Continuous Equivariant Learning with Linearly Many Frames

Florian Frick* Nathaniel Kenschaft*
Department of Mathematical Sciences, Carnegie Mellon University
frick@cmu.edu nkenscha@andrew.cmu.edu
*Joint first authors.

Abstract

Coordinate frames let ordinary neural networks make equivariant predictions by transforming the input into a standard position and transforming the output back. Choosing a single frame can make predictions jump as the input varies. Weighted averaging restores continuity, but its cost grows with the number of frames. For n labeled vectors spanning $\mathbb{R}^d$, with $d \geq 2$ fixed, linearly many frames are necessary and sufficient for a continuous $\mathrm{SO}(d)$- or $\mathrm{O}(d)$-equivariant weighted frame. Both bounds are new, and the lower bounds answer a question of Dym et al. (2024). We prove the lower bounds using Dold's generalization of the Borsuk–Ulam theorem and sharpen their constants using characteristic classes of Grassmannians. An explicit construction from Wronskians attains the upper bound with $d(n-d)+1$ frames. For fixed $d \geq 3$, this improves the n^{d-1} growth of Dym et al.'s construction to linear growth, using 61 instead of 1,012 frames for 24 particles in three dimensions. The same construction supports efficient training with one frame per input, followed by continuous prediction through weighted averaging without retraining. On a synthetic 24-particle experiment learning force predictions, this procedure achieves 48.2% lower error on held-out configurations than a frame-free baseline built on pairwise inner products, at comparable training compute.

1 Introduction

Many AI systems in science predict physical quantities, such as the forces on particles, whose behavior under a change of coordinates is known in advance. Rotating a configuration should rotate its predicted forces by the same amount. This requirement, called equivariance, can be built into the layers of a network, but it can also be obtained by wrapping an ordinary network, which we call the backbone: orient the input, make a prediction, and transform the answer back (Puny et al., 2022; Kaba et al., 2023; Ma et al., 2024). The wrapper makes any architecture, including a pretrained one, exactly equivariant without redesigning it.

A choice of orthonormal axes is a frame. If the columns of a matrix A are the input vectors and Q is a frame, the backbone f receives the coordinates $Q^\top A$, and its vector prediction is transformed back as $Qf(Q^\top A)$. When the input rotates by R and the frame rotates with it, from Q to RQ, the backbone sees the same coordinates, since $(RQ)^\top RA = Q^\top A$, while the returned prediction rotates by R. Frame averaging repeats this orient, predict and return step in several frames and averages the results (Puny et al., 2022). This explains equivariance, but not continuity, since nearby inputs should also receive nearby predictions.

For example, under planar rotations one could use the first nonzero input vector as the positive horizontal axis. As that vector shrinks to zero, its direction can vary arbitrarily, even if the other vectors still span the plane, so the frame must switch to the next vector. At the switch, the

prediction changes by the difference between the predictions transformed back from the two frames. It therefore jumps under an arbitrarily small change of input whenever these two predictions differ, and a continuous backbone does not prevent this. Such jumps matter when a model is used along a path or through its derivatives, for example in a simulation whose trajectory crosses a frame switch, or when forces are computed as gradients of a predicted energy.

Weighted frame averaging blends the predictions from several frames with nonnegative weights that sum to one and depend on the input (Dym et al., 2024; Pozdnyakov and Ceriotti, 2023). Such an input-dependent weighted collection of frames is a *weighted frame*. A frame's weight tends to zero as the input approaches a point where that frame is undefined, suppressing its contribution even when the frame itself has no limit. In the planar example, weights proportional to the squared lengths of the vectors do this, as long as some vector remains nonzero. If the weighted frame transforms with the input and varies continuously, including at inputs where an individual frame is undefined, then averaging preserves both equivariance and continuity for every continuous backbone.

Averaging has a cost, since evaluating each frame with positive weight separately requires another backbone call. The worst-case support of a weighted frame is the largest number of distinct frames with positive weight at any one input, and we ask how small it can be. Support is the part of the cost that continuity constrains. It counts frames, not running time, which also depends on batching and on the cost of scoring the frames (Section 6).

Our main result answers this question for n labeled vectors spanning $\mathbb{R}^d$, with $d \geq 2$ fixed, where labeled means that each vector keeps a fixed identity, such as the index of a particle. For rotations, $\mathrm{SO}(d)$, and for rotations together with reflections, $\mathrm{O}(d)$, linearly many frames are necessary and sufficient (Theorem 1 in Section 2). Both bounds are new. The lower bounds answer a question of Dym et al. (2024, Section 8), who asked for stronger lower bounds on the number of frames that continuous weighted frames for rotations need, and they also apply to their weighted frames, which are defined on all inputs (Corollary 5). The upper bound comes from an explicit construction whose size is linear in n. For planar rotations the answer is exact: $n - 1$ frames are necessary, and the squared-length weights of the example above, on the first $n - 1$ vectors, attain this (Corollary 4).

The lower bound tells a practitioner before training how many frames continuity requires. For 24 particles, represented after centering by $n = 23$ vectors in $\mathbb{R}^3$, every continuous equivariant weighted frame needs at least 11 frames at some input, so evaluating each frame separately requires at least

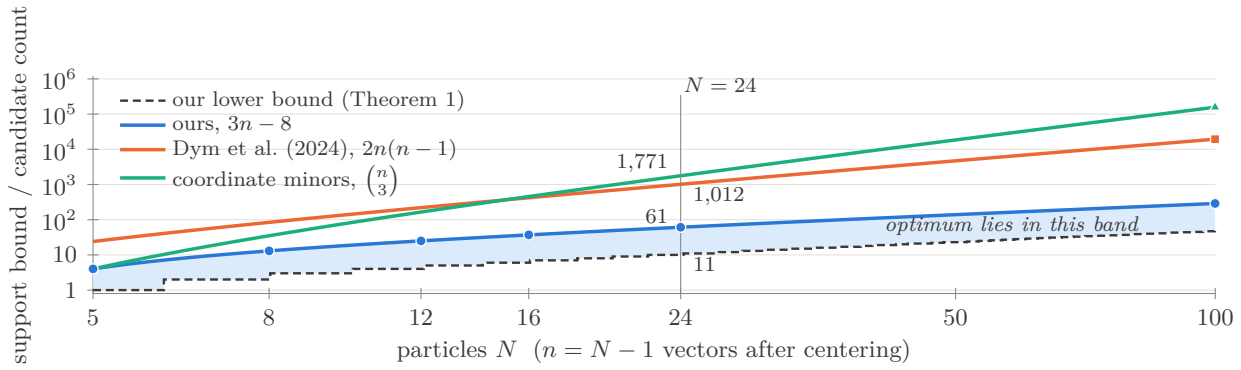


Figure 1: Worst-case support bounds and candidate-bank sizes for O(3)-equivariant weighted frames for N particles, represented after centering by $n = N - 1$ vectors (logarithmic axes). The minimum worst-case support lies in the shaded band. The other curves count frames in the construction of Dym et al. (2024) and frames from all coordinate minors, one per triple of input vectors. Dots mark the experimental sizes. Evaluating each frame with nonzero weight separately uses one backbone call.

11 backbone calls there. Principal-axis frames (Puny et al., 2022), which take the eigenvectors of $AA^\top$ as axes where its eigenvalues are distinct and average over their signs, use at most eight frames for O(3) and four for SO(3). Since $11 > 8$, they cannot be extended to a continuous equivariant weighted frame on all such inputs without using more frames. Our construction uses 61 frames at 24 particles, compared with 1,012 for the O(3) construction of Dym et al. (2024) on the same 23 vectors (Figure 1). For fixed $d \geq 3$ its size grows linearly in n, whereas theirs grows as n^{d-1}.

The construction uses a fixed bank of linear *probes*, matrices P that mix the input vectors into d candidate axes AP. At every input whose vectors span $\mathbb{R}^d$, at least one probe of the bank gives independent axes and hence a frame. The bank adapts Wronskian rank condensers, fixed linear maps from algebraic complexity theory at least one of which preserves the rank of any given input (Forbes and Guruswami, 2015; Forbes et al., 2014), and has $d(n-d)+1$ probes. It supports two evaluation rules. Hard selection evaluates the backbone once, in the frame with the largest determinant score $|\det(AP)|$. Averaging weights each frame by an even power $2k$ of that determinant, for any integer $k \geq 1$, so a frame fades out as its axes become dependent. The averaged prediction then has $2k-1$ continuous derivatives whenever the backbone has them (C^{2k-1} regularity, Theorem 2), and this is sharp for these weights.

The lower bounds use topology to show that a continuously varying equivariant weighted collection must sometimes assign positive weight to many frames. A first linear bound comes from Dold's generalization of the Borsuk–Ulam theorem, a classical result of equivariant topology (Dold, 1983; Appendix A.3). Sharper bounds come from characteristic classes, topological invariants that detect when a family of subspaces admits no continuous choice of basis, combined with a theorem of Daundkar and García-Calcines (2026) that turns nonzero products of such classes into lower bounds, and, for $d = 2, 3$, from a self-contained argument (Theorem 11). The upper bound turns the coverage property of the bank into continuous equivariant weighted frames, and a second, nonconstructive argument gives the sharper bound $\lfloor d(n-d)/2 \rfloor + 1$ for $\mathrm{SO}(d)$ (Schwarz, 1966; Appendix A.2).

Full rank, meaning that the vectors span $\mathbb{R}^d$, is the generic case: rank-deficient inputs form an algebraic set of Lebesgue measure zero, and our construction covers the entire full-rank domain, which is open and dense. The lower bounds already hold on perfectly conditioned inputs, those with $AA^\top = I_d$. Our construction does not impose symmetry under relabeling the vectors (Section 8).

Empirical evidence. We test the construction on synthetic force prediction for 24 labeled particles, against a standard frame-free baseline, a Gram-matrix model that predicts from pairwise inner products of positions. Training with hard selection, one backbone call per input, was the most compute-efficient training rule we tested (Section 7.3). Averaging the trained networks without retraining, with weights that smoothly drop low-scoring frames, gives a continuous predictor that is more accurate than hard selection and has 48.2% lower force error than the Gram baseline on held-out configurations, with slightly less training time than Gram and single-input latency similar to hard selection (Section 7.4). In a separate five-particle experiment with four frames, moving only the frame switch of a fixed trained network shows that the prediction jumps there by the difference between the two frames' predictions and that averaging removes the jump (Section 7.1).

2 Setting and main result

Let $\mathcal{M}^{\mathrm{fr}}_{d,n} = \{A \in \mathbb{R}^{d\times n} : \operatorname{rank} A = d\}$, the full-rank inputs. Its columns are n labeled vectors spanning $\mathbb{R}^d$. An N-particle configuration that spans $\mathbb{R}^d$ after centering is represented, up to translation, by $n = N-1$ such vectors (Appendix B.2). Let G be $\mathrm{SO}(d)$ or $\mathrm{O}(d)$. A weighted frame assigns to each input A finitely many distinct frames with positive weights summing to one,

that is, a finitely supported probability distribution μ_A on G whose support is this set of frames. Equivariance means that replacing A by RA replaces each frame Q by RQ, with the same weight. Continuity means that the μ_A-average of every continuous function on G varies continuously with A, so a frame may disappear where its weight reaches zero. Write $\mathrm{wfr}_G(\mathcal{M}^{\mathrm{fr}}_{d,n})$ for the least m such that some continuous equivariant weighted frame on $\mathcal{M}^{\mathrm{fr}}_{d,n}$ has support size at most m at every input (6).

Theorem 1 (Linearly many frames are necessary and sufficient)**.** *Let $n > d \geq 2$ and $G \in \{\mathrm{O}(d), \mathrm{SO}(d)\}$. Then*

$$1 + \left\lfloor \frac{n-d}{2} \right\rfloor \;\leq\; \mathrm{wfr}_G(\mathcal{M}^{\mathrm{fr}}_{d,n}) \;\leq\; d(n-d)+1. \tag{1}$$

An explicit family of $d(n-d)+1$ fixed linear probes gives the upper bound. For $d = 3$, $\lfloor (n-1)/2 \rfloor \leq \mathrm{wfr}_G(\mathcal{M}^{\mathrm{fr}}_{3,n}) \leq 3n-8$. For even d the lower bound improves to $n-d+1$, and $\mathrm{wfr}_{\mathrm{SO}(2)}(\mathcal{M}^{\mathrm{fr}}_{2,n}) = n-1$. Thus $\mathrm{wfr}_G(\mathcal{M}^{\mathrm{fr}}_{d,n}) = \Theta(n)$ for fixed d.

The lower bounds hold even for weighted frames defined only on inputs with $AA^\top = I_d$ (Appendix A.4). Section 5 proves the lower bounds and the planar equality, and Section 6 constructs the probes for the upper bound.

3 Related work

Frames and canonicalization. Frame averaging (Puny et al., 2022) and canonicalization, which selects one frame per input (Kaba et al., 2023; Ma et al., 2024), evaluate a generic backbone on transformed copies of the input, so pretrained models can be wrapped without redesigning their internals (Mondal et al., 2023). They have been applied to materials (Duval et al., 2023), with explicit frames for point clouds (Baker et al., 2024) and frames of minimal size (Lin et al., 2024). Whether a single continuous frame exists depends on the input domain: for SO(3), Dym et al. (2024) rule out continuous canonicalization on all 3×4 inputs, while Section 5 constructs a global frame on the full-rank subset. Sverdlov et al. (2026) relate the choice of canonical representative to generalization bounds through covering numbers of the canonical image, a question complementary to ours. Probabilistic symmetrization (Kim et al., 2023) and stochastic frame averaging (Duval et al., 2023) sample transformations or frames, so they are exactly equivariant in expectation but a finite sample is equivariant only in distribution, whereas we evaluate a finite weighted sum exactly.

Continuous weighted frames. Dym et al. (2024) introduced continuous weighted frames, which they call robust frames, for permutations, $\mathrm{SO}(d)$ and $\mathrm{O}(d)$. For $\mathrm{SO}(d)$ and $\mathrm{O}(d)$, their conditions average over the orthogonal transformations fixing an input, and only the identity fixes a full-rank input, so on this domain they become the equivariance and continuity of Section 2 (Appendix A.1). Their quantitative support lower bound concerns permutations, and for rotations they asked for stronger lower bounds (their Section 8), which Theorem 1 provides. Their rotation constructions also handle permutation symmetry and degenerate inputs, and use $n(n-1)\cdots(n-d+2)$ frames for $\mathrm{SO}(d)$ and twice as many for $\mathrm{O}(d)$, against our $d(n-d)+1$ on full-rank labeled inputs, which is linear rather than of order n^{d-1} for fixed $d \geq 3$. Both our bounds are new. In $\mathbb{R}^3$, the smooth ensembles of Pozdnyakov and Ceriotti (2023) average models over frames from neighbor pairs, with adaptive cutoffs and an equivariant fallback near collinearity, whereas our fixed bank covers all full-rank inputs.

Regularity of symmetrized models. Zhang et al. (2020) noted discontinuities of canonicalization for sets of vectors. Tahmasebi and Jegelka (2025) characterized when canonicalized models are continuous or differentiable, through compatibility conditions on the backbone across the identifications the canonicalization makes. Our weights need no such condition: determinant powers $2k$ give C^{2k-1} predictions for every C^{2k-1} backbone, and this is sharp for these weights (Theorem 2).

Equivariant architectures and invariant features. Equivariant networks (Cohen and Welling, 2016; Thomas et al., 2018; Satorras et al., 2021; Batzner et al., 2022; Batatia et al., 2022; Geiger and Smidt, 2022) and invariant-feature models built from inner products (Villar et al., 2021) select no frames. Our experiments compare with both kinds.

Topological tools for frame bounds. The Schwarz genus of a map is the fewest sets in an open cover of the base, each carrying a continuous local section, here a local frame (Schwarz, 1966). Reduced analog and distributional sectional categories instead place a continuously varying, finitely supported probability measure in each fiber, and are one less than the least uniform support bound (Knudsen and Weinberger, 2024; Jauhari, 2025). For the compact Stiefel bundles used here the two agree (Knudsen and Weinberger, 2026), so wfr_G is one more than either (Theorem 3). Support can be smaller than the chart count: two weighted frames against four charts for $\mathrm{O}(3)$ on full-rank 3×4 inputs (Section 5). Applied to the Stiefel bundles for all d and n, the rational cup-length bound of Daundkar and García-Calcines (2026) gives the lower bound in Theorem 1, and a characteristic-class argument gives a self-contained proof for $d = 2, 3$.

4 Prediction by selection and averaging

An equivariant weighted frame turns any backbone into an equivariant predictor: evaluate the backbone in each frame's coordinates, transform each output back, and average. Let $f\colon \mathcal{M}^{\mathrm{fr}}_{d,n} \to V$ be a backbone with values in a finite-dimensional space V, and let $\rho\colon G \to \mathrm{O}(V)$ be an orthogonal representation, which specifies how the output transforms under each $R \in G$. For forces, $\rho(R)$ applies R to every force vector (Appendix B.2). For frames $g_j(A) \in G$ and weights $w_j(A)$ of μ_A, summing over frames of positive weight, the prediction is

$$F(A) = \sum_j w_j(A)\, Y_j(A), \qquad Y_j(A) = \rho(g_j(A))\, f\big(g_j(A)^{-1}A\big). \tag{2}$$

We call Y_j the *branch* of frame j, and the distinct frames in the support of μ_A its *atoms*. Replacing A by RA replaces each frame g by Rg with the same weight, possibly in another order. Since $\rho(Rg)f\big((Rg)^{-1}RA\big) = \rho(R)\rho(g)f(g^{-1}A)$, reindexing gives $F(RA) = \rho(R)F(A)$, even when the group permutes the frames. If f is already equivariant, every Y_j equals f, and averaging returns f.

Hard selection and switching. Hard selection evaluates only the frame of largest weight, with one backbone call, whereas evaluating the positive-weight branches of the m probes below separately uses at most m calls. These counts exclude computing scores and frames and are not runtimes. For these fixed indexed probes, invariant scores and tie-breaking by index make hard selection equivariant. If along a path $A(t)$ the largest weight passes from frame i to j at $t = 0$, ties pick a maximizer, and both branches extend continuously to $A(0)$, then

$$\lim_{t\uparrow 0} F_{\mathrm{hard}}(A(t)) - \lim_{t\downarrow 0} F_{\mathrm{hard}}(A(t)) = Y_i(A(0)) - Y_j(A(0)). \tag{3}$$

So a switch produces a jump exactly when the two branches, in common output coordinates, disagree there (Appendix A.7). Ordinary force-error training does not require agreement, and a finite penalty on disagreement does not guarantee it (Section 7.1), whereas averaging requires no such agreement.

Frames from probes. A probe is a fixed matrix $P \in \mathbb{R}^{n\times d}$ with orthonormal columns. The proposed axes AP can be dependent even when A has full rank. On its chart U_P, the set where $\det(AP) \neq 0$, the polar decomposition, an orthogonal times a positive definite factor,

$$AP = Q_P(A)\, S_P(A), \qquad Q_P(A) \in \mathrm{O}(d), \quad S_P(A) \succ 0, \tag{4}$$

gives a frame with $Q_P(RA) = RQ_P(A)$. For $\mathrm{O}(d)$ we use $g_P = Q_P$. For $\mathrm{SO}(d)$ we fix a reflection J and use $g_P(A) = Q_P(A)$ where $\det(AP) > 0$ and $g_P(A) = Q_P(A)J$ where $\det(AP) < 0$, so that the frame has determinant one. Given probes $P_1, \dots, P_m$ whose charts cover $\mathcal{M}^{\mathrm{fr}}_{d,n}$ and an integer $k \geq 1$, we use the determinant weights

$$w_j(A) = \frac{\det(AP_j)^{2k}}{\sum_{\ell=1}^{m} \det(AP_\ell)^{2k}}. \tag{5}$$

They are invariant, vanish exactly outside the chart of P_j, and have a positive denominator because the charts cover. Full averaging in Sections 7.2–7.4 uses $k = 1$, and Section 7.4 also uses compact weights that give low-scoring frames zero weight, with C^2 predictions for a C^2 backbone and support at most m (Remark 20). Section 7.1 uses the variants in Table 2.

Theorem 2 (Continuous and smooth averaging)**.** *Let $P_1, \dots, P_m$ be probes whose charts cover $\mathcal{M}^{\mathrm{fr}}_{d,n}$, let $k \geq 1$, and define F by (2) and (5), omitting frames of weight zero. Then $A \mapsto \mu_A$ is a continuous equivariant weighted frame with at most m atoms, and for every continuous backbone f the prediction F is continuous and G-equivariant. If f is C^{2k-1}, then so is F. This is sharp, since in general F is not C^{2k}, even for a C^∞ backbone.*

Proof outline. Equivariance holds term by term. Near a full-rank input, the inputs $g_j(A)^{-1}A$ seen by the backbone remain in a compact set, so a vanishing weight suppresses a frame even when the frame has no limit. Fix any integer $k \geq 1$ independently of the input. The quantity $s = \sigma_{\min}(AP_j)$ enters only the derivative estimates. Derivatives of the polar factor of order r are $O(s^{-r})$, whereas those of the weighted numerator of order $r < 2k$ are $O(s^{2k-r})$ and so extend by zero across a singular probe. Appendix A.6 proves these estimates and sharpness for $\mathrm{O}(d)$ and for $\mathrm{SO}(d)$ with any fixed J. □

With ρ trivial, a C^{2k-1} scalar backbone gives an invariant C^{2k-1} energy E and a C^{2k-2} equivariant force $-\nabla E$, differentiating through frames and weights. Our experiments predict forces directly.

5 How many frames does continuity require?

This section proves the lower bounds of Theorem 1: every continuous equivariant weighted frame needs many frames of positive weight at some input. Whitening, the map $A \mapsto (AA^\top)^{-1/2}A$, sends full-rank inputs equivariantly to inputs U with orthonormal rows. Two such inputs differ by an element of G exactly when their row spaces agree, for $G = \mathrm{SO}(d)$ together with the orientation given by the ordered rows. On a G-invariant open set of such inputs with a continuous equivariant frame g, the rows of $g(U)^\top U$ form an orthonormal basis of the row space that depends only on the

row space. So the tautological bundle of the Grassmannian, which assigns to each row space that subspace (oriented for $\mathrm{SO}(d)$), has a continuous basis over these row spaces, and its characteristic classes, invariants that obstruct such a basis, vanish there. If a product of c such classes is nonzero, c such sets cannot cover all inputs, and at least $c+1$ charts are needed. Section 5.2 extends this to weighted frames.

5.1 Charts, frames, and sectional category

Whitening is an equivariant retraction of $\mathcal{M}_{d,n}^{\mathrm{fr}}$ onto the Stiefel manifold $\mathrm{St}_{d,n}^{\mathrm{row}} = \{U \in \mathbb{R}^{d\times n} : UU^\top = I_d\}$, and the straight-line homotopy from A to its whitening makes it a deformation retraction. Restriction and composition with whitening show that both spaces need the same number of frames (Appendix A.2). The quotients of $\mathrm{St}_{d,n}^{\mathrm{row}}$ by $\mathrm{O}(d)$ and $\mathrm{SO}(d)$ are the Grassmannian $\mathrm{Gr}_d(\mathbb{R}^n)$ of row spaces and the oriented Grassmannian $\mathrm{Gr}_d^+(\mathbb{R}^n)$. For a free action of a compact Lie group G on a space X, the definition of Section 2 becomes

$$\mathrm{wfr}_G(X) = \min\{k : \exists\, \mu\colon X \to \mathcal{P}_{\le k}(G) \text{ continuous with } \mu_{gx} = g_*\mu_x\}, \tag{6}$$

where $\mathcal{P}_{\le k}(G)$ is the space of finitely supported probability measures on G with at most k atoms, in the weak topology, in which averages of continuous functions vary continuously.

