Generating Physically Consistent Molecules with Energy-Based Models

Christoph Griesbacher[†]
Graz University of Technology
griesbacher@tugraz.at

Lea Bogensperger
University of Zurich
lea.bogensperger@uzh.ch

Andreas Habring
Graz University of Technology
andreas.habring@tugraz.at

Thomas Pock
Graz University of Technology
thomas.pock@tugraz.at

Abstract

Molecules in equilibrium follow a Boltzmann distribution, making the underlying energy landscape a physically grounded modeling objective. However, such landscapes are difficult to learn from data and, once learned, hard to sample from. Diffusion and flow-matching models sidestep these difficulties by learning a time-conditional score or transport field between noise and data, losing the energy inductive bias in exchange for a more tractable training objective. We introduce EBMol, an energy-based model (EBM) that restores this inductive bias by learning an atom-additive scalar potential without explicit simulation during training. Our method employs a flow-inspired Restoring Field Matching objective to approximate the energy landscape. We adopt the Mirror-Langevin algorithm for sampling, enabling unified updates of atomic positions and types, and incorporate parallel tempering for inference-time compute scaling. EBMol is the first EBM for 3D molecular generation to achieve state-of-the-art performance on QM9 and GEOM-Drugs. Moreover, we show that the learned energy landscape serves as a principled quality metric for ranking and filtering configurations, and demonstrate controllable generation without retraining through shape-steered sampling via potential composition and zero-shot linker design.[*]

1 Introduction

Molecular states at thermal equilibrium are governed by the Boltzmann distribution $p(\mathbf{x}) \propto \exp(-E(\mathbf{x})/(k_B T))$, in which the probability of a configuration is determined entirely by its potential energy. The standard datasets for molecular generation, QM9 [1] and GEOM [2], are not drawn from this distribution. Instead, they contain geometry-optimized structures, each relaxed to a local energy minimum by density functional theory or semi-empirical methods. What is physically meaningful in these datasets is where the energy minima lie. The relative frequencies at which different molecular structures appear reflect enumeration strategies and computational budgets, not thermodynamic weights [3]. A scalar energy landscape whose minima coincide with stable configurations is therefore the natural modeling target: it captures exactly the physical content of the data and takes the functional form of a Boltzmann distribution, $p_\theta(\mathbf{x}) \propto \exp(-E_\theta(\mathbf{x})/\tau)$.

Learning and sampling from such landscapes, however, have remained open problems. Training a physically-calibrated energy function requires Boltzmann-distributed data [4] such as SPICE [5]

[†]Corresponding author.
[*]Code available at https://github.com/griesbchr/EBMol.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

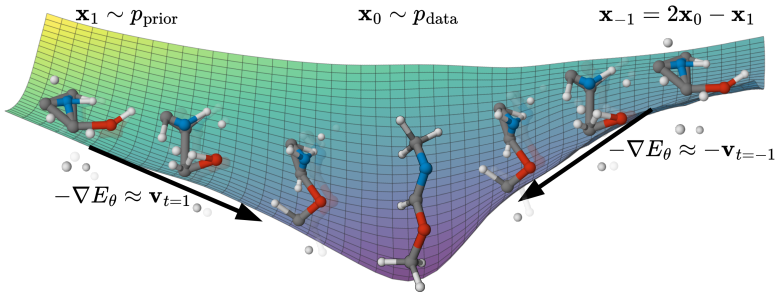


Figure 1: Overview of Restoring Field Matching. The RFM loss shapes a scalar energy landscape with data points $x_0 \sim p_{\text{data}}$ as local minima by training the negative gradient $-\nabla E_\theta$ to point toward data from both sides: from the prior (interpolation, $t \in [0, 1]$) and from beyond the data (extrapolation, $t \in [-1, 0]$).

or MD17 [6], but datasets with this property target interatomic interactions and conformational sampling rather than chemical diversity and are unsuitable for generation. Sampling is equally difficult: generating molecules from an energy function typically relies on Markov Chain Monte Carlo (MCMC) methods, which must traverse the extremely rugged, unphysical regions between noise and valid structures [7]. Diffusion [8, 9] and flow matching models [10, 11] sidestep both barriers by directly learning noise-to-data transport maps through time-conditional vector fields, achieving state-of-the-art results. However, the resulting models provide no explicit energy function and no mechanism to evaluate, rank, or compose configurations at inference time.

In this work, we introduce *EBMol*, an Energy-Based Model (EBM) for Molecules that addresses both barriers. For training, we propose Restoring Field Matching (RFM, Figure 1), a flow-matching-inspired objective that shapes the energy landscape with data points as local minima, without requiring simulation or Boltzmann-distributed data. For sampling, we adapt the Mirror Langevin Algorithm (MLA) [12] to molecular state spaces, applying Euclidean updates on coordinates and simplex-constrained updates on atom types jointly within each step. Given that the learned energy is an explicit, persistent scalar function rather than an implicit mapping from noise to data, capabilities that time-conditional models cannot naturally provide follow directly: energy-based ranking and filtering of generated samples, composable generation through potential composition, and zero-shot conditional generation by fixing molecular fragments.

Our contributions are as follows:

- We introduce Restoring Field Matching (RFM), a simulation-free training objective for tractable learning of per-atom additive energy landscapes, and establish theoretical guarantees for the resulting optimization problem.
- We adapt the Mirror Langevin Algorithm to the mixed continuous-simplex state space of molecules, unifying coordinate and atom types updates in a single framework, combined with parallel tempering enabling inference-time compute scaling.
- EBMol is the first EBM for 3D molecular generation to achieve state-of-the-art performance on QM9 and GEOM-Drugs. This demonstrates that the benefits of energy-based modeling, including physically meaningful scoring, inference-time ranking, composable shape steering, and zero-shot linker design, are attainable without sacrificing generation quality.

2 Related work

Energy-based modeling and related time-unconditional frameworks. Energy-based models [13] have traditionally been trained via MCMC-based maximum likelihood [14] or score matching [15], both of which sidestep the partition function but at substantial cost: MCMC is expensive and prone to mode collapse [16], while score matching either requires Hessian traces or re-introduces a time variable through noise conditioning [15]. Despite the recent popularity of time-conditional diffusion [8, 9] and flow matching methods [10, 11], several recent works have revisited time-

unconditional EBMs. Energy Matching [17] learns an energy landscape in two stages, combining transport and contrastive-divergence. Equilibrium Matching [18] is conceptually closest to RFM, also enforcing stationarity at data by scaling the target field to zero, but learns an implicit non-conservative vector field rather than an explicit scalar energy. Also, neither method addresses mixed discrete-continuous state spaces. Within the molecular domain specifically, time-unconditional EBM training has previously been attempted only on highly restricted settings by [19], who train an equivariant EBM via simulation-based contrastive divergence on a single-stoichiometry subset of QM9, thus neglecting variable molecule size, mixed discrete-continuous state, or modern benchmarks.

Generative modeling of 3D molecules. Various different generative modeling paradigms have been explored for 3D molecular generation. Besides autoregressive models [20–22], field-based approaches [23, 24], and, most prominently, diffusion and flow matching models. These methods differ primarily in their treatment of bonds: bond-explicit methods use dedicated supervision [25–30], while bond-implicit approaches, which include our work, rely on emergent geometric accuracy to recover connectivity post-hoc via OpenBabel [31, 26] or predefined lookup tables [32]. The seminal EDM [32] introduced equivariant diffusion via the EGNN [33] backbone to molecular modeling. Subsequent work refined this line through latent-space variants [34], alternative architectures and training [35], flow matching [36–38], and Bayesian flow networks [39]. Several works that incorporate energy-inspired components into their generative pipeline are related to our approach. These include using Amber and MACE force fields [40, 7], interpreting self-consistency residuals as energies in iteratively refined flow models [41], or employing energy-driven objectives for conformer generation [42]. Crucially, however, none of these methods learn a persistent, scalar energy landscape that drives end-to-end sampling.

Mirror Langevin algorithm and constrained sampling. The Mirror Langevin Algorithm [12] extends mirror descent [43, 44], originally developed for constrained optimization, to sampling on constrained domains. Theoretical analyses on convex domains have established convergence rates [45, 46] and Metropolis-adjusted extensions [47]. While MLA has recently been adapted for generative modeling on homogeneous constrained domains [48], EBMol is, to the best of our knowledge, the first unified mixed-geometry Langevin framework for joint continuous-discrete deep generative modeling.

3 Method

We propose *EBMol*, a time-unconditional energy-based model that learns a scalar potential $E_\theta(\mathbf{x})$ over molecular configurations. EBMol combines two largely independent components: Restoring Field Matching (RFM) for tractable shaping of the energy landscape, and the Mirror Langevin Algorithm (MLA) [12] for sampling from it.

3.1 Problem formulation

A molecule with N atoms is represented as a joint state $\mathbf{x} = (\mathbf{c}, \mathbf{p})$, where $\mathbf{c} \in \mathbb{R}^{N\times 3}$ are atomic coordinates in Euclidean space and $\mathbf{p} = (\mathbf{p}_0, \ldots, \mathbf{p}_{N-1})$ are per-atom categorical distributions with $\mathbf{p}_i \in \Delta^{K-1} = \{\mathbf{p}_i \in \mathbb{R}^K \mid p_{i,k} \geq 0, \sum_k p_{i,k} = 1\}$ over K chemical elements. Bonds are not explicitly modeled; instead, connectivity and bond orders are inferred post-hoc from $(\mathbf{c}, \mathbf{p})$, placing the burden of chemical validity primarily on geometric consistency. We parameterize a time-unconditional energy $E_\theta(\mathbf{c}, \mathbf{p})$ over this mixed state space. To handle variable-sized molecules, we impose a permutation-invariant additive decomposition over atoms $E_\theta(\mathbf{c}, \mathbf{p}) = \sum_{i=1}^{N} E_\theta^{(i)}(\mathbf{c}, \mathbf{p})$.

3.2 Learning with the Restoring Field Matching objective

Learning a meaningful energy landscape is challenging. Maximum-likelihood training relies on MCMC [14], and score matching either requires Hessian traces [49] or introduces a time variable [8, 9]. We instead construct the landscape *geometrically*: using flow-matching-style interpolation, we regress the energy gradient toward a field that points back to the data from both sides of the manifold, making data points local minima.

Constructing the restoring field. Standard flow matching defines an interpolant $\mathbf{x}_t$ between $\mathbf{x}_1 \sim p_{\text{prior}}$ at $t = 1$ and $\mathbf{x}_0 \sim p_{\text{data}}$ at $t = 0$, with a target velocity that is nonzero everywhere on the interval. Data is thus a trajectory endpoint, not a stationary point of any field. We begin by extending the interpolation to $t \in [-1, 1]$, so the interpolant $\mathbf{x}_t$ passes *through* the data rather than terminating at it. We construct this extended interpolant separately for the two components of $\mathbf{x} = (\mathbf{c}, \mathbf{p})$. For continuous coordinates, we first align $\mathbf{c}_1$ to $\mathbf{c}_0$ using the equivariant optimal-transport coupling to reduce trajectory curvature. The aligned noise is denoted $\mathbf{c}_1$, and we define a linear trajectory $\mathbf{c}_t = \mathbf{c}_0 + t(\mathbf{c}_1 - \mathbf{c}_0), \quad t \in [-1, 1]$. For categorical features, linear extrapolation leaves the simplex. We instead use a reflection, $\mathbf{p}_t = \mathbf{p}_0 + |t|(\mathbf{p}_1 - \mathbf{p}_0)$, which remains in Δ^{K-1} and is symmetric around $t = 0$.

Along each interpolant, we define a target field that points toward the data $\mathbf{x}_0$. Building on the natural choice $\text{sign}(t)(\mathbf{x}_0 - \mathbf{x}_1)$, we directly encode stationarity at the data by scaling the target with a smoothing function $b : [0, 1] \to [0, 1]$ satisfying $b(0) = 0$ and $b(1) = 1$, so that the target vanishes at $t = 0$:

$$\mathbf{u}^{\mathbf{c}} = b(|t|) \cdot \text{sign}(t) \cdot (\mathbf{c}_0 - \mathbf{c}_1), \qquad \mathbf{u}^{\mathbf{p}} = b(|t|) \cdot (\mathbf{p}_0 - \mathbf{p}_1). \tag{1}$$

The model is trained to match $-\nabla_{\mathbf{x}} E_\theta(\mathbf{x}_t)$ against this target, as described below.

Training objective. Denoting $\mathbf{v} := -\nabla_{\mathbf{x}} E_\theta(\mathbf{x}_t)$ and $\mathbf{u} := (\mathbf{u}^{\mathbf{c}}, \mathbf{u}^{\mathbf{p}})$, the RFM objective is given by

$$\mathcal{L}_{\text{RFM}} = \mathbb{E}_{\mathbf{x}_0, \mathbf{x}_1, t \sim U[-1,1]} \left[\left\| \mathbf{v} - \mathbf{u} \right\|^2 \right]. \tag{2}$$

While the restoring target specifies the energy *gradient*, it leaves the energy *value* at the data unconstrained. Without regularization, different data points may sit at different absolute energies, biasing the sampler toward artificially deeper basins. We therefore add a per-atom regularization term that anchors clean-data energies toward zero:

$$\mathcal{L}_{\text{reg}} = \mathbb{E}_{\mathbf{x}_0 \sim p_{\text{data}}} \left[\frac{1}{N} \sum_{i=1}^{N} E_\theta^{(i)}(\mathbf{x}_0)^2 \right]. \tag{3}$$

The full training loss is then given by

$$\mathcal{L} = \mathcal{L}_{\text{RFM}} + \lambda_{\text{reg}} \, \mathcal{L}_{\text{reg}}, \tag{4}$$

where λ_{reg} controls the regularization strength. Training is summarized in Algorithm 1.

3.3 Analysis

We establish two theoretical properties of the RFM objective that characterize the minimization problem defined by (4). The analysis considers E as a general element of a function space rather than a specific network architecture. For sufficiently expressive networks, these properties are expected to hold approximately in practice, which our experiments in Section 4 support. Formal assumptions and complete proofs for both properties are given in Appendix A.

The first result ensures that the training problem is well-posed.

Proposition 1 (Existence of solutions, informal)**.** *Under appropriate assumptions on p_{prior} and p_{data}, for every λ_{reg} the following problem admits a unique solution:*

$$\min_E \mathcal{L}(E).$$

Given existence, we characterize the solution's behavior at the two extremes of the regularization parameter.

