TUNABLE EQUIVARIANCE VIA LEARNABLE ORDER-DEPENDENT SCALING

Anonymous authors
Paper under double-blind review

ABSTRACT

Equivariance is a powerful inductive bias in geometric deep learning, but real-world data often only approximately respects known symmetries. Strictly equivariant models can therefore be overly restrictive, while unconstrained models fail to exploit useful geometric structure. Existing relaxed-equivariance approaches mainly interpolate between equivariant and non-equivariant components, without modulating the internal structure of the equivariant features. We introduce a *tunable-equivariance* framework in which a single learnable parameter $\delta \in [0, 1]$ controls both the unconstrained branch and the relative strength of irrep components, weighting order ℓ by δ^ℓ. This yields a structured, continuous modulation of equivariance, motivated by a torus–hub geometric interpretation within an *associative operator algebra*. We learn δ jointly with the task objective under an empirical equivariance constraint. We instantiate the framework on $\mathrm{SO}(3)$-equivariant architectures for N-body dynamics, molecular property prediction, 3D shape classification, and LiDAR pose estimation. Across these benchmarks, the learned degree of equivariance is task-dependent. It generally improves performance over strict-equivariant baselines and outperforms learned relaxed-equivariance baselines where directly compared. Higher-order irrep components strengthen with δ, and the single-layer equivariance error is consistent with the predicted $O(1 - \delta)$ bound. Code will be released upon acceptance.

1 INTRODUCTION

Equivariance to geometric transformations is a central principle in geometric deep learning, improving sample efficiency, generalization, and robustness in domains such as robotics, vision, physics, and chemistry. A model is equivariant if transformations applied to the input produce corresponding transformations in the output. Formally, equivariance requires that a mapping commutes with the action of a transformation group. In practice, equivariant neural networks Cohen & Welling (2016); Worrall et al. (2016); Fuchs et al. (2020); Brandstetter et al. (2021); Cesa et al. (2022a;b); Kim et al. (2024); Xu et al. (2024) enforce this through representation-theoretic constraints, organizing features into irreducible representation (*irreps*) components and designing layers that preserve transformation rules. However, these methods assume that data strictly follows the underlying symmetry, which is often violated due to noise, partial observability, or task-specific asymmetries Wang et al. (2022); Yang et al. (2023); T. et al. (2024); Eric & S (2024).

This motivates *relaxed equivariance* Finzi et al. (2021); van der Ouderaa et al. (2022); Huang et al. (2023); Gruver et al. (2023); Stefanos et al. (2024); Ashman et al. (2024); Xie & Smidt (2024), where symmetry acts as a soft inductive bias. Existing approaches relax equivariance by blending equivariant features or weights with unconstrained ones. The blending scalar is either heuristic, hand-tuned Elhag et al. (2025), or learned as in ACE Manolache et al. (2025). However, these approaches relax the equivariant part uniformly. They do not explicitly control which of its components, such as irreducible orders, are trusted.

This raises a key question. Can a single parameter control both the blend and the internal structure of the equivariant part? Such control would let the model match its degree of symmetry to the task. We address this with a representation-driven framework termed as *tunable equivariance*. A single global parameter $\delta \in [0, 1]$ controls both the equivariant–unconstrained blend and the weight of each

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.

geometric component within the equivariant branch. We realize this through order-dependent scaling and learn δ with the task objective under an equivariance constraint.

Our contributions are threefold. **First**, we introduce a *single global* learnable parameter δ with two distinct roles: δ^ℓ modulates the relative contribution of equivariant irrep orders while preserving equivariance, whereas $(1-\delta)$ controls the unconstrained branch and hence the model's equivariance relaxation. **Second**, the proposed δ^ℓ scaling preserves $\ell = 0$ and progressively attenuates higher orders as δ decreases, without modifying the underlying equivariant operators. This scaling is motivated by a geometric interpretation on a torus–hub space $\mathcal{M}$, whose toroidal directions interact through a shared Euclidean hub. **Third**, we evaluate the framework across multiple geometric learning tasks, including N-body dynamics, molecular property prediction, 3D shape classification, and LiDAR pose estimation, analyzing predictive performance, sample efficiency, equivariance dynamics, and irrep modulation. Across these tasks, the learned coupling is task-dependent and generally improves over strictly equivariant baselines, while outperforming learned relaxed-equivariance baselines where directly compared.

2 Related Work

Equivariant neural networks have been studied via group convolutions for discrete Luo et al. (2019); Worrall & Brostow (2018); Wang et al. (2021); Carlos et al. (2019) and continuous Xie et al. (2023) groups. Representation theory later extended this line by organizing features into irreps Brandstetter et al. (2021); Cesa et al. (2022b). Layers are built from tensor products and Clebsch–Gordan projections Kondor et al. (2018), as in Steerable CNNs Cesa et al. (2022b); Kim et al. (2024). Group-convolution and harmonic Worrall et al. (2016) methods achieve exact equivariance. However, they require discretization or costly integration, which limits scalability. Gauge-based approaches Cohen et al. (2019) enforce local consistency but rely on predefined geometric structures.

Relaxed equivariance Elhag et al. (2025); van der Ouderaa et al. (2022); Huang et al. (2023); Ashman et al. (2024); Gruver et al. (2023); Finzi et al. (2021); Xie & Smidt (2024) addresses symmetry violations. It adds non-equivariant components or regularization van der Ouderaa et al. (2022); Moskalev et al. (2022); Lee et al. (2026); Tran et al. (2026). However, these methods generally cannot adapt the degree of symmetry to the task automatically.

A parallel line of work reduces symmetry at the *input* level. ASEN Goel et al. (2026) feeds an equivariant backbone a fixed auxiliary input. The stabilizer of this input is a chosen subgroup, and ASEN is instantiated only for permutation groups. Rapidash Vadgama et al. (2025) breaks symmetry or equivariance via reference frames or geometric inputs. In both, the degree of equivariance is a discrete choice made before training.

Pertigkiozoglou et al. Stefanos et al. (2024) relax equivariance only during optimization. They use a scheduled, not learned, unconstrained term and project back to exact equivariance at inference. This is an optimization aid rather than tunable, retained equivariance. ACE Manolache et al. (2025) instead learns the blending parameters via a *primal-dual update* rule Andreani et al. (2008); Chamon et al. (2020); Cherukuri et al. (2016), with a homotopy-based geometric interpretation. These parameters vary across layers. However, each layer still blends only at the branch level. Lin et al. Lin et al. (2026) study the symmetry increase inherent to strictly equivariant maps. They mitigate it through irrep selection, which remains a fixed feature choice. To our knowledge, none of these methods grades the internal structure of the equivariant branch.

In contrast, our framework uses a single continuous δ. It interpolates between strict and non-equivariant regimes. It also attenuates each irrep order separately. The degree of symmetry is thus tuned within one model. It is not fixed by a subgroup, a reference frame, or a feature choice.

3 Tunable Equivariance via a Torus-Hub operator transformation

Having positioned our approach relative to existing relaxation mechanisms, we now formalize the tunable-equivariance layer and then develop the torus–hub operator interpretation that motivates its order-dependent scaling.

3.1 Equivariant Training and Controlled Relaxation

Strict equivariance. Equivariant neural networks enforce that feature transformations commute with a group action. Let $\rho, \hat{\rho}$ be representations of $G \subset \mathrm{SO}(3)$ on the input/output spaces. A layer L is equivariant if

$$L(\rho(g)x) = \hat{\rho}(g)L(x), \quad \forall g \in G, \tag{1}$$

where x is the raw input (e.g. a point), drawn from a data distribution $\mathfrak{D}$. It is lifted into irreducible representation (*irreps*) blocks via spherical harmonics, giving features $\psi_{\text{in}} = \bigoplus_\ell \psi_{\text{in}}^{(\ell)}$, each block transforming under the corresponding Wigner-D matrix $\mathcal{D}^\ell(g)$: scalars ($\ell = 0$) transform trivially, $\mathcal{D}^0(g) \equiv 1$; vectors ($\ell = 1$) as $\mathcal{D}^1(g) = R(g)$, the rotation matrix of g. For the first layer, $\rho(g)x$ corresponds concretely to ψ_{in}, capped at $\ell \in \{0, 1\}$; $\hat{\rho}(g)L(x)$ corresponds to $\psi_{\text{out}} = \bigoplus_\ell \psi_{\text{out}}^{(\ell)}$, reaching $\ell \in \{0, 1, 2\}$ after the subsequent layer's tensor-product operations (Algo. 1, L4): the previous layer's ψ_{out} becomes the next layer's ψ_{in}. $\rho, \hat{\rho}$ are fixed throughout, determined entirely by the irreducible content of ψ_{in} and ψ_{out} respectively, and never modified during training. This assumes the data strictly follows the prescribed symmetry.

Relaxed equivariance. A common relaxation Gruver et al. (2023); Finzi et al. (2021); Xie & Smidt (2024); Stefanos et al. (2024) of Eq. 1 allows controlled deviation:

$$L(\rho(g)x) = \hat{\rho}(g)L(x) + r(x, g), \tag{2}$$

where $r(x, g)$ is the vector-valued *equivariance residual*, and its norm $\epsilon(x, g) = \|r(x, g)\|$ is the scalar *model* equivariance error, which measures the model's failure to be equivariant and largely arises from the unconstrained/non-equivariant part. Concretely, $L(x) \equiv \beta L_{\text{eq}}(x) + \gamma L_{\text{non-eq}}(x)$, where $L_{\text{eq}}(x)$ satisfies Eq. 1 and $L_{\text{non-eq}}(x)$ is unconstrained; $r(x, g)$, and hence ϵ, arises from $L_{\text{non-eq}}(x)$. ACE Manolache et al. (2025) fixes $\beta = 1$ and treats γ as a learnable per-layer parameter, optimized via *primal-dual update* alongside the task loss: in its equality-constrained variant, γ is driven toward 0; in its resilient (inequality-constrained) variant, γ instead settles at a learned, generally nonzero equilibrium, retaining partial equivariance where beneficial.

Our method: Our work aligns with ACE Manolache et al. (2025) and modifies Eq. 2 as:

$$L_\delta(\rho(g)x) = \hat{\rho}(g)\, L_\delta(x) + r_\delta(x, g), \tag{3}$$

where $\delta \in [0, 1]$ and $r_\delta(x, g)$ is the residual. The scalar equivariance error of L_δ is $\epsilon_\delta(x, g) = \|r_\delta(x, g)\|$, written $\epsilon(\delta)$ when (x, g) is implicit. In practice, each layer maps ψ_{in} to ψ_{out} as

$$\psi_{\text{out}} = \bigoplus_{\ell \in \{0,1,2\}} \delta^\ell\, \psi_{\text{in}}^{(\ell)} + (1 - \delta)\, \mathbf{W}(\psi_{\text{in}}), \tag{4}$$

with $\mathbf{W}$ an unconstrained dense weight matrix. L_δ denotes this layer; the network composes such layers. Here $L_{\text{eq}}(x) \equiv \bigoplus_\ell \delta^\ell\, \psi_{\text{in}}^{(\ell)}$ and we keep $\beta = 1$. The residual $r_\delta(x, g)$ arises entirely from the $(1 - \delta)\mathbf{W}(\psi_{\text{in}})$ term.

Interpretation of δ. As $\delta \to 1$, the unconstrained branch vanishes and each layer recovers the strict $\mathrm{SO}(3)$-equivariant representation. As $\delta \to 0$, the non-scalar equivariant orders ($\ell \geq 1$) are suppressed, the invariant $\ell = 0$ component is retained (since $\delta^0 = 1$), and the unconstrained branch receives full weight. Intermediate values attenuate each order geometrically in ℓ, so higher-order components are relaxed first. δ is a single, global scalar parameter learned via *primal-dual update* Andreani et al. (2008); Chamon et al. (2020); Cherukuri et al. (2016) enforcing $\epsilon(\delta) \leq \tau$ for a target threshold τ, jointly with the task loss. Thus $L_\delta(x)$ *tunes* itself to the degree of symmetry enforced by the task at hand via δ.

Theoretical operator. Eq. 4 is motivated by the operator transformation

$$\psi_{\text{out}} \approx \exp(\mathbf{A}(\delta))\, \psi_{\text{in}} = \underbrace{\bigoplus_{\ell \in \{0,1,2\}} \delta^\ell\, \tilde{\psi}_\ell(\delta)}_{\text{realized as } \bigoplus_\ell \delta^\ell \psi_{\text{in}}^{(\ell)}} + \underbrace{(1 - \delta)\, \psi_{\text{flat}}(\delta)}_{\text{realized as } (1-\delta)\mathbf{W}(\psi_{\text{in}})}, \tag{5}$$

obtained by a Taylor expansion of $\exp(\mathbf{A}(\delta))$ (Sec. 3.4). Each term maps onto its counterpart in Eq. 4. The equivariant term follows Proposition 2. The unconstrained term follows Sec. 3.5. Eq. 4 is thus a practical realization of Eq. 5, not an exact evaluation.

Algorithm 1 A high-level pseudocode of our framework

Require: Points $\{x_i\}$, irrep features ψ_i, coupling δ $\{x_i$ denotes the coordinates of point/node $i\}$
1: Build k-NN graph $\mathcal{G}$
2: Compute edge geometry: r_{ij}, $Y_{ij} = Y_\ell(\hat{r}_{ij})$ $\{Y_{ij}$ denotes spherical harmonics of the edge direction; gives $\psi_{\text{in}}^{\ell=0,1}$; r_{ij} is used in the radial weight $w(\|r_{ij}\|)$.$\}$
3: **for** each layer **do**
4: Messages: $m_{ij} = \text{TP}(\psi_j, Y_{ij}, w(\|r_{ij}\|))$ $\{\text{TP}$(tensor product) produces $\ell = 0, 1, 2$, per Eq. 10.$\}$
5: Aggregate and set layer input: $\psi_{\text{in},i} \leftarrow \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} m_{ij}$ $\{\mathcal{N}(i)$ denotes the neighbors of node i.$\}$
6: Irrep scaling: $\tilde{\psi}_{\text{in},i}^{(\ell)} \leftarrow \delta^\ell \, \psi_{\text{in},i}^{(\ell)}$ {Scaled copy of each irrep block; the original $\psi_{\text{in},i}$ remains unscaled.}
7: Update: $\psi_i \leftarrow \tilde{\psi}_{\text{in},i} + (1 - \delta)\,\text{MLP}(\psi_{\text{in},i})$ {Eq. 4; MLP $\equiv$ $\mathbf{W}$ acts on the unscaled layer input (Sec. 3.5).$\}$
8: **end for**
9: Pool and apply task head
10: **return** output

3.2 Model Equivariance Error

We analyze $\epsilon_\delta(x, g)$ for a single implemented layer, Eq. 4. The equivariant branch contributes no error: scaling each irrep block by δ^ℓ commutes exactly with $\mathcal{D}^\ell(g)$. The residual therefore comes entirely from the unconstrained branch: $r_\delta(x, g) = (1-\delta)\big[\mathbf{W}(\rho(g)x) - \hat{\rho}(g)\mathbf{W}(x)\big]$. The derivation is given in App. A.4.

Assumption 1 (Lipschitz continuity)**.** Assume $\mathbf{W}$ is L-Lipschitz, i.e. for any two features f_1, f_2 of matching shape, $\|\mathbf{W}(f_1) - \mathbf{W}(f_2)\| \le L\|f_1 - f_2\|$.

Theorem 1 (Linear error bound)**.** Under Assumption 1, for a single layer and any x, g, $\epsilon_\delta(x, g) \le C(1 - \delta) = \mathcal{O}(1 - \delta)$, with $C := 2\big(L\|x\| + \|\mathbf{W}(x)\|\big)$ independent of g and δ. The bound vanishes linearly as $\delta \to 1$.

Proof is given in Appendix A.4.

3.3 Learning the Coupling Parameter δ

We learn $\delta \in [0, 1]$ jointly with network parameters by explicitly coupling it to an empirical equivariance constraint, adopting this learning paradigm from ACE Manolache et al. (2025). We measure equivariance violation using stochastic group actions. Given an input x and a randomly sampled transformation $g \sim G$, we define the residual and the scalar equivariance error

$$r_\delta(x, g) = \psi(\rho(g)x; \delta) - \hat{\rho}(g)\,\psi(x; \delta), \qquad \epsilon_\delta(x, g) = \|r_\delta(x, g)\|\,, \tag{6}$$

and enforce a soft equivariance constraint $\epsilon(\delta) \le \tau$ through an augmented Lagrangian Andreani et al. (2008):

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda\, \max\big(0, \epsilon(\delta) - \tau\big) + \mu \left(\max(0, \epsilon(\delta) - \tau)\right)^2. \tag{7}$$

Here, $\psi(\cdot; \delta)$ denotes the final-layer output feature, i.e. ψ_{out} of the last layer L_δ as a function of the input, so r_δ is the network-level residual of Eq. 3. The error is computed in the output space using the group action $\hat{\rho}(g)$ appropriate for the task. The constraint is imposed on the expected error $\epsilon(\delta) = \mathbb{E}_{x \sim \mathfrak{D},\, g \sim G}\big[\epsilon_\delta(x, g)\big]$, estimated at each step as a minibatch average over inputs and sampled transformations. Here $\lambda \ge 0$ is a dual variable, $\max(0, \epsilon(\delta) - \tau)$ is the constraint violation, $\mu > 0$ sets the penalty stiffness, and τ is the threshold. The dual variable λ is updated Cherukuri et al. (2016); Chamon et al. (2020) via gradient ascent on the violation, so violations increase the penalty during training. In practice, λ is optimized jointly with the network parameters θ using Adam Kingma & Ba (2015). Adam ascends in λ and descends in θ and δ, and λ is projected to remain non-negative. The full procedure is summarized in Algo. 2.

