THE SYMMETRY STRESS-TEST: VIOLATION-AWARE BENCHMARKING FOR CALIBRATED EQUIVARIANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Approximate symmetry is by now uncontroversial: gravity, contacts, boundaries, and sensor pipelines routinely break the groups our models assume, and soft or relaxed equivariance can absorb the mismatch. What the field cannot yet do is *measure* when a symmetry prior helps, hurts, or is untested, because aggregate IID accuracy hides all three. We make it measurable with *violation-aware benchmarking*: sweep a controlled symmetry-breaking parameter, verify the parameter is a valid proxy for violation (monotone, stable, and predictive of the strict-versus-calibrated gap), and account for total cost. The three steps fit in a one-paragraph *Symmetry Stress-Test Card* that a reviewer can require. Across three settings, the same recipe exposes the same pattern. Our primary setting is N-body dynamics under an external field, which contracts $O(3)$ to a horizontal subgroup, so the violation comes from the environment rather than the labels. There a strictly equivariant EGNN degrades from median MSE 0.006 to 1.56 as the field grows, while an otherwise identical calibrated model holds at 0.006: a $\mathbf{249}\times$ gap for $1.05\times$ the cost. Rotation augmentation *backfires*, re-imposing the very symmetry the field removed, and the strict model catches up only when handed the field axis, which the calibrated model never needs. Two synthetic sweeps reproduce the pattern across group and output type ($O(3)$ tensor, $\mathbf{134}\times$; $O(5)$ scalar, $\sim\mathbf{1046}\times$), and calibration keeps equivariance's data efficiency where the symmetry holds. The lesson is conditional: hard-code symmetry where the pipeline preserves it, calibrate where it does not, and report the card so the difference is visible rather than assumed.

1 INTRODUCTION

Symmetry has powered many of the field's largest jumps in geometric and graph machine learning. In vision, CNNs that embed translation equivariance displaced handcrafted features; in geometric deep learning, SE(3)- or E(n)-equivariant GNNs such as Cormorant (Anderson et al., 2019), TFN (Thomas et al., 2018), SE(3)-Transformer (Fuchs et al., 2020), EGNN (Satorras et al., 2022), and NequIP (Batzner et al., 2022) set accuracy records on molecular, physical simulation, and robotics benchmarks. Theory explains these gains: imposing a symmetry group shrinks the hypothesis space, and universality results show that little expressive power is lost (Keriven & Peyré, 2019; Finzi et al., 2021; Bronstein et al., 2021). More recently, large models trained at scale, not all of them equivariant, have begun to set the state of the art in materials discovery and weather forecasting benchmarks (Merchant et al., 2023; Lam et al., 2023).

When both a strictly equivariant model and a large unconstrained one top the same leaderboard, aggregate accuracy can no longer tell us why: whether the symmetry prior is doing real work, whether scale is just papering over a misspecified prior, or whether the benchmark never probes the regime where the prior would break. The field is left arguing "physics vs. scale" with no instrument to settle it. That symmetries are often only approximate, and that soft or relaxed equivariance can help, is by now widely accepted (Wang et al., 2023; Kim et al., 2023; Finzi et al., 2021); the open problem is how to *measure* when it matters. **We supply that instrument: a falsifiable, standardized test of when a symmetry prior helps, and evidence that it exposes regime-dependent behavior that aggregate accuracy conceals.**

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.

The prevailing approach, choosing a group G and enforcing G-equivariance everywhere, is brittle outside idealized settings (Wang et al., 2023; Kaba & Ravanbakhsh, 2024). The fixed-group paradigm (Finzi et al., 2021) implicitly assumes that the chosen symmetry is (a) known a priori, (b) globally valid, and (c) preserved throughout the data pipeline and environment. These assumptions fail in structured, predictable ways (Section 2): for example, gravity contracts $\mathrm{SE}(3)$ to a subgroup; contact events introduce localized anisotropy; and sensor pipelines can break symmetry even when the underlying physical law remains symmetric. Recent architectural advances (Abramson et al., 2024; Kim et al., 2023) have shifted toward softer symmetry constraints, yet papers are still evaluated primarily on aggregate IID accuracy on idealized benchmarks, which says nothing about violation regimes or computational cost. We close this gap with three contributions.

(1) A three-axis organization of the calibration methods: We organize the existing "soft symmetry" methods into three mechanisms, each with a distinct evidence standard: *restricted discovery* of transformation generators (M1), *relaxation* of equivariance via soft penalties or gates (M2), and *localization* of symmetry across space, time, or node roles (M3). We do not claim these mechanisms as new; the contribution is mapping each to the evidence that would validate it. M1 also covers symmetries that are latent, visible only in a learned representation (Park et al., 2022). We do not propose relearning permutation equivariance, which message passing already provides; what we calibrate is the continuous geometric symmetry, that is, how strongly and where to enforce it.

(2) Violation-aware evaluation standards and a Symmetry Stress-Test Card: We define three reporting standards (a validated violation proxy with stratified reporting (S1), controlled violation sweeps (S2), and full cost accounting (S3)) and consolidate them into a compact *Symmetry Stress-Test Card*. The card turns "symmetry helps" into a falsifiable, context-dependent claim rather than an assumed prior.

(3) New evidence that the card surfaces what aggregate benchmarks hide: Running the protocol across three settings (Section 5) we find the same signature every time: as the violation grows, the strict model stalls at a *bias floor*, an error level its hypothesis class cannot go below, while its calibrated twin holds at negligible extra cost. In the primary setting, N-body dynamics under an external field, the violation is *physical* rather than label-induced, and there we also run the controls a skeptic would ask for. Augmentation over the full rotation group *backfires* as the field strengthens, augmentation over the correct subgroup stops the damage but stays $2.5\times$ behind calibration, and a strict model catches up with calibration only once it is given the field axis. The gap is the price of assuming the wrong group, which is what the card is built to expose. Two further settings vary the group and the output type, so the effect is not an artifact of one task.

Our claim is conditional and falsifiable. Strict equivariance keeps its sample-efficiency advantage in low-violation regimes; calibrated models are more accurate at similar cost in high-violation regimes. We do not claim calibration is universally better: Card 2 (Appendix F) shows why NequIP's strict $E(3)$ prior suits MD-17, whose pipeline preserves the symmetry. Sections 2–4 develop the failure modes, the calibration framework, and the card; Sections 5 and 6 report the evaluation and weigh alternative views.

2 Failure Modes and Violation Measures

The fixed-group paradigm commits to one group G (e.g., $\mathrm{SE}(3)$) enforced everywhere. This works in idealized regimes, but in real scientific and engineering pipelines the relevant symmetry is often *conditional*: symmetries in data are frequently only approximate due to measurement noise or external effects (Kim et al., 2023). In physical settings, environmental fields such as gravity can break full $\mathrm{E}(3)/\mathrm{SE}(3)$ assumptions, effectively contracting the applicable symmetry to a weaker subgroup (Han et al., 2022). Symmetry can also hold only on certain substructures (e.g., local/cluster symmetries in otherwise asymmetric graphs) (Cen et al., 2024), or be altered by boundary conditions that induce discrete point-group symmetries (subgroups of $O(n)$) (Zheng et al., 2024).

We use *symmetry* to mean equivariance under a specified group G acting through a representation ρ: $f(g{\cdot}x){=}\rho(g)f(x)$ for all $g{\in}G$ (invariance is the special case $\rho{\equiv}\mathrm{id}$). Except for discovery (M1), which learns the group itself through its Lie algebra $\mathfrak{g} \subset \mathfrak{gl}(n)$, "calibrating symmetry" means adjusting the strength or scope of G-equivariance for a fixed G. A natural measure of how far a

predictor f deviates from G-equivariance is the *equivariance residual*

$$\Delta_g(x) \;=\; \|f(g{\cdot}x) - \rho(g)f(x)\|, \tag{1}$$

with ρ acting on the output space. When $\Delta_g(x) \approx 0$ uniformly, strict equivariance is warranted; when $\Delta_g(x)$ is large and structured, calibrated enforcement is preferable.

In physical domains, symmetry is often correct in principle but violated in practice by the environment or the measurement pipeline. These violations are *structured*, concentrated in specific regimes rather than spread i.i.d., and each maps onto one assumption of the fixed-group paradigm. Assumption (a), the right group, fails under environment fields: a system that is $\mathrm{SE}(3)$-symmetric in free space contracts to a horizontal subgroup under gravity, and a model locked to the full group cannot use the gravity axis as a cue, so it settles onto a bias floor (Li et al., 2024; Fuchs et al., 2020). Assumption (b), global validity, fails at contacts and other discontinuities, where friction cones, impacts, and switching dynamics inject anisotropy exactly where a globally strict model has no way to turn equivariance off. Assumption (c), pipeline preservation, fails through the measurement itself: camera-frame biases, anisotropic noise, and occlusions break the symmetry of the data even when the underlying law keeps it, and a strict model averages these effects away rather than compensating for them (McClellan et al., 2025; Kim et al., 2023).