Proposition 2 (Convergence for small and large regularization parameters, informal)**.** *Let $E^*(\mathbf{x}, \lambda)$ be the minimizer of $\mathcal{L}$ for $\lambda_{reg} = \lambda$. The following holds true:*

- *(i) $\underline{\lambda = \infty}$: For every sequence $\lambda_n \to \infty$, there exists a subsequence $(\lambda_{n_k})_k$ such that $E^*(\cdot, \lambda_{n_k}) \to 0$ p_0-a.e.; figuratively: $E^*(\mathbf{x}_0, \lambda) \approx 0$ for $\lambda \gg 0$ and $\mathbf{x}_0 \sim p_{data}$.*
- *(ii) $\underline{\lambda = 0}$: For every sequence $\lambda_n \to 0$, there exists a subsequence $(\lambda_{n_k})_k$ such that $\nabla E^*(\cdot, \lambda_{n_k})$ converges to the projection of the prescribed restoring field $\mathbf{u}$ onto the space*

of all conservative vector fields. That is, $\nabla E^*(\cdot, \lambda_{n_k})$ converges to the minimizer of the functional

$$E \mapsto \mathbb{E}_{\mathbf{x}_0, \mathbf{x}_1, t\sim U[-1,1]}[\left\|-\nabla E(\mathbf{x}_t) - \mathbf{u}\right\|^2] \tag{5}$$

Moreover, if $\mathbf{x} \mapsto \mathbb{E}[\mathbf{u}|x_t = x]$ is conservative

$$\lim_{k\to\infty} \nabla E^*(\mathbf{x}, \lambda_{n_k}) = -\mathbb{E}[\mathbf{u}|x_t = x] \quad p_x\text{-a.e.}. \tag{6}$$

The above results confirm our modeling approach theoretically: strong regularization anchors the energy to zero at the data, while the gradient matching term leads to the L2 projection of the target restoring field onto the space of conservative vector fields. At any finite λ_{reg}, the minimizer interpolates between these two extremes. However, the central modeling claim, that data points are local minima of E_θ, is not established by the propositions above. It rests on the target construction ($b(0) = 0$ encodes zero gradient at $t = 0$) and is verified empirically in Appendix B.

3.4 Sampling with annealed Mirror-Langevin algorithm

We sample from $p_\theta(\mathbf{x}) \propto \exp(-E_\theta(\mathbf{x})/\tau)$ via Langevin dynamics. While standard Langevin suffices for coordinates, it pushes atom types outside the simplex. The Mirror Langevin Algorithm (MLA) [12] resolves this by performing updates in an unconstrained dual space and projecting back through a mirror map.

Let $\Phi : \mathcal{D} \to \mathbb{R}$ be a strictly convex, twice-differentiable mirror potential on the constrained domain $\mathcal{D}$ such that its gradient $\nabla\Phi$ maps $\mathcal{D}$ to an unconstrained dual space, and the inverse map $(\nabla\Phi)^{-1}$ projects back to $\mathcal{D}$. The MLA update with step size η, temperature τ, and noise $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ reads

$$\mathbf{y}^{(k+1)} = \nabla\Phi(\mathbf{x}^{(k)}) - \eta\, \nabla_{\mathbf{x}} E_\theta(\mathbf{x}^{(k)}) + \sqrt{2\eta\tau}\, \left(\nabla^2\Phi(\mathbf{x}^{(k)})\right)^{1/2} \boldsymbol{\xi}, \tag{7}$$

$$\mathbf{x}^{(k+1)} = (\nabla\Phi)^{-1}(\mathbf{y}^{(k+1)}). \tag{8}$$

The factor $(\nabla^2\Phi)^{1/2}$ is the Riemannian metric correction induced by the mirror potential, ensuring the update is consistent with the underlying mirror Langevin SDE.

Continuous coordinates and categorical features. The choice of mirror potential Φ specializes MLA to each component of $\mathbf{x} = (\mathbf{c}, \mathbf{p})$. For $\mathbf{c} \in \mathbb{R}^{N\times 3}$, the Euclidean potential $\Phi^{\mathbf{c}}(\mathbf{c}) = \frac{1}{2}\|\mathbf{c}\|^2$ reduces MLA to the Unadjusted Langevin Algorithm. For atom types $\mathbf{p} \in \Delta^{K-1}$, the negative entropy $\Phi^{\mathbf{p}}(\mathbf{p}) = \sum_i p_i \log p_i$ maps the simplex to logit space via $\nabla\Phi^{\mathbf{p}} = \log \mathbf{p} + \mathbf{1}$, with $(\nabla\Phi^{\mathbf{p}})^{-1} = \mathrm{softmax}$ projecting back exactly. Since the negative-entropy mirror map does not satisfy the self-concordance condition of [12] near the simplex boundary, we clamp $\mathbf{p} \geq \epsilon$ before each update (see Appendix H for details). The full update is given in Algorithm 2.

Algorithm 1 RFM training step

Require: Energy E_θ, smoothing function $b(t)$
1: Sample $\mathbf{x}_0 = (\mathbf{c}_0, \mathbf{p}_0) \sim p_{\text{data}}$
2: Sample $\mathbf{x}_1 = (\mathbf{c}_1, \mathbf{p}_1) \sim p_{\text{prior}}$
3: Sample $t \sim \mathcal{U}(-1, 1)$
4: $T_{\text{OT}} \leftarrow \mathrm{OTSolver}(\mathbf{c}_1, \mathbf{c}_0)$
5: $\mathbf{c}_t \leftarrow \mathbf{c}_0 + t\,(T_{\text{OT}}\,\mathbf{c}_1 - \mathbf{c}_0)$
6: $\mathbf{p}_t \leftarrow \mathbf{p}_0 + |t|\,(\mathbf{p}_1 - \mathbf{p}_0)$
7: $\mathbf{v} \leftarrow -\nabla_{\mathbf{x}} E_\theta(\mathbf{c}_t, \mathbf{p}_t)$
8: $\mathbf{u}_{\mathbf{c}} \leftarrow b(|t|)\,\mathrm{sign}(t)\,(\mathbf{c}_0 - \mathbf{c}_1)$
9: $\mathbf{u}_{\mathbf{p}} \leftarrow b(|t|)\,(\mathbf{p}_0 - \mathbf{p}_1)$
10: $\mathcal{L}_{\text{reg}} \leftarrow \frac{\lambda_{\text{reg}}}{N} \sum_{i=1}^{N} E_\theta^{(i)}(\mathbf{x}_0)^2$
11: $\mathcal{L} \leftarrow \|\mathbf{v} - \mathbf{u}\|^2 + \mathcal{L}_{\text{reg}}$
12: **Update** $\theta \leftarrow \theta - \eta\, \nabla_\theta \mathcal{L}$

Algorithm 2 Mirror-Langevin sampling step

Require: State $(\mathbf{c}, \mathbf{p})$, E_θ, step size η, temperature τ, noise scales $\sigma_{\mathbf{c}}, \sigma_{\mathbf{p}}$, clipping ϵ
1: $\mathbf{g}_{\mathbf{c}} \leftarrow \nabla_{\mathbf{c}} E_\theta(\mathbf{c}, \mathbf{p}), \quad \mathbf{g}_{\mathbf{p}} \leftarrow \nabla_{\mathbf{p}} E_\theta(\mathbf{c}, \mathbf{p})$
Continuous update:
2: $\boldsymbol{\xi}_{\mathbf{c}} \sim \mathcal{N}(\mathbf{0}, \sigma_{\mathbf{c}}^2 \mathbf{I})$
3: $\mathbf{c}' \leftarrow \mathbf{c} - \eta\, \mathbf{g}_{\mathbf{c}} + \sqrt{2\eta\tau}\, \boldsymbol{\xi}_{\mathbf{c}}$
4: $\mathbf{c}' \leftarrow \mathbf{c}' - \frac{1}{N}\sum_i \mathbf{c}'_i$
Categorical update:
5: $\tilde{\mathbf{p}} \leftarrow \max(\mathbf{p}, \epsilon)$
6: $\boldsymbol{\xi}_{\mathbf{p}} \sim \mathcal{N}(\mathbf{0}, \sigma_{\mathbf{p}}^2 \mathbf{I})$
7: $\mathbf{y} \leftarrow \log \tilde{\mathbf{p}} - \eta\, \mathbf{g}_{\mathbf{p}} + \sqrt{2\eta\tau}\, \tilde{\mathbf{p}}^{-1/2} \odot \boldsymbol{\xi}_{\mathbf{p}}$
8: $\mathbf{p}' \leftarrow \mathrm{Softmax}(\mathbf{y})$
9: **return** $(\mathbf{c}', \mathbf{p}')$

3.5 Parallel tempering

We run M parallel chains at a fixed ladder of temperatures $\tau_1 > \tau_2 > \cdots > \tau_M$, with periodic Metropolis-Hastings swaps between adjacent levels according to energies E with acceptance probability

$$\min\Big(1,\, \exp\Big[\big(E(\mathbf{x}_j) - E(\mathbf{x}_i)\big)\big(\tfrac{1}{\tau_i} - \tfrac{1}{\tau_j}\big)\Big]\Big)\,. \tag{9}$$

Hot chains explore broadly, cold chains refine, and swaps propagate discoveries down the ladder. Periodically, we harvest samples from the coldest chain and replace them with prior samples at the same coldest level, which are swapped upward by the energy-based exchange mechanism. Subsequently, harvested samples undergo a short $\tau = 0$ relaxation. Parallel tempering offers a built-in mechanism for inference-time compute scaling by increasing the number of tempering rounds between harvests. Additional swaps and exploration before samples are extracted yields progressively higher-quality molecules at the cost of longer sampling.

4 Experiments

4.1 Setup

Datasets. We evaluate EBMol on the two standard benchmarks for unconditional 3D molecular generation, QM9 [1] and GEOM-Drugs [2]. QM9 contains 134k small organic molecules with up to nine heavy atoms across five element types (H, C, N, O, F), exhaustively enumerating all stable molecules within this space at DFT level. GEOM-Drugs poses a more challenging benchmark, comprising roughly 430k drug-like molecules with an average size of 44 atoms (up to 181) across 15 element types. This dataset is substantially more diverse than QM9.

Implementation. To ensure SE(3) invariance of the energy, we employ a standard EGNN [33] network, propagating equivariant messages across a fully connected atomic graph. We compute the per-atom energies by applying a two-layer MLP to the invariant node features of the final layer, omitting the final equivariant feature update. The total molecule energy is calculated by summing over the node energies, preserving SE(3) invariance of E_θ. During training, we use a diagonal anisotropic Gaussian position prior whose per-molecule covariance is set to the principal-component eigenvalues of a training molecule with the same atom count, introducing a mild data-dependent shape bias that improves training stability and performance. For atom types, we use a Dirichlet distribution with concentration $1/K$, where K is the number of atom types. We employ $\tanh(\cdot)$ as smoothing function. Pseudocode for the training and sampling algorithms is presented in Algorithms 1 and 2. During generation, we sample the atom count N from the empirical training distribution and base the parallel tempering swaps on the largest per-atom energy $\max_i E_\theta^{(i)}(\mathbf{x})$. A full list of hyperparameters, more experiment details and ablations on design choices can be found in Appendices G and H.

Baselines and evaluation. We compare against a variety of established bond-implicit baselines for molecular generation [32, 40, 36, 50, 34, 38, 39, 35]. All baselines share the same EGNN network architecture [33], with the exception of SLDM, employing a modified EGNN variant based on UniGEM [51] with more parameters (2.2M vs 4.2M for GEOM-Drugs). On GEOM-Drugs, we restrict baselines to methods with publicly available checkpoints to ensure consistent re-evaluation under our revised evaluation protocol.

On QM9 we follow the standard protocol of [32], reporting atom and molecule stability, validity, uniqueness, and novelty. On GEOM-Drugs, recent models have begun to saturate this protocol: its QM9-calibrated bond and valency lookup tables yield a data baseline of only 86.5% atom stability, a ceiling already exceeded by SLDM [35]. To restore this benchmark's discriminative power, we adopt a revised protocol. We perform bond inference via OpenBabel [31], which additionally infers formal charges and aromaticity. This additional information unlocks the usage of the valency tables proposed by [52], which are specifically calibrated for the element and bond-type diversity encountered in GEOM-Drugs. Furthermore, we use the OpenBabel-inferred molecules for RDKit sanitization with stereochemistry assignment for more expressive validity, uniqueness, and novelty metrics. Under this revised protocol, the data baseline rises to 99.6% atom stability (from 86.5%) and 88.2% molecule stability (previously ~0%), indicating a much more faithful assessment of chemical validity. We additionally report the Vendi score [53] for molecular diversity and the physical metric suite of [52],

which assesses differences in bond lengths, bond angles and torsion angles between generated molecules and their GFN2-xTB-optimized [54] counterparts, alongside the associated mean and median relaxation energies.

4.2 Unconditional generation

Table 1: Standard molecular generation metrics on QM9, computed on 10k generated samples per model. Results marked with * are from our own experiments. Unless otherwise noted in the subscript, models use 1000 NFEs.

Method	Atom ↑ Stability	Molecule ↑ Stability	Validity ↑	Uniqueness ↑	Valid & Unique ↑	Novelty ↑
Data	99.40	95.30	97.70	99.97	97.60	–
EDM	98.70	82.00	91.90	–	90.70	65.70
EDM-Bridge	98.80	84.60	92.00	–	90.70	–
EquiFM_{210}	98.90	88.30	94.70	–	93.50	–
END	98.90	89.10	94.80	–	92.60	–
GEOLDM	98.90	89.40	93.80	–	92.70	57.00*
GOAT	99.20	–	92.90	99.00	–	**78.60**
GeoBFN_{2000}	99.31	93.32	96.88	–	**92.41**	65.30
SLDM	99.43	95.42	97.07	–	90.42	–
SLDM*	99.07	92.32	95.53	95.28	91.02	42.46
EBMol_{930}	99.35	92.34	96.88	86.61	83.91	45.68
EBMol_{1370}	99.65	95.58	98.32	82.78	81.39	43.37
EBMol_{1810}	**99.76**	**97.09**	**98.93**	82.30	81.42	42.08

The results for unconditional generation on QM9 are summarized in Table 1. We benchmark EBMol for three different compute-settings, where the lowest setting of 930 network function evaluations (NFEs) matches the baseline models. The results show that EBMol_{930} is competitive with the currently best bond-implicit model SLDM, while using approximately half the parameters.

