DISTANCE DOES NOT PREDICT COMPETENCE: A TWO-AXIS DECOMPOSITION OF OUT-OF-DISTRIBUTION FAILURE

Anonymous authors
Paper under double-blind review

ABSTRACT

Out-of-distribution generalization is usually measured in a single aggregate metric, which assumes that all directions away from the training data have identical effects on extrapolation. We show they do not. We introduce a two-axis evaluation protocol that separates the directions that are governed by a model's imposed structure from those governed by the law it has to learn. Invariant architectures generalize exactly along their built-in symmetries but fail as badly as generic Neural Networks when the law itself must be extrapolated, and the representation that is most accurate in-distribution can be the least accurate out-of-distribution.

We show that distance from the training data does not predict competence. At matched distance, errors differ by three to four orders of magnitude depending only on the direction taken, and a model given the most distant test set can outperform the same model given closer test sets. On distant queries whose answers the recovered law already provides, six standard uncertainty and distance metrics rank these answerable points as the least answerable. Being far from the data and being undetermined by it are distinct conditions.

The shape of out-of-distribution failure, whether a model diverges or flattens, follows whether its representation is bounded, and the pattern holds in Transformer sequence models without hand-built invariances. We argue that a model's competence should be predicted by whether it computes a representation structurally equivalent to the generating law, which training data combined with the chosen representation can signal, and not by taking into account its proximity to the training set. Reliable abstention requires knowing what the hypothesis class determines, a signal that distance metrics lack.

1 INTRODUCTION

A model that fails Out-of-Distribution (OOD) is usually characterized by one number: an aggregate error on a shifted test set. That practice embeds an assumption, rarely stated, that the directions away from the training data are equivalent, so a single scalar summarizes what the model has lost. This paper is an argument that the assumption is false, and that several things which are invisible under aggregation become sharp and measurable once the directions are separated.

The setting where this is easiest to see is one in which structure has been deliberately imposed. An equivariant architecture is constrained along the orbits of a symmetry group and unconstrained elsewhere, so a problem carrying both a symmetry and a functional law presents at least two qualitatively different ways to leave the training set: move along the symmetry group, where the constraint holds algebraically, or move along the scalar law, where the constraint says nothing. A radially symmetric potential is the simplest case and the general one: interatomic potentials, physical simulators, and most scientific regression targets have this form.

We introduce a protocol that separates these axes and performs evaluation on each independently, and we report four findings:

(i) **The axes come apart, and stay apart at increasing realism** (§3, §7). An invariant model's error is unchanged three quarters of the way around an annulus into unseen angles and degrades

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

on the radial axis to the level of a generic Neural Network; one body order up, on a Lennard-Jones dimer (Jones, 1924) in $\mathbb{R}^3$, a SchNet-style (Schütt et al., 2017) Radial Basis Function (RBF) Neural Network is the best of three models in distribution and the worst on the law axis in all twenty-four batches; and with no hand-built structure at all, a Transformer's (Vaswani et al., 2017) positional encoding fixes the axis.

(ii) **Distance from the training data does not order the outcome** (§4), the paper's central negative result. Leaving one training set along two directions, error at matched nearest-neighbour distance differs by between three and four orders of magnitude ($744\times$ to $6787\times$) across four architectures; the most distant regime incurs a normalized error of 0.28 for the LayerNorm model where the nearest OOD regime incurs a normalized error of 3679, in every seed; and from $d = 2$ to $d = 32$ the catastrophic compressing direction becomes the nearer of the two in nearest-neighbour distance while the ratio holds. Feature-space scores recover the ordering for three of the four architectures and fail on the $\tanh$-activated Neural Network in the way §4 predicts.

(iii) **The shape of failure follows representational boundedness** (§5). The literature reports both unbounded continuation of a trend (Hein et al., 2019; Xu et al., 2020) and reversion toward a constant (Kang et al., 2024), with different tasks and architectures. When the task, optimizer and capacity budget are held fixed, the way a model fails OOD depends on whether its trunk is bounded: an unnormalized ReLU Neural Network diverges, its predictions growing by $69\times$, whereas a LayerNorm Neural Network, a Transformer and an unnormalized $\tanh$ Neural Network all plateau. Because the $\tanh$ Neural Network flattens with no normalization layer, the operative property is boundedness, not normalization.

(iv) **The model's own uncertainty has the same representational boundary, and abstention needs a different quantity** (§6). Ensemble disagreement tracks error along the axis a prior acts on and decouples from it along the law axis, where the ensemble is both inaccurate and confident; on a target whose recovered law provides the correct answer at 56% of the distant queries, six standard metrics all rank those as the least answerable. Determination relative to a hypothesis class, which we define, is what an abstention rule needs, and no geometric or dispersion score supplies it.

Each result is invisible in an aggregate; together they point at the reading of §8, that what predicts competence is whether the model computes a representation structurally equivalent to the generating law. Which axes decouple is predicted by the inference-path rule of Rocha et al. (2026): a quantity whose path contains only exact stages is exact, and a fitted stage on the path contributes an error bounded by its residual times the downstream Lipschitz factors, so a coordinate that an invariant projection discards before the Neural Network (Villar et al., 2021; Bronstein et al., 2021) leaves the model exact along that axis (Appendix K).

2 A TWO-AXIS EVALUATION PROTOCOL

Let the target be $f(\mathbf{x}) = g(\rho(\mathbf{x}))$, with ρ a maximal invariant of a group $\mathfrak{G}$ acting on the input and g a scalar law. Every input direction is then of exactly one of two kinds: it either moves along an orbit of $\mathfrak{G}$, leaving ρ and hence the target unchanged, or it changes ρ and therefore $g(\rho)$. Training data covers a restricted range of the orbit coordinate and of ρ, and a test point can leave that range in three ways, which we evaluate separately: an unseen orbit coordinate, the angle in the radial case, at a trained ρ (the **symmetry axis**), an unseen ρ at a trained orbit coordinate (the **law axis**), and both together.