Algorithm 2: Learning with Equivariance Constraint

Require: Data x, model $\psi(\cdot; \delta)$, threshold τ
for each training step **do**
  Sample transformation $g \sim G$
  $y = \psi(x; \delta)$, $y_g = \psi(\rho(g)x; \delta)$
  Equivariance error: $\epsilon = \|y_g - \hat{\rho}(g)y\|$
  Violation: $s = \max(0, \epsilon - \tau)$
  Loss: $\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda s + \mu s^2$
  Update θ, δ via gradient descent
  Update λ via gradient ascent on s; project to $\lambda \ge 0$
end for

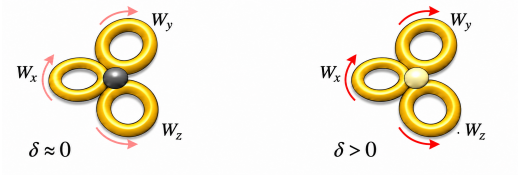


Figure 1: **Conceptual illustration of the torus–hub space $\mathcal{M}$.** Each torus is a toroidal direction, carrying a rotation generator $\mathbf{W}_x$, $\mathbf{W}_y$, or $\mathbf{W}_z$. The central sphere is the Euclidean hub, where the three tori meet and their generators can interact. $\mathbf{A}(\delta)$ realizes this picture as an operator in the *associative operator algebra*. The hub's shade shows the coupling strength. For $\delta \approx 0$ (left), the hub is inactive (dark). The generators act nearly independently, and the system is approximately factorized (faded arrows). As δ increases (right), the hub activates (light). The generators become coupled through it (bold arrows), inducing higher-order interactions.

3.4 Torus–hub operator transformation and Clebsch–Gordan decomposition

Here we define and derive Eq. 5. We define $\mathbf{A}(\delta)$ as:

$$\mathbf{A}(\delta) = \mathbf{A}_0 + \delta\,\mathbf{C} + (1-\delta)\,\mathbf{W}, \qquad \mathbf{A}_0 = \mathbf{W}_x + \mathbf{W}_y + \mathbf{W}_z. \tag{8}$$

$\mathbf{A}_0$ sums the rotation generators $\mathbf{W}_i$, one per toroidal direction of $\mathcal{M}$. It is fixed, bounded, and antisymmetric ($\ell = 1$). $\mathbf{C} = \sum_{i<j}[\mathbf{W}_i, \mathbf{W}_j]$ is their commutator. It is nonzero only at the Euclidean hub, where the directions interact, and is also antisymmetric ($\ell = 1$). $\mathbf{W}$ is a dense, learned matrix with no equivariance constraint. δ gates only the hub terms $\delta\mathbf{C}$ and $(1-\delta)\mathbf{W}$. At $\delta \approx 0$, $\mathbf{A}(\delta)$ remains factorized and there is no interaction among the generators. As δ increases, interactions increases through the hub. Since $\mathbf{W}$ is unconstrained, $\mathbf{A}(\delta)$ is not a Lie-algebra element. It lies in the *associative operator algebra*, closed under (associative) matrix multiplication. The conceptual diagram is shown in Fig. 1. The construction of $\mathcal{M}$ and the full component properties are given in App. A.5 and App. A.5.6.

$\mathbf{A}_0 + \delta\mathbf{C}$ is the connection-generator part of $\mathbf{A}(\delta)$. To obtain a transformation, we exponentiate, $\exp(\mathbf{A}(\delta)) = \sum_{k\geq 0} \mathbf{A}(\delta)^k/k!$. Each power $\mathbf{A}(\delta)^k$ expands into words over $\{\mathbf{A}_0, \mathbf{C}, \mathbf{W}\}$. The words without $\mathbf{W}$ sum to $(\mathbf{A}_0 + \delta\mathbf{C})^k$, and every word containing $\mathbf{W}$ carries a factor $(1-\delta)$. This gives the exact split

$$\exp(\mathbf{A}(\delta)) = \psi_{\text{eq}}(\delta) + (1-\delta)\,\psi_{\text{flat}}(\delta), \qquad \psi_{\text{eq}}(\delta) = \exp(\mathbf{A}_0 + \delta\mathbf{C}), \tag{9}$$

where ψ_{flat} collects all words containing $\mathbf{W}$ and is the non-equivariant part. Up to $k = 2$: $(1-\delta)\psi_{\text{flat}}(\delta) = (1-\delta)(\mathbf{W} + \frac{1}{2}(\mathbf{A}_0\mathbf{W} + \mathbf{W}\mathbf{A}_0) + \frac{\delta}{2}(\mathbf{C}\mathbf{W} + \mathbf{W}\mathbf{C}) + \frac{1-\delta}{2}\mathbf{W}^2 + \cdots)$, reflecting the non-commutativity of $\mathbf{W}$ with $\mathbf{A}_0$ and $\mathbf{C}$. The terms of $\psi_{\text{eq}}(\delta)$ admit the following Clebsch–Gordan decomposition(showing upto $k = 2$ example):

$$\psi_{\text{eq}}(\delta) = \underbrace{\mathbf{I}}_{\ell=0} + \underbrace{\mathbf{A}_0}_{\ell=1} + \tfrac{1}{2}\underbrace{\mathbf{A}_0^2}_{\ell=0,2} + \delta\Big(\underbrace{\mathbf{C}}_{\ell=1} + \tfrac{1}{2}\underbrace{(\mathbf{A}_0\mathbf{C} + \mathbf{C}\mathbf{A}_0)}_{\ell=0,2}\Big) + \tfrac{1}{2}\delta^2\underbrace{\mathbf{C}^2}_{\ell=0,2} + \cdots. \tag{10}$$

Each term of $\psi_{\text{eq}}(\delta)$ is a fixed 3×3 matrix in $\text{End}(\mathcal{D}^1)$, the 9-dimensional space of linear operators on $\mathcal{D}^1$. These terms arise from matrix multiplication, not from a tensor product $u \otimes v$, so the Clebsch–Gordan rule does not apply to the product directly. However, $\mathcal{D}^1$ is real and orthogonal, hence self-dual, so $\text{End}(\mathcal{D}^1) \cong \mathcal{D}^1 \otimes \mathcal{D}^1$. By closure of the associative operator algebra, every term, for any Taylor order $k \geq 1$, stays in this space. Its Clebsch–Gordan decomposition, $\mathcal{D}^1 \otimes \mathcal{D}^1 \cong \mathcal{D}^0 \oplus \mathcal{D}^1 \oplus \mathcal{D}^2$, therefore applies to every term, giving $\ell = \{0, 1, 2\}$ as in Eq. 10 (detailed explanation with an example is given in App. A.1). Post-decomposition, Eq. 9 becomes:

$$\exp(\mathbf{A}(\delta)) = (1-\delta)\,\psi_{\text{flat}}(\delta) + \bigoplus_{\ell=\{0,1,2\}} \psi_{\text{eq},\ell}(\delta). \tag{11}$$

In practice, the irrep contributions arising from different algebraic orders in Eq. 10 are not uniformly distributed across irreducible components. Degree $\ell \neq 0$ is reached only by products of at least ℓ generators, so higher irreps arise from higher-order interactions. We further assume the following.

Assumption 2 (Order–Irrep Alignment)**.** Irrep order ℓ corresponds to angular frequency on $\text{SO}(3)$ (spherical harmonics of degree ℓ are eigenfunctions of the angular Laplacian, eigenvalue $\ell(\ell+1)$); by the spectral bias of neural networks Rahaman et al. (2019) toward low-frequency content, higher-ℓ components are learned less readily than lower-ℓ components.

Table 1: **Comparison of our method with existing approaches on the N-body and QM9 datasets. Left:** Test MSE $\times 10^{-3}$ results on N-body. **Right:** MAE val. (↓) results for the 12 targets of QM9 on unconstrained PET Pozdnyakov & Ceriotti (2023), invariant Schütt et al. (2017),strict-equivariant Brandstetter et al. (2021),SOTA relaxed-equivariant Manolache et al. (2025) and ours. ACE and ours both use SEGNN as the backbone. Best results are shown in **bold**, and second-best results are underlined. In N-Body experiment, the baselines without Uncertainties: the upper block are official published scores (no uncertainties reported), reproduced as published for fair comparison.

Method	MSE
Linear Brandstetter et al. (2021)	81.9
SE(3)-Tr. Fuchs et al. (2020)	24.4
RF Klein et al. (2023)	10.4
ClofNet weitao Du et al. (2022)	6.5
EGNN Satorras et al. (2021)	7.1
EGNO Xu et al. (2024)	5.4
SEGNN Brandstetter et al. (2021)	5.6 ± 0.14
$\text{SEGNN}_{\text{REL}}$ Stefanos et al. (2024)	4.8 ± 0.15
$\text{SEGNN}_{\text{ACE}}$ Manolache et al. (2025)	4.0 ± 0.17
$\text{SEGNN}_{\text{Ours}}$	**3.5 ± 0.15**

Target	PET	SchNet	SEGNN	$\text{SEGNN}_{\text{ACE}(f_{eq})}$	$\text{SEGNN}_{\text{Ours}}$
μ	0.0562 ± 0.0029	0.0641 ± 0.0021	0.0371 ± 0.0003	0.0369 ± 0.0005	**0.0351 ± 0.0005**
α	0.0995 ± 0.0049	0.1311 ± 0.0035	0.0991 ± 0.0069	0.0956 ± 0.0024	**0.0944 ± 0.0001**
ϵ_{homo}	0.0592 ± 0.0006	0.0607 ± 0.0003	0.0250 ± 0.0008	0.0244 ± 0.0013	**0.0221 ± 0.0062**
ϵ_{lumo}	0.0508 ± 0.0007	0.0521 ± 0.0005	**0.0244 ± 0.0005**	0.0259 ± 0.0002	0.0258 ± 0.0133
$\Delta\epsilon$	0.0694 ± 0.0054	0.0858 ± 0.0012	**0.0450 ± 0.0035**	0.0545 ± 0.0052	0.0528 ± 0.0251
$\langle R^2 \rangle$	1.4634 ± 0.0023	1.6486 ± 0.0007	1.2028 ± 0.0145	1.1681 ± 0.0724	**1.1410 ± 0.0062**
ZPVE	0.0034 ± 0.0003	0.0035 ± 0.0005	0.0030 ± 0.0006	0.0026 ± 0.0008	**0.0024 ± 0.0004**
U_0	7.8451 ± 1.3410	9.7429 ± 2.8625	2.6163 ± 0.9513	2.5597 ± 0.1632	**2.4326 ± 0.0007**
U	6.4276 ± 1.5196	8.6311 ± 2.5427	2.5645 ± 1.5634	2.4796 ± 1.6231	**2.4441 ± 0.0057**
H	5.5493 ± 1.0433	9.5985 ± 3.0542	2.7577 ± 1.0542	2.4969 ± 1.0846	**2.4721 ± 1.0785**
G	6.8267 ± 2.0231	8.9455 ± 3.1436	2.7505 ± 1.6842	2.5954 ± 0.8451	**2.3237 ± 0.0006**
c_v	0.0401 ± 0.0132	0.0479 ± 0.0013	0.0399 ± 0.02163	0.0410 ± 0.0089	**0.0395 ± 0.01249**

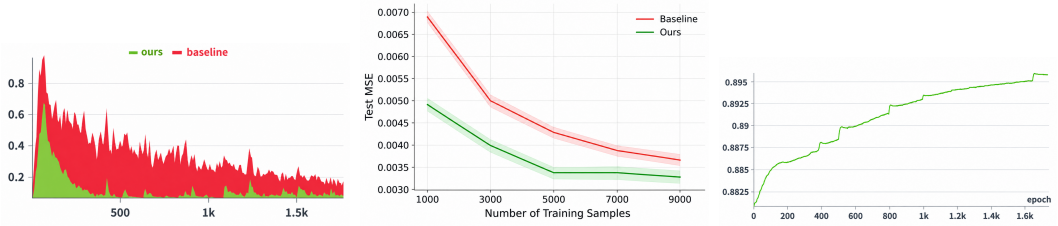


Figure 2: **N-body analysis. Left:** Equivariance error comparison of baseline SEGNN Brandstetter et al. (2021) and ours over the training epochs. **middle:** Test MSE vs number of training samples comparing baseline SEGNN Brandstetter et al. (2021) and our method. **right:** δ variation across epochs.

Assumption 2 motivates attenuating higher degrees more strongly. The following result gives the form of this attenuation.

Proposition 2 (Degree-wise attenuation)**.** Each degree-ℓ block admits the surrogate

$$\psi_{\text{eq},\ell}(\delta) \approx \delta^{\ell}\,\tilde{\psi}_{\ell}, \qquad \tilde{\psi}_{\ell} := \psi_{\text{eq},\ell}(1), \tag{12}$$

which is exact at $\delta = 1$, with fidelity error $\mathcal{O}(1-\delta)$ (App. A.3).

In the implementation we set $\tilde{\psi}_{\ell} = \psi_{\text{in}}^{(\ell)}$ (Sec. 3.5). Substituting Eq. 12 in Eq. 11 gives Eq. 5. In practice, we directly scale the irrep features of ψ_{in} (obtained by wrapping the input features x, with Spherical harmonics; Algo. 1, L1–2), and replace $\psi_{\text{flat}}(\delta)$ with $\mathbf{W}$ alone, applied to the unscaled aggregate(L7); this substitution is explained in Sec. 3.5.

3.5 Limitations and Scope

In Eq. 4, we directly scale the irrep features of ψ_{in}. As seen in Eq. 10, the irreps ℓ arise from distinct mixing paths (e.g. $\mathbf{A}_0\mathbf{C}$, $\mathbf{C}^2$) that current PyTorch Paszke et al. (2019) libraries do not expose as composable operator products; consequently, a literal implementation of Eqs. 8–11 is not possible, so we apply Proposition 2. As a result, we replace $\tilde{\psi}_{\ell}$ with $\psi_{in}^{(\ell)}$ in Eq. 4. For the same reason, we replace $\psi_{\text{flat}}(\delta)$ (Eq. 9) with $\mathbf{W}$ alone: $\mathbf{W}$, an unconstrained dense linear map, already breaks equivariance on its own, making the cross-terms with $\mathbf{A}_0$ and $\mathbf{C}$ unnecessary to recover this behavior. Eq. 8—12 are the mathematical scaffold to reach Eq. 5. Also, we used TP in Algo. 1, L4 to produce $\ell \in \{0, 1, 2\}$ as opposed to ordinary matrix multiplication for the same reason as stated here.

4 Experimental Setup and Results

In this section, we aim to answer the following six research questions:

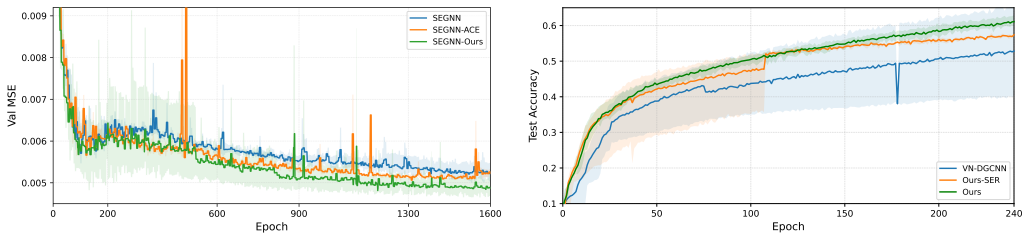


Figure 3: **Left:** N-body Val MSE result for strict SEGNN(baseline), SEGNN-ACE and ours. **Right:** Classification accuracy curves on ModelNet40 with VN-DGCNN as baseline and ours ablation on a relaxed-equivariant framework Lee et al. (2026).

RQ1. *Performance:* Does learning δ outperform fixed and relaxed-equivariance baselines?

RQ2. *Equivariance control:* How does the learned δ shape the network's equivariance error during training?

RQ3. *Task dependence:* Does the preferred level of equivariance vary across tasks?

RQ4. *Irrep composition:* Are the observed *irrep* components consistent with the graded scaling in Proposition 2?

RQ5. *Theory–empirical alignment:* Does the measured single-layer equivariance error follow the $\mathcal{O}(1-\delta)$ behavior predicted by Theorem 1?

RQ6. *Lipschitz assumption:* Is Assumption 1 empirically supported across tasks?

4.1 Benchmark Performance and Sample Efficiency

N-Body simulations. To answer **RQ1**, we evaluate on the N-Body Kipf et al. (2018) dataset. Fig. 3 (left) compares the validation MSE (mean $\pm$ std) over 1600 epochs between vanilla SEGNN Brandstetter et al. (2021), ACE Manolache et al. (2025) and our tunable-equivariant approach built upon SEGNN backbone. Our method converges faster and exhibits greater stability than the other baselines. Tab. 1 (left) reports test MSE ($\times 10^{-3}$) across strict equivariant, non-equivariant, and relaxed-equivariant baselines. Our model consistently outperforms all baselines, including the relaxed-equivariance method ACE Manolache et al. (2025). Fig. 2 (middle) reports test MSE ($\times 10^{-3}$) across different training set sizes. While the baseline approaches our performance at 5000 samples and nearly converges at 9000 samples, our method achieves comparable performance with roughly 40% fewer training samples, demonstrating improved sample efficiency.

Molecular Property Regression. We evaluate on the QM9 R. Ramakrishnan & von Lilienfeld (2014); Ruddigkeit L & JL. (2012) dataset using the strict equivariant SEGNN Brandstetter et al. (2021). Even though SEGNN is more flexible than a strict invariance architecture like SchNet Schütt et al. (2017), it enforces strict equivariance in. Tab. 1(right). Our method on top of SEGNN boosts the performance of SEGNN, and it outperforms the SOTA architecture ACE Manolache et al. (2025) on relaxed equivariance. Even though our method is able to achieve the best performance in 11 of the 13 target metrics, there are 2 metrics where our method failed to beat the baseline performance. We also compared with an unconstrained model, PET Pozdnyakov & Ceriotti (2023). Among all the baselines, PET performs poorly. Overall, our method's performance is largely comparable.

