OPERATING-POINT LOCALITY OF SOFT CONSTRAINTS IN FLOW-MATCHING GENERATORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Few-step flow-matching generators such as MeanFlow expose two operating modes—a one-function-evaluation (1-NFE) map and a multi-step Euler sampler—and neither respects, by construction, the continuous symmetries of the physical point clouds they are increasingly used to generate. We study *soft* constraints (a symmetry-encouraging equivariance penalty, and a physics sum-rule residual) added to the training loss, and ask whether a constraint enforced during training governs the *samples*. It does, but only locally: a soft constraint lowers the target violation only at the sampling operating point—the (t, r) pair, and even the step count K—at which it was enforced, and relaxes elsewhere. Constraining the 1-NFE endpoint makes one-step samples equivariant ($16\times$ on $\mathrm{SO}(3)$ molecules, $\sim 600\times$ on $\mathrm{SO}(2)$ jets) while multi-step samples stay at the unconstrained level; constraining the sampler's velocity step does the reverse. A per-step audit locates the cause in a state mismatch, and a diffusion control isolates it. Enforcing the constraint at both operating points makes the generator equivariant at both ($81\times$ at 1-step and $3.1\times$ at 10-step on molecules; $96\times$ and $2.6\times$ on jets), an equal-budget ablation shows this is coverage rather than regularization strength, a relocation control moves the effect by moving the training step count, and marginalizing over K yields one model equivariant at every sampling budget. The same law governs a physics constraint, and we measure its boundary: under a non-compact Lorentz boost the fix holds at one step only. This is a mechanism paper; it makes no sample-fidelity claim.

1 INTRODUCTION

Generative models are increasingly used as fast surrogates for expensive physical simulation—jet generation for collider experiments, conformer generation for molecules—and the defining property of those systems is *symmetry*: a jet is physically unchanged by a Lorentz boost or a rotation about the beam axis, a molecule's energy by a global rotation. A faithful generator should be *equivariant*: rotating the latent noise should rotate the sample. Hard equivariance builds this into the architecture (Satorras et al., 2021; Hoogeboom et al., 2022; Gong et al., 2022; Spinner et al., 2024), guaranteeing the symmetry but constraining the model class. Soft equivariance adds a training penalty that encourages but does not guarantee it, leaving the architecture free (Finzi et al., 2021; Wang et al., 2022; Hebbar et al., 2025). Soft constraints are attractive precisely because they compose with any backbone and any other loss, including physics-informed residuals (Raissi et al., 2019; Bastek et al., 2025).

We study soft constraints on *MeanFlow* (Geng et al., 2025), a flow-matching variant (Lipman et al., 2023; Liu et al., 2023; Albergo et al., 2023) whose defining feature is that it learns an *average*-velocity field $u(z, t, r)$, so that a single function evaluation maps noise to data while the same network can also be run as a multi-step Euler sampler for higher fidelity. This dual mode—shared with consistency, shortcut, and flow-map models (Song et al., 2023; Kim et al., 2024; Frans et al., 2025; Boffi et al., 2024; Sabour et al., 2025)—makes a question first-class that is invisible for single-mode models: *a training-time constraint is enforced at some operating point; which operating points does it govern at sample time?*

The answer we find is that it governs only the one it was enforced at. We call this *operating-point locality*, and the paper establishes it, explains it, and prescribes the fix (Figure 1). Concretely:

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.

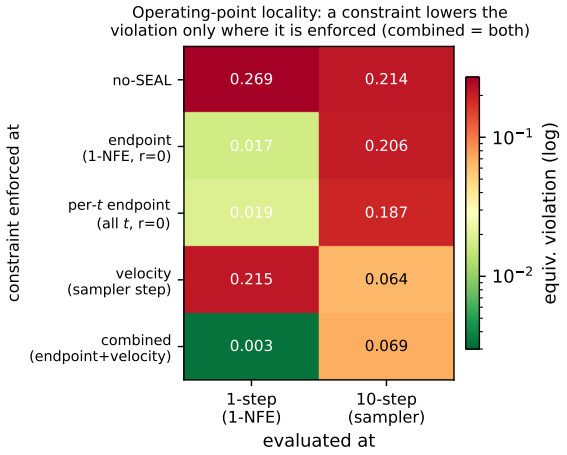


Figure 1: **Operating-point locality in one view** (SPICE $\mathrm{SO}(3)$, masked probe, seed means). A soft constraint lowers the equivariance violation only at the operating point it is *enforced* at: endpoint and per-t endpoint fix 1-step but not 10-step; velocity-at-the-step fixes 10-step but not 1-step; only the *combined* constraint is low in both columns.

- **The law.** A soft equivariance penalty on the 1-NFE endpoint makes one-step samples strongly equivariant but leaves multi-step samples at the unconstrained level; a penalty on the sampler's velocity step does the reverse. Sweeping the evaluation step count K, each constraint is low only at the K it was matched to, and a relocation control moves the low point by moving the training K (§4.1, §4.4).
- **The mechanism.** A per-step trajectory audit shows the penalized velocity is equivariant at the states the penalty sampled (Gaussian noise) and not at the evolved states the sampler visits, where the violation is $\sim 6\times$ larger. A diffusion model, whose denoiser is trained at all noise levels, is state-matched by construction and retains a $4.5\times$ multi-step benefit that MeanFlow does not; its residual erosion is measured exposure bias (§4.2).
- **The fix.** Enforcing the constraint at both the endpoint and the sampler-matched velocity step makes the generator equivariant at both operating points; an equal-budget ablation shows the gain is coverage, not penalty weight; and drawing the step count per sample from a set gives one model equivariant at every K (§4.3, §4.4).
- **Generality and boundary.** The same locality governs a physics sum-rule residual on jets, and holds under a compact ($\mathrm{SO}(3)$, molecules; $\mathrm{SO}(2)$, jets) and a non-compact (genuine $\mathrm{SO}(1,3)$ boost) group. The fix itself has a boundary: under the boost it holds at one step and not at ten, which we trace to the unbounded orbit (§4.5, §4.6).

Every table row is a mean over 3–5 seeds. We make no state-of-the-art sample-fidelity claim: the generators are deliberately plain, and the contribution is the law, its mechanism, and the enforcement recipe, which applies to any few-step generator regardless of its absolute fidelity.

2 Background and Method

MeanFlow. Given data x and noise $\epsilon \sim \mathcal{N}(0, I)$, define the interpolant $z_t = (1-t)x + t\epsilon$ and target velocity $v = \epsilon - x$. MeanFlow trains an average-velocity field $u(z, t, r)$ over $[r, t]$ via the identity $u_{\mathrm{tgt}} = v - (t - r)\,\partial_t u$, evaluated with a Jacobian–vector product (JVP). One-step generation is $f(z) = z - u(z, 1, 0)$; K-step generation applies $z_{k+1} = z_k - (t_k - r_k)\,u(z_k, t_k, r_k)$ on a grid $1 = t_0 > t_1 > \cdots > t_K = 0$ with $r_k = t_{k+1}$.

