and performing implicit bond modeling. In contrast, **VEDA-S** derives from the SemlaFlow framework (Irwin et al. 2025), adopting a Semla backbone for explicit bond modeling.

Methods	NFE ↓	Atom Sta(%)↑	Mol Sta(%)↑	Valid (%)↑	Valid & Unique(%)↑
EDM	1000	98.7	82	91.9	90.9
GeoLDM	1000	98.9	89.4	93.8	92.6
UniGEM	1000	99.0	89.8	95	93.2
EquiFM	210	98.9	88.3	94.7	93.5
GeoBFN	2000	99.3	93.3	96.9	92.4
GOAT	90	99.2	-	92.9	92.0
VEDA-E	**50**	**99.4**	**93.7**	**98.1**	**97.9**
VEDA-E	**30**	99.2	91.5	97.0	96.8
MiDi	500	97.5	97.9	95.5	-
FlowMol2	100	99.9	99.6	99.5	-
EQGAT-diff	500	99.9	98.7	99.0	98.8
SemlaFlow	100	99.9	99.7	99.4	98.7
VEDA-S	100	**100.0**	**99.9**	**99.6**	**98.9**
VEDA-S	50	**100.0**	99.7	99.4	98.4
VEDA-S	30	99.9	99.5	99.5	98.9

Table 1: Results on QM9. *Top:* bond-implicit methods (bonds inferred from coordinates); *Bottom:* bond-explicit methods (bonds generated directly). NFE = number of function evaluations. Best in each group in **bold**.

Datasets and Metrics We evaluate our models on QM9 (Ramakrishnan et al. 2014) and GEOM-DRUGS (Axelrod and Gomez 2022) using distinct protocols. For QM9, we report standard metrics for chemical correctness: Atom Stability (Atom Sta.), Molecule Stability (Mol Sta.), Validity, and Unique percentage. For the more complex GEOM-DRUGS dataset, our main analysis followed the protocol proposed by Nikitin et al. (2025), which prioritizes physical realism, reporting the fixed version of molecular stability and validity alongside key metrics from GFN2-xTB geometry optimization, such as relaxation energy (ΔE_{relax}), post-optimization RMSD, and the structural difference after optimization. A supplementary analysis using the MMFF94 (Halgren 1996) force field to assess conformational energy and strain is detailed in the extended version.

Baselines We compare VEDA against leading 3D molecular generative models, grouped by bond treatment. For methods that do not explicitly generate bond structures, we consider EDM (Hoogeboom et al. 2022), GeoLDM (Xu et al. 2023), EquiFM (Song et al. 2023b), GeoBFN (Song et al. 2023c), UniGEM (Feng et al. 2024) and GOAT (Hong, Lin, and Tan 2024), all of which use equivariant architectures. For methods with explicit bond modeling, we compare to SemlaFlow (Irwin et al. 2025), FlowMol2 (Dunn and Koes 2024), and JODO (Huang et al. 2023), EQGAT-diff (Le et al. 2024) and Megalodon (Reidenbach et al. 2025).

Main Results

Results on QM9 As shown in Table 1, our model VEDA achieves state-of-the-art performance on the QM9 dataset. In the bond-implicit setting, VEDA-E achieves the highest valid and unique rate (97.9%) with only 50 NFE, significantly outperforming methods like EDM (90.9% at 1000 NFE) and GeoLDM (92.6% at 1000 NFE). In the bond-

Model	NFE	Time (s)	MS↑	V&C↑	Bond Length ↓ ($\times 10^{-2}$)	Bond Angles↓	Tor.↓	Median ΔE_{relax} ↓	Mean ΔE_{relax} ↓	Median RMSD↓	Mean RMSD↓
EQGAT	500	3468	0.899	0.834	1.00	1.15	8.58	6.40	11.1	0.915	0.975
JODO	500	673	0.963	0.879	0.77	0.83	6.01	4.74	7.04	-	-
Megalodon-quick	500	-	0.944	0.900	**0.66**	0.71	5.58	3.19	5.76	-	-
FlowMol2	100	126.3	0.944	0.746	1.30	1.62	15.0	17.9	24.3	0.992	1.038
SemlaFlow	100	150.3	0.969	0.920	3.10	2.06	6.05	32.3	91.0	0.273	0.358
Megalodon-flow	100	-	0.987	0.948	2.30	1.62	5.58	20.9	46.9	-	-
VEDA-S	100	164.8	**0.995**	**0.988**	0.69	**0.41**	**2.71**	**1.72**	**2.93**	**0.096**	**0.197**
VEDA-S	**50**	**84.1**	0.968	0.966	0.86	0.60	5.10	2.99	8.38	0.236	0.355