A related failure is *scope mismatch*: even the correct group may not hold uniformly across time, space, or node type. Contact graphs come and go, so the relevant equivariance shifts with the dynamical regime; and while message passing is permutation-equivariant over node indices, the weight sharing one wants depends on role, since a bottleneck node or a distinct agent type is not interchangeable with the rest. The response in both cases is to localize or gate enforcement, that is, to learn *partial* exchangeability rather than impose a global symmetry (Huang et al., 2023). Enforcing the full symmetry regardless is not just inaccurate but wasteful, since the spherical harmonics and Clebsch–Gordan products that implement it are expensive (Frank et al., 2024; Wang et al., 2024; Duval et al., 2023); comparisons must therefore count total cost (training, inference, augmentation) (Brehmer et al., 2025). This applies only when symmetry is a modeling choice; a physical law such as energy conservation or gauge invariance must be enforced whatever it costs.

Each failure mode names the metric that would catch it: a controlled violation sweep for group misspecification, stratified reporting for localized breakage, and cost accounting for enforcing a prior where it does not apply. Section 4 turns these into reporting standards; Section 5 shows the sweep exposing a $134\times$ to $\sim 1046\times$ median gap that aggregate benchmarks hide.

3 The Inductive Bias Trilemma

Whether to impose, learn, or relax a symmetry is not a free choice among equals: the three options trade off against one another. Concretely, a model rarely secures all three of the following at once.

1. *Data efficiency:* strong generalization from limited data.
2. *Robustness under violations:* high performance on high-violation strata and along violation sweeps (Section 4).
3. *Design simplicity:* low architectural and tuning overhead (few extra hyperparameters; easy deployment).

Three architectural regimes, used as running labels throughout, each buy two of these axes at the expense of the third. Strict equivariance (A) hard-codes the group: it is data-efficient (1) and simple (3), but it gives up robustness (2), because its error stalls at a bias floor the moment the assumed symmetry is violated (Section 5). Unconstrained scale and augmentation (C) sit at the opposite corner: robust (2) and simple (3) once enough data and compute are available, but data-*inefficient* (1), since every invariance must be relearned from examples. Calibrated symmetry (B) is the interior point that tries to keep (1) and (2) together by making the strength and scope of equivariance learnable, paying only in simplicity (3) through the extra gating or localization machinery. No regime holds all three, so the choice is really which axis to give up. Our position is that in scientific and engineering settings, where data are scarce *and* violations are structured, (3) is the one to spend: a little design complexity buys both data efficiency and robustness. We pair each calibration mechanism below with the evidence standard (Section 4) that would show it works.

(M1) Discovery, learning *what* symmetry applies: Instead of hard-coding a group G, discovery parameterizes the symmetry through learnable generators, e.g., a Lie-algebra basis (Yang et al., 2023; Finzi et al., 2020). This matters when the transformation family is unknown, varies across regimes, or is task-dependent. Because unconstrained discovery can collapse or become ill-posed, it should be restricted to a stable hypothesis class (e.g., learning $\mathfrak{g} \subset \mathfrak{gl}(n)$ with regularization) (Finzi et al., 2020). Discovery recovers *data efficiency (1)* by finding the right hypothesis class but raises tuning overhead (3), and should be validated by controlled transformation tests and equivariance residuals, not aggregate IID accuracy.

(M2) Relaxation, learning *how strongly* to enforce symmetry: Relaxation treats equivariance as a soft constraint, so the model can leave the bias floor when the symmetry is violated, gaining robustness (2). One form interpolates via a learnable gate $\alpha \in [0, 1]$:

$$h' = \alpha\, f_{\mathrm{eq}}(h) + (1 - \alpha)\, \mathrm{Proj}_{\rho}\big(f_{\mathrm{free}}(h)\big), \tag{2}$$

where α is the relaxation coefficient and Proj_{ρ} enforces type consistency even when full equivariance is relaxed. If the output decomposes into isotypic blocks $\rho = \bigoplus_k \rho_k$ (scalars, vectors, higher-order tensors), Proj_{ρ} is the block-diagonal orthogonal projection onto that type structure, so the free branch can change a component's *value* but not its representation type (a scalar cannot leak into a vector channel). Alternatives to Eq. 2 include soft penalties on the ℓ_2-distance of weights to the equivariant subspace (Kim et al., 2023); residual pathway priors (RPP), which add a free residual layer to each equivariant layer and regularize its magnitude (Finzi et al., 2021), the mechanism our experiments use; and layer-wise learned equivariance via differentiable marginal-likelihood optimization (van der Ouderaa et al., 2023). The gate α may be global, per-layer, per-node, or conditioned on a violation proxy $\hat{v}(x)$, a measured quantity that tracks how badly the symmetry is broken (Section 4). Relaxation improves *robustness (2)* at modest cost to *simplicity (3)*, while preserving most of the *data efficiency (1)* benefit through the equivariant branch, and should be evaluated with violation sweeps or stratified reporting plus total-cost accounting.

(M3) Localization, learning *where and when* symmetry applies: Localization makes the scope of symmetry conditional, applying constraints only to certain subgraphs, time windows, or node roles. *Regional localization* (McClellan et al., 2025) uses a spatially-conditioned α that reduces equivariance in boundary regions where symmetry is broken (e.g., road edges or safety zones). *Cluster-based localization* (Huang et al., 2023) scopes symmetry to subgraph clusters whose automorphisms define a local exchangeability group, rather than global S_n interchangeability. For discrete structure this is again *partial exchangeability* (role/type-conditioned weight sharing), not full permutation symmetry. Localization addresses both *data efficiency (1)* (correct prior where valid) and *robustness (2)* (no enforcement where invalid), at the cost of conditioning mechanisms (3), and is validated by showing gains concentrate on high-violation strata and by ablating the conditioning.

Whether to calibrate at all comes down to two questions. If the output type or a physical law fixes the symmetry (forces must rotate, energies must be invariant), enforce it strictly; there is nothing to calibrate. Otherwise, ask whether the violation is *structured enough to learn from*. If a proxy $\hat{v}(x)$ tracks it, condition relaxation or localization on that proxy and report stratified metrics; if the violation is unstructured noise with no usable proxy, strict equivariance is the safer default and calibration should be added only with care. We make this concrete in Section 5.6, where removing the structure from our violation collapses the calibration advantage almost entirely.

4 Evidence and Evaluation Standards

The framework of Section 3 makes a falsifiable prediction, that strict models keep their edge at low violation and lose it at high violation, but a single aggregate accuracy number cannot check it: the same number can hide graceful degradation, a bias floor, or a task the benchmark never stressed. We therefore ask for three kinds of evidence, which together separate these cases.

The first is a *violation proxy* and stratified reporting (S1). One picks a benchmark-specific quantity $\hat{v}(x)$ that tracks how badly the symmetry is broken, such as distance-to-boundary, heterophily, anisotropy, or a controlled knob λ, and reports error stratified by it, so a gain is credited to the high-violation strata where it occurs rather than diluted into the symmetric core. A proxy is only worth reporting if it is *monotone* (larger $\hat{v}$ really does mean more violation), *stable* (unmoved by nuisance variation unrelated to the symmetry), and *predictive* (correlated with the strict-versus-calibrated

gap). The three are jointly necessary, not a menu: a monotone and stable but non-predictive proxy establishes that a violation exists without showing it matters for the model. We verify three real-dataset proxies against all three criteria in Appendix G, and show in Section 5 that our knob λ qualifies (Spearman r=0.996 on N-body, 0.966 on Inertia, 0.97 on O5Synthetic).

The second is a *controlled sweep* (S2). Where a continuous symmetry-breaking parameter can be constructed, plotting error against it reads off the failure signature directly: a calibrated model bends smoothly, a misspecified strict model bias-floors or collapses, and the two curves diverge in a way a single IID point cannot express. We apply this reading retrospectively to PEnGUiN, NequIP, and PER in Appendix F.

The third is *total-cost accounting* (S3). A claim that scale or augmentation "matches equivariance" means little unless it is made at equal total cost, counting training, inference, and the extra epochs or passes augmentation consumes, because in the low-data regime augmentation tends to close the accuracy gap only by spending far more compute (Brehmer et al., 2025; Gerken et al., 2022).

Table 3 (Appendix C) summarizes five studies comparing strict (A), calibrated (B), and unconstrained (C) baselines in one pipeline: strict wins when symmetry holds, calibrated when it is systematically violated, and none runs a controlled sweep (S2) or full cost accounting (S3).

4.1 The Symmetry Stress-Test Card

The card condenses S1–S3 into a minimal reporting template that reviewers can require as a paragraph or a short table (Table 1); without all five fields, robustness and efficiency conclusions should be read as preliminary.

Table 1: The Symmetry Stress-Test Card. A compliant submission provides all five fields. Missing fields indicate gaps in evidence, not model flaws.

Field	**Required content**
Assumed group & representation	The group G and representation ρ enforced or relaxed by the architecture.
Calibration mechanism	Which mechanism is used: Discovery (M1), Relaxation (M2), Localization (M3), or none (strict baseline).
Baselines	Results for (A) strictly equivariant, (B) calibrated/relaxed, and (C) unconstrained baselines under the same pipeline.
Violation evidence	Either a sweep over a violation parameter λ (S2), or test metrics stratified by a violation proxy $\hat{v}(x)$ (S1).
Total-cost reporting	Training compute (FLOPs or wall-clock), inference cost, and augmentation cost (extra epochs or passes) (S3).

These fields are diagnostic. A strict model reported on QM9 or MD-17 in isolation fills the group and mechanism fields but not baselines, violation evidence, or cost; the card marks these as gaps in *evidence*, not model flaws, so robustness claims beyond the symmetric regime stay unsupported.