3D shape classification. On the rotated, noisy ModelNet40 Wang et al. (2019) classification benchmark, we apply our method through a message-passing formulation. We compare against VN-DGCNN Congyue et al. (2021), which enforces strict equivariance, and KPConv Thomas et al. (2019), a non-equivariant point-cloud CNN. We additionally build an internal ablation, **Ours**$_{\text{SER}}$, adapting the depth-wise equivariant/invariant decoupling principle of Lee et al. (2026) to our own architecture: the first few layers are held strictly equivariant (δ=1), the remaining layers are left unconstrained (δ=0), and a paired-rotation alignment loss is applied at the cutoff. Tab. 2 (right) shows that our method (continuous, per-irrep δ) outperforms both baselines and this hard-cutoff ablation on classification and instance accuracy, gaining +2.54 points in class accuracy and +2.86 points in

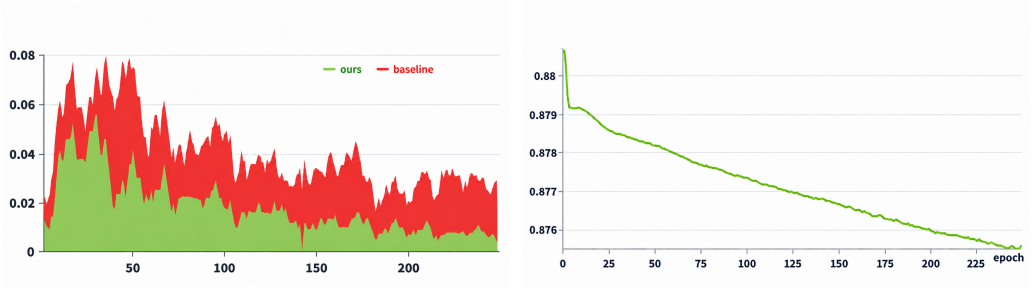


Figure 4: **Equivariance error comparison with baseline and delta variation of our model on ModelNet40.** From left to right: Equivariance error comparison of baseline and ours. The delta variation across the epochs.

Table 2: **Left: OxfordRobotCar Maddern et al. (2017) pose estimation analysis in 6 DoF on Lidar scans.** Rotation error in ∘ and translation error in metres. **right: ModelNet40 classification results.** Test set accuracy comparison across methods. Best results in **Bold** and second-best results in underlined. In Ours$_{\text{SER}}$, we have attached SER Lee et al. (2026) regularization to our model.

(a) Pose estimation

Method	RE ∘ ↓	TE (m.) ↓
PointNet++ Qi et al. (2017)	13.0 ± 0.013	87.0 ± 0.029
Atlas-D Bhat et al. (2025)	44.23 ± 9.14	194.55 ± 3.46
VN-DGCNN Congyue et al. (2021)	12.201 ± 0.003	13.131 ± 0.005
VN-DGCNN-Quasi Tran et al. (2026)	11.04 ± 0.006	12.63 ± 0.013
Ours	**10.361 ± 0.004**	**12.216 ± 0.002**

(b) Classification

Method	Class Acc. ↑	Instance Acc. ↑
KPConv Thomas et al. (2019)	0.7717 ± 0.0056	0.8200 ± 0.0076
VN-DGCNN Congyue et al. (2021)	0.7700 ± 0.0151	0.8323 ± 0.0147
Ours$_{\text{SER}}$ Lee et al. (2026)	0.7877 ± 0.0034	0.8503 ± 0.0043
Ours	**0.7954 ± 0.0016**	**0.8609 ± 0.0035**

instance accuracy over VN-DGCNN. Fig. 3 (right) shows the test-accuracy trajectories of **Ours**, **Ours**$_{\text{SER}}$, and VN-DGCNN over training, with **Ours** converging to the highest accuracy throughout.

Pose Estimation. We evaluate lidar-scan-based equivariant pose estimation on the Oxford RobotCar dataset Maddern et al. (2017) (Tab. 2(left)). We compare against VN-DGCNN Congyue et al. (2021), a strictly equivariant baseline, and VN-DGCNN-Quasi Tran et al. (2026), which relaxes VN-DGCNN via a learned per-input rotation applied to its output. VN-DGCNN-Quasi improves over strict VN-DGCNN, confirming that relaxing equivariance helps on this task. Our approach improves further, outperforming all baselines in both rotation and translation error.

4.2 Analysis on δ

For **RQ2**, learning δ reduces the task-level error of Eq. 6 below the strict baseline (Fig. 2, left; Fig. 4, left). The error rises early but then decreases gracefully. Theorem 1 bounds each iterate, not the curve's shape, since $\mathbf{W}$ and δ evolve during training.

To answer **RQ3**, Fig. 2(right) and Fig. 4(right) show the evolution of δ across epochs for N-body Kipf et al. (2018) and rotated noisy ModelNet40 Wang et al. (2019). We observe opposite trends across the two tasks: in N-body, δ increases monotonically, indicating that stronger equivariant coupling benefits the dynamics, whereas in ModelNet40, δ decreases during training, suggesting that weaker symmetry constraints better accommodate noisy shape variations. This shows that the optimal degree of equivariance is task-dependent.

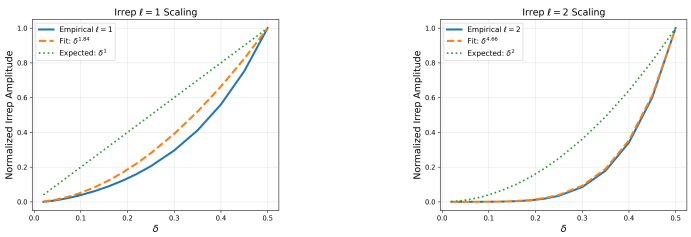


Figure 5: **Irreps signals with $\ell = 1, 2$ variation over discrete δ at test time on QM9.** This demonstrates δ modulation of representation structure, validating prop. 2.

Table 3: **Lipschitz analysis and computational cost.**

(a) Analytical and empirical Lipschitz estimates across tasks.

Task	Analytical L_a	Empirical L_e	Ratio L_a/L_e	Bound Satisfied
ModelNet40 Classification	33.72	2.05	16.45	True
N-body Dynamics	1.56	0.10	15.22	True
RobotCar Pose Estimation	13.54	1.61	8.41	True
QM9	2.32	0.16	14.12	True

(b) Parameter count and runtime comparison.

Model	# Parameters	Train (s/epoch)	Val (s/epoch)
SEGNN – N-body	147,424	1.368 ± 0.004	0.505 ± 0.005
$\text{SEGNN}_{\text{ACE}}$ – N-body	937,960	1.510 ± 0.005	0.525 ± 0.007
$\text{SEGNN}_{\text{Ours}}$ – N-body	937,962	1.512 ± 0.004	0.527 ± 0.004
SEGNN – QM9	1,033,525	118.403 ± 0.108	11.152 ± 0.082
$\text{SEGNN}_{\text{ACE}}$ – QM9	9,869,667	128.027 ± 0.034	11.423 ± 0.049
$\text{SEGNN}_{\text{Ours}}$ – QM9	9,869,668	129.039 ± 0.043	11.402 ± 0.089

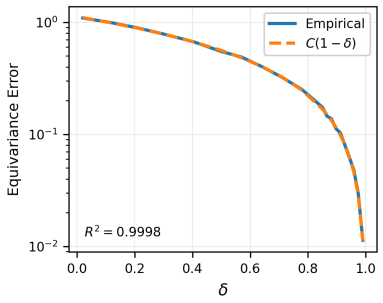


Figure 6: **Single-layer equivariance error.** Measured $\epsilon(\delta)$ compared with the $C(1-\delta)$ dependence predicted by Theorem 1. $C = 1.12$, $R^2 = 0.9998$.

To answer **RQ4**, we vary δ at test time and measure how feature energy is distributed across irrep orders. Fig. 5 shows the energy of the $\ell = 1, 2$ features over $\delta \in [0, 0.5]$ on QM9, normalized at $\delta = 0.5$. Increasing δ strengthens higher-order features, and higher orders remain more suppressed than lower ones. The fitted exponents (1.84 for $\ell = 1$, 4.66 for $\ell = 2$) exceed the per-layer δ^ℓ shown as "Expected", because the scaling of Proposition 2 compounds across depth. At the operator level, the coupling-driven part (terms inside $\psi_{\text{eq},\ell}$ with at least ℓ factors of $\mathbf{C}$) and is exact at $\delta = 1$.) scales as δ^ℓ (Fig. 16, App. A.3). Lower orders thus dominate the features at small δ, and higher orders emerge as δ increases. This shows that δ provides graded control over the irrep composition of the features.

RQ5. Theorem 1 bounds a single layer $F = F_{\text{eq}} + (1-\delta)\mathbf{W}$, not the composed network. We isolate one layer and measure $\|\epsilon(\delta)\| = \|F(\rho(g)x) - \hat{\rho}(g)F(x)\|$ over the test set for $g \in \text{SO}(3)$, across 20 values of $\delta \in [0.02, 0.99]$ densely sampled near $\delta \to 1$. The k-NN graph is built once (unrotated) and held fixed across both forward passes, so $\epsilon(\delta)$ reflects only δ, not rotation-induced graph changes. Figure 6 fits the result to $C(1-\delta)$ via least squares (no intercept), giving $R^2 = 0.9998$. The observed scaling therefore provides strong empirical support for the single-layer dependence predicted by Theorem 1. Additional analysis is provided in App. A.2.2.

4.3 Ablation Study

To support the claim of **RQ1** in Tab. 3b, we report the comparison against ACE below. Parameter count is matched, since both methods wrap the same equivariant backbone. At this matched budget, our training and validation times coincide with ACE to within noise, with a marginal increase in QM9 training time. Our accuracy improves on ACE beyond combined error bars on both datasets. Tunability is thus obtained at essentially matched parameter count and compute cost relative to ACE.

To answer **RQ6** we took trained models of our framework across all the benchmarks and computed the theoretical and empirical (99th percentile) values of the Lipschitz constant in Assumption 1 to validate the boundedness. Tab. 3a shows that across all the datasets, the Lipschitz constant is bounded.

5 Conclusion

We introduced a framework for tunable equivariance in which a learnable parameter δ controls the strength of symmetry, enabling a continuous transition between non-equivariant and equivariant representations. Instead of enforcing symmetry as a hard constraint, the model adapts the degree of equivariance to the task. Across multiple benchmarks, the approach consistently improves predictive accuracy and training stability over strictly equivariant baselines. Empirically, the model adjusts δ to balance expressivity and inductive bias, activating structured interactions when beneficial while retaining flexibility otherwise. These results suggest that equivariance can be treated as a tunable property rather than a binary design choice. This perspective opens directions for extending the framework to other symmetry groups, introducing adaptive coupling mechanisms, and designing architectures that more directly reflect underlying algebraic structure.

AI USE STATEMENT

In this work, we used generative AI tools for polishing and editing the manuscript. We have not used generative AI tools for result generation or writing code. Additionally, we used generative AI tools for searching related works and understanding mathematical concepts of differential geometry and equivariance. We have reviewed all AI-assisted work. For example, we checked LLM-generated research ideas for potential plagiarism through a manual literature survey. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work focuses on improving the training of equivariant neural networks, with potential applications in areas such as drug discovery and materials science. We do not anticipate direct negative societal impacts from this work; however, broader implications depend on downstream applications.

REPRODUCIBILITY STATEMENT

The details of implementation are provided in the Appendix section: Additional experimental details, Sec A.6. All datasets are publicly available and open-source. Hardware usage and details regarding the implementation across all experiments are provided in the appendix along with the time complexity analysis.

REFERENCES

R. Andreani, E. G. Birgin, J. M. Martínez, and M. L. Schuverdt. On augmented lagrangian methods with general lower-level constraints. *SIAM Journal on Optimization*, 2008.

Matthew Ashman, Cristiana Diaconu, Adrian Weller, Wessel Bruinsma, and Richard E. Turner. Approximately equivariant neural processes. In *Advances in Neural Information Processing Systems*, 2024.

Manu Bhat, Jonghyun Park, Jianke Yang, Nima Dehmamy, Robin Walters, and Rose Yu. Atlasd: Automatic local symmetry discovery. In *Forty-Third International Conference on Machine Learning*, 2025.

Johannes Brandstetter, Rob Hesselink, Elise van der Pol, Erik Bekkers, and Max Welling. Geometric and physical quantities improve e(3) equivariant message passing. In *International Conference on Learning Representations*, 2021.

Esteves Carlos, Xu Yinshuang, Allen-Blanchette Christine, and Kostas Daniilidis. Equivariant multi-view networks. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019.

Gabriele Cesa, Leon Lang, and Maurice Weiler. A program to build e(n)-equivariant steerable cnns. In *International Conference on Learning Representations*, 2022a.

Gabriele Cesa, Leon Lang, and Maurice Weiler. A program to build E(N)-equivariant steerable CNNs. In *International Conference on Learning Representations*, 2022b.

Luiz F. O. Chamon, Santiago Paternain, and Alejandro Ribeiro. Counterfactual programming for optimal control. In *2nd Annual Conference on Learning for Dynamics and Control*, 2020.

Ashish Cherukuri, Enrique Mallada, and Jorge Cortes. Asymptotic convergence of constrained primal-dual dynamics. *Systems and Control Letters*, 2016.

Taco S. Cohen and Max Welling. Group equivariant convolutional networks. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning*, 2016.

Taco S. Cohen, Maurice Weiler, Berkay Kicanaoglu, and Max Welling. Gauge equivariant convolutional networks and the icosahedral cnn. In *Proceedings of the 36 th International Conference on Machine Learning*, 2019.

Deng Congyue, Litany Or, Duan Yueqi, Poulenard Adrien, Tagliasacchi Andrea, and Guibas Leonidas. Vector neurons: A general framework for so(3)-equivariant networks. In *International Conference on Computer Vision*, 2021.

Ahmed Elhag, T. Rusch, Francesco Di Giovanni, and Michael Bronstein. Relaxed equivariance via multitask learning. In *Proceedings of the Fourth Learning on Graphs Conference*, 2025.

Qu Eric and Krishnapriyan Aditi S. The importance of being scalable: Improving the speed and accuracy of neural network interatomic potentials across chemical domains. In *Advances in Neural Information Processing Systems*, 2024.

Matthias Fey and Jan E. Lenssen. Fast graph representation learning with pytorch geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

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

Fabian B. Fuchs, Daniel E. Worrall, Volker Fischer, and Max Welling. Se(3)-transformers: 3d roto-translation equivariant attention networks. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, 2020.

Abhinav Goel, Derek Lim, Hannah Lawrence, Stefanie Jegelka, and Ningyuan Huang. Any-subgroup equivariant networks via symmetry breaking. In *International Conference on Learning Representations*, 2026.

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

Ningyuan Huang, Ron Levie, and Soledad Villar. Approximately equivariant graph networks. In *Advances in Neural Information Processing Systems*, 2023.

Jaein Kim, Hee Bin Yoo, Dong-Sig Han, Yeon-Ji Song, and Byoung-Tak Zhang. Continuous so (3) equivariant convolution for 3d point cloud analysis. In *European Conference on Computer Vision*, 2024.

Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Thomas Kipf, Ethan Fetaya, Kuan-Chieh Wang, Max Welling, and Richard Zemel. Neural relational inference for interacting systems. In *International conference on machine learning*, 2018.

Leon Klein, Andreas Krämer, and Frank Noe. Equivariant flow matching. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Risi Kondor, Zhen Lin, and Shubhendu Trivedi. Clebsch–gordan nets: a fully fourier space spherical convolutional neural network. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS*, 2018.

John M. Lee. *Introduction to Smooth Manifolds*, volume 218 of *Graduate Texts in Mathematics*. Springer, New York, NY, 2nd edition, 2012. ISBN 978-1-4419-9982-5. doi: 10.1007/978-1-4419-9982-5.

Joohyung Lee, Changhun Kim, Hyunsu Kim, Kwanhyung Lee, and Juho Lee. Soft equivariance regularization for invariant self-supervised learning. In *International Conference on Learning Representations*, 2026.

Ning Lin, Jiacheng Cen, Anyi Li, Wenbing Huang, and Hao Sun. Reducing symmetry increase in equivariant neural networks. In *International Conference on Learning Representations*, 2026.

Junren Luo, Wannpeng Zhang, Jiongming Su, and Fengtao Xiang. Hexagonal convolutional neural networks for hexagonal grids. *IEEE Access*, 2019.

Will Maddern, Geoff Pascoe, Chris Linegar, and Paul Newman. 1 year, 1000km: The oxford robotcar dataset. *The International Journal of Robotics Research (IJRR)*, 2017.

Andrei Manolache, Luiz F. O. Chamon, and Mathias Niepert. Learning (approximately) equivariant networks via constrained optimization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Artem Moskalev, Anna Sepliarskaia, Ivan Sosnovik, and Arnold Smeulders. Liegg: Studying learned lie group generators. In *Advances in Neural Information Processing Systems*, 2022.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lererand James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lina, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high performance deep learning library. In *Annual Conference on Neural Information Processing Systems*, 2019.

Sergey Pozdnyakov and Michele Ceriotti. Smooth, exact rotational symmetrization for deep learning on point clouds. In *Advances in Neural Information Processing Systems*, 2023.

Charles R. Qi, Li Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Conference on Neural Information Processing Systems (NIPS)*, 2017.

P. M. Rupp R. Ramakrishnan, O. Dral and O. Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data Nature*, 2014.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred A. Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Blum LC Ruddigkeit L, van Deursen R and Reymond JL. Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17. *ACS*, 2012.

Víctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

K. T. Schütt, P.-J. Kindermans, H. E. Sauceda, S. Chmiela, A. Tkatchenko, and K.-R. Müller. Schnet: a continuous-filter convolutional neural network for modeling quantum interactions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017.

Pertigkiozoglou Stefanos, Chatzipantazis Evangelos, Trivedi Shubhendu, and Daniilidis Kostas. Improving equivariant model training via constraint relaxation. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, 2024.

Nguyen Quynh T., Schatzki Louis, Braccia Paolo, Ragone Michael, Coles Patrick J., Sauvage Frédéric, Larocca Martín, and Cerezo M. Theory for equivariant quantum neural networks. *PRX Quantum*, 2024.

Hugues Thomas, Charles R. Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J. Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

Viet-Hoang Tran, An Nguyen, Benoˆıt Guerand, Thieu N. Vo, and Tan M. Nguyen. Quasi-equivariant metanetworks. In *International Conference on Learning Representations*, 2026.

Sharvaree Vadgama, Mohammad Mohaiminul Islam, Domas Buracas, Christian Shewmake, Artem Moskalev, and Erik Bekkers. Probing equivariance and symmetry breaking in convolutional networks. In *Advances in Neural Information Processing Systems*, 2025.

Tycho F.A. van der Ouderaa, David W. Romero, and Mark van der Wilk. Relaxing equivariance constraints with non-stationary continuous filters. In *36th Conference on Neural Information Processing Systems*, 2022.

Rui Wang, Robin Walters, and Rose Yu. Incorporating symmetry into deep dynamics models for improved generalization. In *International Conference on Learning Representations*, 2021.