When scaling compute to 1810 NFEs, EBMol exceeds all baselines in stability and validity. This includes the strong baseline GeoBFN that also achieves high quality samples through inference-time compute adaptation. Uniqueness and novelty are lower, at 82% and 42%, respectively. This is expected on QM9 because the dataset exhaustively enumerates its composition space [3], so higher-quality molecules are more likely to already appear in the training set and novel compositions tend to violate stability constraints [32, 40]. The quality–diversity tradeoff observed in Tables 1 and 2 is a direct consequence of the parallel tempering sampler concentrating probability mass on the lowest-energy modes. While this ensures high chemical validity, it naturally limits diversity.

Shifting towards the more challenging GEOM-Drugs benchmark, we report the revised metrics in Table 2. At a compute-matched setting, EBMol already surpasses the SLDM baseline, a performance gap that widens substantially as compute scales. While we observe a slight decrease in uniqueness as stability improves, this shift is consistent with the expected quality–diversity tradeoff. The model still produces 96.4% unique samples among 10k generations, suggesting a slow decay in diversity rather than a catastrophic mode collapse. Randomly selected samples are visualized in Appendix I and a failure mode analysis can be found in Appendix D.

The physical metrics presented in Table 3 clearly demonstrate the impact of the energy inductive bias. In the medium-compute configuration, EBMol_{1960} already outperforms bond-implicit baselines and is the first such model to surpass the MMFF [55] force field. When compute is scaled, EBMol_{7240} achieves a median energy decrease of 1.78 kcal/mol, an order of magnitude lower than previous bond-implicit models. Notably, EBMol remains competitive with the bond-explicit Megalodon (3.19 kcal/mol, 40.6M parameters), despite having 18× fewer parameters and no explicit bond supervision. This trend extends to bond length and angle distributions, where EBMol_{7240} matches or exceeds Megalodon's performance.

Table 2: Revised stability, validity, uniqueness, novelty, and diversity metrics on GEOM-Drugs, computed on 10k generated samples per model. Compute budgets (NFEs) are indicated in each method's subscript.

Method	Atom ↑ Stability	Molecule ↑ Stability	Validity ↑	Uniqueness ↑	Novelty ↑	Diversity ↑
Data	99.60	88.62	97.91	97.89	–	901.7
Noise	37.27	00.00	71.89	100.00	100.00	1873.8
EDM_{1000}	87.07	1.03	84.43	**99.99**	99.94	**1621.6**
GEOLDM_{1000}	91.61	6.08	86.93	99.96	99.88	1582.5
SLDM_{1000}	95.27	20.59	86.98	99.95	99.93	1474.0
EBMol_{1080}	95.86	24.68	95.36	99.67	99.69	1096.8
EBMol_{1960}	97.17	37.40	97.02	99.39	99.50	927.0
EBMol_{3720}	98.15	52.31	98.61	98.47	99.42	783.0
EBMol_{7240}	**98.68**	**61.42**	**99.03**	96.42	99.34	645.0

Table 3: Structural and energy metrics on GEOM-Drugs, computed on valid and connected molecules only. Bond lengths (Å), angles (degrees), energies (kcal/mol). The best bond-explicit models from GEOM-Drugs Revisited [52] are included for reference.

Method	Bond Lengths ↓ ($\times 10^{-2}$)	Bond Angles ↓	Torsions ↓	Avg Energy ↓ Diff.	Med Energy ↓ Diff.
Data	$0.00^{\pm 0.00}$	$0.00^{\pm 0.00}$	$0.00^{\pm 0.01}$	$0.00^{\pm 0.01}$	$0.00^{\pm 0.00}$
MMFF → GFN2-xTB	$1.12^{\pm 0.01}$	$1.22^{\pm 0.00}$	$4.89^{\pm 0.10}$	$11.40^{\pm 0.20}$	$9.84^{\pm 0.06}$
FlowMol2_{100} (4.3M)	$1.30^{\pm 0.04}$	$1.62^{\pm 0.02}$	$15.00^{\pm 0.30}$	$24.30^{\pm 0.80}$	$17.90^{\pm 0.50}$
Megalodon_{500} (40.6M)	$0.66^{\pm 0.02}$	$0.71^{\pm 0.01}$	$5.58^{\pm 0.11}$	$5.76^{\pm 0.27}$	$3.19^{\pm 0.12}$
EDM_{1000} (2.2M)	$5.26^{\pm 0.03}$	$5.73^{\pm 0.05}$	$30.82^{\pm 0.18}$	$175.34^{\pm 2.44}$	$155.58^{\pm 1.87}$
GEOLDM_{1000} (2.2M)	$3.20^{\pm 0.11}$	$3.90^{\pm 0.06}$	$24.51^{\pm 0.19}$	$99.25^{\pm 4.49}$	$80.13^{\pm 2.13}$
SLDM_{1000} (4.2M)	$1.66^{\pm 0.03}$	$1.34^{\pm 0.02}$	$10.07^{\pm 0.08}$	$23.77^{\pm 0.87}$	$17.90^{\pm 0.30}$
EBMol_{1080} (2.2M)	$1.77^{\pm 0.02}$	$1.67^{\pm 0.03}$	$8.46^{\pm 0.13}$	$40.77^{\pm 9.16}$	$6.67^{\pm 0.30}$
EBMol_{1960} (2.2M)	$1.22^{\pm 0.04}$	$1.19^{\pm 0.04}$	$6.53^{\pm 0.19}$	$26.13^{\pm 7.87}$	$4.12^{\pm 0.07}$
EBMol_{3720} (2.2M)	$0.83^{\pm 0.01}$	$0.81^{\pm 0.01}$	$4.56^{\pm 0.07}$	$10.36^{\pm 1.55}$	$2.35^{\pm 0.05}$
EBMol_{7240} (2.2M)	$\mathbf{0.65^{\pm 0.01}}$	$\mathbf{0.62^{\pm 0.01}}$	$\mathbf{3.65^{\pm 0.06}}$	$\mathbf{5.27^{\pm 0.39}}$	$\mathbf{1.78^{\pm 0.04}}$

4.3 Composable and conditional generation

Shape-steered generation. We compose the learned potential E_θ with a differentiable potential U representing molecular shapes. This potential is defined via PCA eigenvalues of atom positions and targets either linear ($U_{\text{lin}} = \lambda_2 + \lambda_3$), planar ($U_{\text{disk}} = \lambda_3 + \lambda_1 - \lambda_2$) or spherical ($U_{\text{sphere}} = \lambda_1 - \lambda_3$) shapes during generation. Here $\lambda_1 \geq \lambda_2 \geq \lambda_3$ are the PCA eigenvalues of the covariance matrix of $\mathbf{c}$, measuring the stretch along each principal axis. For example, minimizing $U_{\text{lin}} = \lambda_2 + \lambda_3$ suppresses the two minor axes, thus collapsing the molecule onto a single axis. Steered molecules successfully

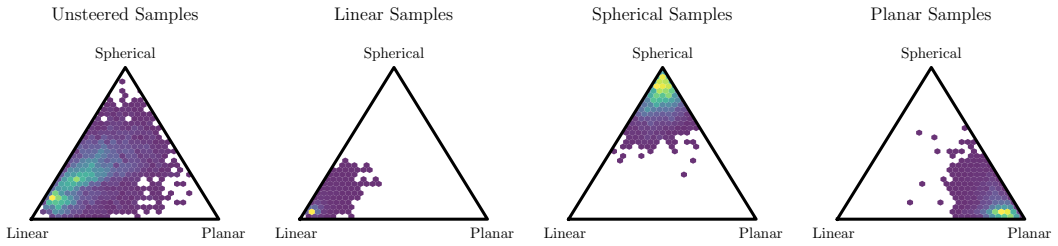


Figure 2: Shape distribution of 5000 generated molecules. Left: unsteered generation. Right: generation with composable shape potentials targeting linear, planar, and spherical geometries. Steering successfully redirects sampling toward underrepresented shapes.

Table 4: Zero-shot conditional generation results. **Left:** Linker design across six problems of varying complexity, each evaluated over 100 attempts. Success requires a single connected component passing RDKit sanitization. **Right:** Shape-steered generation across three target shape classes alongside unsteered generation, each over 5000 samples. No task-specific training is required.

Linker design

Problem	**Free Atoms**	**Succ.** ↑	**Uniq.** ↑
P1 (easy)	5	95.2	16.4
P2 (medium)	7	77.9	48.1
P3 (medium)	12	88.5	84.5
P4 (hard, 3 frags)	10	55.8	51.5
P5 (hard, 2 locs)	5 + 5	75.0	83.3
P6 (hard, 2 locs)	8 + 6	51.0	93.3
Average		73.9	62.85

Shape-steered generation

Target	**Mol. Stab.** ↑	**Bond Stab.** ↑	**Validity** ↑
Linear	20.76	95.57	88.28
Planar	8.96	92.90	88.80
Spherical	31.66	96.59	96.70
Unsteered	37.52	97.09	96.68

shift toward the target shape (Figure 2) while maintaining reasonable stability and validity (Table 4), confirming that the learned energy landscape contains valid basins even for shapes underrepresented in the training data.

Case study on linker design. The additive energy composition naturally enables conditioning on arbitrary inputs, without any retraining. To test this, we conduct a linker design case study where we fix a subset of atoms and sample the remaining ones. We evaluate this zero-shot linker design by removing linking fragments between stable parts of molecules like ring structures and replacing the removed atoms with Gaussian noise matched to the removed atoms spatial statistics. We subsequently run a modified parallel tempering sampler for 2000 iterations. Following the evaluation protocol of [56], EBMol achieves an average success rate of 73.9% across six conditioning tasks of varying complexity (Table 4), all without task-specific training.

4.4 Physical meaningfulness of the learned energy

The energy inductive bias not only improves generation quality but produces a physically meaningful scoring function. We evaluate the correlation between EBMol energy and the energy decrease upon GFN2-xTB [54] relaxation for molecules generated by the independently trained EDM (Figure 6 in Appendix C). The correlation ($r = 0.65, p < 0.0001$) emerges despite EBMol never receiving any energy-related supervision: the model is trained solely on atomic positions and atom types, yet learns to assign high energy to physically strained geometries. As a result, E_θ can serve as a lightweight first-pass filter for screening generated molecules by physical quality, providing order-of-magnitude speedups over GFN2-xTB optimization when screening large molecular batches (Appendix F).

5 Conclusion and limitations

Contributions. EBMol demonstrates the potential of energy-based models to serve as a versatile and practical framework for 3D molecular generation, achieving state-of-the-art results on QM9 and GEOM-Drugs while providing critical capabilities that standard distribution-matching frameworks lack: energy-based ranking, steering via potential composition, and conditional generation from a single model. The core components are Restoring Field Matching, which shapes a scalar energy landscape without simulation, and Mirror-Langevin sampling, unifying continuous and discrete updates within a principled framework. Perhaps most notably, EBMol generates structures surprisingly close to physical local minima: its median relaxation energy of 1.78 kcal/mol is an order of magnitude below previous bond-implicit methods. This performance is competitive with bond-explicit models at 18× the parameter count, despite EBMol requiring no bond or energy supervision.

Limitations. While we establish the design space for learning an energy landscape combined with Mirror-Langevin sampling, we do not systematically explore advanced sampler designs beyond the three algorithms in our sampler ablation (Table 13). Currently, achieving the best results requires a significant number of NFEs, and the employed sampling algorithms are sensitive to hyperparameter

choices. On the theoretical side, our analysis establishes well-posedness of the training problem and characterizes the solution at the regularization extremes, but the claim that data points correspond to local minima in the learned landscape is motivated by the target construction and verified only empirically.

Outlook. A natural extension to EBMol is integrating energy calibration by replacing the zero-anchor objective with a loss that aligns $E_\theta(\mathbf{x}_0)$ to a physical quantity such as DFT or GFN2-xTB energy values. A calibrated energy would also enable coupled generative-simulation workflows in which the same E_θ both generates candidate molecules and serves as an approximate molecular dynamics potential. On the sampling side, we expect significant improvements from more sophisticated samplers such as accelerated Langevin methods [57, 58] for faster sampling or from repulsive interaction potentials [17] for improved sample diversity.

6 Acknowledgements

The financial support by the LEAD project DigiBioTech of Graz University of Technology is gratefully acknowledged.

References

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

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

[3] Jean-Louis Reymond, Lars Ruddigkeit, Lorenz Blum, and Ruud van Deursen. The enumeration of chemical space. *WIREs Computational Molecular Science*, 2(5):717–733, 2012.

[4] Oliver T. Unke, Stefan Chmiela, Huziel E. Sauceda, Michael Gastegger, Igor Poltavsky, Kristof T. Schütt, Alexandre Tkatchenko, and Klaus-Robert Müller. Machine Learning Force Fields. *Chemical Reviews*, 121(16):10142–10186, August 2021.

[5] Peter Eastman, Pavan Kumar Behara, David L. Dotson, Raimondas Galvelis, John E. Herr, Josh T. Horton, Yuezhi Mao, John D. Chodera, Benjamin P. Pritchard, Yuanqing Wang, Gianni De Fabritiis, and Thomas E. Markland. SPICE, A Dataset of Drug-like Molecules and Peptides for Training Machine Learning Potentials. *Scientific Data*, 10(1):11, January 2023.

[6] Stefan Chmiela, Alexandre Tkatchenko, Huziel E. Sauceda, Igor Poltavsky, Kristof T. Schütt, and Klaus-Robert Müller. Machine learning of accurate energy-conserving molecular force fields. *Science Advances*, 3(5):e1603015, May 2017.

[7] Rokas Elijošius, Fabian Zills, Ilyes Batatia, Sam Walton Norwood, Dávid Péter Kovács, Christian Holm, and Gábor Csányi. Zero shot molecular generation via similarity kernels. *Nature Communications*, 16(1):5991, July 2025.

[8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Conference on Neural Information Processing Systems*, 2020.

[9] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

[10] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

[11] Xingchao Liu, Chengyue Gong, et al. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

[12] Kelvin Shuangjian Zhang, Gabriel Peyré, Jalal Fadili, and Marcelo Pereyra. Wasserstein control of mirror langevin monte carlo. In *Conference on learning theory*, pages 3814–3841. PMLR, 2020.

[13] Yann LeCun, Sumit Chopra, Raia Hadsell, Marc'Aurelio Ranzato, and Fu Jie Huang. A Tutorial on Energy-Based Learning. *Predicting structured data*, 2006.

[14] Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

[15] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Conference on Neural Information Processing Systems*, 2019.

[16] Yilun Du, Shuang Li, Joshua Tenenbaum, and Igor Mordatch. Improved contrastive divergence training of energy-based models. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2837–2848. PMLR, 18–24 Jul 2021.