Equivariance and the penalty. For a group with Lie-algebra generators $\{L_a\}$ acting linearly on the feature axis, a map g is equivariant iff $\partial g / \partial z \cdot (L_a z) = L_a\, g(z)$. The deltaSEAL penalty (Hebbar et al., 2025) is the squared violation of this identity, evaluated by JVP (the Lie-derivative view

of Gruver et al. (2023)). Generators act on the kinematic block (4-momentum for jets, position for molecules) and trivially on the remaining features.

Three axes of a soft constraint. A soft equivariance penalty has three design choices that are usually conflated and, we will show, matter decisively for iterative samplers. *(1) Target*: *endpoint* SEAL constrains the 1-NFE map $f(z)$; *velocity* SEAL constrains $u(z, t, r)$ directly. *(2) Time*: for velocity SEAL, (t, r) can follow the MeanFlow training distribution or be *sampler-matched* to the Euler grid $\{(t_k, t_{k+1})\}$ the sampler uses. *(3) State*: the latent z at which the penalty is evaluated can be Gaussian noise, the interpolant z_t, or the sampler's own evolved state z_k.

Why state matters. If $u(\cdot, t, r)$ is equivariant *at the states the sampler visits*, linearity of the action R makes each Euler step commute with it, $g_k(Rz) = Rz - (t_k - r_k)u(Rz, t_k, r_k) = R\, g_k(z)$, and the composition is exactly equivariant for any K. The catch is the quantifier: a soft penalty enforces equivariance only in expectation over the distribution it samples, so that distribution must match the sampler's state distribution at the sampler's (t, r).

A multi-step equivariance probe. We measure equivariance of the *generator*, not the training loss. For a K-step sampler $\mathcal{S}_K$ and a random group element R, draw $z \sim \mathcal{N}(0, I)$ and report the relative violation $\|\mathcal{S}_K(Rz) - R\,\mathcal{S}_K(z)\|_2 / \|R\,\mathcal{S}_K(z)\|_2$, averaged over samples and group elements, on the kinematic block of real (unpadded) particles. $K = 1$ probes the 1-NFE map, $K = 10$ the multi-step sampler; the gap is our central diagnostic. For jets we probe beam-axis $\mathrm{SO}(2)$ rotations and a genuine $\mathrm{SO}(1, 3)$ boost (random direction, rapidity $\xi \sim \mathcal{U}(0, 0.3)$, metric-verified to 10^{-7}).

Physics residuals. To test whether the law extends beyond symmetry, we also study a PIDM-style (Bastek et al., 2025) *residual*: a penalty on the generated jet enforcing aggregate sum rules (total E, p_T, mass), and a two-point form matching the generated energy–energy correlator (EEC) (Basham et al., 1978; Komiske et al., 2023) to the real batch (Appendix B). Sample quality is evaluated with 1-Wasserstein distances on jet observables and the EEC distance.

3 Experimental setup

JetClass (Qu et al., 2022): $H \to b\bar{b}$ and hadronic-top jets, 14 per-particle features whose first four are $(E, p_x, p_y, p_z)/100$ GeV, Lorentz symmetry on that block, jet-level conditioning. *SPICE* (Eastman et al., 2023): small-molecule conformers, 17 per-atom features (positions, charge, element one-hots), $\mathrm{SO}(3)$ on the position block. The model is a Particle Transformer (Qu et al., 2022) in per-particle output mode wrapped for MeanFlow (sinusoidal t, r embeddings; BatchNorm replaced by Identity and attention forced to the math kernel so that forward-mode JVP is well defined; pairwise features disabled). All variants share the backbone and differ only in the loss. SPICE models use $\lambda = 10$ unless stated; JetClass endpoint models use $\lambda = 200$ because their normalized violation is far smaller. Every reported number is a mean $\pm$ std over independently trained seeds (3–5; counts in each table), probed with a fixed latent/rotation set so that seeds are the only source of variance.

4 Results

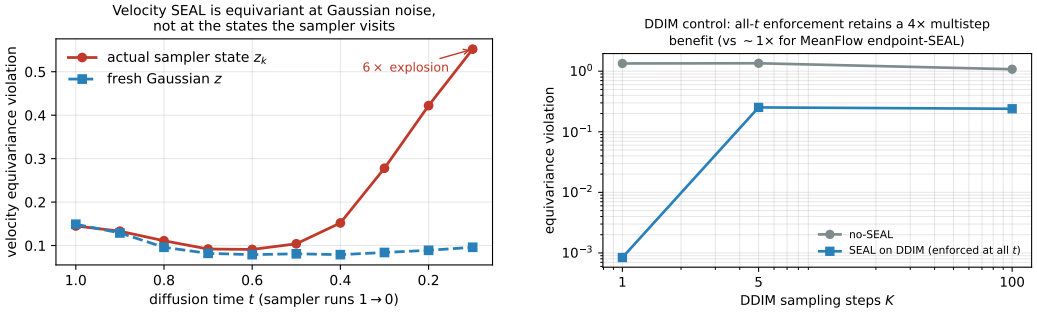
4.1 One-step equivariance, and its erosion under multi-step sampling

Constraining the 1-NFE map works, and it works in every domain (Table 1). On SPICE the one-step violation falls from 0.269 to 0.017 at $\lambda = 2$ and to 0.004 at $\lambda = 10$; on JetClass it falls from 0.230 to 0.0004 under beam-axis rotations and from 0.214 to 0.011 under a genuine Lorentz boost. Since one-step generation is MeanFlow's headline mode, this is a directly useful positive result: a soft penalty buys near-equivariant one-shot generation from a non-equivariant backbone. It also improves the rotation-sensitive observable most: the jet pseudorapidity 1-Wasserstein distance improves $2.7\times$ ($0.65 \to 0.24$) while the p_T scale, which the constraint does not touch, is unchanged (Appendix Figure 12).

The same constraint does *not* make multi-step samples equivariant. At $K = 10$ the SPICE violation is 0.206, a $12\times$ erosion from one step and indistinguishable from no-SEAL (0.214); JetClass shows

Table 1: Endpoint SEAL makes 1-NFE generation equivariant in every domain and group, and 10-step sampling erodes it back to the no-SEAL level in every case. Relative violation (lower is better), masked probe, mean ± std over seeds.

Domain (group)	variant	1-step	10-step	1-step reduction	seeds
SPICE (SO(3))	no-SEAL	0.269 ± 0.026	0.214 ± 0.012	—	4
	endpoint, λ=2	$\mathbf{0.017 \pm 0.016}$	0.206 ± 0.031	16×	3
	endpoint, λ=10	$\mathbf{0.004 \pm 0.001}$	0.196 ± 0.036	62×	3
JetClass (SO(2))	no-SEAL	0.230 ± 0.022	0.159 ± 0.013	—	5
	endpoint, λ=200	$\mathbf{0.0004 \pm 0.0003}$	0.158 ± 0.011	∼600×	3
JetClass (boost)	no-SEAL	0.214 ± 0.037	0.243 ± 0.038	—	5
	endpoint, λ=200	$\mathbf{0.011 \pm 0.009}$	0.272 ± 0.006	19×	3



(a) Per-step audit along the real sampler trajectory.