5 Experiments

We fill the card from controlled violation sweeps in *three* settings chosen to vary the group, the output geometry, and the source of the violation. The **primary** setting (Section 5.2) is N-body motion under an external field, where the violation comes from the environment rather than the labels and where we also run augmentation and correct-subgroup controls; two synthetic tasks then test generality across group and output type (Section 5.4). Every setting uses the identical protocol: the three core models (A, B, C) of Section 3, one continuous knob $\lambda \in [0, 1]$ swept over 11 levels at 10 seeds, and medians with interquartile ranges (IQR) throughout.

5.1 Setup: Three Settings, One Protocol

In every setting model (A) is a strictly equivariant backbone. Model (B) is its calibrated counterpart: it adds a free, non-equivariant residual pathway with strong weight decay on the free weights (RPP-style), so it defaults to equivariant behavior when the prior is correct but can deviate when the data demand. Model (C) is unconstrained. On the synthetic tasks these are EMLP, MixedEMLP, and MLP (Finzi et al., 2021); on N-body they are EGNN, a MixedEGNN that adds one weight-decayed free MLP at the output, and a plain GNN (Satorras et al., 2022) (Appendix H gives architectures and hyperparameters).

The primary task puts five charged particles in 3D under a Coulomb force: from their initial positions, velocities, and charges the model predicts where they land after T integration steps. Without the field the dynamics are $O(3)$-equivariant; we break them with a uniform external field of magnitude $\lambda\bar{s}$ along a fixed axis $\hat{z}$, where $\bar{s}$ is a fixed per-task scale (Appendix H). The field contracts the effective symmetry to $SO(2)_{\hat{z}}$, the horizontal subgroup and exactly the gravity example of Section 2, and λ sets its strength.

The two synthetic tasks inject the violation into the labels instead. The moment-of-inertia task (Finzi et al., 2021) maps five point masses to the 3×3 inertia matrix, an $O(3)$-equivariant *tensor*, and we break it to $O^x(2)$, the transformations that fix the axis $\hat{x}$, by

$$\tilde{y} \;=\; y \;-\; \lambda \cdot \bar{s} \cdot \hat{x}\hat{x}^\top, \tag{3}$$

matching stratum 2 of Kim et al. (2023). O5Synthetic maps two 5-vectors to an $O(5)$-*invariant scalar*, broken to $O(4)$ by subtracting $\lambda\bar{s}\,(x_1{\cdot}e_1)$. Across the three the group ($SO(2)_{\hat{z}}$, $O(3)$, $O(5)$), the output (dynamics, tensor, scalar), and the source of the violation (a physical field or a label perturbation) all change, while the protocol does not.

5.2 Primary Result: N-body Under an External Field

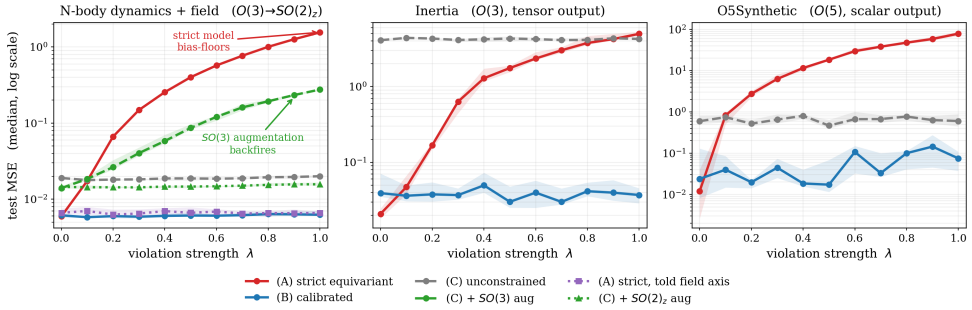

Figure 1: **Violation sweep across three settings** (median over 10 seeds, shaded = IQR, log y-axis). *Left:* N-body + field (primary), with the controls of Section 5.2: rotation augmentation over $SO(3)$ (C+aug) and over $SO(2)_{\hat{z}}$, and a strict EGNN given the field axis, which tracks B. *Middle:* Inertia ($O(3)$ tensor). *Right:* O5Synthetic ($O(5)$ scalar). In all three, strict (A) degrades monotonically to a bias floor while calibrated (B) stays low. On N-body, $SO(3)$ augmentation *backfires*: it imposes the very symmetry the field breaks.

Figure 1 (left) shows the result. At zero field (λ=0) the strict EGNN (A) is best (median MSE 0.006), confirming equivariance is the right prior when it holds; the calibrated MixedEGNN (B) matches it (0.006) because weight decay keeps its free pathway near zero. As the field grows, A climbs monotonically to 1.56 at λ=1. The reason is structural, not a matter of capacity: an $O(3)$-equivariant map commutes with every rotation, so it cannot single out the fixed axis $\hat{z}$ the field selects, and the field's contribution, a common drift along $\hat{z}$, is outside its hypothesis class. B, whose free pathway can express that direction, stays flat at $\approx$0.006 throughout: a $\mathbf{249}\times$ median advantage at full field for $1.05\times$ A's training cost. The gap grows with λ (Spearman r=0.996, $p<10^{-100}$), so λ is a valid proxy for it.

The same setting lets us test the usual rejoinder that augmentation makes architectural symmetry unnecessary. The plain GNN trained with random-rotation augmentation for the same number of

epochs (C+aug) improves on plain C at λ=0 (0.014 vs. 0.019), as expected when the symmetry holds. As the field grows, though, it degrades to 0.28 at λ=1, 14× *worse* than the un-augmented GNN and 45× worse than B, for the reason that sinks the strict model: it teaches the model to treat all rotations alike, although the field makes them differ. Restricting augmentation to the subgroup the field preserves, rotations about $\hat{z}$ (C+aug$_{\hat{z}}$), removes the damage but not the gap. It stays near 0.015 across the sweep, above B in all 110 runs and 2.5× behind it at λ=1.

The sharper control asks whether calibration beats strict equivariance itself or only strict equivariance to the *wrong* group. We give the EGNN the axis $\hat{z}$ as an extra equivariant input, the standard way to break a symmetry through the input (Kaba & Ravanbakhsh, 2024); the resulting model (A$_{\hat{z}}$) is exactly equivariant to rotations about $\hat{z}$ (residual $<10^{-6}$). It tracks B at every λ (medians within 3–21%) and is indistinguishable from it at λ=1 (0.0066 vs. 0.0062, paired Wilcoxon p=1.0), at 1.15× A's training cost against 1.06× for B in the same timing run. This supports the conditional claim rather than undercutting it. Once the breaking direction is known, hard-coding the correct subgroup is the right call, and the 249× gap is the price of the wrong one; B pays nothing for not knowing the axis, and telling the two situations apart is what the sweep and the card are for.

5.3 Mechanism: Equivariance Residuals

To see *why* calibration holds up, we measure the equivariance residual $\Delta_g(x) = \|f(g{\cdot}x) - \rho(g)f(x)\|$ (Eq. 1) over 20 random group elements (10 on N-body) and all test samples, per model and λ (for an invariant target this reduces to $|f(g{\cdot}x) - f(x)|$). Figure 2 shows the same story in all three settings. Strict (A) holds a near-zero residual regardless of λ (N-body: $\sim 10^{-6}$): the hard constraint forbids any deviation even when the task is no longer symmetric, so it is equivariant but wrong. Calibrated (B) instead raises its residual monotonically with λ, learning to break the symmetry in the direction the data demand while its MSE stays flat. That A's residual stays near zero while its MSE explodes points to the misspecified prior rather than limited capacity, and a parameter-matched ablation confirms it (Appendix D): a deeper, larger strict model does not close the gap and trains unstably.

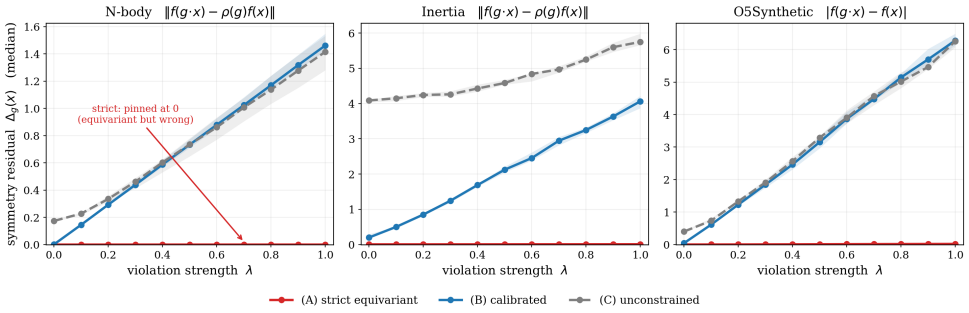

Figure 2: **Mechanism across three settings** (median over 10 seeds, shaded = IQR; panels ordered N-body, Inertia, O5 as in Figure 1). Strict (A) stays pinned near zero (forced symmetric, unable to adapt) while its MSE explodes; calibrated (B) raises its residual with λ while keeping MSE low. Where full relaxation is optimal (O5 and N-body), B's residual approaches the unconstrained model's.