An atlas of local frames, that is, invariant open sets U_i covering X with continuous $g_i\colon U_i \to G$ satisfying $g_i(gx) = gg_i(x)$, blends into a weighted frame through a partition of unity. Hence $\mathrm{wfr}_G(X)$ is at most the fewest charts, the Schwarz genus of $X \to X/G$ (Schwarz, 1966). For $\mathrm{O}(d)$ on $\mathcal{M}_{d,n}^{\mathrm{fr}}$ with $n > d$ the genus lies between $\max\{d, n-d\}+1$ and $d(n-d)+1$ (Theorem 8).

Theorem 3 (Topological characterization)**.** *If a compact Lie group G acts freely on a compact metrizable space X, then*

$$\mathrm{wfr}_G(X) = \mathrm{asecat}(X \to X/G) + 1 \le \mathrm{genus}(X \to X/G),$$

where asecat*, the reduced analog sectional category, is one less than the fewest atoms of a continuous family of measures on the fibers (Knudsen and Weinberger, 2024).*

A fiber of $X \to X/G$ is one orbit, and a weighted frame at x gives a measure on it by sending each frame g in its support to $g^{-1}x$ with the same weight. Replacing x by hx replaces g by hg, and $(hg)^{-1}hx = g^{-1}x$, so this measure does not depend on the chosen point of the orbit. The correspondence preserves the number of atoms and continuity (Appendix A.1). The distributional sectional category (Dranishnikov and Jauhari, 2024) topologizes these measures differently, but on these compact metrizable spaces the two topologies agree for each fixed support bound, so the categories coincide (Knudsen and Weinberger, 2026, Proposition 5.2 and Lemma 5.3).

Frames can be fewer than charts. The quotient of $\mathrm{St}_{3,4}^{\mathrm{row}}$ by $\mathrm{O}(3)$ is $\mathrm{Gr}_3(\mathbb{R}^4) \cong \mathbb{R}\mathrm{P}^3$, and whitening transfers both counts to full-rank 3×4 inputs, where

$$\mathrm{genus} = 4, \qquad \mathrm{wfr}_{\mathrm{O}(3)} = 2. \tag{7}$$

The row space of $U \in \mathrm{St}_{3,4}^{\mathrm{row}}$ has a unit normal $q(U)$, and quaternion multiplication turns it into a second orthonormal basis of the row space (Appendix A.8). The change of basis $c(U) \in \mathrm{SO}(3)$ satisfies $c(RU) = \det(R)\, R\, c(U)$. So c is a global frame for $\mathrm{SO}(3)$, and $\mu_U = \frac{1}{2}\delta_{c(U)} + \frac{1}{2}\delta_{-c(U)}$ is a continuous $\mathrm{O}(3)$-equivariant weighted frame. Reflections exchange its two atoms. Four coordinate-minor charts suffice, and four are necessary, because the generator of $H^1(\mathbb{R}\mathrm{P}^3;\mathbb{F}_2)$ vanishes on the Stiefel manifold and has nonzero cube (Appendix A.8). A lower bound on the number of charts therefore does not bound weighted support, so Section 5.2 uses rational classes and measures.

5.2 Lower bounds from equivariant cohomology

Let $n > d \geq 2$, let $q\colon \mathrm{St}^{\mathrm{row}}_{d,n} \to \mathrm{Gr}^{+}_{d}(\mathbb{R}^n)$ be the $\mathrm{SO}(d)$-bundle, and let γ_d be its oriented tautological bundle. The rows of U trivialize $q^*\gamma_d$, so its positive-degree rational classes lie in $\ker q^*$.

For a graded ideal I, its rational cup length is the largest number of positive-degree elements of I with nonzero product. The rational kernel cup-length theorem of Daundkar and García-Calcines (2026, Theorem 4.17), stated as Lemma 13, together with Theorem 3, gives

$$\mathrm{wfr}_{\mathrm{SO}(d)}(\mathcal{M}^{\mathrm{fr}}_{d,n}) \geq 1 + \mathrm{cl}_{\mathbb{Q}}(\ker q^*). \tag{8}$$

Appendix A.4 checks the lemma's hypotheses for q, including that q is a fibration of path-connected spaces of locally finite CW type with metric total space. Let $c = n - d$. The first Pontryagin class and, for even d, the Euler class are standard characteristic classes with $p_1(\gamma_d)^{\lfloor c/2 \rfloor} \neq 0$ and $e(\gamma_d)^c \neq 0$ (Lemma 12). Equation (8) therefore yields $1 + \lfloor c/2 \rfloor$ in general and $c + 1$ for even d.

An $\mathrm{SO}(d)$-equivariant pushforward transfers these bounds to $\mathrm{O}(d)$, and Theorem 11 gives an independent proof for $d = 2, 3$ from characteristic-class degree (Appendix A.4). For $N \geq 5$ labeled particles in $\mathbb{R}^3$, every continuous $\mathrm{O}(3)$- or $\mathrm{SO}(3)$-equivariant weighted frame on all full-rank centered configurations has at least $1 + \lfloor (N-4)/2 \rfloor$ frames of positive weight at some input.

Corollary 4 (Planar rotations)**.** *For every $n \geq 2$,*

$$\mathrm{wfr}_{\mathrm{SO}(2)}(\mathcal{M}^{\mathrm{fr}}_{2,n}) = n - 1.$$

Proof. Identify the columns with $z_1, \ldots, z_n \in \mathbb{C}$ and rotations with unit complex numbers. Full row rank implies $S(A) = \sum_{j=1}^{n-1} |z_j|^2 > 0$, since otherwise only the last column could be nonzero. Set

$$\mu_A = \sum_{\substack{1 \leq j < n \\ z_j \neq 0}} \frac{|z_j|^2}{S(A)}\, \delta_{z_j/|z_j|}. \tag{9}$$

Rotating A rotates every atom and preserves every weight. When a selected column tends to zero, its weight tends to zero, so the measure is weakly continuous. Its support has size at most $n - 1$, matching the lower bound. □

On all nonzero $2 \times n$ matrices the minimum support is n. Restriction to the Hopf bundle gives the lower bound, and using all n columns in (9) gives the upper bound (Appendix A.4). This explains the domain difference from the n-frame construction of Dym et al. (2024).

Corollary 5 (Robust frames on all inputs)**.** *Let $n > d \geq 2$. Every robust frame of Dym et al. (2024) for $\mathrm{O}(d)$ or $\mathrm{SO}(d)$ on $\mathbb{R}^{d \times n}$, finitely supported at every full-rank input, attains support at least $1 + \lfloor (n-d)/2 \rfloor$ there. For even d this lower bound improves to $n - d + 1$.*

Proof. Only the identity fixes a full-rank input, so there the conditions of Dym et al. (2024) reduce to the equivariance and continuity of Section 2 (see Section 3). If every full-rank input had smaller support, the restriction would contradict Theorem 1. No global support bound is needed in advance. □

6 An explicit linear atlas

Theorem 2 turns any bank of probes whose charts cover every full-rank input, so that each such A has $\det(AP_j) \neq 0$ for some j, into a continuous equivariant weighted frame with support at

most the bank size. Wronskian rank condensers from algebraic complexity give such a bank of $d(n-d)+1$ probes: for every d-dimensional subspace, at least one of these fixed maps to $\mathbb{R}^d$ preserves its dimension (Forbes et al., 2014; Forbes and Guruswami, 2015). We prove this directly. Choose polynomials $\psi_0, \dots, \psi_{n-1}$ with $\deg \psi_r = r$, put $b(t) = (\psi_0(t), \dots, \psi_{n-1}(t))^\top$, and let

$$\widetilde{B}(t) = \big[b(t), b'(t), \dots, b^{(d-1)}(t)\big] \in \mathbb{R}^{n\times d}. \tag{10}$$

Read each row a_i of A as the polynomial $f_i(t) = a_i b(t)$. Entry (i,r) of $A\widetilde{B}(t)$ is $f_i^{(r-1)}(t)$, so $W_A(t) = \det(A\widetilde{B}(t))$ is their Wronskian, and the probe at t fails exactly where $W_A(t) = 0$.

Theorem 6 (Universal coverage)**.** *Let $n > d \geq 1$ and $D = d(n-d)$. For distinct nodes $t_1, \dots, t_{D+1} \in \mathbb{R}$, let P_j be obtained by orthonormalizing the columns of $\widetilde{B}(t_j)$. Then*

$$\max_{1\leq j\leq D+1} |\det(AP_j)| > 0 \qquad \textit{for all } A \in \mathcal{M}^{\mathrm{fr}}_{d,n}. \tag{11}$$

Proof. Full row rank makes $f_1, \dots, f_d$ linearly independent, so their Wronskian is a nonzero polynomial. After an invertible row operation, the f_i have distinct degrees $e_1 < \cdots < e_d \leq n-1$. The leading coefficient of the Wronskian is then a nonzero multiple of $\prod_{i<j}(e_j - e_i)$, and its degree is $\sum_i e_i - d(d-1)/2 \leq d(n-d)$ (Lemma 14). A nonzero polynomial of degree at most D cannot vanish at $D+1$ points, and orthonormalization multiplies $\det(A\widetilde{B}(t_j))$ by a nonzero constant. □

With Theorem 2, this proves the upper bound $\mathrm{wfr}_G(\mathcal{M}^{\mathrm{fr}}_{d,n}) \leq d(n-d)+1$ of Theorem 1. For instance, take $d = 2$, $n = 3$, $b(t) = (1, t, t^2)^\top$ and nodes $t = -1, 0, 1$, so $D = 2$. Along the path of full-rank matrices $A(s)$ with rows $(1,0,0)$ and $(0,s,1)$, encoding $f_1 = 1$ and $f_2 = st + t^2$, the Wronskian is $W(t) = s + 2t$. At the nodes it equals $s-2$, s and $s+2$, so at $s = 0$ the probe at $t = 0$ fails. As $s \to 0$ its weight, of order s^{2k}, fades to zero, while the probes at $t = \pm 1$ keep weights near $1/2$ and carry the average. Appendix A.5 interprets probe failure through the row space and relates the bank to rank condensers.

Comparison. For $\mathrm{O}(d)$ and $n > d$ every atlas needs at least $\max\{d, n-d\}+1$ charts, so our $d(n-d)+1$ charts are within a factor d of the fewest. Dym et al. (2024) cover all inputs with $n(n-1)\cdots(n-d+2)$ frames for $\mathrm{SO}(d)$, twice as many for $\mathrm{O}(d)$. For fixed $d \geq 3$ our count on full-rank inputs is linear in n rather than of order n^{d-1}. At $n = 23$ ($N = 24$ particles, as in Section 7) this is 61 probes against 506 frames for $\mathrm{SO}(3)$ and 1,012 for $\mathrm{O}(3)$. With identical untrained backbones on one A40, full averaging over our probes had 4.1 and 10.3 times lower median latency than our implementation of their $\mathrm{O}(3)$ frames at batch sizes 1 and 32 (Appendix B.12). For planar rotations, Corollary 4 gives the optimal $n-1$, below our bank's $2n-3$.

Cost. Scoring all m probes and forming frames where defined takes $O(mnd^2 + md^3)$ operations for $m = d(n-d)+1$, quadratic in n for fixed d, excluding the backbone. At almost every full-rank input all m weights are positive, so averaging evaluates the m branches separately, hard selection one (Section 4), and compact blending drops low-weight terms (Remark 20). We use normalized Legendre polynomials and Chebyshev nodes (Appendix A.5).

7 Experiments

On synthetic force prediction, we test whether frame switches cause jumps and whether averaging removes them (Section 7.1), compare the bank with a frame-free baseline at equal training updates

and at comparable training time (Sections 7.2–7.3), and make a network trained with hard selection continuous after training (Section 7.4). An input is N labeled particles in $\mathbb{R}^3$, and the target is each particle's force under a fixed rotation-invariant potential. The $N = 5$ studies of Section 7.1 use a separate trajectory potential (Appendices B.4 and B.5); the others use that of Appendix B.1, with ratio of smallest to largest singular value of the centered positions in $[0.15, 1]$. Error is NRMSE, the RMS force error divided by the RMS target force, on held-out configurations. Our methods evaluate a Transformer over particles in frames from a bank. We compare with a Gram-matrix model, a frame-free baseline that predicts from pairwise inner products of positions, and at $N = 5$ also with an e3nn equivariant network and the 24-frame construction of Dym et al. (2024). Seeds and test inputs are paired across methods (Appendix B).

7.1 Frame switches cause jumps

With $N = 5$ particles we use four simple frames, each built from three of the four difference vectors, and train networks with averaging, with a single frame, and with a penalty on disagreement between frames. Keeping each network fixed, we move only its single-frame switch. As two inputs straddling the switch approach each other, the residual change tends to the disagreement between the two frames' predictions. Without a switch, or with averaging, the change vanishes (Figure 2). Seven constant score offsets move the single-frame switch to seven locations on each of 504 paths. All seven variants are evaluated on the same input pair, and exactly one switches between them. At separation 0.0006, the switching variant's residual change exceeds the mean residual change of the six others by 0.033 of the RMS target force for networks trained with a single frame, 0.016 with the penalty, and 0.165 with averaging, which does not require its branches to agree. For averaging on the same inputs, the same comparison gives differences of magnitude below 4×10^{-7} (Appendix B.5). The two-frame construction (7) and that of Dym et al. (2024) show no response pattern following the moved switch.

Jumps remain at the final checkpoints of the $N = 24$ hard-selection networks of Section 7.3. In a post hoc check at 80 selected frame switches on paths between random configurations, shared across five seeds, they jump by a median 0.11% of the RMS target force (maximum 0.31%), about half their median error at the same inputs and about seven times less than at the early 37,808-update checkpoint (Appendix B.9).

Simulations through frame switches. At the qualifying first frame switches of 302 of 1,200 held-out five-particle trajectories, averaging over the four frames rather than using a single frame lowers the force NRMSE by 39.3% (paired 95% interval 34.4–43.9%), with eight training seeds (Appendix B.4). In learned-force simulations through these switches, it lowers the mean across seeds of the median final centered-position RMSD by 15.0% (10.2–19.6%). The gain is concentrated near frame switches. Gram, e3nn and Dym et al. (2024) reach lower overall NRMSE (0.0015, 0.0039 and 0.0060, three seeds each) than averaging (0.0068, eight seeds, Table 6). These separately trained models do not isolate continuity's contribution. This study uses four simple frames, not our bank.

7.2 Comparison with the Gram baseline at equal training updates

At $N \in \{5, 8, 12, 16, 24\}$ we train each method for 40,000 updates with five seeds and test on 20,000 held-out configurations. Our network is a 0.53M-parameter Transformer over particles, evaluated in the $3N - 11$ frames of the Wronskian bank of Section 6, with averaging or with hard selection. The 0.73M-parameter Gram baseline predicts rotation-invariant coefficients from the inner products of positions and outputs the corresponding combination of position vectors. It is equivariant

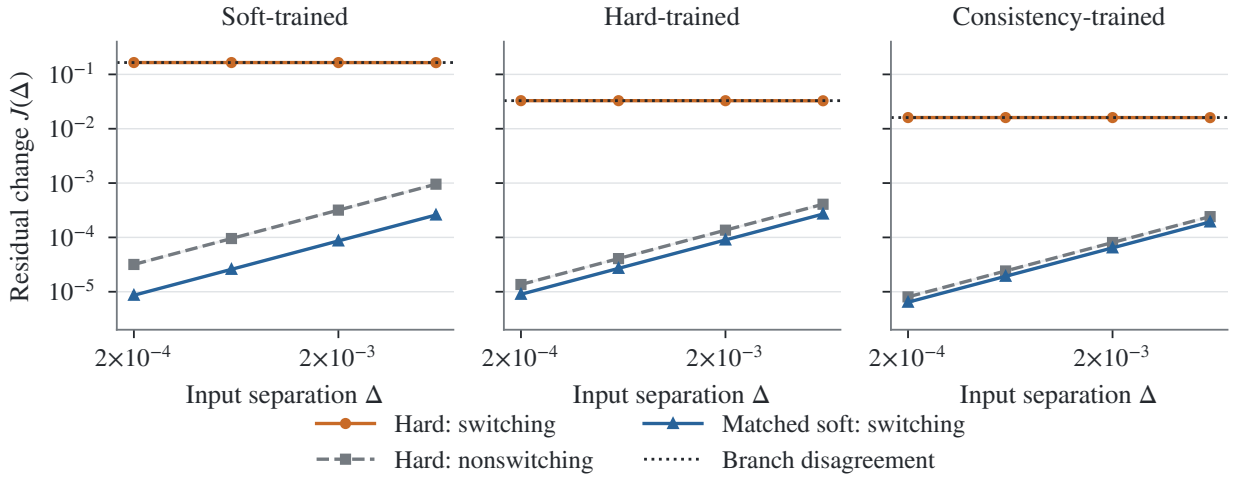


Figure 2: Jumps at a frame switch ($N = 5$, four frames). $J(\Delta)$ is the norm of the change in the prediction residual between two inputs at distance Δ on either side of the switch, divided by the RMS target force. The networks stay fixed and only the switch moves. Panels: networks trained with averaging (soft-trained), with a single frame (hard-trained), and with a penalty on disagreement between frames (consistency-trained). For the single-frame rule that switches between the two inputs (hard: switching), J approaches the branch disagreement (dotted), the norm of the difference between the two frames' predictions, also normalized, as $\Delta \to 0$. For single-frame rules that do not switch there and for averaging (matched soft), J vanishes as $\Delta \to 0$. Shading: paired 95% intervals.

without frames, and its inputs determine every continuous equivariant force on full-rank inputs (Appendix A.9). With averaging, our construction has 14.5%, 36.1% and 54.5% lower NRMSE than Gram at $N = 12, 16, 24$, so its advantage grows with N. Gram has lower NRMSE at $N = 5, 8$ (Figure 3). Averaging also has 13–23% lower NRMSE than hard selection on the same bank, and lower NRMSE than hard selection among all frames built from three input vectors (Table 9). It costs more per update, because it evaluates the network once per frame. At $N = 24$ one training run on one NVIDIA A40 GPU, including development-set evaluation every 500 updates, takes 12.32 hours with averaging, 2.08 with Gram and 1.90 with hard selection. All runs use their final checkpoint.

7.3 Comparison with the Gram baseline at equal training time

Because the methods differ in cost per update, we also retrain them from scratch, targeting comparable GPU time per run (Appendix B.7). The long budget is the time averaging needs for 40,000 updates at $N = 24$, in which hard selection completes 302,459 updates and Gram 336,426. The short budget is the time Gram needs for 40,000 updates, at $N = 24$ and $N = 12$. Learning-rate schedules are scaled to each run's length, and the four comparisons summarized by the ratios of Table 1 were fixed before evaluating the test sets. With the long budget at $N = 24$, hard selection has about half the error of the Gram baseline (NRMSE 0.00296 against 0.00555) and 68% lower NRMSE than training with averaging, in 6.9% less time (Table 1). Gram has 41% lower NRMSE than averaging, and lower NRMSE at the short budgets. In the long $N = 24$ comparison, hard selection is therefore the most compute-efficient of the tested ways to train the bank, and Section 7.4 makes it continuous.

Table 1: Training at equal GPU time. NRMSE pools five seeds on the test sets of Section 7.2. Hours are mean GPU time per run on one NVIDIA A40, without the development-set evaluations of Section 7.2. Ratios divide each method's NRMSE by that of our construction trained with averaging, with simultaneous two-sided 95% intervals over the four comparisons fixed in advance. The short-budget averaging/Gram ratios and intervals are inverted here.

N	Budget	Method	Updates	Hours	NRMSE	Ratio to averaging
24	long	Ours, averaging	40,000	11.49	0.00938	
		Ours, hard selection	302,459	10.70	0.00296	0.316 [0.299, 0.333]
		Gram baseline	336,426	11.48	0.00555	0.592 [0.556, 0.630]
24	short	Ours, averaging	4,633	1.37	0.0256	
		Gram baseline	40,000	1.38	0.0206	0.806 [0.766, 0.848]
12	short	Ours, averaging	10,889	1.08	0.0165	
		Gram baseline	40,000	1.00	0.00836	0.506 [0.495, 0.517]

7.4 Continuous predictions after training

The networks trained with hard selection jump at frame switches. Because the bank supplies all frames, we can average after training. We evaluate the five networks of Table 1 on 20,000 new held-out configurations at $N = 24$, with the procedure fixed before these were generated (Appendix B.8), and compare with hard routing, which uses their highest-scoring frame. Full blending averages over all 61 frames with the weights (5). It is continuous by Theorem 2, and its NRMSE is 6.3% lower than that of hard routing (simultaneous 95% ratio interval $[0.934, 0.941]$). It is our most accurate $N = 24$ predictor (Table 11). Compact blending also gives zero weight to frames whose score is small relative to the largest, through a cutoff as in Dym et al. (2024) and Pozdnyakov and Ceriotti (2023). Its predictions are C^2 for a C^2 network. Its NRMSE is 48.2% lower than that of the separately trained Gram baseline (simultaneous 95% interval 44.9–51.4%) and 2.7% lower than that of hard routing. On one A40 at batch size 1 it takes about as long per input as hard routing. At batch size 32 it takes 11.3 ms, against 4.9 ms for hard routing, 3.1 ms for Gram and 35.8 ms for full blending. In a post hoc control, a uniform average over the eight highest-weight frames recovers 92% of full blending's reduction over hard routing. It can jump when its selected set changes, and no continuous weighted frame can have support at most eight at every full-rank input, since $8 < 11$ (Theorem 1). This control does not isolate continuity's contribution (Appendix B.8). Averaging the same networks gives continuity without retraining and, on this task, lower NRMSE than hard routing.

8 Conclusions and open questions

For rotations and reflections in fixed dimension $d \geq 2$, continuous equivariant weighted frames on full-rank labeled inputs need worst-case support linear in n, and our explicit bank attains this order (Theorem 1). For planar rotations the optimum is exactly $n - 1$ (Corollary 4). The lower bounds, which hold already on perfectly conditioned inputs and apply to robust frames defined on all inputs (Corollary 5), answer the question of Dym et al. (2024). For fixed $d \geq 3$ our full-rank count is linear in n, against order n^{d-1} for their all-input construction. Our determinant-power weights (5) give C^{2k-1} predictions for C^{2k-1} backbones, which is sharp for these weights. On the synthetic $N = 24$ task, compact blending of networks trained with a single frame per input has 48.2% lower NRMSE than the separately trained Gram baseline at comparable training time.

Open questions are sharp constants for $d \geq 3$, optimal support bounds with permutation symmetry, explicit linear-size frames on ranks at least $d-1$, and coverage and conditioning bounds as n grows (Proposition 15). Long simulations and energy-based force models are the next tests of the benefit of smoothness.

Reproducibility statement

The appendices give proofs, probe parameters, study-specific data generators, models, training rules, endpoints, and interval constructions.