(b) Diffusion control: all-t enforcement retains a multi-step benefit.

Figure 2: **The mechanism.** (a) For a velocity-SEAL checkpoint trained at Gaussian states, the velocity-equivariance violation measured at the states the 10-step sampler *actually visits* (red) explodes ∼6× toward low t, while at fresh Gaussian states (blue) it stays flat: the penalty held exactly where it was sampled. (b) Adding the same penalty to a DDIM denoiser, which trains at all t, makes the constraint state-matched by construction and retains a $4.5\times$ benefit at K=100 within its own setup (toy SO(3) data, 3 seeds; the K=1 point is tautological and not used as evidence); MeanFlow endpoint-SEAL retains $\sim 1.15\times$ within its setup.

the same collapse ($0.0004 \rightarrow 0.158$ against 0.159). Naively constraining the velocity field on the sampler's time grid but at Gaussian states does not help either: the 10-step violation stays at 0.19–0.16 even as the *training* penalty is driven down three orders of magnitude (Appendix Figure 11). The training loss says the model is equivariant; the sampler says it is not.

4.2 The mechanism: a state mismatch

Why does driving the velocity-equivariance loss to 10^{-5} not yield equivariant samples? We audit a trained velocity-SEAL checkpoint, measuring the velocity violation at each state the 10-step sampler visits (Figure 2a). It is small at the Gaussian start (0.145) but grows $\sim 6\times$ along the trajectory to 0.55 at $t = 0.1$, whereas at fresh Gaussian states it is flat at ~ 0.09 for all t. The penalty made the velocity equivariant *where it sampled the constraint* and nowhere else; the sampler evaluates it on evolved states partway to the data manifold that the penalty never saw. This is not a failure of the commutation argument in §2 but of the distribution over which the soft penalty was enforced.

A controlled experiment isolates the cause from the loss form. Standard diffusion trains its denoiser at t drawn over the whole schedule, so an equivariance penalty on its x_0-estimate is state-matched by construction. On SO(3) hypersphere data, SEAL on a DDIM sampler (Song et al., 2021) retains a $4.5\times$ multi-step benefit at $K = 100$ (0.24 vs. 1.08), where MeanFlow endpoint-SEAL retains essentially none (Figure 2b; these are two within-setup ratios, not a cross-setup comparison). The diffusion constraint still erodes mildly with K, and a per-t audit shows why: the reverse-sampler states

Table 2: The fix (SPICE $\mathrm{SO}(3)$, masked probe, mean $\pm$ std over seeds; lower is better). Each single constraint helps only its own operating point; the combined constraint is equivariant at both.

SEAL variant (λ=10 unless noted)	1-step	10-step	seeds
no-SEAL	0.269 ± 0.026	0.214 ± 0.012	4
endpoint (λ=2)	0.017 ± 0.016	0.206 ± 0.031	3
per-t endpoint	0.019 ± 0.004	0.187 ± 0.047	3
velocity, sampler-matched, interpolant state	0.215 ± 0.011	0.064 ± 0.008	4
combined (endpoint + velocity)	$\mathbf{0.0033 \pm 0.0023}$	$\mathbf{0.069 \pm 0.005}$	3
Equal budget (total λ in parentheses)			
endpoint (20)	0.003 ± 0.002	0.207 ± 0.016	3
velocity (20)	0.191 ± 0.020	0.052 ± 0.010	3
combined 5+5 (10)	0.022 ± 0.029	0.073 ± 0.004	3

carry $2.0\times$ the violation of the forward-diffused training states (Appendix Figure 5), the exposure-bias gap (Ning et al., 2023; 2024; Li et al., 2024) measured on an equivariance observable. The unified statement is qualitative but robust: *a soft constraint transfers in proportion to the overlap between the states it is enforced on and the states the sampler visits.*

4.3 The fix: enforce the constraint at both operating points

The diagnosis prescribes the fix: enforce the constraint at the (t, r) and the states the sampler evaluates. Three candidates sharpen the law (Table 2). *(i) Velocity at the sampler step*: constraining $u(z_t, t, r)$ at the interpolant state with $r = t - 1/K$ recovers the 10-step violation ($0.214 \to 0.064$) but degrades the 1-step map (0.215), which is not its operating point. *(ii) Per-t endpoint*: constraining $\hat{x}(z_t, t) = z_t - t\,u(z_t, t, 0)$ across all t—the MeanFlow analogue of a diffusion x_0-penalty—does the opposite (0.019 / 0.187). Spanning all t is not enough, because it uses $r = 0$ rather than the sampler's small step: it is the (t, r) *pair* that must match, not the state alone. *(iii) Combined*: enforcing both in one loss makes the generator equivariant at both operating points at once, 0.0033 at 1-step ($81\times$) and 0.069 at 10-step ($3.1\times$).

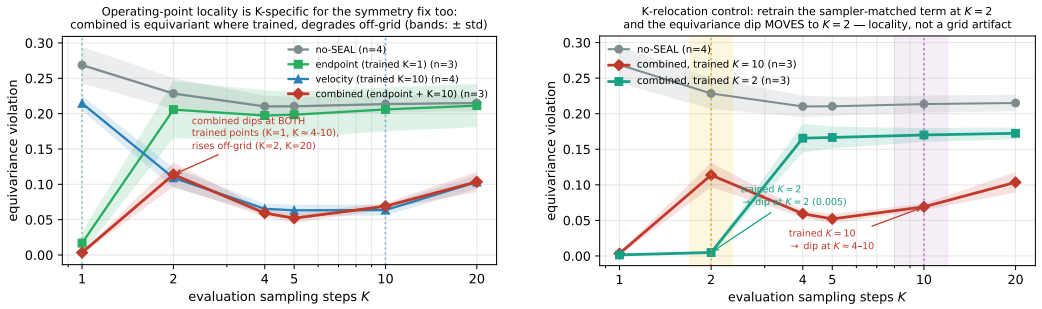
It is the combination, not the budget. Doubling a single constraint's weight does not buy the other operating point (endpoint $\lambda = 20$ still leaves 10-step at 0.21; velocity $\lambda = 20$ leaves 1-step at 0.19), whereas the combined constraint at *half* weight, the same total as a single $\lambda = 10$, reaches both (0.022 / 0.073). The benefit is covering both operating points.

Cross-domain. On JetClass the combined constraint reaches both operating points under $\mathrm{SO}(2)$: 0.0024 ± 0.0006 at 1-step and 0.060 ± 0.017 at 10-step against no-SEAL 0.230 / 0.159 ($96\times$ and $2.6\times$), and endpoint-only and velocity-only again each fix one point (0.0004 / 0.158 and 0.183 / 0.061; Appendix Figure 9).