5.4 Generality: Tensor and Scalar Synthetic Tasks

The two synthetic sweeps (Figure 1, middle and right) reproduce the pattern under a different group and output type, with the violation injected into labels rather than the environment. On **Inertia** ($O(3)$ tensor), strict EMLP is best at λ=0 (median MSE 0.02) and rises to a bias floor of 4.97 at λ=1, while MixedEMLP holds 0.037 (IQR $[0.029, 0.045]$): a **134**× gap; A crosses above B near $\lambda \approx 0.1$, a clean violation threshold. On **O5Synthetic** ($O(5)$ scalar) the effect is sharper: A climbs to 78.4 while B holds 0.075, a $\sim$**1046**× gap. On both tasks λ satisfies the S1 desiderata (r=0.97), and B costs 1.03–1.04× the strict training time.

Two caveats sharpen rather than weaken the conclusion. First, on O5 the unconstrained MLP is a non-trivial competitor (it learns the smooth scalar map once the prior is wrong), though its median never beats B; the stress-test thus *surfaces* that the right fallback baseline is task-dependent, which aggregate accuracy hides. Second, on O5 the calibrated free pathway occasionally diverges (4 of 110 runs reach NaN or $\sim 10^{10}$), an M2 instability we return to in Section 6; this is why we report medians and IQR throughout.

5.5 Sample Efficiency

Does calibration sacrifice the data efficiency that motivates equivariance? No. Sweeping the N-body training-set size from 200 to 5000 samples at $\lambda{=}0$ (Appendix B), the calibrated model (B) tracks the strict model (A) exactly, and both sit 3–7$\times$ below the unconstrained GNN (C) at every size: (B) inherits equivariance's data efficiency where the symmetry holds and adds robustness where it does not (Sections 5.2, 5.4).

5.6 Robustness Checks: Proxy Validity and Violation Structure

Two checks confirm the sweep measures what we claim. The first asks whether λ is a legitimate proxy in the sense S1 requires, that is, monotone, stable, and predictive (Section 4). It is: the strict-versus-calibrated gap ($\mathrm{MSE}_A - \mathrm{MSE}_B$) climbs monotonically from ≈ 0 at $\lambda{=}0$ to its maximum at $\lambda{=}1$, its IQR stays bounded relative to the median, and it is strongly rank-correlated with λ (Spearman $r{=}0.97$, $p{=}1.8{\times}10^{-65}$ over all 110 per-seed points on Inertia; $r{=}0.97$ over the 106 non-divergent runs on O5), so λ tracks the part of the violation that matters to the model (Appendix A).

The second check asks whether calibration's advantage really comes from the *structure* of the violation rather than its magnitude alone. When we redraw the breaking axis uniformly at random for each sample, leaving no fixed direction to exploit, the Inertia B/A gap collapses from $134\times$ to $\approx 1.7\times$ at $\lambda{=}1$ (median MSE $\approx 5.0/3.0/6.4$ for A/B/C). With no fixed direction to learn, strict equivariance is the safer default and calibration buys almost nothing, as the framework predicts: the gain scales with how structured the violation is.

5.7 Cost Accounting (S3) Results

Table 2: **Total-cost accounting** (per-run training time, mean $\pm$ s.d.; median MSE at full violation $\lambda{=}1$, IQR in brackets). Calibration adds about 5% or less training overhead while cutting error by two to three orders of magnitude. †N-body controls, timed in a separate 3-seed run at $\lambda{=}1$ (s.d. ≤ 0.1 s) with cost relative to A in that run; their MSE is over the full 10-seed sweep.

Setting	Model	Time / run (s)	Cost vs. A	MSE at $\lambda{=}1$
N-body + field	(A) EGNN — strict	91.8 ± 1.5	$1.00\times$	1.56 $[1.54, 1.59]$
	(B) MixedEGNN — calib.	96.8 ± 2.4	$1.05\times$	0.0062 $[0.0061, 0.0071]$
	(C) GNN — unconstr.	52.3 ± 1.1	$0.57\times$	0.020 $[0.019, 0.021]$
	($\mathrm{A}_{\hat{z}}$) EGNN + $\hat{z}$ input†	106.6	$1.15\times$	0.0066 $[0.0063, 0.0069]$
	(C+aug) $SO(3)$ aug.†	66.8	$0.72\times$	0.276 $[0.267, 0.282]$
	(C+aug$_{\hat{z}}$) $SO(2)_{\hat{z}}$ aug.†	55.4	$0.60\times$	0.016 $[0.015, 0.017]$
Inertia	(A) EMLP — strict	56.4 ± 2.2	$1.00\times$	4.97 $[4.74, 5.30]$
	(B) MixedEMLP — calib.	58.5 ± 1.8	$1.04\times$	0.037 $[0.029, 0.045]$
	(C) MLP — unconstr.	4.9 ± 1.1	$0.09\times$	4.25 $[4.16, 4.66]$
O5Synthetic	(A) EMLP — strict	48.6 ± 0.8	$1.00\times$	78.4 $[69.3, 87.6]$
	(B) MixedEMLP — calib.	50.1 ± 0.8	$1.03\times$	0.075 $[0.037, 0.108]$
	(C) MLP — unconstr.	4.0 ± 0.2	$0.08\times$	0.60 $[0.45, 0.88]$

Table 2 reports the cost accounting required by standard S3 across all three settings. Calibration (B) adds only a lightweight free pathway, costing 1.03–$1.05\times$ the strict model's training time, yet cuts median MSE at full violation by $249\times$ (N-body), $134\times$ (Inertia), and $\sim 1046\times$ (O5). At inference B runs A's forward pass plus its free branch (one MLP on N-body, one dense branch per layer in MixedEMLP). The only N-body model that ties B, $\mathrm{A}_{\hat{z}}$, costs more and needs the axis; the augmented GNNs are cheaper but 2.5–$45\times$ worse. On Inertia the parameter overhead is modest (B $918{,}450$ vs.

A 572,121), and the parameter-matched ablation of Appendix D rules out capacity as the explanation. Each setting fills all five card fields (group, mechanism, A/B/C baselines, λ sweep, cost); the completed card for Inertia appears in Table 4 (Appendix F).

6 Discussion: Alternative Views

We address four standpoints that might argue calibration is unnecessary, each with the empirical discriminator that would settle it; Appendix E gives the extended treatment.

"Physics is known, so hard-code it." When the pipeline preserves the symmetry, strict equivariance is correct (our $\lambda{=}0$ result; NequIP, Card 2), and so is strict equivariance to the right subgroup once the breaking axis is known ($\mathrm{A}_{\hat{z}}$). The problem is misspecification: "physics is $\mathrm{SE}(3)$" does not make the *dataset* $\mathrm{SE}(3)$-consistent after boundaries, constraints, or sensor artifacts, and the sweep shape is what reveals it.

"Scale and augmentation are enough." On N-body at matched epochs (Section 5.2), full-rotation augmentation *degrades* as the field grows ($14\times$ worse than none at $\lambda{=}1$) because it enforces the broken symmetry, and even augmentation over the correct subgroup stays $2.5\times$ behind calibration. Augmentation helps only as far as its group is right, and closing the gap to equivariance takes extra compute (Brehmer et al., 2025).

"Invariants are safer than equivariants." Invariants suit scalar targets but are lossy for geometric outputs. This axis is orthogonal to calibration, which relaxes *either* an invariant or an equivariant prior; we stress-test both output types (O5 scalar, Inertia tensor) and calibration helps in both.

"Learning symmetry is too hard." This targets *discovery* (M1), which can be ill-posed (Goodfellow et al., 2014; Yang et al., 2023; Ahuja et al., 2021; Schmidt et al., 2025; Karjol et al., 2023). We use *relaxation* (M2), a regularized free pathway with no symmetry search, stable on all 10 Inertia seeds for $\lambda{\geq}0.1$ at $1.04\times$ cost; the O5 divergences (Section 5.4) mark where it needs care. Discovery, if used, should be restricted to a small generator set $\mathfrak{g} \subset \mathfrak{gl}(n)$ (Finzi et al., 2020).

Our study has two main limitations. It exercises a single calibration mechanism, M2 relaxation through residual pathways, and all three settings are controlled simulations rather than experimental data, even though the N-body violation is physically induced. Extending the protocol to the M1 and M3 mechanisms and to a real measured pipeline, such as MD-17 under an applied field, is left to future work.

7 Conclusion

Strict equivariance is the right prior exactly when the pipeline preserves the group it assumes; the difficulty is that real pipelines rarely do, and aggregate IID accuracy cannot tell us when they stop. Our answer is an instrument, not another architecture: a violation sweep, a proxy-validity check, and total-cost accounting, packaged as a one-paragraph Stress-Test Card. Across three settings, the same recipe gives the same answer. On N-body dynamics under an external field, a strict model degrades by $249\times$ while its calibrated twin does not, for $1.05\times$ the cost, and rotation augmentation makes things *worse* because it re-imposes the broken symmetry. The synthetic sweeps repeat the effect across group and output type ($134\times$ on an $O(3)$ tensor task, $\sim 1046\times$ on an $O(5)$ scalar task), and calibration gives up none of equivariance's data efficiency when the symmetry holds.

The controls also show what the gap measures. A strict model told the field axis catches up with the calibrated one, so the $249\times$ gap is the cost of assuming the wrong group, not of equivariance itself. When the breaking axis is instead redrawn at random for every sample, the advantage shrinks from $134\times$ to about $1.7\times$: calibration pays off only when the violation has structure to learn. The prescription is therefore deliberately conditional. Hard-code the group when the pipeline preserves it or the breaking direction is known, calibrate when the violation is structured but unmodeled, and fall back to strict equivariance when it is unstructured noise. The natural next test is a measured pipeline, such as MD-17 under an applied field, where the violation is not of our own making. Reporting the card is what makes these regimes visible, so that "symmetry helps" becomes a claim one can check rather than one we assume.