AI use statement

LLMs assisted with manuscript restructuring and editing, literature checks, development and checking of mathematical arguments, interpretation of reported results, and preparation of figures from reported values. LLM-based coding agents also substantially wrote and orchestrated the experimental code, including data generation, model implementations, training and evaluation pipelines, and statistical analysis. The authors specified the experimental protocols and decision criteria. The authors are responsible for the mathematical arguments, experimental design, numerical claims, and final text.

References

Justin Baker, Shih-Hsin Wang, Tommaso de Fernex, and Bao Wang. An explicit frame construction for normalizing 3D point clouds. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 2456–2473. PMLR, 2024.

Ilyes Batatia, Dávid Péter Kovács, Gregor N. C. Simm, Christoph Ortner, and Gábor Csányi. MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P. Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E. Smidt, and Boris Kozinsky. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications*, 13:2453, 2022.

Pavle V. M. Blagojević, Aleksandra S. Dimitrijević Blagojević, and Günter M. Ziegler. The topological transversal Tverberg theorem plus constraints. In Gergely Ambrus, Imre Bárány, Károly J. Böröczky, Gábor Fejes Tóth, and János Pach, editors, *New Trends in Intuitive Geometry*, volume 27 of *Bolyai Society Mathematical Studies*, pages 49–64. Springer, 2018. doi: 10.1007/978-3-662-57413-3_3. URL `https://arxiv.org/abs/1604.02814v2`. Preprint: arXiv:1604.02814v2.

Stefan Chmiela, Huziel E. Sauceda, Klaus-Robert Müller, and Alexandre Tkatchenko. Towards exact molecular dynamics simulations with machine-learned force fields. *Nature Communications*, 9:3887, 2018. doi: 10.1038/s41467-018-06169-2.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48, pages 2990–2999, 2016.

Octav Cornea, Gregory Lupton, John Oprea, and Daniel Tanré. *Lusternik–Schnirelmann Category*, volume 103 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2003.

Navnath Daundkar and José Manuel García-Calcines. New bounds on distributional sectional category and applications to distributional homotopic distance. arXiv:2608.08298v2, 2026. URL `https://arxiv.org/abs/2608.08298v2`. Version 2, 26 October 2026; cited theorem numbers refer to this version.

Albrecht Dold. Simple proofs of some Borsuk–Ulam results. In *Proceedings of the Northwestern Homotopy Theory Conference*, volume 19 of *Contemporary Mathematics*, pages 65–69. American Mathematical Society, 1983. doi: 10.1090/conm/019/711043.

Alexander Dranishnikov and Ekansh Jauhari. Distributional topological complexity and LS-category. In Michael Farber and Jesús González, editors, *Topology and AI: Topological Aspects of Algorithms for Autonomous Motion*, volume 4 of *EMS Series in Industrial and Applied Mathematics*, pages 363–385. EMS Press, 2024.

Alexander N. Dranishnikov. The Lusternik–Schnirelmann category and the fundamental group. *Algebraic & Geometric Topology*, 10(2):917–924, 2010. doi: 10.2140/agt.2010.10.917.

Alexandre Duval, Victor Schmidt, Alex Hernández-García, Santiago Miret, Fragkiskos D. Malliaros, Yoshua Bengio, and David Rolnick. FAENet: Frame averaging equivariant GNN for materials modeling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 9013–9033. PMLR, 2023.

Nadav Dym, Hannah Lawrence, and Jonathan W. Siegel. Equivariant frames and the impossibility of continuous canonicalization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 12228–12267, 2024.

Alexandre Eremenko and Andrei Gabrielov. Degrees of real Wronski maps. *Discrete & Computational Geometry*, 28:331–347, 2002.

Michael Farber. Topological complexity of motion planning. *Discrete & Computational Geometry*, 29(2):211–221, 2003.

Michael A. Forbes and Venkatesan Guruswami. Dimension expanders via rank condensers. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2015)*, volume 40 of *Leibniz International Proceedings in Informatics*, pages 800–814, 2015. doi: 10.4230/LIPIcs.APPROX-RANDOM.2015.800. URL `https://arxiv.org/abs/1411.7455`. Full version: arXiv:1411.7455.

Michael A. Forbes, Ramprasad Saptharishi, and Amir Shpilka. Hitting sets for multilinear read-once algebraic branching programs, in any order. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, pages 867–875. ACM, 2014. doi: 10.1145/2591796.2591816. URL `https://arxiv.org/abs/1309.5668`. Full version: Pseudorandomness for Multilinear Read-Once Algebraic Branching Programs, in any Order, arXiv:1309.5668 (2013).

Mario Geiger and Tess Smidt. e3nn: Euclidean neural networks. *arXiv preprint arXiv:2207.09453*, 2022.

Andrew M. Gleason. Spaces with a compact Lie group of transformations. *Proceedings of the American Mathematical Society*, 1(1):35–43, 1950. doi: 10.2307/2032430. Erratum, volume 1, number 6, page 826 (1950).

Allen Hatcher. *Algebraic Topology*. Cambridge University Press, 2002.

Nicholas J. Higham. Computing the polar decomposition—with applications. *SIAM Journal on Scientific and Statistical Computing*, 7(4):1160–1174, 1986.

Dale Husemoller. *Fibre Bundles*. Springer, 3 edition, 1994.

Ekansh Jauhari. On sequential versions of distributional topological complexity. *Topology and its Applications*, 363:109271, 2025.

S.-O. Kaba, Arnab Kumar Mondal, Yan Zhang, Yoshua Bengio, and Siamak Ravanbakhsh. Equivariance with learned canonicalization functions. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 15546–15566, 2023.

Jinwoo Kim, Tien Dat Nguyen, Ayhan Suleymanzade, Hyeokjun An, and Seunghoon Hong. Learning probabilistic symmetrization for architecture agnostic equivariance. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Ben Knudsen and Shmuel Weinberger. Analog category and complexity. *SIAM Journal on Applied Algebra and Geometry*, 8(3):713–732, 2024.

Ben Knudsen and Shmuel Weinberger. On the analog category of finite groups. *Algebraic & Geometric Topology*, 26(4):1585–1596, 2026.

Yuchao Lin, Jacob Helwig, Shurui Gui, and Shuiwang Ji. Equivariance via minimal frame averaging for more symmetries and efficiency. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 30042–30079. PMLR, 2024.

George Ma, Yifei Wang, Derek Lim, Stefanie Jegelka, and Yisen Wang. A canonicalization perspective on invariant and equivariant learning. In *Advances in Neural Information Processing Systems*, volume 37, pages 60936–60979, 2024.

John W. Milnor and James D. Stasheff. *Characteristic Classes*. Princeton University Press, 1974.

Arnab Kumar Mondal, Siba Smarak Panigrahi, Sékou-Oumar Kaba, Sai Rajeswar, and Siamak Ravanbakhsh. Equivariant adaptation of large pretrained models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Art B. Owen. The pigeonhole bootstrap. *The Annals of Applied Statistics*, 1(2):386–411, 2007. doi: 10.1214/07-AOAS122.

Sergey N. Pozdnyakov and Michele Ceriotti. Smooth, exact rotational symmetrization for deep learning on point clouds. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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

Víctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 9323–9332. PMLR, 2021.

Albert S. Schwarz. The genus of a fiber space. *American Mathematical Society Translations, Series 2*, 55:49–140, 1966.

Frank Sottile. *Real Solutions to Equations from Geometry*, volume 57 of *University Lecture Series*. American Mathematical Society, Providence, RI, 2011.

Edwin H. Spanier. *Algebraic Topology*. McGraw-Hill, New York, 1966.

Yonatan Sverdlov, Benjamin Friedman, Snir Hordan, and Nadav Dym. When and how to canonize: A generalization perspective. arXiv:2605.11008, 2026. URL `https://arxiv.org/abs/2605.11008`.

Edward Szpilrajn. La dimension et la mesure. *Fundamenta Mathematicae*, 28(1):81–89, 1937. doi: 10.4064/fm-28-1-81-89.

Behrooz Tahmasebi and Stefanie Jegelka. Regularity in canonicalized models: A theoretical perspective. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 4789–4797. PMLR, 2025.

Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation- and translation-equivariant neural networks for 3D point clouds. arXiv:1802.08219, 2018.

Soledad Villar, David W. Hogg, Kate Storey-Fisher, Weichi Yao, and Ben Blum-Smith. Scalars are universal: Equivariant machine learning, structured like classical physics. In *Advances in Neural Information Processing Systems*, volume 34, pages 28848–28863, 2021.

Yan Zhang, Jonathon Hare, and Adam Prügel-Bennett. FSPool: Learning set representations with featurewise sort pooling. In *International Conference on Learning Representations*, 2020.

A Theory details and proofs

Throughout this appendix, G is a compact Lie group acting on the left. The matrix domain is $\mathcal{M}_{d,n}^{\mathrm{fr}} = \{A \in \mathbb{R}^{d\times n} : \operatorname{rank} A = d\}$.

A.1 Gauges and fiberwise probability measures

Let $q\colon P \to B = P/G$ be a principal left G-bundle over a paracompact Hausdorff base. A gauge is a continuous map $\gamma\colon P \to G$ satisfying $\gamma(gp) = g\gamma(p)$.

A gauge determines a section by $s([p]) = \gamma(p)^{-1}p$. Conversely, a section determines the unique $\gamma(p)$ satisfying $p = \gamma(p)s(q(p))$. This map is continuous in local bundle coordinates and satisfies the gauge identity. Thus gauges and sections are equivalent.

For an orthogonal representation $\rho\colon G \to \mathrm{O}(V)$, use the convention

$$P \times_G V = (P \times V)/((p, v) \sim (gp, \rho(g)v)).$$

A continuous equivariant map $F\colon P \to V$ corresponds to the section $[p] \mapsto [p, F(p)]$ of this associated bundle. Canonicalization chooses an orbit representative. Prediction chooses an output over each orbit. An obstruction to the former need not obstruct the latter, as the zero prediction already illustrates.

Measures and their topology. Write $\mathcal{P}(Z)$ for probability measures on Z with the weak topology and $\mathcal{P}_{\le k}(Z)$ for those supported on at most k points. If Z is compact metrizable, then $\mathcal{P}(Z)$ is compact metrizable. Let $\Delta^{k-1} = \{(a_1, \dots, a_k) : a_i \ge 0,\ \sum_i a_i = 1\}$. The map

$$Z^k \times \Delta^{k-1} \longrightarrow \mathcal{P}_{\le k}(Z), \qquad ((z_i), (a_i)) \longmapsto \sum_{i=1}^{k} a_i \delta_{z_i} \tag{12}$$

is a continuous surjection from a compact space to a Hausdorff space. Consequently it is a quotient map. This topology allows atoms to be permuted, to merge, and to disappear when their weights go to zero. A continuous measure therefore need not have continuously labeled atoms.

Theorem 7 (Weighted frames as fiberwise probability measures)**.** *Let $q\colon P \to B$ be a principal left G-bundle with P, B, G compact metrizable. The space of probability measures supported on at most k points in a single fiber of q is homeomorphic over B to*

$$P \times_G \mathcal{P}_{\le k}(G), \qquad (p, \mu) \sim (hp, h_*\mu), \tag{13}$$

where h_* is pushforward by left translation. Sections of this bundle correspond to continuous equivariant maps $\mu\colon P \to \mathcal{P}_{\le k}(G)$. Hence, with reduced analog sectional category,

$$\mathrm{wfr}_G(P) = \mathrm{asecat}(q) + 1.$$

Proof. Let

$$D_k(q) = \{\nu \in \mathcal{P}_{\le k}(P) : q_*\nu = \delta_b \text{ for some } b \in B\}.$$

The condition $q_*\nu = \delta_b$ says that all mass lies in the fiber over b. This defines a compact Hausdorff space, because pushforward is continuous and the Dirac measures form a closed subset of $\mathcal{P}(B)$. For each $p \in P$, the map $\theta_p\colon G \to q^{-1}(q(p))$, $\theta_p(a) = a^{-1}p$, is a homeomorphism. Define

$$T\colon P \times_G \mathcal{P}_{\le k}(G) \longrightarrow D_k(q), \qquad T([p, \mu]) = (\theta_p)_*\mu.$$

The identity $\theta_{hp}(ha) = \theta_p(a)$ makes T well defined. For every continuous test function $u\colon P \to \mathbb{R}$,

$$(p, \mu) \longmapsto \int_G u(a^{-1}p)\, d\mu(a)$$

is continuous, since compactness gives uniform continuity in the parameter p. Thus T is continuous. Freeness and transitivity of the action in each fiber make it bijective. A continuous bijection from a compact space to a Hausdorff space is a homeomorphism.

An equivariant map $p \mapsto \mu_p$ determines the section $q(p) \mapsto (\theta_p)_*\mu_p$, independently of the chosen lift p. Conversely, suppose $b \mapsto \nu_b$ is a section of $D_k(q)$. Let $d(p, z) \in G$ be the unique element such that $z = d(p, z)^{-1}p$, for $q(p) = q(z)$. The division map d is continuous in bundle coordinates. The measures

$$\mu_p = (d(p, \cdot))_*\nu_{q(p)}$$

vary continuously and satisfy $\mu_{hp} = h_*\mu_p$. These constructions are inverse and preserve support size. The least admissible k is precisely one plus the reduced analog sectional category (Knudsen and Weinberger, 2024). □

For a free action of a compact Lie group on a compact metrizable space, the quotient is a principal bundle (Gleason, 1950, Theorem 3.3). Thus the preceding theorem proves Theorem 3.

From a cover to a measure. For a finite sectional cover $U_i \subset B$ of a numerable principal bundle, let g_i be its local gauges and choose a subordinate partition of unity ϕ_i with $\operatorname{supp} \phi_i \subset U_i$. Then

$$\mu_p = \sum_{i:\phi_i(q(p))>0} \phi_i(q(p))\delta_{g_i(p)}$$

is equivariant and has support at most the number of cover elements. It is weakly continuous. Each continuous test function, bounded on compact G, gives summands that tend to zero where their weights vanish. Consequently $\operatorname{wfr}_G(P) \le \operatorname{genus}(q)$.

Comparison with related definitions. The Schwarz genus also underlies Lusternik–Schnirelmann category and topological complexity (Farber, 2003). Knudsen and Weinberger (2024) give finitely supported measures the quotient topology from $\coprod_{k\ge1} Z^k \times \Delta^{k-1}$. On each $\mathcal{P}_{\le k}(Z)$ for compact metrizable Z, this agrees with the weak topology (Knudsen and Weinberger, 2026, Lemma 5.3). The Lévy–Prokhorov topology used for distributional sectional category agrees with it here as well. In particular, the analog and distributional sectional categories of our compact principal bundles coincide (Knudsen and Weinberger, 2026, Proposition 5.2).

On a free metrizable domain, stabilizer averaging in the robust-frame definition of Dym et al. (2024) has no effect. Robust continuity and weak equivariance therefore become weak continuity and equivariance. Among frames finitely supported at each input, wfr is the minimum worst-case support in that framework. The finite-support qualification in Corollary 5 is essential. On $\mathrm{O}(d)$ or $\mathrm{SO}(d)$ with $d \ge 2$, the constant Haar probability measure is a robust frame with no atoms, but it has infinite support.

A.2 The matrix bundle and deterministic chart count

Define the row and column Stiefel manifolds by

$$\mathrm{St}^{\mathrm{row}}_{d,n} = \{U \in \mathbb{R}^{d\times n} : UU^\top = I_d\}, \qquad \mathrm{St}^{\mathrm{col}}_{n,d} = \{P \in \mathbb{R}^{n\times d} : P^\top P = I_d\}.$$

Whitening, $U(A) = (AA^\top)^{-1/2}A$, is an equivariant retraction to $\mathrm{St}^{\mathrm{row}}_{d,n}$. Indeed, if $H(A) = (AA^\top)^{1/2}$, then

$$H_t(A) = \big((1-t)H(A) + tI_d\big)U(A), \qquad 0 \le t \le 1,$$

is an equivariant deformation retraction through full-row-rank matrices. The resulting compact quotient bundles are

$$\mathrm{St}^{\mathrm{row}}_{d,n}/\mathrm{O}(d) \cong \mathrm{Gr}_d(\mathbb{R}^n), \qquad \mathrm{St}^{\mathrm{row}}_{d,n}/\mathrm{SO}(d) \cong \mathrm{Gr}^+_d(\mathbb{R}^n).$$

Here Gr^+_d denotes the oriented Grassmannian. For the full-rank domain, these identifications are deformation retractions of the quotients, not homeomorphisms to the Grassmannians.

If i is inclusion and $r = U$ is whitening, then $r \circ i = \mathrm{id}$. Pulling local gauges back along r and restricting them along i gives

$$\mathrm{genus}(\mathcal{M}^{\mathrm{fr}}_{d,n} \to \mathcal{M}^{\mathrm{fr}}_{d,n}/G) = \mathrm{genus}(\mathrm{St}^{\mathrm{row}}_{d,n} \to \mathrm{St}^{\mathrm{row}}_{d,n}/G). \tag{14}$$

The same pullback and restriction preserve weighted support. Together with Theorem 7, this gives

$$\mathrm{wfr}_G(\mathcal{M}^{\mathrm{fr}}_{d,n}) = \mathrm{wfr}_G(\mathrm{St}^{\mathrm{row}}_{d,n}) = \mathrm{asecat}(\mathrm{St}^{\mathrm{row}}_{d,n} \to \mathrm{St}^{\mathrm{row}}_{d,n}/G) + 1. \tag{15}$$

Theorem 8 (Deterministic orthogonal-atlas bounds)**.** *For $n > d \ge 1$,*

$$\max\{d, n-d\} + 1 \le \mathrm{genus}(\mathrm{St}^{\mathrm{row}}_{d,n} \to \mathrm{Gr}_d(\mathbb{R}^n)) \le \min\left\{d(n-d)+1, \binom{n}{d}\right\}. \tag{16}$$

For $n = d$, the genus is one.

Proof. Set $m = n - d$. The fiber of the tautological bundle γ_d over a d-plane is that plane itself. Its pullback to the frame bundle has a specified orthonormal basis in each fiber, so it is trivial. Since $\gamma_d \oplus \gamma_d^\perp \cong \mathbb{R}^n$, the pullback of its complement is stably trivial. The Whitney product formula therefore gives

$$w(q^*\gamma_d) = w(q^*\gamma_d^\perp) = 1.$$

In particular, every positive-degree Stiefel–Whitney class of either bundle belongs to $\ker q^*$.

Two nonzero products give the lower bound. Projecting a fixed vector of $\mathbb{R}^n$ onto a d-plane defines a section of γ_d. For m independent vectors $v_1, \dots, v_m$, these sections vanish simultaneously at the unique d-plane V orthogonal to their span. In a graph chart $T\colon V \to V^\perp$, the derivative of the combined section is $T \mapsto (T^\top v_1, \dots, T^\top v_m)$. Since the v_i form a basis of $V^\perp$, this is an isomorphism. The mod–2 zero count gives $\langle w_d(\gamma_d)^m, [\mathrm{Gr}_d(\mathbb{R}^n)]\rangle = 1$. Projecting d independent generic vectors onto the complementary bundle gives, in the same way, $\langle w_m(\gamma_d^\perp)^d, [\mathrm{Gr}_d(\mathbb{R}^n)]\rangle = 1$, because the unique common zero is the plane spanned by those vectors. These are the usual top Stiefel–Whitney zero counts (Milnor and Stasheff, 1974). Schwarz's kernel cup-length bound now gives both $m+1$ and $d+1$ (Schwarz, 1966).

For the coordinate-minor upper bound, let E_I select the columns in a d-element subset $I \subset \{1, \dots, n\}$. The sets $\{U : \det(UE_I) \neq 0\}$ cover by full row rank, and the polar factor of UE_I gives a continuous gauge on each set. There are $\binom{n}{d}$ such sets. Appendix A.5 supplies the $d(n-d)+1$ Wronskian cover. When $n = d$, the polar factor is a global gauge. □

Nonconstructive upper bounds from classical topology. Applying classical dimension and connectivity bounds to the Stiefel bundles gives another proof that linearly many frames suffice. We record this deduction to compare its constants with our explicit construction. Below, the genus and the Lusternik–Schnirelmann category $\mathrm{cat}(B)$ are unreduced. Each counts open sets, and $\mathrm{cat}(B)$ is the fewest open sets covering B whose inclusions into B are null-homotopic.

Proposition 9 (Upper bounds from sectional category)**.** *Let $n > d \geq 2$. Then*

$$\mathrm{wfr}_{\mathrm{SO}(d)}(\mathcal{M}^{\mathrm{fr}}_{d,n}) \leq \mathrm{genus}(\mathrm{St}^{\mathrm{row}}_{d,n} \to \mathrm{Gr}^{+}_{d}(\mathbb{R}^n)) \leq \left\lfloor \frac{d(n-d)}{2} \right\rfloor + 1,$$

and $\mathrm{wfr}_{\mathrm{O}(d)}(\mathcal{M}^{\mathrm{fr}}_{d,n}) \leq \mathrm{genus}(\mathrm{St}^{\mathrm{row}}_{d,n} \to \mathrm{Gr}_{d}(\mathbb{R}^n)) \leq d(n-d)+1$.

Proof. Support is at most genus. A sectional cover with k sets gives an equivariant weighted frame with at most k atoms through a partition of unity (Appendix A.1). By (14) and (15), the genus and the weighted support are the same on $\mathcal{M}^{\mathrm{fr}}_{d,n}$ and on $\mathrm{St}^{\mathrm{row}}_{d,n}$.

Genus is at most the category of the base. Let $q\colon P \to B$ be either Stiefel bundle, and let $U \subseteq B$ be open with a homotopy $H\colon U \times [0,1] \to B$ from a constant map at b_0 to the inclusion. Since q is locally trivial over a paracompact base, it is a Hurewicz fibration. Lifting H with constant initial value $x_0 \in q^{-1}(b_0)$ gives $\tilde{H}$ with $q \circ \tilde{H} = H$, and $\sigma = \tilde{H}(\cdot, 1)$ is a section over U. Writing $x = \gamma(x)\,\sigma(q(x))$ for $x \in q^{-1}(U)$ defines a continuous equivariant gauge γ, because the action is free and q is locally trivial. Hence $\mathrm{genus}(q) \leq \mathrm{cat}(B)$.

Category from dimension and connectivity. If a CW complex B of dimension D is c-connected, then $\mathrm{cat}(B) \leq \lfloor D/(c+1) \rfloor + 1$. For $c = 0$ this is the elementary bound by dimension, and for simply connected B it is the theorem of Grossman and of G. W. Whitehead (Cornea et al., 2003); see also Dranishnikov (2010, Section 1, p. 917). Both Grassmannians are compact smooth manifolds of dimension $d(n-d)$, hence finite CW complexes. The Grassmannian $\mathrm{Gr}_d(\mathbb{R}^n)$ is connected. The oriented Grassmannian is simply connected. In the exact sequence of the fibration $\mathrm{SO}(d) \times \mathrm{SO}(n-d) \to \mathrm{SO}(n) \to \mathrm{Gr}^{+}_{d}(\mathbb{R}^n)$, the group $\pi_0(\mathrm{SO}(d) \times \mathrm{SO}(n-d))$ is trivial, and $\pi_1(\mathrm{SO}(d) \times \mathrm{SO}(n-d)) \to \pi_1(\mathrm{SO}(n))$ is onto, because for $n \geq 3$ a loop of rotations in one coordinate plane generates $\pi_1(\mathrm{SO}(n))$ and lies in $\mathrm{SO}(d)$ when $d \geq 2$. Taking $c = 1$ for $\mathrm{Gr}^{+}_{d}(\mathbb{R}^n)$ and $c = 0$ for $\mathrm{Gr}_d(\mathbb{R}^n)$ gives the two bounds. □

This argument (Schwarz, 1966) gives nonconstructive upper bounds of linear order for fixed d. For $\mathrm{O}(d)$ the resulting bound equals the Wronskian count $d(n-d)+1$. Theorem 6 supplies explicit charts and Theorem 2 controls the regularity of the resulting prediction. For $\mathrm{SO}(d)$, the resulting bound is about half the Wronskian count and also nonconstructive. For $d = 2$ it equals $n-1$, which is exact by Corollary 4, whose explicit weights attain it.