[17] Michal Balcerak, Tamaz Amiranashvili, Antonio Terpin, Suprosanna Shit, Lea Bogensperger, Sebastian Kaltenbach, Petros Koumoutsakos, and Bjoern Menze. Energy Matching: Unifying Flow Matching and Energy-Based Models for Generative Modeling. In *Conference on Neural Information Processing Systems*, October 2025.

[18] Runqian Wang and Yilun Du. Equilibrium Matching: Generative Modeling with Implicit Energy-Based Models, October 2025.

[19] Priyank Jaini, Lars Holdijk, and Max Welling. Learning Equivariant Energy Based Models with Equivariant Stein Variational Gradient Descent. In *Conference on Neural Information Processing Systems*, November 2021.

[20] Ameya Daigavane, Song Eun Kim, Mario Geiger, and Tess Smidt. Symphony: Symmetry-Equivariant Point-Centered Spherical Harmonics for 3D Molecule Generation. In *International Conference on Representation Learning*, October 2024.

[21] Austin H. Cheng, Chong Sun, and Alán Aspuru-Guzik. Scalable Autoregressive 3D Molecule Generation, May 2025.

[22] Niklas W. A. Gebauer, Michael Gastegger, Stefaan S. P. Hessmann, Klaus-Robert Müller, and Kristof T. Schütt. Inverse design of 3d molecular structures with conditional generative neural networks. *Nature Communications*, 13(1):973, February 2022.

[23] Matthieu Kirchmeyer, Pedro O. Pinheiro, and Saeed Saremi. Score-based 3D molecule generation with neural fields. In *Conference on Neural Information Processing Systems*, November 2024.

[24] Matthieu Kirchmeyer, Pedro O. Pinheiro, Emma Willett, Karolis Martinkus, Joseph Kleinhenz, Emily K. Makowski, Andrew Martin Watkins, Vladimir Gligorijevic, Richard Bonneau, and Saeed Saremi. Unified all-atom molecule generation with neural fields. In *Conference on Neural Information Processing Systems*, October 2025.

[25] Xingang Peng, Jiaqi Guan, Qiang Liu, and Jianzhu Ma. MolDiff: Addressing the Atom-Bond Inconsistency Problem in 3D Molecule Diffusion Generation. In *International Conference on Machine Learning*, pages 27611–27629. PMLR, July 2023.

[26] Clement Vignac, Nagham Osman, Laura Toni, and Pascal Frossard. MiDi: Mixed Graph and 3D Denoising Diffusion for Molecule Generation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, May 2023.

[27] Han Huang, Leilei Sun, Bowen Du, and Weifeng Lv. Learning Joint 2D & 3D Diffusion Models for Complete Molecule Generation, June 2023.

[28] Ross Irwin, Alessandro Tibo, Jon Paul Janet, and Simon Olsson. SemlaFlow – Efficient 3D Molecular Generation with Latent Attention and Equivariant Flow Matching. In *International Conference on Artificial Intelligence and Statistics*. arXiv, 2025.

[29] Ian Dunn and David Ryan Koes. Mixed Continuous and Categorical Flow Matching for 3D De Novo Molecule Generation, April 2024.

[30] Danny Reidenbach, Filipp Nikitin, Olexandr Isayev, and Saee Paliwal. Applications of Modular Co-Design for De Novo 3D Molecule Generation, 2025. Version Number: 1.

[31] Noel M. O'Boyle, Michael Banck, Craig A. James, Chris Morley, Tim Vandermeersch, and Geoffrey R. Hutchison. Open Babel: An open chemical toolbox. *Journal of Cheminformatics*, 3(1):33, October 2011.

[32] Emiel Hoogeboom, Victor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant Diffusion for Molecule Generation in 3D. In *Conference on Neural Information Processing Systems*. arXiv, June 2022.

[33] Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E (n) equivariant graph neural networks. In *International Conference on Machine Learning*, pages 9323–9332. PMLR, 2021.

[34] Minkai Xu, Alexander S. Powers, Ron O. Dror, Stefano Ermon, and Jure Leskovec. Geometric Latent Diffusion Models for 3D Molecule Generation. In *International Conference on Machine Learning*, pages 38592–38610. PMLR, July 2023.

[35] Yuyan Ni, Shikun Feng, Haohan Chi, Bowen Zheng, Huan-ang Gao, Wei-Ying Ma, Zhi-Ming Ma, and Yanyan Lan. Straight-Line Diffusion Model for Efficient 3D Molecular Generation. In *Conference on Neural Information Processing Systems*, October 2025.

[36] Yuxuan Song, Jingjing Gong, Minkai Xu, Ziyao Cao, Yanyan Lan, Stefano Ermon, Hao Zhou, and Wei-Ying Ma. Equivariant Flow Matching with Hybrid Probability Transport for 3D Molecule Generation. In *Conference on Neural Information Processing Systems*, 2023.

[37] Leon Klein, Andreas Krämer, and Frank Noé. Equivariant flow matching. In *Conference on Neural Information Processing Systems*, 2023.

[38] Haokai Hong, Wanyu Lin, and Kay Chen Tan. Accelerating 3D Molecule Generation via Jointly Geometric Optimal Transport, March 2025.

[39] Yuxuan Song, Jingjing Gong, Hao Zhou, Mingyue Zheng, Jingjing Liu, and Wei-Ying Ma. Unified Generative Modeling of 3D Molecules with Bayesian Flow Networks. In *International Conference on Learning Representations*, October 2023.

[40] Lemeng Wu, Chengyue Gong, Xingchao Liu, Mao Ye, and Qiang Liu. Diffusion-based Molecule Generation with Informative Prior Bridges. In *Conference on Neural Information Processing Systems*, October 2022.

[41] Wenyin Zhou, Christopher Iliffe Sprague, Vsevolod Viliuga, Matteo Tadiello, Arne Elofsson, and Hossein Azizpour. Energy-Based Flow Matching for Generating 3D Molecular Structure. In *International Conference on Machine Learning*, June 2025.

[42] Guikun Xu, Xiaohan Yi, Peilin Zhao, and Yatao Bian. Energy-Guided Flow Matching Enables Few-Step Conformer Generation and Ground-State Identification, December 2025.

[43] Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, May 2003.

[44] Aharon Ben-Tal, Tamar Margalit, and Arkadi Nemirovski. The Ordered Subsets Mirror Descent Optimization Method with Applications to Tomography. *SIAM Journal on Optimization*, 12(1):79–108, January 2001.

[45] Ruilin Li, Molei Tao, Santosh S. Vempala, and Andre Wibisono. The Mirror Langevin Algorithm Converges with Vanishing Bias. In *International Conference on Algorithmic Learning Theory*, pages 718–742. PMLR, March 2022.

[46] Kwangjun Ahn and Sinho Chewi. Efficient constrained sampling via the mirror-Langevin algorithm. In *Conference on Neural Information Processing Systems*, 2021.

[47] Vishwak Srinivasan, Andre Wibisono, and Ashia Wilson. Fast sampling from constrained spaces using the metropolis-adjusted mirror langevin algorithm. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 4593–4635. PMLR, 2024.

[48] Guan-Horng Liu, Tianrong Chen, Evangelos Theodorou, and Molei Tao. Mirror Diffusion Models for Constrained and Watermarked Generation. In *Conference on Neural Information Processing Systems*, November 2023.

[49] Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

[50] François R. J. Cornet, Grigory Bartosh, Mikkel N. Schmidt, and Christian A. Naesseth. Equivariant Neural Diffusion for Molecule Generation. In *Conference on Neural Information Processing Systems*, June 2024.

[51] Shikun Feng, Yuyan Ni, Lu Yan, Zhi-Ming Ma, Wei-Ying Ma, and Yanyan Lan. UniGEM: A Unified Approach to Generation and Property Prediction for Molecules. In *International Conference on Learning Representations*, October 2024.

[52] Filipp Nikitin, Ian Dunn, David Ryan Koes, and Olexandr Isayev. GEOM-drugs revisited: toward more chemically accurate benchmarks for 3D molecule generation. *Digital Discovery*, 4(11):3282–3291, 2025.

[53] Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*, 2022.

[54] Christoph Bannwarth, Sebastian Ehlert, and Stefan Grimme. GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *Journal of Chemical Theory and Computation*, 15(3):1652–1671, March 2019.

[55] Thomas A Halgren. Merck molecular force field. i. basis, form, scope, parameterization, and performance of mmff94. *Journal of computational chemistry*, 17(5-6):490–519, 1996.

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

[57] Alexander Falk, Andreas Habring, Christoph Griesbacher, and Thomas Pock. An Inertial Langevin Algorithm, October 2025.

[58] Benedict J. Leimkuhler, Daniel Paulin, and Peter A. Whalley. Contraction and Convergence Rates for Discretized Kinetic Langevin Dynamics. *SIAM Journal on Numerical Analysis*, 62(3):1226–1258, June 2024.

[59] Santosh Vempala and Andre Wibisono. Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. *Advances in neural information processing systems*, 32, 2019.

[60] Dominique Bakry, Franck Barthe, Patrick Cattiaux, and Arnaud Guillin. A simple proof of the poincaré inequality for a large class of probability measures. 2008.

[61] Matthias Fey, Jinu Sunil, Akihiro Nitta, Rishi Puri, Manan Shah, Blaž Stojanovič, Ramona Bendias, Alexandria Barghi, Vid Kocijan, Zecheng Zhang, Xinwei He, Jan Eric Lenssen, and Jure Leskovec. PyG 2.0: Scalable Learning on Real World Graphs. In *Temporal Graph Learning Workshop @ KDD 2025*, August 2025.

A Proofs

In this section we provide rigorous formulations and proofs of the theoretical results in section 3.3. Let us first fix some notation. We denote the marginal densities of $\mathbf{x}_t$ by $p_t = \mathrm{law}(\mathbf{x}_t)$ and $p_x(\mathbf{x}) = \int p_t(\mathbf{x})\, p(t)\, \mathrm{d}t$. We denote the domain of $\mathbf{x}$ as $\Omega = \mathbb{R}^{N\times 3} \times \Delta^{K-1}$. We define the Sobolev space $W^{1,2}(\Omega) = \{f \in L^2(\Omega) \,|\, \nabla f \in L^2(\Omega)\}$ where the gradient is understood in the usual weak sense. Moreover, as the underlying solution space for our training problem we define the weighted Sobolev space

$$\mathcal{W}^{1,2}(\Omega) = \Big\{ f : \Omega \to \mathbb{R} \;\Big|\; \nabla f \text{ exists weakly and } \int_\Omega \big(|f(\mathbf{x})|^2 + |\nabla f(\mathbf{x})|^2\big)\, p_x(\mathbf{x})\, \mathrm{d}\mathbf{x} < \infty \Big\}. \quad (10)$$

We impose the following assumptions on the occurring distributions, respectively their densities.

Assumption 1.

1. *The domain* $\mathrm{dom}(p_x)$ *is connected,* p_x *is continuous, bounded above and strictly positive on* Ω*. Moreover,* p_x *satisfies a Poincaré inequality, that is, there exists* $C > 0$ *such that*

$$\|f - \bar{f}\|_{L^2(p_x)} \le C \|\nabla f\|_{L^2(p_x)}, \quad f \in \mathcal{W}^{1,2}(\Omega) \qquad \text{(PI)}$$

   where $\bar{f} = \int f(\mathbf{x}) p_x(\mathbf{x}) \mathrm{d}\mathbf{x}$.

2. *It holds* $p_0 \ll p_x$ *with* $\frac{p_0}{p_x}$ *locally bounded. Moreover, there exists* $A \subset \mathbb{R}^d$ *such that for* $\mathbf{x} \in A$, $p_0(\mathbf{x}) = 0$ *implies* $p_x(\mathbf{x}) = 0$ *and* $\int_A p_x(\mathbf{x}) \mathrm{d}\mathbf{x} > 0$.

Remark 1. *Let us briefly comment on assumption 1.*

1. *In the first assumption the most interesting part is the Poincaré inequality. This assumptions appears frequently in the context of diffusion processes and Langevin sampling [59]. A Poincaré inequality is, in particular, satisfied for a density p, if*

$$\langle \nabla \log p(\mathbf{x}), \mathbf{x} \rangle \leq -a\|\mathbf{x}\| + b$$

for some $a > 0$, $b \in \mathbb{R}$ [60]. That is, the Poincaré inequality imposes in some sense sufficiently fast decaying tails on the density.

2. *Regarding the second assumption, $p_0 \ll p_x$ is to be expected, as p_x is effectively a perturbed version of p_0. The converse absolute continuity on the set A is, for instance, satisfied if there exists a set A of positive measure p_x with $p_0 > 0$ on A.*

Before considering the main results, we prove the following equivalence of norms in $\mathcal{W}^{1,2}(\Omega)$.

