BIT: BOLTZMANN-INSPIRED TRAINING FOR ONE-STEP MOLECULAR GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Molecular forces provide rich local information about the Boltzmann distribution. To harness this information for direct one-step molecular generation, we introduce Boltzmann-Inspired Training (BIT), an attraction–repulsion framework with adaptive training targets. BIT projects reference forces onto atom-pair directions and uses their alignment with differences in pairwise distances between generated and reference conformations to reweight the local attraction. It combines the resulting target with a data-only target using a weight that varies across generated conformations, and trains the generator by regression. On seven MD17 molecules, BIT reduces mean pooled-distance total variation distance (TVD) by 55% and mean atom-pair TVD by 24% relative to the same generator trained without forces. On average across molecules, BIT also achieves lower pooled-distance TVD and distance-histogram mean absolute error (MAE) than AvgFlow and ET-Flow. At inference, the trained generator produces each conformation with a single network evaluation, while force information is used only during training.

1 INTRODUCTION

A molecular conformation generator should capture not only the structures a molecule can adopt, but also how frequently those structures occur. We study these distributions for individual molecules using molecular dynamics (MD) trajectories, where each reference conformation is accompanied by a physical force describing the local variation of the energy landscape. For one-step generation, the model must learn a noise-to-coordinate mapping that reproduces the conformational distribution in a single network evaluation. Can the force information available during training improve this mapping without sacrificing one-step sampling?

For a Boltzmann density $p_B(x) \propto \exp[-E(x)/(k_BT)]$, where E is the potential energy and k_BT is the thermal energy, the physical force $F(x) = -\nabla_x E(x)$ determines the score:

$$\nabla_x \log p_B(x) = \frac{F(x)}{k_BT}. \tag{1}$$

Thus, forces complement the observed conformations with local information about the target density. This connection has motivated force-supervised score models and energy-informed generative methods (Guo et al., 2025; De Bortoli et al., 2024; Akhound-Sadegh et al., 2024). Here, we ask how the same information can be incorporated into the training targets of a coordinate generator.

We introduce **Boltzmann-Inspired Training (BIT)**, a force-informed extension of attraction–repulsion training (Deng et al., 2026). This objective attracts generated samples toward reference conformations and repels them from one another. Geometric proximity determines the data-only attraction weights, although nearby references may provide different local force information. BIT projects reference forces onto atom-pair directions and uses their alignment with differences in pairwise distances between generated and reference conformations to reweight the attraction, retaining the same repulsive component. A statistic comparing local force and geometry information then controls how the data-only and force-informed targets are combined for each generated conformation. The combined target is learned by regression. All force-dependent computation is confined to training; the trained generator maps noise to molecular coordinates in one network evaluation.

On seven MD17 molecules, BIT reduces mean pooled-distance total variation distance (TVD) by 55% and mean atom-pair TVD by 24% relative to the same generator trained without forces. It also

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

achieves lower mean pooled-distance TVD and histogram MAE than locally trained AvgFlow and ET-Flow. We further evaluate joint distance distributions and geometric state occupancy. Additional controls test the importance of pairing projected forces with their reference conformations and allowing mixing weights to vary across generated samples.

Contributions. Our main contributions are:

1. **Reference-force reweighting.** We incorporate projected reference forces into the attraction targets of a direct molecular generator.
2. **Sample-dependent target combination.** A local statistic adjusts the weight of the force-informed correction for each generated conformation.
3. **Improved distance-distribution matching with one-step sampling.** Experiments on seven MD17 molecules show improved distance-distribution matching over the force-free counterpart, with supporting evaluations of joint distances and geometric state occupancy.

2 Related Work

Force- and energy-informed molecular generation. Score-based generative models learn gradients of log densities and sample through iterative updates (Song et al., 2021; Ho et al., 2020). In molecular settings, Potential Score Matching uses forces associated with reference conformations to supervise molecular score models (Guo et al., 2025), while Target Score Matching and Iterated Denoising Energy Matching incorporate target-score or energy information into diffusion-based training (De Bortoli et al., 2024; Akhound-Sadegh et al., 2024). Other approaches, including Boltzmann generators and Flow Annealed Importance Sampling Bootstrap (FAB), use energy information to train normalizing flows for direct sampling (Noé et al., 2019; Midgley et al., 2023). BIT uses recorded reference forces to construct adaptive attraction targets for training a one-step coordinate generator by regression.

One-step generation and attraction–repulsion training. Distillation, consistency models, rectified flows, and mean flows reduce generation to one or a few network evaluations (Salimans and Ho, 2022; Song et al., 2023; Liu et al., 2023; Geng et al., 2025). For molecular generation, ET-Flow learns an equivariant flow evaluated over multiple steps (Hassan et al., 2024), whereas AvgFlow combines reflow and distillation to obtain a one-step sampler (Cao et al., 2025). More recently, drifting objectives train one-step generators by attracting generated samples toward data while repelling generated samples from one another (Deng et al., 2026); with Gaussian kernels, the resulting field is related to score differences between smoothed distributions (Turan et al., 2026). BIT builds on this attraction–repulsion perspective, but augments the attraction term with reference-force information and combines data-only and force-informed targets in a sample-dependent manner. Kernel score identities motivate the local statistic used for this combination (Liu et al., 2016; Liu and Wang, 2016; Gorham and Mackey, 2017).

Molecular conformation distributions. Molecular conformation generation has also been studied through a range of geometric representations and sampling procedures, including GeoMol, ConfGF, GeoDiff, and Torsional Diffusion (Ganea et al., 2021; Shi et al., 2021; Xu et al., 2022; Jing et al., 2022). Our setting focuses on matching conformational distributions represented by MD trajectories. We therefore evaluate atom-pair distance marginals, joint distance structure, and geometric state occupancy.

3 Boltzmann-Inspired Training

3.1 Direct generation and data-only attraction–repulsion

The training data provide physical reference coordinates $Y_j \in \mathbb{R}^{3n}$ and forces F_j for a molecule with n atoms. Throughout this section, coordinates are centered and normalized as

$$y_j = \frac{Y_j - \langle Y_j \rangle}{\sigma_{\text{coord}}},$$

where $\langle\cdot\rangle$ denotes the mean over atoms and σ_{coord} is the training-coordinate scale. The corresponding scaled force and score are $\widetilde{F}(y_j) = \sigma_{\mathrm{coord}}(F_j - \langle F_j\rangle)$ and $s(y_j) = \widetilde{F}(y_j)/(k_{\mathrm{B}}T)$, following Eq. (1). Appendix A specifies the scaling and units. The generator directly produces normalized coordinates $x = G_\theta(\xi, m)$, where $\xi \sim \mathcal{N}(0, I)$ and m contains atom-, graph-, and molecule-level features.

We construct the training targets in pairwise-distance space. Let $\phi(x) \in \mathbb{R}^P$ list all $P = n(n-1)/2$ unordered interatomic distances in a fixed order. This representation is invariant to rigid translations and rotations. For a reference conformation y_j, define $\delta_j(x) = \phi(y_j) - \phi(x)$. With bandwidth $\tau > 0$, the data-only attraction weights are

$$w_j(x) = \mathrm{softmax}_j\left[-\frac{\|\delta_j(x)\|^2}{2\tau^2}\right]. \tag{2}$$

The softmax is normalized over a shared reference batch, so references closer to the generated query in distance space receive larger attraction weights.

We additionally repel each generated sample from nearby samples in the current generated batch. For the generated batch $\{x_k^-\}$, let

$$w_k^-(x) \propto \exp\left[-\frac{\|\phi(x_k^-) - \phi(x)\|^2}{2\tau^2}\right],$$

with the weights normalized to sum to one over k, and define

$$\begin{aligned} \Delta_\phi^-(x) &= \sum_k w_k^-(x)\big[\phi(x_k^-) - \phi(x)\big], \\ t_0(x) &= \phi(x) + \sum_j w_j(x)\delta_j(x) - \Delta_\phi^-(x). \end{aligned} \tag{3}$$

The generated batch includes x itself: its displacement contribution is zero, although its weight still enters the normalization. Thus, t_0 attracts each query toward nearby reference conformations while repelling it from nearby generated samples.

3.2 Force-informed reference reweighting

The data-only weights distinguish reference conformations through geometric proximity alone. To incorporate local force information, we project the scaled reference force through the Jacobian $J_\phi(y)$ of the distance map:

$$\widetilde{P}(y) = J_\phi(y)\widetilde{F}(y), \qquad \widetilde{P}_{ab}(y) = \frac{y_a - y_b}{\|y_a - y_b\|} \cdot \big(\widetilde{F}_a(y) - \widetilde{F}_b(y)\big). \tag{4}$$

This quantity describes the first-order change of the distance representation under a displacement along $\widetilde{F}(y)$; numerical distance floors are specified in Appendix A.

For a generated query x and reference y_j, define the force–geometry alignment

$$a_j(x) = \widetilde{P}(y_j)^\top \delta_j(x).$$

It measures the alignment between the projected reference force and the difference between reference and generated distance vectors, incorporating both direction and magnitude. Consequently, geometrically nearby references can still receive different contributions when their projected-force information differs.

To reduce sensitivity to the overall scale of the alignments, we normalize them by their sample standard deviation across the reference batch,

$$v(x) = \max\left\{\mathrm{sd}_j[a_j(x)],\, 10^{-8}\right\}.$$

With force-alignment coefficient $g \geq 0$, the force-informed attraction weights are

$$w_j^F(x) = \mathrm{softmax}_j\left[-\frac{\|\delta_j(x)\|^2}{2\tau^2} + g\frac{a_j(x)}{v(x)}\right]. \tag{5}$$

Both the data-only and force-informed branches normalize over the same reference conformations. Keeping the generated-sample repulsion unchanged gives

$$t_F(x) = \phi(x) + \sum_j w_j^F(x)\delta_j(x) - \Delta_\phi^-(x). \tag{6}$$

Hence,

$$t_F(x) - t_0(x) = \sum_j \big(w_j^F(x) - w_j(x)\big)\delta_j(x),$$

so force information changes only how individual references contribute to the attraction, while leaving the repulsive component unchanged.

Figure 1 summarizes target construction and one-step sampling.