4.4 Locality is specific to the sampler step count

Sweeping the evaluation step count $K \in \{1, 2, 4, 5, 10, 20\}$ resolves the law at finer grain than the 1-vs-10 summary (Figure 3a). The combined constraint is equivariant at both of its trained operating points, the $K = 1$ endpoint (0.003) and the neighbourhood of its trained $K = 10$ (0.052–0.069 for $K \in \{4, 5, 10\}$), yet rises to 0.114 at $K = 2$ and 0.104 at $K = 20$, step counts it never saw. The endpoint constraint dips only at $K = 1$, the velocity constraint only in the $K \approx 4$–10 band, and no-SEAL is flat and high. The constraint governs the specific sampler grid it was matched to.

A relocation control rules out a discretization artifact. Is $K = 2$ high because of locality, or because a coarse two-step grid is simply a harder state distribution? We falsify the latter: retraining the same combined loss with its sampler-matched term at $K = 2$ moves the dip (Figure 3b). The K=2-trained model reaches 0.005 at $K = 2$ (where the K=10 model sits at 0.114) and gives up $K \geq 4$ (0.170 at $K = 10$ against 0.069). Each model is lowest at its own trained step count; the curves cross. The low violation is a property of the enforced operating point, not of the Euler grid.



(a) Violation vs. evaluation step count K.

(b) Relocation control: the dip moves with the training K.

Figure 3: **Locality in** K (SPICE SO(3), masked probe, log-K axis, bands ± std over seeds). (a) Each constraint is low only at the step counts it was matched to: endpoint at K=1, velocity in the K≈4–10 band, combined at both; all rise at the off-grid K=2 and K=20. (b) Retraining the same combined loss with its sampler term at K=2 instead of K=10 moves the dip to K=2; each model is lowest at its own trained K and the curves cross.

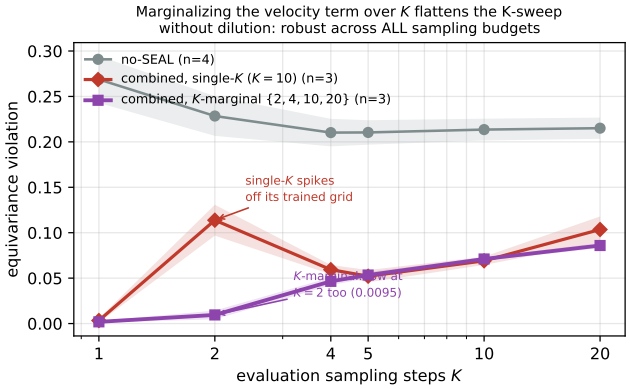


Figure 4: **The deployable recipe** (SPICE SO(3), 3 seeds, bands ± std). Drawing the sampler-matched term's K per sample from $\{2, 4, 10, 20\}$ (purple) flattens the sweep without dilution: it matches or beats the single-K combined fix (red) at every K.

The deployable recipe: marginalize over K**.** For a generator meant to run at any budget, we draw the sampler term's step count per sample from $\{2, 4, 10, 20\}$ each batch, endpoint term unchanged. One might fear dilution, but the opposite happens (Figure 4): the K-marginal model matches or beats the single-K fix at every step count and removes its off-grid spikes ($K = 2$: $0.114 \to 0.0095$; $K = 20$: $0.104 \to 0.086$; $K = 1$ improves to 0.0019). One model is equivariant at every sampling budget.

4.5 Physics constraints obey the same law

The law is not specific to symmetry. A sum-rule residual on total jet E/p_T/mass, enforced on the 1-NFE generation, cuts the jet-p_T Wasserstein distance from 392 to 163 GeV at one step and reverts to the no-SEAL level at every $K \geq 2$ (Table 3). Enforced instead on a $K = 4$ rollout it gives a $6.6\times$ improvement at exactly $K = 4$ and reverts at $K = 10$. Enforcing both broadens the improvement across the low-K range and removes the $K = 2$ spike that trips each single residual, but, consistent with locality, does not reach the off-grid $K = 10$. Two honest differences from the symmetry case: the K=4 residual already helps at $K = 1$, so the single-point dichotomy is sharper for equivariance than for an aggregate sum rule; and every residual variant improves p_T at the cost

Table 3: A physics sum-rule residual on JetClass jets is operating-point-local at the granularity of the step count. Jet-p_T 1-Wasserstein distance to real jets (GeV, lower is better) vs. evaluation step count K; the residual is enforced on the 1-NFE endpoint, on a $K{=}4$ differentiable rollout, or both (3 seeds for combined).

residual enforced at	$K{=}1$	$K{=}2$	$K{=}4$	$K{=}10$
none (no-SEAL)	392	400	416	411
endpoint	**163**	436	429	435
$K{=}4$ rollout	79	229	**63**	434
endpoint + $K{=}4$ rollout	124	124	93	322

of the jet-mass distribution (Appendix B). The transferable conclusion is the locality, which holds for both constraint families.

4.6 The boundary of the fix: a non-compact group

Under the genuine Lorentz boost the combined constraint nails one step (0.0057 ± 0.0008 against 0.214, $38\times$) but at ten steps its violation (0.229 ± 0.055) is statistically indistinguishable from no-SEAL (0.243 ± 0.038), and endpoint-only is slightly *worse* than no-SEAL there (0.272; Appendix Figure 9). The orbit is unbounded: binning the 10-step violation by rapidity, it grows $\sim 3.5\times$ across $\xi \in [0, 0.3]$ for every variant (no-SEAL $0.108 \to 0.385$, combined $0.098 \to 0.344$; the combined constraint sits $\sim 10\%$ below no-SEAL in every bin but within one standard deviation). The sampler is carried along the orbit to states outside the constrained distribution, and a soft penalty sampled over a bounded region cannot follow it. We report this as the boundary of the recipe rather than a partial success.

5 Related work

Few-step generators and where they are trained. MeanFlow (Geng et al., 2025), consistency models (Song et al., 2023; Kim et al., 2024), shortcut models (Frans et al., 2025), and flow-map matching (Boffi et al., 2024; Sabour et al., 2025) all learn a map indexed by a (t, r) pair, and progressive distillation (Salimans & Ho, 2022) and rectified flow (Liu et al., 2023) target one-step generation by re-training on the sampler's own outputs. That the states a sampler visits differ from the states a denoiser was trained on is the exposure-bias problem of diffusion (Ning et al., 2023; 2024; Li et al., 2024); we measure it on an equivariance observable and show that for dual-mode models the mismatch is operating-point-specific, down to the step count, which single-mode diffusion analyses do not expose.