AI USE STATEMENT

In this work, we used generative AI tools to help implement and debug parts of the experiment, analysis, and plotting code, to check reported numbers against the raw result files, and to edit parts of the manuscript for clarity and consistency. We did not use generative AI tools to generate the research idea, the evaluation protocol, or the main experiments, which the authors conceived and implemented. We have reviewed all AI-assisted work: the authors read and revised every AI-suggested edit, inspected all AI-assisted code, and confirmed that every reported number can be recomputed from the logged results. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All three settings are simulated and fully specified in Section 5 and Appendix H, which lists architectures, optimizers, weight decays, dataset sizes, simulator constants, hardware, and the seed protocol. Every sweep result is a median (with IQR) over 10 seeds per violation level, computed from per-run logs; timings are means over runs, and the ablations state their own seed counts; the N-body pipeline is deterministic given the seed (a rerun of one baseline reproduced all 110 original runs bit for bit). Anonymized code for every sweep, control, and figure, together with the raw per-run results behind every table and figure, is included in the supplementary material; its README maps each result in the paper to the command that produces it.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I. Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B. Fuchs, Hannah Gladman, Rishub Jain, Yousuf A. Khan, Caroline M. R. Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michal Zielinski, Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M. Jumper. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, Jun 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07487-w. URL https://doi.org/10.1038/s41586-024-07487-w.

Kartik Ahuja, Jason Hartford, and Yoshua Bengio. Properties from mechanisms: An equivariance perspective on identifiable representation learning, 2021. URL https://arxiv.org/abs/2110.15796.

Brandon Anderson, Truong-Son Hy, and Risi Kondor. Cormorant: Covariant molecular neural networks, 2019. URL https://arxiv.org/abs/1906.04015.

Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P. Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E. Smidt, and Boris Kozinsky. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications*, 13 (1):2453, May 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-29939-5. URL https://doi.org/10.1038/s41467-022-29939-5.

Johann Brehmer, Sönke Behrends, Pim de Haan, and Taco Cohen. Does equivariance matter at scale?, 2025. URL https://arxiv.org/abs/2410.23179.

Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges, 2021. URL https://arxiv.org/abs/2104.13478.

Jiacheng Cen, Anyi Li, Ning Lin, Yuxiang Ren, Zihe Wang, and Wenbing Huang. Are high-degree representations really unnecessary in equivariant graph neural networks? In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in*

Neural Information Processing Systems, volume 37, pp. 26238–26266. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/2e4531a45b99f61947b23ccdd608303b-Paper-Conference.pdf.

Junwoo Chang, Minwoo Park, Joohwan Seo, Roberto Horowitz, Jongmin Lee, and Jongeun Choi. Partially equivariant reinforcement learning in symmetry-breaking environments, 2025. URL https://arxiv.org/abs/2512.00915.

Alexandre Duval, Victor Schmidt, Alex Hernandez Garcia, et al. Faenet: Frame averaging equivariant gnn for materials modeling. In *International Conference on Machine Learning (ICML)*, 2023. URL https://proceedings.mlr.press/v202/duval23a/duval23a.pdf.

Marc Finzi, Samuel Stanton, Pavel Izmailov, and Andrew Gordon Wilson. Generalizing convolutional neural networks for equivariance to lie groups on arbitrary continuous data. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Marc Finzi, Max Welling, and Andrew Gordon Gordon Wilson. A practical method for constructing equivariant multilayer perceptrons for arbitrary matrix groups. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3318–3328. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/finzi21a.html.

J. T. Frank et al. A euclidean transformer for fast and stable machine learned force fields. *Nature Communications*, 2024. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC11303804/.

Fabian Fuchs, Daniel Worrall, Volker Fischer, and Max Welling. Se(3)-transformers: 3d roto-translation equivariant attention networks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1970–1981. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/15231a7ce4ba789d13b722cc5c955834-Paper.pdf.

Jan E. Gerken, Oscar Carlsson, Hampus Linander, Fredrik Ohlsson, Christoffer Petersson, and Daniel Persson. Equivariance versus augmentation for spherical images, 2022. URL https://arxiv.org/abs/2202.03990.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. URL https://arxiv.org/abs/1406.2661.

Jiaqi Han, Wenbing Huang, Tingyang Xu, and Yu Rong. Equivariant graph hierarchy-based neural networks, 2022. URL https://arxiv.org/abs/2202.10643.

Ningyuan Huang, Ron Levie, and Soledad Villar. Approximately equivariant graph networks, 2023. URL https://arxiv.org/abs/2308.10436.

Sékou-Oumar Kaba and Siamak Ravanbakhsh. Symmetry breaking and equivariant neural networks, 2024. URL https://arxiv.org/abs/2312.09016.

Pavan Karjol, Rohan Kashyap, Aditya Gopalan, and Prathosh A. P. A unified framework for discovering discrete symmetries, 2023. URL https://arxiv.org/abs/2309.02898.

Nicolas Keriven and Gabriel Peyré. Universal invariant and equivariant graph neural networks, 2019. URL https://arxiv.org/abs/1905.04943.

Hyunsu Kim, Hyungi Lee, Hongseok Yang, and Juho Lee. Regularizing towards soft equivariance under mixed symmetries. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, and Peter Battaglia. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023. doi: 10.1126/science.adi2336. URL https://www.science.org/doi/abs/10.1126/science.adi2336.

Bohao Li, Bowen Du, Junchen Ye, Jiajing Huang, Leilei Sun, and Jinyan Feng. Learning solid dynamics with graph neural network. *Information Sciences*, 676:120791, 2024. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2024.120791. URL https://www.sciencedirect.com/science/article/pii/S0020025524007059.

Joshua McClellan, Greyson Brothers, Furong Huang, and Pratap Tokekar. Penguin: Partially equivariant graph neural networks for sample efficient marl, 2025. URL https://arxiv.org/abs/2503.15615.

Amil Merchant, Simon Batzner, Samuel S. Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85, Dec 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06735-9. URL https://doi.org/10.1038/s41586-023-06735-9.

Jung Yeon Park, Ondrej Biza, Linfeng Zhao, Jan-Willem van de Meent, and Robin Walters. Learning symmetric representations for equivariant world models, 2022. URL https://openreview.net/forum?id=D637S6zBRLD.

Victor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks, 2022. URL https://arxiv.org/abs/2102.09844.

Tobias Schmidt, Steffen Schneider, and Matthias Bethge. Equivariance by contrast: Identifiable equivariant embeddings from unlabeled finite group actions. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=kvI0QTVRQD.

Nathaniel Thomas, Tess E. Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation- and translation-equivariant neural networks for 3d point clouds. *CoRR*, abs/1802.08219, 2018. URL http://arxiv.org/abs/1802.08219.

Tycho F.A. van der Ouderaa, Alexander Immer, and Mark van der Wilk. Learning layer-wise equivariances automatically using gradients. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=bNIHdyunFC.

Dian Wang, Jung Yeon Park, Neel Sortur, Lawson L. S. Wong, Robin Walters, and Robert Platt. The surprising effectiveness of equivariant models in domains with latent symmetry, 2023. URL https://arxiv.org/abs/2211.09231.

Y. Wang et al. Enhancing geometric representations for molecules with equivariant vector-scalar interactive message passing. *Nature Communications*, 2024. URL https://www.nature.com/articles/s41467-023-43720-2.

Jianke Yang, Robin Walters, Nima Dehmamy, and Rose Yu. Generative adversarial symmetry discovery. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Zinan Zheng, Yang Liu, Jia Li, Jianhua Yao, and Yu Rong. Relaxing continuous constraints of equivariant graph neural networks for broad physical dynamics learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, pp. 4548–4558, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704901. doi: 10.1145/3637528.3671957. URL https://doi.org/10.1145/3637528.3671957.

A Proxy Verification Plot (S1)

Figure 3 plots the strict-versus-calibrated performance gap against λ for the Inertia sweep, verifying the three S1 desiderata referenced in Section 5.

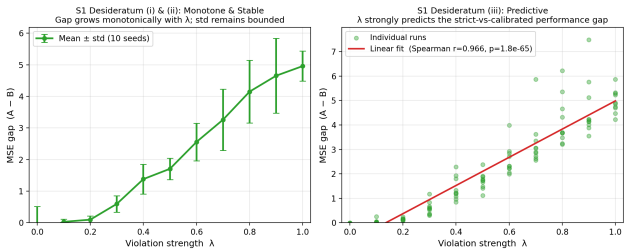

Figure 3: λ **satisfies all three S1 desiderata** as a violation proxy (10 seeds $\times$ 11 levels $=$ 110 points). The gap ($\text{MSE}_A - \text{MSE}_B$) grows monotonically with λ (monotone) with bounded spread (stable), and is strongly rank-correlated with λ (Spearman r=0.966, p=1.8×10^{-65}; predictive).

B Sample-Efficiency Curve

Figure 4 accompanies Section 5.5: at λ=0, strict (A) and calibrated (B) coincide and lie 3–7$\times$ below the unconstrained GNN (C) at every training-set size, so calibration retains equivariance's data efficiency where the symmetry holds.

