Elign: Equivariant Diffusion Model Alignment from Foundational Machine Learning Force Fields

Anonymous Authors[1]

Abstract

Generative models for 3D molecular conformations must respect Euclidean symmetries and concentrate probability mass on thermodynamically favorable, mechanically stable structures. However, E(3)-equivariant diffusion models often reproduce biases from semi-empirical training data rather than capturing the equilibrium distribution of a high-fidelity Hamiltonian. While physics-based guidance can correct this, it faces two computational bottlenecks: expensive quantum-chemical evaluations (e.g., DFT) and the need to repeat such queries at every sampling step. We present Elign, a post-training framework that amortizes both costs. First, we replace expensive DFT evaluations with a faster, pretrained foundational machine-learning force field (MLFF) to provide physical signals. Second, we eliminate repeated run-time queries by shifting physical steering to the training phase. To achieve the second amortization, we formulate reverse diffusion as a reinforcement learning problem and introduce Force–Energy Disentangled Group Relative Policy Optimization (FED-GRPO) to fine-tune the denoising policy. FED-GRPO includes a potential-based energy reward and a force-based stability reward, which are optimized and group-normalized independently. Experiments show that Elign generates conformations with lower gold-standard DFT energies and forces, while improving stability. Crucially, inference remains as fast as unguided sampling, since no energy evaluations are required during generation.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

The generation of realistic three-dimensional molecular conformations is a central problem in computational chemistry, materials science, and drug discovery (Xu et al., 2023; Hoogeboom et al., 2022). For practical use, a generative model must satisfy two requirements. First, it must respect the symmetries of physics, most notably invariance to rigid body translations and rotations. Second, it must generate samples that correspond to low energy and physically stable configurations.

Score-based diffusion models with E(3)-equivariant architectures have shown notable performance in molecular generation (Xu et al., 2023; Hoogeboom et al., 2022; Cornet et al., 2025). In the idealized case where the training data are drawn from a true thermodynamic equilibrium, the data distribution itself would follow the Boltzmann law (Car & Parrinello, 1985). Under this condition, maximizing likelihood naturally coincides with favoring mechanically plausible, low-energy configurations. In practice, however, this rarely holds. Standard datasets are typically constructed via semi-empirical relaxations (Bannwarth et al., 2019) or heuristic enumeration (Ramakrishnan et al., 2014; Isert et al., 2022), approximating equilibrium only coarsely. Consequently, the resulting model replicates the biases of the dataset generation rather than the equilibrium distribution of a high-fidelity Hamiltonian (Hohenberg & Kohn, 1964; Kohn & Sham, 1965). This gap highlights a limitation of diffusion models: the likelihood objective does not, by design, inherently enforce physical stability.

A direct way to mitigate this is to introduce physical guidance or rewards during the generation process (Wu et al., 2022). In this setting, an oracle approximating the potential energy surface penalizes mechanically unstable configurations during sampling. While prior work (Zhou et al., 2025) has attempted to use first-principles oracles like density functional theory (DFT) (Hohenberg & Kohn, 1964; Kohn & Sham, 1965), the computational cost is often high. As a result, most approaches are forced to rely on tractable surrogates and compromises, such as applying rewards at terminal sampling stages (Zhou et al., 2025), relying on post-processing (Wu et al., 2022), or utilizing semi-empirical potential methods (Shen et al., 2024; Zhou et al., 2025). No-

tably, many guidance-based approaches (Shen et al., 2024) require *run-time alignment*, meaning that energy or force evaluations must be performed during sampling, incurring additional computational overhead at inference. Another line of work reframes reward-guided diffusion as stochastic optimal control, leading to algorithms such as Adjoint Matching (Havens et al., 2025) and Adjoint Schrödinger Bridge Sampler (Liu et al., 2025a). These methods rely on adjoint equations that assume reward differentiability with respect to the state. In molecular diffusion models with mixed discrete and continuous variables, this assumption could fail, requiring gradients to be approximated using zeroth-order or surrogate estimators, e.g., Simultaneous Perturbation Stochastic Approximation (SPSA) (Spall, 1992; Shen et al., 2024).

MLFFs offer a natural mechanism for injecting physical constraints into generative models (Chmiela et al., 2017; Schütt et al., 2017). By amortizing the cost of expensive first-principles calculations (Car & Parrinello, 1985) into the training phase, inference reduces to efficient neural network evaluations of energies and forces (Behler, 2021; Unke et al., 2021). We see this as the *first level of amortization* in our framework. Moreover, a suggestive connection exists between MLFFs and diffusion models: the denoising objective used to train diffusion models (Hoogeboom et al., 2022; Xu et al., 2023) has also proven effective for MLFF pretraining (Zaidi et al., 2022; Feng et al., 2023). It is also worth noting that historically, MLFFs are typically trained on narrow, system-specific datasets (Chmiela et al., 2017; 2018) such as single molecules, small chemical families, or homogeneous materials. Thus, they do not define a unified potential across the vast chemical space spanned by modern diffusion models. Recent "foundation" MLFFs go a long way toward addressing these limitations (Amin et al., 2025; Mannan et al., 2025; Unke et al., 2021; Wood et al., 2025): trained on large, chemically diverse quantum-mechanical datasets (Chanussot et al., 2021; Tran et al., 2023; Eastman et al., 2023; Unke et al., 2024), they approximate potential energy surfaces across wide molecular classes while retaining amortized efficiency.

MLFFs could enable physically grounded simulation pipelines, such as molecular dynamics or equilibrium sampling. It can also provide guidance that is more accurate than purely data-driven generators. However, those methods remain expensive when many diverse conformations must be generated, as it requires long trajectories and repeated force evaluations. This motivates a *second level of amortization*: compiling physical constraints into the diffusion model itself, shifting computation from inference to a post-training stage. By *post-training*, we mean an additional training stage applied after the likelihood-based pretraining. Together, MLFFs amortize quantum-mechanical calculations into reusable potentials, while post-training amortizes repetitive inference into the denoiser, eliminating run-time overhead. Based on these, we introduce Elign, a framework that fine-tunes E(3)-equivariant diffusion models using MLFF-derived rewards. Elign employs potential-based reward shaping derived from MLFF energies. Both energy and force signals are combined under a disentangled group-relative policy optimization scheme (FED-GRPO), analogous to MLFF training. Experiments on QM9 and GEOM-drugs show that Elign generates low-energy, mechanically stable conformations, outperforming runtime-guided methods while matching the inference speed of unguided samplers.

2. Preliminaries

Denoising diffusion models. Diffusion models define a generative distribution $p_\theta(\boldsymbol{z}_0)$ over data $p_{\text{data}}(\boldsymbol{z}_0)$ on a space $\mathcal{Z}$ by learning to reverse an iterative forward noising process. The forward process is defined by the SDE $d\boldsymbol{z}_t = b_t(\boldsymbol{z}_t)\,dt + \sigma_t\,d\mathbf{B}_t$ with initial condition $\boldsymbol{z}_0 \sim p_{\text{data}}$. Let $p_t = \text{Law}(\boldsymbol{z}_t)$. For the general forward diffusion $d\boldsymbol{z}_t = -\frac{1}{2}\beta_t \boldsymbol{z}_t\,dt + \sqrt{\beta_t}\,d\mathbf{B}_t$, the transition distribution is $p(\boldsymbol{z}_t|\boldsymbol{z}_0) = \mathcal{N}(\boldsymbol{z}_t; \alpha_t \boldsymbol{z}_0, \bar{\sigma}_t^2 \mathbf{I})$ with $\alpha_t = \exp(-\frac{1}{2}\int_0^t \beta_s ds)$ and $\bar{\sigma}_t^2 = 1 - \exp(-\int_0^t \beta_s ds)$. With a constant noise schedule $\beta_t = 2$, we recover the Ornstein–Uhlenbeck forward process $d\boldsymbol{z}_t = -\boldsymbol{z}_t\,dt + \sqrt{2}\,d\mathbf{B}_t$ with $\alpha_t = e^{-t}$ and $\bar{\sigma}_t^2 = 1 - e^{-2t}$. This gives a signal-to-noise ratio $\text{SNR}(t) := \alpha_t^2/\bar{\sigma}_t^2$ that decreases monotonically in t, ensuring that as $t \to T$, $p_T \approx \mathcal{N}(0, \mathbf{I})$. To define the reverse process, let $\boldsymbol{z}_t^{\leftarrow} = \boldsymbol{z}_{T-t}$ denote the time-reversed trajectory with law $p_t^{\leftarrow} = p_{T-t}$. By Nelson's theorem (Nelson, 1967) with mild regularity conditions, the reverse-time process satisfies the drift relation $b_t(\boldsymbol{z}) + b_{T-t}^{\leftarrow}(\boldsymbol{z}) = a_t \nabla \log p_t(\boldsymbol{z})$ where $a_t = \sigma_t \sigma_t^\top$, yielding the reverse SDE $d\boldsymbol{z}_t^{\leftarrow} = (-b_{T-t}(\boldsymbol{z}_t^{\leftarrow}) + a_{T-t}\nabla \log p_{T-t}(\boldsymbol{z}_t^{\leftarrow}))\,dt + \sigma_{T-t}\,d\mathbf{B}_t$. In practice, the score function is approximated by a neural network $s_\theta : \mathcal{Z} \times [0, T] \to \mathcal{Z}$ trained via denoising score matching by minimizing $\mathbb{E}\left[\|s_\theta(\boldsymbol{z}_t, t) - \nabla_{\boldsymbol{z}_t} \log p(\boldsymbol{z}_t|\boldsymbol{z}_0)\|^2\right]$.

Equivariant diffusion models. Let $\mathcal{G}$ be a group acting on a space $\mathcal{Z}$. A function $f : \mathcal{Z} \to \mathcal{Z}$ is $\mathcal{G}$-equivariant if $f(g \cdot \boldsymbol{z}) = g \cdot f(\boldsymbol{z})$ for all $g \in \mathcal{G}, \boldsymbol{z} \in \mathcal{Z}$, and a distribution p on $\mathcal{Z}$ is $\mathcal{G}$-invariant if its density satisfies $p(g \cdot \boldsymbol{z}) = p(\boldsymbol{z})$. To construct a diffusion model that generates samples from a $\mathcal{G}$-invariant law, both the forward and reverse SDEs must respect this symmetry. For the forward process, the Fokker–Planck equation $\partial_t p_t = \frac{1}{2}\langle \sigma_t \sigma_t^\top, \nabla^2 p_t \rangle - \nabla \cdot (b_t p_t)$ preserves invariance when b_t is equivariant and $\sigma_t \sigma_t^\top$ commutes with the group action. By defining a $\mathcal{G}$-equivariant forward process, the true reverse-time drift is automatically $\mathcal{G}$-equivariant, since it is composed of an equivariant drift b_t and an equivariant score function $\nabla \log p_t$ (for a $\mathcal{G}$-invariant distribution p_t, $\nabla \log p_t(g \cdot \boldsymbol{z}) = g \cdot \nabla \log p_t(\boldsymbol{z})$). In prac-

tice, the unknown true score is estimated by a parameterized model $s_\theta(\boldsymbol{z}_t, t)$. To preserve this symmetry, the score model must be parameterized as a $\mathcal{G}$-equivariant neural network. The reverse process must also be initiated by drawing samples from a $\mathcal{G}$-invariant terminal distribution p_T.

N-body data and subspace diffusion. An N-body system is defined by state $\boldsymbol{z} = [\mathbf{x}, \mathbf{h}] \in \mathbb{R}^{N\times 3} \times \mathbb{R}^{N\times d_h}$, comprising positions $\mathbf{x} \in \mathcal{X}$ and invariant features $\mathbf{h} \in \mathcal{H}$ (e.g., atom types). The physical distribution $p(\boldsymbol{z})$ is invariant under the Euclidean group $\mathrm{E}(3)$ acting on $\mathcal{X}$. Translation invariance implies that $p(\boldsymbol{z})$ is not normalizable on the full space $\mathbb{R}^{3N}$. We therefore restrict the diffusion to the linear subspace of center-of-mass (CoM) zero configurations (Hoogeboom et al., 2022), $\mathcal{M} = \{\mathbf{x} \in \mathbb{R}^{N\times 3} \mid \sum_i \mathbf{x}_i = \mathbf{0}\}$. We define the projection operator $\mathbf{P}_{\mathrm{CoM}}$ as a block-diagonal matrix: on positions it is the $3N \times 3N$ centering projector, while on features it acts as identity. Explicitly, $\mathbf{P}_{\mathrm{CoM}} = \mathrm{diag}(\mathbf{P}_{\mathcal{M}}, \mathbf{I}_{Nd_h})$ where $\mathbf{P}_{\mathcal{M}}$ projects onto $\mathcal{M}$. The forward SDE is modified to diffuse only within this subspace: $d\boldsymbol{z}_t = -\frac{1}{2}\beta_t \boldsymbol{z}_t\, dt + \sqrt{\beta_t}\, \mathbf{P}_{\mathrm{CoM}}\, d\mathbf{B}_t$. Because $\mathbf{P}_{\mathcal{M}}$ is idempotent (that is, $\mathbf{P}_{\mathcal{M}} \circ \mathbf{P}_{\mathcal{M}} = \mathbf{P}_{\mathcal{M}}$) and commutes with the $\mathrm{E}(3)$ action restricted to rotations and reflections, the marginal distributions p_t remain supported on $\mathcal{M}$ and retain $\mathrm{E}(3)$-invariance. The score network s_θ is parameterized to be $\mathrm{E}(3)$-equivariant and to output vectors in the tangent space of $\mathcal{M}$ (i.e., satisfying the zero-CoM constraint). This could be realized by projecting coordinate updates with $\mathbf{P}_{\mathcal{M}}$ at each denoising step.

Machine-learning force fields. Machine-learning force fields (MLFFs) serve as efficient surrogates for quantum mechanical (QM) calculations, enabling dynamics simulations at scales inaccessible to *ab initio* methods. We assume access to a *quantum chemistry oracle* which, for any given molecular configuration $\boldsymbol{z}$, computes a reference energy E^{ref} and reference atomic forces $\mathbf{F}^{\mathrm{ref}}$. As direct queries to such oracles are computationally expensive, often scaling cubically with the number of atoms, MLFFs amortize this cost by training a neural network to approximate the potential energy function $E_\phi(\boldsymbol{z})$. Atomic forces are typically obtained via automatic differentiation as the negative gradient of this potential, $\mathbf{F}_\phi^{(i)} = -\nabla_{\mathbf{x}^{(i)}} E_\phi(\boldsymbol{z})$, though they may also be parameterized directly via a dedicated force regression head. In practice (e.g., (Wood et al., 2025)), one can first train a direct force-regression head and then fine-tune via automatic differentiation. During training, the network parameters ϕ are optimized by minimizing a combined energy-and-force matching loss over a dataset of reference calculations:

$$\mathcal{L} = \lambda_E \left(E_\phi(\boldsymbol{z}) - E^{\mathrm{ref}}\right)^2 + \lambda_F \sum_{i=1}^{N} \left\| \mathbf{F}_\phi^{(i)}(\boldsymbol{z}) - \mathbf{F}^{\mathrm{ref},(i)}(\boldsymbol{z}) \right\|^2, \tag{1}$$

where λ_E and λ_F are weighting coefficients. Physical consistency imposes strict symmetry constraints: E_ϕ must be $\mathrm{E}(3)$-invariant (unaffected by global rotation, translation, and inversion), while the derived forces must be $\mathrm{E}(3)$-equivariant. These constraints are enforced via specialized invariant or equivariant network architectures. Recently, foundational MLFFs have scaled this paradigm to massive heterogeneous datasets, learning universal potentials that are *transferable* across diverse chemical systems.