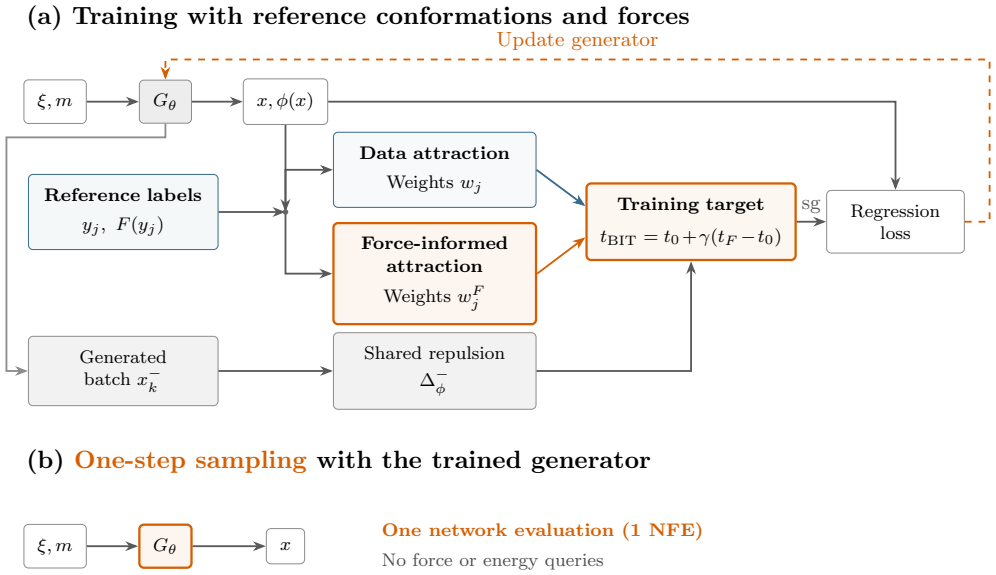

Figure 1: BIT training and one-step sampling. Data-only and force-informed attraction share the same generated-sample repulsion. A local score–geometry statistic sets their sample-dependent mixture, producing a regression target held fixed during backpropagation. At sampling time, noise and molecular features pass through the trained generator once to produce a conformation.

3.3 Adaptive combination using a local score–geometry statistic

We next allow the contribution of the force-informed correction to vary across generated conformations. Define the distance-space kernel

$$k_\tau(x, y) = \exp\left[-\frac{\|\phi(y) - \phi(x)\|^2}{2\tau^2}\right],$$

and let p_{B} denote the target density in the normalized coordinate convention. For each fixed query x, under the regularity and boundary conditions given in Appendix B, integration by parts yields

$$\tau^2 \mathbb{E}_{y\sim p_{\mathrm{B}}}\left[k_\tau(x, y)s(y)\right] = \mathbb{E}_{y\sim p_{\mathrm{B}}}\left[k_\tau(x, y)J_\phi(y)^\top\big(\phi(y) - \phi(x)\big)\right]. \tag{7}$$

This identity relates a kernel-weighted score term to a geometric term pulled back from pairwise-distance space to coordinate space. Importantly, the Jacobian on the right-hand side is evaluated at each reference y, whereas regression later backpropagates through $J_\phi(x)$ at the generated query.

We use this population identity to construct a finite-batch statistic around each generated sample. With the normalized data-only weights $w_j(x)$, define the empirical residual

$$r(x) = \sum_j w_j(x) \left[\tau^2 s(y_j) - J_\phi(y_j)^\top \delta_j(x)\right], \qquad D(x) = \frac{\left[\|r(x)\|^2 - \widehat{\sigma}^2(x)\right]_+}{S^2(x) + \varepsilon}. \tag{8}$$

Here $[u]_+ = \max\{u, 0\}$, $\varepsilon = 10^{-8}$, $\widehat{\sigma}^2(x)$ is a finite-reference variance correction, and $S^2(x)$ sets the local scale. Their full expressions are given in Appendix A. The residual compares locally weighted force-derived score and geometric pullback terms estimated from the reference batch. To obtain $D(x)$, we subtract the variance correction from the squared residual, clip at zero, and normalize by the local scale. This statistic controls the target mixture. For $c > 0$, define

$$t_{\mathrm{BIT}}(x) = t_0(x) + \gamma(x)\big(t_F(x) - t_0(x)\big), \qquad \gamma(x) = \frac{D(x)}{D(x) + c}. \tag{9}$$

When $D(x) \ll c$, the combined target remains close to the data-only target t_0; when $D(x) \gg c$, it places greater weight on the force-informed target t_F. Appendix B.3 derives this mixing rule by minimizing conditional squared error under additional assumptions on the errors of the two targets.

3.4 Regression training and one-step sampling

For a generated batch $\mathcal{B}$, we regress the distance representation toward the combined target:

$$\mathcal{L}(\theta) = \frac{1}{|\mathcal{B}|} \sum_{x \in \mathcal{B}} \|\phi(x) - \mathrm{sg}[t_{\mathrm{BIT}}(x)]\|_2^2 . \tag{10}$$

The stop-gradient operator sg holds the entire target fixed during backpropagation, including the attraction weights, force projections, residual statistics, and mixing coefficients. For a single summand $\mathcal{L}_x$, the gradient with respect to the generated coordinates is

$$\nabla_x \mathcal{L}_x = 2 J_\phi(x)^\top \big(\phi(x) - t_{\mathrm{BIT}}(x)\big). \tag{11}$$

The distance-space residual is therefore pulled back to Cartesian coordinates and then propagated through G_θ to update the generator parameters. The target distance vector guides training but need not correspond to a valid three-dimensional conformation; the generator itself always outputs Cartesian coordinates.

All force-dependent computation is confined to training. At sampling time, we draw $\xi \sim \mathcal{N}(0, I)$ and return $\sigma_{\mathrm{coord}} G_\theta(\xi, m)$ using the trained generator and fixed molecular conditions. Each conformation therefore requires a single network evaluation.

4 Experiments

4.1 Experimental setup

We evaluate seven MD17 molecules (Chmiela et al., 2017). Each locally trained method uses the first 5,000 frames, batch size 256, 200k optimizer updates, and three training seeds. BIT, No-force, and Ungated share the generator, Adam optimizer, and cosine learning-rate schedule, with 512 reference conformations sampled with replacement for each generated batch. No-force uses t_0, Ungated uses t_F throughout training, and BIT combines them by Eq. (9). BIT and Ungated are tuned independently using training-frame TVD (Appendix A.5). AvgFlow and ET-Flow use the same frames, batch size, and update count; their different architectures mean that total training compute is not matched.

Each final checkpoint generates 1,000 conformations. Primary evaluation uses the 50,000 reference frames in $[5000, 55000)$. Pooled-distance TVD compares histograms of all unordered atom-pair distances after excluding values outside the reference range and renormalizing the counts; Appendix D.2 reports checks retaining this mass. Atom-pair TVD averages separate histogram errors across fixed pairs. Results for learned methods show means and sample SDs across training seeds, with equal weighting of molecules in aggregate. Appendix A gives the architecture and training details; Appendices D–G describe the evaluation protocols.

4.2 DISTANCE-DISTRIBUTION MATCHING

Comparing training targets. BIT reduces the seven-molecule mean pooled-distance TVD from 0.1056 for No-force to 0.0472 (55%), with independently tuned Ungated at 0.0800 (Table 1). BIT has lower TVD than both controls for every molecule and training seed.

Table 1: Training-target comparison on MD17 (TVD ↓, mean ± sample SD over three training seeds). All columns share architecture, data, and schedule. Independent tuning selects the same g for BIT and Ungated, which share the force-reweighting formulation; Ungated uses t_F, and No-force uses t_0. The Mean row averages molecules within each seed before computing SD. Bold marks the lowest mean per row.

Molecule	No-force	Ungated	BIT
ethanol	0.0754 ± 0.0046	0.0898 ± 0.0018	$\mathbf{0.0492} \pm 0.0008$
malonaldehyde	0.1216 ± 0.0035	0.0798 ± 0.0020	$\mathbf{0.0645} \pm 0.0035$
uracil	0.0662 ± 0.0031	0.0586 ± 0.0031	$\mathbf{0.0434} \pm 0.0019$
toluene	0.2197 ± 0.0046	0.1140 ± 0.0007	$\mathbf{0.0521} \pm 0.0015$
salicylic acid	0.0589 ± 0.0010	0.0459 ± 0.0030	$\mathbf{0.0250} \pm 0.0015$
naphthalene	0.1027 ± 0.0171	0.0989 ± 0.0008	$\mathbf{0.0599} \pm 0.0014$
aspirin	0.0948 ± 0.0044	0.0732 ± 0.0022	$\mathbf{0.0358} \pm 0.0007$
Mean	0.1056 ± 0.0031	0.0800 ± 0.0014	$\mathbf{0.0472} \pm 0.0009$

Comparison with generative baselines. BIT achieves mean pooled-distance TVD of 0.0472, compared with 0.0709 for AvgFlow and 0.0752 for ET-Flow (Table 2), with lower error on five and six of seven molecules, respectively. Mean histogram MAE is likewise lower: 0.0698 for BIT, 0.0949 for AvgFlow, and 0.1037 for ET-Flow (Appendix Table 3). Published DSM and PSM results retain their original evaluation protocols and are reported separately in Appendix Table 4.

Table 2: Pooled-distance TVD on MD17 (↓). The local methods use the same 5,000 training frames per molecule, training batch size 256, and 200k updates. Values are mean ± sample SD over three training seeds; the Mean row first averages molecules within each seed. Bold marks the lowest mean per row.

Molecule NFE	BIT 1	AvgFlow 1	ET-Flow 50
ethanol	$\mathbf{0.0492} \pm 0.0008$	0.0906 ± 0.0057	0.1155 ± 0.0026
malonaldehyde	0.0645 ± 0.0035	$\mathbf{0.0544} \pm 0.0034$	0.0690 ± 0.0003
uracil	$\mathbf{0.0434} \pm 0.0019$	0.0775 ± 0.0011	0.0553 ± 0.0012
toluene	$\mathbf{0.0521} \pm 0.0015$	0.0927 ± 0.0041	0.0910 ± 0.0048
salicylic acid	$\mathbf{0.0250} \pm 0.0015$	0.0727 ± 0.0015	0.0648 ± 0.0033
naphthalene	$\mathbf{0.0599} \pm 0.0014$	0.0794 ± 0.0022	0.0971 ± 0.0014
aspirin	0.0358 ± 0.0007	$\mathbf{0.0287} \pm 0.0004$	0.0335 ± 0.0003
Mean	$\mathbf{0.0472} \pm 0.0009$	0.0709 ± 0.0020	0.0752 ± 0.0004

Agreement at the atom-pair level. Mean atom-pair TVD decreases from 0.2372 for No-force to 0.1809 for BIT, with improvements on all seven molecules. The mean ordering BIT $<$ Ungated $<$ No-force persists across 64, 128, and 256 bins per pair and for pooled TVD with out-of-range mass retained (Appendices D.4 and D.2). Figure 2 illustrates closer agreement with the toluene and ethanol reference profiles.

Training resampling also gives a low mean pooled-distance TVD (0.0412; Appendix C.5). We therefore examine joint distances and geometric state occupancy as well.

4.3 BEYOND MARGINAL DISTANCE MATCHING

Geometric state occupancy. We fit PCA on training distance vectors (ten components per molecule; retained variances in Appendix F.2) and construct a shared $K = 10$ partition using K-means on 3,000 non-training reference frames. Generated samples participate in neither fit.