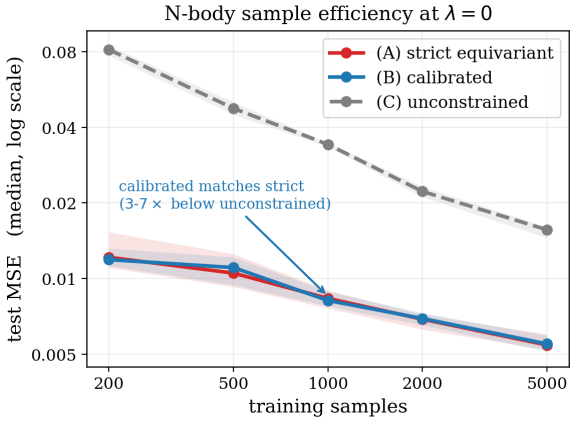

Figure 4: **Sample efficiency at** λ=0 (N-body, median $\pm$ IQR). Calibrated (B) tracks strict (A) and both sit well below the unconstrained GNN (C) at every training-set size.

C Within-Study Evidence from the Literature

Table 3 summarizes five studies that compare strict (A), calibrated (B), and unconstrained (C) baselines within a single pipeline, providing the prior evidence that motivates the evaluation standards in Section 4.

D Parameter-Matched Ablation (Inertia)

One might attribute the strict-versus-calibrated gap on the Inertia task to capacity: MixedEMLP (B) has 918,450 parameters vs. EMLP (A)'s 572,121, a 1.61$\times$ difference. To rule this out, we train a deeper EMLP with 5 layers at the same channel width (A+: 1,056,057 parameters, 1.15$\times$ of B) on the same sweep (6 λ levels, 10 seeds, 2000 epochs).

A+ does not close the gap. For every $\lambda \geq 0.4$, 9 or 10 of its 10 seeds diverge (to NaN or to MSE between 10^2 and 10^{14}), and at λ=1 none converges (9 NaN, one at 7.6×10^3). B instead reaches MSE 0.079 ± 0.088 (mean $\pm$ s.d.) on all 10 seeds at λ=1. Depth does not lower the bias floor, and it makes training unstable: every layer must satisfy the same misspecified constraint, so extra layers add parameters but not the missing direction. The advantage of MixedEMLP is therefore

Table 3: **Within-study evidence for calibrated symmetry from the literature.** (A) strictly equivariant; (B) calibrated/relaxed; (C) unconstrained. Only results reported within the same study and pipeline are included. None of these studies includes a controlled violation sweep or full cost accounting; Section 5 fills this gap.

Study	Violation type	Key result (A vs. B vs. C)	Where B helps
Kim et al. (2023) *PER*	Controlled mixed symmetry ($O^x(2), O^y(2), O^z(2)$); natural motion forecasting	High-violation ($O^x(2)$ only): PER (B) MSE $\mathbf{0.25}$ vs. $O(3)$-EMLP (A) 62.36 vs. MLP (C) 2.84; under full $O(3)$ symmetry PER matches best	High-violation / broken-symmetry regimes; robust across all Waymo normalization schemes
McClellan et al. (2025) *PEnGUiN*	Controlled sensor bias / safety regions; natural road layout	In biased scenarios, strict EGNN (A) suffers significant reward degradation; PEnGUiN (B) remains near-optimal; GNN (C) underperforms in symmetric cases	High-violation MARL; robustness under spatial asymmetry
Huang et al. (2023) *Approx. equivariant graphs*	Approximate graph automorphisms (partial symmetry)	Human3.6M: Relax-$\mathcal{S}_{16}$ (B) MPJPE 39.87 vs. $\mathcal{S}_{16}$ (A) 42.55 vs. trivial (C) 41.60	Intermediate symmetry strength (bias–variance sweet spot)
Chang et al. (2025) *Partially equivariant RL*	Natural reward/dynamics asymmetry (gravity, contact, region rewards)	Ant-v2: PE-SAC (B) reward $\sim$6000 vs. equivariant SAC (A) $\sim$5000 vs. SAC (C) 4500; strict agent can fail under strong symmetry breaking	Asymmetry increases; prevents collapse of strict equivariant agent
Brehmer et al. (2025) *Equivariance at scale*	Symmetric simulation; architecture vs. augmentation	Equivariant Transformer (A) achieves $\sim 2\times$ lower error at equal FLOPs; augmentation closes the gap only with many extra epochs in the small-data regime	Low-data regime; compute-efficient learning of symmetry

architectural, since its residual pathway gives the model a way around the constraint, and not a consequence of parameter count.

E Alternative Views: Extended Treatment

This appendix expands the four standpoints of Section 6. For each, we state the argument in its strongest form, give our response, and name the empirical discriminator that would decide the question, so that the position remains falsifiable rather than rhetorical.

"Physics is known, so hard-code it." *The argument.* In many scientific domains, symmetries are not a modeling choice but a property of the underlying laws (e.g., $\mathrm{SE}(3)$ for isolated mechanics). Hard-coding the correct group yields sample efficiency, stability, and interpretability; learning symmetry may introduce unnecessary degrees of freedom and failure modes. *Response.* We agree that when the pipeline and environment preserve the symmetry, strict equivariance is the appropriate default. Card 2 (Appendix F) confirms this for NequIP on MD-17, and our own sweep confirms it at $\lambda{=}0$. The practical problem is misspecification: "physics is $\mathrm{SE}(3)$" does not imply the dataset distribution is $\mathrm{SE}(3)$-consistent after boundary conditions, constraints, contacts, or sensor artifacts. Calibration does not disregard physical law; it conditions enforcement strength and scope on empirical evidence, often contracting to a subgroup or relaxing locally where violations concentrate. *Empirical discriminator.* Under violation sweeps, strict equivariance should win on low-violation strata. Our results confirm this: EMLP (A) outperforms MixedEMLP (B) at $\lambda{=}0$, and the two cross near $\lambda{\approx}0.1$. If calibrated models never help and just add variance, hard-coding is sufficient; if their gains concentrate on high-violation strata while they behave like the strict model elsewhere, as observed, calibration is warranted.

"Scale and augmentation are enough." *The argument.* Large models with data augmentation can learn approximate invariances without architectural constraints, avoiding brittle inductive bias and keeping engineering simple. Given enough data and compute, unconstrained models can match or exceed equivariant ones. *Response.* This view is plausible in large-scale regimes, but many graph and physics settings are compute- and data-limited, and the relevant question is not whether scale can match performance but at what total cost (training + inference + augmentation). Augmentation also carries a hidden assumption: it enforces whatever group it samples from. On N-body (Section 5.2), random $SO(3)$ rotations re-impose the symmetry the field breaks, and the augmented GNN degrades to 0.28 at $\lambda{=}1$, $14\times$ worse than the same GNN without augmentation. Sampling only rotations about $\hat{z}$, the subgroup the field preserves, avoids the damage but stays at 0.016, $2.5\times$ behind the calibrated model at $\lambda{=}1$ and above it in all 110 runs. Unconstrained models are also not uniformly weak: on O5Synthetic the MLP reaches 0.60 at $\lambda{=}1$, far below the strict model's 78.4, and on Inertia it edges past the strict model (4.25 vs. 4.97), though in both cases it stays well behind the calibrated model (0.075 and 0.037). Calibrated symmetry is the more compute-efficient route to robustness under structured violations (Brehmer et al., 2025), evaluated under explicit cost accounting (Section 4). *Empirical discriminator.* At matched budget, compare A/B/C and augmented baselines under violation sweeps. If augmentation-only baselines match calibrated models on high-violation strata without large extra compute, the scale view is strengthened. Our results show the opposite: on Inertia the MLP at $0.09\times$ the cost of EMLP reaches median MSE 4.25 at $\lambda{=}1$ vs. MixedEMLP's 0.037, a $\sim 115\times$ gap, and on N-body at $\lambda{=}1$ neither augmented GNN comes within $2.5\times$ of calibration.

"Invariants are safer than equivariants." *The argument.* Invariant scalar features are stable, cheap, and avoid the complexity of equivariant vector/tensor representations. For many tasks (e.g., scalar property prediction), invariants may be sufficient and less error-prone. *Response.* We agree that invariants are excellent for scalar targets, but they are fundamentally lossy for tasks whose outputs are geometric objects (forces, velocities, trajectories, policies): they force the model to reconstruct geometry from compressed descriptors and can harm out-of-distribution behavior. The invariant-vs-equivariant debate is *orthogonal* to the strict-versus-calibrated one: calibration relaxes whichever prior is assumed, invariant or equivariant. Our experiments make this explicit by stress-testing an invariant-scalar task (O5Synthetic) and an equivariant-tensor task (Inertia) under the same protocol; calibration helps under structured violation in both. *Empirical discriminator.* Report performance separately for scalar-target versus vector/tensor-field tasks, and stratify by violation. If invariant-only models match calibrated equivariant models on field-prediction tasks under distribution shift at lower cost, invariants suffice; otherwise, structured equivariance remains necessary.