Table 2: Performance comparison on the GEOM-DRUGS dataset. Our model, VEDA-S (bottom), is compared against denoising-based (top) and flow-based (middle) methods. **Bold** values indicate the best result. Metrics include Molecular Stability (MS), Validity & Connectivity (V&C), and Mean Absolute Error (MAE) for bond lengths, angles, and torsions. All metrics are lower-is-better except for MS and V&C. Time (s) reports the average wall-clock time for generating one molecule. NFE is the Number of Function Evaluations. RMSD values for some models are missing as they were not in the original benchmark. For each model, 5000 molecules were evaluated, the full results with confidence intervals and sampling at other steps are included in extended version.

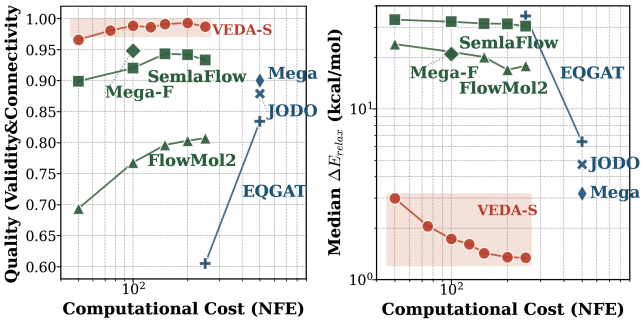

Figure 3: Trade-off between Generation Quality and Computational Cost. The figure compares our model, VEDA-S (red), against flow-based (green) and denoising-based (blue) models. The horizontal axis represents the computational cost, measured by the Number of Function Evaluations (NFE). (Left) Quality measured by molecule Validity and Connectivity, where higher values are better. (Right) Quality measured by the median energy difference (ΔE_{relax}), where lower values are better.

explicit setting, VEDA-S achieves nearly perfect atom stability and molecular stability, and a valid and unique score of 98.9% at just 100 NFE. Even at 30 NFE, it remains competitive with most baselines, highlighting its efficiency.

Results on GEOM-DRUGS We evaluate our model on the GEOM-DRUGS dataset using the protocol from Nikitin et al. (2025), with detailed metrics provided in Table 2. As the results show, our VEDA-S model achieves a new state-of-the-art performance across nearly all metrics. Notably, this is accomplished in just 100 function evaluations (NFE), significantly outperforming computationally expensive 500-step models like Megalodon, and surpassing previous flow-based approaches by a large margin in geometric accuracy.

This superior performance is not just in final quality but also in the trade-off between sampling efficiency and generation quality, as shown in Figure 3. The figure plots generation quality as a function of NFE. Quality is assessed by both validity and geometric stability, measured via median relaxation energy. The results are striking: VEDA-S (in red) consistently occupies the optimal region in both plots. It achieves the high efficiency of flow-based models (green) while delivering a geometric accuracy (lower ΔE_{relax}) that is an order of magnitude better than all competitors.

To further validate this comparison, we also tested noise-injected sampling for SemlaFlow , but observed no significant gains. However, this modification did not lead to significant improvements. This finding suggests that VEDA's superior performance arises not merely from the presence of noise, but from its systematic framework design. Even on saturated metrics like molecular stability, our model pushes the scores closer to 100%. Furthermore, even with just 50 steps, VEDA-S remains highly competitive, demonstrating its excellent trade-off between sampling efficiency and generation quality.

Ablation Studies

Component Ablation Table 3 presents an ablation study on the GEOM-DRUGS dataset. We start from the full VEDA model and progressively remove key components to assess their impact. Disabling preconditioning refers to setting α_t to be 0. It results in a small drop in validity and modest increases in RMSD and relaxation error, showing the stabilizing effect of preconditioning. Removing the noise injection leads to a substantial drop in all metrics, highlighting its importance in guiding the denoising process. Introducing the optimal transport (OT) alignment mechanism, as used in SemlaFlow (Irwin et al. 2025), increases validity but significantly worsens energy and RMSD. We attribute this to OT alignment violating the independence assumption between noise and original molecular coordinates, which causes the denoising direction to become overly correlated with the aligned noise. This results in a collapse of coordinate vari-