Equivariance: hard, soft, and measured. Hard-equivariant architectures (Satorras et al., 2021; Hoogeboom et al., 2022; Xu et al., 2022; Jing et al., 2022; Gong et al., 2022; Bogatskiy et al., 2022; Spinner et al., 2024), frame averaging (Puny et al., 2022), and learned canonicalization (Kaba et al., 2023) guarantee symmetry by construction; soft approaches (Finzi et al., 2021; Wang et al., 2022; Hebbar et al., 2025) penalize its violation and leave the architecture free. Gruver et al. (2023) measure learned equivariance with the Lie derivative, the same quantity our penalty minimizes; our probe measures the equivariance of the *sampler*, and our contribution is that the two can disagree by an order of magnitude depending on the operating point. We do not benchmark against hard-equivariant conformer generators (Hassan et al., 2024): they are a different model class with a fidelity objective, whereas we study when a soft penalty transfers.

Physics-informed generation and HEP generative models. Physics-informed residuals (Raissi et al., 2019; Bastek et al., 2025; Shu et al., 2023) penalize a generated sample's violation of known laws; we show such residuals obey the same locality as symmetry penalties and extend them to a two-point EEC constraint (Basham et al., 1978; Komiske et al., 2023). Point-cloud generators for jets and showers (Kansal et al., 2021; Buhmann et al., 2023b; Leigh et al., 2024; Mikuni et al., 2023; Buhmann et al., 2023a; Krause et al., 2024) and their evaluation (Kansal et al., 2023; Komiske et al., 2018) motivate our domain; concurrent work finds the flow-matching loss an unreliable fidelity

indicator (JetPrism collaboration, 2026), which we corroborate (Appendix C) but do not make a contribution.

6 Discussion and limitations

The central message is a cautionary one for any soft regularizer in an iterative generator: *a soft penalty constrains only the distribution it is sampled over*. For a dual-mode model that distribution is a (t, r) pair and a state distribution, and for a multi-step sampler it must be the sampler's own evolving state at the sampler's own step, not the convenient Gaussian prior or the training-time schedule. The recipe that follows is cheap and general: enforce at the endpoint and at the sampler step, and marginalize the latter over the step counts you intend to use. It applies to any few-step generator, and to physics residuals as much as to symmetry, regardless of the generator's absolute fidelity; that is why the result matters even though our generators are deliberately plain.

Limitations. Our generators are not sample-quality competitive, by design, and we make no fidelity claim. The clean both-points fix is a compact-group result: under the non-compact boost it holds at one step only, and closing that gap likely needs a state distribution that follows the orbit (an open problem we state, not solve). The combined physics residual is only partially transferable and trades jet mass for p_T. Results are on two datasets and one backbone; the mechanism argument is architecture-agnostic, but the magnitudes are not.

7 Conclusion

A soft constraint on a few-step flow-matching generator is operating-point-local: it governs the samples at the (t, r) and step count where it was enforced and relaxes elsewhere, because it was enforced over a state distribution the sampler does not visit. We establish this for symmetry and physics constraints, in two domains, with a diffusion control that isolates the mechanism and a relocation control that moves the effect. Enforcing the constraint at both operating points fixes both, marginalizing over the step count fixes all of them, and the recipe has a measured boundary at non-compact groups.

AI use statement

Large language models (Anthropic Claude and OpenAI Codex, via their coding assistants) were used throughout this project as coding and writing assistants: to port and refactor training and evaluation code, to draft and condense text under the authors' direction, to act as adversarial reviewers of intermediate drafts, and to audit experiment scripts for correctness. All experimental design decisions, all numbers reported, and the final text were verified by the authors, who take full responsibility for the content.

Reproducibility statement

All code (the generative-training extension to an open jet-physics training framework, the loss with its target/time/state axes, the equivariance probe, samplers and metrics), data configurations, cluster launchers, per-seed result files for every table and figure, and the figure generator are provided in the supplementary material and will be released publicly. JetClass and SPICE are public datasets; the SPICE conversion script is included. Every reported number is a mean over independently seeded training runs, probed with a fixed latent and group-element set; seed counts are given in each table and figure.

References

Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.

Simon Axelrod and Rafael Gómez-Bombarelli. GEOM, energy-annotated molecular conformations for property prediction and molecular generation. *Scientific Data*, 9:185, 2022.

C. Louis Basham, Lowell S. Brown, Stephen D. Ellis, and Sherwin T. Love. Energy correlations in electron-positron annihilation: Testing quantum chromodynamics. *Physical Review Letters*, 41: 1585, 1978.

Jan-Hendrik Bastek, WaiChing Sun, and Dennis M. Kochmann. Physics-informed diffusion models. In *International Conference on Learning Representations (ICLR)*, 2025.

Nicholas M. Boffi, Michael S. Albergo, and Eric Vanden-Eijnden. Flow map matching. *arXiv preprint arXiv:2406.07507*, 2024.

Alexander Bogatskiy, Timothy Hoffman, David W. Miller, and Jan T. Offermann. PELICAN: Permutation equivariant and Lorentz invariant or covariant aggregator network for particle physics. *arXiv preprint arXiv:2211.00454*, 2022.

Erik Buhmann, Cedric Ewen, Darius A. Faroughy, Tobias Golling, Gregor Kasieczka, Matthew Leigh, Guillaume Quétant, John Andrew Raine, Debajyoti Sengupta, and David Shih. EPiC-ly fast particle cloud generation with flow-matching and diffusion. *arXiv preprint arXiv:2310.00049*, 2023a.

Erik Buhmann, Gregor Kasieczka, and Jesse Thaler. EPiC-GAN: Equivariant point cloud generation for particle jets. *SciPost Physics*, 15:130, 2023b.

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*, 10:11, 2023.

Marc Finzi, Gregory Benton, and Andrew Gordon Wilson. Residual pathway priors for soft equivariance constraints. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One step diffusion via shortcut models. In *International Conference on Learning Representations (ICLR)*, 2025.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J. Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. *arXiv preprint arXiv:2505.13447*, 2025.

Shiqi Gong, Qi Meng, Jue Zhang, Huilin Qu, Congqiao Li, Sitian Qian, Weitao Du, Zhi-Ming Ma, and Tie-Yan Liu. An efficient Lorentz equivariant graph neural network for jet tagging. *Journal of High Energy Physics*, 2022(7):30, 2022.

Nate Gruver, Marc Finzi, Micah Goldblum, and Andrew Gordon Wilson. The lie derivative for measuring learned equivariance. In *International Conference on Learning Representations (ICLR)*, 2023.

Majdi Hassan, Nikhil Shenoy, Jungyoon Lee, Hannes Stärk, Stephan Thiede, and Dominique Beaini. ET-Flow: Equivariant flow-matching for molecular conformer generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Pradyun Hebbar, Thandikire Madula, Vinicius Mikuni, Benjamin Nachman, Nadav Outmezguine, and Inbar Savoray. SEAL: A symmetry-encouraging loss for high energy physics. *arXiv preprint arXiv:2511.01982*, 2025.

Emiel Hoogeboom, Víctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3D. In *International Conference on Machine Learning (ICML)*, 2022.