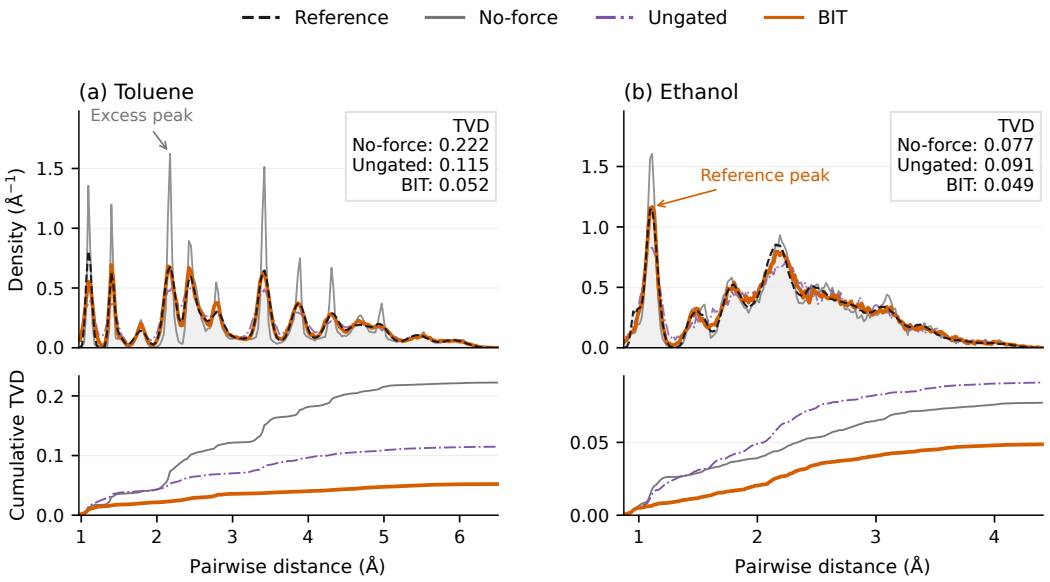

Figure 2: Generated distance distributions for toluene and ethanol at training seed 42. Top: pooled-distance densities, with the MD reference shown in black. Bottom: cumulative TVD contributions $C_b = \frac{1}{2}\sum_{\ell \le b} |p_\ell - q_\ell|$, whose final value equals TVD. The corresponding three-seed results appear in Table 1.

BIT has lower occupancy TVD than No-force on six of seven molecules and than AvgFlow on five of seven (Figure 3). The seven-molecule mean decreases from 0.2544 for No-force to 0.2424 for BIT (4.7%). BIT also improves over training-set resampling on four of seven molecules. Results at $K = 5$ and $K = 20$ show a similar overall pattern (Appendix F). These clusters define geometric partitions, not kinetic metastable states or free-energy basins.

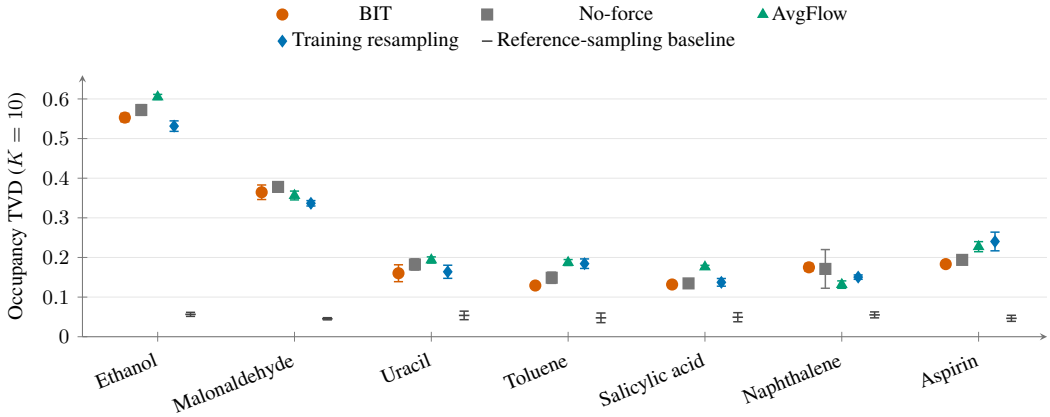

Figure 3: Geometric state-occupancy TVD at $K = 10$ (↓). Points show means; error bars show sample SD across three training seeds for BIT, No-force, and AvgFlow, and across three sampling repeats for training resampling and the reference-sampling baseline. Each run first averages comparisons with three common reference sets. All methods share the fixed partition. Full values and sensitivity summaries appear in Appendix F.

Joint distance structure. Using complete distance vectors and a Gaussian kernel with fixed bandwidth 0.25σ, where σ is a training-derived distance scale, BIT has lower three-seed mean MMD^2 than No-force on all seven molecules. Comparisons with AvgFlow and ET-Flow vary by molecule (Appendix E).

Proximity to the training set. To characterize proximity to training data, we compare nearest-neighbor distances with those of non-training reference frames. BIT has similar median distance scales on toluene, salicylic acid, and naphthalene, but is closer to the training set on ethanol and malonaldehyde. No evaluated BIT sample matches a training distance vector within 10^{-10} Å. Full distributions appear in Appendix F.3; these comparisons describe training-set proximity rather than establish conformational novelty.

4.4 Role of force information and adaptive combination

We test the roles of projected-force pairing and sample-dependent mixing.

Projected-force pairing. We shuffle projected forces among references only in the reweighting branch of Eq. (5); the residual retains its original score–conformation pairs, and the data-only branch is unchanged. Across ethanol, toluene, and aspirin, mean pooled-distance TVD increases from 0.0457 for BIT to 0.1375 under shuffling. Correct pairing yields lower TVD for every tested molecule and training seed (Appendix Table 5).

Sample-dependent mixing. Batch-mean replaces each sample-dependent mixing coefficient with its batch average, preserving the mean coefficient while removing within-batch variation. On naphthalene, ethanol, and aspirin, restoring sample-dependent mixing reduces mean pooled-distance TVD by 8.40%, 5.99%, and 2.22%, respectively (Appendix Table 6).

4.5 Sampling efficiency

BIT generates each conformation with one network evaluation and no force or energy queries. On one RTX 5090 with sampling batch size 256, generating 1,000 conformations takes 7.14 ms for BIT, 48.21 ms for one-step AvgFlow, and 6.62 s for 50-NFE ET-Flow (Figure 4). These measurements characterize end-to-end latency under the respective architectures and software stacks; Appendix G provides the timing protocol and per-molecule results.

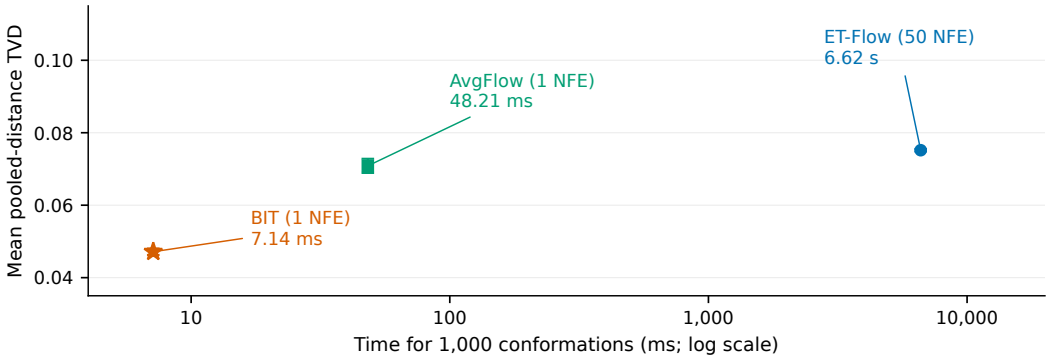

Figure 4: Distance-distribution accuracy and sampling time on MD17. The horizontal axis shows the mean of the seven per-molecule median times for generating 1,000 conformations. The vertical axis shows the seven-molecule mean pooled-distance TVD from Table 2; error bars give the sample SD of the three training-seed means. NFE denotes network evaluations per sample.

5 DISCUSSION

The results suggest that reference-force reweighting and sample-dependent combination make complementary contributions to BIT. Improvements over Ungated and Batch-mean support adaptive mixing, while the projected-force pairing control highlights the importance of associating local force information with its reference conformation. Together, these observations help explain the gains on the evaluated molecules.

The Boltzmann force–score relation motivates using reference forces to modify local attraction. The kernel identity in Eq. (7) motivates the score–geometry statistic, and the conditional-risk analysis in Appendix B.3 explains the mixing rule under additional assumptions on target errors. These connections motivate the training targets without guaranteeing exact sampling from the Boltzmann distribution.

Our experiments assess static conformation distributions on seven MD17 molecules. Pairwise distances cannot distinguish mirror images, and the mixing statistic can depend on the relative orientations of reference conformations (Appendix A.2). Molecular kinetics lies outside the present task. Further studies on larger and more flexible molecules could explore representations that distinguish mirror-image conformations and evaluate metastable-state populations and energy or free-energy statistics.

6 CONCLUSION

BIT incorporates reference forces into attraction targets and adaptively combines force-informed and data-only targets. Across seven MD17 molecules, it improves distance-distribution matching over its force-free counterpart, with supporting evidence from evaluations of joint distances and geometric state occupancy. Under the matched local-training protocol, BIT also achieves lower mean pooled-distance TVD and histogram MAE than AvgFlow and ET-Flow. The trained generator produces each conformation in one network evaluation without force or energy queries.

REPRODUCIBILITY STATEMENT

The appendices specify the training algorithm, architecture, coordinate conventions, data splits, hyperparameter-selection procedure, distance metrics, occupancy and nearest-neighbor protocols, and timing measurements. We will release the core BIT implementation, including single-molecule MD17 training, one-step sampling, fixed experimental configurations, and distance-distribution evaluation code.

AI USE STATEMENT

We used large language models to assist with literature discovery and to edit and polish the manuscript. The authors checked the cited literature against the original sources and reviewed and revised the AI-assisted edits. All final research and editorial decisions were made by the authors, who take full responsibility for the content, accuracy, originality, and integrity of this work.

REFERENCES

Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, B. Poole. Score-based generative modeling through stochastic differential equations. ICLR 2021.

J. Ho, A. Jain, P. Abbeel. Denoising diffusion probabilistic models. NeurIPS 2020.

L. Guo, Z. Wang, C. Liu, J. Li, P. Hu, Y. Zhu, T. Qin. Potential score matching: debiasing molecular structure sampling with potential energy guidance. Transactions on Machine Learning Research, 2025; arXiv:2503.14569.

T. Akhound-Sadegh et al. Iterated denoising energy matching for sampling from Boltzmann densities. ICML 2024.

V. De Bortoli, M. Hutchinson, P. Wirnsberger, A. Doucet. Target score matching. arXiv:2402.08667, 2024.

L. Midgley, V. Stimper, G. N. C. Simm, B. Schölkopf, J. M. Hernández-Lobato. Flow annealed importance sampling bootstrap. ICLR 2023.

T. Salimans, J. Ho. Progressive distillation for fast sampling of diffusion models. ICLR 2022.

Y. Song, P. Dhariwal, M. Chen, I. Sutskever. Consistency models. ICML 2023.

X. Liu, C. Gong, Q. Liu. Flow straight and fast: learning to generate and transfer data with rectified flow. ICLR 2023.

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

M. Hassan et al. ET-Flow: equivariant flow matching for molecular conformer generation. NeurIPS 2024.

M. Deng, H. Li, T. Li, Y. Du, K. He. Generative modeling via drifting. arXiv:2602.04770, 2026.

E. Turan, N. Dufour, M. Ovsjanikov. Generative drifting is secretly score matching: a spectral and variational perspective. arXiv:2603.09936v2, 2026.