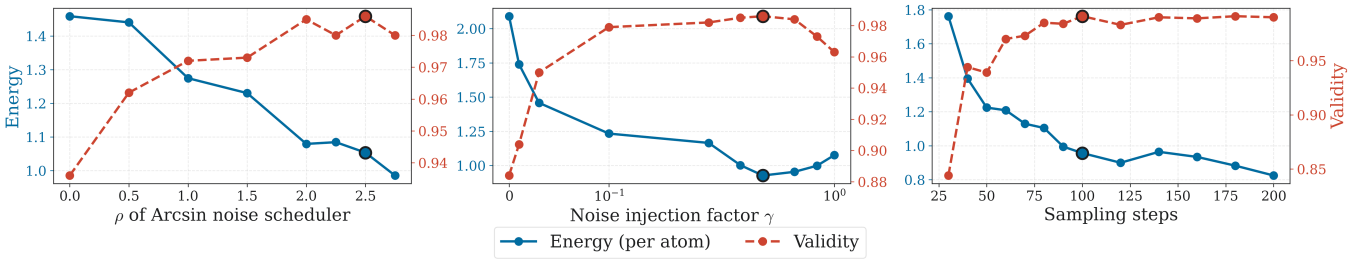

Figure 4: Ablation study of key hyperparameters on GEOM-DRUGS: We report MMFF94 energy and validity when varying (left) arcsin noise factor ρ, (middle) noise injection level γ, and (right) sampling steps. Black circles indicate selected values (ρ=2.5, γ=0.4, 100 steps).

Variant	OT	Noise	Pre	V&C ↑	M.S.↑	E. ↓	R. ↓
VEDA (Full)	✗	✓	✓	98.8	99.5	1.72	0.096
– Pre	✗	✓	✗	97.9	98.5	2.14	0.153
– Noise	✗	✗	✓	95.9	97.6	6.29	0.382
– Noise & Pre	✗	✗	✗	93.8	92.9	10.1	0.409
+ OT Align	✓	✗	✗	91.7	96.0	23.5	0.557
SemlaFlow	✓	✗	✗	92.0	96.9	32.3	0.273

Table 3: Ablation study on GEOM-DRUGS. Each variant enables (✓) or disables (✗) key components of VEDA: OT alignment (OT), noise injection (Noise), and preconditioning (Pre). V&C ↑ and M.S.↑ are reported as percentage-based metrics. The last two rows share the same component configuration and differ only in their time parameterization (VE vs. FM).

ance during sampling and degrades overall sample quality. These results show that each module in VEDA provides complementary improvements to accuracy and validity.

Hyperparameter Sensitivity We performed a sensitivity analysis to support our final hyperparameter choices, as shown in Figure 4. For the arcsin scheduler factor, ρ, the model performs robustly across $\rho \in [2.0, 2.75]$, approaching the theoretical limit where the scheduler function remains monotonic. For the noise injection level γ follows the expected annealing behavior, where an intermediate value achieves the best trade-off. Finally, validity saturates around 100 sampling steps, offering an efficient trade-off between quality and cost.

Discussion

Our findings regarding optimal transport alignment, used in prior works like SemlaFlow(Irwin et al. 2025), reveals a critical design principle: training strategies must not only improve the loss but also consider their impact on downstream sampling. Our ablation study shows that seemingly beneficial optimizations can violate core statistical assumptions—like noise-data independence—that are crucial for generating diverse and valid molecules.

While two-step (e.g., 2D or SMILES-to-3D) generation pipeline are often considered simpler and more scalable, we argue that native 3D generative capability is essential for designing functional molecules in constrained environments like protein pockets. This necessity is underscored by recent benchmarks showing that current target-aware diffusion models frequently produce invalid geometric structures, despite their high docking scores (Baillif et al. 2024). Therefore, enhancing the foundational 3D generative model, as we have focused on in this work, is critical for future success in structure-based design.

Conclusions

In this work, we introduced **VEDA**, a novel equivariant generative model for 3D molecular structures that unifies continuous and discrete generative processes within a single framework. Through comprehensive experiments on QM9 and GEOM-DRUGS, we show that VEDA not only surpasses strong baselines like GeoBFN and SemlaFlow in structural validity and stability metrics, but also delivers efficient generation thanks to its unified diffusion framework. Ablations studies validate that both the amplified VE noise schedule and the discrete sampling mechanism are critical to VEDA's performance.