Rui Wang, 1 Robin Walters, and Rose Yu. Approximately equivariant networks for imperfectly symmetric dynamics. In *Proceedings of the 39 th International Conference on Machine Learning*, 2022.

Sijie Wang, Qiyu Kang, Rui She, Wei Wang, Kai Zhao, Yang Song, and Wee Peng Tay. Hypliloc: Towards effective lidar pose regression with hyperbolic fusion. In *Computer Vision and Pattern Recognition*, 2023.

Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. Dynamic graph cnn for learning on point clouds. *ACM Trans. Graph.*, 2019.

weitao Du, He Zhang, Yuanqi Du, Qi Meng, Wei Chen, Bin Shao, and Tie-Yan Liu. Equivariant vector field network for many-body system modeling. In *International Conference on Learning Representations*, 2022.

Daniel Worrall and Gabriel Brostow. Cubenet: Equivariance to 3d rotation and translation. In *European Conference on Computer Vision*, 2018.

Daniel E. Worrall, Stephan J. Garbin, Daniyar Turmukhambetov, and Gabriel J. Brostow. Harmonic networks: Deep translation and rotation equivariance. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Qi Xie, Qian Zhao, Zongben Xu, and Deyu Meng. Fourier series expansion based filter parametrization for equivariant convolutions. *IEEE Transactions on Pattern Analysis and Machine Intelligence.*, 2023.

YuQing Xie and Tess Smidt. Equivariant symmetry breaking sets. *Transactions on Machine Learning Research*, 2024.

Minkai Xu, Jiaqi Han, Aaron Lou, Jean Kossaifi, Arvind Ramanathan, Kamyar Azizzadenesheli, Jure Leskovec, Stefano Ermon, and Anima Anandkumar. Equivariant graph neural operator for modeling 3d dynamics. In *International Conference on Machine Learning*, 2024.

Jianke Yang, Robin Walters, Nima Dehmamy, and Rose Yu. Generative adversarial symmetry discovery. In *Fortieth International Conference on Machine Learning*, 2023.

APPENDIX

A.1 BACKGROUND CONCEPTS

This section summarizes key concepts used throughout the paper.

Irreducible Representations (Irreps) A representation $\rho : G \to \mathrm{GL}(V)$ is called *irreducible* if it admits no non-trivial invariant subspaces. Any finite-dimensional unitary representation of a compact group (such as $\mathrm{SO}(3)$) representation can be decomposed as a direct sum of irreducible components:

$$V \cong \bigoplus_{\ell} V_\ell. \tag{A.1.1}$$

In the case of $\mathrm{SO}(3)$, irreps are indexed by $\ell \in \mathbb{N}$ and correspond to spherical harmonic components. Features transforming under irreps provide a structured basis for equivariant networks.

Clebsch–Gordan Projection When combining two irreducible representations, their tensor product decomposes into a direct sum of irreps:

$$V_{\ell_1} \otimes V_{\ell_2} \;\cong\; \bigoplus_{\ell=|\ell_1-\ell_2|}^{\ell_1+\ell_2} V_\ell. \tag{A.1.2}$$

The coefficients defining this decomposition are known as *Clebsch–Gordan coefficients*. In practice, they are used to project tensor products of features back into irreducible components, forming the basis of equivariant message passing and convolution.

Partition of Unity A *partition of unity* subordinate to an open cover $\{U_i\}$ of a manifold is a collection of smooth functions $\{\lambda_i\}$ such that

$$\lambda_i \in C^\infty, \quad 0 \le \lambda_i \le 1, \quad \sum_i \lambda_i = 1, \tag{A.1.3}$$

with the support of each λ_i contained in U_i. It combines local constructions into a globally smooth one. In our setting, the cover has two sets, U_{hub} and U_{torus}. The weights are $\lambda_{\mathrm{hub}} = \alpha$ and $\lambda_{\mathrm{torus}} = 1 - \alpha$, where α is the hub weight (App. A.5.2).

Smoothness via Partition of Unity Let χ and σ be smooth maps for the toroidal and Euclidean regions. Their weighted combination

$$\Gamma = \lambda_{\mathrm{torus}}\, \chi + \lambda_{\mathrm{hub}}\, \sigma \tag{A.1.4}$$

is smooth, since it is a finite sum of products of smooth functions. The transition between the torus and the hub therefore introduces no discontinuities.

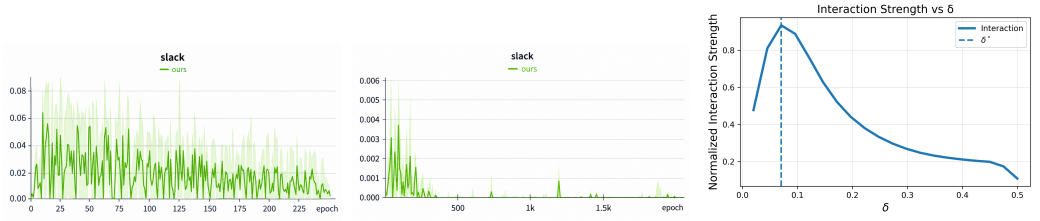


Figure 7: **(left:) Slack variation at training on Oxford RobotCar.(Middle:) Slack variation at training on N-body. (Right:) Interaction strength vs δ. Interaction is minimal in the factorized regime (low δ), peaks in an intermediate regime, and decreases as the model approaches a structured transformation at high δ.**

Table 4: **Analysis of the coupling parameter δ.** (a) Fixed vs. learned δ on N-body. (b) Final learned δ per dataset.

(a) Test MSE ($\times 10^{-3}$) on N-body under fixed and learned δ.

δ **config.**	**Test MSE**
Fixed $\delta = 0.2$	11.5 ± 0.17
Fixed $\delta = 0.8$	5.9 ± 0.13
Learned δ	$\mathbf{3.5 \pm 0.15}$

(b) Final learned δ. Higher δ is closer to strict equivariance. Trend shows the direction of δ during training.

Dataset	**Final** δ	**Trend**
ModelNet40 (rot., noisy)	≈ 0.875	$\downarrow$
N-body	≈ 0.964	$\uparrow$
Oxford RobotCar	≈ 0.862	$\downarrow$
QM9	≈ 0.910	$\uparrow$

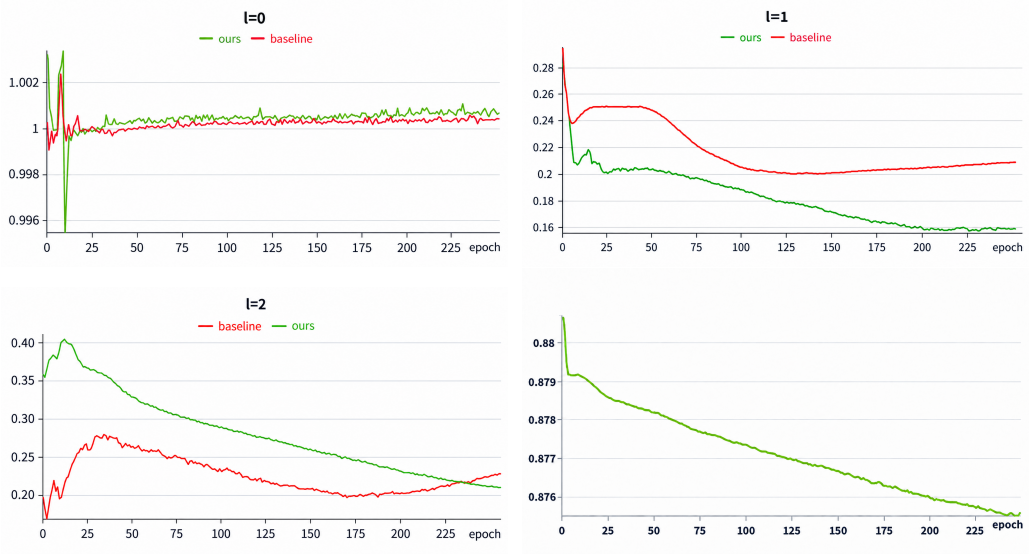


Figure 8: **Irreps signals with ℓ=0,1,2 variation of our model on ModelNet40 Wang et al. (2019).** From left to right clockwise: Irreps variation of ℓ=0 that is scalar, then ℓ=1 that is vector, then δ variation across epochs and ℓ=2 that is tensor. Here baseline denotes taking $\delta = 1.0$ as a strict equivariant model.

Generation of $\ell = 0, 2$ from $\ell = 1$. The argument has three steps.

(i) Self-duality: $\mathcal{D}^1$ is real and orthogonal, so $R(g)^{-T} = R(g)$. Its dual therefore transforms identically, $(\mathcal{D}^1)^* \cong \mathcal{D}^1$. Hence $\mathrm{End}(\mathcal{D}^1) \cong \mathcal{D}^1 \otimes (\mathcal{D}^1)^* \cong \mathcal{D}^1 \otimes \mathcal{D}^1$, with g acting on matrix $\mathbf{M} \in \mathrm{End}(\mathcal{D}^1)$ by conjugation, $\mathbf{M} \mapsto R(g)\mathbf{M}R(g)^T$.

(ii) Clebsch–Gordan: This 9-dimensional space splits into three invariant subspaces: $\mathcal{D}^0$ (multiples of $\mathbf{I}$), $\mathcal{D}^1$ (antisymmetric matrices), and $\mathcal{D}^2$ (traceless symmetric matrices), with dimensions $1+3+5 = 9$.

(iii) Closure and equivariance of the product: $\mathrm{End}(\mathcal{D}^1)$ is closed under matrix multiplication. Conjugation also respects products, since $R(\mathbf{AB})R^T = (R\mathbf{A}R^T)(R\mathbf{B}R^T)$. The product is therefore an equivariant bilinear map, and its output decomposes into the irreps of step (ii).

Explicit form. Write each $\ell = 1$ operator as $\hat{\mathbf{a}}$, with $\hat{\mathbf{a}}\mathbf{v} = \mathbf{a} \times \mathbf{v}$. Then

$$\hat{\mathbf{a}}\hat{\mathbf{b}} = \underbrace{-\tfrac{2}{3}(\mathbf{a}\cdot\mathbf{b})\,\mathbf{I}}_{\ell=0} + \underbrace{\tfrac{1}{2}\widehat{\mathbf{a}\times\mathbf{b}}}_{\ell=1} + \underbrace{\tfrac{1}{2}(\mathbf{b}\mathbf{a}^T + \mathbf{a}\mathbf{b}^T) - \tfrac{1}{3}(\mathbf{a}\cdot\mathbf{b})\,\mathbf{I}}_{\ell=2}. \tag{A.1.5}$$

These are exactly the Clebsch–Gordan components of $\mathbf{a} \otimes \mathbf{b}$: the dot product, the cross product, and the symmetric traceless outer product. Matrix multiplication of $\ell = 1$ operators thus realizes the tensor-product decomposition. For a square, $\mathbf{a} = \mathbf{b}$, the cross product vanishes: $\mathbf{A}_0^2 = -\frac{2}{3}\|\mathbf{a}\|^2\mathbf{I} + \left(\mathbf{a}\mathbf{a}^T - \frac{1}{3}\|\mathbf{a}\|^2\mathbf{I}\right)$, so $\mathbf{A}_0^2$ contains only $\ell = 0$ and $\ell = 2$. The $\ell = 2$ part is nonzero for any $\mathbf{a} \neq \mathbf{0}$.

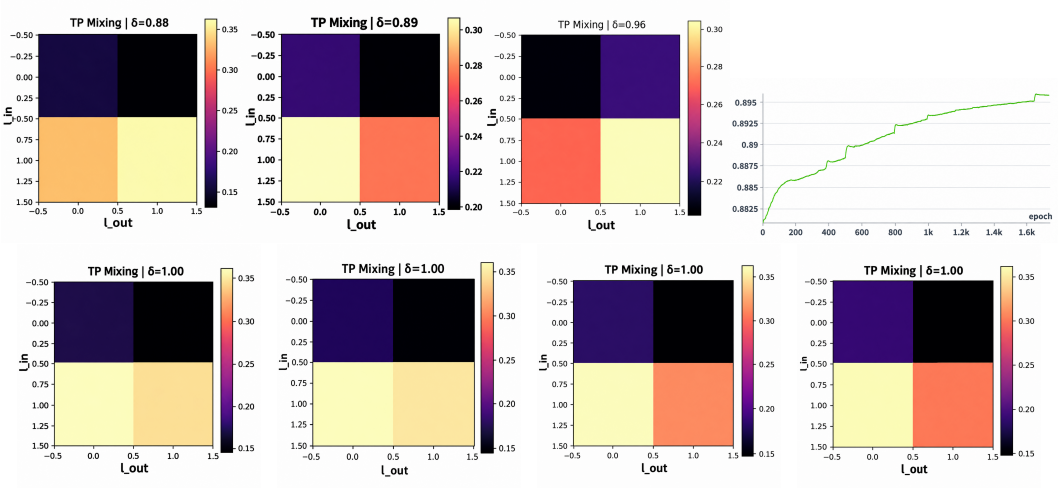


Figure 9: **Control of interaction strength via δ on N-body.** Tensor product (CG-TP) mixing matrices between irreps ($\ell_{\text{in}} \to \ell_{\text{out}}$) during training on the N-body dataset. **Top row (ours):** Evolution of mixing at epochs 5, 800, and 1600, together with the learned δ trajectory.**Bottom row (SEGNN baseline, $\delta = 1$):** Mixing matrices at corresponding epochs.

The same holds for $\mathbf{C}^2$. The symmetrized product $\mathbf{A}_0\mathbf{C} + \mathbf{C}\mathbf{A}_0$ cancels the $\ell = 1$ term, which explains the labels in Eq. 10.

Hadamard Lemma:

Lemma 1 (Hadamard, one-variable form)**.** If $f : \mathbb{R} \to V$ (V a normed vector space) is C^∞, then for any $a \in \mathbb{R}$,

$$f(\delta) = f(a) + (\delta - a)\, g(\delta), \qquad g(\delta) := \int_0^1 f'\big(a + t(\delta - a)\big)\, dt, \tag{A.1.6}$$

with g smooth and $g(a) = f'(a)$.

Extreme Value Theorem Let K be a compact set and $f : K \to \mathbb{R}$ continuous. Then f is bounded on K and attains its maximum and minimum:

$$\exists\, x_{\min}, x_{\max} \in K : \quad f(x_{\min}) \le f(x) \le f(x_{\max}) \quad \forall x \in K. \tag{A.1.7}$$

We apply it on the compact sets $G = \mathrm{SO}(3)$ and $\delta \in [0, 1]$. Any continuous quantity defined on them, such as an operator norm, therefore has a finite supremum. This makes the constants in our bounds finite.

Parity of Words in the Taylor Expansion The exponential of $\mathbf{A}(\delta) = \mathbf{A}_0 + \delta\mathbf{C}$ expands as

$$e^{\mathbf{A}_0 + \delta\mathbf{C}} = \sum_{k=0}^{\infty} \frac{1}{k!} (\mathbf{A}_0 + \delta\mathbf{C})^k = \sum_{k=0}^{\infty} \frac{1}{k!} \sum_{j=0}^{k} \delta^j\, \mathbf{S}_{k,j}, \tag{A.1.8}$$

where $\mathbf{S}_{k,j}$ is the sum of all length-k words containing j copies of $\mathbf{C}$ and $k - j$ copies of $\mathbf{A}_0$. Since $\mathbf{A}_0$ and $\mathbf{C}$ are antisymmetric, transposing a word reverses its order and multiplies it by $(-1)^k$:

$$(\mathbf{X}_1\mathbf{X}_2 \cdots \mathbf{X}_k)^T = (-1)^k\, \mathbf{X}_k \cdots \mathbf{X}_2\mathbf{X}_1, \qquad \mathbf{X}_i \in \{\mathbf{A}_0, \mathbf{C}\}. \tag{A.1.9}$$

Reversal maps the set of words in $\mathbf{S}_{k,j}$ onto itself. Hence $\mathbf{S}_{k,j}^T = (-1)^k\, \mathbf{S}_{k,j}$. Recall that antisymmetric matrices form the $\ell = 1$ subspace, and symmetric matrices form $\ell = 0 \oplus \ell = 2$. Therefore odd k contributes only to $\ell = 1$, and even k contributes only to $\ell = 0$ and $\ell = 2$. Grouping by ℓ and by power of δ gives

$$\big[e^{\mathbf{A}(\delta)}\big]_{\ell=1} = \sum_{k\ \text{odd}} \frac{1}{k!} \sum_{j=0}^{k} \delta^j\, \mathbf{S}_{k,j}, \qquad \big[e^{\mathbf{A}(\delta)}\big]_{\ell=0,2} = \sum_{k\ \text{even}} \frac{1}{k!} \sum_{j=0}^{k} \delta^j\, \mathbf{S}_{k,j}. \tag{A.1.10}$$

The $\ell = 0$ and $\ell = 2$ parts are then separated by the trace: $[\mathbf{M}]_{\ell=0} = \frac{1}{3}\mathrm{tr}(\mathbf{M})\,\mathbf{I}$ and $[\mathbf{M}]_{\ell=2} = \mathbf{M}_{\mathrm{sym}} - \frac{1}{3}\mathrm{tr}(\mathbf{M})\,\mathbf{I}$.

Schur's Lemma: A linear map $\mathcal{T} : V \to V'$ between two representations of G is *equivariant* if $\mathcal{T}\,\rho(g) = \rho'(g)\,\mathcal{T}$ for all $g \in G$.

Lemma 2 (Schur)**.** Let V and V' be irreducible representations of $\mathrm{SO}(3)$, and let $\mathcal{T} : V \to V'$ be equivariant.

(i) If $V \ncong V'$, then $\mathcal{T} = 0$.

(ii) If $V = V'$, then $\mathcal{T} = \lambda\,\mathrm{id}$ for some scalar $\lambda \in \mathbb{R}$.

Part (i) means equivariant maps cannot mix different degrees. Part (ii) means that an operator commuting with all rotations acts as a single scalar on each irreducible component. In Step 2 of App. A.3, $\mathrm{End}(\mathcal{D}^{\bar{\ell}})$ contains each degree ℓ exactly once. Hence the Casimir $\mathcal{K}$, which commutes with all rotations, acts as a scalar on each degree.