The lower and upper bounds of Theorem 1 are both new. For the lower bounds, an application of Dold's theorem (Dold, 1983) to measures with bounded support gives one (Appendix A.3). The characteristic-class arguments of Section 5.2 improve its slope from $1/(d(d-1)/2+1)$ to $1/2$, or to 1 for even d. For the upper bounds, Theorems 6 and 2 give the explicit construction of weighted frames with $d(n-d)+1$ atoms and sharp regularity, and the proposition above gives the nonconstructive alternative.

A.3 Measures with bounded support

Set $\mathcal{Y}_k = \mathcal{P}_{\leq k}(G)$ and $m_G = \dim G$. Choose a faithful orthogonal embedding of G and restrict the Frobenius metric to obtain a bounded bi-invariant metric d_G of Hausdorff dimension m_G. The

associated Wasserstein metric W_1 metrizes weak convergence on compact G. For paired atoms,

$$W_1\left(\sum_i a_i\delta_{g_i}, \sum_i b_i\delta_{h_i}\right) \le \sum_i a_i d_G(g_i, h_i) + \frac{\operatorname{diam} G}{2}\sum_i |a_i - b_i|. \tag{17}$$

Indeed, first move mass a_i from g_i to h_i, and then redistribute the excess mass among the h_i.

Lemma 10 (Geometry of the measure space)**.** *For $k \ge 1$, the space $\mathcal{Y}_k$ is compact metrizable, and*

$$\dim_{\mathrm{cov}}(\mathcal{Y}_k/G) \le (k-1)(m_G+1).$$

A measure with r support points has finite stabilizer of order dividing r.

Proof. Compactness and metrizability follow from (12). Left translation acts isometrically on measures, so

$$\bar{d}([\mu],[\nu]) = \min_{g\in G} W_1(\mu, g_*\nu)$$

is a metric inducing the orbit-space topology. Translating one support point to the identity gives a Lipschitz surjection

$$G^{k-1}\times\Delta^{k-1} \longrightarrow \mathcal{Y}_k/G,$$
$$((h_2,\dots,h_k),(a_1,\dots,a_k)) \longmapsto \left[a_1\delta_e + \sum_{i=2}^{k} a_i\delta_{h_i}\right].$$

Padding with zero weights includes smaller supports, and (17) gives the Lipschitz property. The domain has Hausdorff dimension $(k-1)(m_G+1)$. A Lipschitz map cannot increase Hausdorff dimension, which bounds covering dimension for metric spaces (Szpilrajn, 1937, Theorem 1).

The stabilizer of a measure permutes its r support points faithfully. An element fixing a point g satisfies $hg = g$, hence $h = e$. It is therefore finite, and its free action on the support shows that its order divides r. □

A weaker bound from connectivity. For $G \in \{\mathrm{O}(d), \mathrm{SO}(d)\}$ and $n > d \ge 2$, Dold's obstruction (Dold, 1983, p. 68) gives the weaker estimate

$$\mathrm{wfr}_G(\mathcal{M}^{\mathrm{fr}}_{d,n}) \ge \left\lfloor \frac{n-d}{d(d-1)/2+1} \right\rfloor + 1.$$

Indeed, a planar rotation subgroup of prime order $p > k$ acts freely on $\mathcal{Y}_k$, whose covering dimension is at most $k(m_G+1)-1$ by (12) and (17). The $(n-d-1)$-connectivity of $\mathrm{St}^{\mathrm{row}}_{d,n}$ (Husemoller, 1994) excludes an equivariant map when $k(m_G+1) \le n-d$. The cohomological arguments below improve this estimate.

A.4 Rational cohomology and characteristic classes

The three lower-bound arguments use different information:

Argument	Input	Result for fixed d
Dold (A.3)	Connectivity and target dimension	Linear lower bound, with weaker constants.
Degree (below)	Degree of one nonzero characteristic class	Recovers the main estimates for $d = 2, 3$.
Cup length (Section 5.2)	Number of factors in a nonzero product	Main estimates, including $n-d+1$ for even d.

The direct degree argument compares the class degree with the orbit-space dimension in Lemma 10.

Theorem 11 (Cohomological degree bound)**.** *Let $n > d \geq 2$, let $m = d(d-1)/2$, and suppose the characteristic-class map $H^D(B\mathrm{SO}(d);\mathbb{Q}) \to H^D(\mathrm{Gr}_d^+(\mathbb{R}^n);\mathbb{Q})$ is nonzero. Then*

$$\mathrm{wfr}_{\mathrm{O}(d)}(\mathcal{M}_{d,n}^{\mathrm{fr}}) \geq \mathrm{wfr}_{\mathrm{SO}(d)}(\mathcal{M}_{d,n}^{\mathrm{fr}}) \geq 1 + \left\lceil \frac{D}{m+1} \right\rceil.$$

One can take $D = 4\lfloor (n-d)/2 \rfloor$, and $D = d(n-d)$ for even d.

We prove this bound using a finite Stiefel manifold in place of the universal bundle. Finite stabilizers make the relevant fibers rationally acyclic through the required degree, so the characteristic class must vanish if that degree exceeds the orbit-space dimension. This gives the same estimates as (8) for $d = 2, 3$ with the classes above. For $d \geq 4$, cup length gives at least as strong a bound.

Lemma 12 (Nonzero characteristic-class powers)**.** *Let $n > d \geq 2$, set $c = n - d$, and let γ_d be the tautological oriented bundle on $\mathrm{Gr}_d^+(\mathbb{R}^n)$. Then*

$$p_1(\gamma_d)^{\lfloor c/2 \rfloor} \neq 0 \quad \textit{in rational cohomology.}$$

For even d, one also has $e(\gamma_d)^c \neq 0$.

Proof. Fix an oriented $(d-2)$-plane and use its orthogonal complement to define

$$\iota \colon \mathrm{Gr}_2^+(\mathbb{R}^{c+2}) \longrightarrow \mathrm{Gr}_d^+(\mathbb{R}^n), \qquad V \longmapsto \mathbb{R}^{d-2} \oplus V.$$

The pullback of γ_d is a trivial bundle plus γ_2. Hence $\iota^* p_1(\gamma_d) = p_1(\gamma_2) = e(\gamma_2)^2$ (Milnor and Stasheff, 1974).

For an oriented orthonormal basis (u, v) of a two-plane, the complex line $[u + iv]$ is independent of that basis. This identifies $\mathrm{Gr}_2^+(\mathbb{R}^{c+2})$ with the smooth quadric

$$Q^c = \{[z_0 : \cdots : z_{c+1}] \in \mathbb{CP}^{c+1} : z_0^2 + \cdots + z_{c+1}^2 = 0\}.$$

Indeed, the real and imaginary parts of an isotropic vector are orthogonal and have equal length, so they recover the oriented plane. For $z = u + iv$, the map $\lambda z \mapsto \mathrm{Re}(\lambda z)$ identifies the tautological complex line, as a real plane, with $\mathrm{span}(u, v)$, up to orientation. Its Euler class is therefore $\pm h$, where h is the hyperplane class restricted to Q^c. Since a quadric has degree two, $\langle h^c, [Q^c] \rangle = 2$. Every power h^a with $0 \leq a \leq c$ is consequently nonzero. Taking $a = 2\lfloor c/2 \rfloor$ proves the Pontryagin assertion. For $d = 2$, taking $a = c$ also proves the Euler assertion.

For general even d, nonvanishing of the Euler power is classical (Blagojević et al., 2018, Lemma 3.3(2)). The following zero count gives its integral evaluation. Choose orthonormal vectors $v_1, \ldots, v_c$ and consider the section

$$s(V) = (\pi_V v_1, \ldots, \pi_V v_c) \quad \text{of } \gamma_d^{\oplus c},$$

where π_V is orthogonal projection onto V. Its two zeros are the two orientations of $V_0 = \mathrm{span}(v_1, \ldots, v_c)^\perp$. In a graph chart $T \colon V_0 \to V_0^\perp$, identify the nearby plane with V_0. The derivative is $T \mapsto (T^\top v_1, \ldots, T^\top v_c)$, an isomorphism, so both zeros are transverse.

The oriented Grassmannian is an orientable manifold of dimension dc. The involution reversing the orientation of V reverses the orientations of both V and $V^\perp$. On the tangent space $\mathrm{Hom}(V, V^\perp)$ its orientation sign is $(-1)^{d+c} = (-1)^n$, while on $\gamma_d^{\oplus c}$ it is $(-1)^c$. The two local indices thus differ by the factor $(-1)^d$ and agree when d is even. Consequently,

$$\langle e(\gamma_d)^c, [\mathrm{Gr}_d^+(\mathbb{R}^n)] \rangle = \pm 2,$$

which proves rational nonvanishing. □

Proof of Theorem 11. Set $X = \mathrm{St}_{d,n}^{\mathrm{row}}$ and $G = \mathrm{SO}(d)$, and suppose there is an equivariant map $\mu\colon X \to \mathcal{Y}_k$. Let $c \in H^D(BG;\mathbb{Q})$ be a class whose image $c(\gamma^{(n)})$ on $\mathrm{Gr}_d^+(\mathbb{R}^n)$ is nonzero, where $\gamma^{(N)}$ denotes the tautological oriented bundle on $\mathrm{Gr}_d^+(\mathbb{R}^N)$. Choose $N \geq \max\{n, D+d+2\}$ and set $E = \mathrm{St}_{d,N}^{\mathrm{row}}$. This is a compact $(N-d-1)$-connected manifold with a free G-action (Husemoller, 1994). Appending zero columns gives an equivariant inclusion $\iota_N\colon X \to E$. On quotients it pulls $\gamma^{(N)}$ back to $\gamma^{(n)}$.

The map $x \mapsto (\mu_x, \iota_N(x))$ induces the factorization

$$X/G \xrightarrow{\bar{\mu}} (\mathcal{Y}_k \times E)/G \xrightarrow{\bar{\pi}_2} E/G, \qquad \bar{\pi}_2 \circ \bar{\mu} = \bar{\iota}_N.$$

The other projection, $\bar{\pi}_1\colon (\mathcal{Y}_k \times E)/G \to \mathcal{Y}_k/G$, is a closed surjection between compact metrizable spaces. Its fiber over $[y]$ is E/G_y, where G_y is finite by Lemma 10. Transfer for the finite covering gives $H^*(E/G_y;\mathbb{Q}) \cong H^*(E;\mathbb{Q})^{G_y}$ (Hatcher, 2002, Proposition 3G.1). Thus the fibers have vanishing reduced rational cohomology through degree $D+1$. They are closed manifolds, so their singular and Čech cohomology agree. The Vietoris–Begle theorem (Spanier, 1966, Theorem 6.9.15) therefore gives

$$\bar{\pi}_1^*\colon \check{H}^D(\mathcal{Y}_k/G;\mathbb{Q}) \xrightarrow{\cong} \check{H}^D((\mathcal{Y}_k \times E)/G;\mathbb{Q}).$$

If $D > (k-1)(m+1)$, the group on the left vanishes by Lemma 10. Indeed, finite open covers then admit refinements whose nerves have dimension less than D, so the defining direct limit for Čech cohomology is zero. Naturality now gives

$$c(\gamma^{(n)}) = \bar{\mu}^* \bar{\pi}_2^* c(\gamma^{(N)}) = 0,$$

a contradiction. On the Grassmannians, Čech and singular cohomology agree. Hence $D \leq (k-1)(m+1)$, proving the support bound, and Lemma 12 supplies the stated degrees.

To compare the two groups, fix a reflection J and define $r\colon \mathrm{O}(d) \to \mathrm{SO}(d)$ by $r(g) = g$ for $\det g = 1$ and $r(g) = gJ$ otherwise. This map is continuous on the two components and commutes with left multiplication by $\mathrm{SO}(d)$. Pushing an orthogonal weighted frame forward by r gives a special-orthogonal weighted frame with no more atoms. Finally, whitening transfers both bounds to the matrix domain. □

The cup-length bound in Theorem 1. The lower bound in Theorem 1 uses the following result, quoted from version 2 of Daundkar and García-Calcines (2026), where it is Theorem 4.17 (Theorem 4.21 in version 1). Here dsecat is the reduced distributional sectional category of Dranishnikov and Jauhari (2024): $\mathrm{dsecat}(p)$ is the least $k \geq 0$ for which there is a continuous map $b \mapsto \nu_b$ from B to the space $\mathcal{B}_{k+1}(E)$ of probability measures on E with at most $k+1$ atoms, in the Lévy–Prokhorov topology, such that $\operatorname{supp}\nu_b \subseteq p^{-1}(b)$ for every $b \in B$. It is infinite if no such k exists (Daundkar and García-Calcines, 2026, Definition 2.2). For a graded ring map p^*, $\mathrm{cl}_{\mathbb{Q}}(\ker p^*)$ is the rational cup length of Section 5.2.

Lemma 13 (Rational kernel cup length)**.** *Let $p\colon E \to B$ be a fibration whose total space E is metric, and suppose that E and B are path-connected and have the homotopy type of locally finite CW complexes. Then*

$$\mathrm{dsecat}(p) \geq \mathrm{cl}_{\mathbb{Q}}\big(\ker(p^*\colon H^*(B;\mathbb{Q}) \to H^*(E;\mathbb{Q}))\big).$$

Checking the hypotheses for $q\colon \mathrm{St}_{d,n}^{\mathrm{row}} \to \mathrm{Gr}_d^+(\mathbb{R}^n)$. For $n > d$, the map q is a principal $\mathrm{SO}(d)$-bundle over a compact manifold, so it is locally trivial over a paracompact base and hence a Hurewicz fibration. Its total space and base are connected compact smooth manifolds. They are therefore

metrizable and path-connected, and, being triangulable, they have the homotopy type of finite CW complexes. Lemma 13 therefore gives $\operatorname{dsecat}(q) \geq \operatorname{cl}_{\mathbb{Q}}(\ker q^*)$.

From distributional category to weighted support. The map q is proper, since its total space is compact and its base Hausdorff, and its base is compact Hausdorff, hence convenient in the sense of Knudsen and Weinberger (2024). Proposition 5.2 of Knudsen and Weinberger (2026) then identifies the distributional and analog sectional categories of q. For each fixed support bound, the analog and Lévy–Prokhorov topologies on measures over the compact Stiefel manifold agree with the weak topology (Appendix A.1). Theorem 3, applied to the free $\mathrm{SO}(d)$-action on $\mathrm{St}^{\mathrm{row}}_{d,n}$, gives $\operatorname{wfr}_{\mathrm{SO}(d)}(\mathrm{St}^{\mathrm{row}}_{d,n}) = \operatorname{asecat}(q) + 1 = \operatorname{dsecat}(q) + 1$. The added one converts the reduced category, which counts atoms beyond the first, into a number of atoms. Whitening transfers the bound to $\mathcal{M}^{\mathrm{fr}}_{d,n}$, which yields (8).

The classes. The tautological bundle is trivial after pullback, so $p_1(\gamma_d)$ and $e(\gamma_d)$ belong to $\ker q^*$. Lemma 12 gives products of lengths $\lfloor (n-d)/2 \rfloor$ in general and $n-d$ for even d. Equation (8) now yields the lower bounds in Theorem 1. The map r just constructed transfers them to $\mathrm{O}(d)$. Theorems 6 and 2 give its upper bound, and Corollary 4 gives the planar equality.

Well-conditioned inputs and support. The lower bounds already hold on $\mathrm{St}^{\mathrm{row}}_{d,n}$, where all nonzero singular values equal one. They therefore apply to any continuous equivariant weighted frame defined on a neighborhood of that manifold, irrespective of its behavior near rank loss. There is also a local consequence. For any weakly continuous measure map $A \mapsto \mu_A$, the set

$$\{A : |\operatorname{supp} \mu_A| \geq L\}$$

is open for each positive integer L. Indeed, at an input with at least L support points, choose L nonnegative continuous functions on G with disjoint supports and positive integrals against that measure. The integrals remain positive nearby, forcing at least one support point in each of these disjoint sets.

Take L to be the lower bound in Theorem 1. Some Stiefel input requires at least L atoms. For every $\varepsilon > 0$, a nonempty open neighborhood of that input therefore requires at least L atoms and has condition number less than $1 + \varepsilon$.

For example, the PCA frame of Puny et al. (2022) has at most eight atoms for $\mathrm{O}(3)$ and four for $\mathrm{SO}(3)$ on the simple-spectrum domain. With $n = N - 1$, Theorem 1 rules out continuous equivariant extensions with those support bounds to all full-rank centered inputs when $N \geq 20$ and $N \geq 12$, respectively.

The planar domain distinction. The space of all nonzero $2 \times n$ matrices retracts equivariantly to $S^{2n-1} \subset \mathbb{C}^n$. Its rotation quotient is the Hopf bundle over $\mathbb{CP}^{n-1}$. The Euler class of this circle bundle has nonzero $(n-1)$-st power. The same factorization, using $E = S^{2N-1}$ for sufficiently large N, $m = 1$, and $D = 2(n-1)$, gives support at least n. Conversely, (9), with all n columns included, defines a continuous frame there with at most n atoms. The improvement to $n - 1$ on the full-rank domain is possible because every selection of $n - 1$ columns then contains a nonzero column.

A.5 The Wronskian probe cover

Geometric interpretation. Probe j fails exactly when the row space of A contains a nonzero polynomial vanishing to order d at t_j, that is, on the Schubert divisor of the osculating flag at t_j. Theorem 6 is the classical fact that $D + 1$ such divisors have no common point (Eremenko and Gabrielov, 2002; Sottile, 2011).

Relation to rank condensers. A lossless rank condenser supplies a family $E_j \in \mathbb{R}^{d\times n}$ such that, for every rank-d matrix $M \in \mathbb{R}^{n\times d}$, some E_jM has rank d. Take thin QR, $E_j^\top = P_jR_j$, and set $M = A^\top$. Then

$$E_jA^\top = R_j^\top (AP_j)^\top,$$

so this property is exactly coverage by the probes P_j. Ordinary Wronskians are classical rank condensers in characteristic zero (Forbes et al., 2014, Section 3.1 of the full version). A folded-Wronskian construction of Forbes and Guruswami (2015, Proposition 6.2 and Corollary 6.3) gives the same count. For example, over $\mathbb{R}$ one can take $E(t)_{ab} = (2^at)^b$, $0 \le a < d$, $0 \le b < n$. After removing a power of t, $\det(E(t)M)$ is a nonzero polynomial of degree at most $d(n-d)$, so $d(n-d)+1$ distinct nonzero nodes suffice. This folded bank and our derivative bank need not coincide.

We build on this algebraic coverage principle to construct a continuous equivariant weighted frame with an explicit linear support bound, and our determinant weights give the sharp regularity guarantee. The support lower bounds address the complementary topological question. We include the short proof to make the passage from probes to weighted frames explicit.

Derivative probes. Let $\psi_0, \dots, \psi_{n-1}$ be degree-graded real polynomials, so $\deg \psi_j = j$, and write $b(t) = (\psi_0(t), \dots, \psi_{n-1}(t))^\top$. As in (10), form the $n \times d$ derivative matrix

$$\widetilde{B}(t) = \big[b(t), b'(t), \dots, b^{(d-1)}(t)\big].$$

Its first d rows form a triangular block with nonzero diagonal, so $\widetilde{B}(t)$ has rank d for every real t.

Lemma 14 (Polynomial Wronskian bound)**.** *If $f_1, \dots, f_d$ are linearly independent real polynomials of degree less than n, their Wronskian is nonzero and has degree at most $d(n-d)$.*

Proof. A constant invertible row operation replaces the family by polynomials with distinct leading degrees $0 \le e_1 < \cdots < e_d \le n-1$ and nonzero leading coefficients c_i. It multiplies the Wronskian by a nonzero constant. For the new family, the coefficient of the highest possible power of t is

$$\left(\prod_i c_i\right) \det\big[(e_i)_{r-1}\big]_{i,r=1}^{d} = \left(\prod_i c_i\right) \prod_{i<j} (e_j - e_i) \neq 0,$$

where $(e)_s = e(e-1)\cdots(e-s+1)$ and $(e)_0 = 1$. The determinant is a Vandermonde determinant because the falling factorials are monic polynomials of successive degrees. Thus

$$\deg W = \sum_i e_i - \frac{d(d-1)}{2} \le \sum_{r=0}^{d-1} (n-d+r) - \frac{d(d-1)}{2} = d(n-d).$$

□

Proof of Theorem 6. Choose a positive diagonal matrix $\Lambda(t)$ and take thin QR,

$$\widetilde{B}(t)\Lambda(t) = P(t)R(t), \qquad P(t) \in \mathrm{St}_{n,d}^{\mathrm{col}},$$

with positive diagonal in $R(t)$. For a row frame U with rows u_i, set $f_i(t) = u_ib(t)$. The basis coefficient map is an isomorphism, so the f_i are linearly independent. Their Wronskian is $W_U(t) = \det(U\widetilde{B}(t))$, a nonzero polynomial of degree at most $D = d(n-d)$. Hence it cannot vanish at all $D+1$ distinct real nodes t_j. Moreover,

$$\det(U\widetilde{B}(t_j)\Lambda(t_j)) = \det(UP(t_j)) \det R(t_j),$$

and both $\det \Lambda(t_j)$ and $\det R(t_j)$ are positive. Thus the fixed probes $P_j = P(t_j)$ include at least one with $\det(UP_j) \neq 0$ for every U. Whitening transfers coverage to every full-row-rank A. Local polar gauges and determinant weights give an equivariant measure supported on at most $D+1$ atoms. □

The experiments use normalized Legendre polynomials,

$$\psi_k(t) = \sqrt{\frac{2k+1}{2}}\, P_k(t), \qquad k = 0, \dots, n-1, \tag{18}$$

where P_k is the usual Legendre polynomial. Each derivative column is multiplied by its positive inverse Euclidean norm, followed by unpivoted thin Householder QR with positive diagonal in R. For $K = d(n-d)+1$, the nodes are

$$t_j = \cos\left(\frac{(2j-1)\pi}{2K}\right), \qquad j = 1, \dots, K. \tag{19}$$

The proof applies to any degree-graded polynomial basis, positive column scalings, and distinct real nodes. Equations (18)–(19), together with the stated QR convention, fix the probes used in the experiments.

A.6 Continuity, Lipschitz dependence, and higher derivatives

Fix a finite family of probes covering the full-rank domain and an integer $k \geq 1$. The determinant scores before and after whitening are

$$s_j(A) = \det(AP_j)^{2k}, \qquad \bar{s}_j(A) = \det(U(A)P_j)^{2k}.$$

Since $A = H(A)U(A)$ with $H(A) = (AA^\top)^{1/2}$, $s_j(A) = \det(H(A))^{2k}\bar{s}_j(A)$. Thus they give identical normalized weights. Their polar factors generally differ:

$$\operatorname{polar}(H(A)U(A)P_j) \neq \operatorname{polar}(U(A)P_j) \quad \text{in general.}$$

The particle-count study uses $k = 1$, the raw polar factor $Q_j(A) = \operatorname{polar}(AP_j)$, and the whitened scores for the weights.