The success of VEDA in unconditional generation establishes a robust foundation, opening a clear and immediate path toward property-guided molecular design. A natural extension is to condition the generative process on target chemical properties such as binding affinity or solubility, leveraging VEDA's unified framework to guide generation toward desired molecular profiles. Beyond conditional generation, we identify a fundamental limitation shared across current diffusion models: reliance on implicit velocity prediction through score matching. Future equivariant architectures should be designed to explicitly output the time-dependent vector field $v_\theta(\mathbf{x}, t)$. This would align training objectives with advanced samplers like Consistency Flow Matching (Lu and Song 2025) and MeanFlow (Geng et al. 2025), further improving sampling efficiency and unlocking one- or few-step integration.

The power of VEDA lies in its core approach: it uses VE diffusion to smooth the geometric landscape with controlled noise, while seamlessly integrating discrete atomic features. This principled fusion of dynamics and geometry offers an effective blueprint for the future of molecular design.

Acknowledgments

We thank the anonymous reviewers for their constructive comments, which significantly improved the quality of this paper. This work was partially conducted while Dr. Minghu Song was at the University of Connecticut.

References

Abramson, J.; Adler, J.; Dunger, J.; Evans, R.; Green, T.; Pritzel, A.; Ronneberger, O.; Willmore, L.; Ballard, A. J.; Bambrick, J.; et al. 2024. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 1–3.

Austin, J.; Johnson, D. D.; Ho, J.; Tarlow, D.; and Van Den Berg, R. 2021. Structured denoising diffusion models in discrete state-spaces. *Advances in neural information processing systems*, 34: 17981–17993.

Axelrod, S.; and Gomez, R., Bombarelli. 2022. GEOM, energy-annotated molecular conformations for property prediction and molecular generation. *Scientific Data*, 9(1): 185.

Baillif, B.; Cole, J.; McCabe, P.; and Bender, A. 2024. Benchmarking structure-based three-dimensional molecular generative models using GenBench3D: ligand conformation quality matters. *arXiv preprint arXiv:2407.04424*.

Campbell, A.; Yim, J.; Barzilay, R.; Rainforth, T.; and Jaakkola, T. 2024. Generative Flows on Discrete State-Spaces: Enabling Multimodal Flows with Applications to Protein Co-Design. In *International Conference on Machine Learning*, 5453–5512. PMLR.

Chang, H.; Zhang, H.; Jiang, L.; Liu, C.; and Freeman, W. T. 2022. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11315–11325.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. 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)*, 4171–4186.

Dunn, I.; and Koes, D. R. 2024. Exploring Discrete Flow Matching for 3D De Novo Molecule Generation. *ArXiv*, arXiv–2411.

Feng, S.; Ni, Y.; Lu, Y.; Ma, Z.-M.; Ma, W.-Y.; and Lan, Y. 2024. UniGEM: A Unified Approach to Generation and Property Prediction for Molecules. *arXiv preprint arXiv:2410.10516*.

Gao, R.; Hoogeboom, E.; Heek, J.; Bortoli, V.; Murphy, K. P.; and Salimans, T. 2024. Diffusion meets flow matching: Two sides of the same coin. *The Internet*.

Gat, I.; Remez, T.; Shaul, N.; Kreuk, F.; Chen, R. T.; Synnaeve, G.; Adi, Y.; and Lipman, Y. 2024. Discrete flow matching. *Advances in Neural Information Processing Systems*, 37: 133345–133385.

Geng, Z.; Deng, M.; Bai, X.; Kolter, J. Z.; and He, K. 2025. Mean flows for one-step generative modeling. *arXiv preprint arXiv:2505.13447*.

Halgren, T. A. 1996. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *Journal of computational chemistry*, 17(5-6): 490–519.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hong, H.; Lin, W.; and Tan, K. C. 2024. Accelerating 3D Molecule Generation via Jointly Geometric Optimal Transport. In *The Thirteenth International Conference on Learning Representations*.

Hoogeboom, E.; Satorras, V. G.; Vignac, C.; and Welling, M. 2022. Equivariant diffusion for molecule generation in 3d. In *International conference on machine learning*, 8867–8887. PMLR.

Huang, H.; Sun, L.; Du, B.; and Lv, W. 2023. Learning Joint 2D & 3D Diffusion Models for Complete Molecule Generation. *CoRR*.

Irwin, R.; Tibo, A.; Janet, J. P.; and Olsson, S. 2025. SemlaFlow – Efficient 3D Molecular Generation with Latent Attention and Equivariant Flow Matching. In *The 28th International Conference on Artificial Intelligence and Statistics*.

Jing, B.; Corso, G.; Chang, J.; Barzilay, R.; and Jaakkola, T. 2022. Torsional diffusion for molecular conformer generation. *Advances in neural information processing systems*, 35: 24240–24253.