A.2 Additional Experiments

Here we further answer the following RQs along with 2 new RQs:

RQ2: *Equivariance control:* How does the learned δ shape the network's equivariance error during training?

RQ3: *Task dependence:* Does the preferred level of equivariance vary across tasks in Sec.?

RQ4: *Irrep composition:* Are the observed *irrep* components consistent with the graded scaling in Proposition 2?

RQ5: *Theory–empirical alignment:* Does the measured single-layer equivariance error follow the $\mathcal{O}(1-\delta)$ behavior predicted by Theorem 1 analyzed with a different metric in Sec. A.2.2?

RQ7. *Does δ control interaction strength:* Factorized vs coupled behavior. Precisely we try to measure $E_{\mathrm{comm}}(\delta) = \mathbb{E}_x\left[\|\psi(\rho_1\rho_2 x;\delta) - \psi(\rho_2\rho_1 x;\delta)\|\right]$.

RQ8. Is fixed δ beneficial over learnable δ. We try to understand whether the δ learning explained in Sec. 3.3 is required for a given task or not.

To further answer **RQ2**, does learning δ predictably reduce equivariance error? The equivariance constraint is enforced by the slack, which is $\max\big(0, \epsilon(\delta) - \tau\big)$. From Fig. 7, we can observe the slack variation over epochs for Oxford RobotCar Maddern et al. (2017) and N-body Kipf et al. (2018) datasets. We observe the gradual decrease of slack across epochs, which can indirectly validate the decrease in equivariance error of our model across different benchmarks.

To further answer **RQ4**, does varying delta activate higher-order equivariant components? From Fig. 8, we can observe that as δ follows a monotonic decreasing trend in the ModelNet40 classification task, ℓ =0 remains unaffected by δ as per proposition 2 and almost follows the same trend as the baseline. While ℓ=1,2 follows a decreasing trend across epochs, with ℓ=2 decreasing more sharply than ℓ=1. This validates that the activation of higher-order structured components follows the modulation of δ.

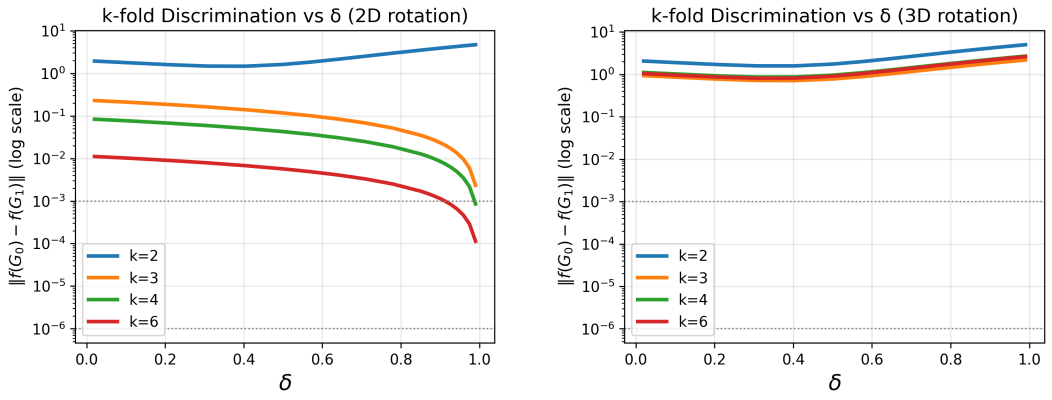
Another validation for **RQ4** the fig. 9(top row) shows that at initial epoch 5, ℓ=0,1 contribution is higher in our model than in the baseline, where higher ℓ=1 is already active. In the middle stage at epoch 800, we observe a gradual increase of ℓ=1 contribution in our model, while in the baseline, a decreased level of contribution of higher order ℓ. In the final stage at epoch 1600, the ℓ=1 is highly active in our model, roughly matching contributions with the baseline. Together, with these visual comparisons, we can say that our approach shows a continuous progress from factorized to coupled behaviour as δ increases across the epochs. This also validates the assumption 2,proposition 2.

To answer **RQ7**, we investigate whether δ controls the transition from factorized to coupled transformations. We measure $E_{\mathrm{comm}}(\delta)$, which quantifies the sensitivity of the learned representation to transformation ordering. Since $\mathrm{SO}(3)$ is intrinsically non-commutative, this metric does not isolate group commutators directly; instead, it measures the δ-controlled interaction strength of the

learned representation over ordering effects, serving as a proxy for interactions among model-induced transformation components.

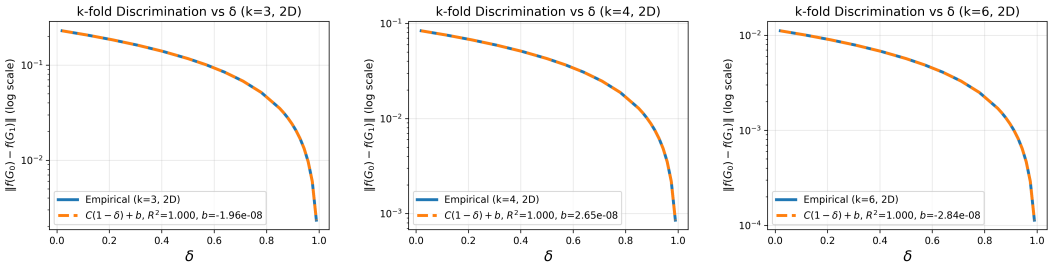
As shown in Fig. 7(right), interaction strength is low for small δ, peaks near the intermediate threshold δ^*, and decreases again as $\delta \to 1$. This reveals three regimes: (i) a factorized regime at low δ, where transformations act nearly independently, (ii) a hybrid regime near δ^*, where structured and non-equivariant components interact maximally, and (iii) a structured regime at high δ, where representations become more coherent and less order-sensitive.

This behavior is consistent with Eq. 8: for small δ, the factorized generators $\mathbf{W}$ dominate, while increasing δ strengthens the coupled component $\mathbf{C}$. The interaction peak near δ^* emerges from the coexistence of both terms.



(a) 2D rotation (about the shape's own axis). (b) 3D rotation (generic, non-coplanar).

Figure 10: Aggregate k-fold discrimination $\|f(G_0) - f(G_1)\|$ vs. δ, all $k \in \{2, 3, 4, 6\}$. Dotted lines mark the distinguishable (10^{-3}) and numerically-collapsed (10^{-6}) thresholds of Lin et al. Lin et al. (2026). In 2D mode, $k \geq 3$ decays monotonically toward the collapse floor as $\delta \to 1$; k=2 does not. In 3D mode all four curves share the same non-monotonic shape.



(a) k=3: R^2=1.000. (b) k=4: R^2=1.000. (c) k=6: R^2=1.000.

Figure 11: Exact confirmation: $k \geq 3$, 2D mode. The zero-intercept fit $C(1 - \delta)$ — the same functional form as Theorem 1 — matches the empirical curve across the *entire* δ range, not just near δ=1, with fitted intercepts of order 10^{-8}.

To answer **RQ8**, we analyze whether learning δ is necessary, or whether fixed coupling configurations can achieve comparable performance. Tab. 4a compares the proposed learnable formulation against fixed low- and high-coupling regimes on the N-body benchmark. While stronger fixed coupling ($\delta = 0.8$) significantly improves performance over weak coupling ($\delta = 0.2$), the learnable formulation achieves the lowest test error overall. This suggests that adaptive modulation of equivariant coupling provides additional flexibility beyond static interpolation settings.

A.2.1 ADDITIONAL ABLATION STUDY

The Tab. 4b shows the final converged δ values for every task. The table also shows that the equivariance is task-dependent, as indicated by the "Trend" column supporting the claim of **RQ3** that equivariance is task dependent.

A.2.2 RQ5: THEORY–EMPIRICS ALIGNMENT

RQ5. *Theory–empirics alignment. Validating Theorem 1.*

The main text validates Theorem 1 by measuring the equivariance defect $\epsilon(\delta) = \|F(\rho(g)x) - \hat{\rho}(g)F(x)\|$ directly at a single layer. Here we test whether the same $(1-\delta)$ law governs a second, structurally unrelated quantity: the model's ability to *discriminate* rotationally symmetric inputs. This gives an independent, falsifiable check on the mechanism behind Theorem 1 rather than a repeated measurement of the same one.

Setup. Following the symmetry-increase framework of Lin et al. (Lin et al., 2026), we construct k-fold planar point clouds ($k \in \{2, 3, 4, 6\}$): a center point at the origin plus k points on the unit circle at angles $2\pi i/k$. For each k we generate a rotated copy G_1 of the base structure G_0 under two protocols: a *2D* rotation about the shape's own axis (excluding angles that coincide with a genuine C_k symmetry element), and a *3D* rotation, a generic $\mathrm{SO}(3)$ element constrained to tilt the shape's plane by at least $15°$. We pass both through a single layer ($\texttt{irreps} = \ell \in \{0, 1, 2\}$) across the same δ grid used to validate Theorem 1, and report $\|f(G_0) - f(G_1)\|$, mean-pooled over nodes and averaged over 10 trials per (k, mode, δ).

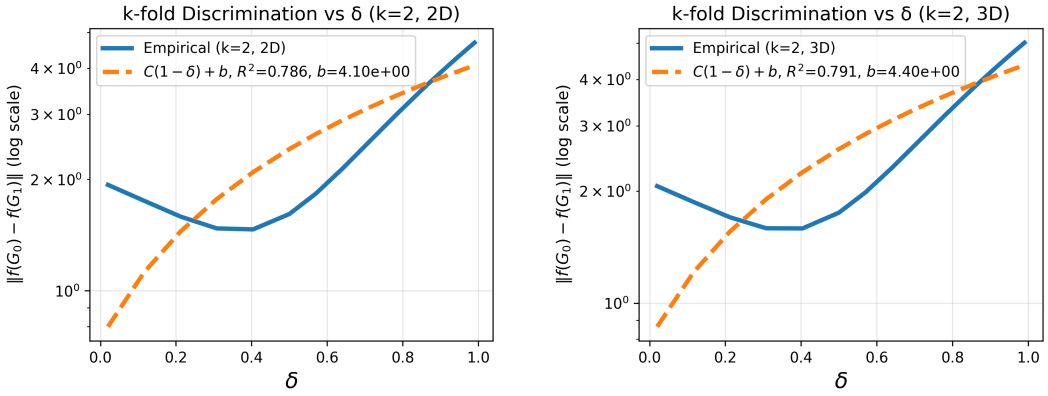
Prediction. Only proper rotations are applied in this sweep, so the relevant input isotropy is $C_k \leq \mathrm{SO}(3)$, not the full dihedral group. Using the $\mathrm{SO}(3)$-only symmetry-infimum table of Lin et al. and the direct-sum rule for representations of $\mathrm{SO}(3)$ (every closed subgroup there satisfies the bottleneck condition, so the infimum of a direct sum is the minimum over its summands): the ℓ=0 block is always fully degenerate; the ℓ=1 block's infimum is C_∞ for every $k \geq 2$; the ℓ=2 block resolves C_2 *exactly* when k=2 (l_0=2 $\geq k$) but only reaches D_∞ for $k \geq 3$. Combined, the equivariant branch's symmetry infimum is exactly C_2 for k=2 (no degeneration) and exactly C_∞ for $k \geq 3$ — invariance under *every* rotation about the shape's own axis, at *every* δ, since the infimum depends on which degrees are present, not their magnitude. Because our 2D protocol is exactly such a rotation, this yields a sharp, δ-independent prediction: for $k \geq 3$, 2D mode, the equivariant branch's contribution to $\|f(G_0) - f(G_1)\|$ is identically zero at every δ, so the *entire* observed curve must be carried by the $(1-\delta)W$ branch — i.e. it should obey the same linear law as Theorem 1, forced through the origin at δ=1, across the full domain rather than only asymptotically. For k=2 the opposite is predicted (flat, non-vanishing discrimination at every δ); for the 3D protocol, C_∞-invariance about one axis does not constrain a generic out-of-plane rotation, so no equally sharp claim applies and 3D mode serves as a descriptive control.

Results: exact confirmation for $k \geq 3$, 2D. Figure 10 shows the aggregate discrimination curves across all k for both rotation modes. Figure 11 isolates the three cases the theory makes an exact claim about: $k \in \{3, 4, 6\}$, 2D mode. Fitting $\|f(G_0) - f(G_1)\| = C(1-\delta)$ (zero intercept, matching the functional form of Theorem 1) gives $R^2 = 1.000$ for all three, with fitted intercepts at δ=1 of order 10^{-8} — the float32 noise floor, not a residual trend. This confirms the prediction to numerical precision: the equivariant branch contributes *nothing* to in-plane discrimination for $k \geq 3$, at any δ, and $(1-\delta)$ alone governs the entire signal. Table 5 reports the complete fit summary across all (k, mode) combinations.

k=2 and all 3D-mode panels: correctly predicted, wrong functional family for the straight-line diagnostic. As predicted, k=2 does not collapse: Figure 12 shows flat, non-vanishing discrimination across the full δ range in both rotation modes, and Table 5 reports the negative R^2_0 that follows from forcing a zero-intercept line onto a curve the theory never claimed was linear through the origin. The free-intercept fit ($R^2_{\text{free}} \approx 0.79$) is a better diagnostic but still visibly does not track the curve's interior: every one of these five panels (k=2 in both modes, plus $k \in \{3, 4, 6\}$ in 3D mode) is non-monotonic — a shallow dip near $\delta \approx 0.3$–0.4 followed by a steep, convex rise toward δ=1 — a shape no straight line can represent (Figure 13).

Table 5: Linear-law fit summary, $\|f(G_0)-f(G_1)\| \approx C_0(1-\delta)$ (zero intercept) and $\approx C_1(1-\delta)+b$ (free intercept), across all (k, mode) pairs. b/peak is the fitted $\delta{=}1$ intercept as a fraction of the curve's own peak value.

k	mode	C_0	R_0^2	b (at $\delta{=}1$)	b/peak	R^2_{free}	verdict
2	2D	3.3745	−5.067	4.1027	0.875	0.786	matches theory (flat, nonzero)
2	3D	3.6285	−5.122	4.4043	0.877	0.791	no sharp prediction (descriptive)
3	2D	0.2338	**1.000**	-1.96×10^{-8}	0.000	1.000	**matches theory (exact zero)**
3	3D	1.6162	−5.337	1.9183	0.876	0.782	no sharp prediction (descriptive)
4	2D	0.0853	**1.000**	2.65×10^{-8}	0.000	1.000	**matches theory (exact zero)**
4	3D	1.9673	−5.216	2.3756	0.879	0.796	no sharp prediction (descriptive)
6	2D	0.0114	**1.000**	-2.84×10^{-8}	0.000	1.000	**matches theory (exact zero)**
6	3D	1.8715	−4.996	2.3088	0.879	0.801	no sharp prediction (descriptive)



(a) $k{=}2$, 2D mode.

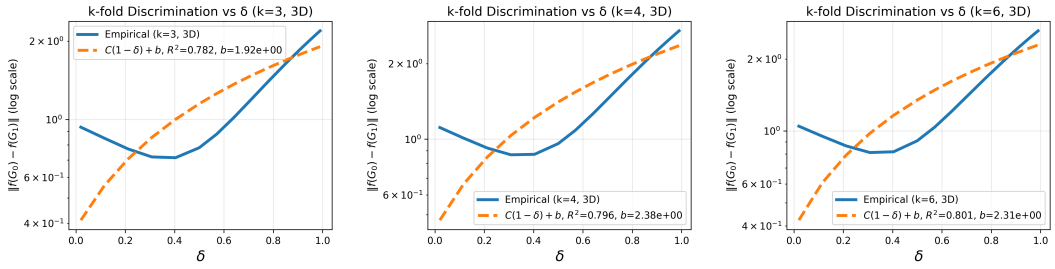
(b) $k{=}2$, 3D mode.

Figure 12: $k{=}2$: correctly predicted exception. The equivariant branch's $\ell{=}2$ block resolves C_2 exactly ($l_0{=}2 \geq k{=}2$), so discrimination does not vanish as $\delta \to 1$. The straight-line overlay is shown only as a diagnostic, not a claimed functional form; see Figure 15 for the correct fit.

Per-degree decomposition. Figure 14 decomposes discrimination by irrep degree ℓ for representative cases. The $\ell{=}0$ and $\ell{=}1$ blocks decay smoothly toward the numerical floor as $\delta \to 1$ in every case; the $\ell{=}2$ block alone carries the non-monotonic, U-shaped signature, consistent with it being the sole surviving channel for $k{=}2$ and the dominant channel elsewhere.

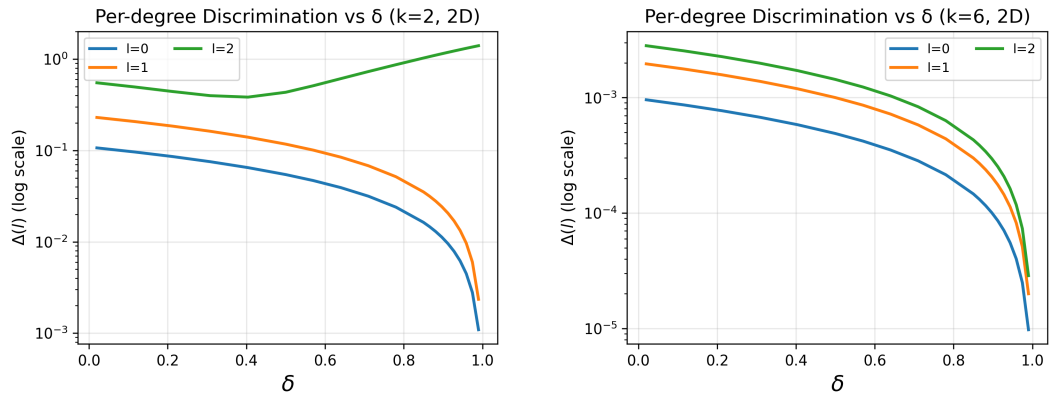
Quadratic refit for the non-collapsing panels. The observed U-shape is consistent with a two-term competition: the surviving $\ell{=}2$ equivariant content, attenuated by the model's δ^ℓ scaling (Proposition 2), against the $(1-\delta)$-gated non-equivariant branch, i.e. $\|f(G_0) - f(G_1)\| \approx A\delta^2 + B(1-\delta)$. Refitting the five non-collapsing panels against this form improves R^2 from 0.78–0.80 (linear) to 0.97–0.98 uniformly (Table 6, Figure 15). Notably, the fitted minimum location $\delta^* = B/(2A)$ is nearly constant across all five panels (0.169–0.184) despite A and B individually varying by a factor of $\sim$2.4$\times$ across different k and rotation modes — consistent with δ^* reflecting an intrinsic property of the trained layer (the relative typical magnitude of its $\ell{=}2$ tensor-product output versus its non-equivariant branch) rather than a property of any particular test geometry.