F. Noé, S. Olsson, J. Köhler, H. Wu. Boltzmann generators: sampling equilibrium states of many-body systems with deep learning. Science, 365(6457):eaaw1147, 2019.

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

Z. Cao, M. Geiger, A. dos Santos Costa, D. Reidenbach, K. Kreis, T. Geffner, F. Pellegrini, G. Zhou, E. Kucukbenli. Efficient molecular conformer generation with SO(3)-averaged flow matching and reflow. ICML 2025; arXiv:2507.09785.

O. Ganea et al. GeoMol: torsional geometric generation of molecular 3D conformer ensembles. NeurIPS 2021.

C. Shi, S. Luo, M. Xu, J. Tang. Learning gradient fields for molecular conformation generation. ICML 2021.

M. Xu, L. Yu, Y. Song, C. Shi, S. Ermon, J. Tang. GeoDiff: a geometric diffusion model for molecular conformation generation. ICLR 2022.

B. Jing, G. Corso, J. Chang, R. Barzilay, T. Jaakkola. Torsional diffusion for molecular conformer generation. NeurIPS 2022.

Q. Liu, J. Lee, M. Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. ICML 2016.

Q. Liu, D. Wang. Stein variational gradient descent. NeurIPS 2016.

J. Gorham, L. Mackey. Measuring sample quality with kernels. ICML 2017.

A TRAINING AND IMPLEMENTATION DETAILS

Distance-metric definitions are collected in Appendix D.

A.1 BIT TRAINING ALGORITHM

Algorithm 1 BIT training step

Require: generator G_θ; reference conformations with scores s and distance-space force projections $\widetilde{P}$ in the coordinate convention below; bandwidth τ_t (constant for MD17); alignment coefficient g; mixing constant c

1: Sample latent noise and generate a batch $\mathcal{B}$; compute its distance features
2: Draw a shared reference batch $\{y_j\}$
3: **for** each generated query $x \in \mathcal{B}$ **do**
4: Base weights $w_j \propto \exp\bigl(-\|\phi(y_j) - \phi(x)\|^2/(2\tau_t^2)\bigr)$
5: Force-aligned weights w_j^F by Eq. (5); negative weights w_k^- from the generated batch
6: Form data-only t_0 and force-informed t_F using displacements and weights w, w^F
7: Compute residual $r(x)$ and variance-corrected statistic $D(x)$ by Eq. (8)
8: Mixing coefficient $\gamma(x) = D(x)/(D(x) + c)$; target $t_{\mathrm{BIT}} = t_0 + \gamma\,(t_F - t_0)$
9: **end for**
10: Compute the batch-averaged loss in Eq. (10); take a gradient step on θ

All target-side quantities (weights, forces, residual, mixing coefficient) are stop-gradient.

Adaptive combination statistic. Equation (8) defines the residual $r(x)$ and normalized statistic $D(x)$. Write $z_j = \tau^2 s(y_j) - J_\phi(y_j)^\top \delta_j$, so that $r(x) = \sum_j w_j z_j$. The local scale and variance correction are

$$S^2(x) = \tau^4 \sum_j w_j \|s(y_j)\|^2 + \sum_j w_j \bigl\| J_\phi(y_j)^\top (\phi(y_j) - \phi(x)) \bigr\|^2,$$

and

$$\widehat{\sigma}^2(x) = \frac{\sum_j w_j^2 \|z_j - r(x)\|^2}{\max\{1 - \sum_j w_j^2,\, 10^{-8}\}}, \qquad \varepsilon = 10^{-8}.$$

For independent, identically distributed z_j and uniform weights, the correction estimates the trace of the covariance of their sample mean. With the data-dependent weights used here, it serves as an empirical finite-reference correction. Appendix B gives the population identity motivating the residual.

A.2 COORDINATE AND FORCE CONVENTIONS

We use reference coordinates and forces in the Cartesian frame provided by the dataset, applying centering and scalar rescaling without additional rotational alignment.

Because $r(x)$ averages Cartesian vectors across references, independently rotating the references and their forces can change $D(x)$ and $\gamma(x)$. A common rotation leaves both unchanged, and the projected-force weights are invariant to either transformation.

Let Y and F denote a reference conformation and its forces in physical units. The scalar σ_{coord} is the standard deviation of all centered training-coordinate components. Consistent with Section 3, write $y = (Y - \langle Y \rangle)/\sigma_{\mathrm{coord}}$ and $\widetilde{F}(y) = \sigma_{\mathrm{coord}}(F - \langle F \rangle)$, where brackets denote arithmetic averages over atoms within each conformation. The score used by the mixing component is $s(y) = \widetilde{F}(y)/(k_{\mathrm{B}}T)$. Following the MD17 energy and coordinate convention of kcal/mol and Å (Guo et al., 2025), the implementation fixes the thermal-energy scale at 1 kcal/mol as a numerical approximation to the molar thermal energy of 0.9936 kcal/mol at 500 K. Scaling transforms the score to normalized coordinates, and centering projects forces onto the translation-free subspace. Force centering cancels in atom-pair differences, giving the projection below; its denominator is floored at 10^{-6} in normalized coordinates:

$$\widetilde{P}_{ab}(y) = \frac{y_a - y_b}{\|y_a - y_b\|} \cdot \sigma_{\mathrm{coord}}(F_a - F_b).$$

A.3 GENERATOR ARCHITECTURE AND CONDITIONING

The generator uses a transformer backbone with width 256, four blocks, eight attention heads per block, and feed-forward width 1,024. Centered coordinate noise enters through a $3 \to 256$ linear projection. After the transformer blocks, a linear head predicts a three-dimensional coordinate update for each atom. These updates are centered and added to the centered input noise to obtain the generated conformation.

Fixed graph and positional features. For each molecule, an undirected edge connects atom pairs whose mean distance over the first 100 training conformations is below 1.8 Å. Edge categories 2, 3, and 1 correspond to mean distances in $[0, 1.25)$, $[1.25, 1.5)$, and $[1.5, 1.8)$ Å, respectively; category 0 denotes nonedges and self-pairs. These categories encode distance ranges, not chemical bond orders. A learned 4×8 embedding maps categories to one additive attention bias per head, shared across all four transformer blocks. From the binary adjacency matrix, we construct the unnormalized graph Laplacian and take the eight eigenvectors following the first in ascending eigenvalue order, zero-padding if fewer are available. Their signs are randomized during training, and an $8 \to 256$ linear projection gives the positional features.

Local chemical-environment features. Separate 256-dimensional embedding tables encode the central atom type, neighboring atom types, and incident edge categories. For each atom, we concatenate its central embedding with the mean neighbor-type embedding and mean incident-edge embedding, then apply a $768 \to 256 \to 256$ MLP with a GELU between the linear layers. A second MLP, $512 \to 256 \to 256$ with GELU, takes the concatenation of this representation and its mean over neighboring atoms. Both means use the fixed graph, with zero aggregates for atoms without neighbors. The resulting 256-dimensional local feature is added to the coordinate projection, atom-type embedding, and projected Laplacian features at each input token.

Molecular conditioning. The descriptor contains ten components: five element fractions (C, H, O, N, S), three fractions of inferred edges in categories $1/2/3$, the atom count divided by 24, and the non-hydrogen atom fraction. A $10 \to 256 \to 256$ MLP with SiLU projects this descriptor. Its output is added to the learned step embedding at index zero for one-step generation. This condition supplies the shifts, scales, and residual gates in the transformer blocks and the shifts and scales of the final layer normalization.

A.4 MD17 TRAINING PROTOCOL

No-force, Ungated, and BIT share the coordinate generator, Adam optimizer, an initial learning rate of 10^{-3} with a 200k-step cosine schedule, and batch size 256. Within each training step, all generated queries share one batch of 512 reference conformations sampled uniformly with replacement from the training frames. Gradient-norm clipping uses a threshold of 10. The constant bandwidth is the median pairwise Euclidean distance between the distance-feature vectors of the first 1,000 training frames, with multiplier one. Training uses the zero-based frame interval $[0, 5000)$; primary evaluation uses reference frames $[5000, 55000)$, 1,000 generated samples, generation seed 42, and the final 200k checkpoint.

For BIT, No-force, and Ungated, reported sample SDs summarize runs initialized with seeds 42/43/44; implementation details, including random-number handling, are documented in the code README.

A.5 HYPERPARAMETER SELECTION

For BIT, the force-alignment coefficient g and mixing constant c are selected separately for each molecule using the nine candidate pairs in the Cartesian product

$$g \in \{0.05, 0.1, 0.2\}, \qquad c \in \{10^{-6}, 10^{-5}, 10^{-4}\}.$$

Each candidate is trained on the first 5,000 frames, $[0, 5000)$, with training seed 45 for 200k optimizer updates. We use its fixed final 200k checkpoint to generate 1,000 conformations and compute 200-bin pooled-distance TVD against the same 5,000 training frames. For molecule m, we select

$$(g_m, c_m) = \operatorname*{arg\,min}_{(g,c)\in\mathcal{G}\times\mathcal{C}} \mathrm{TVD}_m(g, c),$$

where $\mathcal{G}$ and $\mathcal{C}$ are the candidate sets above. The selected pair is then fixed for the main runs with training seeds 42/43/44, whose results provide the reported means and sample SDs. The evaluation interval $[5000, 55000)$ is not used to select g or c. All reported main runs use the final 200k checkpoint.

For BIT, $g = 0.2$ on all seven molecules except salicylic acid, which uses $g = 0.1$. The mixing constant is $c = 10^{-6}$ for malonaldehyde and toluene, $c = 10^{-4}$ for naphthalene, and $c = 10^{-5}$ for ethanol, uracil, salicylic acid, and aspirin. Each molecule uses the same g and c across training seeds 42/43/44.

Ungated independently selects g for each molecule from $\{0.05, 0.1, 0.2\}$ using the same selection protocol: each candidate is trained with seed 45 on $[0, 5000)$ for 200k updates, and its final checkpoint generates 1,000 conformations. We select the g with the lowest 200-bin pooled-distance TVD against those 5,000 training frames, then fix it across training seeds 42/43/44. These independently selected values coincide with BIT's: $g = 0.1$ for salicylic acid and $g = 0.2$ for the other six molecules. The evaluation interval $[5000, 55000)$ is excluded from this selection. Ungated uses t_F throughout training and has no mixing constant c.

A.6 Local baseline configurations

The local AvgFlow and ET-Flow configurations are taken from their official codebases and run under the common training protocol specified here. BIT, AvgFlow, and ET-Flow use the same first 5,000 frames per molecule, training batch size 256, 200k optimizer updates, and training seeds 42/43/44. Evaluation uses each method's final endpoint and the same sample count, generation seed, reference frames, and 200-bin pooled TVD evaluator. AvgFlow uses eight layers of width 512, with 100k teacher, 50k reflow, and 50k distillation updates; its final sampler uses one NFE. It uses AdamW with a per-stage cosine schedule: a 500-step warmup from 10^{-5} to 10^{-4}, followed by decay to 10^{-5}. Sampling uses the final distillation weights with exponential moving average (EMA) decay 0.999. ET-Flow uses 20 layers of width 160; its sampler uses 50 NFE. It uses AdamW with weight decay 10^{-8} and a 200k-step cosine schedule with maximum learning rate 5×10^{-4} and minimum 10^{-8}. Sampling uses its final 200k weights without EMA. Latency measurements also use a sampling batch size of 256.