Karras, T.; Aittala, M.; Aila, T.; and Laine, S. 2022. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35: 26565–26577.

Kim, J.; Shah, K.; Kontonis, V.; Kakade, S. M.; and Chen, S. 2025. Train for the Worst, Plan for the Best: Understanding Token Ordering in Masked Diffusions. In *Forty-second International Conference on Machine Learning*.

Le, T.; Cremer, J.; Noe, F.; Clevert, D.-A.; and Schütt, K. T. 2024. Navigating the Design Space of Equivariant Diffusion-Based Generative Models for De Novo 3D Molecule Generation. In *The Twelfth International Conference on Learning Representations*.

Liao, Y.-L.; and Smidt, T. 2023. Equiformer: Equivariant Graph Attention Transformer for 3D Atomistic Graphs. In *The Eleventh International Conference on Learning Representations*.

Lipman, Y.; Chen, R. T.; Ben-Hamu, H.; Nickel, M.; and Le, M. 2023. Flow Matching for Generative Modeling. In *11th International Conference on Learning Representations, ICLR 2023*.

Lu, C.; and Song, Y. 2025. Simplifying, Stabilizing and Scaling Continuous-time Consistency Models. In *The Thirteenth International Conference on Learning Representations*.

Miao, Y.; Feher, V. A.; and McCammon, J. A. 2015. Gaussian accelerated molecular dynamics: unconstrained enhanced sampling and free energy calculation. *Journal of chemical theory and computation*, 11(8): 3584–3595.

Nie, S.; Zhu, F.; You, Z.; Zhang, X.; Ou, J.; Hu, J.; ZHOU, J.; Lin, Y.; Wen, J.-R.; and Li, C. 2025. Large Language Diffusion Models. In *ICLR 2025 Workshop on Deep Generative Model in Machine Learning: Theory, Principle and Efficacy*.

Nikitin, F.; Dunn, I.; Koes, D. R.; and Isayev, O. 2025. GEOM-Drugs Revisited: Toward More Chemically Accurate Benchmarks for 3D Molecule Generation. *arXiv preprint arXiv:2505.00169*.

Ramakrishnan, R.; Dral, P. O.; Rupp, M.; and Von Lilienfeld, O. A. 2014. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data*, 1(1): 1–7.

Reidenbach, D.; Nikitin, F.; Isayev, O.; and Paliwal, S. 2025. Applications of Modular Co-Design for De Novo 3D Molecule Generation. *arXiv preprint arXiv:2505.18392*.

Satorras, V. G.; Hoogeboom, E.; and Welling, M. 2021. E (n) equivariant graph neural networks. In *International conference on machine learning*, 9323–9332. PMLR.

Song, Y.; Gong, J.; Xu, M.; Cao, Z.; Lan, Y.; Ermon, S.; Zhou, H.; and Ma, W.-Y. 2023a. Equivariant flow matching with hybrid probability transport for 3d molecule generation. *Advances in Neural Information Processing Systems*, 36: 549–568.

Song, Y.; Gong, J.; Xu, M.; Cao, Z.; Lan, Y.; Ermon, S.; Zhou, H.; and Ma, W.-Y. 2023b. Equivariant flow matching with hybrid probability transport for 3d molecule generation. *Advances in Neural Information Processing Systems*, 36: 549–568.

Song, Y.; Gong, J.; Zhou, H.; Zheng, M.; Liu, J.; and Ma, W.-Y. 2023c. Unified generative modeling of 3d molecules with bayesian flow networks. In *The Twelfth International Conference on Learning Representations*.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*.

Watson, J. L.; Juergens, D.; Bennett, N. R.; Trippe, B. L.; Yim, J.; Eisenach, H. E.; Ahern, W.; Borst, A. J.; Ragotte, R. J.; Milles, L. F.; et al. 2023. De novo design of protein structure and function with RFdiffusion. *Nature*, 620(7976): 1089–1100.

Xu, M.; Powers, A. S.; Dror, R. O.; Ermon, S.; and Leskovec, J. 2023. Geometric latent diffusion models for 3d molecule generation. In *International Conference on Machine Learning*, 38592–38610. PMLR.

Zhang, P.; Baker, D.; Song, M.; and Bi, J. 2025. Unraveling the potential of diffusion models in small-molecule generation. *Drug Discovery Today*, 104413.

Zhang, Z.; Chen, Y.; and Chu, S. 2025. D3MES: Diffusion Transformer with multihead equivariant self-attention for 3D molecule generation. *arXiv preprint arXiv:2501.07077*.