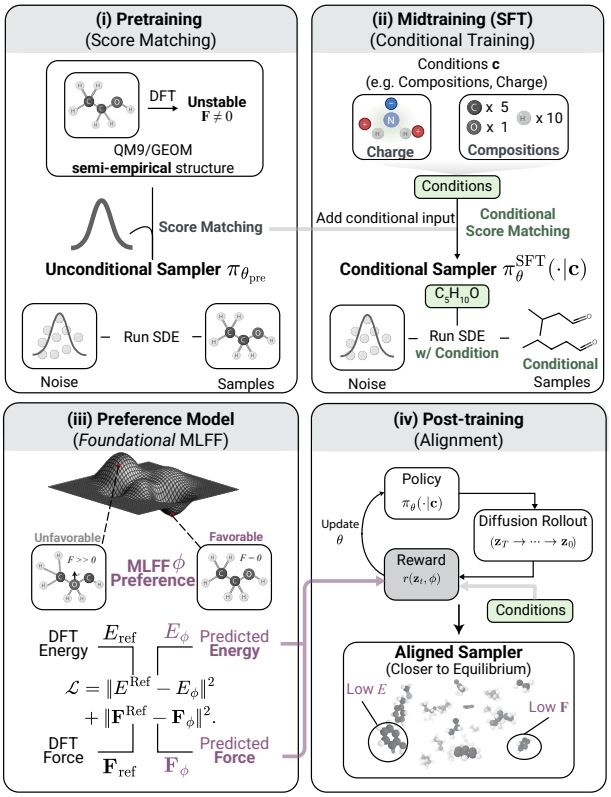

Figure 1. **Staged pipeline for equilibrium molecular generation.** (i) **Pretraining:** score-matching trains an E(3)-equivariant diffusion model $\pi_\theta^{\mathrm{pre}}$ on approximate structures. (ii) **Optional SFT:** conditional fine-tuning improves adherence to discrete specifications. (iii) **Preference model:** a foundation MLFF ϕ provides energy/force signals. (iv) **Post-training:** RL fine-tunes the sampler to improve stability without run-time oracle calls.

3. Diffusion post-training with a force-field alignment model

Setup. As shown in Figure 1, we organize equilibrium molecular generation into a staged pipeline analogous to LLM training: base model training, optional supervised conditioning, preference modeling, and post-training alignment. (i) **Pretraining (score matching):** We train an equivariant diffusion model $\pi_{\theta_{\mathrm{pre}}}$ via score matching on large collections of approximate equilibrium structures (e.g., QM9-style semi-empirical pipelines). This stage learns a broad generative distribution but does not enforce physical stability. (ii) **Optional SFT (conditional training):** When explicit constraints are required, one could optionally fine-tune the model to condition on variables such as by training a conditional diffusion model $\pi_{\theta_{\mathrm{sft}}}(\cdot \mid \mathbf{c})$ on paired data $(\mathbf{c}, \boldsymbol{z})$ using the same denoising objective. This step corresponds to in-

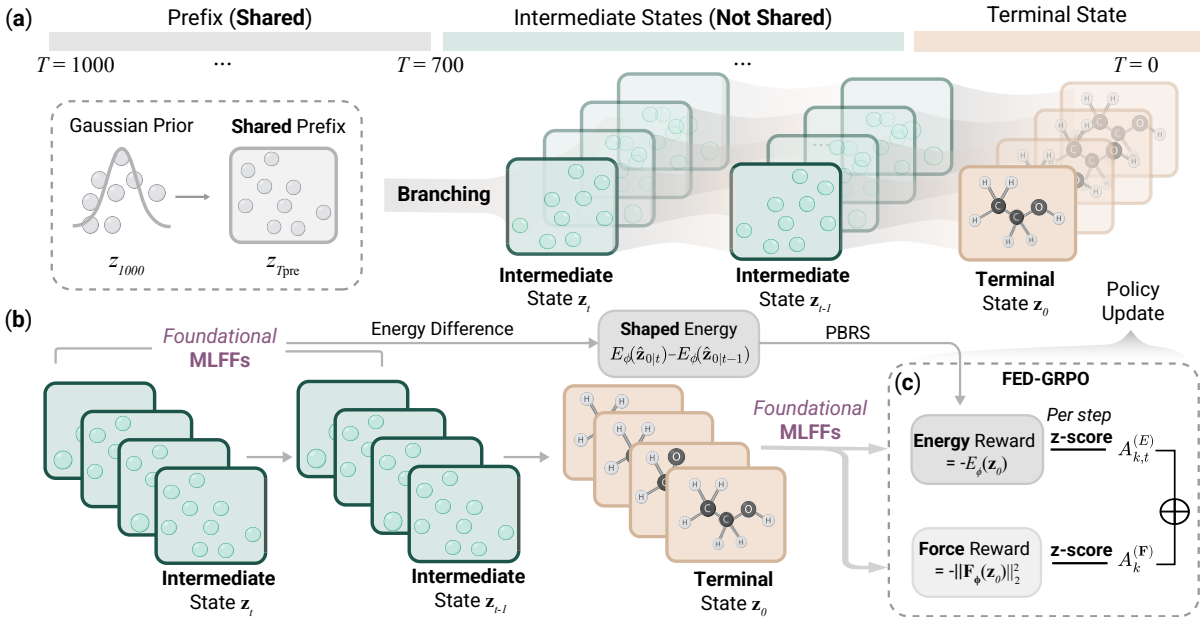

Figure 2. Overview of Elign. **(a)** Rollout branching with shared prefix: starting from a CoM-free Gaussian prior at $t = T$, an EGNN policy denoises to $t = 0$. At $t = T_{\text{prefix}}$, we cache a shared prefix state and branch into K rollouts with independent noise. Each trajectory is propagated to its terminal state $\boldsymbol{z}_0$ and scored by a foundational MLFF. **(b)** Energy-based reward shaping: intermediate predicted clean geometries $\hat{\boldsymbol{z}}_{0|t}$ are evaluated by the MLFF, and local energy differences $E_\phi(\hat{\boldsymbol{z}}_{0|t}) - E_\phi(\hat{\boldsymbol{z}}_{0|t-1})$ provide dense shaping signals that bias the policy toward lower-energy conformations. **(c)** FED-GRPO: terminal energy and force rewards are z-score normalized separately per timestep, then combined to compute the final advantage for policy updates.

struction SFT in LLMs and improves constraint adherence without introducing preference optimization[1]. (iii) **Preference model (foundation MLFF):** We use a pretrained foundation MLFF ϕ as a preference model, providing energies and forces that quantify thermodynamic stability and mechanical equilibrium. (iv) **Post-training (alignment):** We treat reverse diffusion as a trajectory-generating process and apply reinforcement learning to align π_θ with the MLFF-defined preferences, while limiting deviation from the pretrained or SFT-initialized model via trust-region regularization.

Diffusion as an MDP. We formulate the iterative denoising process as a finite-horizon Markov Decision Process (MDP). We index reverse time with t decreasing from T to 0. We initialize our method with a pretrained equivariant diffusion model, $\pi_{\theta_{\text{pre}}}$, which serves as the reference base policy. The state is defined by $S_t = (\mathbf{z}_t, t)$, comprising the latent molecular geometry $\mathbf{z}_t$ and the discrete reverse-diffusion time index t. The *policy* $\pi_\theta(\mathbf{z}_{t-1} \mid \mathbf{z}_t, t)$ executes a *one-step discretization* of the reverse SDE update (e.g., Euler-Maruyama. For simplicity, we assume first-order solvers so the unaugmented state S_t is Markov.). This induces a Gaussian policy parameterized by the score network: $\pi_\theta(\boldsymbol{z}_{t-1} \mid \boldsymbol{z}_t, t) = \mathcal{N}(\boldsymbol{z}_{t-1}; \mu_\theta(\boldsymbol{z}_t, t), \Sigma_t)$, where μ_θ is the learned reverse mean and Σ_t is given. The *action* is defined as a realization sampled from the Gaussian policy.

[1]Due to the page limit, we include results on conditional generation in Appendix E.2.

It is important to distinguish the per-step policy, denoted $\pi(\mathbf{z}_{t-1} \mid \mathbf{z}_t, t)$, from the resulting distribution ρ, which arises from the sequential chaining of these policy steps. All stochasticity in the trajectory arises from the Gaussian perturbation inside the policy itself. There are two apparent sources of randomness: (i) the noise term of the discretized reverse-time SDE and (ii) sampling from a stochastic policy. In our case, these two sources *coincide* because the policy itself is the SDE kernel. Here, we offer an alternative interpretation of this MDP through the lens of stochastic optimal control (of which maximum-entropy RL is a special case). In this view, the goal of RL is to learn a policy that acts as a "control knob," modulating the drift term $\mu_\theta(\cdot)$ of the diffusion process to maximize the expected reward. This is also analogous to a *learnable guidance function* in diffusion guidance.

3.1. Alignment Objective

The objective of RL post-training is to ensure the final generated structures are physically valid. Validity requires two properties: thermodynamic stability (low potential energy) and mechanical equilibrium (zero net force). To design the reward, we draw inspiration from training MLFFs, which utilize supervision from both energy and force labels to learn the potential energy surface (Eq. 1). Because force correlates with the gradient of energy, these objectives are coupled but distinct; a structure can have low energy but high unstable forces if it sits on a steep slope of the potential energy surface. We therefore define the *terminal alignment*

reward, assigned only at the final step when $t = 0$. Specifically:

$$r_0^{(E)}(\boldsymbol{z}_0) := -E_\phi(\boldsymbol{z}_0); r_0^{(\mathbf{F})}(\boldsymbol{z}_0) := -\|\mathbf{F}_\phi(\boldsymbol{z}_0)\|_F^2. \quad (2)$$

Here, E_ϕ and $\mathbf{F}_\phi$ are the energy and force predictions from the pretrained MLFF. For implementation, we normalize the energy and force rewards using formation energy per atom and RMS force.

Energy-based potential shaping. Optimizing solely for a terminal alignment reward yields a sparse learning signal over long reverse diffusion horizons. To facilitate credit assignment while preserving the optimal policy of the original terminal-reward MDP, we add an intermediate signal using *potential-based reward shaping* (PBRS). Traditionally, PBRS relies on a potential function $\Psi(S)$ that estimates the "closeness" of a current state to the goal, much like an admissible heuristic in A* search (Hart et al., 1968). In our setting, this role is naturally played by physics: because our objective is thermodynamic stability, the physical potential energy E_ϕ directly quantifies the distance from equilibrium.

The MLFF energy $E_\phi(\cdot)$ is physically meaningful only on clean atomic coordinates, not on the noisy intermediates $\boldsymbol{z}_t$. At each reverse step t we therefore reconstruct a *predicted clean geometry* $\boldsymbol{z}_{0|t}$ from $\boldsymbol{z}_t$ using the diffusion posterior mean estimate: $\hat{\boldsymbol{z}}_{0|t} = \frac{1}{\alpha_t}(\boldsymbol{z}_t + \bar{\sigma}_t^2 s_\theta(\boldsymbol{z}_t, t))$. We define a shaping potential on this reconstructed estimate: $\Psi(S_t) := -E_\phi(\hat{\boldsymbol{z}}_{0|t}), S_t = (\boldsymbol{z}_t, t)$. The intermediate shaping reward for a transition $S_t \to S_{t-1}$ is then

$$r_t^{\text{shape}} := \gamma\,\Psi(S_{t-1}) - \Psi(S_t), \qquad t = T_{\text{prefix}}, \dots, 1. \quad (3)$$

This is the canonical PBRS form and preserves the set of optimal policies under the discounted return. In particular, if we define the (discounted) shaped return-to-go from state S_t as $G_t^{\text{shape}} := \sum_{u=1}^{t} \gamma^{t-u} r_u^{\text{shape}}$, then the shaping contribution telescopes:

$$G_t^{\text{shape}} = \sum_{u=1}^{t} \gamma^{t-u}(\gamma\Psi(S_{u-1}) - \Psi(S_u)) = \gamma^t \Psi(S_0) - \Psi(S_t). \quad (4)$$

Intuitively, the return-to-go is the discounted difference between terminal and current state potentials. This shaping provides informative per-step feedback while preserving policy optimality under the discounted return (Ng et al., 1999). We also considered a force-based potential, e.g., $\Psi_F(S_t) = -\|\mathbf{F}_\phi(\hat{\boldsymbol{z}}_{0|t})\|_F^2$, and apply PBRS analogously. We found it unstable in practice, possibly because forces (first-order derivatives) are more sensitive than energies (zeroth-order) to small geometric errors in $\hat{\boldsymbol{z}}_{0|t}$.

3.2. Theoretical View of the Alignment Objective

The rewards above specify what properties we desire in the final sample. We now adopt a distributional perspective to characterize the *terminal law* that post-training encourages. Because the diffusion policy is parameterized by an E(3)-equivariant architecture, post-training preserves invariance of the terminal distribution. For analytical clarity, we focus on the energy objective and ignore PBRS.

Let $\mathcal{Z}$ denote the (CoM-free) configuration space of clean molecular structures, and identify the terminal denoised sample $\boldsymbol{z}_0 \in \mathcal{Z}$. Let $\rho_{\theta_{\text{pre}}} \in \mathcal{P}(\mathcal{Z})$ denote the terminal distribution induced by the pretrained diffusion model $\pi_{\theta_{\text{pre}}}$, i.e., $\boldsymbol{z}_0 \sim \rho_{\theta_{\text{pre}}}$ when sampling from $\pi_{\theta_{\text{pre}}}$. Any fine-tuned policy π_θ induces its own terminal law ρ_θ over $\mathcal{Z}$.

Theorem 1 (Energy-aligned terminal distribution)**.** *Assume the policy class is rich enough that any $\rho \ll \rho_{\theta_{\text{pre}}}$ can be realized as the terminal law of some admissible reverse diffusion policy, and for a trust-region regularized reward maximization objective: $\mathcal{J}(\rho) := \mathbb{E}_{\boldsymbol{z}_0\sim\rho}[-E_\phi(\boldsymbol{z}_0)] - w_{\text{KL}}\,\text{KL}(\rho\|\rho_{\theta_{\text{pre}}})$, the maximum is attained. Then the maximizer is unique and equals*

$$\rho^\star(\boldsymbol{z}) = \frac{1}{Z_\phi}\,\rho_{\theta_{\text{pre}}}(\boldsymbol{z}) \exp\Big(-\beta_{\text{eff}}\,E_\phi(\boldsymbol{z})\Big), \beta_{\text{eff}} := \frac{1}{w_{\text{KL}}}, \quad (5)$$

with normalizer $Z_\phi = \int_{\mathcal{Z}} \rho_{\theta_{\text{pre}}}(\mathbf{u}) \exp(-\beta_{\text{eff}} E_\phi(\mathbf{u}))\, d\mathbf{u}$.

Proof sketch. This is the standard variational form of KL-regularized optimization: enforce normalization with a Lagrange multiplier, take first-order optimality conditions in ρ, and normalize. Strict concavity in ρ (from the negative entropy term) gives uniqueness.