B Mathematical background for adaptive target combination

B.1 Kernel score identities and proofs

Kernel score identities connect the reference scores to local attraction and motivate the mixing statistic used in BIT. Define the kernel-averaged score

$$s_{p,\tau}(x) = \frac{\int k_\tau(x,y)\nabla_y \log p(y)\, p(y)\, dy}{\int k_\tau(x,y) p(y)\, dy}.$$

For a Gaussian kernel on the ambient coordinates,

$$\tau^2\, s_{p,\tau}(x) \;=\; \frac{\int k_\tau(x,y)\,(y-x)\, p(y)\, dy}{\int k_\tau(x,y)\, p(y)\, dy}. \tag{12}$$

For the distance kernel, Eq. (7) holds with p_{B} replaced by any smooth density p and $s = \nabla \log p$, under the boundary assumptions below. It uses the Jacobian at each *reference* y. The regression gradient in Eq. (11) uses the Jacobian at the *generated query* x; in general,

$$J_\phi(x)^\top \sum_j w_j(\phi(y_j) - \phi(x)) \neq \sum_j w_j J_\phi(y_j)^\top(\phi(y_j) - \phi(x)). \tag{13}$$

Proposition 1. *Assume a smooth, strictly positive density p_{B} on the coordinate space under consideration, with vanishing integration-by-parts boundary terms and its exact score $s = \nabla \log p_{\mathrm{B}}$. For centered coordinates, the density and derivatives are defined on the linear subspace where the arithmetic mean of the atomic coordinates is zero. Then for* every *query x, $\mathbb{E}_{y\sim p_{\mathrm{B}}}\big[k_\tau(x,y)\big(\tau^2 s(y) - J_\phi(y)^\top(\phi(y) - \phi(x))\big)\big] = 0$.*

Ambient-coordinate identity. For the Gaussian kernel $k_\tau(x,y) = \exp\!\big(-\|x-y\|^2/2\tau^2\big)$ one has $\nabla_y k_\tau(x,y) = k_\tau(x,y)\,(x-y)/\tau^2$. Assume p is smooth, strictly positive, and decays fast enough that boundary terms vanish. Then

$$\begin{aligned}\int k_\tau(x,y)\,\nabla_y \log p(y)\,p(y)\,dy &= \int k_\tau(x,y)\,\nabla_y p(y)\,dy = -\int \nabla_y k_\tau(x,y)\,p(y)\,dy \\ &= \frac{1}{\tau^2}\int k_\tau(x,y)\,(y-x)\,p(y)\,dy,\end{aligned}$$

where the middle step is integration by parts. Dividing both sides by $\int k_\tau(x,y)\,p(y)\,dy$ yields Eq. (12).

Proof of Proposition 1. On the centered-coordinate subspace, use its induced Lebesgue measure and tangential gradient; translation-invariance makes the distance pullback tangent to that subspace. The distance kernel is locally Lipschitz, and its weak gradient agrees with the ordinary gradient away from pairwise coincidences. We assume the integrals below are finite and boundary terms vanish. For $k_\tau(x,y) = \exp\!\big(-\|\phi(y)-\phi(x)\|^2/2\tau^2\big)$, the chain rule gives $\nabla_y k_\tau(x,y) = -k_\tau(x,y)\,J_\phi(y)^\top\big(\phi(y)-\phi(x)\big)/\tau^2$. Then

$$\begin{aligned}\mathbb{E}_{p_{\mathrm{B}}}\big[k_\tau\, s\big] &= \int k_\tau(x,y)\,\nabla_y p_{\mathrm{B}}(y)\,dy = -\int \nabla_y k_\tau(x,y)\,p_{\mathrm{B}}(y)\,dy \\ &= \frac{1}{\tau^2}\,\mathbb{E}_{p_{\mathrm{B}}}\big[k_\tau\, J_\phi^\top(\phi(y)-\phi(x))\big],\end{aligned}$$

which is the claim.

B.2 Residual under a different reference population

Let p be the density associated with exact force score s_p, and let q be a smooth reference-population density with score s_q on the same coordinate space; q is neither the generator distribution nor the discrete empirical measure. Write $Z_q(x) = \mathbb{E}_{y\sim q}[k_\tau(x,y)] > 0$. The normalized population residual is

$$\begin{aligned}r_{p,q}(x) &= \frac{\mathbb{E}_{y\sim q}\big[k_\tau(x,y)\,\{\tau^2 s_p(y) - J_\phi(y)^\top(\phi(y)-\phi(x))\}\big]}{Z_q(x)} \\ &= \tau^2\,\frac{\mathbb{E}_{y\sim q}[k_\tau(x,y)(s_p(y)-s_q(y))]}{Z_q(x)}.\end{aligned}$$

To obtain the second equality, use the same boundary and integrability assumptions as in Proposition 1. Integration by parts under q gives

$$\begin{aligned}\mathbb{E}_q[k_\tau s_q] &= \int k_\tau(x,y)\,\nabla_y q(y)\,dy = -\int \nabla_y k_\tau(x,y)\,q(y)\,dy \\ &= \tau^{-2}\mathbb{E}_q\big[k_\tau(x,y) J_\phi(y)^\top(\phi(y)-\phi(x))\big]\,.\end{aligned}$$

Substitution expresses the residual as a kernel-weighted score difference (Liu et al., 2016, Lemma 2.3). It vanishes for every query when $q = p$, but symmetry can also make it vanish when the distributions differ. For example, let p and q be zero-mean isotropic Gaussians with different variances on the centered-coordinate subspace. Since $\phi(-y) = \phi(y)$, the kernel and q are even in y, while $s_p - s_q$ is odd; the numerator therefore vanishes for every query despite the different distance distributions. Accordingly, BIT uses D as a finite-reference mixing statistic.

B.3 Conditional risk and assumptions for target mixing

Let t_* be an abstract target for the risk analysis, $e = t_* - t_0$, and $\Delta = t_F - t_0$. Conditioning on the information $\mathcal{I}$ used by the mixing rule fixes D and γ; expectations average over the remaining variation in e and Δ. Assuming finite second moments, write $a = \mathbb{E}[e^\top\Delta \mid \mathcal{I}]$ and $b = \mathbb{E}[\|\Delta\|^2 \mid \mathcal{I}]$. Expanding $R(\gamma) = \mathbb{E}[\|e-\gamma\Delta\|^2 \mid \mathcal{I}]$ gives

$$R(0) - R(\gamma) = 2\gamma a - \gamma^2 b, \qquad \gamma_* = \mathrm{clip}_{[0,1]}\,\frac{a}{b} \quad (b>0). \tag{14}$$

For $b > 0$ and $0 < \gamma \leq 1$, improvement occurs precisely when $a > 0$ and $\gamma < 2a/b$. If $b = 0$, the correction vanishes almost surely and all weights have the same risk. When conditioning on D, a nondecreasing ratio $a(D)/b(D)$ suffices for a nondecreasing optimal bounded weight; an increase in residual magnitude alone does not establish that ratio.

Mixing coefficient under additional assumptions. To connect Eq. (14) to the implemented mixing coefficient, set $e_0 = t_0 - t_*$ and $e_F = t_F - t_*$. Under the same conditioning, define $A = \mathbb{E}[\|e_0\|^2 \mid \mathcal{I}]$, $B = \mathbb{E}[\|e_F\|^2 \mid \mathcal{I}]$, and $C = \mathbb{E}[e_0^\top e_F \mid \mathcal{I}]$. These errors may be biased and correlated. Since $a = A - C$ and $b = A + B - 2C$, the equivalent risk and minimizer are

$$R(\gamma) = (1-\gamma)^2 A + \gamma^2 B + 2\gamma(1-\gamma)C, \qquad \gamma_* = \mathrm{clip}_{[0,1]} \frac{A - C}{A + B - 2C}. \tag{15}$$

The minimizer follows by differentiation when $A + B - 2C = \mathbb{E}[\|e_F - e_0\|^2 \mid \mathcal{I}] > 0$; otherwise every mixture has the same risk. Under the additional assumptions $A - C = \kappa D$ and $B - C = \kappa c$, with $\kappa > 0$ and $c > 0$,

$$\gamma_* = \frac{D}{D + c}, \qquad \frac{\partial \gamma_*}{\partial D} = \frac{c}{(D + c)^2} > 0. \tag{16}$$

One sufficient model is $e_0 = u + \eta_0$, $e_F = u + \eta_F$, with u, η_0, and η_F pairwise orthogonal conditional on $\mathcal{I}$, and second moments $\mathbb{E}[\|\eta_0\|^2 \mid \mathcal{I}] = \kappa D$, $\mathbb{E}[\|\eta_F\|^2 \mid \mathcal{I}] = \kappa c$.

The result describes the squared error of the mixed target under these assumptions.

C Supplementary MD17 results and controls

C.1 Generation-seed replication

To assess dependence on the primary generation seed 42, we repeat sampling with generation seeds 42/43/44. After averaging these draws within each training seed, the seven-molecule mean TVDs are 0.1054 for No-force, 0.0799 for Ungated, and 0.0471 for BIT. Per-molecule results agree with the ordering in Table 1.

C.2 Alternative distance metric and published context

Histogram MAE is defined in Appendix D.1.

Table 3: MD17 pairwise-distance histogram MAE (↓). Local results use the protocol of Table 2: per-molecule values are mean ± sample SD over training seeds 42/43/44; the Mean row averages molecules. Published DSM/PSM values are from Guo et al. (2025) under their original evaluation protocols. Bold marks the lowest mean among locally trained methods per row.

	One-step, matched updates		Multi-step, matched updates	Multi-step, published		
Molecule	BIT (ours)	AvgFlow	ET-Flow	DSM	PSM-P	PSM-W
NFE	1	1	50	∼2000	∼2000	∼2000
ethanol	**0.0720** ± 0.0009	0.0925 ± 0.0036	0.1462 ± 0.0060	0.1193	0.0845	0.0911
malonaldehyde	0.0911 ± 0.0013	**0.0775** ± 0.0032	0.0985 ± 0.0040	0.1003	0.0843	0.0630
uracil	**0.0613** ± 0.0021	0.1078 ± 0.0012	0.0776 ± 0.0014	0.0747	0.0705	0.0560
toluene	**0.0705** ± 0.0064	0.1269 ± 0.0071	0.1284 ± 0.0082	0.0970	0.0769	0.0948
salicylic acid	**0.0383** ± 0.0016	0.1062 ± 0.0021	0.0931 ± 0.0050	0.0799	0.0554	0.0584
naphthalene	**0.1009** ± 0.0015	0.1084 ± 0.0029	0.1396 ± 0.0030	0.1257	0.0701	0.1164
aspirin	0.0548 ± 0.0004	0.0449 ± 0.0014	**0.0424** ± 0.0022	0.1240	0.0570	0.0510
Mean	**0.0698**	0.0949	0.1037	0.1030	0.0712	0.0758

Published sampling results. Table 4 retains the published evaluation protocols; these values provide context for the local comparisons.