"Learning symmetry is too hard." *The argument.* Learning group structure from data typically adds difficult optimization: many discovery approaches optimize non-convex objectives (e.g., minimax/adversarial formulations) and can be unstable in practice (Goodfellow et al., 2014; Yang et al., 2023). Discovery can also introduce identifiability issues, since equivariances are a fundamental source of non-identifiability in representation learning (Ahuja et al., 2021; Schmidt et al., 2025). These challenges are amplified for large discrete groups, where the search space ranges over subgroups of S_n (Yang et al., 2023; Karjol et al., 2023). *Response.* This argument targets discovery (M1) and does not carry over to relaxation (M2) or localization (M3). Our experiments use MixedEMLP, a relaxation mechanism with no symmetry discovery at all: it adds a regularized free pathway to an existing equivariant architecture. This is tractable, stable across all 10 Inertia seeds under violation, and yields a $134\times$ median MSE improvement at full violation. When discovery is used, it should be restricted: learning a small set of Lie-algebra generators $\mathfrak{g} \subset \mathfrak{gl}(n)$ rather than searching an unconstrained space (Finzi et al., 2020). For discrete structure we do not advocate learning full S_n from scratch; the target is partial exchangeability across roles and types. *Empirical discriminator.* A calibrated method should be evaluated not only on accuracy but on stability (variance across seeds), hyperparameter sensitivity, and whether the structure it learns is interpretable. Our MixedEMLP results show tight interquartile spread at high violation (Inertia IQR $[0.029, 0.045]$ at $\lambda{=}1$) and stable behavior across all 10 Inertia seeds for $\lambda{\geq}0.1$ (one seed misbehaves at $\lambda{=}0$); the O5 divergences (Section 5.4) mark where M2 relaxation needs care, which is exactly the kind of stability reporting this discriminator asks for. If calibration adds brittleness without robust gains on high-violation regimes, this view prevails.

F SYMMETRY STRESS-TEST CARDS: WORKED EXAMPLES

Table 4 presents the completed Stress-Test Card for our own Inertia violation sweep (Section 5), the worked example of a fully compliant submission referenced in the main text.

Table 4: **Symmetry Stress-Test Card for the Inertia violation sweep.** All five fields are filled; see Table 1 for field definitions.

Field	Content
Assumed group & representation	$O(3)$ acting on 3D positions; output is a rank-2 tensor (inertia matrix).
Calibration mechanism	(A) None — strict $O(3)$-equivariance hard-coded via constrained weights (EMLP). (B) Relaxation (M2) via residual pathway priors (MixedEMLP): free pathway added per layer, with L2 penalties 10^2 on the free and 10^{-3} on the equivariant weights.
Baselines	(A) EMLP, (B) MixedEMLP, (C) MLP, all matched in depth (3 layers) and width (384 channels).
Violation evidence	Controlled sweep over $\lambda \in [0, 1]$ (Eq. 3), 11 levels, 10 seeds. Spearman r=0.966 confirms λ satisfies all three S1 desiderata.
Total-cost reporting	Training: (A) 56.4 s/run, (B) 58.5 s/run (1.04× overhead), (C) 4.9 s/run. Inference: identical batch sizes. Augmentation: none for any model.

To demonstrate that the Stress-Test Card is immediately actionable, we apply it retrospectively to three recent works. Card 2 validates strict equivariance in a low-violation regime, showing the card is not a critique-only tool. Cards 1 and 3 demonstrate calibrated models in high-violation regimes, each using a different M2 relaxation mechanism (Card 1 also localizes it in space, M3). In each case, the card surfaces information that aggregate IID reporting would hide: the violation regime, the calibration mechanism, and the compute context.

Card 1: PEnGUiN (McClellan et al., 2025) - Relaxation (M2), localized in space (M3), via a learned symmetry score.

- **Assumed group & representation:** $O(n)$ acting on 2D/3D positional embeddings of agents in Multi-Agent Reinforcement Learning (MARL).
- **Calibration mechanism:** Relaxation (M2) with spatial localization (M3). A learnable *symmetry score* $\alpha \in [0, 1]$ per node blends a strictly equivariant coordinate update with a non-equivariant update:
$$\mathbf{u}_i^{l+1} = \alpha\, \mathbf{u}_{i,\text{eq}}^{l+1} + (1 - \alpha)\, \mathbf{u}_{i,p}^{l}.$$
α is output by an *Equivariance Estimator* MLP conditioned on node position and type, enabling spatially- and entity-dependent modulation. By construction, α=1 recovers a fully equivariant EGNN and α=0 recovers a standard GNN.
- **Baselines (A/B/C):** (A) EGNN / E2GN2 (strictly $O(n)$-equivariant); (B) PEnGUiN (calibrated, learned α); (C) Standard GNN (non-equivariant).
- **Violation sweep / strata:** Four controlled asymmetry conditions in Multi-Particle Environments (MPE): *None* (symmetric baseline), *Sensor Bias* (constant positional offset on observed entities - approximate equivariance), *Safety Region* (reward penalty in upper-right quadrant - regional equivariance), and *Decoy* (non-rewarding distractor - feature-wise equivariance). Highway-env racetrack and roundabout also test regional equivariance via road-boundary constraints.
- **Results on violation strata:** In the *None* (symmetric) condition, EGNN/E2GN2 (A) and PEnGUiN (B) achieve comparable rewards, validating that calibration does not hurt in symmetric settings. In the *Bias* and *Safety* conditions, strictly equivariant models (A) suffer significant reward degradation, while PEnGUiN (B) maintains near-optimal performance across all asymmetry types. Standard GNN (C) is less sensitive to bias but consistently underperforms both equivariant models in symmetric settings and falls short of PEnGUiN

in asymmetric ones. The learned α spatially tracks the violation: in the Safety Region scenario, α concentrates low values in the upper-right quadrant exactly where equivariance is broken, meeting the predictive desideratum of S1.

- **Total-cost reporting:** All architectures trained under identical PPO hyperparameters (batch size 2000, lr 3×10^{-4}, 10 SGD iterations, 10 seeds). PEnGUiN adds only a lightweight Equivariance Estimator MLP (2 layers, width 32) over the EGNN backbone; no spherical harmonics or tensor products are introduced.

Card verdict: Card 1 shows the canonical calibration pattern: strict (A) and calibrated (B) tie under no violation, B wins under structured violation, and unconstrained (C) loses in both regimes. The learned α provides a spatially-resolved, interpretable violation proxy that meets the predictive desideratum of S1.

Card 2: NequIP (Batzner et al., 2022) - Strict equivariance in a low-violation regime.

- **Assumed group & representation:** $E(3)$ acting on atomic positions; forces transform as $O(3)$ vectors (obtained as energy gradients, guaranteeing energy conservation and equivariant forces by construction).
- **Calibration mechanism:** None, strict $E(3)$-equivariance is hard-coded via tensor products of irreducible representations and spherical harmonic filters. This card validates strict models where their assumptions are met.
- **Baselines (A/B/C):** (A) NequIP ($E(3)$-equivariant, $l_{\max} = 3$); (B) Calibrated alternatives not evaluated in this study; (C) SchNet, DimeNet++ (invariant GNN-IPs operating on scalar distances only).
- **Violation sweep / strata:** MD-17 benchmark (aspirin, ethanol, malonaldehyde, naphthalene, etc.) in vacuum, a *low-violation* regime where the data pipeline preserves $E(3)$ symmetry: no gravity field, no boundaries, no sensor artifacts, and periodic boundary conditions handled separately. Training budget fixed at N=1,000 reference DFT structures.
- **Results on violation strata:** NequIP achieves state-of-the-art force MAE across all molecules (e.g., aspirin: 8.0 meV/Å vs. SchNet 58.5 meV/Å at the same training budget on the original MD-17 dataset; 8.2 meV/Å vs. FCHL19 20.9 meV/Å on the revised dataset). On the water data-efficiency experiment, equivariant NequIP ($l \geq 1$) follows a steeper log-log learning curve than invariant baselines (l=0): $E(3)$ equivariance changes the *slope* of the learning curve, not just its intercept, a qualitative benefit in this low-violation regime. No bias floor is observed because the violation level is near zero by construction.
- **Total-cost reporting:** Training is more expensive per epoch (Clebsch–Gordan tensor products, spherical harmonics) than invariant SchNet-style baselines. However, NequIP reaches target accuracy with far fewer training structures (e.g., 1,000$\times$ fewer than DeepMD on the water/ice benchmark at matched or better force error), yielding a favorable *total-cost* frontier in this low-violation regime (Brehmer et al., 2025).

Card verdict: In a low-violation regime, strict equivariance (A) is the right choice, since it maximizes sample efficiency and reaches better accuracy at lower total cost. This is consistent with our conditional claim: calibration is not universally better; it helps *specifically when violations are structured and consequential.*

Card 3: PER (Kim et al., 2023) - Relaxation (M2) via projection-based equivariance regularization.

- **Assumed group & representation:** Mixed symmetries $\{O^x(2), O^y(2), O^z(2)\}$ acting on 3D particle coordinates for moment-of-inertia prediction; candidate groups are provided as a hypothesis class.
- **Calibration mechanism:** Relaxation (M2) via soft regularization penalties, an alternative form to architectural gating. For each candidate group G_k, the Projection-based Equivariance Regularizer (PER) penalizes the ℓ_2-distance of network weights to the G_k-equivariant

subspace:

$$\mathcal{R}_k^{\mathrm{PER}}(W, b) = \tfrac{\lambda_k}{2}\|\mathrm{vec}(W) - Q_k Q_k^\top \mathrm{vec}(W)\|^2 + \tfrac{\lambda_k}{2}\|b - R_k R_k^\top b\|^2.$$

Regularization coefficients λ_k are *automatically tuned* during training based on the observed equivariance errors, allowing the model to discover which symmetry types are strongly vs. weakly violated without explicit supervision. Note: this form of relaxation operates on the *loss function*, not the architecture, distinguishing it from gating (Eq. 2) or residual pathways (Finzi et al., 2021).