JetPrism collaboration. JetPrism: Diagnosing convergence for generative simulation and inverse problems in nuclear physics. *arXiv preprint arXiv:2604.01313*, 2026.

Bowen Jing, Gabriele Corso, Jeffrey Chang, Regina Barzilay, and Tommi Jaakkola. Torsional diffusion for molecular conformer generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Sékou-Oumar Kaba, Arnab Kumar Mondal, Yan Zhang, Yoshua Bengio, and Siamak Ravanbakhsh. Equivariance with learned canonicalization functions. In *International Conference on Machine Learning (ICML)*, 2023.

Raghav Kansal, Javier Duarte, Hao Su, Breno Orzari, Thiago Tomei, Maurizio Pierini, Mary Touranakou, Jean-Roch Vlimant, and Dimitrios Gunopulos. Particle cloud generation with message passing generative adversarial networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Raghav Kansal, Anni Li, Javier Duarte, Nadezda Chernyavskaya, Maurizio Pierini, Breno Orzari, and Thiago Tomei. Evaluating generative models in high energy physics. *Physical Review D*, 107:076017, 2023.

Dongjun Kim, Chieh-Hsin Lai, Wei-Hsiang Liao, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yutong He, Yuki Mitsufuji, and Stefano Ermon. Consistency trajectory models: Learning probability flow ODE trajectory of diffusion. In *International Conference on Learning Representations (ICLR)*, 2024.

Patrick T. Komiske, Eric M. Metodiev, and Jesse Thaler. Energy flow polynomials: A complete linear basis for jet substructure. *Journal of High Energy Physics*, 2018(4):13, 2018.

Patrick T. Komiske, Ian Moult, Jesse Thaler, and Hua Xing Zhu. Analyzing N-point energy correlators inside jets with CMS open data. *Physical Review Letters*, 130:051901, 2023.

Claudius Krause et al. CaloChallenge 2022: A community challenge for fast calorimeter simulation. *arXiv preprint arXiv:2410.21611*, 2024.

Matthew Leigh, Debajyoti Sengupta, Guillaume Quétant, John Andrew Raine, Knut Zoch, and Tobias Golling. PC-JeDi: Diffusion for particle cloud generation in high energy physics. *SciPost Physics*, 16:018, 2024.

Mingxiao Li, Tingyu Qu, Ruicong Yao, Wei Sun, and Marie-Francine Moens. Alleviating exposure bias in diffusion models through sampling with shifted time steps. In *International Conference on Learning Representations (ICLR)*, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023.

Vinicius Mikuni, Benjamin Nachman, and Mariel Pettee. Fast point cloud generation with diffusion models in high energy physics. *Physical Review D*, 108:036025, 2023.

Mang Ning, Enver Sangineto, Angelo Porrello, Simone Calderara, and Rita Cucchiara. Input perturbation reduces exposure bias in diffusion models. In *International Conference on Machine Learning (ICML)*, 2023.

Mang Ning, Mingxiao Li, Jianlin Su, Albert Ali Salah, and Itir Onal Ertugrul. Elucidating the exposure bias in diffusion models. In *International Conference on Learning Representations (ICLR)*, 2024.

Omri Puny, Matan Atzmon, Heli Ben-Hamu, Ishan Misra, Aditya Grover, Edward J. Smith, and Yaron Lipman. Frame averaging for invariant and equivariant network design. In *International Conference on Learning Representations (ICLR)*, 2022.

Huilin Qu, Congqiao Li, and Sitian Qian. Particle transformer for jet tagging. In *International Conference on Machine Learning (ICML)*, 2022.

Maziar Raissi, Paris Perdikaris, and George E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.

Amirmojtaba Sabour, Sanja Fidler, and Karsten Kreis. Align your flow: Scaling continuous-time flow map distillation. *arXiv preprint arXiv:2506.14603*, 2025.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations (ICLR)*, 2022.

Víctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In *International Conference on Machine Learning (ICML)*, 2021.

Dule Shu, Zijie Li, and Amir Barati Farimani. A physics-informed diffusion model for high-fidelity flow field reconstruction. *Journal of Computational Physics*, 478:111972, 2023.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning (ICML)*, 2023.

Jonas Spinner, Victor Bresó, Pim de Haan, Tilman Plehn, Jesse Thaler, and Johann Brehmer. Lorentz-equivariant geometric algebra transformers for high-energy physics. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Rui Wang, Robin Walters, and Rose Yu. Approximately equivariant networks for imperfectly symmetric dynamics. In *International Conference on Machine Learning (ICML)*, 2022.

Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. GeoDiff: A geometric diffusion model for molecular conformation generation. In *International Conference on Learning Representations (ICLR)*, 2022.

A DIFFUSION CONTROL AND MEASURED EXPOSURE BIAS

If the state mismatch is the cause, a model that enforces equivariance across all noise levels should largely escape the erosion. We add a SEAL penalty on the x_0-prediction at the sampled t of a DDIM model (Song et al., 2021) on $\mathrm{SO}(d)$-symmetric hypersphere data ($d = 3$, 3 seeds). The honest evidence is the within-DDIM retention ratio read along a single column: the $K = 1$ number is near-zero but tautological (a one-step DDIM returns exactly the x_0-map that the penalty constrains) and is not used. At $K = 100$ SEAL-on-DDIM retains a $4.5\times$ benefit (0.24 vs. 1.08), whereas MeanFlow endpoint-SEAL retains $\sim 1.15\times$ at $K = 10$ within its setup. The residual erosion has a measurable source: per diffusion time t, the x_0-map violation on the reverse-DDIM trajectory (what generation visits) is $2.0\times$ that on the forward-diffused training states (what the denoiser saw), up to $3.7\times$ (Figure 5). This is exposure bias measured on an equivariance observable, and it is exactly the gap the sampler-matched fix is designed to close.

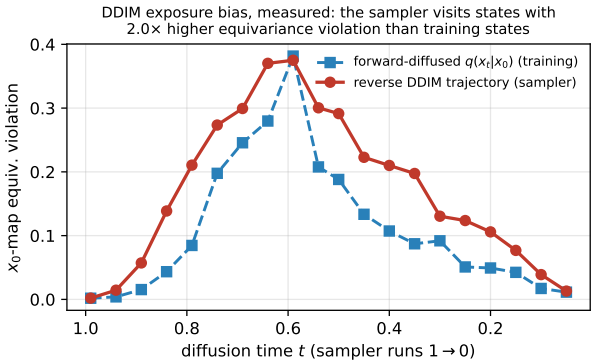


Figure 5: Measured exposure bias in the diffusion control: per-t x_0-map equivariance violation on the reverse-DDIM trajectory vs. on forward-diffused training states.

B PHYSICS-INFORMED RESIDUALS AND THE EEC

Vanilla MeanFlow reproduces *aggregate* jet observables poorly (Figure 7): the generated jet energy, transverse momentum and mass all peak too low (jet-p_T 1-Wasserstein $\sim$400 GeV; the mass shows a spurious peak at zero rather than the W/top scale), and the EEC distance is 1.05 of a maximum ~ 2. The per-particle marginals are far better reproduced than these sums: the model learns local structure but not global sum rules, which is what a physics residual targets.