Summary. The $(1-\delta)$ mechanism underlying Theorem 1 reproduces exactly, to numerical precision, in a second, structurally independent quantity: for every input class the theory identifies as unresolvable by the equivariant branch alone ($k \geq 3$, in-plane rotation), discrimination follows the identical linear law, forced through the origin at $\delta{=}1$, across the entire domain. Where the theory does not predict a collapse ($k{=}2$; any out-of-plane rotation), none is observed, and the deviation from a straight-line fit is itself explained by a second, physically motivated two-term law rather than left as an unexplained residual.



(a) k=3, 3D. (b) k=4, 3D. (c) k=6, 3D.

Figure 13: 3D mode, $k \geq 3$: no sharp claim is made here (see Setup), and none is violated. All three panels share the same non-monotonic shape as k=2's, essentially independent of k ($R^2_{\text{free}} \approx 0.78$–$0.80$, intercept fraction $\approx$ 0.876–0.879 throughout) — consistent with C_∞-invariance about the shape's own axis simply not constraining a generic out-of-plane rotation.



(a) k=2, 2D: ℓ=2 non-monotonic, ℓ=0, 1 decay. (b) k=6, 2D: all degrees decay to the numerical floor.

Figure 14: Per-degree discrimination $\Delta(\ell)$ vs. δ. Left: k=2, where ℓ=2 alone survives structurally and rises toward δ=1. Right: k=6, where every degree — including ℓ=2 — decays monotonically, consistent with the exact-zero result in Figure 11. Additional per-degree panels for $k \in \{3, 4\}$ and all 3D-mode cases follow the same two qualitative patterns and are omitted here for space.

A.3 Justification of Proposition 2

We motivate Proposition 2 and bound its error. Irrep degree ℓ corresponds to angular frequency on $\mathrm{SO}(3)$. Spherical harmonics of degree ℓ are eigenfunctions of the angular Laplacian with eigenvalue $\ell(\ell+1)$, which increases with ℓ. By the spectral bias of neural networks toward low frequencies Rahaman et al. (2019), higher-ℓ components are learned less readily. They are therefore natural candidates for stronger attenuation as δ decreases.

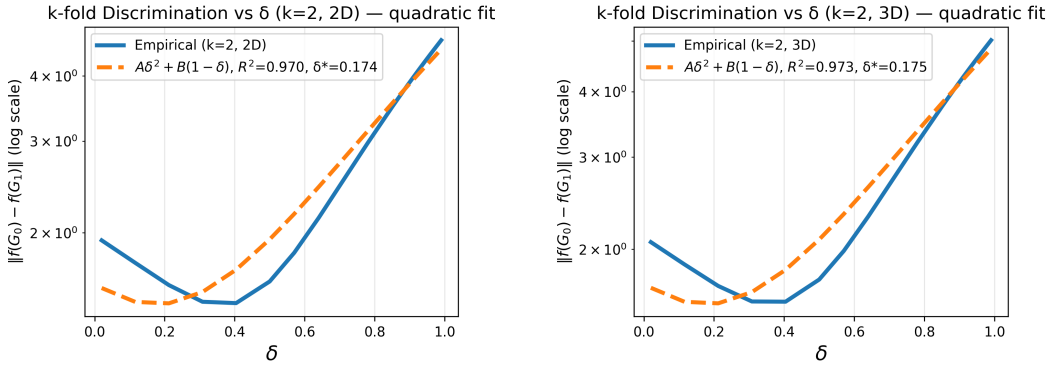
We derive the result for a general representation $\mathcal{D}^{\bar{\ell}}$. The main paper uses $\bar{\ell} = 1$, which gives degrees $\ell \leq 2$ (Remark 3).

Setup. We write $\bar{\ell}$ for the degree of the representation on which the operators act, and ℓ for the degree of a component of their products. $\bar{\ell}$ is fixed, while ℓ ranges over $0, \ldots, 2\bar{\ell}$. Let $\mathbf{J}_x, \mathbf{J}_y, \mathbf{J}_z$ be the real antisymmetric generators of $\mathfrak{so}(3)$ on $\mathcal{D}^{\bar{\ell}}$, of size $n \times n$ with $n = 2\bar{\ell} + 1$, satisfying $[\mathbf{J}_i, \mathbf{J}_j] = \epsilon_{ijk}\mathbf{J}_k$. At each point, $\mathbf{A}_0$ and $\mathbf{C}$ are linear combinations of $\mathbf{J}_x, \mathbf{J}_y, \mathbf{J}_z$, so both are antisymmetric. $\mathbf{W}$ is an unconstrained $n \times n$ matrix. We study

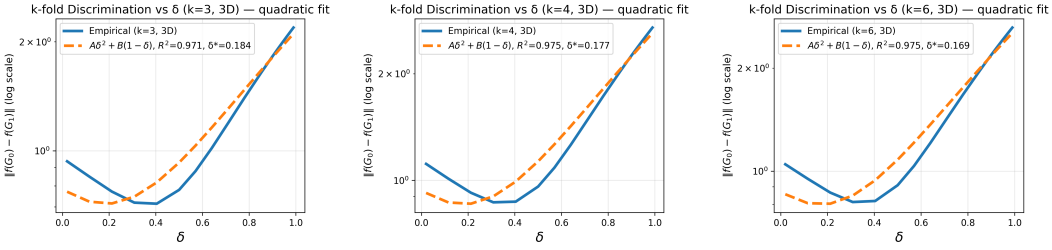
$$\psi(\delta) = \exp\big(\mathbf{A}(\delta)\big), \qquad \mathbf{A}(\delta) = \mathbf{A}_0 + \delta\mathbf{C} + (1-\delta)\mathbf{W}. \tag{A.3.1}$$

Table 6: Quadratic-law fit, $\|f(G_0) - f(G_1)\| \approx A\delta^2 + B(1-\delta)$, for the five panels where the zero-intercept linear law does not apply. $\delta^* = B/(2A)$ is the fitted interior minimum.

k	mode	A	B	R^2	δ^*
2	2D	4.5772	1.5969	0.970	0.174
2	3D	4.9117	1.7210	0.973	0.175
3	3D	2.1396	0.7852	0.971	0.184
4	3D	2.6482	0.9388	0.975	0.177
6	3D	2.5741	0.8718	0.975	0.169



(a) k=2, 2D mode: R^2=0.970. (b) k=2, 3D mode: R^2=0.973.



(c) k=3, 3D mode: R^2=0.971. (d) k=4, 3D mode: R^2=0.975. (e) k=6, 3D mode: R^2=0.975.

Figure 15: Quadratic refit $A\delta^2 + B(1-\delta)$ vs. the empirical curve, k=2 (both modes). R^2 improves from 0.786/0.791 (linear, Figure 12) to 0.970/0.973. Figures for $k \in \{3, 4, 6\}$, 3D mode are structurally identical (see Table 6 for their numbers) and can be added here once the remaining plot files are available.

Throughout, $\|\cdot\|$ is the Frobenius norm and $\|\cdot\|_{\mathrm{op}}$ the operator norm. We use $\|\mathbf{MN}\| \le \|\mathbf{M}\|_{\mathrm{op}}\|\mathbf{N}\|$, $\|\mathbf{M}\|_{\mathrm{op}} \le \|\mathbf{M}\|$, and $\|\mathbf{UMV}\| = \|\mathbf{M}\|$ for orthogonal $\mathbf{U}, \mathbf{V}$. by cyclicity of the trace. We write $a = \|\mathbf{A}_0\|$ and $c = \|\mathbf{C}\|$. All series below converge absolutely, so regrouping their terms is valid.

Step 1: Separating the equivariant and flat parts. Expanding $\psi(\delta) = \sum_{k\ge 0} \mathbf{A}(\delta)^k/k!$, each $\mathbf{A}(\delta)^k$ is a sum of 3^k ordered words over $\{\mathbf{A}_0, \mathbf{C}, \mathbf{W}\}$. The words without $\mathbf{W}$ sum to $(\mathbf{A}_0 + \delta\mathbf{C})^k$. Every word containing $\mathbf{W}$ carries at least one factor $(1-\delta)$. Hence

$$\psi(\delta) = \psi_{\mathrm{eq}}(\delta) + (1-\delta)\,\psi_{\mathrm{flat}}(\delta), \qquad \psi_{\mathrm{eq}}(\delta) = \exp(\mathbf{A}_0 + \delta\mathbf{C}). \tag{A.3.2}$$

Since $\mathbf{A}_0 + \delta\mathbf{C} \in \mathfrak{so}(3)$, $\psi_{\mathrm{eq}}(\delta)$ is orthogonal and exactly equivariant for every δ. All equivariance error lies in $(1-\delta)\psi_{\mathrm{flat}}$, which Theorem 1 bounds.

Step 2: Degree decomposition of $\mathrm{End}(\mathcal{D}^{\bar\ell})$. Every irrep of $\mathrm{SO}(3)$ is real. In a real orthonormal basis, $\rho_{\bar\ell}(g)$ is orthogonal, so $\rho_{\bar\ell}(g)^{-T} = \rho_{\bar\ell}(g)$ and $(\mathcal{D}^{\bar\ell})^* \cong \mathcal{D}^{\bar\ell}$. A rotation acts on $n \times n$ matrices

by conjugation, $\mathbf{M} \mapsto \rho_{\bar{\ell}}(g)\mathbf{M}\rho_{\bar{\ell}}(g)^T$. Under this action,

$$\mathrm{End}(\mathcal{D}^{\bar{\ell}}) \cong \mathcal{D}^{\bar{\ell}} \otimes (\mathcal{D}^{\bar{\ell}})^* \cong \mathcal{D}^{\bar{\ell}} \otimes \mathcal{D}^{\bar{\ell}} \cong \bigoplus_{\ell=0}^{2\bar{\ell}} \mathcal{D}^{\ell}. \tag{A.3.3}$$

For $\bar{\ell} = 1$, the trace separates the two symmetric degrees $\ell = 0, 2$ (Remark 3). For $\bar{\ell} \geq 2$, the symmetric part contains all even degrees up to $2\bar{\ell}$, and we need a general way to separate them.

Generators of the conjugation action. Differentiating the conjugation by $e^{t\mathbf{J}_i}$ at $t = 0$ gives

$$\frac{d}{dt}\, e^{t\mathbf{J}_i}\, \mathbf{M}\, e^{-t\mathbf{J}_i}\Big|_{t=0} = [\mathbf{J}_i, \mathbf{M}]. \tag{A.3.4}$$

So the maps $[\mathbf{J}_i, \cdot\,]$ are the infinitesimal rotations acting on matrices.

Casimir. We define

$$\mathcal{K}(\mathbf{M}) = -\sum_i [\mathbf{J}_i, [\mathbf{J}_i, \mathbf{M}]]. \tag{A.3.5}$$

This is the matrix analogue of the angular Laplacian. $\mathcal{K}$ commutes with every rotation, so by Schur's lemma it acts as a scalar on each irreducible component. On degree ℓ this scalar is $\ell(\ell+1)$, as for spherical harmonics of degree ℓ. For example, $\mathcal{K}(\mathbf{I}) = \mathbf{0}$, since $\mathbf{I}$ commutes with every $\mathbf{J}_i$. This is the degree-0 component.

Projectors. The eigenvalues $\ell(\ell+1)$ are distinct for different ℓ. Hence the polynomial in $\mathcal{K}$ that equals 1 at $\ell(\ell+1)$ and 0 at every other eigenvalue projects onto degree ℓ:

$$\pi_\ell = \prod_{\substack{\ell'=0 \\ \ell' \neq \ell}}^{2\bar{\ell}} \frac{\mathcal{K} - \ell'(\ell'+1)}{\ell(\ell+1) - \ell'(\ell'+1)}. \tag{A.3.6}$$

On degree ℓ, every factor equals 1. On any other degree ℓ', the factor for ℓ' equals 0.

Each $[\mathbf{J}_i, \cdot\,]$ is skew-adjoint under the Frobenius inner product, since $\mathbf{J}_i$ is antisymmetric. Hence $\mathcal{K}$ is self-adjoint, and the π_ℓ are orthogonal projectors, with $\|\pi_\ell(\mathbf{M})\| \leq \|\mathbf{M}\|$ and $\sum_\ell \pi_\ell = \mathrm{id}$. Since $\mathcal{K}$ commutes with conjugation, so does each π_ℓ. We define the degree-ℓ part $\psi_{\mathrm{eq},\ell}(\delta) = \pi_\ell\big(\psi_{\mathrm{eq}}(\delta)\big)$.

Step 3: Which words reach which degree. Let $\mathbf{S}_{k,j}$ be the sum of all length-k words over $\{\mathbf{A}_0, \mathbf{C}\}$ with exactly j copies of $\mathbf{C}$. Then

$$\psi_{\mathrm{eq}}(\delta) = \sum_{k \geq 0} \frac{1}{k!} \sum_{j=0}^{k} \delta^j\, \mathbf{S}_{k,j}. \tag{A.3.7}$$

Two rules fix which degrees $\mathbf{S}_{k,j}$ can reach.

Degree bound. Conjugation respects products, so matrix multiplication is an equivariant bilinear map. $\mathbf{A}_0$ and $\mathbf{C}$ lie in the degree-1 subspace spanned by the $\mathbf{J}_i$. By the Clebsch–Gordan rule $\mathcal{D}^{\ell_1} \otimes \mathcal{D}^1 \subset \bigoplus_{\ell \leq \ell_1+1} \mathcal{D}^\ell$ and induction on k, every length-k word has degree at most k.

Parity. Since $\mathbf{A}_0$ and $\mathbf{C}$ are antisymmetric, transposing a word reverses it and multiplies it by $(-1)^k$. Reversal preserves the count of $\mathbf{C}$'s, so $\mathbf{S}_{k,j}^T = (-1)^k \mathbf{S}_{k,j}$. Symmetric matrices correspond to $\mathrm{Sym}^2(\mathcal{D}^{\bar{\ell}})$, which contains exactly the even degrees. Antisymmetric matrices correspond to $\Lambda^2(\mathcal{D}^{\bar{\ell}})$, which contains exactly the odd degrees. This follows from the Clebsch–Gordan symmetry $\langle \bar{\ell} m_1 \bar{\ell} m_2 | \ell m \rangle = (-1)^{2\bar{\ell}-\ell} \langle \bar{\ell} m_2 \bar{\ell} m_1 | \ell m \rangle$.

Together, $\pi_\ell(\mathbf{S}_{k,j}) = \mathbf{0}$ unless $k \geq \ell$ and $k \equiv \ell \pmod 2$. Therefore

$$\psi_{\mathrm{eq},\ell}(\delta) = \sum_{j \geq 0} \delta^j\, T_j^{(\ell)}, \qquad T_j^{(\ell)} = \sum_{\substack{k \geq \max(j,\ell) \\ k \equiv \ell \pmod 2}} \frac{1}{k!}\, \pi_\ell(\mathbf{S}_{k,j}). \tag{A.3.8}$$

Parity of k fixes the degree. The number j of $\mathbf{C}$'s fixes the power of δ. The two are independent, so every power δ^j can appear in every degree ℓ.

Step 4: The floor. Setting $\delta = 0$ keeps only $j = 0$:

$$\psi_{\mathrm{eq},\ell}(0) = T_0^{(\ell)} = \pi_\ell\big(\exp(\mathbf{A}_0)\big). \tag{A.3.9}$$

We call this the *floor* at degree ℓ. It is built from $\mathbf{A}_0$ alone and is nonzero in general. For $\ell \geq 2$, the terms with $1 \leq j < \ell$ also carry fewer than ℓ powers of δ. The next step isolates all of them.