Lemma 1. *Under assumption 1 for every $u \in \mathcal{W}^{1,2}(\Omega)$ we have*

$$\|u\|_{L^2(p_x)} \leq \|u\|_{L^2(p_0)} + \|\nabla u\|_{L^2(p_x)}. \tag{11}$$

Remark 2. *The important aspect above is that the first L2 norm on the right-hand side is with respect to p_0.*

Proof. Assume to the contrary, we find a sequence $(u_n)_n \subset \mathcal{W}^{1,2}(\Omega)$ with

$$\|u_n\|_{L^2(p_x)} \geq n(\|u_n\|_{L^2(p_0)} + \|\nabla u_n\|_{L^2(p_x)}). \tag{12}$$

Define

$$v_n = \frac{u_n}{\|u_n\|_{L^2(p_x)}} \tag{13}$$

so that $\|v_n\|_{L^2(p_x)} = 1$ and by (12)

$$\frac{1}{n} \geq (\|v_n\|_{L^2(p_0)} + \|\nabla v_n\|_{L^2(p_x)}). \tag{14}$$

In particular, we can extract a subsequence $(v_{n_k})_k$ such that

$$\begin{aligned} v_{n_k} &\rightharpoonup v \\ \nabla v_{n_k} &\rightharpoonup g \end{aligned} \tag{15}$$

both in $L^2(p_x)$ and by the assumed positivity and continuity of p_x we find in particular, that $v_{n_k} \rightharpoonup v$ in $W^{1,2}_{\text{loc}}(\Omega)$, that is, with respect to the regular Lebesgue measure, and $g = \nabla v$. Moreover, $v \in \mathcal{W}^{1,2}(\Omega)$. Now, (14) implies that $\nabla v_{n_k} \to 0$ and since weak limits are unique we have $\nabla v = g = 0$. The Poincaré inequality, *cf.* assumption 1, then implies

$$\|v - \bar{v}\|_{L^2(p_x)} \leq C\|\nabla v\|_{L^2(p_x)} = 0,$$

that is, $v(\mathbf{x})$ is equal to a constant, say $c \in \mathbb{R}$, p_x-a.e. Similarly, (14) yields $v_{n_k} \to 0$ in $L^2(p_0)$ so that $v(\mathbf{x}) = 0$ for p_0-a.e. $\mathbf{x}$. Using the set A from assumption 1 we find

$$c \int_A p_x(\mathbf{x}) \mathrm{d}\mathbf{x} = \int_A v(\mathbf{x}) \frac{p_x(\mathbf{x})}{p_0(\mathbf{x})} p_0(\mathbf{x}) \mathrm{d}\mathbf{x} = 0 \tag{16}$$

where the last equality follows from $v(\mathbf{x}) = 0$, p_0-a.e. Since $\int_A p_x(\mathbf{x}) \mathrm{d}\mathbf{x} > 0$ it follows $c = 0$, that is, $v = 0$, p_x-a.e. Thus, we have that $v_{n_k} \rightharpoonup 0$ in $L^2(p_x)$. Since constants are elements of $L^2(p_x)$ this also implies that

$$\bar{v}_{n_k} = \int v_{n_k}(\mathbf{x}) p_x(\mathbf{x}) \mathrm{d}\mathbf{x} \to 0. \tag{17}$$

Then, the Poincaré inequality implies

$$\begin{aligned} \|v_{n_k}\|_{L^2(p_x)} \leq& \|v_{n_k} - \bar{v}_{n_k}\|_{L^2(p_x)} + \|\bar{v}_{n_k}\|_{L^2(p_x)} \\ \leq& \|v_{n_k} - \bar{v}_{n_k}\|_{L^2(p_x)} + \bar{v}_{n_k} \\ \leq& C\|\nabla v_{n_k}\|_{L^2(p_x)} + \bar{v}_{n_k} \to 0 \end{aligned} \tag{18}$$

We, therefore, find

$$1 = \|v_{n_k}\|_{L^2(p_x)} \to 0 \tag{19}$$

which is a contradiction. □

We can now prove existence and uniqueness of solutions of the training problem. Let us first rigorously reformulate the result.

Proposition 3 (Existence of solutions)**.** *Under assumption 1 the problem*

$$\min_{E\in\mathcal{W}^{1,2}(\Omega)} \mathcal{L}(E)$$

admits a unique solution.

Proof. Denoting $\mathbf{u} = (\mathbf{u^c}, \mathbf{u^p})$ we can rewrite the objective as

$$\begin{aligned}\mathcal{L}(E) &= \iiint \left\{ \left\| -\nabla E_\theta(\mathbf{x}_t) - \mathbf{u} \right\|^2 + \lambda_{\text{reg}} E(\mathbf{x}_0)^2 \right\} p(\mathbf{x}_0, \mathbf{x}_t, t) \mathrm{d}\mathbf{x}_t \mathrm{d}\mathbf{x}_0 \mathrm{d}t \\ &= \iiint \left\| -\nabla E_\theta(\mathbf{x}_t) - \mathbf{u} \right\|^2 p(\mathbf{x}_0, \mathbf{x}_t, t) \mathrm{d}\mathbf{x}_t \mathrm{d}\mathbf{x}_0 \mathrm{d}t + \lambda_{\text{reg}} \int E(\mathbf{x})^2 p_0(\mathbf{x}) \mathrm{d}\mathbf{x}\end{aligned} \tag{20}$$

Now let $(E_n)_n \subset \mathcal{W}^{1,2}(\Omega)$ be a minimizing sequence, that is,

$$\lim_{n\to\infty} \mathcal{L}(E_n) = \inf_{E\in\mathcal{W}^{1,2}(\Omega)} \mathcal{L}(E).$$

By definition of $\mathcal{L}$ it follows that $\|E_n\|_{L^2(p_0)}$ and $\|\nabla E_n\|_{L^2(p_x)}$ are bounded. The Poincaré inequality, *cf.* lemma 1 then yields that also $\|E_n\|_{L^2(p_x)}$ is bounded so that we can extract a subsequence $(E_{n_k})_k$ and functions $E^* \in L^2(p_x)$, $g \in L^2(p_x)$ such that $E_{n_k} \rightharpoonup E^*$, $\nabla E_{n_k} \rightharpoonup g$ both in $L^2(p_x)$. Moreover, by the assumed properties of p_x we also find that the weak convergence is in $W^{1,2}(\Omega)$ and that $\nabla E^* = g$ and $E^* \in \mathcal{W}^{1,2}(\Omega)$. By potentially extracting another subsequence, E_{n_k} converges also weakly in $L^2(p_0)$ and the limit is again E^* due to the assumed boundedness of p_0/p_x as the latter implies that weak convergence in $L^2(p_x)$ includes also weak convergence in $L^2(p_0)$. Note that

$$G \mapsto \iiint \left\| -G(\mathbf{x}_t) - \mathbf{u} \right\|^2 p(\mathbf{x}_0, \mathbf{x}_t, t) \mathrm{d}\mathbf{x}_t \mathrm{d}\mathbf{x}_0 \mathrm{d}t. \tag{21}$$

is convex and continuous in $L^2(p_x)$ and therefore also lower semi-continuous with respect to weak convergence in $L^2(p_x)$. Similarly,

$$E \mapsto \iiint \lambda_{\text{reg}} E(\mathbf{x}_0)^2 p(\mathbf{x}_0) \mathrm{d}\mathbf{x}_0 \tag{22}$$

is weakly lower semi-continuous in $L^2(p_0)$. Thus,

$$\mathcal{L}(E^*) \le \liminf_{n\to\infty} \mathcal{L}(E_n) = \inf_E \mathcal{L}(E) \tag{23}$$

showing existence of a solution. Uniqueness is a consequence of strict convexity of the objective. □

Having established well-posedness we can consider the concise formulation and proof of proposition 2.

Proposition 4 (Convergence for small and large regularization parameters)**.** *Under assumption 1, let $E^*(x, \lambda)$ be the minimizer of $\mathcal{L}$ for $\lambda_{reg} = \lambda$. It holds true that*

(i) There exists a weak accumulation point $E^*_\infty \in \mathcal{W}^{1,2}(\Omega)$ of $E^*(\cdot, \lambda)$ as $\lambda \to \infty$ and every such accumulation point satisfies $E^*_\infty(x) = 0$ for p_0-a.e. x.

(ii) There exists $E^*_0 \in \mathcal{W}^{1,2}(\Omega)$ such that ∇E^*_0 is a weak accumulation point of $\nabla E^*(\cdot, \lambda)$ as $\lambda \to 0$ and every such E^*_0 minimizes

$$\mathcal{W}^{1,2}(\Omega) \ni G \mapsto \iiint \left\| -\nabla G(\mathbf{x}_t) - \mathbf{u} \right\|^2 p(\mathbf{x}_0, \mathbf{x}_t, t) \mathrm{d}\mathbf{x}_t \mathrm{d}\mathbf{x}_0 \mathrm{d}t, \tag{24}$$

that is, ∇E^*_0 is precisely the projection of $\mathbf{u}$ onto the space of all gradient fields. Moreover, if there exists $F \in \mathcal{W}^{1,2}(\Omega)$ with

$$\nabla F(x) = -\mathbb{E}[\mathbf{u}|\mathbf{x}_t = \mathbf{x}], \quad p_x - a.e.\ \mathbf{x}. \tag{25}$$

then $\nabla E^*(\mathbf{x}) = -\mathbb{E}[\mathbf{u}|\mathbf{x}_t = \mathbf{x}]$, p_x-a.e.

(iii) If, in addition it holds true that $v(\mathbf{x}_0, \mathbf{x}_1, t) = v(\mathbf{x}_t)$, that is, v depends solely on $\mathbf{x}_t$, then there exists a weak accumulation point $E_0^* \in \mathcal{W}^{1,2}(\Omega)$ of $E^*(\cdot, \lambda)$ as $\lambda \to 0$ and every such accumulation point satisfies $\nabla E_0^* = -\mathbf{u}$, p_x-a.e.

Proof. We begin with the limit for $\lambda \to \infty$. Assume without loss of generality $\lambda \geq 1$. By optimality of $E^*(\cdot, \lambda)$

$$\begin{aligned}
&\int E^*(\mathbf{x}_0, \lambda)^2 p_0(x_0) \mathrm{d}\mathbf{x}_0 \\
&\quad \leq \iiint \left\| -\nabla E^*(\mathbf{x}_t, \lambda) - \mathbf{u} \right\|^2 p(\mathbf{x}_0, \mathbf{x}_t, t) \mathrm{d}\mathbf{x}_t \mathrm{d}\mathbf{x}_0 \mathrm{d}t + \lambda \int E^*(\mathbf{x}_0, \lambda)^2 p_0(\mathbf{x}_0) \mathrm{d}\mathbf{x}_0 \\
&\quad \leq \iiint \left\| 0 - \mathbf{u} \right\|^2 p(\mathbf{x}_0, \mathbf{x}_t, t) \mathrm{d}\mathbf{x}_t \mathrm{d}\mathbf{x}_0 \mathrm{d}t + \lambda \int 0 \mathrm{d}\mathbf{x}_0 \\
&\quad \leq \iiint \left\| \mathbf{u} \right\|^2 p(\mathbf{x}_0, \mathbf{x}_t, t) \mathrm{d}\mathbf{x}_t \mathrm{d}\mathbf{x}_0 \mathrm{d}t < \infty
\end{aligned} \tag{26}$$

which implies that $\|E^*(\cdot, \lambda)\|_{L^2(p_0)}$ and $\|\nabla E^*(\cdot, \lambda)\|_{L^2(p_x)}$ are bounded for all $\lambda \geq 1$. As above, we can therefore extract a weakly convergent subsequence $E^*(\cdot, \lambda_n) \to E_\infty^*$ in $\mathcal{W}^{1,2}(\Omega)$ as well as in $L^2(p_0)$. By weak lower semi-continuity of $E \mapsto \int E(\mathbf{x})^2 p_0(\mathbf{x}) \mathrm{d}\mathbf{x}$ and (26) it follows

$$\begin{aligned}
\int E_\infty^*(\mathbf{x}_0)^2 p_0(\mathbf{x}_0) \mathrm{d}\mathbf{x}_0 &\leq \liminf_{n \to \infty} \int E^*(\mathbf{x}_0, \lambda_n)^2 p_0(\mathbf{x}_0) \mathrm{d}\mathbf{x}_0 \\
&\leq \frac{1}{\lambda_n} \iiint \left\| \mathbf{u} \right\|^2 p(\mathbf{x}_0, \mathbf{x}_t, t) \mathrm{d}\mathbf{x}_t \mathrm{d}\mathbf{x}_0 \mathrm{d}t \to 0
\end{aligned} \tag{27}$$

as $n \to \infty$ implying $E_\infty^* = 0$, $p(\mathbf{x}_0)$-a.e.

Regarding the convergence as $\lambda \to 0$, as above we find that

$$\begin{aligned}
&\iiint \left\| -\nabla E^*(\mathbf{x}_t, \lambda) - \mathbf{u} \right\|^2 p(\mathbf{x}_0, \mathbf{x}_t, t) \mathrm{d}\mathbf{x}_t \mathrm{d}\mathbf{x}_0 \mathrm{d}t \\
&\qquad \leq \iiint \left\| \mathbf{u} \right\|^2 p(\mathbf{x}_0, \mathbf{x}_t, t) \mathrm{d}\mathbf{x}_t \mathrm{d}\mathbf{x}_0 \mathrm{d}t < \infty.
\end{aligned} \tag{28}$$

Thus, we can only deduce boundedness of $\nabla E^*(\cdot, \lambda)$ in $L^2(p_x)$ as $\lambda \to 0$. By the Poincaré inequality (*cf.* assumption 1), however, we also have that $F^*(\cdot, \lambda) = E^*(\cdot, \lambda) - \bar{E}^*(\lambda)$ is bounded in $L^2(p_x)$. Therefore, we may extract a subsequence such that $F^*(\cdot, \lambda_n) \rightharpoonup F_0^*$ as $n \to \infty$. In particular, it follows that $\nabla E^*(\cdot, \lambda_n) = \nabla F^*(\cdot, \lambda_n) \rightharpoonup \nabla F_0^*$. Once again, by weak lower semicontinuity and optimality we find for any $G \in \mathcal{W}^{1,2}(\Omega)$

$$\begin{aligned}
&\iiint \left\| -\nabla F_0^*(\mathbf{x}_t) - \mathbf{u} \right\|^2 p(\mathbf{x}_0, \mathbf{x}_t, t) \mathrm{d}\mathbf{x}_t \mathrm{d}\mathbf{x}_0 \mathrm{d}t \\
&\quad \leq \liminf_{n \to \infty} \iiint \left\| -\nabla E^*(\mathbf{x}_t, \lambda_n) - \mathbf{u} \right\|^2 p(\mathbf{x}_0, \mathbf{x}_t, t) \mathrm{d}\mathbf{x}_t \mathrm{d}\mathbf{x}_0 \mathrm{d}t \\
&\quad \leq \liminf_{n \to \infty} \iiint \left\| -\nabla E^*(\mathbf{x}_t, \lambda_n) - \mathbf{u} \right\|^2 p(\mathbf{x}_0, \mathbf{x}_t, t) \mathrm{d}\mathbf{x}_t \mathrm{d}\mathbf{x}_0 \mathrm{d}t + \lambda_n \int E^*(\mathbf{x}_0, \lambda)^2 p(\mathbf{x}_0) \mathrm{d}\mathbf{x}_0 \\
&\quad \leq \liminf_{n \to \infty} \iiint \left\| -\nabla G(\mathbf{x}_t) - \mathbf{u} \right\|^2 p(\mathbf{x}_0, \mathbf{x}_t, t) \mathrm{d}\mathbf{x}_t \mathrm{d}\mathbf{x}_0 \mathrm{d}t + \lambda_n \int G(\mathbf{x}_0)^2 p(\mathbf{x}_0) \mathrm{d}\mathbf{x}_0 \\
&\quad \leq \iiint \left\| -\nabla G(\mathbf{x}_t) - \mathbf{u} \right\|^2 p(\mathbf{x}_0, \mathbf{x}_t, t) \mathrm{d}\mathbf{x}_t \mathrm{d}\mathbf{x}_0 \mathrm{d}t
\end{aligned} \tag{29}$$

implying the desired result. The representation, $\nabla E^*(\mathbf{x}) = -\mathbb{E}[\mathbf{u} | \mathbf{x}_t = \mathbf{x}]$ follows directly by the fact that

$$w \mapsto \iiint \left\| w(\mathbf{x}_t) - \mathbf{u} \right\|^2 p(\mathbf{x}_0, \mathbf{x}_t, t) \mathrm{d}\mathbf{x}_t \mathrm{d}\mathbf{x}_0 \mathrm{d}t \tag{30}$$

is minimized for $w(\mathbf{x}) = \mathbb{E}[\mathbf{u} | \mathbf{x}_t = \mathbf{x}]$.