One-point residual. We penalize the generated jet's total E, p_T and mass against targets derived from the per-event conditioning. Enforced on the 1-NFE map it improves the jet-p_T distance from 392 to 163 GeV at one step and does not transfer to any $K \geq 2$ (Table 3). Enforced on a $K = 4$ differentiable Euler rollout it improves $K = 4$ by $6.6\times$ ($416 \to 63$) and reverts at $K = 10$ (Figure 6). Enforced on both (3 seeds) it is most robust across low K ($124/124/93$ at $K = 1/2/4$) with no $K = 2$ spike, but does not reach $K = 10$ (322 vs. 411). All residual variants raise the jet-mass distance from ~ 85 to ~ 120 GeV, a sum-rule tension we do not claim to resolve. A methodological note: an early version of this residual inverted the input standardization and trained toward targets compressed toward the mean; the corrected residual tracks per-event conditioning.

Two-point residual (EEC). The energy–energy correlator $\mathrm{EEC}(\Delta R) = \sum_{i<j} E_i E_j \, \delta(\Delta R - \Delta R_{ij}) / (\sum_i E_i)^2$ is a differentiable function of the generated point cloud; we bin it softly and penalize the L^1 distance of the generated histogram to the real batch, and use the same distance as a metric. It is implemented and validated (the distance is zero for real-vs-real batches) but we report no trained EEC-residual result.

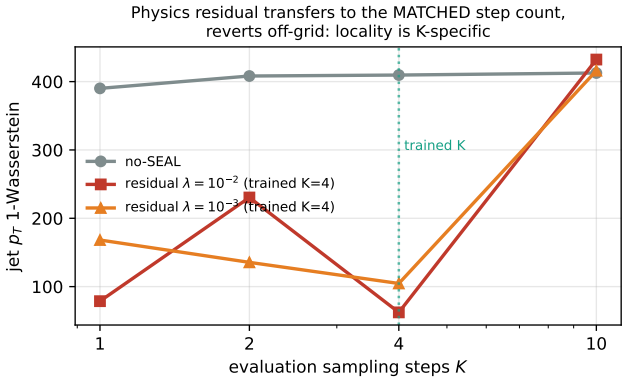


Figure 6: The physics residual is operating-point-local at the granularity of the step count: trained with a $K{=}4$ rollout it gives a $6.6\times$ jet-p_T improvement at $K{=}4$ and reverts to the no-SEAL level at $K{=}10$.

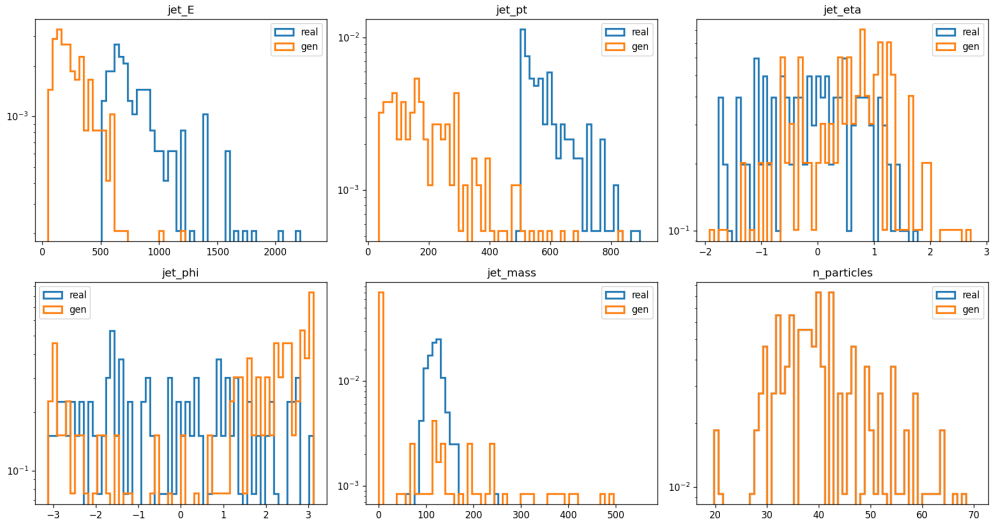


Figure 7: Generated (orange) vs. real (blue) jet observables for vanilla MeanFlow on JetClass. Local structure is roughly captured; aggregate E, p_T and mass are not. (n-particles is set by the conditioning mask and matches by construction.)

C Validation loss is a poor fidelity proxy and mis-selects checkpoints

A data-scaling study of plain MeanFlow on JetClass finds the validation loss nearly flat across a $25\times$ range of training samples (best 0.428 ± 0.002) and a $9\times$ range of model size. Sharper than "the loss does not improve": selecting checkpoints by validation loss actively mis-picks. Sweeping every saved epoch of each run and comparing the best-observable epoch (the quality frontier) against the loss-selected epoch, the frontier wins on all nine runs by 12–30% in jet-p_T Wasserstein (Figure 8). All comparisons in the paper therefore use quality-selected checkpoints, and equivariance is probed at the same checkpoints.

D On the use of SPICE, and a reverse direction

We use SPICE as a second symmetry group ($\mathrm{SO}(3)$) to test whether the findings are domain-general, and they are. SPICE (Eastman et al., 2023) is a dataset of quantum-chemical energies and forces

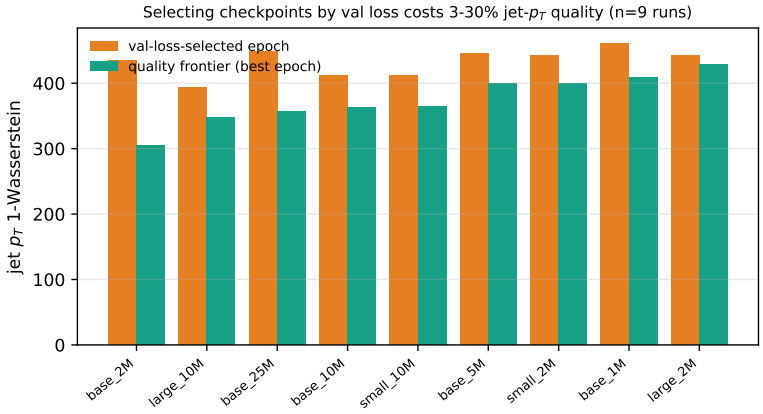


Figure 8: Validation loss mis-selects checkpoints: loss-selected epoch (orange) vs. best-over-epochs quality frontier (teal) in jet-p_T Wasserstein, nine runs.

built to train interatomic potentials; its standard evaluation is energy/force accuracy, not generative fidelity. The standard conformer-generation benchmark is GEOM (Axelrod & Gómez-Bombarelli, 2022), scored by coverage and average-minimum-RMSD, where strong equivariant baselines exist (Xu et al., 2022; Jing et al., 2022; Hassan et al., 2024). Our SPICE generator is a from-scratch pilot and is not competitive as a conformer generator; its per-atom marginals, in particular the discrete element one-hots, are visibly imperfect. The molecular results concern equivariance *transfer*, a property of the generator map that is meaningful regardless of absolute sample quality. The natural follow-up inverts the direction: SPICE's energies and forces become the *constraint* of a physics-informed generator, with a residual $\|\nabla E(\hat{x})\|^2$ under a differentiable energy model guiding generation and SPICE evaluating the energy of the generated conformers. We leave this to future work.