Step 5: Exact split by repeated Hadamard. Fix $\ell \geq 1$. Applying Lemma 1 ℓ times at $\delta = 0$ gives Taylor's theorem with integral remainder:

$$\psi_{\mathrm{eq},\ell}(\delta) = \underbrace{\sum_{j<\ell} \delta^j T_j^{(\ell)}}_{\mathbf{D}_\ell(\delta)} + \delta^\ell\, G_\ell(\delta), \qquad G_\ell(\delta) = \frac{1}{(\ell-1)!}\int_0^1 (1-t)^{\ell-1}\, \psi_{\mathrm{eq},\ell}^{(\ell)}(t\delta)\, dt. \tag{A.3.10}$$

This split is exact. $\mathbf{D}_\ell$ collects every term with fewer than ℓ powers of δ, including the floor. We further split the kept part at its leading coefficient:

$$\delta^\ell G_\ell(\delta) = \delta^\ell\, T_\ell^{(\ell)} + \sum_{j>\ell} \delta^j T_j^{(\ell)}. \tag{A.3.11}$$

This gives the monomial law $\psi_{\mathrm{eq},\ell}(\delta) \approx \delta^\ell\, T_\ell^{(\ell)}$: δ^ℓ times a δ-independent coefficient. Its error is

$$\mathbf{R}_\ell(\delta) = \mathbf{D}_\ell(\delta) + \sum_{j>\ell} \delta^j T_j^{(\ell)}. \tag{A.3.12}$$

Step 6: Bounding each part. A length-k word with j copies of $\mathbf{C}$ has norm at most $a^{k-j}c^j$ for $k \geq 1$, and there are $\binom{k}{j}$ such words. By Step 3, $T_j^{(\ell)}$ collects only words with $k \geq \ell \geq 1$, and π_ℓ is contractive. Using $\binom{k}{j}/k! = 1/\big(j!\,(k-j)!\big)$ and $\sum_{m\geq p} a^m/m! \leq e^a a^p/p!$,

$$\|T_j^{(\ell)}\| \leq e^a\, \frac{c^j}{j!}\, \frac{a^{(\ell-j)^+}}{\big((\ell-j)^+\big)!}, \qquad (x)^+ = \max(x, 0). \tag{A.3.13}$$

Dropped part. Summing over $j < \ell$ with the binomial theorem,

$$\|\mathbf{D}_\ell(\delta)\| \leq \frac{e^a}{\ell!}\Big[(a+\delta c)^\ell - (\delta c)^\ell\Big]. \tag{A.3.14}$$

Higher powers. Summing over $j > \ell$ with $(\ell+i)! \geq \ell!\, i!$,

$$\Big\|\sum_{j>\ell} \delta^j T_j^{(\ell)}\Big\| \leq e^a\, \frac{(\delta c)^\ell}{\ell!}\big(e^{\delta c} - 1\big). \tag{A.3.15}$$

Kept coefficient. By parity, the words in $T_\ell^{(\ell)}$ are $\mathbf{C}^\ell$ and words with at least two copies of $\mathbf{A}_0$. Define

$$\gamma_\ell = \frac{\|\pi_\ell(\mathbf{C}^\ell)\|}{\|\mathbf{C}\|^\ell}. \tag{A.3.16}$$

Both norms are rotation-invariant and homogeneous of degree ℓ in $\mathbf{C}$, so $\gamma_\ell > 0$ depends only on ℓ and $\bar{\ell}$. For instance, $\gamma_1 = 1$ for every $\bar{\ell}$, and $\gamma_2 = 1/\sqrt{6}$ for $\bar{\ell} = 1$. Hence

$$\|T_\ell^{(\ell)}\| \geq \frac{c^\ell}{\ell!}\Big(\gamma_\ell - \tfrac{1}{2}a^2 e^a\Big). \tag{A.3.17}$$

Step 7: Relative error of the monomial law. Let $r = a/(\delta c)$ and $\eta = \delta c$, so that $a = r\eta$. Dividing the bounds of Step 6, whenever the denominator is positive,

$$\frac{\|\mathbf{R}_\ell(\delta)\|}{\|\delta^\ell T_\ell^{(\ell)}\|} \leq \frac{e^a\Big[(1+r)^\ell - 1 + e^\eta - 1\Big]}{\gamma_\ell - \frac{1}{2}a^2 e^a} = \frac{\ell\, r + \eta}{\gamma_\ell} + O\big(r^2 + \eta^2\big). \tag{A.3.18}$$

The term ℓr measures the dropped lower powers, including the floor. The term η measures the higher powers of δ. For fixed $\ell \leq 2\bar{\ell}$, the bound tends to zero as $r, \eta \to 0$. Then $\psi_{\mathrm{eq},\ell}(\delta) = \delta^\ell\, T_\ell^{(\ell)}\big(1+o(1)\big)$, and δ^ℓ is the leading power in degree ℓ. The degree $\ell = 0$ is not attenuated. Its variation is bounded by $\|\psi_{\mathrm{eq},0}(\delta) - \psi_{\mathrm{eq},0}(0)\| \leq \delta c$, by the derivative bound of Step 8.

Step 8: A global bound anchored at $\delta = 1$. The following bound holds on all of $[0, 1]$, for any r and η. Define $\mathbf{E}_\ell(\delta) = \psi_{\text{eq},\ell}(\delta) - \delta^\ell \psi_{\text{eq},\ell}(1)$, which vanishes at $\delta = 1$. Lemma 1 at $\delta = 1$ gives, exactly,

$$\mathbf{E}_\ell(\delta) = (\delta - 1) \int_0^1 \mathbf{E}'_\ell\big(1 + t(\delta - 1)\big)\, dt. \tag{A.3.19}$$

With $\mathbf{X}_\delta = \mathbf{A}_0 + \delta\mathbf{C}$, the derivative of the matrix exponential is

$$\psi'_{\text{eq}}(\delta) = \int_0^1 e^{s\mathbf{X}_\delta}\, \mathbf{C}\, e^{(1-s)\mathbf{X}_\delta}\, ds. \tag{A.3.20}$$

Both exponentials are orthogonal, so $\|\psi'_{\text{eq},\ell}(\delta)\| \le c$. Also $\|\psi_{\text{eq},\ell}(1)\| \le \|\psi_{\text{eq}}(1)\| = \sqrt{n}$. With $\frac{d}{d\delta}\delta^\ell \le \ell$ on $[0, 1]$,

$$\big\|\psi_{\text{eq},\ell}(\delta) - \delta^\ell\, \psi_{\text{eq},\ell}(1)\big\| \le (1 - \delta)\big(c + \ell\sqrt{n}\big), \qquad \delta \in [0, 1]. \tag{A.3.21}$$

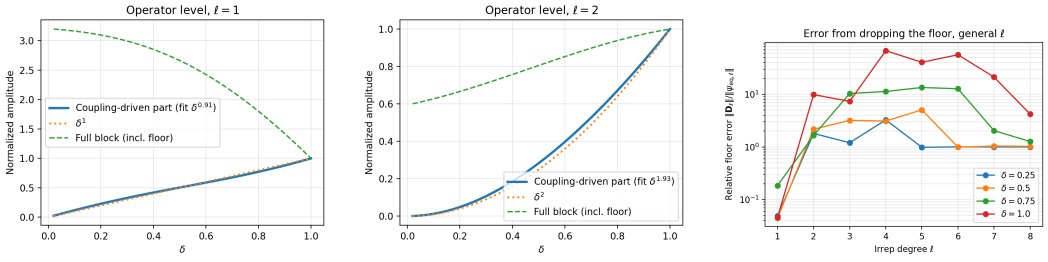
Summing over the mutually orthogonal degrees,

$$\Big\|\psi_{\text{eq}}(\delta) - \sum_{\ell=0}^{2\bar{\ell}} \delta^\ell\, \psi_{\text{eq},\ell}(1)\Big\| \le (1 - \delta)\Big(\sum_{\ell=0}^{2\bar{\ell}} \big(c + \ell\sqrt{n}\big)^2\Big)^{1/2}. \tag{A.3.22}$$

Step 7 gives a relative error that is small when r and η are small. Step 8 gives an absolute error that vanishes at $\delta = 1$ and grows linearly as δ decreases.

Step 9: The approximation costs fidelity, not equivariance. $\psi_{\text{eq},\ell}(\delta)$ is equivariant, because $\psi_{\text{eq}}(\delta)$ is (Step 1) and π_ℓ commutes with the group action (Step 2). Scaling each degree by a real scalar also commutes with the group action. Hence both $\delta^\ell T_\ell^{(\ell)}$ and $\delta^\ell \psi_{\text{eq},\ell}(1)$ are equivariant for every δ. The bounds equation A.3.18 and equation A.3.21 measure fidelity to the idealized construction, not a loss of equivariance. The only source of equivariance error remains $(1 - \delta)\psi_{\text{flat}}$ in equation A.3.2, bounded by Theorem 1.

Step 10: The attenuation law. Steps 5–7 identify δ^ℓ as the leading power in degree ℓ once the lower powers are dropped. Step 8 bounds the error of the law on all of $[0, 1]$. We therefore use $c_\ell(\delta) = \delta^\ell$ for $\ell = 0, \ldots, 2\bar{\ell}$. It uses a single global scalar. It is exact at $\delta = 1$. It leaves $\ell = 0$ unscaled. For $\delta \in (0, 1)$, δ^ℓ decreases strictly in ℓ, so higher angular frequencies are attenuated more strongly, consistent with spectral bias. At $\delta = 0$, it suppresses the floors of all degrees $\ell \ge 1$.



(a) $\ell = 1$, $\mathcal{D}^1$ (b) $\ell = 2$, $\mathcal{D}^1$ (c) Floor size, general ℓ ($\mathcal{D}^4$)

Figure 16: **Operator-level behavior of the degree blocks and the dropped floor.** All curves use the exact Taylor coefficients of $\exp(\mathbf{A}_0 + \delta\mathbf{C})$. (a, b) Solid: the coupling-driven part of $\psi_{\text{eq},\ell}(\delta)$, with at least ℓ factors of $\mathbf{C}$, normalized at $\delta = 1$. It closely follows δ^ℓ (dotted), with fitted exponents 0.91 and 1.93. Dashed: the full block, including the floor $\mathbf{D}_\ell$. (c) Relative floor size $\|\mathbf{D}_\ell(\delta)\|/\|\psi_{\text{eq},\ell}(\delta)\|$ for $\ell = 1, \ldots, 8$, with $\mathbf{A}_0, \mathbf{C}$ represented in $\mathcal{D}^4$. Values above 1 indicate cancellation between the floor and the coupling-driven part.

remark 3 (The case $\bar{\ell} = 1$)**.** For $\bar{\ell} = 1$, $n = 3$ and the degrees are $\ell \in \{0, 1, 2\}$. The projectors reduce to $\pi_0(\mathbf{M}) = \frac{1}{3}\text{tr}(\mathbf{M})\,\mathbf{I}$, $\pi_1(\mathbf{M}) = \frac{1}{2}(\mathbf{M} - \mathbf{M}^T)$, and $\pi_2(\mathbf{M}) = \frac{1}{2}(\mathbf{M} + \mathbf{M}^T) - \frac{1}{3}\text{tr}(\mathbf{M})\,\mathbf{I}$.

The series equation A.3.8 then has a closed form. Writing $\mathbf{X}_\delta = \hat{\mathbf{v}}$, with $\mathbf{v} = \mathbf{v}(\delta)$ and $\theta = \|\mathbf{v}\|_2$, Rodrigues' formula gives

$$\psi_{\text{eq}}(\delta) = \underbrace{\tfrac{1+2\cos\theta}{3}\,\mathbf{I}}_{\ell=0} + \underbrace{\tfrac{\sin\theta}{\theta}\,\hat{\mathbf{v}}}_{\ell=1} + \underbrace{\tfrac{1-\cos\theta}{\theta^2}\left(\mathbf{v}\mathbf{v}^T - \tfrac{1}{3}\theta^2\mathbf{I}\right)}_{\ell=2}. \tag{A.3.23}$$

When $\mathbf{A}_0$ is small relative to $\delta\mathbf{C}$ and θ is small, $\hat{\mathbf{v}} \approx \delta\hat{\mathbf{c}}$, $\sin\theta/\theta \approx 1$ and $(1-\cos\theta)/\theta^2 \approx \frac{1}{2}$. The $\ell = 1$ and $\ell = 2$ parts then scale as δ and δ^2, matching Step 7. Outside this regime, δ enters through $\sin$ and $\cos$, and Step 8 controls the error. For $\bar{\ell} \geq 2$, no such three-term form exists, and equation A.3.8 is the general description.

Fig. 16 examines Proposition 2 at the operator level. Figs. 16a and 16b show that the coupling-driven part of each block scales as δ^ℓ, with fitted exponents 0.91 and 1.93. The full block, however, also contains the floor $\mathbf{D}_\ell$. Fig. 16c quantifies this floor for general ℓ, with $\mathbf{A}_0$ and $\mathbf{C}$ represented in $\mathcal{D}^4$. The floor is small only for $\ell = 1$. For $\ell \geq 2$, it is comparable to or larger than the full block, and it does not grow monotonically in ℓ. Dropping the floor is therefore not a small perturbation of the operator. We thus treat δ^ℓ as a design choice: it keeps the coupling-driven part and discards the uncoupled content by construction. Its effect on the trained network is validated empirically in **RQ4**.

A.4 Proof of Error Bounds

We prove Theorem 1. Here $\rho(g)$ acts on the layer's *input* space and $\hat{\rho}(g)$ on its *output* space. For hidden-to-hidden layers they coincide; for the first layer they do not. The bound below does not assume $\rho = \hat{\rho}$.

Step 0: Isolate the residual. Let $F \equiv \psi(\cdot;\delta) = F_{\text{eq}} + (1-\delta)\mathbf{W}$ denote the layer, with $F_{\text{eq}} \equiv \bigoplus_\ell \delta^\ell \psi^{(\ell)}$. Since the sum is additive,

$$r_\delta(x,g) = \big[F_{\text{eq}}(\rho(g)x) - \hat{\rho}(g)F_{\text{eq}}(x)\big] + (1-\delta)\big[\mathbf{W}(\rho(g)x) - \hat{\rho}(g)\mathbf{W}(x)\big]. \tag{A.4.1}$$

The first bracket vanishes: F_{eq} scales each irrep block by δ^ℓ, a scalar multiple of the identity, which commutes with $\mathcal{D}^\ell(g)$. Hence

$$\epsilon_\delta(x,g) = \|r_\delta(x,g)\| = (1-\delta)\big\|\mathbf{W}(\rho(g)x) - \hat{\rho}(g)\mathbf{W}(x)\big\|. \tag{A.4.2}$$

Step 1: Insert and subtract $\mathbf{W}(x)$. By the triangle inequality,

$$\begin{aligned}\big\|\mathbf{W}(\rho(g)x) - \hat{\rho}(g)\mathbf{W}(x)\big\| &= \big\|\big[\mathbf{W}(\rho(g)x) - \mathbf{W}(x)\big] + \big[\mathbf{W}(x) - \hat{\rho}(g)\mathbf{W}(x)\big]\big\| \\ &\leq \big\|\mathbf{W}(\rho(g)x) - \mathbf{W}(x)\big\| + \big\|\mathbf{W}(x) - \hat{\rho}(g)\mathbf{W}(x)\big\|. \end{aligned}\tag{A.4.3}$$

Step 2: Bound the first term via Lipschitz continuity. By Assumption 1,

$$\big\|\mathbf{W}(\rho(g)x) - \mathbf{W}(x)\big\| \leq L\,\big\|\rho(g)x - x\big\|. \tag{A.4.4}$$

Since $\rho(g)x - x = (\rho(g) - I)x$, the operator norm on the *input* space gives

$$\big\|\rho(g)x - x\big\| = \big\|(\rho(g) - I)x\big\| \leq \|\rho(g) - I\|\,\|x\|. \tag{A.4.5}$$

Combining equation A.4.4–equation A.4.5, using $\|\rho(g) - I\|$ throughout (no assumption that $\rho = \hat{\rho}$ is needed here):

$$\big\|\mathbf{W}(\rho(g)x) - \mathbf{W}(x)\big\| \leq L\,\|\rho(g) - I\|\,\|x\|. \tag{A.4.6}$$

Step 3: Bound the second term directly. Factoring $(I - \hat{\rho}(g))$ out of the second term of equation A.4.3, using the operator norm on the *output* space:

$$\big\|\mathbf{W}(x) - \hat{\rho}(g)\mathbf{W}(x)\big\| = \big\|(I - \hat{\rho}(g))\mathbf{W}(x)\big\| \leq \|\hat{\rho}(g) - I\|\,\|\mathbf{W}(x)\|. \tag{A.4.7}$$

Step 4: Combine. Substituting equation A.4.6 and equation A.4.7 into equation A.4.3:

$$\big\|\mathbf{W}(\rho(g)x) - \hat{\rho}(g)\mathbf{W}(x)\big\| \le L\,\|\rho(g) - I\|\,\|x\| + \|\hat{\rho}(g) - I\|\,\|\mathbf{W}(x)\|. \tag{A.4.8}$$

Since G is compact, $\rho(g)$ and $\hat{\rho}(g)$ are orthogonal in the chosen (real spherical-harmonic) basis. Hence $\|\rho(g)\| = \|\hat{\rho}(g)\| = 1$ in the operator norm, and by the triangle inequality

$$\|\rho(g) - I\| \le \|\rho(g)\| + \|I\| = 2, \qquad \|\hat{\rho}(g) - I\| \le 2. \tag{A.4.9}$$

This bound is tight: the eigenvalues of $\rho(g) - I$ are $e^{i\theta} - 1$, with $|e^{i\theta} - 1| = 2\,|\sin(\theta/2)| \le 2$, and a rotation by π attains equality. It holds whether or not $\rho = \hat{\rho}$.

Applying the fixed prefactor $(1 - \delta)$ from equation A.4.2 and $\|\hat{\rho}(g) - I\| \le 2$ from Eq. A.4.9:

$$\begin{aligned}\|\epsilon(\delta)\| &\le (1-\delta)\Big[L\,\|\rho(g) - I\|\,\|x\| + \|\hat{\rho}(g) - I\|\,\|\mathbf{W}(x)\|\Big] \\ &\le (1-\delta)\cdot 2\big(L\|x\| + \|\mathbf{W}(x)\|\big) = C(1-\delta),\end{aligned} \tag{A.4.10}$$

where $C := 2\big(L\|x\| + \|\mathbf{W}(x)\|\big)$ is finite and δ-independent, since L is finite (Assumption 1) and $x, \mathbf{W}(x)$ are evaluated at a fixed input. This establishes $\|\epsilon(\delta)\| = \mathcal{O}(1-\delta)$, vanishing linearly as $\delta \to 1$, uniformly across every layer in the network regardless of whether its input and output irrep spaces coincide.

A.5 Construction of the Manifold $\mathcal{M}$

In this section, we provide a precise construction of the manifold $\mathcal{M}$, verify that it admits a C^∞ smooth structure, and show how it determines the fixed operators $\mathbf{A}_0, \mathbf{C}$ used in $\mathbf{A}(\delta)$ (Eq. 8).

A.5.1 Base Space and Coordinates

We begin with the product manifold $\mathbb{T}^3 = S^1_x \times S^1_y \times S^1_z$, where each S^1_k is identified with $\mathbb{R}/\mathbb{Z}$ and parameterized by a coordinate $\theta_k \in [0, 1)$ with periodic identification $\theta_k \sim \theta_k + 1$. Let $\theta = (\theta_x, \theta_y, \theta_z) \in \mathbb{T}^3$. We extend this space by introducing auxiliary Euclidean coordinates $\mathbf{u} = (u_1, u_2) \in \mathbb{R}^2$. The underlying set of $\mathcal{M}$ is

$$\mathcal{M} := \mathbb{T}^3 \times \mathbb{R}^2.$$

$\mathcal{M}$ is fixed once and for all, independent of the coupling parameter δ; only δ is learned or varied.