In the last case, we note that

$$\begin{aligned}&\iiint \left\|w(\mathbf{x}_t)-\mathbf{u}\right\|^2 p(\mathbf{x}_0,\mathbf{x}_t,t)\mathrm{d}\mathbf{x}_t\mathrm{d}\mathbf{x}_0\mathrm{d}t\\&\quad=\int\left(\iint\left\|w(\mathbf{x}_t)-\mathbf{u}\right\|^2 p(\mathbf{x}_0,t|\mathbf{x}_t)\mathrm{d}\mathbf{x}_0\mathrm{d}t\right)p(\mathbf{x}_t)\mathrm{d}\mathbf{x}_t\end{aligned}\tag{31}$$

which is zero for $w(\mathbf{x}_t)=\mathbf{u}(\mathbf{x}_t)$. Therefore, under the assumption that there exists $F\in\mathcal{W}^{1,2}(\Omega)$ with $\nabla F(\mathbf{x}_t)=\mathbf{u}(\mathbf{x}_t)$, we find by optimality of $E^*(\cdot,\lambda)$

$$\begin{aligned}&\iiint\left\|-\nabla E^*(\mathbf{x}_t,\lambda)-\mathbf{u}\right\|^2 p(\mathbf{x}_0,\mathbf{x}_t,t)\mathrm{d}\mathbf{x}_t\mathrm{d}\mathbf{x}_0\mathrm{d}t+\lambda_n\int E^*(\mathbf{x}_0,\lambda)^2p_0(\mathbf{x}_0)\mathrm{d}\mathbf{x}_0\\&\quad\le\iiint\left\|\nabla F(\mathbf{x}_t)-\mathbf{u}\right\|^2 p(\mathbf{x}_0,\mathbf{x}_t,t)\mathrm{d}\mathbf{x}_t\mathrm{d}\mathbf{x}_0\mathrm{d}t+\lambda_n\int F(\mathbf{x}_0)^2p_0(\mathbf{x}_0)\mathrm{d}\mathbf{x}_0\\&\quad=\lambda_n\int F(\mathbf{x}_0)^2p_0(\mathbf{x}_0)\mathrm{d}\mathbf{x}_0\end{aligned}\tag{32}$$

so that in this case obtain again boundedness of $\nabla E^*(\cdot,\lambda)$ and $E^*(\cdot,\lambda)$ both in $L^2(p_x)$. Extracting weakly convergent subsequences again with limit E_0^* we find by weak lower semi-continuity

$$\begin{aligned}&\iiint\left\|\nabla E_0^*(\mathbf{x}_t)-\mathbf{u}\right\|^2 p(\mathbf{x}_0,\mathbf{x}_t,t)\mathrm{d}\mathbf{x}_t\mathrm{d}\mathbf{x}_0\mathrm{d}t\\&\quad\le\liminf_{n\to\infty}\iiint\left\|\nabla E^*(\mathbf{x}_t,\lambda_n)-\mathbf{u}\right\|^2 p(\mathbf{x}_0,\mathbf{x}_t,t)\mathrm{d}\mathbf{x}_t\mathrm{d}\mathbf{x}_0\mathrm{d}t\\&\quad\le\liminf_{n\to\infty}\iiint\left\|\nabla E^*(\mathbf{x}_t,\lambda_n)-\mathbf{u}\right\|^2 p(\mathbf{x}_0,\mathbf{x}_t,t)\mathrm{d}\mathbf{x}_t\mathrm{d}\mathbf{x}_0\mathrm{d}t+\lambda_n\int E^*(\mathbf{x}_0,\lambda)^2p_0(\mathbf{x}_0)\mathrm{d}\mathbf{x}_0\\&\quad\le\liminf_{n\to\infty}\lambda_n\int F(\mathbf{x}_0,\lambda)^2p_0(\mathbf{x}_0)\mathrm{d}\mathbf{x}_0=0\end{aligned}\tag{33}$$

concluding the proof. □

B Empirical Verification of Local Minima

The RFM construction aims to place data points at local minima of the learned energy, as the smoothing function ensures pointwise stationarity at $t=0$, and the symmetric restoring targets induce non-negative curvature around data. The theoretical analysis in Section 3.3 establishes existence and uniqueness of the loss minimizer and characterizes its behavior under varying regularization strength, but does not formally prove the local-minima property for the marginalized predictor. Here we verify empirically that the trained network E_θ realizes this intended behavior through three complementary experiments.

Deterministic gradient relaxation from data. We initialize 1000 molecules taken from the training and test sets and run deterministic gradient descent ($\tau=0$) for $N=500$ steps with step size $\eta=0.01$ on the learned energy landscape. We evaluate atom displacement via mean RMSD, the energy change during relaxation, and the change in molecule stability. We conduct this experiment for two models: EBMol trained with the full RFM objective, and the OTFM ablation from Table 12, which is trained without the extrapolation construction.

Table 5: Deterministic gradient relaxation initialized at data. Energy values, RMSD (in), and stability after N steps of gradient descent on E_θ.

Method	**Split**	**Median $\Delta E\downarrow$**	**Mean RMSD $\downarrow$**	**ΔMol. Stability $\downarrow$**
EBMol (RFM)	Train	-0.59	0.121	2.4
EBMol (RFM)	Test	-0.67	0.124	2.1
OTFM (12)	Train	-154.98	143.365	63.1

Results are reported in Table 5. Three observations confirm the role of the RFM construction. First, on training data, EBMol configurations remain effectively stationary (RMSD 0.121 , stability loss

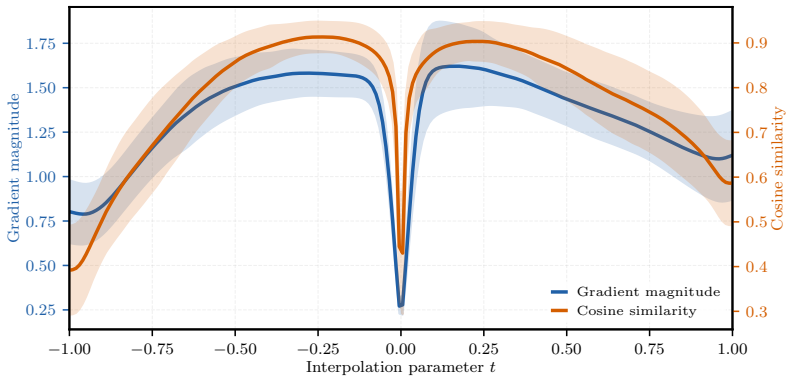


Figure 3: Gradient (wrt. position) magnitude (blue) and cosine similarity between the learned restoring field and the prescribed target direction (orange) along interpolation paths from prior ($t = \pm 1$) through data ($t = 0$). Shaded regions indicate one standard deviation over 1000 molecules. The gradient vanishes near data, confirming stationarity, while the high cosine similarity away from data confirms that the learned field consistently points toward training molecules.

2.4 percentage points), verifying that the RFM loss produces energy minima at training samples as intended. Second, the same holds on held-out test data (RMSD 0.124, stability loss 2.1 percentage points), demonstrating that the local-minima property generalizes to molecules unseen during training. Third, under the OTFM loss, configurations diverge dramatically (RMSD 143.4) into spurious minima that do not correspond to valid molecular structures, with molecule stability dropping by 63 percentage points. This confirms that the extrapolation construction of RFM is responsible for placing data at local minima.

Energy and gradient profiles along interpolation paths. We evaluate the gradient magnitude with respect to the position $\|\nabla_c E_\theta(\mathbf{x}_t)\|$ and the cosine similarity between $\nabla_c E_\theta(\mathbf{x}_t)$ and the unsmoothed restoring direction $\mathbf{u}^{\mathbf{c}} = \text{sign}(t) \cdot (\mathbf{c}_0 - \mathbf{c}_1)$, along linear interpolation paths between training molecules ($t = 0$) and corresponding optimal-transport-matched prior samples ($t = \pm 1$).

Figure 3 shows the results averaged over 1000 molecules. The gradient magnitude drops sharply near $t = 0$, consistent with data sitting at stationary points of the learned energy. Away from data, the gradient grows monotonically toward the prior, indicating increasingly steep energy walls around each basin. The cosine similarity remains high (> 0.6) throughout most of the interval, confirming that the learned gradient field points back toward data across the interpolation range, and drops near $t = 0$ where the vanishing gradient magnitude makes the direction ill-defined. The symmetric profile across $t < 0$ and $t > 0$ reflects the symmetric construction of the RFM target.

Response to Gaussian perturbation. We perturb training molecules by adding isotropic Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 I)$ to atomic coordinates for increasing noise levels σ and evaluate E_θ at the perturbed configurations. Figure 4 shows the learned energy as a function of σ, averaged over 1000 molecules. The energy increases smoothly and monotonically with perturbation magnitude, starting near zero at clean data. This confirms that data points sit at the center of well-shaped energy basins, with the learned landscape assigning progressively higher energy to configurations that deviate from valid molecular geometries.

C Structural and energetic analysis

Bond distance distributions. Figure 5 provides a per-element decomposition of the aggregate bond length errors reported in Table 3, comparing bond distance distributions for molecules generated by SLDM and EBMol_{7240} against the training data on GEOM-Drugs. EBMol recovers the C–C distribution with high fidelity, accurately capturing peak locations, relative heights, and the separation between single, aromatic, and double bond modes. For C–N and C–O, EBMol correctly resolves relative peak proportions, which SLDM fails to achieve. EBMol also successfully recovers the bimodality of C–H bonds.

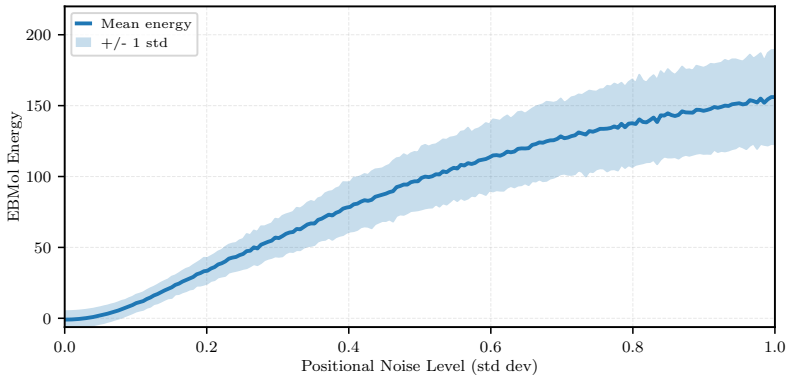


Figure 4: Learned energy E_θ as a function of Gaussian noise magnitude σ applied to training molecule coordinates. Shaded regions indicate one standard deviation over 1000 molecules. The monotonic increase confirms that data points occupy local minima of the learned landscape.

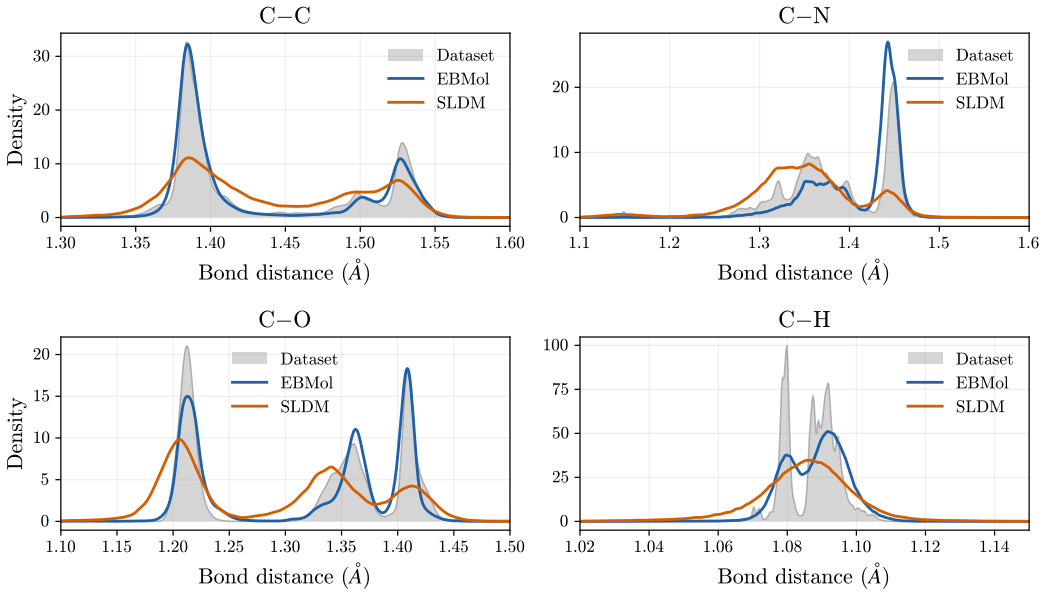


Figure 5: Bond distance distributions for C–C, C–N, C–O, and C–H pairs on GEOM-Drugs, comparing EBMol_{7240} (blue) and SLDM (orange) against the dataset (gray). Distinct peaks correspond to different bond orders. EBMol recovers the multimodal structure of the reference distributions, with sharper peak separation and more realistic proportions than SLDM.