E Additional figures and tables

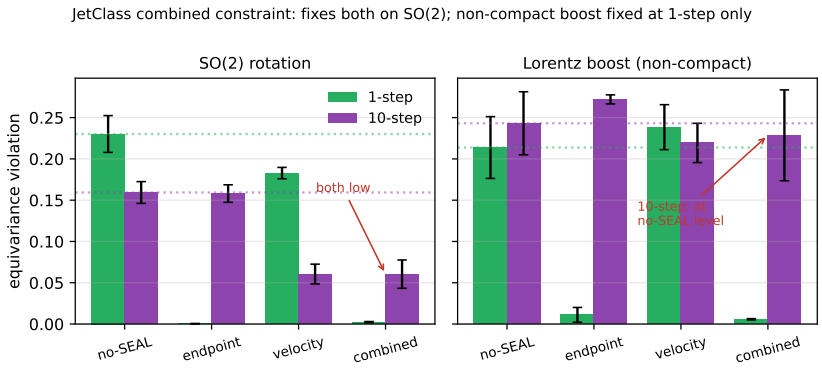


Figure 9: JetClass combined constraint, 1-step vs. 10-step (masked probe; dotted lines mark no-SEAL; bars seed means, whiskers std; no-SEAL 5 seeds, others 3). On $\mathrm{SO}(2)$ the combined constraint reaches both operating points (left); under the non-compact boost it nails 1-step but the 10-step violation stays at the no-SEAL level (right), the boundary of the fix.

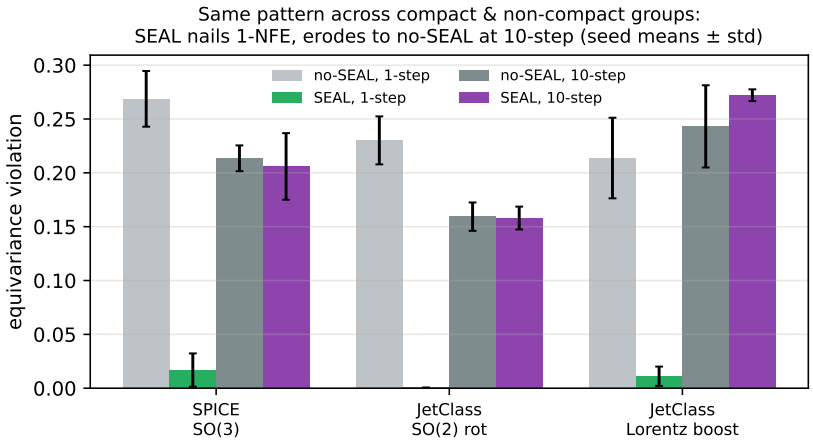


Figure 10: The 1-NFE-equivariant / multi-step-eroding pattern is group-general: $\mathrm{SO}(3)$ (molecules), beam-axis $\mathrm{SO}(2)$ (jets) and a genuine $\mathrm{SO}(1,3)$ boost (jets); seed means $\pm$ std.

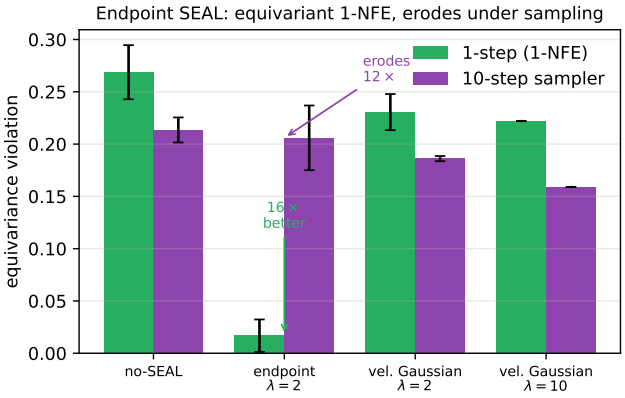


Figure 11: Endpoint SEAL reduces the 1-step violation $16\times$ and erodes $12\times$ under 10-step sampling; velocity SEAL at Gaussian states helps at neither (SPICE $\mathrm{SO}(3)$, masked probe, seed means $\pm$ std).

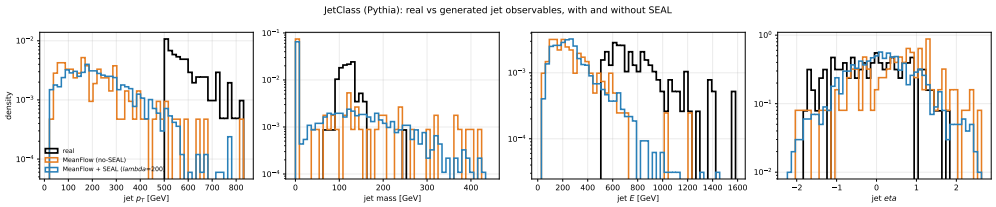


Figure 12: JetClass generated vs. real jet observables with and without endpoint SEAL (λ=200). SEAL most improves the angular observable η (Wasserstein $0.65 \rightarrow 0.24$) and jet mass ($106 \rightarrow 89$ GeV); the aggregate p_T/energy scale remains mis-modeled by both.

Table 4: Per-Euler-step velocity-equivariance violation along the actual sampler trajectory vs. at fresh Gaussian states, for a velocity-SEAL (sampler-matched, Gaussian-state) SPICE checkpoint.

step k	0	4	6	7	8	9
t_k	1.0	0.6	0.4	0.3	0.2	0.1
violation at trajectory z_k	0.145	0.091	0.152	0.278	0.422	**0.552**
violation at Gaussian z	0.149	0.079	0.079	0.084	0.089	0.096

Table 5: Early, unmasked, single-seed SPICE probe (its scale differs from the masked seed-averaged probe used elsewhere), kept for the training-penalty column: the training loss is driven to 10^{-5} while the 10-step sampler stays non-equivariant.

SEAL variant	1-step	10-step	training penalty (final)
no-SEAL	0.307	0.254	—
endpoint (λ=2)	**0.017**	0.220	5×10^{-6}
velocity, sampler-matched, Gaussian (λ=2)	0.255	0.206	1.4×10^{-4}
velocity, sampler-matched, Gaussian (λ=10)	0.235	0.172	4.0×10^{-5}
velocity, state-matched (λ=10)	0.230	**0.076**	8×10^{-4}