Table 4: Published MD17 pooled-distance TVD (↓), reported by Guo et al. (2025). PSM-P and PSM-W denote the Piecewise and Weighted variants. These values retain the original evaluation protocols.

Molecule NFE	DSM ∼2000	PSM-P ∼2000	PSM-W ∼2000
ethanol	0.0663	0.0508	0.0501
malonaldehyde	0.0515	0.0492	0.0518
uracil	0.0478	0.0421	0.0426
toluene	0.0536	0.0428	0.0490
salicylic acid	0.0450	0.0424	0.0337
naphthalene	0.0716	0.0365	0.0592
aspirin	0.0662	0.0373	0.0290
Mean	0.0574	0.0430	0.0451

C.3 Projected-force pairing control

This three-molecule diagnostic was run on ethanol, toluene, and aspirin with training seeds 42/43/44. We permute projected force vectors $\widetilde{P}(y_j)$ among the 512 positive references before computing their alignments in Eq. (5). The Cartesian scores used to compute the residual retain their original pairing; the data-only computation is unchanged. A separate random-number stream drives the permutation, leaving the other training draws matched to BIT. All runs use the final 200k checkpoint and the primary evaluation protocol.

Table 5: Projected-force pairing control: pooled-distance TVD (↓), mean ± sample SD across training seeds 42/43/44. Paired differences are Permuted-force minus BIT. The final row averages the three molecules within each seed before aggregation. Bold marks the lowest mean among the two generative methods per row; the difference column is not ranked.

Molecule	BIT	Permuted-force	Paired difference
ethanol	$\mathbf{0.0492} \pm 0.0008$	0.0892 ± 0.0101	$+0.0399 \pm 0.0094$
toluene	$\mathbf{0.0521} \pm 0.0015$	0.2310 ± 0.0109	$+0.1788 \pm 0.0097$
aspirin	$\mathbf{0.0358} \pm 0.0007$	0.0923 ± 0.0050	$+0.0565 \pm 0.0044$
Mean	$\mathbf{0.0457} \pm 0.0002$	0.1375 ± 0.0013	$+0.0917 \pm 0.0014$

C.4 Batch-mean target combination

For each generated batch B, the Batch-mean variant computes the original coefficients and replaces them by $\bar{\gamma}_B = |B|^{-1} \sum_{x \in B} \gamma(x)$ before constructing the targets. This three-molecule diagnostic was run on naphthalene, ethanol, and aspirin with training seeds 42/43/44. All runs use final 200k checkpoints, 1,000 generated samples per run, and generation seed 42, with the same initialization, training data, and evaluation references as BIT.

Table 6: Batch-mean control: pooled-distance TVD (↓), mean ± sample SD over training seeds 42/43/44. Relative reduction is $100 \times (\mathrm{Batch\text{-}mean} - \mathrm{BIT})/\mathrm{Batch\text{-}mean}$, computed from unrounded TVD means. The last column counts seeds with lower BIT TVD. Bold marks the lower mean per row.

Molecule	Batch-mean	BIT	Relative reduction	Seeds with lower BIT TVD
naphthalene	0.0654 ± 0.0048	$\mathbf{0.0599} \pm 0.0014$	8.40%	3/3
ethanol	0.0524 ± 0.0045	$\mathbf{0.0492} \pm 0.0008$	5.99%	2/3
aspirin	0.0366 ± 0.0012	$\mathbf{0.0358} \pm 0.0007$	2.22%	3/3

C.5 Reference sampling and training-set reuse

Table 7 evaluates sets of 1,000 sampled frames using the primary distance metric. Sampling from the evaluation reference characterizes finite-sample differences; sampling from the training set provides a comparison with direct data reuse. The training-set comparison uses a single draw without replacement from the 5,000 training frames. It has an equally weighted mean TVD of 0.0412 across the seven molecules, below BIT's 0.0472. At the reported precision, training-set samples have lower TVD on four molecules, equal TVD on two, and higher TVD on ethanol.

Table 7: Reference sampling and training-set reuse: $1{,}000$-frame draws scored as sample sets under the primary distance evaluator (↓).

	Distance TVD		Distance MAE	
Molecule	Reference draw	Training-set draw	Reference draw	Training-set draw
ethanol	0.0242	0.0615	0.0174	0.0537
malonaldehyde	0.0277	0.0411	0.0245	0.0514
uracil	0.0171	0.0434	0.0136	0.0525
toluene	0.0152	0.0316	0.0138	0.0406
salicylic acid	0.0125	0.0250	0.0138	0.0303
naphthalene	0.0115	0.0559	0.0115	0.0932
aspirin	0.0111	0.0297	0.0134	0.0377
Mean	0.0170	0.0412	0.0154	0.0513

D Distance-distribution evaluation and robustness

D.1 Primary distance metrics

The primary TVD pools all unordered atom-pair distances. Its 200 bins span the distances in the 50k-frame reference, and TVD is one half the ℓ_1 difference between normalized histogram counts. The primary evaluator normalizes after excluding out-of-range generated distances; Appendix D.2 reports the corresponding calculation with this mass retained. Atom-pair-resolved evaluation is specified in Appendix D.3.

We also report scaled histogram MAE, following the existing evaluation convention (Guo et al., 2025). Its fixed grid of $K = 200$ bins on $[10^{-3}, 20]$ Å over the full distance matrix differs from the reference-derived grid used for primary TVD. For density-normalized histogram heights h_b and h_b^{ref} in Å^{-1}, the dimensionless metric is

$$\mathrm{MAE}_h = \frac{20\ \text{Å}}{K} \sum_{b=1}^{K} |h_b - h_b^{\mathrm{ref}}|.$$

D.2 Retaining out-of-range distances

We evaluate the final-200k samples of No-force, Ungated, and BIT on seven MD17 molecules and three training seeds, for 63 sets of 1,000 conformations. Each comparison uses the reference-derived edges of the same 200 bins and the same reference probabilities, with the primary evaluator's float32 upper-triangle distance calculation. Let n_b be the interior counts, n_- and n_+ the counts below and above the reference range, N their total, and p_b the reference probability in bin b. Keeping the original inclusive rightmost edge, the TVD with two overflow bins is

$$\mathrm{TVD}_{\mathrm{all}} = \tfrac{1}{2}\left[\frac{n_- + n_+}{N} + \sum_{b=1}^{200} \left|\frac{n_b}{N} - p_b\right|\right].$$

Table 8: Pooled-distance TVD (↓) with and without out-of-range mass. Mean ± sample SD of seven-molecule means across three training seeds. Out-of-range mass is the percentage of generated distances outside the reference range. Bold marks the lowest mean in each TVD column.

Method	TVD out-of-range excluded	TVD out-of-range retained	Out-of-range mass (%)
No-force	0.1056 ± 0.0031	0.1059 ± 0.0031	0.069 ± 0.006
Ungated	0.0800 ± 0.0014	0.0815 ± 0.0015	0.290 ± 0.010
BIT	$\mathbf{0.0472} \pm 0.0009$	$\mathbf{0.0478} \pm 0.0009$	0.132 ± 0.016

D.3 Atom-pair-resolved distributions

For each unordered atom pair, 128 equal-width bins span that pair's minimum and maximum distance in the 5,000 training frames, and two additional bins retain underflow and overflow mass. All methods and the 50k-frame evaluation reference use the same edges. TVD is computed separately for each pair and then averaged equally across pairs. We reuse the final-200k sample sets from the primary experiment. For the Reference column, each of three 1,000-frame draws is compared with the full 50k-frame evaluation reference, and the resulting TVDs are averaged to characterize sample-size effects. Each draw samples without replacement from that reference (RNG 20260920).

Table 9: Atom-pair-resolved distance TVD (↓); 1,000 generated conformations per run. Generated-method values are mean ± sample SD across three training seeds. Bold marks the lowest mean among generated methods per row; the reference column is not ranked.

Molecule	Reference	No-force	Ungated	BIT
ethanol	0.1206	0.2502 ± 0.0037	0.2954 ± 0.0033	$\mathbf{0.2433} \pm 0.0039$
malonaldehyde	0.1279	0.2824 ± 0.0065	0.2464 ± 0.0035	$\mathbf{0.2159} \pm 0.0023$
uracil	0.1225	0.1782 ± 0.0035	0.1772 ± 0.0012	$\mathbf{0.1529} \pm 0.0015$
toluene	0.1265	0.3133 ± 0.0036	0.2066 ± 0.0006	$\mathbf{0.1635} \pm 0.0035$
salicylic acid	0.1226	0.1709 ± 0.0020	0.1575 ± 0.0023	$\mathbf{0.1424} \pm 0.0012$
naphthalene	0.1208	0.2004 ± 0.0220	0.2038 ± 0.0014	$\mathbf{0.1683} \pm 0.0014$
aspirin	0.1267	0.2646 ± 0.0020	0.2450 ± 0.0006	$\mathbf{0.1802} \pm 0.0011$
Mean	0.1239	0.2372 ± 0.0030	0.2189 ± 0.0009	$\mathbf{0.1809} \pm 0.0011$

D.4 Histogram-resolution sensitivity

Table 10 varies the number of interior bins per atom pair, using the same training-range construction and two overflow bins as Appendix D.3.

Table 10: Atom-pair TVD across histogram resolutions (↓). Values average the seven molecule means over training seeds 42/43/44. All resolutions retain underflow and overflow mass. Bold marks the lowest value per row.

Interior bins per pair	No-force	Ungated	BIT
64	0.2181	0.1947	**0.1535**
128	0.2372	0.2189	**0.1809**
256	0.2685	0.2587	**0.2239**

E Joint distance-vector evaluation

Evaluation protocol. The joint evaluation covers all seven MD17 molecules with training seeds 42/43/44. We compare BIT, No-force, AvgFlow (1 NFE), and ET-Flow (50 NFE). Each trained model provides 1,000 generated conformations, evaluated against three fixed reference sets of 1,000 frames each from $[20000, 55000)$. For each molecule, we uniformly permute the 35,000 frame indices with

NumPy `default_rng(20260922)` and split the first 3,000 indices into three disjoint sets. The same reference sets are used across methods and training seeds. We first average the three reference comparisons for each model, then report the mean and sample SD of these averages across the three training seeds in Table 11.

Kernel and estimator. Let $\phi(x) \in \mathbb{R}^P$ contain all unordered atom-pair distances in a fixed atom-pair order, with distances measured in Å and P atom pairs. The evaluation kernel is

$$k(x, y) = \exp\left[-\frac{\|\phi(x) - \phi(y)\|_2^2}{2P(0.25\sigma)^2}\right].$$

For each molecule, we sample 1,024 of the 5,000 training conformations without replacement (RNG 20260921). The scale σ is the median of the nonzero values of $\|\phi(y_i) - \phi(y_j)\|_2/\sqrt{P}$ between these conformations. All methods use the fixed evaluation bandwidth 0.25σ, which is distinct from the training bandwidth τ. For a generated set $X = \{x_i\}_{i=1}^N$ and a reference set $Y = \{y_i\}_{i=1}^N$, with $N = 1000$, we use the unbiased-form MMD^2 estimator

$$\widehat{\mathrm{MMD}}_u^2(X, Y) = \frac{\sum_{i\neq j} k(x_i, x_j) + \sum_{i\neq j} k(y_i, y_j)}{N(N-1)} - \frac{2}{N^2}\sum_{i,j} k(x_i, y_j).$$

The within-set sums exclude diagonal terms, and negative estimates are not clipped.