This result admits an energy decomposition view: defining a "prior energy" $E_{\text{prior}}(\boldsymbol{z}) := -\beta_{\text{eff}}^{-1} \log \rho_{\theta_{\text{pre}}}(\boldsymbol{z})$, the optimal density satisfies

$$\rho^\star(\boldsymbol{z}) \propto \exp\Big(-\beta_{\text{eff}}[E_\phi(\boldsymbol{z}) + E_{\text{prior}}(\boldsymbol{z})]\Big).$$

The terminal distribution is therefore determined by two components: the physical energy $E_\phi(\boldsymbol{z})$ from the MLFF and an implicit prior induced by the pretrained model. Note that we do not compute E_{prior} explicitly, this is a conceptual decomposition. In this view, alignment sharpens thermodynamic plausibility while remaining confined to the learned data manifold, with the pretrained model acting as both a regularizer and a support constraint (via the absolute-continuity condition). This highlights the importance of high-quality pretraining. Moreover, post-training can be viewed as amortized sampling: instead of applying energy guidance at inference time via repeated E_ϕ evaluations, RL *compiles* the corresponding Gibbs reweighting into the reverse diffusion policy. We next relate MLFF approximation error to the true physical Boltzmann law:

MLFF approximation and distribution error. Let $E_\star : \mathcal{Z} \to \mathbb{R}$ be a target potential energy and E_ϕ its MLFF approximation. Fix a reference terminal law $\rho_{\theta_{\text{pre}}}$ and

define $\rho^\star_\star(\boldsymbol{z}) \propto \rho_{\theta_{\text{pre}}}(\boldsymbol{z})\, e^{-\beta_{\text{eff}} E_\star(\boldsymbol{z})}$, $\rho^\star_\phi(\boldsymbol{z}) \propto \rho_{\theta_{\text{pre}}}(\boldsymbol{z})\, e^{-\beta_{\text{eff}} E_\phi(\boldsymbol{z})}$. We relate the MLFF error to the distributional error via the total variation distance:

Theorem 2 (Energy error implies distribution error)**.** *Assume the MLFF uniformly approximates the target energy,* $\sup_{\boldsymbol{z}\in\mathcal{Z}} |E_\phi(\boldsymbol{z}) - E_\star(\boldsymbol{z})| \le \delta$ *for some* $\delta \ge 0$*. Then the aligned terminal laws in Eq.* (3.2) *satisfy*

$$\|\rho^\star_\phi - \rho^\star_\star\|_{\text{TV}} \le \tanh(\beta_{\text{eff}}\delta). \tag{6}$$

The bound in Eq. (6) implies that the terminal distribution error is controlled by the product $\beta_{\text{eff}}\delta$. In the small-error or moderate-temperature regime, $\tanh(\beta_{\text{eff}}\delta) \approx \beta_{\text{eff}}\delta$, so the mismatch scales linearly with the uniform energy error. In contrast, when β_{eff} is large, even small MLFF biases can induce substantial deviations in the terminal law. This highlights a practical trade-off: aggressive alignment toward low effective temperatures requires either higher MLFF accuracy in the relevant region or stronger trust-region regularization (smaller β_{eff}) to avoid amplifying model miscalibration. We would also like to highlight that this uniform error is conservative, providing a worst-case guarantee; in practice we expect smaller errors in high-density regions. In the next section, we describe how this theoretical objective is implemented via reinforcement learning.

3.3. RL Training

Force–energy disentangled GRPO (FED-GRPO). To fine-tune the diffusion model, we require a policy gradient algorithm that is both sample-efficient and computationally inexpensive. Standard actor–critic methods are costly, as they would require training a separate value function alongside the score model. Moreover, in our setting the value function must be E(3)-invariant and defined over high-dimensional molecular states; learning such a critic via bootstrapping can introduce instability, in addition to extra compute. We instead adopt *Group Relative Policy Optimization* (GRPO) (Shao et al., 2024), a critic-free algorithm, and introduce a *disentangled* advantage formulation that separately accounts for energy and force signals; we refer to the resulting method as Force–Energy Disentangled GRPO (**FED-GRPO**).

Shared-prefix grouping. We first define a *group* over a diffusion trajectory. In large language models, all members of a group share the same prompt; for example, given a fixed math question, multiple responses are sampled in parallel and compared to identify higher-quality answers. Analogously, we extend this idea to diffusion by rolling out a common reverse-diffusion prefix from $T \to T_{\text{prefix}}$ (for a chosen prefix time $T_{\text{prefix}} \in \{1,\dots,T\}$) and caching $\boldsymbol{z}_{T_{\text{prefix}}}$. From this shared state we branch into K stochastic continuations with fresh noise. This shared initialization has also been used in concurrent work (e.g., FlowGRPO (Liu et al., 2025b)) and lowers variance by making early denoising dynamics comparable.

Two-channel advantages. For each branched rollout k, we distinguish between dense energy feedback and sparse force feedback: (i) **Step-wise energy advantage:** For energy, we utilize the dense PBRS signal. We first compute the raw *return-to-go* at each step t: $G^{(E)}_{k,t} = \sum_{u=1}^{t} \gamma^{t-u} r^{E,\text{shape}}_{k,u}$. Crucially, we do **not** normalize $r_{k,u}$ per step; we first sum raw shaped rewards to preserve telescoping (Eq. (4)), then normalize only at the advantage level. (ii) **Trajectory-level force advantage:** For force, we assign a single scalar return to the entire trajectory: $G^{(\mathbf{F})}_k = -\left\|\mathbf{F}_\phi(\boldsymbol{z}^{(k)}_0)\right\|^2_F$.

Disentangled group-relative normalization. Directly summing the raw rewards is undesirable, as their differing scales and temporal structures can lead to interference, with one signal dominating the optimization dynamics. We therefore standardize the two channels differently to mitigate these effects and to disentangle their contributions according to granularity. For **force**, we compute a single group-level mean μ_F and standard deviation σ_F. For **energy**, we compute statistics **per time-step** t, denoted $(\mu_{E,t}, \sigma_{E,t})$, across the K rollouts. This ensures that the advantage at step t is relative to the group's performance *at that specific step*. The total advantage for rollout k at step t is:

$$\hat{A}_{k,t} = w_E \underbrace{\frac{G^E_{k,t} - \mu_{E,t}}{\sigma_{E,t} + \eta}}_{\text{Step-wise Energy Adv.}} + w_{\mathbf{F}} \underbrace{\frac{G^F_k - \mu_F}{\sigma_F + \eta}}_{\text{Global Force Adv.}}. \tag{7}$$

Here, $\eta > 0$ is a small constant added for numerical stability, and the force term is broadcast to the entire trajectory.

Clipped FED-GRPO objective. Let the probability ratio be $\xi_{k,t}(\theta) = \frac{\pi_\theta(\boldsymbol{z}^{(k)}_{t-1}|\boldsymbol{z}^{(k)}_t, t)}{\pi_{\theta_{\text{old}}}(\boldsymbol{z}^{(k)}_{t-1}|\boldsymbol{z}^{(k)}_t, t)}$, which determines how much more (or less) likely the current policy is to take a given denoising step relative to the old policy. The clipped surrogate objective aggregates over all steps using the per-step advantage $\hat{A}_{k,t}$: $L(\theta) = \mathbb{E}_k[\sum_{t=1}^{T_{\text{prefix}}} \min\Big(\xi_{k,t}(\theta)\,\hat{A}_{k,t},\ \text{clip}(\xi_{k,t}(\theta), 1-\varepsilon, 1+\varepsilon)\,\hat{A}_{k,t}\Big)]$, This objective maximizes (or minimizes) the likelihood of denoising trajectories proportional to their physics-based advantage, with clipping enforcing a *local* (per-step) trust region that stabilizes training; a *global* trust region can also be added via a KL penalty against the pretrained policy $\pi_{\theta_{\text{pre}}}$. Concretely, this is given as: $\mathcal{L}_{\text{reg}}(\theta) := \mathcal{L}(\theta) + w_{\text{KL}}\,\text{KL}\big(\pi_\theta \,\|\, \pi_{\theta_{\text{pre}}}\big)$, where $w_{\text{KL}} > 0$ controls the strength of the global trust region. Note that this is a tractable proxy for teminal-law trust region defined in Sec.3.2.

Finally, we summarize the FED-GRPO procedure in Alg. 1–2.

Algorithm 1 FED-GRPO post-training

Input: $\pi_{\theta_{\text{old}}}$, MLFF ϕ, T_{prefix}, K, $(w_E, w_{\mathbf{F}})$, γ, clip ε.
for each iteration **do**
  Rollout $T \to T_{\text{prefix}}$ with $\pi_{\theta_{\text{old}}}$; cache $\boldsymbol{z}_{\text{start}}$. ▷ Shared.
  for $k = 1, \ldots, K$ **do**
    $\tau_k \leftarrow \textsc{Rollout}(\boldsymbol{z}_{\text{start}}, \pi_{\theta_{\text{old}}})$
    $(\{G_{k,t}^{(E)}\}_{t=1}^{T_{\text{prefix}}}, G_k^{(\mathbf{F})}) \leftarrow \textsc{Reward}(\tau_k, \phi, \pi_{\theta_{\text{old}}})$.
  $\hat{A}_k^{(\mathbf{F})} \leftarrow (G_k^{(\mathbf{F})} - \mu_F)/(\sigma_F + \eta)$ over k.
  $\hat{A}_{k,t}^{(E)} \leftarrow (G_{k,t}^{(E)} - \mu_{E,t})/(\sigma_{E,t} + \eta)$ over k for each t.
  $\hat{A}_{k,t} \leftarrow w_E \hat{A}_{k,t}^{(E)} + w_{\mathbf{F}} \hat{A}_k^{(\mathbf{F})}$;
  $\xi_{k,t} \leftarrow \pi_\theta(\boldsymbol{z}_{t-1}^{(k)} | \boldsymbol{z}_t^{(k)}, t)/\pi_{\theta_{\text{old}}}(\boldsymbol{z}_{t-1}^{(k)} | \boldsymbol{z}_t^{(k)}, t)$.
  Update θ with the GRPO objective using $(\xi_{k,t}, \hat{A}_{k,t})$; ▷ Optionally add a KL Term.
  set $\theta_{\text{old}} \leftarrow \theta$.

Algorithm 2 REWARD: energy PBRS return-to-go + terminal force

function $\textsc{Reward}(\tau, \phi, \pi_{\theta_{\text{old}}})$
  Extract $\{\boldsymbol{z}_t\}_{t=0}^{T_{\text{prefix}}}$ from τ.
  for $t = T_{\text{prefix}}, \ldots, 0$ **do**
    Compute $\hat{\boldsymbol{z}}_{0|t}$ from $\boldsymbol{z}_t$ ▷ Posterior from $\pi_{\theta_{\text{old}}}$.
    $\Psi_t \leftarrow -E_\phi(\hat{\boldsymbol{z}}_{0|t})$.
  for $t = 1, \ldots, T_{\text{prefix}}$ **do**
    $G_t^{(E)} \leftarrow \gamma^t \Psi_0 - \Psi_t$. ▷ Eq. (4)
  $G^{(\mathbf{F})} \leftarrow -\|\mathbf{F}_\phi(\boldsymbol{z}_0)\|_F^2$.
  return $\{G_t^{(E)}\}_{t=1}^{T_{\text{prefix}}}, G^{(\mathbf{F})}$.

4. Results

Implementation. We use the public EDM (Hoogeboom et al., 2022) checkpoint as our pretrained base model and UMA (Wood et al., 2025) as the preference model. UMA is a machine-learning force field trained on OMol25 (Levine et al., 2025), OC20 (Chanussot et al., 2021), ODAC23 (Sriram et al., 2024), OMat24 (Barroso-Luque et al., 2024), etc., that predicts per-molecule energies and per-atom forces from atom types and 3D coordinates. UMA has two variants: UMA-1p1-S and UMA-1p1-M. Unless stated otherwise, Elign uses UMA-1p1-M with potential-based reward shaping (PBRS). Although UMA's training data may include molecules from QM9 or GEOM-Drugs, the specific 3D conformations generated by EDM are novel and not seen during UMA training. Moreover, we use UMA only as a fixed reward oracle without fine-tuning, and our key evaluations (RDKit validity, DFT energies/forces) are computed independently of UMA. We evaluate on QM9 (Ramakrishnan et al., 2014) and GEOM-Drugs (Axelrod & Gómez-Bombarelli, 2022). QM9 contains 130k small molecules with up to 9 heavy atoms (29 atoms including hydrogens). GEOM-Drugs consists of larger organic compounds with up to 181 atoms (44.2 on average) across 37 million conformations for around 450k molecules. Following Xu et al. (2023), we report atom stability (A), molecule stability (M), RDKit validity (V), and Validity×Uniqueness (V×U). QM9 results average three runs of 10,000 samples; GEOM-Drugs uses 1,024 samples. For GEOM-Drugs, we omit M and V×U due to limitations in ground-truth bond inference (bond inference is unreliable for larger molecules with 60–200 atoms) and near-saturated uniqueness ($\approx 100\%$). This is in line with community standards (Xu et al., 2023). Baseline results are taken from published work whenever possible.

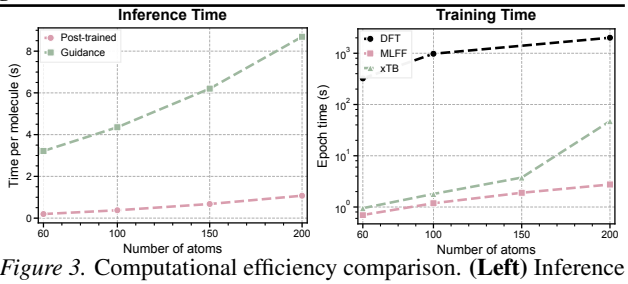

Figure 3. Computational efficiency comparison. **(Left)** Inference time per molecule for post-trained models versus guidance-based methods. **(Right)** Training epoch time comparing different reward oracles on a log scale. For a fair comparison, all methods use terminal-only rewards (no PBRS). Evaluation was conducted on a single NVIDIA H100 GPU.

QM9. Table 1 reports the benchmark results on QM9. Elign improves molecule stability from 82.00% (EDM) to 93.70% and increases V×U from 90.70% to 95.31%. Compared with DFT-based RL (EDM+DFT), Elign matches atom stability and achieves higher V×U (95.31% vs. 92.87%) without DFT queries during training or sampling.

GEOM-drug. Table 3 evaluates larger drug-like molecules. Elign reaches 89.43% atom stability and 99.40% validity, improving over EDM (81.3% / 91.9%) and over RLPF (xTB rewards) on atom stability (87.94% vs. 87.52%).