For $G = \mathrm{O}(d)$, put $g_j^G = Q_j$. For $G = \mathrm{SO}(d)$, fix $J \in \mathrm{O}(d)$ with $\det J = -1$ and set

$$g_j^G(A) = \begin{cases} Q_j(A), & \det(AP_j) > 0, \\ Q_j(A)J, & \det(AP_j) < 0. \end{cases}$$

The two sign components of the valid chart are disjoint and open. On each, g_j^G is continuous, belongs to G, and satisfies $g_j^G(RA) = Rg_j^G(A)$ for $R \in G$. Define

$$Y_j(A) = \rho(g_j^G(A))f((g_j^G(A))^{-1}A), \qquad N_j(A) = \det(AP_j)^{2k}Y_j(A) \tag{20}$$

on the valid chart and extend N_j by zero elsewhere.

Continuity and equivariance in Theorem 2. On a valid chart, the gauge law gives $Y_j(RA) = \rho(R)Y_j(A)$, and the determinant score is invariant. If $A_r \to A_0$ while chart j becomes singular, then $\det(A_rP_j)^{2k} \to 0$. The gauges lie in compact G. Canonicalizing inputs in a sufficiently small compact neighborhood of A_0 therefore keeps them in a compact subset of the full-rank domain. Continuity of f makes Y_j bounded there, so $N_j(A_r) \to 0$. This proves that its zero extension is continuous, without assigning a polar factor at a singular matrix. Coverage gives $Z(A) = \sum_j \det(AP_j)^{2k} > 0$, so $F_{\text{soft}} = \sum_j N_j/Z$ is continuous. Equivariance also holds at chart failures because the zero extensions preserve it. The same argument with a continuous test function $\varphi(g_j(A))$ in place of $Y_j(A)$ proves weak continuity of the measure μ_A itself. □

A.6.1 Lipschitz dependence away from rank loss

Proposition 15 (Rank-margin Lipschitz estimate)**.** *Fix $\varepsilon, R > 0$ and let*

$$K_{\varepsilon,R} = \{A : \sigma_{\min}(A) \geq \varepsilon,\ \|A\|_F \leq R\}.$$

If f is bounded and Lipschitz on $K_{\varepsilon/2,R}$, then F_{soft} is Lipschitz on $K_{\varepsilon,R}$. Its Lipschitz constant depends only on ε, R, the backbone bound and Lipschitz constant, the fixed representation, probe bank, and exponent k, and the coverage margin

$$\Gamma_* = \min_{U \in \mathrm{St}^{\mathrm{row}}_{d,n}} \left(\sum_j \det(UP_j)^{2k} \right)^{1/2} > 0.$$

Proof. Compactness of the Stiefel manifold and coverage give $\Gamma_* > 0$. The main issue is that the polar factor becomes sensitive near a singular projection. We show that the determinant weight compensates for this sensitivity. For a differentiable curve of invertible matrices $Z = QS$, polar perturbation gives

$$\|\dot{Q}\|_F \leq \frac{\|\dot{Z}\|_F}{\sigma_{\min}(Z)} \qquad \text{(Higham, 1986)}.$$

The fixed-reflection correction has the same derivative bound on each sign component. The representation ρ is smooth with bounded derivative on compact G.

Take $A, B \in K_{\varepsilon,R}$ with $\|A - B\|_F < \varepsilon/2$. Along their line segment, Weyl's inequality gives rank margin at least $\varepsilon/2$, and the Frobenius norm is at most R. Thus the segment and all its canonicalized inputs lie in $K_{\varepsilon/2,R}$. Convexity of $K_{\varepsilon,R}$ is not required. For a probe P, the only potentially unbounded term in the derivative of $N(A) = \det(AP)^{2k} Y(A)$ comes from differentiating the polar factor. Its coefficient is bounded because

$$\frac{\det(AP)^{2k}}{\sigma_{\min}(AP)} = \left(\prod_{i=1}^{d-1} \sigma_i(AP)^{2k} \right) \sigma_d(AP)^{2k-1} \leq R^{2kd-1}.$$

The determinant and its derivative are also bounded on the radius-R ball. On each nonsingular subinterval, the canonical input varies smoothly and the backbone is Lipschitz. The product rule holds almost everywhere and, with the polar estimate, gives a uniform bound on the derivative of N.

Along the segment, $\det(AP)$ is a polynomial of degree at most d. It either vanishes identically, in which case $N = 0$, or has finitely many zeros. Integrating the uniform derivative bound on the intervening subintervals and using the continuous zero extension at their endpoints gives a Lipschitz bound across the entire segment.

Finally, on $K_{\varepsilon,R}$,

$$Z(A) = \det(AA^\top)^k \sum_j \det(U(A)P_j)^{2k} \geq \varepsilon^{2kd} \Gamma_*^2.$$

The denominator is a polynomial with bounded derivative on the radius-R ball. Subtracting two quotients yields the Lipschitz bound for close pairs. For pairs at distance at least $\varepsilon/2$, let M bound the backbone. Since $\|F_{\text{soft}}\| \leq M$, $\|F_{\text{soft}}(A) - F_{\text{soft}}(B)\| \leq (4M/\varepsilon)\|A - B\|_F$. Together these estimates cover all pairs. □

In particular, a locally Lipschitz backbone on the full-rank domain meets the proposition's hypothesis on every compact $K_{\varepsilon/2,R}$. A finite neighborhood cover controls nearby pairs, and boundedness controls pairs a fixed distance apart.

The estimate depends on both the input rank margin ε and the probe coverage margin Γ_*. Coverage ensures $\Gamma_* > 0$ at each fixed n, but the polynomial-root argument of Theorem 6 does not bound its decay as n grows. Appendix B.2 describes the numerical checks used for the implemented probes.

A.6.2 Higher derivatives and sharpness

The same compensation extends to higher derivatives. We give the polar estimate first, since it identifies the possible singularity of a branch.

Lemma 16 (Derivatives of the polar factor)**.** *For every integer $r \geq 1$ there is C_r such that, for invertible real $d \times d$ matrices Z,*

$$\|D^r \operatorname{polar}(Z)\| \leq C_r \sigma_{\min}(Z)^{-r}.$$

The norm on the left is the operator norm of the multilinear derivative.

Proof. Consider the symmetric matrix

$$H(Z) = \begin{pmatrix} 0 & Z \\ Z^\top & 0 \end{pmatrix}.$$

Its eigenvalues are $\pm\sigma_i(Z)$, and the upper-right block of $\operatorname{sign} H(Z) = H(Z)(H(Z)^2)^{-1/2}$ is $\operatorname{polar}(Z)$. The resolvent formula

$$\operatorname{sign} H = \frac{1}{\pi} \int_0^\infty \left((H - itI)^{-1} + (H + itI)^{-1}\right) dt$$

follows by diagonalizing H. If $s = \sigma_{\min}(Z)$, each resolvent has norm at most $(s^2 + t^2)^{-1/2}$. Its r-th derivative is a sum of products with $r + 1$ resolvents and r perturbation matrices. Differentiation under the integral is valid locally on the invertible locus, and gives

$$\|D^r \operatorname{sign} H\| \leq C_r \int_0^\infty (s^2 + t^2)^{-(r+1)/2} \, dt = C_r' s^{-r}.$$

Taking the indicated block proves the assertion. Multiplication by the fixed reflection J gives the same estimate on either sign component of a special-orthogonal probe frame. □

Lemma 17 (Zero extension)**.** *Let $\Omega \subseteq \mathbb{R}^p$ be open, $S \subseteq \Omega$ relatively closed, and $r \geq 0$ an integer. Suppose $N \colon \Omega \to V$ vanishes on S and is C^r on $\Omega \setminus S$. If $D^j N(x) \to 0$ as $x \to x_0$ through $\Omega \setminus S$ for every $x_0 \in S$ and $0 \leq j \leq r$, then N is C^r on Ω, with all these derivatives zero on S.*

Proof. We induct on r. The case $r = 0$ is continuity at S. For $r \geq 1$, fix $x_0 \in S$ and $\varepsilon > 0$. Choose a closed ball about x_0 inside Ω on whose complement of S we have $\|DN\| \leq \varepsilon$. For a sufficiently small h with $x_0 + h \notin S$, let t_* be the last $t \in [0, 1]$ for which $x_0 + th \in S$. Along the remaining segment, N is continuous up to t_*, vanishes there, and has derivative in the segment parameter bounded by $\varepsilon\|h\|$. Integrating gives $\|N(x_0 + h)\| \leq \varepsilon\|h\|$. The same bound holds when $x_0 + h \in S$. Thus $DN(x_0) = 0$. Its zero extension is continuous and satisfies the hypotheses with $r - 1$, so induction completes the proof. □

Smoothness in Theorem 2. Fix a sufficiently small compact neighborhood K of a full-rank input. The set $G \cdot K$ is a compact subset of the full-rank domain. The derivatives of a C^{2k-1} backbone are bounded there, as are all derivatives of the representation on G.

For one probe, write $h(A) = \det(AP)$ and $s(A) = \sigma_{\min}(AP)$. The integer exponent $k \geq 1$ is fixed in advance, independently of the input, and s enters only the estimates below. Since $s(A) \leq \|A\|$ is bounded above on K, a bound Cs^{-j} with $j \leq r$ implies a bound $C's^{-r}$ there. We use this to absorb lower-order terms.

On the valid chart of the probe, write $g = g_P$, $u(A) = g(A)^{-1}A$ and $v(A) = \rho(g(A))$, so that $Y = v\,(f \circ u)$. We show

$$\|D^r Y(A)\| \leq C_r s(A)^{-r}, \qquad 0 \leq r \leq 2k-1, \tag{21}$$

where the case $r = 0$ means a uniform bound.

1. *The frame.* The map $A \mapsto AP$ is linear with norm at most one, so Lemma 16 gives $\|D^b g(A)\| \leq C_b s(A)^{-b}$ for $b \geq 1$, and $\|g(A)\| = 1$.

2. *The transport factor.* Extend ρ smoothly to a neighborhood of the compact submanifold $G \subset \mathbb{R}^{d\times d}$. Its derivatives are bounded near G. By the Faà di Bruno formula, $D^a v$ for $a \geq 1$ is a sum of terms $D^j\rho(g)[D^{b_1}g, \ldots, D^{b_j}g]$ with $1 \leq j \leq a$, $b_i \geq 1$ and $b_1 + \cdots + b_j = a$, so each term is at most Cs^{-a}; and v itself is orthogonal.

3. *The canonical input.* Here $u(A) = g(A)^\top A$, and $A \mapsto A$ is linear. The Leibniz rule writes $D^b u[H_1, \ldots, H_b]$ as $(D^b g^\top[H_1, \ldots, H_b])\,A$ plus b terms of the form $(D^{b-1}g^\top[\ldots])\,H_i$, with $D^0 g^\top = g^\top$. Hence $\|D^b u\| \leq C_b(s^{-b}\|A\| + s^{-(b-1)}) \leq C'_b s^{-b}$ on K for $b \geq 1$.

4. *The backbone.* By Faà di Bruno again, $D^c(f \circ u)$ is a sum of terms $D^j f(u)[D^{b_1}u, \ldots, D^{b_j}u]$ with $1 \leq j \leq c$, $b_i \geq 1$ and $b_1 + \cdots + b_j = c$. The point $u(A) = g(A)^{-1}A$ lies in the compact set $G \cdot K$, where $D^j f$ is bounded for $j \leq 2k-1$. Each term is therefore at most Cs^{-c}.

5. *The product.* By the Leibniz rule, $D^r Y$ is a sum of terms $D^a v \cdot D^c(f \circ u)$ with $a + c = r$, each at most $Cs^{-a}s^{-c} = Cs^{-r}$. This proves (21).

All derivatives of h are bounded on K, and $|h(A)| \leq \|AP\|^{d-1}s(A)$. In differentiating h^{2k} a total of a times, at least $2k - a$ factors remain undifferentiated. Hence

$$\|D^a(h^{2k})(A)\| \leq C_a s(A)^{2k-a}, \qquad 0 \leq a \leq 2k-1.$$

The product rule therefore yields

$$\|D^r(h^{2k}Y)(A)\| \leq C_r s(A)^{2k-r}, \qquad 0 \leq r \leq 2k-1. \tag{22}$$

Every one of these derivatives tends to zero at a singular probe.

Lemma 17 now shows that each weighted numerator is C^{2k-1} across its singular set. Summing and dividing by the positive smooth denominator Z preserves this regularity. □

A sharp example. Take $d = 2$, $n = 3$, $G = \mathrm{O}(2)$, scalar output with trivial representation, and the three coordinate probes P_{12}, P_{13}, P_{23}. They cover the full-rank domain. Let $f(B) = B_{13}$, a smooth linear backbone, and set

$$A(t) = \begin{pmatrix} t & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

This matrix has full row rank for every t. For $t \neq 0$, the 12 probe has determinant t, polar factor $\operatorname{diag}(\operatorname{sign} t, 1)$, and transported prediction $\operatorname{sign} t$. The 13 probe is singular, and the 23 probe has determinant -1, a constant polar factor, and transported prediction zero. Thus the entire averaged prediction is

$$F(A(t)) = \frac{t^{2k} \operatorname{sign} t}{1 + t^{2k}}, \qquad F(A(0)) = 0.$$

Its $2k$-th one-sided derivatives are $(2k)!$ and $-(2k)!$, so it is not C^{2k}. This proves sharpness even with a C^∞ backbone.

Sharpness for $\mathrm{SO}(d)$. The orientation correction does not improve the regularity guarantee. For any $d \geq 2$ and fixed $J \in \mathrm{O}(d)$ with $\det J = -1$, take $n = d + 1$ and let P_i omit column i. These coordinate probes cover the full-rank domain. For $s \in \{1, -1\}$, set

$$A_s(t) = [\, I_d \;\; te_1 + se_2 \,].$$

Along this path $D_1 := \det(A_s(t)P_1) = (-1)^{d-1}t$, $|D_2| = D_{d+1} = 1$, and all other determinants vanish. Hence

$$F(A_s(t)) = \frac{t^{2k} Y_1(A_s(t)) + Y_2(A_s(t)) + Y_{d+1}(A_s(t))}{t^{2k} + 2},$$

where the last two branches are smooth for a smooth backbone.

The only nonconstant block of the polar factor of $A_s(t)P_1$ is that of $M_s(t) = \left(\begin{smallmatrix} 0 & t \\ 1 & s \end{smallmatrix}\right)$, from $\operatorname{span}(e_1, e_d)$ to $\operatorname{span}(e_1, e_2)$. For $t \neq 0$ it is

$$\operatorname{polar} M_s(t) = \frac{1}{\sqrt{1 + (1 + |t|)^2}} \begin{pmatrix} -s \operatorname{sign} t & t + \operatorname{sign} t \\ 1 + |t| & s \end{pmatrix}.$$

Thus the full polar factor has analytic extensions from each side of zero. Its limits $Q_\pm$ agree on $u_s^\perp$ and satisfy $Q_+ = Q_-(I - 2u_s u_s^\top)$, where $u_s = (e_1 - se_d)/\sqrt{2}$ spans $\ker(A_s(0)P_1)$. After the fixed correction, the frame limits $g_\pm$ coincide exactly when $J = I - 2u_s u_s^\top$, for either parity of d. Since $u_1 \perp u_{-1}$, choose s for which they differ.

The matrices $B_\pm = g_\pm^\top A_s(0)$ are distinct because $A_s(0)$ has full rank, and have equal Frobenius norm. With trivial scalar representation and the linear backbone $f(B) = \langle B_+, B \rangle_F$, the one-sided limits of Y_1 therefore differ by

$$\|B_+\|_F^2 - \langle B_+, B_- \rangle_F = \tfrac{1}{2}\|B_+ - B_-\|_F^2 > 0.$$

The $2k$-th one-sided derivatives of $F(A_s(t))$ differ by $(2k)!/2$ times this quantity. Thus F need not be C^{2k} for $\mathrm{SO}(d)$ either.

Smooth cutoffs. The sharpness examples concern determinant-power weights. A different weight rule can give a smooth average for every smooth backbone. Set $a_j(A) = \det(AP_j)^2 / \sum_\ell \det(AP_\ell)^2$ and choose $0 < \tau < 1/m$. Let $\chi \colon \mathbb{R} \to [0, \infty)$ be smooth, zero on $(-\infty, \tau]$, and positive on (τ, ∞). Then

$$\widetilde{w}_j(A) = \frac{\chi(a_j(A))}{\sum_\ell \chi(a_\ell(A))}$$

is defined everywhere on the full-rank domain. Some a_j is at least $1/m$. Near a failed probe, its weight vanishes identically, so its weighted contribution extends smoothly by zero. The resulting average is C^∞ for a C^∞ backbone. Relative cutoffs also occur in Dym et al. (2024); Pozdnyakov and Ceriotti (2023). They can suppress valid but low-weight branches. Table 11 and Appendix B.12 measure costs for the relative cutoff of (38). The cubic cutoff used in the first two studies is only C^1 at its transition points.

A.7 The jump at a hard-selection boundary

Suppose $A(t)$ crosses at $t = 0$ between two regular chart scores, their difference has nonzero derivative, no third chart ties, and Y_i, Y_j extend continuously to $A(0)$. Label the sides so hard selection returns $Y_i(A(t))$ for $t < 0$ and $Y_j(A(t))$ for $t > 0$. Then

$$\lim_{t\uparrow 0} F_{\mathrm{hard}}(A(t)) - \lim_{t\downarrow 0} F_{\mathrm{hard}}(A(t)) = Y_i(A(0)) - Y_j(A(0)).$$

This is (3). The tie convention affects only the value at $t = 0$, not the one-sided limits. With a tie rule selecting either of the two maximizers, the prediction is continuous at the crossing if and only if their transported predictions agree. This selection tie occurs where both probes remain invertible.

A.8 Four deterministic charts but two weighted frames

We prove (7) on $\mathrm{St}^{\mathrm{row}}_{3,4}$. Whitening transfers both quantities to $\mathcal{M}^{\mathrm{fr}}_{3,4}$. For a row frame U with rows u_1, u_2, u_3, define the signed cofactor normal by

$$q_j(U) = (-1)^{j-1} \det U_{\widehat{j}}, \qquad j = 1, \dots, 4,$$

where $U_{\widehat{j}}$ omits column j. Laplace expansion and Cauchy–Binet give

$$Uq(U)^\top = 0, \qquad \|q(U)\|_2 = 1, \qquad q(RU) = \det(R) q(U).$$

These signs make the ordered basis $(q(U), u_1, u_2, u_3)$ positively oriented, since expansion of its determinant along its first row gives $\sum_j \det(U_{\widehat{j}})^2 = 1$.

Orthogonal complementation identifies $\mathrm{Gr}_3(\mathbb{R}^4)$ with $\mathbb{R}\mathrm{P}^3$. Its four coordinate-minor charts give genus at most four. If $a \in H^1(\mathbb{R}\mathrm{P}^3; \mathbb{F}_2)$ is the generator, then $a^3 \neq 0$ and $q^* a = 0$ for the frame bundle. Schwarz's cup-length bound therefore gives genus four.

To construct two weighted frames, identify $\mathbb{R}^4$ with the quaternions $\mathbb{H}$ in the oriented basis $(1, i, j, k)$. For a unit quaternion q, let $T(q)$ have rows given by the coordinate vectors of iq, jq, kq. Right multiplication by q is an orientation-preserving isometry, so (q, iq, jq, kq) is a positive orthonormal basis. Consequently U and $T(q(U))$ are equally oriented orthonormal bases of $q(U)^\perp$, and

$$c(U) = UT(q(U))^\top \in \mathrm{SO}(3).$$

Since $T(-q) = -T(q)$,

$$c(RU) = \det(R) R\, c(U).$$

It follows that

$$\mu_U = \tfrac{1}{2}\delta_{c(U)} + \tfrac{1}{2}\delta_{-c(U)} \tag{23}$$

is a continuous O(3)-equivariant measure. A reflection exchanges the two transformed atoms. The two maps c and $-c$ are continuous, but neither is an O(3)-equivariant gauge. The weights remain constant. Allowing reflections to exchange the two choices makes their measure equivariant.

Support one would give a continuous equivariant gauge and therefore a global section, contradicting genus four. Thus $\mathrm{wfr}_{\mathrm{O}(3)}(\mathcal{M}^{\mathrm{fr}}_{3,4}) = 2$. For SO(3), the map c itself is a global gauge. This example isolates the obstruction caused by reflections and shows why deterministic chart lower bounds do not automatically lower-bound weighted support.

A.9 The Gram quotient and equivariant vector outputs

Let

$$\mathcal{S}^+_{n,d} = \{\Sigma \in \mathbb{R}^{n\times n} : \Sigma = \Sigma^\top,\ \Sigma \succeq 0,\ \operatorname{rank}\Sigma = d\}.$$

Theorem 18 (Gram quotient and vector outputs)**.** *The Gram map induces a homeomorphism* $\mathcal{M}^{\mathrm{fr}}_{d,n}/\mathrm{O}(d) \cong \mathcal{S}^+_{n,d}$*. Every continuous* $\mathrm{O}(d)$*-equivariant map* $F\colon \mathcal{M}^{\mathrm{fr}}_{d,n} \to \mathbb{R}^{d\times m}$ *has the form* $F(A) = AC(A^\top A)$ *for a continuous* $C\colon \mathcal{S}^+_{n,d} \to \mathbb{R}^{n\times m}$*.*

Proof. An eigendecomposition $\Sigma = V\Lambda V^\top$, with V column-orthonormal and $\Lambda \succ 0$ of size d, gives the preimage $A = \Lambda^{1/2}V^\top$. Thus the induced Gram map is onto.

If $A^\top A = B^\top B$, then $\|Ac\|_2 = \|Bc\|_2$ for all $c \in \mathbb{R}^n$. Consequently $Ac \mapsto Bc$ is a well-defined isometry of $\mathbb{R}^d$, and $B = RA$ for a unique $R \in \mathrm{O}(d)$. This proves injectivity on the quotient.

The induced continuous bijection is proper. Over a compact subset of $\mathcal{S}^+_{n,d}$, the identity $\|A\|_F^2 = \operatorname{tr}(A^\top A)$ bounds the representatives. Any convergent sequence of representatives over that subset has a limit whose Gram matrix still has rank d, so the limit remains in the full-rank domain. Thus the inverse image before taking the quotient is compact, as is its quotient. A proper continuous bijection between these locally compact Hausdorff spaces is a homeomorphism.