Table 11: Joint distance-vector MMD^2 (↓), mean ± sample SD across training seeds 42/43/44 after averaging three fixed reference-set comparisons per trained model. Both means and SDs are multiplied by 10^3 for display. All methods use kernel bandwidth 0.25σ. Bold marks the lowest mean among the four listed methods for each molecule.

Molecule	BIT	No-force	AvgFlow	ET-Flow
ethanol	14.446 ± 0.165	19.422 ± 0.283	29.512 ± 0.469	$\mathbf{9.532} \pm 0.010$
malonaldehyde	23.122 ± 0.663	34.457 ± 2.726	21.429 ± 0.719	$\mathbf{12.805} \pm 0.166$
uracil	$\mathbf{1.232} \pm 0.037$	3.607 ± 0.195	14.827 ± 0.771	5.530 ± 0.420
toluene	$\mathbf{4.513} \pm 0.141$	14.640 ± 0.991	9.731 ± 0.860	7.490 ± 0.454
salicylic acid	$\mathbf{1.287} \pm 0.111$	5.381 ± 0.066	12.692 ± 0.190	12.424 ± 1.838
naphthalene	$\mathbf{2.084} \pm 0.151$	12.166 ± 5.836	10.008 ± 0.361	14.006 ± 0.654
aspirin	7.553 ± 0.652	12.855 ± 0.390	15.668 ± 0.623	$\mathbf{7.369} \pm 0.088$

F Geometric state occupancy and training-set proximity

F.1 Samples and reference construction

The diagnostics reuse the final main-experiment samples for BIT, No-force, and AvgFlow on all seven MD17 molecules at training seeds 42/43/44 (1,000 conformations per checkpoint). BIT uses the archived main checkpoints; no model is retrained and no model samples are regenerated. Training frames occupy $[0, 5000)$. For each molecule, NumPy `default_rng(20260922)` permutes the 35,000 indices in $[20000, 55000)$; the first 6,000 are split into six disjoint 1,000-frame sets. The first three are the common reference sets used in Appendix E; the other three form the reference-sampling baseline. These sets do not overlap training frames and come from the historical evaluation window.

Training resampling draws 1,000 frames with replacement from all 5,000 training frames using seeds 42/43/44. These three draws differ from the single without-replacement training draw in Table 7. Across the three learned methods and two sampling controls, the diagnostics cover 105 sample sets. Learned-model error bars represent sample SD across training seeds; control error bars represent sample SD across sampling repeats.

F.2 PCA representation and geometric state occupancy

PCA uses the complete fixed-atom pairwise-distance vectors, including explicit hydrogens, in Å. It is fitted only on the 5,000 training frames using centered full SVD, without feature scaling or

whitening. We retain $d = \min\{d_{95}, 10\}$ components, where d_{95} is the number required to explain 95% of training variance. All seven molecules have $d = 10$; Table 12 reports their actual cumulative explained variance.

Table 12: Training-only PCA representation used for all geometric partitions. The ten-component cap is active for each molecule.

Molecule	Atom pairs	Components	Retained variance (%)
ethanol	36	10	94.72
malonaldehyde	36	10	94.85
uracil	66	10	86.44
toluene	105	10	94.10
salicylic acid	120	10	83.25
naphthalene	153	10	85.03
aspirin	210	10	92.14

Partition and occupancy metric. The representation is fitted on training frames, whereas the occupancy partition is defined from reference conformations that do not overlap training. In the fixed PCA representation, K-means is fitted on the 3,000 common reference frames, with random seed 20260924, 20 initializations, at most 300 Lloyd iterations, and tolerance 10^{-4}. Generated samples participate in neither fit. The same partition is used across methods and training seeds for each molecule and K.

For occupancy probabilities p_k and q_k, TVD is $\frac{1}{2}\sum_k |p_k - q_k|$. An additional invalid category retains samples without finite features; none occur in these evaluations. Each sample set is compared separately with the three common reference sets, and the TVDs are averaged before aggregation across training seeds or sampling repeats. The primary setting is $K = 10$; $K = 5$ and 20 use the same samples and PCA representation. Table 13 reports every molecule and method at all three resolutions. These are descriptive comparisons, and the clusters are geometric partitions, not kinetic metastable states or free-energy basins.

Table 13: Geometric state-occupancy TVD at every evaluated resolution (↓), mean ± sample SD across three training seeds or sampling repeats, after averaging three reference comparisons per run. The $K = 10$ Mean row averages molecules within each seed/repeat before computing mean and SD.

Molecule	BIT	No-force	AvgFlow	Training resampling	Reference-sampling baseline
$K = 5$					
ethanol	0.2653 ± 0.0042	0.2953 ± 0.0122	0.2837 ± 0.0065	0.2766 ± 0.0207	0.0284 ± 0.0075
malonaldehyde	0.3629 ± 0.0008	0.3617 ± 0.0040	0.3311 ± 0.0176	0.3769 ± 0.0083	0.0279 ± 0.0107
uracil	0.1193 ± 0.0050	0.1314 ± 0.0138	0.1623 ± 0.0096	0.1300 ± 0.0249	0.0274 ± 0.0107
toluene	0.1207 ± 0.0038	0.1373 ± 0.0176	0.1827 ± 0.0038	0.1816 ± 0.0281	0.0283 ± 0.0038
salicylic acid	0.1070 ± 0.0069	0.1076 ± 0.0073	0.1827 ± 0.0095	0.0576 ± 0.0161	0.0254 ± 0.0076
naphthalene	0.0888 ± 0.0170	0.0986 ± 0.0232	0.1019 ± 0.0120	0.0861 ± 0.0015	0.0321 ± 0.0071
aspirin	0.1673 ± 0.0087	0.1617 ± 0.0176	0.0824 ± 0.0110	0.1080 ± 0.0081	0.0308 ± 0.0074
$K = 10$ (primary)					
ethanol	0.5530 ± 0.0106	0.5721 ± 0.0135	0.6050 ± 0.0065	0.5314 ± 0.0132	0.0564 ± 0.0051
malonaldehyde	0.3646 ± 0.0183	0.3780 ± 0.0078	0.3563 ± 0.0111	0.3367 ± 0.0068	0.0453 ± 0.0027
uracil	0.1602 ± 0.0213	0.1823 ± 0.0142	0.1937 ± 0.0078	0.1639 ± 0.0165	0.0538 ± 0.0108
toluene	0.1292 ± 0.0022	0.1488 ± 0.0141	0.1872 ± 0.0075	0.1843 ± 0.0120	0.0477 ± 0.0122
salicylic acid	0.1318 ± 0.0077	0.1346 ± 0.0044	0.1760 ± 0.0042	0.1372 ± 0.0097	0.0492 ± 0.0115
naphthalene	0.1751 ± 0.0098	0.1711 ± 0.0488	0.1316 ± 0.0093	0.1504 ± 0.0056	0.0551 ± 0.0075
aspirin	0.1832 ± 0.0070	0.1939 ± 0.0081	0.2272 ± 0.0127	0.2403 ± 0.0236	0.0464 ± 0.0076
Mean	0.2424 ± 0.0045	0.2544 ± 0.0042	0.2681 ± 0.0031	0.2492 ± 0.0066	0.0506 ± 0.0040
$K = 20$					
ethanol	0.5842 ± 0.0108	0.6166 ± 0.0208	0.6397 ± 0.0068	0.5987 ± 0.0186	0.0729 ± 0.0023
malonaldehyde	0.5107 ± 0.0223	0.5071 ± 0.0118	0.4766 ± 0.0076	0.5402 ± 0.0109	0.0651 ± 0.0064
uracil	0.1586 ± 0.0077	0.1837 ± 0.0074	0.2298 ± 0.0104	0.1640 ± 0.0105	0.0759 ± 0.0081
toluene	0.1871 ± 0.0039	0.2309 ± 0.0164	0.2097 ± 0.0088	0.2204 ± 0.0108	0.0767 ± 0.0058
salicylic acid	0.1361 ± 0.0042	0.1410 ± 0.0172	0.1827 ± 0.0081	0.1196 ± 0.0113	0.0633 ± 0.0044
naphthalene	0.2117 ± 0.0228	0.2479 ± 0.0294	0.1594 ± 0.0053	0.2120 ± 0.0058	0.0727 ± 0.0072
aspirin	0.2517 ± 0.0263	0.2651 ± 0.0183	0.2738 ± 0.0101	0.3506 ± 0.0191	0.0671 ± 0.0025

F.3 Nearest-neighbor distance to training conformations

Let $\phi(x)$ contain all unordered pairwise distances, including explicit hydrogens, with fixed atom identities and distances in Å. For every sample, we compute

$$d_{\mathrm{NN}}(x) = \min_{y \in \mathcal{D}_{\mathrm{train}}} \|\phi(x) - \phi(y)\|_2$$

against all 5,000 training frames, without PCA or an approximate search index. Table 14 reports each sample set's median, followed by the mean and sample SD of these medians across three training seeds or sampling repeats. No evaluated sample from BIT, No-force, or AvgFlow has $d_{\mathrm{NN}} \leq 10^{-10}$ Å; training resampling has zero distance by construction. All 105,000 queries have finite nearest-neighbor distances.

Table 14: Nearest-neighbor distance to all training frames: mean ± sample SD of the three sample-set medians, in Å. Variability is across training seeds for learned methods and sampling repeats for controls. Distances describe proximity and are not ranked as a quality score.

Molecule	BIT	No-force	AvgFlow	Training resampling	Reference-sampling baseline
ethanol	1.0854 ± 0.1426	1.1410 ± 0.0332	0.5070 ± 0.0019	0.0000 ± 0.0000	1.5362 ± 0.0138
malonaldehyde	0.5740 ± 0.0027	0.4960 ± 0.0092	0.5863 ± 0.0104	0.0000 ± 0.0000	0.9003 ± 0.0043
uracil	0.4364 ± 0.0013	0.3703 ± 0.0022	0.3462 ± 0.0018	0.0000 ± 0.0000	0.4268 ± 0.0025
toluene	0.8821 ± 0.0071	0.7217 ± 0.0012	0.7904 ± 0.0067	0.0000 ± 0.0000	0.9265 ± 0.0080
salicylic acid	0.6715 ± 0.0027	0.6086 ± 0.0027	0.5940 ± 0.0018	0.0000 ± 0.0000	0.7409 ± 0.0016
naphthalene	0.7847 ± 0.0024	0.6358 ± 0.0161	0.6438 ± 0.0042	0.0000 ± 0.0000	0.7913 ± 0.0009
aspirin	2.4220 ± 0.0293	2.2353 ± 0.0517	1.8597 ± 0.0171	0.0000 ± 0.0000	2.7190 ± 0.0341

F.4 DISTRIBUTIONS AND NORMALIZATION

Figure 5 shows the nearest-neighbor distributions for every molecule. If P is the number of atom pairs, the per-pair RMS distance is $d_{\mathrm{NN}}/\sqrt{P}$; it has units of Å and preserves within-molecule rankings. Table 15 reports these normalized medians, and Table 16 compares BIT and reference-sampling quantiles. Machine-readable results include all five methods' quantiles, distances to the pooled 3,000-frame reference bank, and full cluster populations at all three resolutions.