Ablation analysis. We analyze the contribution of individual components in Elign through controlled ablations on QM9 (Table 2). Our default configuration achieves the strongest overall validity×uniquenessy tradeoff across metrics, with validity×uniqueness 95.31% (V×U). Replacing dense PBRS with sparse terminal rewards reduces V×U from 95.31% to 93.85%, suggesting that intermediate reward signals improve optimization effectiveness over long diffusion horizons. Substituting the larger MLFF (UMA-1p1-M) with its smaller counterpart (UMA-1p1-S) exposes a stability–diversity tradeoff: although the smaller model slightly improves molecule stability (94.83% vs. 93.70%), it leads to a marked decrease in uniqueness, resulting in a 4.5 percentage point reduction in V×U. A plausible explanation is that the smaller model induces a less smooth energy landscape, yielding higher-variance gradients that concentrate samples around a limited set of local minima. Further decomposing the reward highlights the complementary roles of its components. Energy-only rewards result in low molecule stability (86.00%), while force-only rewards achieve higher stability but reduced diversity (V×U of 90.21%). Together, these results indicate that jointly optimizing energy and

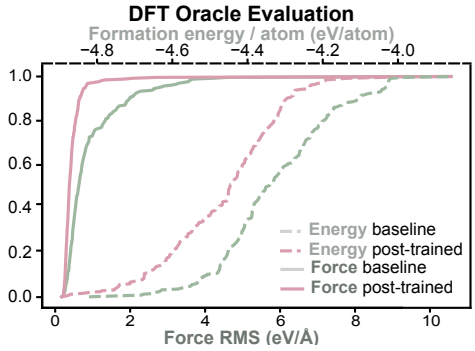

Figure 4. DFT oracle evaluation of generated molecules from QM9. Cumulative distribution functions comparing baseline and post-trained models on force RMS (left) and formation energy per atom (right). The results are obtained from 256 samples.

force objectives, combined with dense reward shaping, is important for balancing stability and diversity.

DFT Oracle Evaluation. To verify that Elign's improvements reflect genuine physical quality rather than exploiting the MLFF reward, we evaluate generated molecules using independent DFT calculations. Specifically, we use PySCF (Sun et al., 2020) with B3LYP/6-31G(d) to compute total energies and nuclear gradients, then report formation energy per atom (total energy minus isolated atom energies, divided by atom count). Figure 4 shows cumulative distributions for baseline EDM and Elign-trained models. Post-training shifts both distributions toward improved values: reduced force RMS indicates conformations closer to stationary points, while lower formation energy reflects increased thermodynamic stability. Because these metrics come from a DFT oracle not used during training, the results confirm that Elign improves physical fidelity beyond standard stability metrics.

Speed. Figure 3 reports wall-clock scaling. DFT evaluation is orders of magnitude slower than MLFF/UMA and scales steeply with system size; xTB also becomes expensive beyond ∼150–200 atoms. This observation supports the first level of amortization in our framework. During inference, run-time guidance is slower because it requires oracle gradients at each denoising step, whereas Elign uses the unguided EDM sampling procedure; across 60–200 atoms, post-trained sampling is about 8–16× faster in our benchmark. This supports the second level of amortization.

5. Related Work.

Diffusion Models for Molecules. For molecular generation, equivariant diffusion have achieved strong geometric fidelity while respecting symmetries (Hoogeboom et al., 2022; Xu et al., 2023; Cornet et al., 2025; Peng et al., 2023; Song et al., 2023; 2024; Feng et al., 2024; Qiang et al., 2023). Extensions incorporating physical guidance or control include energy-informed priors (Wu et al., 2022; Shen et al., 2024), reinforcement learning with DFT (Zhou et al., 2025), and stochastic control formulations such as Adjoint Matching and Adjoint Schrödinger Bridge Samplers (Havens et al., 2025; Liu et al., 2025a). While effective in biasing samples toward desired properties, these methods typically rely on sparse rewards, differentiable objectives, or run-time oracle evaluations, potentially limiting their scalability. We also include an expanded related-work appendix.

Table 1. QM9 3D generation. Atom stability (A), molecule stability (M), RDKit validity (V), and Validity×Uniqueness (V×U).

Model	A [%] ↑	M [%] ↑	V [%] ↑	V×U [%] ↑
EDM (Hoogeboom et al., 2022)	98.70	82.00	91.90	90.70
EDM + DPM Solver++ (Lu et al., 2022)	95.72	72.23	87.82	87.82
EDM-BRIDGE (Wu et al., 2022)	98.80	84.60	92.00	90.70
GeoLDM (Xu et al., 2023)	98.90	89.40	93.80	92.70
EDN (Cornet et al., 2025)	98.90	89.10	94.80	92.60
UniGEM (Feng et al., 2024)	99.00	89.80	95.00	93.20
GeoBFN (Song et al., 2024)	99.08	90.87	95.31	92.96
RLPF (EDM + DFT PPO) (Zhou et al., 2025)	99.08	93.37	98.22	92.87
EDM + Rejection Sampling (Zhou et al., 2025)	98.99	89.47	93.20	92.60
GeoLDM + xTB Guidance (Shen et al., 2024)	99.02	90.60	91.40	91.40
EDM + Soft Metropolis–Hastings (Feng et al., 2025)	**99.56**	91.70	**98.70**	83.20
GeoLDM + Soft Metropolis–Hastings (Feng et al., 2025)	99.09	91.30	95.10	94.90
EDM + UMA-1p1-S Guidance	98.90	87.00	92.54	92.34
EDM + DPM Solver++ + UMA-1p1-S Guidance	97.72	76.23	89.84	89.84
Elign (EDM + UMA-1p1-M)	99.33	**93.70**	98.32	**95.31**
Data (Ground Truth)	99.00	95.20	97.70	97.70

Table 2. QM9 ablations. We vary reward composition (E vs F), reward density (Sparse terminal vs Dense PBRS shaping), and MLFF capacity (UMA-1p1-S vs UMA-1p1-M).

Reward			MLFF					
E	F	PBRS	S	M	A [%] ↑	M [%] ↑	V [%] ↑	V×U [%] ↑
✓	✓	✓		✓	99.33	93.70	**98.32**	**95.31**
✓	✓	✓	✓		**99.42**	94.83	97.53	90.81
✓	✓			✓	99.02	93.75	96.54	93.85
✓				✓	98.70	86.00	91.70	91.70
	✓			✓	99.40	**94.92**	96.84	90.21

Table 3. GEOM-drug 3D generation. Atom stability (A) and RDKit validity (V).

Model	A [%] ↑	V [%] ↑
EDM (Hoogeboom et al., 2022)	81.3	91.9
EDM-BRIDGE (Wu et al., 2022)	82.4	91.9
GeoLDM (Xu et al., 2023)	84.4	99.3
EDN (Cornet et al., 2025)	87.0	92.9
UniGEM (Feng et al., 2024)	85.1	98.4
GeoBFN (Song et al., 2024)	85.6	92.08
RLPF (EDM+xTB (Zhou et al., 2025) PPO)	87.52	99.20
Elign (EDM+ UMA-1p1-M)	**87.94**	**99.40**
Data (Ground Truth)	–	86.5

6. Conclusion & Outlook

We presented Elign, a framework that post-trains equivariant molecular diffusion models with physical preferences using a pretrained foundational MLFF. By casting reverse diffusion as an RL problem and optimizing a force–energy disentangled GRPO objective, Elign improves thermodynamic stability and mechanical equilibrium while preserving fast, unguided inference. In some settings, we observe improved reward values without corresponding gains in chemical validity, suggesting reward hacking; developing reliable diagnostics and mitigation strategies is an important direction for future work.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Amin, I., Raja, S., and Krishnapriyan, A. Towards fast, specialized machine learning force fields: Distilling foundation models via energy Hessians. *arXiv preprint arXiv:2501.09009*, 2025.

Axelrod, S. and Gómez-Bombarelli, R. Geom, energy-annotated molecular conformations for property prediction and molecular generation. *Scientific Data*, 9(1):185, 2022. doi: 10.1038/s41597-022-01288-4. URL https://doi.org/10.1038/s41597-022-01288-4.

Aykent, S. and Xia, T. GotenNet: Rethinking Efficient 3D Equivariant Graph Neural Networks. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=5wxCQDtbMo.

Bannwarth, C., Ehlert, S., and Grimme, S. 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.

Barroso-Luque, L., Shuaibi, M., Fu, X., Wood, B. M., Dzamba, M., Gao, M., Rizvi, A., Zitnick, C. L., and Ulissi, Z. W. Open materials 2024 (omat24) inorganic materials dataset and models. *arXiv preprint arXiv:2410.12771*, 2024.

Batatia, I., Kovacs, D. P., Simm, G. N. C., Ortner, C., and Csanyi, G. MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=YPpSngE-ZU.

Batzner, S., Musaelian, A., Sun, L., Geiger, M., Mailoa, J. P., Kornbluth, M., Molinari, N., Smidt, T. E., and Kozinsky, B. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications*, 13(1):2453, 2022.

Behler, J. Four generations of high-dimensional neural network potentials. *Chemical Reviews*, 121(16):10037–10072, 2021.

Black, K., Janner, M., Du, Y., Kostrikov, I., and Levine, S. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2024.

Car, R. and Parrinello, M. Unified approach for molecular dynamics and density-functional theory. *Physical Review Letters*, 55(22):2471, 1985.

Chanussot, L., Das, A., Goyal, S., Lavril, T., Shuaibi, M., Riviere, M., Tran, K., Heras-Domingo, J., Ho, C., Hu, W., et al. Open catalyst 2020 (oc20) dataset and community challenges. *Acs Catalysis*, 11(10):6059–6072, 2021.

Chen, C., Ye, W., Zuo, Y., Zheng, C., and Ong, S. P. Graph networks as a universal machine learning framework for molecules and crystals. *Chemistry of Materials*, 31(9): 3564–3572, 2019.

Chmiela, S., Tkatchenko, A., Sauceda, H. E., Poltavsky, I., Schütt, K. T., and Müller, K.-R. Machine learning of accurate energy-conserving molecular force fields. *Science Advances*, 3(5):e1603015, 2017.

Chmiela, S., Sauceda, H. E., Müller, K.-R., and Tkatchenko, A. Towards exact molecular dynamics simulations with machine-learned force fields. *Nature Communications*, 9 (1):1–10, 2018.

Cornet, F., Bartosh, G., Schmidt, M. N., and Naesseth, C. A. Equivariant neural diffusion for molecule generation. *arXiv preprint arXiv:2506.10532*, 2025.

Du, W., Zhang, H., Du, Y., Meng, Q., Chen, W., Zheng, N., Shao, B., and Liu, T.-Y. SE(3)-equivariant graph neural networks with complete local frames. In *International Conference on Machine Learning*, pp. 5583–5608. PMLR, 2022.

Eastman, P., Behara, P. K., Dotson, D. L., Galvelis, R., Herr, J. E., Horton, J. T., Mao, Y., Chodera, J. D., Pritchard, B. P., Wang, Y., et al. Spice, a dataset of drug-like molecules and peptides for training machine learning potentials. *Scientific Data*, 10(1):11, 2023.

Feng, H., Qiu, P., Zhang, M.-C., Fan, Y., Tao, Y., and Poczos, B. Soft metropolis-hastings correction for generative model sampling. *bioRxiv preprint*, 2025.

Feng, R., Zhu, Q., Tran, H., Chen, B., Toland, A., Ramprasad, R., and Zhang, C. May the force be with you: Unified force-centric pre-training for 3D molecular conformations. *arXiv preprint arXiv:2308.14759*, 2023.

Feng, S., Zhou, Y., Li, Y., Wang, Y., Quan, Y., Ju, P., Chen, W., Wang, Y., Zheng, M., Liu, J., et al. Unigem: A unified approach to generation and property prediction for molecules. *arXiv preprint arXiv:2410.10516*, 2024.

Fu, X., Wood, B. M., Barroso-Luque, L., Levine, D. S., Gao, M., Dzamba, M., and Zitnick, C. L. Learning smooth and expressive interatomic potentials for physical property prediction. In *Forty-second International Conference on Machine Learning*, 2025.

Fuchs, F., Worrall, D., Fischer, V., and Welling, M. SE(3)-transformers: 3D roto-translation equivariant attention networks. *Advances in Neural Information Processing Systems*, 33:1970–1981, 2020.

Gasteiger, J., Groß, J., and Günnemann, S. Directional message passing for molecular graphs. *arXiv preprint arXiv:2003.03123*, 2020.

Gasteiger, J., Becker, F., and Günnemann, S. Gemnet: Universal directional graph neural networks for molecules. *Advances in Neural Information Processing Systems*, 34: 6790–6802, 2021.

Hart, P. E., Nilsson, N. J., and Raphael, B. A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics*, 4(2):100–107, 1968.

Havens, A. J., Miller, B. K., Yan, B., Domingo-Enrich, C., Sriram, A., Levine, D. S., Wood, B. M., Hu, B., Amos, B., Karrer, B., Fu, X., Liu, G.-H., and Chen, R. T. Q. Adjoint sampling: Highly scalable diffusion samplers via adjoint matching. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=6Eg1OrHmg2.

Hohenberg, P. and Kohn, W. Inhomogeneous electron gas. *Physical Review*, 136(3B):B864, 1964.

Hoogeboom, E., Garcia Satorras, V., Vignac, C., and Welling, M. Equivariant diffusion for molecule generation in 3D. In *International Conference on Machine Learning*, pp. 8867–8887. PMLR, 2022.

Isert, C., Atz, K., Jiménez-Luna, J., and Schneider, G. Qmugs, quantum mechanical properties of drug-like molecules. *Scientific Data*, 9(1):273, 2022.

Kohn, W. and Sham, L. J. Self-consistent equations including exchange and correlation effects. *Physical Review*, 140(4A):A1133, 1965.

Levine, D. S., Shuaibi, M., Spotte-Smith, E. W. C., Taylor, M. G., Hasyim, M. R., Michel, K., Batatia, I., Csányi, G., Dzamba, M., Eastman, P., et al. The open molecules 2025 (omol25) dataset, evaluations, and models. *arXiv preprint arXiv:2505.08762*, 2025.

Li, Y., Wang, Y., Huang, L., Yang, H., Wei, X., Zhang, J., Wang, T., Wang, Z., Shao, B., and Liu, T.-Y. Long-short-range message-passing: A physics-informed framework to capture non-local interaction for scalable molecular dynamics simulation. *arXiv preprint arXiv:2304.13542*, 2023.

Li, Y., Xia, Z., Huang, L., Wei, X., Harshe, S., Yang, H., Luo, E., Wang, Z., Zhang, J., Liu, C., et al. Enhancing the scalability and applicability of Kohn-Sham Hamiltonians for molecular systems. In *The Thirteenth International Conference on Learning Representations*, 2024.

Li, Y., Huang, L., Ding, Z., Wei, X., Wang, C., Yang, H., Wang, Z., Liu, C., Shi, Y., Jin, P., et al. E2former: An efficient and equivariant transformer with linear-scaling tensor products. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/pdf?id=ls5L4IMEwt.

Liao, Y.-L. and Smidt, T. Equiformer: Equivariant graph attention transformer for 3D atomistic graphs. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=KwmPfARgOTD.

Liu, G.-H., Choi, J., Chen, Y., Miller, B. K., and Chen, R. T. Q. Adjoint schrödinger bridge sampler. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a. URL https://openreview.net/forum?id=rMhQBlhh4c.

Liu, J., Liu, G., Liang, J., Li, Y., Liu, J., Wang, X., Wan, P., Zhang, D., and Ouyang, W. Flow-GRPO: Training flow matching models via online RL. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b. URL https://openreview.net/forum?id=oCBKGw5HNf.

Liu, Y., Wang, L., Liu, M., Lin, Y., Zhang, X., Oztekin, B., and Ji, S. Spherical message passing for 3D molecular graphs. In *International Conference on Learning Representations (ICLR)*, 2022.

Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., and Zhu, J. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022.

Mannan, S., Bihani, V., Gonzales, C., Lee, K. L. K., Gosvami, N. N., Ranu, S., Miret, S., and Krishnan, N. Evaluating universal machine learning force fields against experimental measurements. *arXiv preprint arXiv:2508.05762*, 2025.

Musaelian, A., Batzner, S., Johansson, A., Sun, L., Owen, C. J., Kornbluth, M., and Kozinsky, B. Learning local equivariant representations for large-scale atomistic dynamics. *Nature Communications*, 14(1):579, 2023.

Nelson, E. *Dynamical Theories of Brownian Motion*. Princeton University Press, Princeton, NJ, 1967.

Ng, A. Y., Harada, D., and Russell, S. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the International Conference on Machine Learning*, pp. 278–287, 1999.

Passaro, S. and Zitnick, C. L. Reducing SO(3) convolutions to SO(2) for efficient equivariant GNNs. In *International Conference on Machine Learning*, pp. 27420–27438. PMLR, 2023.

Peng, X., Guan, J., Liu, Q., and Ma, J. Moldiff: Addressing the atom-bond inconsistency problem in 3d molecule diffusion generation. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 27611–27629. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/peng23b.html.

Qiang, B., Song, Y., Xu, M., Gong, J., Gao, B., Zhou, H., Ma, W.-Y., and Lan, Y. Coarse-to-fine: a hierarchical diffusion model for molecule generation in 3d. In *International Conference on Machine Learning*, pp. 28277–28299. PMLR, 2023.

Qu, E. and Krishnapriyan, A. The importance of being scalable: Improving the speed and accuracy of neural network interatomic potentials across chemical domains. *Advances in Neural Information Processing Systems*, 37: 139030–139053, 2024.

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

Satorras, V. G., Hoogeboom, E., and Welling, M. E(n)-equivariant graph neural networks. In *International Conference on Machine Learning*, pp. 9323–9332. PMLR, 2021.

Schütt, K. T., Arbabzadah, F., Chmiela, S., Müller, K. R., and Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nature Communications*, 8 (1):1–8, 2017.

Schütt, K. T., Sauceda, H. E., Kindermans, P.-J., Tkatchenko, A., and Müller, K.-R. Schnet–a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24):241722, 2018.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y. K., Wu, Y., and Guo, D. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Shen, Y., Zhang, C., Fu, S., Zhou, C., Washburn, N., and Poczos, B. Chemistry-inspired diffusion with non-differentiable guidance. In *The Thirteenth International Conference on Learning Representations*, 2024.

Simeon, G. and De Fabritiis, G. Tensornet: Cartesian tensor representations for efficient learning of molecular potentials. *Advances in Neural Information Processing Systems*, 36, 2024.

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

Song, Y., Gong, J., Zhou, H., Zheng, M., Liu, J., and Ma, W.-Y. Unified generative modeling of 3D molecules via Bayesian Flow Networks. In *International Conference on Learning Representations*, 2024.

Spall, J. C. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*, 37(3):332–341, 1992.

Sriram, A., Choi, S., Yu, X., Brabson, L. M., Das, A., Ulissi, Z., Uyttendaele, M., Medford, A. J., and Sholl, D. S. The open dac 2023 dataset and challenges for sorbent discovery in direct air capture, 2024.

Sun, Q., Zhang, X., Banerjee, S., Bao, P., Barbry, M., Blunt, N. S., Bogdanov, N. A., Booth, G. H., Chen, J., Cui, Z.-H., et al. Recent developments in the pyscf program package. *The Journal of chemical physics*, 153(2), 2020.

Tran, R., Lan, J., Shuaibi, M., Wood, B. M., Goyal, S., Das, A., Heras-Domingo, J., Kolluru, A., Rizvi, A., Shoghi, N., et al. The open catalyst 2022 (oc22) dataset and challenges for oxide electrocatalysts. *ACS Catalysis*, 13 (5):3066–3084, 2023.

Unke, O. T., Chmiela, S., Sauceda, H. E., Gastegger, M., Poltavsky, I., Schütt, K. T., Tkatchenko, A., and Muüller, K.-R. Machine learning force fields. *Chemical Reviews*, 121(16):10142–10186, 2021.

Unke, O. T., Stöhr, M., Ganscha, S., Unterthiner, T., Maennel, H., Kashubin, S., Ahlin, D., Gastegger, M., Medrano Sandonas, L., Berryman, J. T., et al. Biomolecular dynamics with machine-learned quantum-mechanical force fields trained on diverse chemical fragments. *Science Advances*, 10(14):eadn4397, 2024.

Wang, L., Liu, Y., Lin, Y., Liu, H., and Ji, S. ComENet: Towards complete and efficient message passing for 3D molecular graphs. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=mCzMqeWSFJ.

Wood, B. M., Dzamba, M., Fu, X., Gao, M., Shuaibi, M., Barroso-Luque, L., Abdelmaqsoud, K., Gharakhanyan, V., Kitchin, J. R., Levine, D. S., et al. Uma: A family of universal models for atoms. *arXiv preprint arXiv:2506.23971*, 2025.

Wu, L., Gong, C., Liu, X., Ye, M., and Liu, Q. Diffusion-based molecule generation with informative prior bridges. In *Advances in Neural Information Processing Systems*, volume 35, pp. 36533–36545, 2022.

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

Xue, Z., Wu, J., Gao, Y., Kong, F., Zhu, L., Chen, M., Liu, Z., Liu, W., Guo, Q., Huang, W., and Luo, P. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025.

Zaidi, S., Schaarschmidt, M., Martens, J., Kim, H., Teh, Y. W., Sanchez-Gonzalez, A., Battaglia, P., Pascanu, R., and Godwin, J. Pre-training via denoising for molecular property prediction. In *The Eleventh International Conference on Learning Representations*, 2022.

Zhou, Z., An, J., Liu, Z., Shi, Y., Zhang, X., Cao, F., Qu, C., and Qi, Y. Guiding diffusion models with reinforcement learning for stable molecule generation. *arXiv preprint arXiv:2508.16521*, 2025.

Appendix Table of Contents

A. Hyperparameters

A.1. RL post-training

Table 4. Hyperparameters for RL postraining

Notation	Description	Value
learning_rate	Learning rate for policy optimization	2e-7 to 4e-6
clip_range	PPO clipping threshold	2e-3
train_micro_batch_size	Micro-batch size for policy updates	4
epoch_per_rollout	Number of optimization epochs per rollout	{1, 3}
kl_penalty_weight	Weight of KL regularization that constrains the policy update to stay close to the reference model	0.04
discount	Discount factor used in calculating intermediate shaping rewards	1
weight_energy	Weighting factor for energy reward	{0.05 0.1}
weight_force	Weighting factor for force reward	1
sample_group_size	Number of sample groups generated per prompt	4
each_prompt_sample	Number of trajectories generated per prompt	24
time_step	Number of diffusion timesteps used during sampling and rollout	1000
mlff_model	Identifier of the machine-learned force field (MLFF) model	uma-m-1p1 or uma-s-1p1
mlff_batch_size	Batch size for reward evaluation	8
force_aggregation	Aggregation method for force errors	rms or max
skip_prefix	Number of initial steps skipped when applying reward shaping	{600,700}
scheduler_name	Learning rate scheduler type	cosine
scheduler_warmup_steps	Number of warm-up steps during which the learning rate increases from zero to the initial value	60
scheduler_total_steps	Total number of scheduler steps before learning rate decay completes	1500
scheduler_min_lr_ratio	Final learning rate expressed as a fraction of the initial learning rate after decay	0.3
KL weight	KL divergence with respect to a reference pretrained policy	{0, 0.02 }

A.2. QM9

Table 5. Hyperparameters used for the pretrained model on QM9 molecular dataset

Notation	Description	Value
model	Dynamics model type	egnn_dynamics
probabilistic_model	Probabilistic model type	diffusion
diffusion_steps	Number of diffusion steps	500
diffusion_noise_schedule	Noise schedule used in diffusion	polynomial_2
diffusion_loss_type	Loss function for diffusion training	vlb, l2
diffusion_noise_precision	Minimum diffusion noise precision used for numerical stability	1e-5
n_epochs	Total number of training epochs	200
batch_size	Batch size used for training	128
lr	Learning rate for the optimizer	0.0002
n_layers	Number of EGNN layers	6
inv_sublayers	Number of invariant sublayers per EGNN layer	1
nf	Hidden feature dimension (node feature size)	128
tanh	Whether to use tanh activation in coordinate MLPs	True
attention	Whether to use attention mechanisms in EGNN layers	True
norm_constant	Normalization constant in coordinate updates	1
sin_embedding	Whether to use sinusoidal time embeddings	False
ode_regularization	Regularization strength for ODE-based formulations	1e-3
dataset	Dataset to use	qm9 or qm9_second_half
filter_n_atoms	Restrict dataset to molecules with a fixed number of atoms	None
dequantization	Dequantization strategy	argmax_variational
ema_decay	Amount of Exponential Moving Average (EMA) decay, 0 means off, a reasonable value is 0.999	0.999
n_stability_samples	Number of samples to compute the stability	500
normalize_factors	Normalization factors for x, categorical, and integer features	1, 4, 1
include_charges	Whether to include atom charge	True
normalization_factor	Normalize the sum aggregation of EGNN	1
aggregation_method	Aggregation for the graph network	sum or mean

Table 6. Hyperparameters used for the pretrained model on the GEOM Drugs dataset

Notation	Description	Value
model	Dynamics model type	egnn_dynamics
probabilistic_model	Probabilistic model type	diffusion
diffusion_steps	Number of diffusion steps	500
diffusion_noise_schedule	Noise schedule used in diffusion	polynomial_2
diffusion_loss_type	Loss function for diffusion training	l2
diffusion_noise_precision	Minimum diffusion noise precision used for numerical stability	1e-5
n_epochs	Total number of training epochs	10000
batch_size	Batch size used for training	64
lr	Learning rate for the optimizer	5e-5
n_layers	Number of EGNN layers	6
inv_sublayers	Number of invariant sublayers per EGNN layer	1
nf	Hidden feature dimension (node feature size)	192
tanh	Whether to use tanh activation in coordinate MLPs	True
attention	Whether to use attention mechanisms in EGNN layers	True
norm_constant	Normalization constant in coordinate updates	1
sin_embedding	Whether to use sinusoidal time embeddings	False
ode_regularization	Regularization strength for ODE-based formulations	1e-3
dataset	Dataset to use	geom
filter_n_atoms	Restrict dataset to molecules with a fixed number of atoms	None
dequantization	Dequantization strategy	argmax_ variational
ema_decay	Amount of Exponential Moving Average (EMA) decay, 0 means off, a reasonable value is 0.999	0
n_stability_samples	Number of samples to compute the stability	20
normalize_factors	Normalization factors for x, categorical, and integer features	1, 4, 10
include_charges	Whether to include atom charge	False
normalization_factor	Normalize the sum aggregation of EGNN	100
aggregation_method	Aggregation for the graph network	sum or mean

A.3. GEOM-Drugs

B. Notations

Symbol	Type	Description
Geometric states and molecular structure		
$\mathbf{x}_0$	$\mathbb{R}^{N\times 3}$	Clean atomic coordinates
$\mathbf{x}_t$	$\mathbb{R}^{N\times 3}$	Noisy conformation at diffusion time t
N	$\mathbb{N}$	Number of atoms
$\mathcal{X}$	space	Configuration space of clean (CoM free) molecular structures
z	$\mathcal{X}$	Generic clean structure (integration variable in appendix proofs)
$\mathbf{z}_t$	$\mathbb{R}^{N\times 3}\times\mathbb{R}^{N\times d_h}$	Diffusion state (coordinates plus hidden features)
Diffusion process and generative model		
$p_t(\mathbf{z})$	distribution	Marginal distribution at time t
$p_\theta(\mathbf{z}_{t-1}\mid\mathbf{z}_t)$	distribution	Reverse diffusion transition
$s_\theta(\mathbf{z}_t,t)$	$\mathbb{R}^{N\times 3}$	Score function $\nabla_{\mathbf{z}_t}\log p_t(\mathbf{z}_t)$
$\boldsymbol{\epsilon}$	$\mathbb{R}^{N\times 3}$	Standard Gaussian noise
α_t	$\mathbb{R}$	Signal coefficient
σ_t	$\mathbb{R}_{>0}$	Noise scale
Trajectories and expectations		
τ	sequence	Reverse diffusion trajectory $(\mathbf{z}_T,\dots,\mathbf{z}_0)$
$\mathbb{E}$	operator	Expectation over samples or trajectories
$\mathcal{T}$	set	Set of sampled trajectories
Energy, forces, and physical models		
$E(z)$	$\mathbb{R}$	Potential energy of a clean structure z
$\mathbf{F}(z)$	$\mathbb{R}^{N\times 3}$	Force field $-\nabla_z E(z)$
$E_\star(z)$	$\mathbb{R}$	Target energy (gold standard reference)
$E_\phi(z)$	$\mathbb{R}$	MLFF energy approximation
$\mathbf{F}_\star(z)$	$\mathbb{R}^{N\times 3}$	Target forces $-\nabla_z E_\star(z)$
$\mathbf{F}_\phi(z)$	$\mathbb{R}^{N\times 3}$	MLFF forces $-\nabla_z E_\phi(z)$
Reinforcement learning and FED GRPO		
π_θ	policy	Reverse diffusion policy
r_t^E	$\mathbb{R}$	Energy based reward
r_t^F	$\mathbb{R}$	Force based stability reward
r_t	$\mathbb{R}$	Total reward
$\hat{A}_t$	$\mathbb{R}$	Advantage estimate
$\mathcal{L}_{\text{GRPO}}$	objective	Group Relative Policy Optimization loss
$\mathcal{L}_{\text{FED-GRPO}}$	objective	Force Energy Disentangled GRPO loss
Measure theoretic and variational notation (appendix proofs)		
$\mathcal{P}(\mathcal{X})$	space	Probability measures on $\mathcal{X}$
μ	$\mathcal{P}(\mathcal{X})$	Pretrained terminal law, $\mu:=\rho_{\theta_{\text{pre}}}$
ρ	$\mathcal{P}(\mathcal{X})$	Candidate terminal law
$\rho\ll\mu$	relation	Absolute continuity of ρ with respect to μ
$\mathrm{KL}(\rho\|\mu)$	$\mathbb{R}_{\geq 0}$	Kullback Leibler divergence
$\mathcal{J}(\rho)$	functional	Variational objective over terminal laws
w_E	$\mathbb{R}_{>0}$	Energy weight in advantage calculation
w_{KL}	$\mathbb{R}_{>0}$	KL weight in $\mathcal{J}$
β_{eff}	$\mathbb{R}_{>0}$	Effective inverse temperature, $\beta_{\text{eff}}:=w_E/w_{\text{KL}}$
$\rho^\star$	$\mathcal{P}(\mathcal{X})$	Maximizer of $\mathcal{J}$
Z_ϕ	$\mathbb{R}_{>0}$	Partition function for $\rho^\star$ under E_ϕ
$Z_\star$	$\mathbb{R}_{>0}$	Partition function for the tilt under $E_\star$
f	density	Radon Nikodym derivative $f:=d\rho/d\mu$
λ	$\mathbb{R}$	Lagrange multiplier for normalization of f
$\|\cdot\|_{\text{TV}}$	$\mathbb{R}_{\geq 0}$	Total variation distance
δ	$\mathbb{R}_{\geq 0}$	Uniform energy error bound, $\sup_{z\in\mathcal{X}}\|E_\phi(z)-E_\star(z)\|\leq\delta$
Symmetry groups and spaces		
$E(3)$	group	Euclidean group of rotations and translations
$\mathrm{SO}(3)$	group	Rotation group
$\mathbb{R}^{N\times 3}$	space	Cartesian coordinate space

Table 7. Summary of symbols and notation used throughout the paper and appendix.