Set $\widetilde{C}(A) = A^+F(A)$. The pseudoinverse is continuous at fixed rank and $(RA)^+ = A^+R^\top$, so $\widetilde{C}(RA) = \widetilde{C}(A)$. It descends to a continuous map C on the Gram quotient. Finally, $AA^+ = I_d$ gives $AC(A^\top A) = F(A)$. □

A.10 Inputs of fixed lower rank

Data of a fixed rank $r < d$ reduce to the full-rank problem in dimension r by restricting to their span. Let $\mathcal{M}_r = \{A \in \mathbb{R}^{d\times n} : \operatorname{rank} A = r\}$ with $1 \le r < d$ and $n > r$. A *partial frame* of $A \in \mathcal{M}_r$ is a matrix $Q \in \mathbb{R}^{d\times r}$ with $Q^\top Q = I_r$ whose columns span the column space of A. The group acts by $Q \mapsto RQ$. The reduced input $Q^\top A \in \mathbb{R}^{r\times n}$ has rank r and the same nonzero singular values as A. For a backbone $f_r\colon \mathcal{M}^{\mathrm{fr}}_{r,n} \to (\mathbb{R}^r)^n$ on reduced inputs, the transported prediction $Y_Q(A) = Q\, f_r(Q^\top A)$ satisfies $Y_{RQ}(RA) = RY_Q(A)$ for every $R \in \mathrm{O}(d)$, and its vectors lie in the column space of A, as the stabilizer of A requires for $\mathrm{O}(d)$-equivariant vector outputs. For $\mathrm{SO}(d)$ at rank $d-1$ the stabilizer is trivial and equivariant outputs may leave the span, as $a_1 \times a_2$ does on rank-two inputs in $\mathbb{R}^3$. This construction does not represent such outputs, although part (2) of the proposition below holds for $\mathrm{SO}(d)$ as well. Partial frames are not elements of G. The stabilizer of A fixes them, so they also cover inputs of rank at most $d-2$, whose stabilizers are infinite.

For a probe $P \in \mathbb{R}^{n\times r}$ with orthonormal columns, set $s_P(A) = \det(P^\top A^\top AP) \ge 0$ and, where $s_P(A) > 0$, $Q_P(A) = AP\,(P^\top A^\top AP)^{-1/2}$. Writing $A = UB$ with U an orthonormal basis of the column space and $B = U^\top A$, one has $AP = U(BP)$, so a bank of width-r probes covers $\mathcal{M}_r$ exactly when it covers $\mathcal{M}^{\mathrm{fr}}_{r,n}$. The width-$r$ Wronskian bank of Theorem 6 therefore covers $\mathcal{M}_r$ with $r(n-r)+1$ probes.

Proposition 19 (Fixed-rank inputs)**.** *Let* $1 \le r < d$ *and* $n > r$*, let* k *be a positive integer, and let* $P_1, \dots, P_m$ *be width-*r *probes covering* $\mathcal{M}^{\mathrm{fr}}_{r,n}$*. Define*

$$F_r(A) = \frac{\sum_j s_{P_j}(A)^k\, Y_{Q_{P_j}(A)}(A)}{\sum_j s_{P_j}(A)^k} \qquad (A \in \mathcal{M}_r),$$

where terms with $s_{P_j}(A) = 0$ *are omitted.*

1. *For continuous f_r, the prediction F_r is continuous and $\mathrm{O}(d)$-equivariant on $\mathcal{M}_r$, and if f_r is C^{2k-1}, so is F_r.*

2. *Consider continuous $\mathrm{O}(d)$-equivariant assignments, to each $A \in \mathcal{M}_r$, of a finitely supported probability measure on the partial frames of A. The least worst-case support among them equals $\mathrm{wfr}_{\mathrm{O}(r)}(\mathcal{M}^{\mathrm{fr}}_{r,n})$. It therefore satisfies the bounds of Theorem 1 with d replaced by r when $r \geq 2$, and equals 2 when $r = 1$. The same holds for $\mathrm{SO}(d)$-equivariant assignments.*

Proof. (1) Equivariance holds term by term, since $Q_P(RA) = RQ_P(A)$ and s_P is invariant. Near $A_0 \in \mathcal{M}_r$, choose a smooth local orthonormal basis $U(A)$ of the column space. The column space depends smoothly on A within $\mathcal{M}_r$. With $C = U^\top AP$ one has $Q_P = U\ \mathrm{polar}(C)$, $s_P = \det(C)^2$, and $\sigma_{\min}(C) = \sigma_{\min}(AP)$. The reduced inputs $Q_P(A)^\top A$ have smallest singular value $\sigma_r(A)$, bounded below near A_0, and lie in a compact set. The proofs of Theorem 2 and of (21), applied to C in place of AP, then give continuity, and C^{2k-1} regularity for a C^{2k-1} backbone.

(2) The map $[R, B] \mapsto R\,[B; 0]$ identifies $\mathcal{M}_r$ equivariantly with $\mathrm{O}(d) \times_H \mathcal{M}^{\mathrm{fr}}_{r,n}$, where $H = \mathrm{O}(r) \times \mathrm{O}(d-r)$ acts on B through its first factor. The partial frames of $R[B; 0]$ are the matrices $R[k; 0]$ with $k \in \mathrm{O}(r)$. An equivariant assignment restricted to the slice $\{[B; 0]\}$ is an $\mathrm{O}(r)$-equivariant weighted frame on $\mathcal{M}^{\mathrm{fr}}_{r,n}$, because $\mathrm{O}(d-r)$ fixes both the slice and the frames $[k; 0]$. Conversely, an $\mathrm{O}(r)$ weighted frame ν on $\mathcal{M}^{\mathrm{fr}}_{r,n}$ extends by $\mu_{R[B;0]} = R_*\nu_B$. This is well defined. If $R[B; 0] = R'[B'; 0]$, then $R'^{-1}R = (h_1, h_2) \in H$ with $h_1 B = B'$, and $(h_1)_*\nu_B = \nu_{B'}$. It is continuous because $\mathrm{O}(d) \to \mathrm{O}(d)/H$ has local sections. Both directions preserve the support size. For $r \geq 2$ the bounds are Theorem 1 applied to (r, n). For $r = 1$, a single atom would be a continuous odd function from the connected space $\mathbb{R}^n \setminus 0$ to $\{\pm 1\}$, which does not exist, while $\frac{1}{2}\delta_{+1} + \frac{1}{2}\delta_{-1}$ has support 2. For $\mathrm{SO}(d)$, replace H by $S(\mathrm{O}(r) \times \mathrm{O}(d-r))$, which projects onto $\mathrm{O}(r)$ because $d - r \geq 1$. □

Extending continuously across ranks is a separate problem, since the determinant weights of the full-rank construction vanish at rank loss.

B Experimental methods and results

All three studies use labeled particles in $\mathbb{R}^3$ and predict forces directly. Table 2 distinguishes their atlas implementations. The moving-boundary study reuses trajectory checkpoints and adds consistency-trained models. Appendices B.1–B.3 specify the particle-count protocol, and Appendices B.4–B.5 specify the other two studies. All reported intervals are nominally 95%.

B.1 Particle-count study: synthetic force task

Inputs and targets. An input is a centered, full-row-rank configuration $X = (x_1, \ldots, x_N) \in \mathbb{R}^{3\times N}$. Positions are the only supplied numerical features. Each model also has trainable labeled-slot embeddings. The target is a conservative O(3)-equivariant force with zero total force. All predictions are projected to zero total force.

Regular-core generator. Keyed pseudorandom draws produce a centered Gaussian matrix $G \in \mathbb{R}^{3\times N}$. We whiten its rows to obtain $W = (GG^\top)^{-1/2}G$. The three singular scales are proportional to $(1, q_2, q_3)$ and normalized to unit Euclidean norm. For family identifier f, the stratum $h = f \bmod 4$ has endpoints (L_h, U_h). Independent keyed $u_3, u_2, u_r \sim \mathrm{Unif}(0, 1)$ determine

$$q_3 = L_h + (U_h - L_h)u_3, \qquad q_2 = q_3 + (1 - q_3)u_2, \qquad r = 0.9 + 0.2u_r.$$

Table 2: Atlas implementations in the three studies. Write $A = X_c H$ for fixed orthonormal contrasts of centered positions and $U = (AA^\top)^{-1/2}A$. The first two studies also include Quaternionic-2 and Dym-24 controls, specified in Appendix B.4.

Study	Frames and weights	Backbone and training
Trajectories, $N = 5$	Four coordinate minors; polar(UP_j); squared-minor scores with cubic cutoff for soft averaging, raw-score maximization for hard routing.	Width-96, three-block Transformer; 30,000 updates; separate soft and hard training.
Moving boundary, $N = 5$	Same frames and cutoff; positive score biases move routing boundaries; hard and soft evaluation share cached transported branches.	Reuses trajectory checkpoints; adds eight consistency-trained models after a three-seed coefficient screen.
Particle counts	$3N - 11$ Wronskian probes; polar(AP_j); normalized $\det(UP_j)^2$ weights without a cubic cutoff, subject to numerical guards.	Atlas models use a width-128, four-block Transformer; 40,000 updates and five paired seeds at each N. Gram and spherical-harmonic baselines use separate architectures.

The four intervals are $[0.15, 0.25]$, $[0.25, 0.40]$, $[0.40, 0.70]$, and $[0.70, 1.00]$. With $a = (1, q_2, q_3)/\|(1, q_2, q_3)\|_2$, the input is $X = \sqrt{N}\, r\, Q\, \mathrm{diag}(a) W$, where $Q \in \mathrm{O}(3)$ has Haar distribution conditional on determinant $(-1)^f$. For a shared family identifier, stratum, quantiles, radius, and base orientation are reused across N, while the centered Gaussian is independently keyed by N.

Force oracle. With cyclic indices, write $d_i^{(a)} = x_{i+a} - x_i$. The fixed potential is

$$V_N(X) = \sum_{i=1}^{N} \sum_{a \in \{1,2\}} \alpha_{i,a} \exp\left(-\frac{\|d_i^{(a)}\|^2}{2\ell_a^2} \right) + \sum_{i=1}^{N} \beta_i \exp\left(-\frac{\|d_i^{(1)}\|^2 + \|d_i^{(2)}\|^2}{2\ell_3^2} \right) \tanh\left(\frac{\langle d_i^{(1)}, d_i^{(2)} \rangle}{\ell_3^2} \right), \tag{24}$$

where $(\ell_1, \ell_2, \ell_3) = (1, 1.4, 1.2)$, $\alpha_{i,a} = 0.85 + 0.30 u_{i,a}$, and $\beta_i = -0.25 + 0.50 v_i$, with fixed keyed uniform draws $u_{i,a}, v_i \in (0, 1)$. Keys omit N, so shared labels and offsets reuse coefficients across sizes. The target is $F_N = -\nabla_X V_N$. Cyclic coefficients make labels consequential, while distances and inner products ensure orthogonal symmetry.

Task validation. A prespecified diagnostic on 20,000 states per N checked force scale, local sensitivity, rank strata, and radial reference predictors. Across $N = 5, 8, 12, 16, 24, 32$, a six-scale radial reference using labeled cyclic edge classes had NRMSE 0.136–0.154. Without those classes, error increased from 0.136 to 0.517. All six sizes passed the task checks, and no model was trained at $N = 32$. Different values of N define different prediction tasks, so comparisons of error and cost must be made within each size.

B.2 Numerical realization and validation

For centered X_c, a fixed Helmert contrast gives $A = X_c H \in \mathbb{R}^{3\times(N-1)}$. Its one-indexed convention is

$$H_{rk} = \begin{cases} [k(k+1)]^{-1/2}, & r \le k, \\ -k[k(k+1)]^{-1/2}, & r = k+1, \\ 0, & r > k+1, \end{cases} \qquad k = 1, \dots, N-1. \tag{25}$$

Since $H^\top H = I$ and $HH^\top = I - \frac{1}{N}\mathbf{1}\mathbf{1}^\top$, we have $AA^\top = X_c X_c^\top$ and $X_c = AH^\top$. Thus A has the same singular values as X_c, and the constructions apply with $d = 3$ and $n = N-1$. The backbone uses labeled particle coordinates while probes act on the contrast columns. The ideal scores, weights, frames, and prediction are

$$\begin{aligned} U &= (AA^\top)^{-1/2}A, & a_j &= \det(UP_j)^2, & w_j &= \frac{a_j}{\sum_\ell a_\ell}, \\ AP_j &= Q_j S_j, & \Phi_j(X) &= Q_j h(Q_j^\top X_c), & \widehat{F}_{\text{ideal}} &= \sum_j w_j \Phi_j. \end{aligned} \tag{26}$$

The polar factor uses AP_j, while its score uses U. Whitening multiplies every squared determinant by the same factor $\det(AA^\top)^{-1}$, so the normalized weights agree with (5) at $k = 1$. Thus (26) is the construction of Theorem 2 before numerical guards.

At $N = 5$, the Wronskian bank and the four coordinate-minor probes both have size four, but they are different banks. The condition-number bound for coordinate hard selection does not transfer to Wronskian selection. Indeed, for coordinate probes E_I, Cauchy–Binet gives

$$\sum_{|I|=3} \det(UE_I)^2 = \det(UU^\top) = 1.$$

The largest minor therefore has determinant magnitude at least $\binom{N-1}{3}^{-1/2}$. Every singular value of UE_I is at most one, so the selected minor obeys

$$\kappa(UE_{I_*}) \le \sqrt{\binom{N-1}{3}}.$$

This equals two when $N = 5$.

Empirical conditioning of the Wronskian bank. To measure coverage separately from learned predictions, we performed a posthoc diagnostic on 20,000 states per N, regenerated from the recorded final-test family identifiers with the generator in Appendix B.1 and the serialized binary64 probes. For each whitened input U, define

$$c(U) = \max_j \sigma_{\min}(UP_j), \qquad Z(U) = \sum_j \det(UP_j)^2, \qquad j_* = \arg\max_j \det(UP_j)^2.$$

The probe attaining $c(U)$ need not be the determinant-maximizing j_*. Table 3 summarizes both statistics and the denominator Z. Median c and Z decrease with N. These sampled values do not bound coverage over the Stiefel manifold or measure every low-weight branch. Moreover, the predictor uses AP_j, for which $\kappa(AP_j) \le \kappa(A)\kappa(UP_j)$ on an invertible probe.

Table 3: Posthoc geometric conditioning, 20,000 regenerated states per particle count. Quantiles use linear interpolation. c is the largest smallest singular value among probes, Z the score denominator, and κ_* the condition number of the determinant-maximizing whitened probe. These are empirical values, not worst-case bounds.

N	Probes	$\min c$	$q_{.01}(c)$	Median c	$q_{.01}(Z)$	Median Z	$q_{.99}(\kappa_*)$
5	4	0.0724	0.1567	0.6660	0.03907	0.9984	6.382
8	13	0.0581	0.1675	0.4353	0.02611	0.2974	6.813
12	25	0.0694	0.1690	0.3529	0.01576	0.1215	6.733
16	37	0.0984	0.1661	0.3114	0.01186	0.06642	6.293
24	61	0.1065	0.1574	0.2653	0.00699	0.02959	6.090

Floating-point guards. In binary64, a frame is used only if its singular values are finite, the ratio of its smallest to largest singular value exceeds $64\varepsilon_{64} \approx 1.42 \times 10^{-14}$, and its polar factor passes orthogonality, reconstruction, symmetry and positive-semidefiniteness checks at tolerance 10^{-12}. Frames that fail are dropped and the remaining weights are renormalized, so the continuity theorem applies to the ideal map and the guards are part of its numerical implementation.

On all full-rank samples in the pretraining validation cohort, denominators resolved, the numerical checks met tolerance, and guard-omitted mass was zero. During training, frames and weights were computed in binary64 without gradient tracking and cast to the float32 predictor. Active weights were checked for normalization at tolerance 10^{-6}. The training records do not retain an aggregate diagnostic of mass omitted by the guards. Since the experiments learn forces directly, gradients of frames and weights with respect to positions are not used in training. Computing forces as gradients of an averaged energy would require differentiating through both.

B.3 Models, training, and statistical analysis

Table 4: Methods in the particle-count study. The three trained frame methods share the particle Transformer specified below. Parameter counts vary with N through the slot embeddings. Figure labels full-soft, hard-trained, and native denote Wronskian averaging, Wronskian hard selection, and the spherical-harmonic network, respectively. In this appendix, full-soft also denotes Wronskian averaging, and hard routing denotes evaluation with hard selection, also for models trained otherwise.

Method	Frames evaluated per input	Parameters
Wronskian averaging	All resolved Wronskian frames from the $3N - 11$ probes	0.532–0.534M
Wronskian hard selection	The Wronskian frame of largest weight	0.532–0.534M
Coordinate hard selection	The coordinate frame of largest minor score	0.532–0.534M
Gram model	None; invariant coefficients carried by X	0.733–0.734M
Spherical-harmonic network ($N = 12$)	None; equivariant message passing	0.502M
Evaluation-only variant of Wronskian-averaging checkpoints		
Posthoc hard selection	The Wronskian frame of largest weight	—

Wronskian averaging evaluates resolved frames in chunks of eight. All frame methods score the complete bank before selection. Posthoc hard selection changes only the evaluation of an averaging checkpoint. The following architectures are specific to the particle-count study.

Particle Transformer. A linear map $\mathbb{R}^3 \to \mathbb{R}^{128}$ of each position is added to a learned 128-dimensional slot embedding. Each of four blocks applies LayerNorm before four-head self-attention and before a $128 \to 256 \to 128$ feedforward map with GELU, with a residual connection around each sublayer. Dropout is zero. A final LayerNorm and linear $128 \to 3$ map produce each force, followed by projection to zero total force. All chart methods share this backbone architecture and paired initialization and data-order seeds.

Gram coefficient model. Write $G = X^\top X$. A learned 32-dimensional slot embedding concatenated with G_{ii} initializes a 128-dimensional node state h_i by a linear map. Each of four blocks computes

$$m_i = N^{-1/2} \sum_j M(h_i, h_j, G_{ij}, G_{ii}, G_{jj}), \qquad h_i \leftarrow h_i + U(\mathrm{LayerNorm}(m_i)).$$

Here M is a $259 \to 256 \to 128$ MLP and U a $128 \to 256 \to 128$ MLP, each with SiLU between its linear layers. Blocks have separate parameters. A shared $257 \to 256 \to 1$ SiLU MLP applied to (h_i, h_j, G_{ij}) gives C_{ij}. The output is XC, projected to zero total force. This model learns invariant coefficients using all pairwise inner products. It has no spatial frame selection.

Native equivariant model. The $N = 12$ model uses three complete-graph message-passing blocks with hidden representation $24 \times 0e + 18 \times 1o + 12 \times 2e$ in e3nn notation. Slot embeddings initialize the even scalars. The other channels start at zero. Messages use spherical harmonics of degrees $0, 1, 2$ and an eight-feature radial basis without a distance cutoff, scaled by $1/(1 + r^2)$. Blocks have irrep-wise RMS normalization, equivariant linear self-connections, SiLU scalars, and sigmoid gates on vector and tensor channels. The readout is one $1o$ vector per node, projected to zero total force.

Training and checkpoint selection. At each N, the fixed calibration, training, development, checkpoint-validation, and final-test cohorts contain 20,000, 100,000, 10,000, 10,000, and 20,000 states, respectively. Calibration, training, and development geometries were reused from code development. Model weights, optimizer states, seeds, checkpoint decisions, and pilot metrics were not. Each run used one NVIDIA A40.

Both this study and the trajectory study use AdamW, effective batch size 512 with microbatches of 128, weight decay 10^{-5}, gradient clipping at 1, a 2,000-update warmup, and cosine decay to zero. Training is float32, with mixed precision and TF32 disabled and detached binary64 frame preprocessing. Table 5 gives the study-specific choices.

The 100 main runs (four methods, five sizes, five seeds) and five spherical-harmonic runs at $N = 12$ all selected update 40,000. Relative convergence beyond this budget is therefore unknown. Initialization and minibatch-order streams are keyed by both N and seed index. Within a size, the three Transformer methods share initialization and data order. Other architectures have distinct initialization keys but the same data order. Across sizes, both streams change. This is the dependence structure used below.

Pooled NRMSE. For method m and particle count N, the reported estimate is

$$\widehat{\mathrm{NRMSE}}_{m,N} = \left(\frac{\sum_{s=1}^{5} \sum_i \|\widehat{F}_{m,N,s,i} - F_{N,i}\|_F^2}{5 \sum_i \|F_{N,i}\|_F^2} \right)^{1/2}. \tag{27}$$

Squared errors are pooled over model seeds and final-test geometries before taking the square root.

Table 5: Training and checkpoint selection. Both studies run to their fixed update budget without early stopping. Checkpoint validation is separate from the development split in the particle-count study.

Setting	Particle-count study	Trajectory study
Updates	40,000	30,000
Learning rate	3×10^{-4} for every method	3×10^{-4}; radial e3nn uses .01, selected from $\{.001, .003, .01\}$.
Validation	Development every 500 updates; 16 candidate checkpoints at $2500, 5000, \ldots, 40000$.	Every 1,000 updates.
Selection	Minimum finite NRMSE on the entire checkpoint-validation split; earlier checkpoint wins exact ties.	Replace the incumbent when normalized validation NRMSE improves by more than 10^{-6}.

Paired bootstrap. The analysis uses 20,000 PCG64 replicates with the recorded seed and linear quantiles. At each N, five model seed indices were multinomial-resampled, independently across N and with pairing preserved across methods. Within each of the four rank strata, 5,000 geometry families were resampled. The geometry count draw was shared across methods and coupled values of N. Separately resampling model and geometry factors follows the crossed-factor structure of the pigeonhole bootstrap (Owen, 2007). Within-size method pairing and cross-size geometry coupling preserve the shared factors. Every replicate recomputes (27) from resampled squared-error and target-energy sums before forming contrasts. Intervals condition on the realized training dataset, oracle, architectures, and protocol. With only five independent training realizations per method and N, they are approximate uncertainty summaries for nonlinear NRMSE contrasts, not finite-sample coverage guarantees.

For a prespecified family of contrasts $\theta_1, \ldots, \theta_J$, let $\hat{\theta}_j$ be the reported estimate and θ^*_{bj} its value in bootstrap replicate b. Define

$$s_j = \mathrm{sd}_{b=1}^{20000}(\theta^*_{bj}),$$
$$q = Q_{0.95}\left(\max_{1\le j\le J}\left|\frac{\theta^*_{bj} - \hat{\theta}_j}{s_j}\right|\right), \qquad I_j = [\hat{\theta}_j - q s_j, \hat{\theta}_j + q s_j]. \tag{28}$$

The standard deviation uses divisor 19999; Q_p is the linearly interpolated empirical quantile. The primary one-sided bounds use the following percentile-style lower envelope:

$$q_- = Q_{0.95}\left(\max_j \frac{\hat{\theta}_j - \theta^*_{bj}}{s_j}\right), \qquad L_j = \hat{\theta}_j - q_- s_j. \tag{29}$$

This chooses a common standardized offset so that approximately 95% of the joint bootstrap vectors lie coordinatewise above the lower endpoints. For one component, $L = Q_{0.05}(\theta^*)$; for several, the maximum accounts for simultaneous downward excursions. This calibrates the lower tail of the bootstrap distribution, unlike a basic error-inversion bound, which uses the opposite centered tail. Zero-variance components have point intervals and are omitted from the maximum. Each five-size contrast family has its own critical value. The 13 posthoc top-k contrasts form a separate family. Individual method intervals and the single spherical-harmonic contrast use the 2.5th and 97.5th percentiles directly. The primary slope is equal-weight OLS of R_N against $\log N$, refitted in every replicate, with the 5th percentile as its one-sided lower bound.

For the five R_N, the critical values are $q = 2.44219$ and $q_- = 2.13354$.