- **Baselines (A/B/C):** (A) $O(3)$-EMLP (strictly $O(3)$-equivariant), $O^{(\mathrm{ax})}$-EMLP (strictly axis-equivariant); (B) PER (calibrated, learned λ_k); (C) Standard MLP (unconstrained).
- **Violation sweep / strata:** Five controlled strata on the moment-of-inertia task, each injecting a different perturbation to break $O(3)$ to a subgroup: (1) no error ($O(3)$-equivariant), (2) $-I\hat{x}\hat{x}^\top$ ($O^x(2)$-equivariant, soft $O(3)$), (3) $-I\hat{y}\hat{y}^\top$ ($O^y(2)$-equivariant, soft $O(3)$), (4) $-I\hat{z}\hat{z}^\top$ ($O^z(2)$-equivariant, soft $O(3)$), (5) mixed perturbation (soft $O(3)$ only).
- **Results on violation strata (Table 1 of Kim et al. (2023)):**
  - *Low-violation* (stratum 1, full $O(3)$): strict $O(3)$-EMLP (A) achieves Test MSE 1.13 ± 0.36; PER (B) 0.27 ± 0.23; MLP (C) 4.25 ± 0.17. At low violation PER beats even the strict model.
  - *High-violation* (stratum 2, $O^x(2)$ only): strict $O(3)$-EMLP (A) **fails catastrophically** at MSE 62.36 ± 41.10, a textbook bias floor from group misspecification; PER (B) adapts and achieves MSE $\mathbf{0.25 \pm 0.16}$; MLP (C) MSE 2.84 ± 0.12.
  - Analogous collapse occurs on strata 3–4: strict $O(3)$-EMLP reaches MSE 29.11 and 46.32, while PER achieves 0.56 and 0.32 respectively.

  PER's learned coefficients λ_k correctly track the violation: after automatic adjustment, $\lambda_{O^z(2)}$ remains high (100.0) when the data are $O^z(2)$-equivariant, while $\lambda_{O^x(2)}$ and $\lambda_{O^y(2)}$ drop to ~ 3 when those symmetries are broken, meeting the predictive desideratum of S1. On the Waymo motion forecasting task, PER also robustly outperforms baselines across all three normalization schemes (scale-aware, symmetry-aware, symmetry-scale-aware), even when competing methods are matched to the dominant $O^z(2)$ symmetry of road vehicles.
- **Total-cost reporting:** All competing models use the same 4-layer architecture with widths adjusted to *matched parameter counts*, so performance differences reflect symmetry modeling, not raw capacity. Unlike architectural gating (PEnGUiN) or residual pathways (RPP), PER adds no extra parameters: the regularizer operates on existing weights, making it the most parameter-efficient calibration mechanism of the three cards.

Card verdict: PER shows that calibration via soft regularization (an alternative M2 form to Eq. 2) matches or beats the strict model on low-violation strata and avoids its catastrophic bias floor on high-violation strata. The automatic λ_k tuning gives an interpretable, data-driven violation proxy, and matched parameter counts make the comparison fair in capacity.

G Standard S1 (Violation Proxy) - Concrete Verification

A violation proxy $\hat{v}(x)$ is useful only if it satisfies three desiderata: it must be *monotone* (larger $\hat{v}$ implies greater equivariance violation), *stable* (invariant to nuisance variation unrelated to the symmetry), and *predictive* (correlated with the performance gap between strict and calibrated models). Table 5 verifies three real-world proxies against these criteria.

Table 5: Violation proxies verified against desiderata (S1). ✓= satisfied; $\sim$ = partially.

Dataset	Proxy $\hat{v}(x)$	Monotone	Stable	Predictive
Highway-env (McClellan et al., 2025)	Distance to road boundary	✓	✓	✓
Human3.6M (Huang et al., 2023)	Equivariance error of pretrained GNN on node-pair swaps	✓	✓	$\sim$
MD-17/NequIP (Batzner et al., 2022)	Rotation order l of tensor features	✓	✓	$\sim$

For **Highway-env**, PEnGUiN's learned symmetry score α decreases specifically near road boundaries, the region where regional equivariance breaks, and is stable to agent-speed variation. The

performance gap between EGNN and PEnGUiN grows with proximity to the boundary, confirming predictiveness (McClellan et al., 2025).

For **Human3.6M**, Huang et al. (2023) directly measure the equivariance error of a trained DCNN over all 260 neighboring node pairs (mean relative error 5.17%), validating that the proxy tracks the actual symmetry violation in the skeleton graph. Predictiveness is partial: the best model (Relax-$\mathcal{S}_{16}$, MPJPE 39.87) is not the strictly equivariant baseline ($\mathcal{A}_G = \mathcal{S}_2^2$, MPJPE 42.18), but the graph is small (16 nodes), so variance effects dominate (Huang et al., 2023).

For **NequIP on MD-17** the check runs in reverse. MD-17 has no controlled violation, so the quantity varied is the rotation order l, that is, how much equivariant structure the model uses. Force MAE falls monotonically with l (aspirin: $l{=}0 \to 42.2$, $l{=}1 \to 12.6$, $l{=}3 \to 8.2$ meV/Å on revised MD-17) (Batzner et al., 2022), as expected when the *pipeline* preserves the symmetry: more equivariant structure only helps. Predictiveness is partial in the S1 sense, because the study compares no calibrated model.

H N-BODY DETAILS, CONTROLS, AND REPRODUCIBILITY

Simulator: Five particles with charges $q_i \in \{\pm 1\}$ interact through a Coulomb force (minimum distance 0.1, per-component force clamp 50) and are integrated with semi-implicit Euler for $T{=}1000$ steps of $\Delta t{=}10^{-3}$. Initial positions are $\mathcal{N}(0,1)$ and velocities $\mathcal{N}(0, 0.2^2)$. The field adds acceleration $\lambda \bar{s}\, \hat{z}$ with $\bar{s}{=}3$ to every particle. Because it is uniform, it leaves relative motion unchanged (up to the force clamp) and adds a common drift along $\hat{z}$ to the target displacement, the component no $O(3)$-equivariant map can produce. Each run uses 3000 training and 2000 test samples drawn with separate seeds.

Models and training: All N-body models use 4 message-passing layers of width 64 and predict the per-particle displacement. (A) is an EGNN (Satorras et al., 2022) with charge products and squared distances as edge inputs and a velocity gate (133,768 parameters). (B) adds a free MLP ($7 \to 64 \to 64 \to 3$) on raw positions, velocities, and charges to A's output (138,635 parameters). Since the output is a single vector type, the projection Proj_ρ of Eq. 2 is the identity and needs no separate implementation. (C) is a message-passing GNN on raw coordinates (29,699 parameters). ($\mathrm{A}_{\hat{z}}$) extends each EGNN layer with the fixed input vector $\hat{z}$: edges also receive $(x_i{-}x_j) \cdot \hat{z}$, nodes are embedded from $(q_i, v_i \cdot \hat{z})$, and the coordinate update gains a learned scalar times $\hat{z}$ (150,988 parameters). It is jointly equivariant in $(x, v, \hat{z})$ and hence exactly equivariant to the stabilizer of $\hat{z}$. All models train for 400 epochs with batch size 100 and learning rate $5{\times}10^{-4}$; A and C use Adam, and B uses AdamW with weight decay 1.0 on the free pathway and 10^{-6} on the equivariant weights. C+aug and C+aug$_{\hat{z}}$ apply one random rotation (from $SO(3)$ or from rotations about $\hat{z}$) to each whole batch of inputs and targets. Residuals use 10 random rotations on N-body and 20 on the synthetic tasks.

Free-pathway decay: Table 6 varies B's free-pathway weight decay over three orders of magnitude (3 seeds each). B is insensitive between 0.01 and 1 (the value used throughout). At 10 the decay pins the free pathway so tightly that B can no longer express the field, and its error at $\lambda{=}1$ rises to 0.26, about $36\times$ its value at the default, the expected failure of over-regularized relaxation (M2).

Table 6: Sensitivity of (B) to its free-pathway weight decay on N-body (median MSE, IQR in brackets, 3 seeds).

Free-pathway decay	$\lambda{=}0$	$\lambda{=}1$
0.01	0.0063 [0.0060, 0.0068]	0.0074 [0.0070, 0.0079]
0.1	0.0057 [0.0057, 0.0066]	0.0071 [0.0067, 0.0073]
1 (default)	0.0062 [0.0059, 0.0062]	0.0072 [0.0068, 0.0077]
10	0.0061 [0.0058, 0.0064]	0.26 [0.21, 0.29]

Synthetic tasks: Inertia and O5Synthetic use EMLP, MixedEMLP, and MLP (Finzi et al., 2021) with 3 layers of 384 channels, trained for 2000 epochs with Adam (learning rate $3{\times}10^{-3}$, batch size

256) on 1000 training and 2000 test samples. MixedEMLP adds L2 penalties with coefficients 10^2 on the free weights and 10^{-3} on the equivariant weights. In Eq. 3 and in the O5 perturbation, $\bar{s}$ is the mean absolute label.

Hardware and seeds: All runs use one NVIDIA H200 MIG slice (3g.71gb). Each violation level uses seeds 0–9 for the training data (on N-body also for model initialization) and seeds 10000–10009 for test data; all models at a given (λ, seed) see identical data. N-body wall-clock times in Table 2 include training and test evaluation.