C. Related Works

Diffusion Models for Molecules. E(3)-equivariant diffusion models have quickly become a leading paradigm for 3D molecular generation, as they inherently enforce rotational and translational symmetries and produce geometries with high fidelity. (Hoogeboom et al., 2022) introduced an SE(3)-equivariant diffusion process that generates molecules as sets of atoms with remarkable realism. Building on this foundation, (Xu et al., 2023) developed GeoLDM, a latent 3D diffusion approach that operates in a learned point-cloud latent space composed of both invariant scalars and equivariant tensors, improving sample efficiency and controllable generation of molecular geometries. More recently, (Cornet et al., 2025) proposed an Equivariant Neural Diffusion (END) model that also learns the forward-noising process, achieving state-of-the-art results on standard molecular generation benchmarks. Beyond unconditional generation, many methods aim to guide or bias diffusion models toward molecules with desired properties or constraints. One strategy is to inject domain knowledge via physically-informed priors. (Wu et al., 2022) introduces informative prior bridges that steer the diffusion trajectory toward low-energy conformations by constructing a biased intermediate distribution for the sampler. (Zhou et al., 2025) formulates 3D molecule generation as a Markov decision process and uses proximal policy optimization to fine-tune a pre-trained equivariant diffusion model's sampling policy by providing a reward based on physics evaluations. An alternative formalism is to cast guided generation as a stochastic optimal control problem. (Havens et al., 2025) introduces Adjoint Sampling, which optimizes a matching objective for the diffusion's probability flow by learning an optimal diffusion drift that samples from an unnormalized target density, such as a Boltzmann distribution. (Liu et al., 2025a) generalizes this idea by relaxing previous assumptions on the prior, solving a Schrödinger Bridge between a simple base distribution and the desired Boltzmann-like distribution.

Machine-learning Force-Field. Machine learning force fields (MLFFs) use geometric deep learning to approximate potential energy surfaces and interatomic forces with high accuracy at far lower cost than quantum chemistry. Early MLFFs relied on invariant message passing networks that predict an energy scalar invariant to rotations and translations, with forces obtained via differentiation. SchNet established this paradigm using continuous filter convolutions over interatomic distances (Schütt et al., 2018). More generally, graph network formalisms demonstrated that message passing can unify molecular and crystalline property prediction under a single framework (Chen et al., 2019). A major line of progress then focused on injecting richer geometric structure into invariant architectures. DimeNet introduced directional message passing through angular features and specialized basis expansions (Gasteiger et al., 2020), and GemNet improved higher order interaction modeling to better capture many-body effects relevant for energies and forces (Gasteiger et al., 2021). SphereNet further leveraged a spherical coordinate parameterization for local neighborhoods to encode distances and angles in a unified way (Liu et al., 2022), while ComENet emphasized complete yet efficient geometric message passing to improve accuracy without sacrificing scaling (Wang et al., 2022). Recent analyses underscore that architecture and training data choices critically determine transfer across chemical domains and system sizes, motivating careful study of inductive biases and data coverage (Qu & Krishnapriyan, 2024). In parallel, E(3) equivariant MLFFs enforce symmetry directly in intermediate representations, enabling vector and tensor features that transform correctly under rotations and translations. EGNN provides a lightweight recipe for equivariant message passing with coordinate updates (Satorras et al., 2021), and complete local frame constructions strengthen SE(3) equivariance and stability by anchoring messages in local geometric frames (Du et al., 2022). Tensor-based designs such as TensorNet offer an efficient Cartesian tensor alternative to spherical harmonic pipelines while preserving equivariant structure (Simeon & De Fabritiis, 2024), and newer architectures focus explicitly on practical efficiency, such as GotenNet (Aykent & Xia, 2025). Equivariant graph networks have also been used as surrogates for more demanding electronic structure components, improving the scalability of learned approximations to Kohn-Sham DFT quantities while maintaining symmetry constraints (Li et al., 2024). A particularly influential family of MLFFs builds on $\mathrm{SO}(3)$ representation theory with spherical harmonics. SE(3) Transformer introduced equivariant attention with irreducible representations (Fuchs et al., 2020), and Equiformer extended this transformer style approach with strong results on quantum chemistry benchmarks (Liao & Smidt, 2023). NequIP demonstrated striking data efficiency for learning interatomic potentials with spherical tensor message passing (Batzner et al., 2022), while MACE advanced higher order equivariant interactions for accurate molecular dynamics potentials (Batatia et al., 2022). Scalability has been further improved with locally deployed equivariant potentials such as Allegro (Musaelian et al., 2023), extensions that model both long range and short range interactions (Li et al., 2023), and methods that reduce the cost of equivariant operations (Passaro & Zitnick, 2023; Li et al., 2025). Finally, recent work targets improved physical consistency and broader transfer, including learning objectives tuned for stable molecular dynamics and phonon properties (Fu et al., 2025) and universal pretrained atomistic models that aim to generalize across elements and compounds (Wood et al., 2025).

Reinforcement Learning for Diffusion Guidance. Integrating RL with diffusion models is a recent trend aimed at steering generative models toward complex objectives that are hard to encode in the training likelihood. One representative approach is Denoising Diffusion Policy Optimization (DDPO) (Black et al., 2024). DDPO reframes the multi-step denoising process as a Markov decision process, where each diffusion step is an action, and a reward is obtained at the final sample. This allows applying policy gradient algorithms to fine-tune a pretrained diffusion model for higher rewards. (Xue et al., 2025) adapts GRPO to diffusion and continuous flow models with *DanceGRPO*, achieving robust RL-based fine-tuning on large-scale text-to-image and text-to-video generation tasks. (Liu et al., 2025b) introduces *Flow-GRPO*, an online RL algorithm that trains flow matching models using reward signals by aligning generated trajectories with desired outcomes through policy gradient updates. (Shen et al., 2024) proposes a chemistry-inspired guided diffusion that uses external force-field evaluations as a form of reward to bias the diffusion sampler toward low-energy structures, without requiring those evaluations to be differentiable. (Zhou et al., 2025) applies Reinforcement Learning with Physical Feedback (RLPF) to an equivariant diffusion model, significantly improving stability of generated conformations. Crucially, both works demonstrate that RL can incorporate domain-specific criteria (like force-field energies) that lie outside the scope of the original generative model's training. Another perspective comes from control theory: *Adjoint methods* avoid explicit RL by solving optimal control formulations for diffusion. (Havens et al., 2025) derives an adjoint formulation to directly match a diffusion sampler to an optimal policy in a single backward pass, while (Liu et al., 2025a) develops an adjoint Schrödinger bridge that computes an optimal transport between the prior and target distributions. These methods achieve goals similar to RL-guided diffusion by biasing generation toward certain distributions or rewards, but do so by analytically aligning sampling trajectories, thereby improving efficiency and scalability.

D. Proofs for Section 3.2

Notation. Let $\mathcal{X}$ denote the configuration space of clean (CoM-free) molecular structures. Let $\mu := \rho_{\theta_{\text{pre}}} \in \mathcal{P}(\mathcal{X})$ be the pretrained terminal law. For any $\rho \in \mathcal{P}(\mathcal{X})$, write $\rho \ll \mu$ for absolute continuity. We use $\mathrm{KL}(\rho\|\mu)$ for Kullback–Leibler divergence. Throughout, we fix the energy weight $w_E = 1$ and define

$$\beta_{\text{eff}} := \frac{1}{w_{\text{KL}}}.$$

(Equivalently, any non-unit w_E can be absorbed into w_{KL} by rescaling.)

Theorem 3 (Energy-aligned terminal distribution (restated))**.** *Assume $w_{\text{KL}} > 0$ and consider the functional over $\rho \in \mathcal{P}(\mathcal{X})$,*

$$\mathcal{J}(\rho) := \mathbb{E}_{z\sim\rho}[-E_\phi(z)] - w_{\text{KL}}\,\mathrm{KL}(\rho\|\mu), \qquad \textit{with the convention } \mathcal{J}(\rho) = -\infty \textit{ if } \rho \not\ll \mu.$$

Assume the supremum of $\mathcal{J}$ over $\mathcal{P}(\mathcal{X})$ is attained. Then the maximizer is unique and equals

$$\rho^\star(dz) = \frac{1}{Z_\phi}\exp(-\beta_{\text{eff}}E_\phi(z))\,\mu(dz), \qquad Z_\phi := \int_{\mathcal{X}} \exp(-\beta_{\text{eff}}E_\phi(z))\,\mu(dz).$$

Proof. Let $\rho \in \mathcal{P}(\mathcal{X})$. If $\rho \not\ll \mu$, then $\mathrm{KL}(\rho\|\mu) = +\infty$ and $\mathcal{J}(\rho) = -\infty$, so we restrict to $\rho \ll \mu$. Write the Radon–Nikodym derivative $f := d\rho/d\mu$. Then $f \geq 0$ μ-a.e. and $\int f\,d\mu = 1$. Moreover,

$$\mathrm{KL}(\rho\|\mu) = \int_{\mathcal{X}} f(z)\log f(z)\,\mu(dz), \qquad \mathbb{E}_{z\sim\rho}[E_\phi(z)] = \int_{\mathcal{X}} E_\phi(z)\,f(z)\,\mu(dz).$$

Thus maximizing $\mathcal{J}(\rho)$ over $\rho \ll \mu$ is equivalent to maximizing over densities f:

$$\mathcal{J}(f) = -\int_{\mathcal{X}} E_\phi(z)\,f(z)\,\mu(dz) - w_{\text{KL}}\int_{\mathcal{X}} f(z)\log f(z)\,\mu(dz), \quad \text{s.t.} \int_{\mathcal{X}} f(z)\,\mu(dz) = 1,\ f \geq 0.$$

Introduce a Lagrange multiplier $\lambda \in \mathbb{R}$ for the normalization constraint and define

$$\mathcal{L}(f,\lambda) = -\int E_\phi(z)f(z)\,\mu(dz) - w_{\text{KL}}\int f(z)\log f(z)\,\mu(dz) + \lambda\Big(\int f(z)\,\mu(dz) - 1\Big).$$

Let h be any bounded perturbation with $\int h\,d\mu = 0$ and consider $f_\varepsilon = f + \varepsilon h$. A first-order stationarity condition at an optimum $f^\star$ implies $\frac{d}{d\varepsilon}\mathcal{L}(f^\star + \varepsilon h, \lambda)|_{\varepsilon=0} = 0$ for all such h. Using $\frac{d}{d\varepsilon}[(f^\star + \varepsilon h)\log(f^\star + \varepsilon h)]|_{\varepsilon=0} = h(1 + \log f^\star)$, we obtain

$$0 = \int_{\mathcal{X}} h(z)\Big(-E_\phi(z) - w_{\text{KL}}(1 + \log f^\star(z)) + \lambda\Big)\mu(dz).$$

Since this holds for all h with zero μ-mean, the bracketed term must be μ-a.e. constant, i.e.

$$-E_\phi(z) - w_{\mathrm{KL}}(1 + \log f^\star(z)) + \lambda = 0 \quad \text{for } \mu\text{-a.e. } z.$$

Rearranging gives

$$\log f^\star(z) = -\frac{1}{w_{\mathrm{KL}}} E_\phi(z) + c = -\beta_{\mathrm{eff}} E_\phi(z) + c$$

for some constant $c \in \mathbb{R}$, hence

$$f^\star(z) = \exp(c) \exp(-\beta_{\mathrm{eff}} E_\phi(z)).$$

Imposing $\int f^\star \, d\mu = 1$ yields $\exp(c) = 1/Z_\phi$ with

$$Z_\phi = \int_{\mathcal{X}} \exp(-\beta_{\mathrm{eff}} E_\phi(z)) \, \mu(dz),$$

so $\rho^\star(dz) = f^\star(z)\mu(dz)$ as claimed.

Uniqueness follows because $f \mapsto \int f \log f \, d\mu$ is strictly convex on $\{f \geq 0 : \int f \, d\mu = 1\}$, hence $\mathcal{J}(f)$ is strictly concave when $w_{\mathrm{KL}} > 0$. □

MLFF approximation and distribution error. Let $E_\star : \mathcal{X} \to \mathbb{R}$ be a target energy and E_ϕ its MLFF approximation. Define the two tilted laws (with the same reference μ):

$$\rho^\star_\star(dz) = \frac{1}{Z_\star} \exp(-\beta_{\mathrm{eff}} E_\star(z)) \, \mu(dz), \qquad \rho^\star_\phi(dz) = \frac{1}{Z_\phi} \exp(-\beta_{\mathrm{eff}} E_\phi(z)) \, \mu(dz),$$

where

$$Z_\star := \int_{\mathcal{X}} \exp(-\beta_{\mathrm{eff}} E_\star(z)) \, \mu(dz), \qquad Z_\phi := \int_{\mathcal{X}} \exp(-\beta_{\mathrm{eff}} E_\phi(z)) \, \mu(dz).$$

Theorem 4 (Energy error implies distribution error (restated))**.** *Assume* $\sup_{z \in \mathcal{X}} |E_\phi(z) - E_\star(z)| \leq \delta$ *for some* $\delta \geq 0$*. Then*

$$\|\rho^\star_\phi - \rho^\star_\star\|_{\mathrm{TV}} \leq \tanh(\beta_{\mathrm{eff}} \delta).$$

Lemma 1 (Likelihood-ratio bound implies TV bound)**.** *Let* P, Q *be probability measures with* $P \ll Q$ *and suppose there exists* $\varepsilon \geq 0$ *such that*

$$e^{-\varepsilon} \leq \frac{dP}{dQ} \leq e^{\varepsilon} \quad Q\text{-a.e.}$$

Then $\|P - Q\|_{\mathrm{TV}} \leq \tanh(\varepsilon/2)$.

Proof. Let $r := dP/dQ$, so $r \in [e^{-\varepsilon}, e^{\varepsilon}]$ Q-a.e. and $\mathbb{E}_Q[r] = 1$. Let $A := \{r \geq 1\}$. Then

$$\|P - Q\|_{\mathrm{TV}} = \frac{1}{2} \int |dP - dQ| = \frac{1}{2} \int |r - 1| \, dQ = \int_A (r - 1) \, dQ,$$

where the last equality uses $\int_A (r-1) \, dQ = \int_{A^c} (1 - r) \, dQ$ from $\mathbb{E}_Q[r] = 1$.