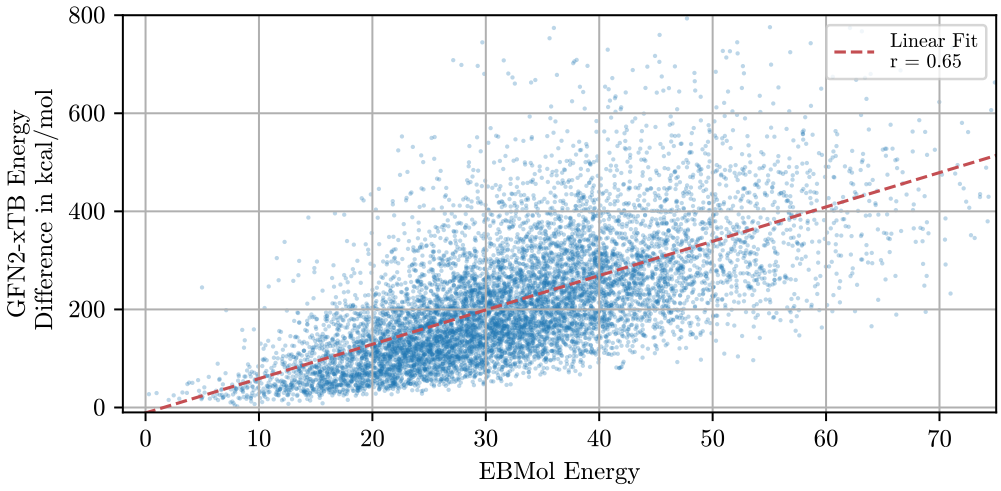


Figure 6: The learned energy correlates with physical quality across models. EBMol energy vs. energy decrease upon GFN2-xTB relaxation for 10k molecules generated by EDM ($r = 0.65$, $p < 0.0001$). Higher EBMol energy predicts larger relaxation energy, indicating physically strained geometries.

Cross-model energy correlation. Figure 6 plots EBMol's learned energy against the energy decrease upon GFN2-xTB relaxation for 10,000 molecules generated by EDM. The positive correlation and its interpretation are discussed in Section 4.4.

D Perception pipeline analysis and failure modes on GEOM-Drugs

Bonds, bond orders, formal charges, and aromaticity are not generated by the model but inferred post-hoc from coordinates and atom types via OpenBabel [31]. A failed stability or validity check can therefore either be caused by an erroneous generated structure or from a limitation of the perception pipeline itself. To separate the two, we first apply the pipeline to the GEOM-Drugs reference conformers, for which the ground-truth SMILES allows each failure to be attributed to its cause, and subsequently report the same diagnostics for generated molecules.

Failure modes of the evaluation pipeline. Although every reference conformer corresponds to a valid molecule, only 88.2% pass the molecule stability check (Table 2). In Table 6, we attribute the remaining 11.8% to its cause. The dominant failure is over-valence, which co-occurs predominantly with formal charges in the reference SMILES. OpenBabel does not reliably recover charged forms from geometry and the uncharged perception then violates the valency tables. Under-valence arises from bonds and aromatic systems that OpenBabel fails to perceive. Notably, OpenBabel issues a kekulization warning for only 1.3% of reference molecules, whereas 3.5% contain an aromatic system that is not perceived, indicating that the warning is not a reliable indicator of kekulization failures.

Failure modes of generated molecules. For generated molecules no reference SMILES is available, so only the pipeline's own signals can be reported. Results are shown in Table 7. EBMol's failures are dominated by under-valence, i.e. atoms whose perceived bond orders fall short of their allowed valence, while disconnected and over-valent molecules are comparatively rare. All valence-related failure rates, as well as disconnectedness and sanitization failures, decrease monotonically with the sampling budget. In contrast, EDM and SLDM fail primarily through disconnected fragments (57.6% and 49.4%).

The kekulization warning should be interpreted with care. OpenBabel must perceive an aromatic system before kekulization can fail on it, so the warning rate is biased downward for models that produce few perceivable aromatic systems. EDM, the least stable model, yields only 4.5% aromatic atoms against 26.2% in the data. We therefore report the kekulization warning as a diagnostic rather than as an aromaticity error rate.

Table 6: Failure modes of the evaluation pipeline on GEOM-Drugs reference conformers. A molecule is over-/under-valent if at least one atom's perceived bond-order sum lies above/below its allowed valence. Multiple failure modes may occur per molecule, so columns do not sum to 100%.

Failure mode	% of all	% of unstable
Disconnected	0.00	0.0
Over-valent	7.70	65.4
Under-valent	6.00	50.9
Formal charge in reference SMILES	5.96	50.4
Aromatic system not perceived	3.54	30.1
RDKit sanitization failure	2.10	17.7
OpenBabel kekulization warning	1.30	11.0
Connectivity differs from reference	0.32	2.7

Table 7: Failure modes of generated molecules on GEOM-Drugs (% of all molecules). Over-/under-valent as in Table 6. Rows overlap and report per-molecule presence.

	Data	EBMol_{1080}	EBMol_{1960}	EBMol_{3720}	EBMol_{7240}	EDM	SLDM
Valid ↑	97.9	95.4	97.0	98.6	99.0	84.4	87.0
Stable ↑	88.2	24.7	37.4	52.3	61.4	1.0	20.6
Disconnected ↓	0.0	14.4	9.1	4.7	3.8	57.6	49.4
Sanitization failure ↓	2.1	4.6	3.0	1.4	1.0	15.6	13.0
Over-valent ↓	7.7	8.8	6.5	4.6	4.1	18.4	25.4
Under-valent ↓	6.0	73.2	60.3	45.5	36.4	98.7	74.6
Aromatic atoms perceived	26.2	17.5	19.1	21.0	22.0	4.5	17.5
Kekulization warning	1.3	21.9	20.8	18.9	17.9	23.6	14.9

E Sampling Cost

Wall-clock sampling cost. Table 8 reports wall-clock sampling cost on a single NVIDIA RTX 4090, measured both per sample and per atom. We additionally report per-atom cost because the parallel-tempering (PT) sampler spends less compute on smaller molecules, which per-sample numbers alone would hide. Regarding implementation, we use a PyTorch-Geometric-based implementation that batches molecules as disconnected graphs, where EDM and SLDM pad every molecule to the GEOM maximum of 181 atoms instead. Although EBMol runs at substantially higher NFE, its throughput is comparable to the diffusion and flow baselines.

Table 8: Sampling throughput on a single NVIDIA RTX 4090. The suffix denotes the number of function evaluations (NFE).

Method	atoms/s	s/sample
EBMol_{1080}	38.4	1.07
EBMol_{1960}	18.5	2.19
EBMol_{3720}	9.4	4.24
EBMol_{7240}	5.2	8.04
SLDM_{200}	10.4	4.49
SLDM_{1000}	4.3	10.91
EDM_{1000}	5.8	8.03

Cost per high-quality sample. To make that tradeoff between sampling cost and sample quality more explicit, we report the cost per high-quality (hq) sample. We define a sample as high-quality if it is stable, valid, connected, and has a GFN2-xTB relaxation energy below a threshold τ. We measure cost in NFEs to decouple it from implementation-specific overheads such as padding (Table 9).

EBMol requires substantially less compute per high-quality sample than SLDM at every threshold. The gap is largely due to SLDM's low hq share, which is mostly caused by disconnected samples.

Table 9: Share of high-quality samples and NFEs per high-quality sample for different relaxation-energy thresholds τ (kcal/mol). Bold marks the lowest cost per threshold.

	$\tau = 15$		$\tau = 5$		$\tau = 2$		$\tau = 1$	
Method	hq (%)	NFE/hq	hq (%)	NFE/hq	hq (%)	NFE/hq	hq (%)	NFE/hq
EBMol_{1080}	22.7	**4766**	18.1	**5964**	11.5	**9412**	4.8	22548
EBMol_{1960}	35.1	5590	30.6	6412	20.8	9433	9.4	20769
EBMol_{3720}	50.2	7413	46.6	7990	35.4	10502	18.1	**20550**
EBMol_{7240}	59.2	12240	56.5	12806	45.5	15928	24.2	29906
SLDM_{1000}	7.6	13081	1.9	53951	0.2	665400	0.04	2495250

Stricter energy thresholds raise the cost per hq sample for every configuration.They also shift the most compute-efficient EBMol configuration toward higher sampling budgets: EBMol-1080 is the most economical for $\tau \geq 2$ kcal/mol, while EBMol-3720 is the cheapest at $\tau = 1$ kcal/mol.

F First-pass filtering experiments

We evaluate whether the learned energy can serve as a fast first-pass filter for identifying low-relaxation-energy molecules among sampled generated by other generative models. Ground-truth labels are GFN2-xTB relaxation energies, and MMFF94 serves as the fast-filter baseline. We filter the outputs of the generative models EDM [32], GeoLDM [34], and SLDM [35].

Throughput. Table 10 reports throughput on 9112 molecules on a single NVIDIA RTX 4090. Energy evaluation with EBMol is a single batched forward pass, so its throughput matches MMFF94 and is roughly $500\times$ higher than GFN2-xTB relaxation.

Table 10: Throughput for first-pass filtering on 9112 molecules.

Method	Throughput (mol/s)	Notes
EBMol	2850	single forward pass, batch size 500 (GPU)
MMFF94	2204	
GFN2-xTB	6.1	16 CPU threads

Retrieval performance. We define the low-energy set as the lowest decile of molecules by GFN2-xTB relaxation energy and measure how well each filter's ranking recovers it. Precision@k is the fraction of the top-k ranked molecules that belong to the low-energy set; recall@k is the fraction of the low-energy set contained in the top-k. A random ranking achieves precision 0.10 at every k and recall@25% of 0.25.

Table 11: First-pass filtering performance (EBMol / MMFF94). The low-energy set is the lowest decile by GFN2-xTB relaxation energy.

Metric	EDM	GeoLDM	SLDM
Precision@1%	**0.90** / 0.47	**0.93** / 0.32	**0.42** / 0.12
Precision@10%	**0.54** / 0.41	**0.54** / 0.29	**0.37** / 0.15
Recall@25%	**0.84** / 0.75	**0.86** / 0.65	**0.71** / 0.40

We identify and test two use-cases and report results in Table 11. When expensive relaxation is applied only to the top-ranked candidates, precision@1% is the relevant metric as it ensures that compute is only spent on the most promising candidates. For EDM and GeoLDM samples, 90% and 93% of EBMol's top 1% belong to the low-energy set, so nearly the entire relaxation budget is spent on true hits. When the goal is to discard poor samples while retaining good ones, recall@25% is the relevant metric: keeping only the top quarter of the pool preserves 84–86% of the low-energy set on EDM and GeoLDM and 71% for SLDM. Across all three generators and all operating points, EBMol outperforms MMFF94 at comparable throughput.

G Training objective and sampler ablation

Training objective ablation. A learned energy landscape requires a sampler to generate samples. We therefore evaluate each training objective under three different samplers (Forward Euler (FWDE), Annealed Langevin Dynamics (ALD), and Parallel Tempering (PT)) to disentangle the contributions of training and sampling. All ablations use a reduced-size model with embedding dimension halved to 128 to keep the search tractable. To ensure a fair comparison, we tune sampler hyperparameters separately for each training objective. For FWDE, we sweep the step size over $[0.05, 0.001]$ and select the configuration with the highest atom stability. For the FM baselines this controls how finely the $[0, 1]$ integration interval is discretized, giving them the best possible performance under their natural sampling procedure. For ALD, we lock in the selected step size and additionally sweep over the noise scale σ, using the same value for coordinates and atom types for simplicity. For PT, we reuse the ALD configuration. We acknowledge this sequential scheme is suboptimal but it keeps the search feasible across all method-sampler combinations. All experiments use a fixed budget of 1080 NFE per sample on average and $n = 500$ samples per experiment. In addition to atom stability and validity, we report two metrics specific to this ablation: W2, the Wasserstein-2 distance between pairwise atom distance distributions of generated and reference molecules restricted to the 1.0–2.0 Å range most relevant for bond lengths, and the number of diverged samples that exceeded plausible energy bounds during sampling and were discarded. Results are reported in Table 12.

We ablate against two baselines that share EBMol's energy architecture but are trained with the flow matching loss: Standard FM (without optimal transport or atom permutation alignment) and Equivariant OT-FM (with SE(3)-equivariant optimal transport coupling and atom permutation). The gap between the two is substantial, confirming that equivariant OT alignment is critical for learning a useful energy landscape in our framework. However, Equivariant OT-FM produces a significant number of diverged samples across all samplers, consistent with the relaxation experiments in Section B: without the RFM construction, data points are not local minima, and gradient-based sampling can escape into unphysical regions. This problem is most severe under PT, where high-temperature chains actively encourage exploration of the landscape. We note that the 1074 diverged samples for only 500 generated samples is possible because our parallel tempering implementation replaces diverged samples with fresh prior samples periodically. This also precludes inference-time compute scaling, since increasing the sampling budget leads to more divergences rather than better samples.

The RFM block tells a different story. No RFM variant produces a single diverged sample, confirming that the restoring-field construction yields a landscape that is globally stable under sampling. Within the RFM variants, several trends emerge. Comparing RFM with and without the smoothing function $b(t)$ shows that smoothing improves the W2 bond-distance metric, suggesting that the smoother basins are easier for the sampler to navigate, though PT partially compensates for the absence of smoothing. Moving from a static isotropic prior to the data-driven anisotropic prior described in the Implementation section yields a further improvement, most visible in the Valid & Connected rate: the full RFM model achieves 37.7% under PT, compared to 22.6% for the static-prior variant. This indicates that the data-driven prior helps the sampler to find not just locally valid atomic environments (atom stability) but globally coherent molecular structures (validity and connectivity).

We ablate sampler choices for our full RFM model used throughout this paper, evaluating on GEOM-Drugs with $n = 200$ samples. We follow the same incremental hyperparameter search as before: sweep step size for FWDE, lock it in and additionally sweep σ for ALD. For PT we use the hyperparameters from Table 14.

Results are reported in Table 13. The results show that noise is critical for sampling quality, as FWDE ($\sigma = 0$) achieves reasonable atom stability but near-zero molecule stability and the highest median energy, indicating that deterministic trajectories get trapped in shallow local minima. ALD improves substantially across all metrics, confirming that stochastic exploration helps escape these shallow basins. However, ALD tends to produce locally valid fragments rather than globally coherent molecules, as evidenced by the gap between its atom stability and Valid & Connected rate. Both PT variants close this gap, indicating that the temperature ladder successfully enables basin-to-basin transitions. A surprising observation is the significant difference between the two PT variants, as Atom Max nearly doubles molecule stability compared to Atom Avg at matched NFE (24.5% vs. 15.5% at 1080 steps). We additionally test inference-time compute scaling by increasing the NFE

Table 12: Ablation of training objectives on GEOM-Drugs using a reduced-size model. ALD: Annealed Langevin Dynamics; PT: Parallel Tempering; FWDE: Forward Euler. No $b(t)$ omits smoothing and static prior uses a single fixed prior during training. Best overall per metric in **bold**, best within each block underlined.