BIT and the reference-sampling baseline have similar median distance scales on several molecules, including toluene, salicylic acid, and naphthalene; BIT is closer to training frames on ethanol and malonaldehyde. Small distances alone do not establish memorization, and large distances do not establish plausible novelty. The exact-match threshold tests nearly identical complete distance vectors and does not rule out approximate memorization. Pairwise distances also cannot distinguish mirror images.

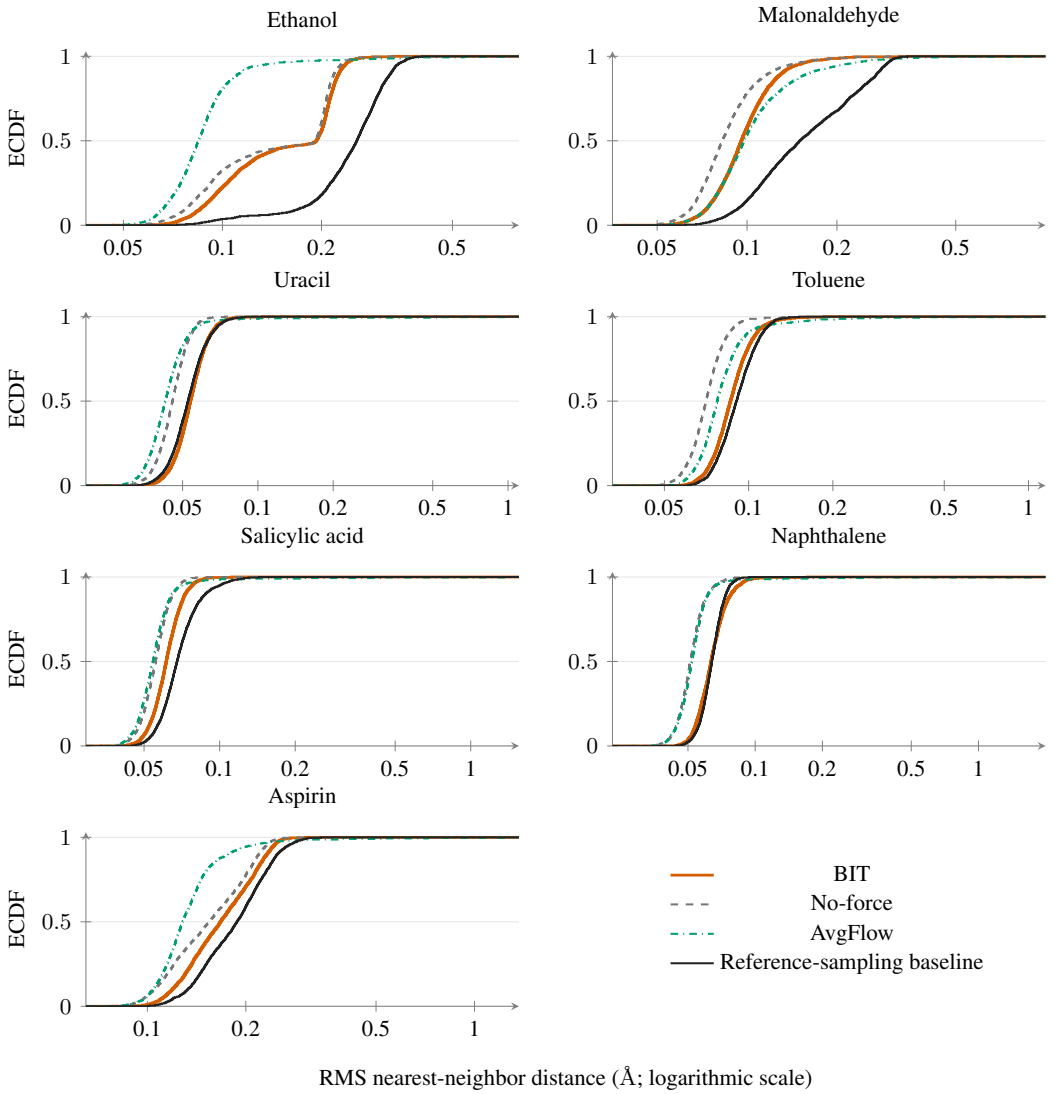

Figure 5: Empirical CDFs of distance to the training set for all seven molecules, expressed as per-pair RMS distance on logarithmic horizontal axes. Each curve averages three equal-sized sample sets. Training resampling has all mass at zero and is omitted from the curves. Horizontal ranges differ by molecule. The distributions characterize proximity rather than a ranking of sample quality.

Table 15: Per-pair RMS nearest-neighbor distance to training frames, $d_{\mathrm{NN}}/\sqrt{P}$, in Å. Entries are mean $\pm$ sample SD of the three medians, with the same aggregation as Table 14.

Molecule	BIT	No-force	AvgFlow	Training resampling	Reference-sampling baseline
ethanol	0.1809 ± 0.0238	0.1902 ± 0.0055	0.0845 ± 0.0003	0.0000 ± 0.0000	0.2560 ± 0.0023
malonaldehyde	0.0957 ± 0.0005	0.0827 ± 0.0015	0.0977 ± 0.0017	0.0000 ± 0.0000	0.1501 ± 0.0007
uracil	0.0537 ± 0.0002	0.0456 ± 0.0003	0.0426 ± 0.0002	0.0000 ± 0.0000	0.0525 ± 0.0003
toluene	0.0861 ± 0.0007	0.0704 ± 0.0001	0.0771 ± 0.0007	0.0000 ± 0.0000	0.0904 ± 0.0008
salicylic acid	0.0613 ± 0.0002	0.0556 ± 0.0002	0.0542 ± 0.0002	0.0000 ± 0.0000	0.0676 ± 0.0001
naphthalene	0.0634 ± 0.0002	0.0514 ± 0.0013	0.0520 ± 0.0003	0.0000 ± 0.0000	0.0640 ± 0.0001
aspirin	0.1671 ± 0.0020	0.1542 ± 0.0036	0.1283 ± 0.0012	0.0000 ± 0.0000	0.1876 ± 0.0024

Table 16: Nearest-neighbor distance quantiles for BIT and the reference-sampling baseline, in Å. Each entry averages the indicated within-sample-set quantile across three training seeds or sampling repeats. All five methods' per-run quantiles accompany the machine-readable results.

Molecule	Method	5%	25%	50%	75%	95%
ethanol	BIT	0.4718	0.6161	1.0854	1.2753	1.4080
ethanol	Reference sampling	0.6962	1.2798	1.5362	1.7588	2.0390
malonaldehyde	BIT	0.4074	0.5000	0.5740	0.6677	0.8804
malonaldehyde	Reference sampling	0.5008	0.6728	0.9003	1.3130	1.7462
uracil	BIT	0.3394	0.3963	0.4364	0.4813	0.5566
uracil	Reference sampling	0.3243	0.3832	0.4268	0.4751	0.5678
toluene	BIT	0.6924	0.7996	0.8821	0.9786	1.1692
toluene	Reference sampling	0.7256	0.8325	0.9265	1.0373	1.2131
salicylic acid	BIT	0.5369	0.6143	0.6715	0.7347	0.8487
salicylic acid	Reference sampling	0.5738	0.6695	0.7409	0.8359	1.0904
naphthalene	BIT	0.6303	0.7111	0.7847	0.8686	1.0373
naphthalene	Reference sampling	0.6488	0.7320	0.7913	0.8544	0.9526
aspirin	BIT	1.6177	1.9795	2.4220	2.9948	3.5201
aspirin	Reference sampling	1.7548	2.1909	2.7190	3.2218	3.8760

G SAMPLING EFFICIENCY

G.1 TIMING PROTOCOL

Timing uses float32 tensors and final seed-42 checkpoints. The timed region includes noise generation, sampling, scale restoration where applicable, and host transfer, with device synchronization. Checkpoint loading, graph preprocessing, and fixed conditioning are prepared before timing. All three methods are measured sequentially on the same GPU, with no concurrent training or competing model computation. Each measurement uses three warmups followed by nine timed repetitions.

BIT uses PyTorch 2.11, and ET-Flow uses PyTorch 2.8; both use high matrix-multiplication precision. AvgFlow uses JAX 0.5.0 and Flax 0.10.2 with the default matrix-multiplication precision. We compile its equivalent one-step Euler update directly in JAX, retaining the original network, weights, conditioning, and random keys. Across all seven molecules, its coordinates differ from the native Diffrax 0.6.2 sampler by at most 4×10^{-6} in maximum absolute error. Compilation completes during warmup. ET-Flow evaluates its network 50 times per sample, corresponding to 200 calls over the four batches.

G.2 LOCAL METHOD TIMING

Table 17: Sampling time in milliseconds (↓) for 1,000 conformations on one RTX 5090, batch 256. Cells report per-molecule medians under the protocol above; the final row averages all seven molecules. Bold marks the lowest time per row.

Molecule NFE	BIT 1	AvgFlow 1	ET-Flow 50
ethanol	**6.72**	34.55	2587.27
malonaldehyde	**6.79**	35.43	2588.42
uracil	**6.85**	45.80	4269.66
toluene	**7.11**	50.36	6811.00
salicylic acid	**7.12**	50.67	7677.04
naphthalene	**7.29**	58.22	9602.25
aspirin	**8.09**	62.41	12777.31
Mean	**7.14**	48.21	6616.14

G.3 DSM AND PSM TIMING CONTEXT

Table 18 adds ethanol measurements from released DSM, PSM-P, and PSM-W checkpoints. Their predictor–corrector sampler uses 1,000 steps, one corrector step per predictor step, and SNR 0.15, for 2,000 NFE per sample and 8,000 calls across the four batches. For these DSM/PSM entries, seven repetitions follow three warmups. The sample count, tensor precision, and timing boundaries match the preceding protocol; XYZ serialization is excluded. These implementations use Python 3.10, PyTorch 2.7, PyTorch Geometric 2.7, e3nn 0.6, and Lightning 2.1. PSM-W uses the server's second RTX 5090; the other measurements use the first. No competing model computation runs on the timed GPU.

Table 18: Representative ethanol sampling time (↓) for 1,000 conformations, batch 256, on RTX 5090 GPUs in the same server. BIT, AvgFlow, and ET-Flow entries come from Table 17; the DSM/PSM entries are seven-repetition medians. Times are in milliseconds, with the lowest value in bold.

Method	NFE	Time (ms)
BIT	1	**6.72**
AvgFlow	1	34.55
ET-Flow	50	2587.27
DSM	2000	486105.32
PSM-P	2000	487368.40
PSM-W	2000	499089.92