B.4 Original seven-method trajectory study

Force task and initial conditions. This study uses a different potential from (24). For unit particle masses, its energy is

$$V(X) = \frac{1}{4}\sum_{i<j} a_{ij}(\|x_i - x_j\|^2 - \ell_{ij}^2)^2 + 0.05 \sum_{i<j<k} (\langle x_j - x_i, x_k - x_i\rangle - c_{ijk})^2, \tag{30}$$

and the force is $-\nabla V$. The reference configuration is

$$X_0 = \begin{pmatrix} .92 & -.48 & -.73 & .07 & .22 \\ .10 & .95 & -.35 & -.80 & .10 \\ 0 & .15 & .85 & -.75 & -.25 \end{pmatrix}.$$

Here ℓ_{ij}^2 and c_{ijk} are its squared distances and indicated inner products. The ten a_{ij}, in lexicographic order of pairs, are fixed NumPy PCG64 uniform draws on $[0.6, 1.2]$.

For independent centered standard Gaussian 3×5 matrices Z_x, Z_v, $a \sim \mathrm{Unif}(0.9, 1.1)$, and Haar $Q \in \mathrm{O}(3)$, set

$$X(0) = Q(aX_0 + \sigma_x Z_x), \qquad \dot{X}(0) = Q\sigma_v Z_v.$$

The three physical regimes use $(\sigma_x^{\min}, \sigma_x^{\max}, \sigma_v)$ equal to $(.08, .30, .15)$, $(.12, .38, .20)$, and $(.15, .45, .25)$, with σ_x uniform on its interval. Initial states must have $\kappa(XH) \le 3$ and minimum pair distance at least 0.4. Velocity Verlet uses step 0.000625 for horizon 10, storing the 100 noninitial states at intervals of 0.1. A trajectory is accepted only if all integration states are finite, full rank, have $\kappa(XH) \le 10$, and minimum pair distance at least 0.25. Numerical rank uses threshold $10^{-10}\max\{\sigma_{\max}, 1\}$. There are at most 50 proposals per trajectory slot. Each regime contributes 1,000 training, 100 validation, and 400 test trajectories. Training uses 32 fixed, approximately equally spaced stored states per trajectory.

Models and training. The shared chart backbone has width 96, three Transformer blocks, four heads, feedforward width 192, GELU, and zero dropout, with the same pre-normalization and position/slot embedding scheme described above. It has 225,699 parameters. The original Gram model instead flattens the ordered 5×5 Gram matrix, projects its 25 entries to width 96, applies four residual LayerNorm–MLP blocks with hidden width 384 and SiLU, and reads out 25 coefficients after a final LayerNorm. It returns $X_c C$, centered over output particles. The radial e3nn model uses the construction above with hidden representation $16 \times 0e + 12 \times 1o + 8 \times 2e$ and 223,628 parameters.

Soft and hard atlas both use polar factors of the four row-whitened coordinate minors UP_j. Soft atlas uses the cubic cutoff $\tau = 10^{-6}$ in (32) with zero bias. Hard atlas maximizes the raw score $\det(UP_j)^2$, before applying any cutoff, with lowest-index tie breaking. In contrast, the particle-count study uses polar factors of AP_j. Augmented raw trains the same backbone on independent Haar orthogonal transformations. Training and checkpoint selection follow Table 5. The learning-rate screen is specific to the radial e3nn baseline.

Crossings and rollouts. Crossings are strict exchanges of the two largest raw squared-minor scores, bracketed between integrator states. Bisection on the interpolated input locates the equality to tolerance 10^{-10}. A qualifying crossing has third-score gap at least 10^{-4} and a stable selected chart on each side over a 0.02 time window. Events lie in $[0.5, 8]$, events within 0.05 are deduplicated, and the first qualifying event is used. The routing margin is the largest score minus the second largest.

Table 6: Original trajectory study. NRMSE and completion are arithmetic means over model seeds. Calls are logical branch or backbone evaluations. The last column averages each seed's median final centered-position RMSD among completed rollouts, without Procrustes alignment.

Method	Seeds	Calls	Overall NRMSE	First-crossing NRMSE	Completion	Mean seed-median RMSD
Soft atlas	8	4	0.00676	0.0110	1.000	0.0143
Quaternionic-2	8	2	0.00817	0.0145	1.000	0.0173
Hard atlas	8	1	0.00747	0.0181	1.000	0.0169
Dym-24	3	24	0.00603	0.00844	1.000	0.0121
Augmented raw	3	1	0.00920	0.0128	1.000	0.0193
Gram	3	1	0.00153	0.00270	1.000	0.00193
e3nn radial MP	3	1	0.00388	0.00628	1.000	0.00548

Near-boundary states have margin at most 0.03, and far-boundary states have margin greater than 0.2.

Learned-force Verlet rollouts start 0.5 before the crossing and end 2 after it, with the same integration step. Final position error is centered RMSD without Procrustes alignment. A numerical failure is a nonfinite position, velocity, or force, or centered RMS radius exceeding ten times that of X_0. The composite event is the first such failure or sampled centered RMSD exceeding 0.25, sampled every 0.1 over the 2.5-unit rollout. Event incidence averages the seed–trajectory indicators. These crossed observations are not independent replicates.

Evaluation and uncertainty. The original $N = 5$ study evaluated fixed checkpoints on 1,200 trajectory-disjoint test trajectories, including 302 qualifying first chart crossings. Soft atlas, Quaternionic-2, and hard atlas each used eight independent training seeds. Dym-24, augmented raw, Gram, and e3nn radial message passing used three each. The primary soft–hard comparison used all eight paired training seeds. Comparisons with three-seed methods used the three corresponding hard-atlas seeds. A 10,000-replicate crossed bootstrap independently resampled paired seeds and trajectories within physical regime, keeping methods paired. Intervals are marginal, unadjusted for multiplicity, and use linearly interpolated 2.5th and 97.5th percentiles. For force-error reductions, each replicate computes NRMSE within each selected seed, averages over seeds, and forms the ratio of these means. The comparisons with three-seed methods have fewer independent training replicates.

Quaternionic-2 implements the explicit two-atom O(3) weighted frame on 3×4 matrices. Dym-24 is the robust weighted frame of Dym et al. (2024): for the columns a_i of the unwhitened reduced matrix A, each ordered pair $i \neq j$ gives a Gram–Schmidt frame and its reflection partner. Each partner has weight

$$\frac{1}{2}\frac{r_i}{\sum_\ell r_\ell}\frac{b_{ij}}{\sum_\ell b_{i\ell}}, \qquad r_i = \|a_i\|, \quad b_{ij} = \|a_i \times a_j\|,$$

after setting r_i to zero at or below $64\varepsilon_{64}\max\{1, \|A\|_F\}$ and b_{ij} to zero at or below $64\varepsilon_{64}\max\{1, \|A\|_F^2\}$. A pair with zero first-level weight is inactive. A positive first-level weight requires a positive area denominator. The twelve ordered pairs and their reflection partners give at most 24 atoms. This comparator uses numerical zero thresholds, not the atlas's cubic cutoff.

The soft–hard improvement is concentrated near selection boundaries. NRMSE falls by 34.42% (paired interval 26.88–40.59%) near the boundary, but increases by 2.64% (0.47–4.92%) far from it. The near-minus-far reduction is 37.05 percentage points (28.83–44.24). Final position error

falls by 15.03% (10.17–19.62%), using the mean across seeds of within-seed median RMSD. These comparisons do not separate continuity, averaging, and learned branch specialization.

Every method completes all rollouts. The sole composite event is a soft rollout whose sampled RMSD exceeds 0.25 at elapsed time 2.3, among $8 \times 302 = 2{,}416$ seed–trajectory pairs (incidence 1/2416). Lower position error therefore does not imply fewer adverse events. Gram and radial e3nn attain lower errors than soft atlas on this task (Table 6).

Numerical limitations. Full-support recombination of soft top-4 and Dym top-24 predictions retains nominal-bound flags, with maximum discrepancies 2.493×10^{-6} and 2.290×10^{-6}. The recombination code renormalizes float32 weights even at full support, while the reference uses their original values. Retained mass can also round above one, making the nominal omitted-mass bound negative.

B.5 Moving-boundary intervention

Paths and selection rules. Starting from a configuration $A_0 = X_{c,0}H$ in a separate cohort generated by (30) and the three regimes above, set

$$A_p(t) = A_0 R_{ij}(t), \qquad X_{c,p}(t) = A_p(t)H^\top, \tag{31}$$

where $R_{ij}(t) \in \mathrm{SO}(4)$ is a right Givens rotation. It acts on contrast columns, not on spatial coordinates, and preserves $AA^\top$. Choose its sign convention so the signed cofactor vector satisfies

$$q_i(t) = q_i(0)\cos t + q_j(0)\sin t, \qquad q_j(t) = -q_i(0)\sin t + q_j(0)\cos t.$$

The other two cofactors are fixed. Each accepted path comes from a different source trajectory. Candidates are ordered by a geometry-only hash before model predictions are computed.

The study uses 504 physical paths: three regimes, six unordered chart pairs, and 28 paths per regime–pair stratum. It evaluates achieved input separations 0.006, 0.002, 0.0006, and 0.0002, with 0.0006 declared primary. Seven biases $\delta \in \{-0.18, -0.12, -0.06, 0, 0.06, 0.12, 0.18\}$ define seven selection rules, also called routers. Each rule has a boundary at a different location along a path. We evaluate all seven rules at all seven locations, giving 49 combinations for each path and model seed.

For row-orthonormal $U \in \mathbb{R}^{3\times 4}$, define the signed cofactors and their squared scores by

$$q_j(U) = (-1)^{j-1} \det U_{\widehat{j}}, \qquad s_j = q_j(U)^2, \qquad \sum_j s_j = 1.$$

For cutoff $\tau = 10^{-6}$, let

$$c_\tau(s) = \begin{cases} 0, & s \le \tau, \\ 3z^2 - 2z^3, & \tau < s < 2\tau, \quad z = (s-\tau)/\tau, \\ 1, & s \ge 2\tau. \end{cases}$$

For a centered bias vector b, effective scores and soft weights are

$$e_j = s_j c_\tau(s_j), \qquad \alpha_j^{(b)} = \frac{e_j \exp(b_j)}{\sum_\ell e_\ell \exp(b_\ell)}. \tag{32}$$

For controlled chart pair (i, j), set $b_i = \delta$, $b_j = -\delta$, and all other entries to zero. Hard selection takes the lowest-index maximizer of $e_j \exp(b_j)$. Matched averaging (matched soft in the tables)

Table 7: Moving-boundary contrasts at separation 6×10^{-4}. Each row measures the switching router's response minus the average response of the six nonswitching routers on the same inputs. "Matched soft" is an evaluation rule, not a training condition. The marginal paired percentile intervals are unadjusted across the six rows.

Checkpoint training	Evaluation	Estimate	Marginal paired 95% interval
Soft atlas	Moved hard	0.1650	$[0.1553, 0.1763]$
Soft atlas	Matched soft	-1.53×10^{-7}	$[-2.95, -0.20] \times 10^{-7}$
Ordinary hard	Moved hard	0.03279	$[0.03091, 0.03475]$
Ordinary hard	Matched soft	3.00×10^{-7}	$[2.32, 3.80] \times 10^{-7}$
Hard + consistency	Moved hard	0.01603	$[0.01511, 0.01700]$
Hard + consistency	Matched soft	7.65×10^{-8}	$[4.10, 12.87] \times 10^{-8}$

combines the same cached transported predictions using $\alpha^{(b)}$. On the active plateau of the cutoff, the boundary solves

$$g_\delta(t) = \log q_i(t)^2 - \log q_j(t)^2 + 2\delta = 0. \tag{33}$$

The implementation scans 1,025 points in $[-0.4, 0.4]$ and locates roots by bracketed float64 rootfinding, to equality tolerance 2×10^{-12}. It retains paths only if all seven roots and their six nonswitching controls exist at every separation. Required bounds are $|g'_\delta| \geq 0.1$, biased third-score gap at least 0.01, minimum pair distance 0.25, oracle force norm at most 100, $\kappa(A) \leq 10$, and displacement from the base at most one third of its Frobenius norm.

Residual change and paired contrast. Let s index model seeds and p paths. For the boundary of selection rule δ and input separation Δ, let $t^\pm = t^\pm_{p\delta}(\Delta)$ denote the two path parameters on either side of that boundary with

$$\Delta = \|X_{c,p}(t^+) - X_{c,p}(t^-)\|_F.$$

Write $R_{sp,\delta'}(t) = \widehat{F}_{sp,\delta'}(X_p(t)) - F(X_p(t))$ for the residual evaluated with rule δ'. Its response at the boundary of rule δ is

$$J_{sp,\delta' \leftarrow \delta}(\Delta) = \frac{\|R_{sp,\delta'}(t^+) - R_{sp,\delta'}(t^-)\|_F}{F_{\mathrm{RMS,cal}}}, \qquad F_{\mathrm{RMS,cal}} = 4.6036908809. \tag{34}$$

This dimensionless change in prediction residual is normalized by force scale, without division by input separation. The contrast is

$$T_{sp\delta}(\Delta) = J_{sp,\delta \leftarrow \delta}(\Delta) - \frac{1}{6} \sum_{\delta' \neq \delta} J_{sp,\delta' \leftarrow \delta}(\Delta). \tag{35}$$

Its first term uses the rule that switches between the two inputs. The second averages the six rules that do not switch there. The comparison therefore holds both inputs and all transported predictions fixed.

We average the seven boundary contrasts within each path, the 28 paths within each of the 18 regime–pair strata, and then the strata and eight model seeds, with equal weights at each stage. The 10,000-replicate paired bootstrap independently resamples eight seeds and 28 paths within each regime–pair stratum, using linear percentile intervals. All selection rules, boundary locations, aggregation rules, and input separations remain together within their clusters.

For soft-trained models at the primary separation, the switching hard response is 0.16513 and direct transported-branch disagreement is 0.16519, a 0.035% relative mismatch using unrounded

values. Figure 2 shows the former approaching the latter as separation contracts, while nonswitching and matched-soft responses shrink. The tiny matched-soft contrast in Table 7 is the difference of two shrinking responses, rather than either response.

Consistency training. All branch outputs in this study include the zero-total-force projection $\Pi(B) = B - \frac{1}{5}(B\mathbf{1})\mathbf{1}^\top$. Write $\Phi_j(x) = \Pi(Q_j f(Q_j^\top X_c))$ for these projected transported predictions. With zero-bias soft weights, the training penalty is

$$\overline{\Phi}(x) = \sum_j \alpha_j(x)\Phi_j(x), \qquad L_{\text{cons}} = \frac{\sum_{x\in\mathcal{B}}\sum_j \alpha_j(x)\|\Phi_j(x) - \overline{\Phi}(x)\|_F^2}{|\mathcal{B}|\,\mathbb{E}_{\text{cal}}\|F\|_F^2}. \tag{36}$$

Here $\mathcal{B}$ is a training batch and $F_{\text{RMS,cal}} = (\mathbb{E}_{\text{cal}}\|F\|_F^2)^{1/2}$ is the same normalization as in (34). The validation screen uses a different quantity. Let $\mathcal{C}$ be the set of geometry-defined unbiased validation first crossings, with switching pair $i(x), j(x)$ at x. For each tuning seed s, it computes

$$D_s = \frac{1}{|\mathcal{C}|F_{\text{RMS,cal}}}\sum_{x\in\mathcal{C}}\|\Phi_{s,i(x)}(x) - \Phi_{s,j(x)}(x)\|_F.$$

Each qualifying crossing has equal weight. This is the norm of the two-branch disagreement at a crossing, averaged over crossings. It is not the four-branch weighted variance in the training penalty. A three-seed development screen selected coefficient $\lambda = 1$, the largest value tested, from $\{0, 0.01, 0.03, 0.1, 0.3, 1\}$ as follows. A coefficient is eligible if its three-seed mean validation NRMSE is at most 1.05 times the contemporaneous zero-control mean and each seed's NRMSE is at most 1.10 times its paired zero-control value. Among eligible coefficients, the screen minimizes the mean of D_s over the three tuning seeds. Coefficients within 1% of the minimum are tied, and the smallest wins. Final training uses eight paired seeds and inherits the trajectory optimizer and validation rule. The selected coefficient is the grid's upper endpoint. The resulting hard contrast remains positive (Table 7).

At the primary separation, Quaternionic-2 and Dym-24 controls had residual changes 3.17×10^{-5} and 1.84×10^{-5}, respectively, with no response pattern following the moved switch. In rollouts, consistency training reduced trajectory error among rollouts completed by both methods. Its hard-routed predictions had exactly the same completion, failure, composite-event incidence, and restricted time as ordinary hard-trained hard routing. Ordinary hard routing minus the soft evaluation of those same hard-trained checkpoints had composite-event difference -6.51×10^{-5}, with interval $[-3.89 \times 10^{-4}, 0]$.

B.6 Results across particle counts

Primary endpoint and decision rule. For the pooled estimates in Equation (27), define

$$R_N = \log\left(\frac{\widehat{\text{NRMSE}}_{\text{Gram,N}}}{\widehat{\text{NRMSE}}_{\text{soft,N}}}\right). \tag{37}$$

Positive R_N favors full-soft. The prespecified joint criterion requires positive one-sided lower bounds for both the slope of R_N against $\log N$ and R_{24}, with the latter simultaneous across the five sizes (Appendix B.3).

The slope is 0.9873, with two-sided percentile interval $[0.936, 1.036]$ and one-sided lower bound 0.9451. Both primary lower bounds are positive. The fit summarizes $N = 5$–24, rather than an asymptotic law. All five paired seeds agree with the within-size ordering in Table 8.

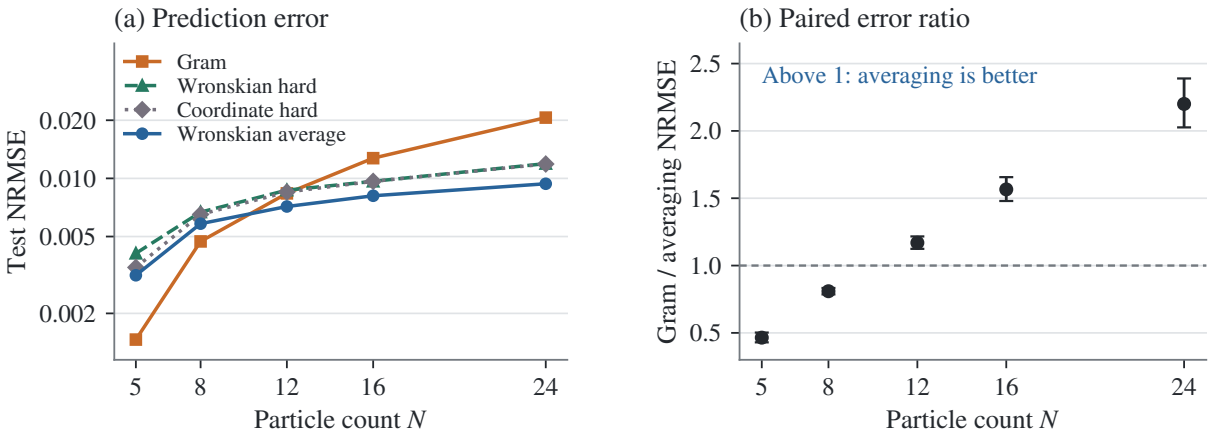


Figure 3: Force error across particle counts after 40,000 updates for every method, which is not equal compute (Section 7.3). Left: pooled test NRMSE for all four trained methods. Right: Gram error divided by Wronskian-averaging error, with simultaneous two-sided 95% intervals. Ratios above one favor averaging.

Table 8: Primary comparison across particle counts. Intervals for the five R_N values are simultaneous two-sided 95% intervals. The last column gives the simultaneous one-sided lower bound.

N	Full-soft pooled NRMSE	Gram pooled NRMSE	R_N [two-sided interval]	One-sided lower bound
5	0.003150	0.001464	-0.7665 $[-0.8447, -0.6884]$	-0.8348
8	0.005828	0.004717	-0.2115 $[-0.2409, -0.1821]$	-0.2372
12	0.007152	0.008363	0.1565 $[0.1171, 0.1959]$	0.1220
16	0.008128	0.012725	0.4483 $[0.3917, 0.5048]$	0.3989
24	0.009380	0.020638	0.7886 $[0.7061, 0.8711]$	0.7165

Hard-routing contrasts. Every simultaneous interval in Table 9 favors averaging, and all five paired seeds agree in each comparison. Posthoc hard selection changes averaging and branch use at a fixed checkpoint. Its contrast is not an isolated effect of continuity.

At $N = 12$, spherical-harmonic baseline minus Wronskian averaging NRMSE is 0.01096 (paired percentile interval $[0.01046, 0.01143]$), with five positive seed directions. This concerns the specified spherical-harmonic architecture under the common learning rate. The trajectory study screens its radial baseline's learning rate separately. It does not compare native equivariant architectures as a class.

B.7 Training at comparable compute

Design. This comparison retrains Wronskian averaging, Wronskian hard selection, and Gram at $N = 24$, and averaging and Gram at $N = 12$, from scratch. Models, optimizer, batch size, data, checkpoint-validation split, and selection rule are those of Appendix B.3. Three things differ: the number of updates, with warmup and cosine decay rescaled to each horizon; the binary64 analytic buffers described below; and the omission of the every-500-update development monitoring, which neither controlled optimization nor checkpoint selection but contributed to the worker times reported in Section 7.2. In those equal-update runs at $N = 24$, mean worker time per run was 12.32 hours for averaging, 2.08 for Gram, 1.90 for Wronskian hard selection and 1.86 for coordinate hard

Table 9: Additive NRMSE contrasts with Wronskian averaging, multiplied by 10^3 for display. Positive values favor averaging. Brackets give simultaneous two-sided 95% intervals within each five-size contrast family. Posthoc hard uses averaging-trained checkpoints. The other columns use separately trained hard models.

N	Posthoc hard	Wronskian hard	Coordinate hard
5	11.430 [10.663, 12.197]	0.941 [0.796, 1.086]	0.309 [0.204, 0.414]
8	3.239 [3.051, 3.427]	0.862 [0.666, 1.057]	0.683 [0.442, 0.923]
12	3.884 [3.659, 4.109]	1.539 [1.169, 1.908]	1.344 [0.939, 1.749]
16	4.634 [4.432, 4.836]	1.527 [1.146, 1.909]	1.524 [1.206, 1.842]
24	6.253 [5.940, 6.565]	2.547 [2.173, 2.921]	2.490 [1.969, 3.011]

selection. A timing pilot on one A40 node type measured each method's synchronized training time per update, together with preparation, validation, and checkpoint overheads, under one protocol. The long budget at $N = 24$ is averaging's projected worker time for 40,000 updates, 11.45 hours. Hard selection and Gram receive the largest numbers of updates whose projected time fits it. The short budgets are Gram's projected times for 40,000 updates, 1.365 hours at $N = 24$ and 1.073 at $N = 12$. Each run warms up over its first 5% of updates and then decays by a cosine to zero at its own horizon, with 16 candidate checkpoints at equal fractions of it. Five seeds use the submitted initialization and data-order streams, so averaging and hard selection share both. Analytic buffers are kept in binary64. In the recovered submitted implementation, casting the model to binary32 had also rounded one of them, the Helmert contrast matrix. The fresh runs are therefore not bitwise replicas. We do not attribute any accuracy difference to this change.