Method	Sampler	Atom Stability ↑	Valid & Connected ↑	W2 Dist. ($\times 10^{-2}$) ↓	Diverged Samples ↓
FM-based baselines					
Standard FM	FWDE	40.4	0.00	5.3	0
Standard FM	ALD	40.3	0.00	5.2	0
Standard FM	PT	40.1	0.00	5.3	0
Equivariant OT-FM	FWDE	60.5	1.6	3.3	40
Equivariant OT-FM	ALD	64.6	2.2	0.6	51
Equivariant OT-FM	PT	53.7	14.9	0.8	1074
RFM ablations					
RFM, no $b(t)$, static prior	FWDE	76.1	9.4	0.8	0
RFM, no $b(t)$, static prior	ALD	79.7	16.8	0.6	0
RFM, no $b(t)$, static prior	PT	**85.2**	22.4	0.4	0
RFM, static prior	FWDE	78.6	15.8	0.4	0
RFM, static prior	ALD	79.1	13.0	0.4	0
RFM, static prior	PT	81.6	22.6	0.5	0
RFM	FWDE	79.6	23.6	0.4	0
RFM	ALD	81.6	33.0	0.3	0
RFM	PT	83.6	**37.7**	**0.2**	0

budget from 1080 to 1960 steps. All samplers improve, with PT + Atom Max showing the largest gains, reaching 40.0% molecule stability.

Table 13: Ablation of sampling algorithms for our RFM-trained model on GEOM-Drugs ($n = 200$ samples). ALD: Annealed Langevin Dynamics; PT: Parallel Tempering; FWDE: Forward Euler.

Algorithm	(Avg.) Steps	Atom Stab. (↑)	Molecule Stab. (↑)	Valid & Connected(↑)	Diversity (↑)	Median Energy (↓)
FWDE	1080	88.2	0.5	48.0	129.8	21.7
FWDE	1960	88.1	0.5	50.0	130.0	20.8
ALD	1080	93.0	18.0	61.0	110.0	7.7
ALD	1960	95.5	23.5	70.5	108.9	4.6
PT + Atom Avg.	1080	92.9	15.5	78.0	116.4	3.8
PT + Atom Avg.	1960	95.5	22.0	82.5	113.0	1.6
PT + Atom Max	1080	95.6	24.5	83.0	114.5	2.5
PT + Atom Max	1960	97.3	40.0	89.5	108.1	1.3

H Hyperparameters and experiment details

Implementation. We implement EBMol in PyTorch using PyTorch Geometric [61] for their efficient batched graph operations. SE(3) and permutation alignment uses the efficient CPU-based optimal-transport solver of [37, 36]. We use exponential moving average (EMA) for evaluation and sampling. Dataset splits for QM9 and GEOM-Drugs follow [32].

Compute. The QM9 model was trained for approximately 4 days on a single NVIDIA RTX 3090 Ti. The GEOM-Drugs model was trained for approximately 3 days on 2 $\times$ NVIDIA L40.

Table 14: Full hyperparameters.

Parameter	QM9	GEOM-Drugs
Network		
EGNN layers	9	4
Embedding dimension	256	128
Energy MLP layers	2	2
Energy MLP hidden dim	256	128
Total parameters	5.2M	2.2M
Training		
Optimization		
Optimizer	Adam	Adam
Learning rate	5e-5	5e-5
LR schedule	Linear	Linear
Batch size	64	16
Training epochs	750	7
EMA decay	0.999	0.999
Gradient clipping	Yes	Yes
RFM loss		
Smoothing function $b(s)$	$\tanh(\gamma s)$	$\tanh(\gamma s)$
Smoothing sharpness γ	25.0	25.0
Energy regularization λ_{reg}	1e-3	1e-3
Prior		
Position prior	Anisotropic Gaussian (PCA-matched)	
Atom-type prior	$\text{Dir}(1/K, \ldots, 1/K)$	
Sampling		
Mirror-Langevin Algorithm		
Step size η	0.05	0.1
Coordinate noise $\sigma_{\mathbf{c}}$	0.6	0.2
Atom-type noise $\sigma_{\mathbf{p}}$	0.6	0.4
Simplex floor ϵ	0.0015	0.0005
Parallel tempering		
Number of chains M	8	8
Temperature levels	11	11
Min/Max Temperature	0.05/1.0	0.05/1.0
Steps between swaps	10	10
Swaps between harvesting	[8, 12, 16]	[8, 16 ,32, 64]
Swap criterion	Atom Max	Atom Max
Relaxation steps ($\tau = 0$)	50	200
Relaxation step size	0.01	0.01

Simplex boundary stabilization. The Mirror Langevin Algorithm requires the mirror potential Φ to satisfy a self-concordance-like condition (A1) of [12]: there exists $\kappa \geq 0$ such that

$$\sqrt{2}\left\|[\nabla^2\Phi(\mathbf{x})]^{1/2} - [\nabla^2\Phi(\mathbf{x}')]^{1/2}\right\|_F \leq \kappa \left\|\nabla\Phi(\mathbf{x}) - \nabla\Phi(\mathbf{x}')\right\|_2. \tag{34}$$

For the negative-entropy potential $\Phi^{\mathbf{p}}(\mathbf{p}) = \sum_i p_i \log p_i$, the Hessian is $\nabla^2\Phi^{\mathbf{p}} = \text{diag}(1/p_i)$ and its square root is $[\nabla^2\Phi^{\mathbf{p}}]^{1/2} = \text{diag}(1/\sqrt{p_i})$. As $p_i \to 0$, the left-hand side of (A1) grows as $p_i^{-1/2}$ while the right-hand side grows as $|\log p_i|$, so $\kappa = \infty$ on the full simplex interior, consistent with Table 1 of [12].

By clamping $\mathbf{p} \geq \epsilon$ before each update, we restrict the domain to $\Delta_\epsilon^{K-1} = \{\mathbf{p} \in \Delta^{K-1} : p_i \geq \epsilon\ \forall i\}$, a compact subset of the simplex interior on which κ is finite. The choice of ϵ involves a tradeoff: too small and the logits $\log p_i$ approach numerical limits, causing instability in the dual-space update; too large and the floor acts as a persistent noise source that prevents atom types from fully committing to a vertex, degrading sample quality. In effect, ϵ controls the minimum categorical noise injected at

each step via the $1/\sqrt{p_i}$ scaling, establishing a noise floor below which the sampler cannot refine atom-type assignments. This makes ϵ a dataset-dependent hyperparameter that requires tuning: datasets with fewer atom types tolerate smaller values, while larger type vocabularies need careful balancing to avoid both numerical instability and premature convergence. We report the values used for each benchmark in Table 14.

I Generated samples

Figure 7 shows random samples for GEOM-Drugs using shape-steering, while Figure 8 shows unconditional samples from the same trained EBMol.

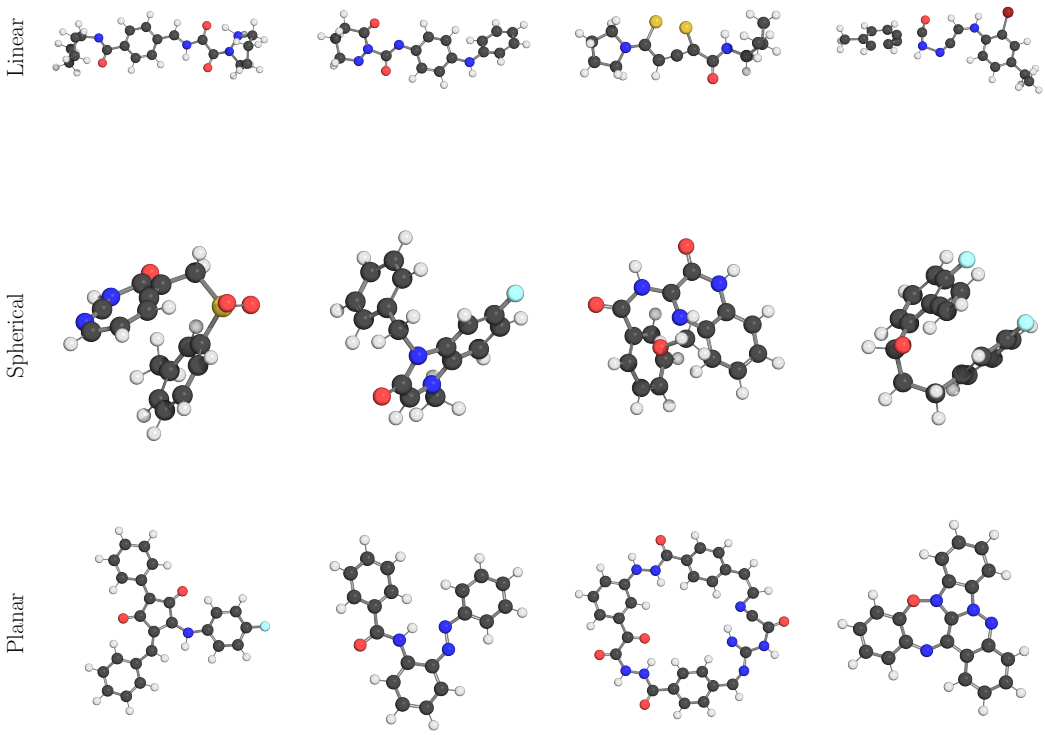


Figure 7: Random samples from an *EBMol* model trained on GEOM-Drugs using potential composition to steer the molecule shape. Samples are generated with a compute budget of 1960 NFEs.

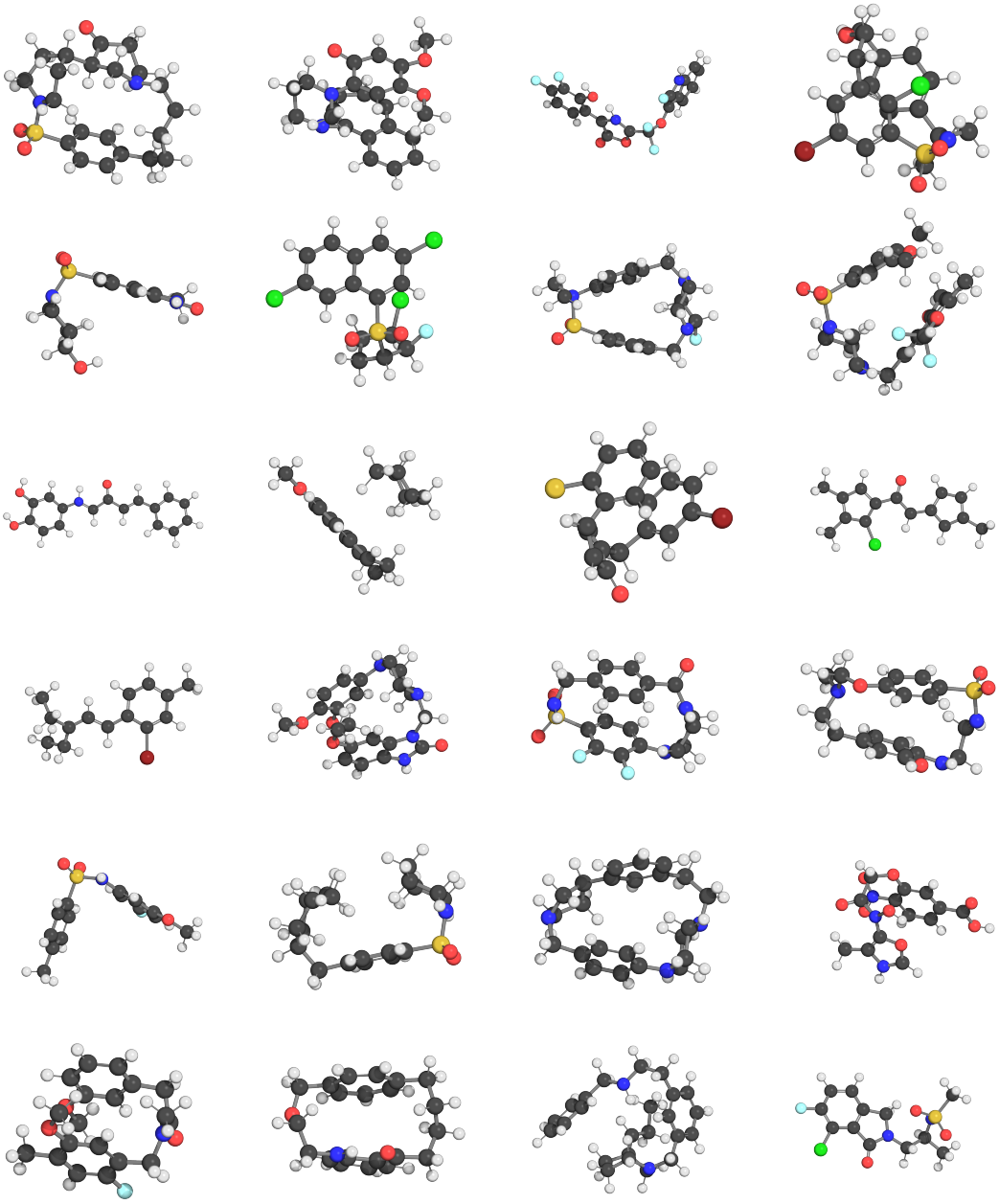


Figure 8: Random samples from an *EBMol* model trained on GEOM-Drugs. Samples are generated with a compute budget of 1960 NFEs.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer:[Yes]

   Justification: The abstract and introduction clearly state the three main contributions, all of which match empirical results 4 and follow formal proofs A.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Key limitations are discussed in the limitations section 5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Comprehensive proofs including assumptions and their discussion is provided in Appendix A.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We present pseudocode for the introduced training 1 and sampling 2 algorithms, details about the experiment setup 4.1, and all relevant hyperparameters with additional information such as compute and utilized packages in Appendix H.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The codebase to reproduce the main results and samples generated by our method will be provided upon acceptance.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Experiment details listed in Appendix H.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Following the experimental protocol of the baseline methods, we do not report confidence intervals for the standard generation metrics (Tables 1 and 2). As these are computed for 10,000 samples, the variance is negligible. For the structural and energy metrics in Table 3, we report the mean $\pm$ standard deviation. Figures in the Appendix include shaded uncertainty regions.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Computational requirements are listed in Appendix H and include the type of hardware used and the training duration.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This paper adheres the code of ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This work investigates the applicability of energy-based generative modeling frameworks to 3D molecular generation. The contributions are methodological and evaluated on public benchmarks. Thus, the results do not have immediate direct societal impact beyond advancing foundational understanding of generative modeling for molecules.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: No such measures are necessary.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Code fragments, utilized packages, models and other methods are properly cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper does not release any new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No such experiments are conducted.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No such experiments are conducted.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not part of our methodology.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.