A.5.2 The Hub Weight α

We fix a hub center $p_{\text{hub}} \in \mathcal{M}$, a hub radius $R_{\text{hub}} > 0$, and a transition width $\Delta R_{\text{hub}} \in (0, R_{\text{hub}})$, and let $d(p) := d(p, p_{\text{hub}})$ be a smooth distance function on $\mathcal{M}$. We define the *shared* partition-of-unity weight $\alpha : \mathcal{M} \to [0, 1]$ as the smooth radial cutoff $\alpha(p) = f_c\big(d(p) \,\big|\, R_{\text{hub}}, \Delta R_{\text{hub}}\big)$

$$f_c(r \mid R, \Delta R) = \begin{cases} 1, & r \le R - \Delta R, \\ \dfrac{1}{2}\Big(\tanh\big(\frac{1}{x+1} + \frac{1}{x-1}\big) + 1\Big),\ x = \dfrac{2(r - R + 0.5\Delta R)}{\Delta R}, & R - \Delta R < r < R, \\ 0, & r \ge R. \end{cases} \tag{A.5.1}$$

This is a C^∞ smooth cutoff, of the same general family as the standard smooth bump functions used to construct partitions of unity on manifolds (see, e.g., Lee (2012), here realized via a $\tanh$-based transition rather than the more common $e^{-1/x}$-based construction. α is a *single* function of position on the whole of $\mathcal{M}$,— this is essential, since only a *shared* weight, applied jointly across directions, can generate a nonzero cross-term when combined with distinct generators (§A.5.5). By construction, $\alpha \equiv 1$ on the hub's interior ($d(p) \le R_{\text{hub}} - \Delta R_{\text{hub}}$, containing p_{hub} itself) and $\alpha \equiv 0$ for $d(p) \ge R_{\text{hub}}$.

A.5.3 Open Cover and Partition of Unity

Define

$$U_{\text{hub}} = \{p \in \mathcal{M} : d(p) < R_{\text{hub}}\}, \qquad U_{\text{torus}} = \{p \in \mathcal{M} : d(p) > R_{\text{hub}} - \Delta R_{\text{hub}}\}.$$

Both are open, and $U_{\text{torus}} \cup U_{\text{hub}} = \mathcal{M}$ (their overlap is exactly the transition shell where $0 < \alpha < 1$). We take α itself as the partition-of-unity weight subordinate to this cover:

$$\lambda_{\text{hub}} := \alpha, \qquad \lambda_{\text{torus}} := 1 - \alpha, \qquad \lambda_{\text{torus}} + \lambda_{\text{hub}} = 1,$$

smooth by construction (§A.5.2), with $\lambda_{\text{hub}} \equiv 0$ on $\mathcal{M} \setminus U_{\text{hub}}$ and $\lambda_{\text{torus}} \equiv 0$ on the hub's interior.

A.5.4 Global Coordinate Map

Let $\chi : \mathbb{T}^3 \to \mathbb{R}^m$ be a smooth embedding (e.g. a Fourier embedding) and $\sigma : \mathbb{T}^3 \times \mathbb{R}^2 \to \mathbb{R}^m$ a smooth map representing the Euclidean patch, both valued in a common ambient space $\mathbb{R}^m$. Define the global coordinate map $\Gamma : \mathcal{M} \to \mathbb{R}^m$ by

$$\Gamma(\theta, \mathbf{u}) = \lambda_{\text{torus}}(p)\, \chi(\theta) + \lambda_{\text{hub}}(p)\, \sigma(\theta, \mathbf{u}), \qquad p = (\theta, \mathbf{u}). \tag{A.5.2}$$

Since $\chi, \sigma, \lambda_{\text{torus}}, \lambda_{\text{hub}}$ are all smooth, Γ is smooth.

Atlas Construction. On U_{torus}, use the standard product chart of $\mathbb{T}^3$; on U_{hub}, use coordinates $(\theta_x, \theta_y, \theta_z, u_1, u_2)$. These charts are compatible, since transitions between them are given by smooth interpolation via $\lambda_{\text{torus}}, \lambda_{\text{hub}}$.

Smooth Structure. Since all local charts are smooth and transition maps are compositions of smooth functions, the atlas defines a C^∞ smooth structure on $\mathcal{M}$.

Dimension. Each point admits local coordinates $(\theta_x, \theta_y, \theta_z, u_1, u_2) \in \mathbb{R}^5$, so $\mathcal{M}$ is a smooth manifold of dimension 5.

Tangent Space Decomposition. At each point $(\theta, \mathbf{u})$, $T_{(\theta,\mathbf{u})}\mathcal{M} \cong \mathbb{R}^5$, decomposing into a 2-dimensional subspace (variations in $\mathbf{u}$) and three 1-dimensional subspaces (variations in $\theta_x, \theta_y, \theta_z$); this decomposition is not canonical but follows from the coordinate structure.

A.5.5 From the Manifold to Operator Space

We now define the fixed operators $\mathbf{W}_i, \mathbf{C}$ used in $\mathbf{A}(\delta)$ (Eq. 8). Fix, independently of $\mathcal{M}$, three standard antisymmetric generators $\mathbf{J}_x, \mathbf{J}_y, \mathbf{J}_z \in \mathfrak{so}(3) \subset \text{End}(\mathcal{D}^1)$. Using the hub weight α (§A.5.2), define, for $p \in \mathcal{M}$,

$$\mathbf{W}_i(p) = \alpha(p)\, \mathbf{J}_i. \tag{A.5.3}$$

Since $\alpha \equiv 1$ on the hub's interior, which contains p_{hub}, we evaluate once, at p_{hub}:

$$\mathbf{W}_i := \mathbf{W}_i(p_{\text{hub}}) = \mathbf{J}_i, \qquad \mathbf{C} := \sum_{i<j} [\mathbf{W}_i(p_{\text{hub}}), \mathbf{W}_j(p_{\text{hub}})] = \sum_{i<j} \varepsilon_{ijk} \mathbf{J}_k. \tag{A.5.4}$$

$\mathbf{W}_i, \mathbf{C}$ are therefore fixed, p-independent and δ-independent matrices. $\mathcal{M}$ and α determine *why* this value of $\mathbf{C}$ is the geometrically motivated one — the bracket's value once fully inside the hub, where the three toroidal directions are jointly weighted rather than independent — but no residual dependence on p enters $\mathbf{A}(\delta)$ itself. Away from the hub ($\alpha \equiv 0$), $\mathbf{W}_i(p) = \mathbf{0}$ and $[\mathbf{W}_i(p), \mathbf{W}_j(p)] = 0$ trivially, matching the intended picture of independent, non-interacting toroidal directions.

$\mathbf{W}$, by contrast, is a learned dense weight matrix, not derived from α or $\mathbf{J}_i$: the Euclidean patch coordinates $\mathbf{u} \in \mathbb{R}^2$ motivate the presence of an unconstrained component at the hub, without literally defining $\mathbf{W}$.

A.5.6 Properties of the Components of $\mathbf{A}(\delta)$

We collect here the properties of $\mathbf{A}_0$, $\mathbf{C}$, and $\mathbf{W}$ used in the main text (Eq. 8).

Toroidal generators. For $p \in \mathcal{M}$, $\mathbf{W}_i(p) = \alpha(p)\,\mathbf{J}_i$ are independent rotation generators scaled by the shared partition-of-unity weight α (§A.5.2). Here $\mathbf{J}_x, \mathbf{J}_y, \mathbf{J}_z$ are the fixed, antisymmetric 3×3 generators of $\mathrm{SO}(3)$ acting on $\mathcal{D}^1$. Each $\mathbf{W}_i(p)$ is bounded, finite-dimensional, and antisymmetric at every point. It vanishes identically away from the hub, where $\alpha \equiv 0$, and is nonzero within it. Away from the hub, the generators therefore commute trivially: $\sum_{i<j}[\mathbf{W}_i(p), \mathbf{W}_j(p)] = 0$, since $\mathbf{W}_i(p) = \mathbf{0}$ there. This matches the picture of independent, non-interacting toroidal directions. Evaluated at p_{hub}, where $\alpha = 1$, we obtain $\mathbf{W}_i = \mathbf{J}_i$ and $\mathbf{A}_0 = \mathbf{J}_x + \mathbf{J}_y + \mathbf{J}_z$.

Hub commutator. Inside the hub, the shared weight produces a nonzero cross-term:

$$\mathbf{C}(p) = \sum_{i<j}[\mathbf{W}_i(p), \mathbf{W}_j(p)] = \alpha(p)^2 \sum_{i<j} \varepsilon_{ijk}\mathbf{J}_k \neq 0 \quad \text{where } \alpha(p) \neq 0. \tag{A.5.5}$$

It is computed via the ordinary matrix commutator on $\mathrm{End}(\mathcal{D}^1)$. The commutator of antisymmetric matrices is antisymmetric, so $\mathbf{C}(p)$ is a bounded, 3×3, antisymmetric $\mathcal{D}^1$-representation matrix. At p_{hub} it reduces to the fixed value $\mathbf{C} = \sum_{i<j} \varepsilon_{ijk}\mathbf{J}_k$ (§A.5.5).

Unconstrained component. $\mathbf{W}$ is a dense, learned linear weight matrix with no equivariance constraint. It is not derived from α or $\mathbf{J}_i$. The Euclidean coordinates $\mathbf{u}$ motivate an unconstrained component at the hub without defining it.

Role of δ. $\mathbf{A}_0$, $\mathbf{C}$, and $\mathbf{W}$ are fixed with respect to δ. δ only gates their relative contribution to $\mathbf{A}(\delta)$ through $\delta\mathbf{C}$ and $(1-\delta)\mathbf{W}$. The Euclidean hub is thus realized as $\delta\mathbf{C} + (1-\delta)\mathbf{W}$.

Boundedness and algebraic setting. All components are bounded, finite-dimensional matrices, so $\mathbf{A}(\delta)$ is finite-dimensional and bounded. Since $\mathbf{W}$ is unconstrained, $\mathbf{A}(\delta)$ is in general not antisymmetric and hence not an element of $\mathfrak{so}(3)$. It is therefore not a Lie-algebra operator. Instead, it lies in the *associative operator algebra*, which is closed under ordinary matrix multiplication. This closure guarantees that every power $\mathbf{A}(\delta)^k$ appearing in the Taylor expansion of $\exp(\mathbf{A}(\delta))$ remains in the same space.

Conclusion. The above construction defines a smooth manifold $\mathcal{M} = \mathbb{T}^3 \times \mathbb{R}^2$, a single shared hub weight α interpolating between toroidal and Euclidean structure, and the fixed operators $\mathbf{W}_i, \mathbf{C} \in \mathrm{End}(\mathcal{D}^1)$ this weight determines, connecting the geometric motivation of $\mathcal{M}$ directly to the operator-space construction of $\mathbf{A}(\delta)$ used in the main text.

A.6 Additional experimental details

All experiments are implemented in PyTorch Paszke et al. (2019) using each model's official code base and default hyperparameters as a starting point. We performed a grid search for both the primal optimizer and, when applicable, the dual optimizer in our constrained formulation. After selecting the best values, we re-ran every configuration with multiple random seeds. Notably, the test results on the QM9 dataset reported in Tab. 1 use different splits for the training, validation, and test datasets. For SchNet, we shuffle the dataset obtained from PyTorch Geometric Fey & Lenssen (2019) and split it into 80% / 10% / 10% train/validation/test ratios. For SEGNN, we use the official splits from Brandstetter et al. (2021). For the experiments on the rotated noisy ModelNet40 Wang et al. (2019) for classification and pose estimation using Lidar point clouds on Oxford RobotCar Maddern et al. (2017) we use the VN-DGCNN Congyue et al. (2021) as our baseline. For the pose estimation task, we used the sequences "2019-01-11-14-02-26-radar-oxford-10k" and "2019-01-14-12-05-52-radar-oxford-10k" for training and "2019-01-17-14-03-00-radar-oxford-10k" for testing.

Pose Estimation Task Loss. Following the homoscedastic-uncertainty formulation of HypLiLoc Wang et al. (2023), the task loss is

$$\mathcal{L}_{\text{task}} = e^{-s_x}\left\|\hat{t} - t\right\|_1 + s_x \;+\; e^{-s_q}\left\|\hat{q} - q\right\|_1 + s_q, \tag{A.6.1}$$

where:

- $\hat{t} \in \mathbb{R}^3$ is the predicted translation, the first three components of the network output;

- $t \in \mathbb{R}^3$ is the ground-truth translation, normalized by the dataset's mean and standard deviation (μ_t, σ_t) computed over the training sequences;
- $\hat{q} \in \mathbb{R}^3$ is the predicted rotation, represented in logarithmic quaternion form (the imaginary part of the unit quaternion, scaled by its rotation angle), the last three components of the network output;
- $q \in \mathbb{R}^3$ is the corresponding ground-truth logarithmic quaternion, obtained from the rotation matrix via $q = \arccos(w)\,\frac{v}{\|v\|}$ for a unit quaternion (w, v), $w \in \mathbb{R}$, $v \in \mathbb{R}^3$;
- $\|\cdot\|_1$ is the elementwise L_1 (mean absolute error) norm;
- $s_x, s_q \in \mathbb{R}$ are learned log-variance (homoscedastic uncertainty) parameters, one per task (translation, rotation), initialized at 0 and optimized jointly with the network weights; the terms e^{-s_x}, e^{-s_q} automatically down-weight a task's loss contribution as its corresponding uncertainty s_x or s_q increases, with s_x, s_q themselves acting as regularizers preventing the uncertainties from growing unbounded.

We set the sampling rate of points for ModelNet40 Wang et al. (2019) as 300 points for all the models, different from 1024 points reported in Deng et al. Congyue et al. (2021), and for Oxford RobotCar Maddern et al. (2017) as 800. We mainly focused on tuning the primal and dual Lr to implement controlled equivariance, and analyse how it gains performance with respect to the baseline without modifying baseline models' components or performing a broader hyperparameter search. Due to these differences, both in input resolution and in the extent of tuning, the reported results are not directly comparable to those in Deng et al Congyue et al. (2021) or in KPConv Thomas et al. (2019).

All of the experiments were performed on internal clusters that contain a mix of Nvidia RTX 2080 Ti, RTX A6000, or GTX 1080 Ti GPUs and Intel Xeon E5-2630, Intel Xeon Silver 4110 CPUs. All experiments consumed a total of approximately 2000 GPU hours, with the longest compute being consumed on the QM9 R. Ramakrishnan & von Lilienfeld (2014) and OxFord RobotCar Maddern et al. (2017) datasets. Runtimes and parameter counts are reported in Table 7 and learning rate grid for both dual and primal with seeds reported in Tab. 8. We also provide a high-level pseudocode of our method in Algorithm 1 and the learning of the coupling parameter in Algorithm 2.

Complexity analysis. We perform time complexity analysis of Algo. 1. Let N be the number of points, k the neighborhood size (k-NN graph), $\ell_{\max}$ the maximum irrep order, $C_{\text{in}}, C_{\text{out}}$ the message channel multiplicities, C_{mid} the equivariant self-interaction channel width, $D_h := \sum_{\ell=0}^{\ell_{\max}} (2\ell{+}1)\, C_\ell$ the total (flattened) per-node/edge irrep feature dimension, and $d_{\mathbf{W}}$ the input/output width of the dense flat-branch weight matrix (L7).

- **Graph construction**(L1): $O(NlogN)$.
- **Edge geometry**(L2): $O(Nk\,\ell_{\max}^2)$.
- **Message construction**(L4): a per-*edge* CG tensor product, $O(Nk\,\ell_{\max}^3\, C_{\text{in}} C_{\text{out}})$.
- **Aggregation**(L5): summing/averaging the D_h-dimensional message per neighbor, $O(Nk\,D_h)$.
- **Irrep scaling** δ^ℓ(L6): a per *node* scalar multiply on the D_h-dimensional feature, $O(ND_h)$.
- **Equivariant branch**(L7): a per-*node* equivariant self-interaction, $O(N\,\ell_{\max}^3\, C_{\text{mid}}^2)$.
- **Flat branch**(L7): a dense linear layer per node, $O(N\,d_{\mathbf{W}}^2)$.

Per-layer total (general):

$$O\Big(Nk\,\ell_{\max}^3\, C_{\text{in}} C_{\text{out}} \;+\; N\,\ell_{\max}^3\, C_{\text{mid}}^2 \;+\; N\,d_{\mathbf{W}}^2\Big). \tag{A.6.2}$$

Comparison. Relative to a strict equivariant baseline (SEGNN), which shares the L4 message and L7 equivariant self-interaction costs, our method's only added terms are the $O(ND_h)$ scaling step (L6) and the $O(Nd_{\mathbf{W}}^2)$ flat branch (L7), neither of which scales with k.

Table 7: Parameter counts and runtimes for equivariant baselines versus ours.

Model	**#Parameters (Train)**	**Train time (s/epoch)**	**Val time (s/epoch)**
SEGNN - N-body	147,424	1.368 ± 0.004	0.505 ± 0.005
Ours - N-body	937,962	1.512 ± 0.004	0.527 ± 0.004
SEGNN - QM9	1,033,525	118.403 ± 0.108	11.152 ± 0.082
Ours - QM9	9,869,668	129.039 ± 0.043	11.402 ± 0.089
VNDGCNN - Oxford RobotCar	2,899,878	1040.51 ± 0.0021	308.04 ± 0.026
Ours - Oxford RobotCar	8,831,845	1240.93 ± 0.0023	308.09 ± 0.032
VNDGCNN - ModelNet40	2,899,830	57.421 ± 0.058	11.15 ± 0.083
Ours - ModelNet40	8,831,612	66.18 ± 0.043	11.87 ± 0.089

Table 8: Learning rates and number of random seeds used for each dataset.

Dataset	**Primal LR grid**	**Dual LR grid**	**Seeds**
ModelNet40	$\{0.10, 0.20\}$	$\{1 \times 10^{-4}, 1 \times 10^{-3}\}$	40
Oxford RobotCar	$\{5 \times 10^{-4}, 1 \times 10^{-3}\ 5 \times 10^{-3}\}$	$\{5 \times 10^{-5}, 2 \times 10^{-4}, 1 \times 10^{-2}\}$	60
N-Body	$\{5 \times 10^{-4}, 8 \times 10^{-4}, 9 \times 10^{-4}, 3 \times 10^{-4}\}$	$\{5 \times 10^{-5}, 1 \times 10^{-3}, 8 \times 10^{-4}\}$	40
QM9	$\{5 \times 10^{-5}, 5 \times 10^{-4}, 5 \times 10^{-3}, 1 \times 10^{-2}\}$	$\{5 \times 10^{-4}, 1 \times 10^{-2}, 1 \times 10^{-3}\}$	80