Write $q := Q(A)$ and the conditional means $m_+ := \mathbb{E}_Q[r \mid A] \in [1, e^{\varepsilon}]$, $m_- := \mathbb{E}_Q[r \mid A^c] \in [e^{-\varepsilon}, 1]$. The constraint $\mathbb{E}_Q[r] = 1$ becomes $qm_+ + (1-q)m_- = 1$, and the TV becomes $q(m_+ - 1)$. The maximum TV under the bounds is achieved at the extreme values $m_+ = e^{\varepsilon}$ and $m_- = e^{-\varepsilon}$. Solving

$$qe^{\varepsilon} + (1-q)e^{-\varepsilon} = 1 \quad \Longrightarrow \quad q = \frac{1}{e^{\varepsilon} + 1},$$

we get

$$\|P - Q\|_{\mathrm{TV}} \leq q(e^{\varepsilon} - 1) = \frac{e^{\varepsilon} - 1}{e^{\varepsilon} + 1} = \tanh(\varepsilon/2).$$

□

Proof of Theorem 4. Let $\beta := \beta_{\text{eff}}$. Define $w_\star(z) := \exp(-\beta E_\star(z))$ and $w_\phi(z) := \exp(-\beta E_\phi(z))$. By $|E_\phi(z) - E_\star(z)| \leq \delta$, for all $z \in \mathcal{X}$,

$$e^{-\beta\delta} w_\star(z) \leq w_\phi(z) \leq e^{\beta\delta} w_\star(z).$$

Integrating w.r.t. μ yields the partition-function sandwich:

$$e^{-\beta\delta} Z_\star \leq Z_\phi \leq e^{\beta\delta} Z_\star.$$

Therefore, for $\rho_\star^\star$-a.e. z,

$$\frac{d\rho_\phi^\star}{d\rho_\star^\star}(z) = \frac{Z_\star}{Z_\phi} \cdot \frac{w_\phi(z)}{w_\star(z)} \in [e^{-\beta\delta}, e^{\beta\delta}] \cdot [e^{-\beta\delta}, e^{\beta\delta}] = [e^{-2\beta\delta}, e^{2\beta\delta}].$$

Applying Lemma 1 with $P = \rho_\phi^\star$, $Q = \rho_\star^\star$ and $\varepsilon = 2\beta\delta$ gives

$$\|\rho_\phi^\star - \rho_\star^\star\|_{\text{TV}} \leq \tanh((2\beta\delta)/2) = \tanh(\beta\delta) = \tanh(\beta_{\text{eff}}\delta).$$

□

E. Supplementary Experiments

E.1. Qualitative analysis.

Figures 5–8 provide a qualitative view of how post-training changes the generated 3D geometries on GEOM-Drugs. In Figure 5, we overlay generated conformers (colored) with the reference structure (gray), highlighting that the post-trained sampler more consistently produces a coherent, well-connected geometry that visually matches the target conformer. Figure **??** complements this visualization by annotating the same examples with MLFF-predicted energies and RMS force norms: the post-trained samples typically shift toward lower energies and smaller residual forces, consistent with configurations closer to local equilibrium. Finally, Figure 8 compares matched-noise reverse trajectories (same initial noise) and shows that the post-trained model corrects failure modes that persist throughout denoising in the baseline (e.g., distorted rings/strained geometries), yielding a chemically valid final structure at $t = 0$. Together, these figures illustrate that alignment not only improves terminal metrics but also changes the denoising dynamics in a way that stabilizes intermediate geometries.

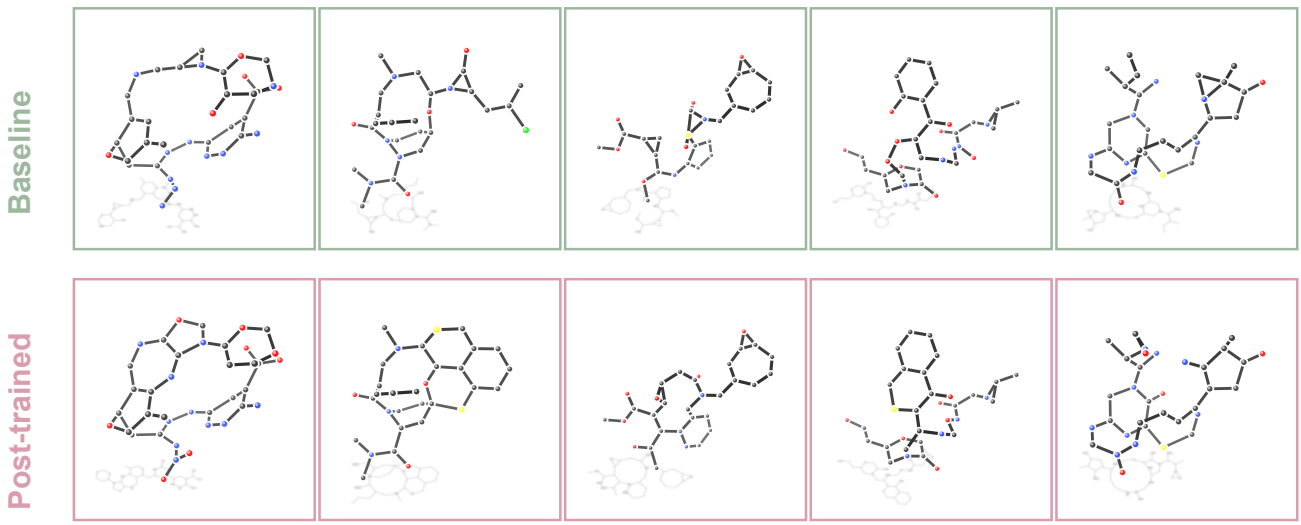

Figure 5. Qualitative comparison of molecular conformers generated by the baseline diffusion model (top row, green) and the post-trained model (bottom row, pink). Each panel displays a generated conformer (dark atoms and bonds) overlaid on the ground-truth reference structure (light gray). The post-trained model produces conformers with improved structural fidelity to the reference geometries.

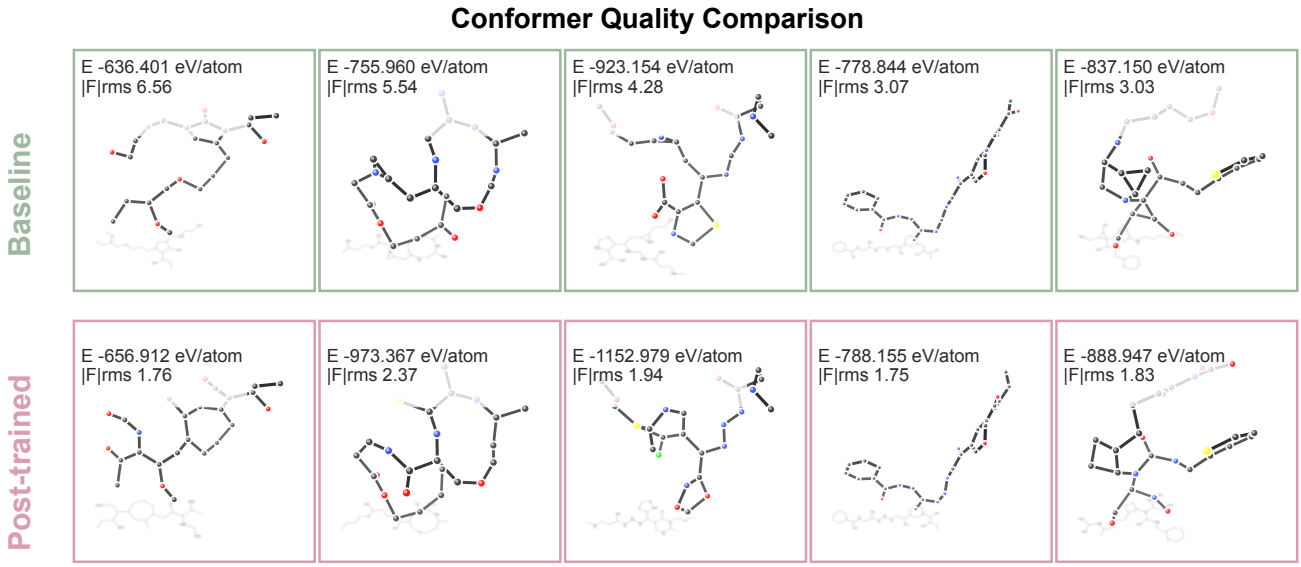

Figure 6. Conformer quality comparison showing energy and force metrics. Top row: baseline model generates structures with higher energies (E) and larger RMS forces, indicating geometries farther from equilibrium. Bottom row: post-trained model consistently achieves lower energies and reduced force magnitudes across all examples, demonstrating improved physical plausibility. Generated conformers (dark) are overlaid on reference structures (light gray).

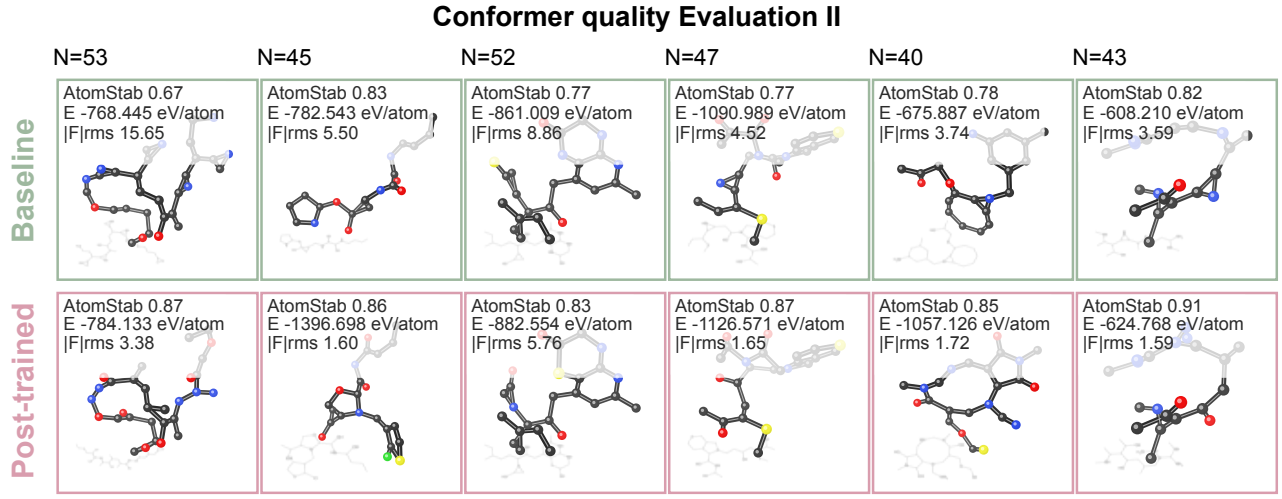

Figure 7. Conformer quality comparison showing energy and force metrics. Top row: baseline model generates structures with higher energies (E) and larger RMS forces, indicating geometries farther from equilibrium. Bottom row: post-trained model consistently achieves lower energies and reduced force magnitudes across all examples, demonstrating improved physical plausibility. Generated conformers (dark) are overlaid on reference structures (light gray).

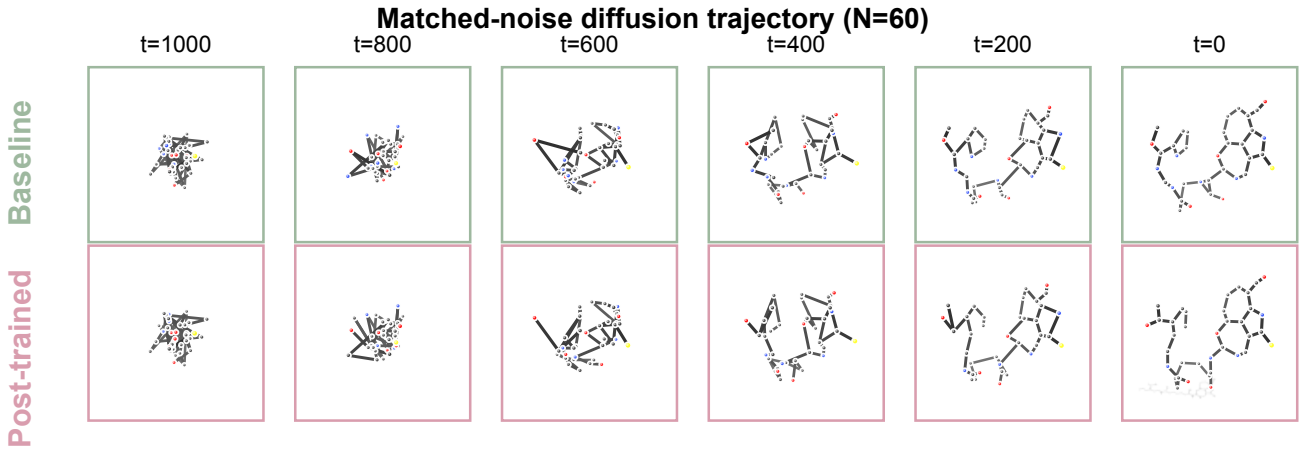

Figure 8. Matched-noise diffusion trajectories comparing the baseline (top row, green) and post-trained (bottom row, pink) models starting from identical initial noise ($N = 60$ atoms, seed=89). Snapshots are shown at diffusion timesteps $t \in \{1000, 800, 600, 400, 200, 0\}$. The baseline model fails to generate a chemically valid molecule, producing a final structure ($t = 0$) with distorted ring geometries and strained bond angles that violate standard valence constraints. In contrast, the post-trained model successfully generates a valid conformer with proper ring planarity and realistic bond geometries, demonstrating improved structural validity through physics-informed fine-tuning.

E.2. Property controllability and distributional fidelity.

We split the QM9 training set into two disjoint halves with 50k samples each. We train the property prediction network ω on the first half, and train conditional generative models on the second half. For a range of target property values $\mathbf{c}$, we draw conditional samples $\boldsymbol{z}_0 \sim \rho_\theta(\cdot \mid \mathbf{c})$ and compute the corresponding predicted properties via $\hat{\mathbf{c}}(\boldsymbol{z}_0) := \omega(\boldsymbol{z}_0)$. We then report the MAE between $\hat{\mathbf{c}}(\boldsymbol{z}_0)$ and the ground-truth QM9 property values.

Property alignment reward. For *conditional generation* with a target property vector $\mathbf{y}$, we add an auxiliary *property alignment reward* at the terminal state:

$$r_0^{(P)}(\boldsymbol{z}_0; \mathbf{y}) := - \|\omega(\boldsymbol{z}_0) - \mathbf{y}\|_1. \tag{8}$$

The total terminal reward combines $r_0^{(P)}$ with the energy/force rewards used by Elign. The property reward is treated as an additional disentangled channel (analogous to energy and force) and is *group-normalized separately* before being weighted and added to the total advantage.

Results. As shown in Table 8, Elign substantially reduces MAE compared to unguided diffusion baselines, narrowing the gap toward the QM9 oracle. Random samples define an upper bound on MAE, while QM9 ground-truth molecules provide a lower bound. Lower MAE indicates that generated samples reside closer to the true data distribution and admit more reliable property prediction. We report results for isotropic polarizability α, HOMO–LUMO gap $\Delta\varepsilon$, and LUMO energy $\varepsilon_{\text{LUMO}}$. Notably, Elign improves all properties simultaneously, indicating that alignment with MLFF-derived energy and force feedback yields samples that are not only structurally stable but also chemically predictive.

Table 8. Mean Absolute Error (MAE) for molecular property prediction. Lower values indicate better controllable generation. Properties are predicted by a pretrained E(3)-equivariant EGNN regressor ω on molecular samples generated by each method. QM9 and Random serve as lower and upper bounds, respectively.