Evaluation. Each selected model was evaluated once on the submitted 20,000-family final-test cohort for its N, after all 35 checkpoint selections had been registered. The evaluator and analysis were committed before the first test access. Because the submitted test sets are reused, this is a prospectively frozen reanalysis rather than an independent sample of test geometries. Pooled NRMSE and the crossed seed/family bootstrap follow Appendix B.3, with 20,000 replicates. The four declared ratios carry simultaneous two-sided 95% intervals on the log scale. The intervals in Table 10 are marginal.

Table 10: All cells of the comparable-compute comparison, with marginal 95% intervals. Hours are mean measured A40 worker time over five seeds.

N	Budget	Method	Updates	Hours	NRMSE	Marginal 95%
24	long	Averaging	40,000	11.494	0.009379	[0.008839, 0.010024]
		Hard selection	302,459	10.703	0.002960	[0.002792, 0.003114]
		Gram	336,426	11.478	0.005550	[0.005489, 0.005613]
24	short	Averaging	4,633	1.371	0.025599	[0.024725, 0.026943]
		Gram	40,000	1.375	0.020638	[0.020244, 0.020978]
12	short	Averaging	10,889	1.078	0.016524	[0.016201, 0.016792]
		Gram	40,000	1.005	0.008363	[0.008213, 0.008526]

Timing. Relative to their targets, the long-budget runs took +0.1% to +0.7% for averaging, −0.3% to +0.9% for Gram, and −13.2% to −3.7% for hard selection. At the short budgets, every averaging run and every $N = 24$ Gram run was within 1.6%, while the $N = 12$ Gram runs took

-11.0% to -3.2%. Six runs fell outside the original $\pm 5\%$ band: two hard-selection runs (-13.2%, -6.9%) and four $N = 12$ Gram runs (-11.0%, -6.7%, -5.5%, -5.3%). All six finished early, and none was excluded. An amendment made during execution turned the band into a reporting rule instead of a reason to halt. Mean worker time for hard selection is 6.9% below averaging's at the long budget, and averaging's is 7.3% above Gram's at $N = 12$. The 35 runs used 192.5 A40 worker-hours, excluding node allocation, qualification, and idle time. Every run selected its final checkpoint, so convergence was not established for any method. Fresh 40,000-update runs of averaging and Gram agree with Section 7.2 to within 0.02% at $N = 24$.

B.8 Blending trained hard-selection models

Design. After the results of Appendix B.7 were known, an exploratory analysis on development data compared inference-time blending rules on the five trained $N = 24$ hard-selection models, and chose the compact-weight parameters below. In that analysis, compact blending made the averaging-trained models 37.8% worse. The confirmatory procedure, with five conditions, four declared ratios, the latency protocol, and all 15 selected $N = 24$ long-budget checkpoints of Appendix B.7 (five each for hard selection, averaging, and Gram), was then frozen before 20,000 new geometries were generated. The regular-core generator was used under a new random key, with four equal strata of the singular-value ratio. No model was retrained or reselected.

Compact weights. With $\sigma_j(A) = \det(AP_j)^2$, m probes, $t(A) = \big(m^{-1}\sum_\ell \sigma_\ell(A)^8\big)^{1/8}$, and $\tau = 1/2$, the compact weights are

$$\tilde{w}_j(A) = \frac{a_j(A)}{\sum_\ell a_\ell(A)}, \qquad a_j(A) = \max\big(\sigma_j(A) - \tau\, t(A),\, 0\big)^3. \tag{38}$$

Scaling every σ_j by a common positive factor leaves $\tilde{w}$ unchanged, so the implementation computes the scores after whitening A.

Remark 20 (Regularity of compact blending). If the charts of $P_1, \dots, P_m$ cover $\mathcal{M}^{\mathrm{fr}}_{d,n}$, the weights (38) are G-invariant and C^2, and $\tilde{F} = \sum_j \tilde{w}_j Y_j$ is G-equivariant, and C^2 whenever the backbone is C^2. Its worst-case support is still at most m.

Proof sketch. Coverage gives $\max_j \sigma_j > 0$. A power mean is at most the maximum, so $\max_j \sigma_j - \tau t \geq \max_j \sigma_j/2 > 0$ and the denominator is positive. Each σ_j is a polynomial, t is smooth where $\sum_\ell \sigma_\ell^8 > 0$, and $x \mapsto \max(x, 0)^3$ is C^2. Hence each $\tilde{w}_j$ is C^2. For a fixed j, the open sets $\{\sigma_j < \tau t\}$ and $\{\sigma_j > 0\}$ cover the domain because $t > 0$. On the first, $\tilde{w}_j Y_j$ vanishes identically. On the second, which is the chart of P_j, the gauge g_{P_j} is smooth, so $\tilde{w}_j Y_j$ is C^2 when f is. The two descriptions agree on the overlap. Invariance of σ_j and t, together with $g_{P_j}(RA) = R g_{P_j}(A)$, gives equivariance term by term. □

Smooth relative cutoffs of frame weights appear in Dym et al. (2024) and Pozdnyakov and Ceriotti (2023), and the latter also symmetrizes trained models. Remark 20 records the corresponding statement for this bank. It concerns the exact map on full-rank inputs. The implementation evaluates in floating point with conditioning guards, and its detached preprocessing does not propagate derivatives through the frames and weights, so it does not supply forces as derivatives of a predicted energy.

Accuracy. Table 11 reports the five conditions. The declared ratios, with simultaneous two-sided 95% intervals, are 0.9375 $[0.9342, 0.9408]$ for full blending over hard routing, 0.9731 $[0.9715, 0.9746]$ for compact blending over hard routing, 0.306 $[0.289, 0.324]$ for hard-trained compact blending over averaging-trained full blending, and 0.518 $[0.486, 0.551]$ for hard-trained compact blending over Gram. The last two compare different trained models and inherit the training costs of Appendix B.7.

Table 11: Blending on 20,000 new $N = 24$ geometries: pooled NRMSE over five seeds, and median A40 inference time in milliseconds at three batch sizes.

Trained as	Evaluated with	NRMSE	Batch 1	Batch 32	Batch 128
Hard selection	hard routing	0.002952	5.47	4.94	6.00
Hard selection	full blend	0.002767	15.52	35.82	125.87
Hard selection	compact blend	0.002872	5.35	11.34	32.67
Averaging	full blend	0.009383	15.47	35.82	125.90
Gram	—	0.005547	1.28	3.13	11.36

Latency. Times are medians over 20 randomized blocks of 10 calls per condition and batch size, for the seed-zero models on one A40 and a separate 512-state cohort. They include scoring, frame construction, gathering, and prediction, but not data transfer. Mixed precision, TF32, and compilation are off. They describe this implementation and host.

Boundary response. Sixteen frame-switch crossings were selected from geometry alone. For each seed we compare prediction changes across a crossing at input separations 10^{-6} and 10^{-4}. Their median ratio is 0.010 for both blends, for averaging, and for Gram, as for a continuous map, and 0.91–0.96 for hard routing, as for a jump. These numerical checks do not prove continuity or measure how often natural inputs cross a switch. Appendix B.9 measures the size of these jumps.

Uniform top-8 control (post hoc). On the same new geometries, averaging the hard-trained backbone uniformly over the eight highest-weight frames lowers NRMSE by 5.74% relative to hard routing (marginal 95% interval for the ratio $[0.9399, 0.9450]$). That is 92% of the full blend's reduction, although this predictor can jump when its selected set changes and has no global continuity guarantee. The full blend is more accurate, with top-8/full ratio 1.0055 $[1.0052, 1.0058]$ (marginal interval). This analysis was added after the confirmatory results were known, and it does not identify how much of either reduction is due to continuity.

Scope. The hard-routing and blending comparisons, and the top-8 control, reuse the same five hard-trained backbones. The averaging and Gram baselines each use five separately trained models with paired seed indices. Reported NRMSE pools the errors of the five models of each condition. It is not an ensemble across seeds. All conditions concern one synthetic task at one particle count, and the fresh geometries provide evidence on new inputs, not independent training replication or real data.

B.9 Jumps of trained hard-selection models

Section 7.1 establishes the jump mechanism at $N = 5$. Here we measure the jumps of the five $N = 24$ hard-selection models of Appendix B.7 and follow them through training. These analyses were specified after the results of Appendices B.7 and B.8 were known. They are descriptive, involve no retraining, and use no final-test states.

Crossings. A crossing is a parameter t^* at which the hard-selection winner changes on the straight path $a + t(b - a)$ between two independent regular-core geometries. It is located from geometry alone, by 48 bisection steps between endpoints with different winners. The bisection predicate (same winner as at $t = 0$) need not be monotone, so the located switch need not be the first on the path. The predictor is evaluated at $a + (t^* \pm \varepsilon)(b - a)$ for $\varepsilon \in \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$. At every ε, both endpoints must have singular-value ratio above 0.01 and different hard winners in the float32 inputs seen by the network. We use the 16 crossings of Appendix B.8 and 64 new ones under a new random key, which required 67 of a fixed budget of 512 path pairs. These crossings are selected, not encountered. They do not estimate how often natural inputs or trajectories cross a switch.

A finite-separation bound. At a crossing, let $p_\pm$ and $y_\pm$ be the hard-selection prediction and the target at the two endpoints. We divide every force difference by the RMS Frobenius norm of the target forces on the 20,000 fresh geometries of Appendix B.8, $D = \sqrt{24}\, F_{\mathrm{rms}}$ with $F_{\mathrm{rms}} = 1.2363$, so that all quantities are on the scale of NRMSE. Write $J = \|p_+ - p_-\|/D$ for the jump, $E_\pm = \|p_\pm - y_\pm\|/D$ for the endpoint errors, $E_{\mathrm{loc}} = \left((E_-^2 + E_+^2)/2\right)^{1/2}$ for the local error, and $T = \|y_+ - y_-\|/D$ for the change of the target. The triangle inequality gives

$$J \;\le\; E_- + E_+ + T \;\le\; 2E_{\mathrm{loc}} + T. \tag{39}$$

The target is continuous, so $T \to 0$ as $\varepsilon \to 0$, and $\max(E_-, E_+) \ge (J - T)/2$. A jump forces an error of at least about half its size on one side of the switch, and a predictor that is accurate there can only jump a little. Recomputing J, $E_\pm$, and T from the saved arrays for all 80 crossings, four separations, five seeds, and 16 checkpoints (25,600 cells), every cell satisfies (39). This checks alignment and normalization, not the inequality itself. At $\varepsilon = 10^{-6}$ and the final checkpoint, $J/(E_- + E_+ + T)$ has median 0.25 and maximum 0.42 (0.33 median at 37,808 updates).

Jump size. The final checkpoints, at 302,459 updates, are the selected models of Table 1. At $\varepsilon = 10^{-6}$, over their 400 seed–crossing cells, the median jump is 0.115% of RMS target force, the 90th and 99th percentiles are 0.155% and 0.206%, and the maximum is 0.313%. Per-seed medians range from 0.107% to 0.123%. The median local error at the same endpoints is 0.222%, and the pooled NRMSE of these models on the fresh geometries is 0.0030 (Table 11). The cells repeat the same 80 geometries for five seeds. They are not 400 independent crossings, and the maximum is not a bound for unseen inputs. The jumps do not shrink with the separation. The median ratio $J(10^{-6})/J(10^{-4})$ is 0.94, whereas the median of T falls by a factor of 100. At $\varepsilon = 10^{-3}$, the target itself changes by a median 0.38%, so we report $\varepsilon = 10^{-6}$. For comparison, across the 16 original crossings full blending changes its prediction by a median 0.0004%, the size of the target change at this separation.

Through training. Figure 4 follows the same 80 crossings through the 16 checkpoints saved at intervals of 18,904 updates in each run. From the second checkpoint, at 37,808 updates and nearest the budget of Section 7.2, to the last, the pooled median jump falls to 0.135 of its value (marginal 95% interval $[0.120, 0.151]$), and the median local error to 0.181 $[0.164, 0.197]$. Their paired ratio is 0.745 $[0.653, 0.828]$, so in these runs jumps shrank faster than the error around them. Intervals use 20,000 bootstrap replicates that resample seeds and crossings as crossed factors, preserving the 16 original and 64 new crossings, with the same draws at every checkpoint so that ratios are paired. The decline is not monotone for every seed, and five seeds give limited information about training randomness. All checkpoints lie on one cosine schedule, so training progress and learning-rate decay are confounded, and the second checkpoint is not a model trained to completion at that budget.

Panel (b) evaluates every checkpoint with full blending and with the uniform top-8 average of Appendix B.8 on the first 2,000 checkpoint-validation states, resampled within strata. The full blend's reduction of NRMSE falls from 12.5% at 37,808 updates to 6.2% at the end, not monotonically, so the gain from blending declines along with the jumps. The uniform top-8 average, which has no global continuity guarantee, recovers 90.4–91.1% of the full blend's reduction across the saved checkpoints. This approximately stable share does not identify a continuity-specific contribution or its evolution, because the two rules also differ in support and weights. These states also select checkpoints, so panel (b) describes trends, not test accuracy. The final reduction agrees with the 6.3% on fresh geometries.

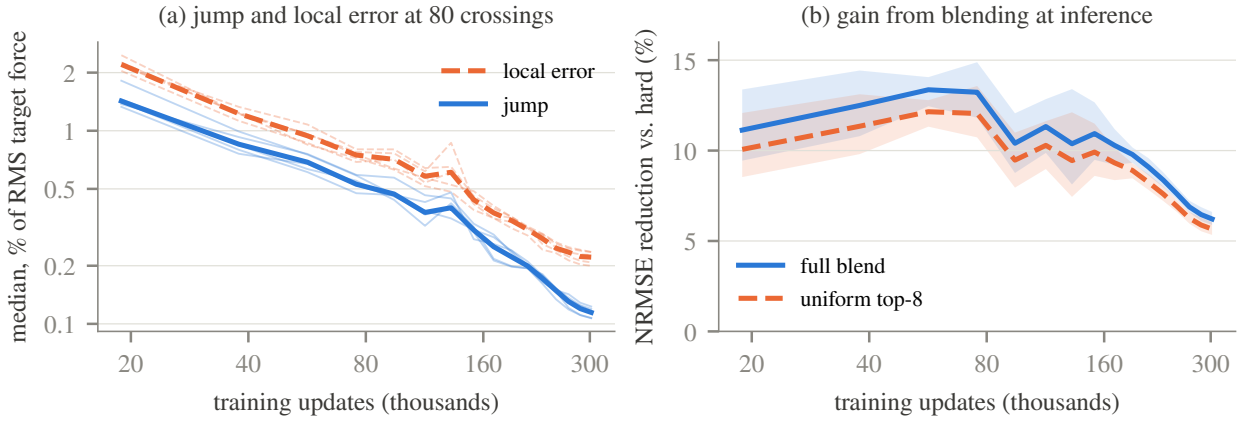


Figure 4: Jumps and blending gains through training of the five $N = 24$ hard-selection runs (post hoc). (a) Median jump and median local error at $\varepsilon = 10^{-6}$ over 80 selected crossings, as a percentage of RMS target force, pooled over seeds (thick) and per seed (thin). (b) NRMSE reduction relative to hard routing when each checkpoint is evaluated with full blending or with a uniform average over its eight highest-weight frames, on 2,000 checkpoint-validation states. Bands are marginal 95% bootstrap intervals. Horizontal axes are logarithmic. One cosine schedule confounds training progress with learning-rate decay.

Scope. Finite separations are not one-sided limits. Small observed jumps are a property of these trained models. The obstruction of Section 5 concerns the choice of frames and does not by itself force a predictor to jump.

B.10 Runtime sweep

Cost per prediction. For m probes, forming AP_j, scoring, and transforming particle inputs and outputs cost $O(mnd^2)$, polar decompositions cost $O(md^3)$, and the backbone calls cost $O(mC_f(N))$. The probes are constructed once. With d fixed and $m = \Theta(N)$, scoring and transforming cost $O(N^2)$. A fixed-width dense-attention backbone costs $O(N^2)$ per call and hence $O(N^3)$ over the full bank. Hard selection scores the full bank and then makes one backbone call.

Measured runtime. Table 12 reports measured costs from a separate microbenchmark at $N \in \{5, 8, 12, 16, 24, 32\}$ on one A40, with TF32 disabled. Weights were untrained, and inputs were resident on the GPU. Each condition used five warmup and 30 measured repetitions, and the table gives medians. A training step uses effective batch 32 with microbatches of 4, unlike the batch of 512 used for training in Appendices B.3 and B.7. At this small batch, the hard-selection step time hardly changes with N, consistent with fixed per-step costs, so these step times should not be

compared with the worker times of Appendix B.7. Inference times are for a whole batch. These numbers are separate from the trained-model latencies of Table 11, which use a different timing protocol.

Table 12: Measured runtime in milliseconds: median time for one training update (effective batch 32) and for inference on a batch of 32, for untrained models on one A40. Averaging evaluates $3N - 11$ frames per input.

Method		$N = 5$	8	12	16	24	32
Averaging	frames	4	13	25	37	61	85
	train	265.7	375.6	457.1	494.1	600.3	661.2
	infer	7.3	10.8	19.2	24.6	37.2	53.1
Hard selection	train	265.3	264.3	265.9	266.9	267.2	267.1
	infer	6.7	6.8	6.1	7.0	7.0	6.9
Top-8	train	—	268.9	258.7	271.9	273.8	279.0
	infer	—	8.5	7.9	9.0	9.2	9.6
Gram	train	158.9	157.3	160.9	161.7	162.9	161.0
	infer	1.6	1.7	1.7	1.6	3.1	4.6
Native equivariant	train	479.2	474.9	480.7	476.8	491.1	493.7
	infer	13.5	13.7	13.9	14.9	30.4	53.0

Over this range, averaging's inference time at batch 32 grows by a factor of 7.2 while its bank grows by a factor of 21, so fixed costs still matter at these sizes. Hard selection stays near 7 ms, although it scores the whole bank. Top-2 and top-4 conditions, and inference at batch size 1, are recorded in the full timing results.

B.11 A worked application and when fewer frames save work

Theorems 1 and 2 give concrete frame-count and regularity consequences for the following molecular setting. This is a theoretical application. Our experiments use synthetic force fields.

Setting. Consider learning the forces on the atoms of one molecular species, for instance to accelerate a molecular-dynamics simulation of its conformations. We assume persistent atom identifiers, so that each input is a labeled configuration in $\mathbb{R}^{3\times N}$, and an isolated target that is invariant under translations and equivariant under O(3), so that rotating or reflecting the molecule transforms every force in the same way. The $n = N - 1$ contrasts remove translations. A backbone that is not equivariant, such as a pretrained model or a sequence model over the atoms in a fixed order, becomes exactly O(3)-equivariant by averaging over the probe bank. Consistently indexed anatomical or robot landmarks fit the same pattern.

Symmetry under relabeling. The averaging wrapper does not enforce equivariance under exchanges of equivalent atoms, which can matter even for a single molecule (Chmiela et al., 2018). The example is therefore a proposed model on a specified labeled operating domain. Consistency under such exchanges, and transfer across molecular sizes or bonding graphs, require additional design and validation.

Rank and conditioning. The continuity and regularity guarantees hold throughout the full-rank domain. Quantitative Lipschitz estimates additionally depend on conditioning and the probe-coverage margin (Proposition 15).

Smoothness. For energy models, let e be a scalar backbone and define the averaged energy

$$E(A) = \sum_j w_j(A)\, e(g_j(A)^{-1}A).$$

With the trivial output representation, Theorem 2 makes E invariant and C^{2k-1} for a C^{2k-1} backbone. Composing with the fixed centering and contrast map and differentiating with respect to the particle positions gives the forces: $k = 1$ with a C^1 backbone gives continuous forces, and $k = 2$ with a C^3 backbone gives C^2 forces. Hard selection applied to a noninvariant e can make the energy jump at frame switches. Computing these forces requires derivatives through both frames and weights, including the preprocessing. Our direct-force experiments do not compute them (Appendix B.2). These are regularity statements, and this energy application was not tested.

Frame counts. At $N = 20$ ($n = 19$), a continuous equivariant weighted frame defined on the entire full-rank domain must use at least 9 frames at some input (Theorem 1). This need not occur on the restricted conformations visited by one molecule. Near a full-rank configuration A_0, any probe P with $\det(A_0P) \neq 0$ gives a single continuous equivariant frame on the invariant open set where $\det(AP) \neq 0$. Our globally defined bank has 49 probes and covers every full-rank shape without a molecule-specific choice of chart. On the same 19-vector representation, the O(3) construction of Dym et al. (2024) has $2n(n-1) = 684$ frames. Its permutation guarantees concern permutations of its input columns, and permutations of atoms generally mix the contrast columns, so this count compares spatial frames only.

When fewer frames reduce cost. Evaluating each positive-weight frame separately costs one backbone call per frame and input. Frame construction and weight scoring add to this. Scoring the full bank takes $O(n^2)$ operations per input at fixed d. A smaller bank reduces the candidates scored by both hard selection and averaging, and the backbone calls required by full averaging. For hard selection the backbone count stays at one. Wall-clock savings are proportional to the number of calls when those calls dominate, for example with a large backbone, a saturated device, or memory-limited batching. Running calls in parallel on otherwise idle hardware reduces elapsed time without reducing the number of calls. In Table 12, averaging's inference time at batch 32 grows by a factor of 7.2 from $N = 5$ to $N = 32$ while its bank grows by a factor of 21. In Table 11, the 61-frame blend takes 7.3 times as long as hard routing at batch 32, not 61 times. The ratio $684/49 \approx 14$ is therefore a ratio of bank sizes, not a predicted speedup.

B.12 Runtime comparison with Dym et al.

Table 13 compares averaging over our bank, with $3(n-3)+1$ frames, against our implementation of the full-rank O(3) weighted frame of Dym et al. (2024), with $2n(n-1)$ frames. Both use the same untrained Transformer backbone with identical parameters, the same inputs on one A40, and a shared implementation that evaluates frame–input pairs in packed chunks of 64. Times include frame construction and are medians of 50 measured forwards after 10 warmups, from a single sweep. The cutoff rows use the relative C^2 rule of (38), with $\tau = 1/2$, on our squared determinant scores and successively on Dym's squared norm, area and volume scores. This implementation choice does

not equalize active frame counts. No accuracy was measured. This implementation is separate from those of Tables 11 and 12, so its times are comparable only with each other.

Table 13: Median inference time in milliseconds per batch and peak allocated CUDA memory in MiB, for averaging over our bank and over our implementation of the O(3) frame of Dym et al. (2024), with identical untrained backbones on one A40. The ratio is Dym's time over ours. With cutoff weights at batch size 1, Dym's frame has 21.8% lower median latency at $N = 24$ and 21.2% at $N = 12$. Our bank uses more peak memory at batch size 32.

	Candidate frames				Time (ms)			Memory (MiB)	
N	Ours	Dym	Weights	Batch	Ours	Dym	Ratio	Ours	Dym
12	25	220	full	1	4.87	6.66	1.37	16.6	13.6
			full	32	25.73	130.64	5.08	214.8	14.0
			cutoff	1	5.69	4.48	0.79	12.3	12.9
			cutoff	32	9.52	22.52	2.36	60.4	13.7
24	61	1,012	full	1	5.04	20.85	4.14	26.3	17.1
			full	32	57.14	590.29	10.33	507.9	24.1
			cutoff	1	5.22	4.08	0.78	16.1	17.1
			cutoff	32	10.84	55.61	5.13	78.6	20.5