Property Units	α Bohr^3	$\Delta\varepsilon$ meV	$\varepsilon_{\text{LUMO}}$ meV
QM9	0.10	64	36
Random	9.01	1470	1457
N_{atoms}	3.86	866	813
EDM (Hoogeboom et al., 2022)	2.76	655	584
GeoLDM (Xu et al., 2023)	2.37	587	522
GeoBFN (Song et al., 2024)	2.34	<u>577</u>	<u>516</u>
Elign (w/ Additional Property Alignment Reward)	**2.32**	**564**	**512**

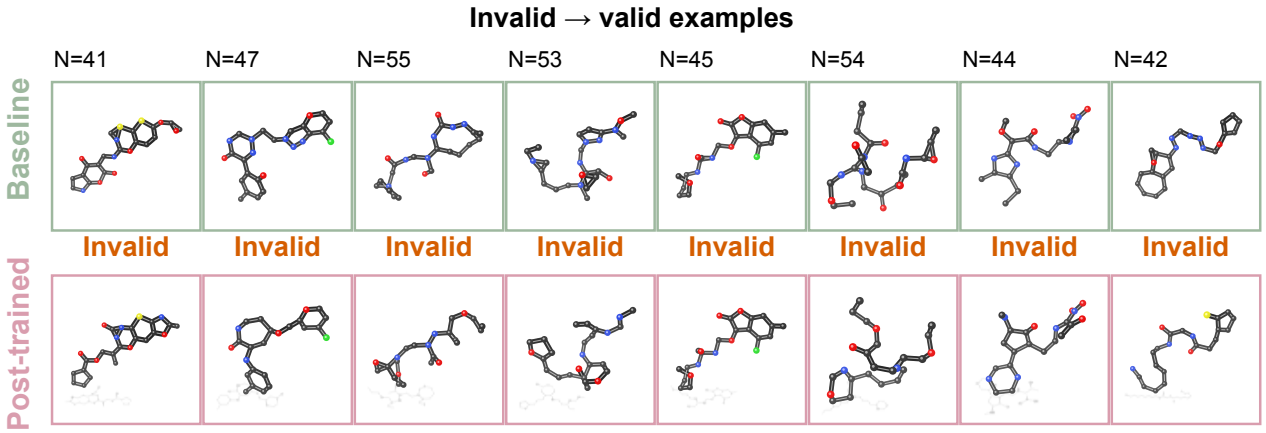

Figure 9. **Qualitative comparison of molecular conformers generated by baseline versus post-trained models.** Each column shows paired conformers for the same molecule (N indicates the number of atoms). The baseline model (top row, green borders) produces chemically invalid structures, while the post-trained model (bottom row, pink borders) generates valid conformers for the same molecular graphs. Atom colors: carbon (gray), nitrogen (blue), oxygen (red), sulfur (yellow), chlorine (green). Hydrogen atoms are shown as small gray spheres.

Table 9. FED-GRPO component ablations.

Setting	Disentangle?	Shared prefix?	PBRS?	A↑	M↑	V×U↑
Full FED-GRPO (Elign)	✓	✓	✓	99.33	93.70	95.31
No disentanglement	×	✓	✓	99.12	92.60	93.80
No shared prefix grouping	✓	×	✓	98.85	83.40	91.20
Naive per-step energy (no PBRS)	✓	✓	×	99.20	92.50	94.60

E.3. More Ablation Studies

FED-GRPO component ablations. Table 9 ablates key ingredients of FED-GRPO. **Disentangled normalization** stabilizes multi-objective optimization by preventing one reward channel (energy, force, or property) from dominating due to scale differences. **Shared-prefix grouping** reduces variance by comparing rollouts that share the same partially denoised prefix, making per-step advantages more meaningful. Finally, **dense PBRS shaping** improves credit assignment over long diffusion horizons, typically improving the stability–diversity trade-off relative to terminal-only rewards.

F. Supplementary Theory: PBRS as an Alchemical Force

While potential-based reward shaping (PBRS) is commonly introduced as a credit-assignment heuristic, in our setting it admits a concrete physical interpretation. We show that PBRS induces an exponential tilting of the one-step reverse diffusion kernel under a KL-regularized local control view, and that in the small-step limit this tilt manifests as an additional drift term proportional to the gradient of the shaping potential. When the potential is chosen as the negative MLFF energy evaluated on the predicted clean geometry, this drift corresponds to a force-like correction acting on the reverse dynamics. Crucially, because the diffusion state includes both coordinates and relaxed feature channels, the induced term decomposes into a physical force on atomic positions and an auxiliary "alchemical" force on non-geometric variables. This result formalizes PBRS as a principled approximation to energy-guided diffusion, clarifying how dense shaping rewards compile physical guidance into the learned reverse policy at training time, rather than injecting oracle forces during inference

Theorem 5 (PBRS induces an approximate (alchemical) force in the reverse drift)**.** *Fix a reverse step t and denote the shaping potential by $\Psi_t(\boldsymbol{z}) := -E_\phi(\hat{\boldsymbol{z}}_{0|t}(\boldsymbol{z}))$ (Eq. (4)). Let the base one-step reverse kernel be Gaussian,*

$$q_t^{\Delta t}(\boldsymbol{z}_{t-\Delta t} \mid \boldsymbol{z}_t) = \mathcal{N}(\boldsymbol{z}_{t-\Delta t};\ \boldsymbol{z}_t + b_t(\boldsymbol{z}_t)\Delta t,\ a_t \Delta t),$$

(e.g. Euler–Maruyama for a reverse SDE; a_t may include $\mathbf{P}_{\mathrm{CoM}}$). Consider the per-step KL-regularized local update with

PBRS reward $r_t^{\text{shape}} = \gamma\Psi_{t-\Delta t}(\boldsymbol{z}_{t-\Delta t}) - \Psi_t(\boldsymbol{z}_t)$ *(Eq. (5)):*

$$\pi_t^{\star,\Delta t}(\cdot \mid \boldsymbol{z}_t) \in \arg\max_{\pi(\cdot|\boldsymbol{z}_t)} \Big\{\mathbb{E}_\pi[r_t^{\text{shape}}] - w_{\text{KL}}\text{KL}(\pi(\cdot \mid \boldsymbol{z}_t)\|q_t^{\Delta t}(\cdot \mid \boldsymbol{z}_t))\Big\}.$$

Assume Ψ_t *is* C^1 *in* $\boldsymbol{z}$ *with locally Lipschitz* $\nabla\Psi_t$*. Then (i) the optimizer is the exponential tilt*

$$\pi_t^{\star,\Delta t}(\boldsymbol{z}_{t-\Delta t} \mid \boldsymbol{z}_t) \propto q_t^{\Delta t}(\boldsymbol{z}_{t-\Delta t} \mid \boldsymbol{z}_t)\, \exp\Big(\frac{\gamma}{w_{\text{KL}}}\Psi_{t-\Delta t}(\boldsymbol{z}_{t-\Delta t})\Big),$$

and (ii) its conditional mean admits the small-step expansion

$$\mathbb{E}_{\pi_t^{\star,\Delta t}}[\boldsymbol{z}_{t-\Delta t} \mid \boldsymbol{z}_t] = \boldsymbol{z}_t + \Big(b_t(\boldsymbol{z}_t) + \frac{\gamma}{w_{\text{KL}}}\, a_t\nabla\Psi_t(\boldsymbol{z}_t)\Big)\Delta t + o(\Delta t).$$

Equivalently, in the $\Delta t \to 0$ *limit the induced controlled reverse drift is*

$$\tilde{b}_t(\boldsymbol{z}) = b_t(\boldsymbol{z}) + \frac{\gamma}{w_{\text{KL}}}\, a_t\nabla\Psi_t(\boldsymbol{z}) = b_t(\boldsymbol{z}) - \frac{\gamma}{w_{\text{KL}}}\, a_t\nabla_{\boldsymbol{z}}E_\phi(\hat{\boldsymbol{z}}_{0|t}(\boldsymbol{z})).$$

Writing $\boldsymbol{z} = [\boldsymbol{x}, \boldsymbol{h}]$*, the* $\boldsymbol{x}$*-block recovers a force-like term* $(-\nabla_{\boldsymbol{x}}E_\phi)$ *while the* $\boldsymbol{h}$*-block yields an "alchemical" preference direction* $(-\nabla_{\boldsymbol{h}}E_\phi)$ *on relaxed types/features.*

Proof. Fix $\boldsymbol{z}_t$. The term $-\Psi_t(\boldsymbol{z}_t)$ is constant in $\boldsymbol{z}_{t-\Delta t}$, so the objective is

$$\max_\pi \Big\{\gamma\,\mathbb{E}_\pi[\Psi_{t-\Delta t}(\boldsymbol{z}_{t-\Delta t})] - w_{\text{KL}}\text{KL}(\pi\|q_t^{\Delta t})\Big\}.$$

A one-line variational calculation (Lagrange multiplier for $\int\pi = 1$) gives the unique maximizer

$$\pi_t^{\star,\Delta t}(u \mid \boldsymbol{z}_t) \propto q_t^{\Delta t}(u \mid \boldsymbol{z}_t)\exp(\eta\Psi_{t-\Delta t}(u)), \qquad \eta := \gamma/w_{\text{KL}}.$$

Now use Gaussian integration by parts (Stein's identity): if $U \sim \mathcal{N}(\mu, \Sigma)$ and g is smooth, $\mathbb{E}[(U-\mu)g(U)] = \Sigma\,\mathbb{E}[\nabla g(U)]$. Apply this with $U \sim q_t^{\Delta t}(\cdot \mid \boldsymbol{z}_t)$ and $g(u) = \exp(\eta\Psi_{t-\Delta t}(u))$ to obtain

$$\mathbb{E}_{\pi_t^{\star,\Delta t}}[U \mid \boldsymbol{z}_t] = \mu + \Sigma\,\eta\,\mathbb{E}_{\pi_t^{\star,\Delta t}}[\nabla\Psi_{t-\Delta t}(U) \mid \boldsymbol{z}_t],$$

where $\mu = \boldsymbol{z}_t + b_t(\boldsymbol{z}_t)\Delta t$ and $\Sigma = a_t\Delta t$. Since $U = \boldsymbol{z}_t + O_{\mathbb{P}}(\sqrt{\Delta t})$ under the Gaussian kernel and $\nabla\Psi$ is locally Lipschitz,

$$\mathbb{E}_{\pi_t^{\star,\Delta t}}[\nabla\Psi_{t-\Delta t}(U) \mid \boldsymbol{z}_t] = \nabla\Psi_t(\boldsymbol{z}_t) + o(1).$$

Substituting and collecting terms yields

$$\mathbb{E}_{\pi_t^{\star,\Delta t}}[\boldsymbol{z}_{t-\Delta t} \mid \boldsymbol{z}_t] = \boldsymbol{z}_t + (b_t(\boldsymbol{z}_t) + \eta\, a_t\nabla\Psi_t(\boldsymbol{z}_t))\Delta t + o(\Delta t),$$

and replacing $\nabla\Psi_t = -\nabla E_\phi(\hat{\boldsymbol{z}}_{0|t})$ gives the stated "force" form. □

From theory to measurement. Theorem 5 provides a local-control view in which energy PBRS induces a force-like correction to the reverse drift. We next empirically probe this effect by measuring the score change induced by post-training and comparing its position component to the MLFF force direction.

F.1. Alchemical force analysis

To understand how post-training compiles physical preferences into the sampler, we measure an *alchemical force* induced by alignment and compare it to the actual force field. Write $\boldsymbol{z} = [\mathbf{x}, \mathbf{h}]$ for positions and atom-type features. Let $s_{\theta_{\text{pre}}}^{(\mathbf{x})}(\boldsymbol{z}_t, t)$ and $s_\theta^{(\mathbf{x})}(\boldsymbol{z}_t, t)$ denote the *position* components of the pretrained and post-trained score networks, respectively. We define the alchemical force proxy (positions only) as

$$\mathbf{F}_{\text{alc}}(\boldsymbol{z}_t, t) := s_\theta^{(\mathbf{x})}(\boldsymbol{z}_t, t) - s_{\theta_{\text{pre}}}^{(\mathbf{x})}(\boldsymbol{z}_t, t), \tag{9}$$

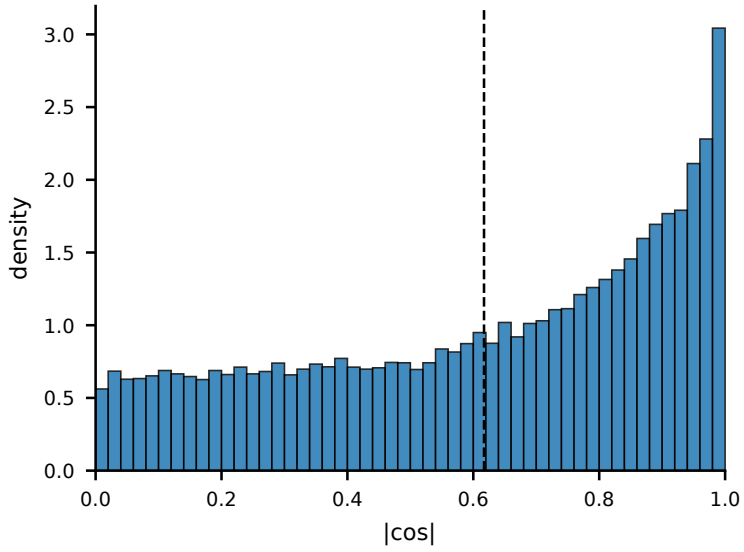

Figure 10. Distribution of $|\cos(\mathbf{F}_{\text{alc}}, \mathbf{F}_\phi)|$ between the position-space alchemical force $\mathbf{F}_{\text{alc}}$ (post-trained minus pretrained score, Eq. 9) and the MLFF force $\mathbf{F}_\phi(\hat{\boldsymbol{z}}_{0|t})$ (Eq. 10). Larger values indicate stronger directional agreement.

so that $\mathbf{F}_{\text{alc}} \in \mathbb{R}^{N\times 3}$ is comparable to a physical force field. Given a force oracle (e.g., the MLFF) $\mathbf{F}_\phi(\hat{\boldsymbol{z}}_{0|t})$, we quantify agreement via cosine similarity

$$\cos(\mathbf{F}_{\text{alc}}, \mathbf{F}_\phi) \;:=\; \frac{\langle \mathbf{F}_{\text{alc}}(\boldsymbol{z}_t, t),\, \mathbf{F}_\phi(\hat{\boldsymbol{z}}_{0|t})\rangle}{\|\mathbf{F}_{\text{alc}}(\boldsymbol{z}_t, t)\|\, \|\mathbf{F}_\phi(\hat{\boldsymbol{z}}_{0|t})\|}. \tag{10}$$

Here $\hat{\boldsymbol{z}}_{0|t}$ is the predicted clean geometry from the diffusion posterior mean. Empirically, we find that $\mathbf{F}_{\text{alc}}$ is strongly aligned with $\mathbf{F}_\phi$. Figure 10 plots a histogram of $|\cos(\mathbf{F}_{\text{alc}}, \mathbf{F}_\phi)|$ computed over sampled diffusion states and molecules; higher values indicate that the post-training update (post-trained minus pretrained score, position component) points in a direction similar to the MLFF force. The distribution concentrates toward large cosine similarity, supporting the interpretation that alignment learns a force-like correction in the reverse dynamics.