Three models see identical data. A **generic** Neural Network on the raw input must learn the symmetry from data. An **invariant** model computes ρ and applies a Neural Network to it to learn g; where ρ is the norm of a single vector under the orthogonal group $\mathrm{O}(d)$, the rotations and reflections of $\mathbb{R}^d$ that fix the origin, this is the canonical instance of the scalar construction Villar et al. (2021) prove universal, not a weakened stand-in. A **symbolic search** receives raw coordinates and a fixed operator set, with ρ deliberately absent, so it must discover the invariant before it can express any law over it; we enumerate bottom-up, merge expressions whose outputs coincide on a probe set of inputs (observational rather than symbolic equivalence), and select on fit error plus a description-length penalty.

Throughout we report RMSE normalized by the RMS of the true signal in each regime, so 1.0 means error as large as the signal; the comparison of §4 spans regimes whose signals differ by orders of

Table 1: Instantiations I to III: normalized RMSE, mean $\pm$ sd over twenty-four disjoint batches, each with its own training draw and model seeds. Bold marks a genuine win; an error of the same order as the signal, whichever model has it, is not one. The RBF model is added in Instantiation III only, where it is the best of the three in-distribution and the worst on the law axis in all twenty-four batches. The symbolic column is a bound: the residual of an exactly recovered symbolic expression is a float64 machine-precision round-off (Reproducibility Statement).

setting	regime	generic	invariant	RBF	symbolic
I: $\mathrm{O}(2)$, $g(r) = r^{-2}$	in-distribution	$0.013 \pm .006$	$0.042 \pm .069$		$< 10^{-15}$
	symmetry axis	$2.689 \pm .835$	$\mathbf{0.042 \pm .069}$		$< 10^{-15}$
	law axis	$2.213 \pm .258$	3.681 ± 2.96		$< 10^{-15}$
	both	22.77 ± 9.52	3.681 ± 2.96		$< 10^{-15}$
II: pairwise, $g(\theta) = \cos^2\theta$	in-distribution	$0.003 \pm .002$	$0.002 \pm .003$		$< 10^{-15}$
	symmetry axis	$0.619 \pm .058$	$\mathbf{0.002 \pm .003}$		$< 10^{-15}$
	law axis	$1.014 \pm .039$	$0.924 \pm .124$		$< 10^{-15}$
	both	$0.695 \pm .042$	$0.924 \pm .124$		$< 10^{-15}$
III: $\mathrm{O}(3)$, Lennard-Jones	in-distribution	$0.041 \pm .007$	$0.023 \pm .011$	$\mathbf{0.002 \pm .003}$	
	symmetry axis	$0.930 \pm .651$	$\mathbf{0.023 \pm .011}$	$\mathbf{0.002 \pm .003}$	
	law axis	$\mathbf{0.921 \pm .013}$	$0.959 \pm .021$	$1.027 \pm .002$	
	both	$\mathbf{0.927 \pm .031}$	$0.959 \pm .021$	$1.027 \pm .002$	

magnitude, and so we normalize there by a fixed scale, the standard deviation (sd) of the training targets, and state it.

Why this evaluation is fair. The setup is not biased toward the invariant model: it succeeds along the symmetry axis because its architecture discards that coordinate, so it cannot see the shift, and along the law axis it is a generic Neural Network on a lower-dimensional input, with no structural advantage. Were the setup tautological, it would win on both axes or both models would fail identically on both axes.

3 THE AXES COME APART

Instantiation I, a single vector under $\mathrm{O}(2)$, the rotations and reflections of the plane. The target is $f(\mathbf{x}) = g(r)$ on $\mathbb{R}^2$ with $r = \|\mathbf{x}\|$; training covers one wedge of an annulus, $r \in [1, 3]$ and $\theta \in [0, \pi/2]$. Table 1 reports the dissociation. Moved three quarters of the way around the annulus into angles it never saw, the invariant model's error stays identical, since it computes r and never sees the angle, while the generic Neural Network degrades by more than two orders of magnitude, by at least $82\times$ in every batch. On the radial axis both fail, with errors two to four times the signal, and the invariant model is not better at all, being worse in 17 of 24 batches. The symbolic search recovers the law and is exact in every regime.

Instantiation II, one body order up. For $\mathrm{O}(d)$ acting diagonally on a tuple of vectors the invariant ring is generated by pairwise inner products. Two unit bond vectors at angle θ with orientation φ free to rotate give $\rho = x_1 \cdot x_2$, and we take $g(\theta) = \cos^2\theta$, proportional to the $\ell{=}2$ angular basis function higher body-order architectures build from spherical harmonics (Gasteiger et al., 2020; Batatia et al., 2022). The shape repeats exactly (Table 1).

Instantiation III, a Lennard-Jones dimer under $\mathrm{O}(3)$**.** Both laws above are dilation-covariant, so a scale-equivariant architecture could in principle absorb them (Sosnovik et al., 2020); a load-bearing test needs an intrinsic scale. The Lennard-Jones potential (Jones, 1924) $V(r) = 4\epsilon[(\sigma/r)^{12} - (\sigma/r)^6]$ on $x_1, x_2 \in \mathbb{R}^3$ supplies one, is non-monotonic, and is the two-body term that equivariant potentials reduce to, at short range (Batzner et al., 2022; Batatia et al., 2022). Training covers $r \in [1.0, 1.5]$ in reduced units, spanning the repulsive wall and the minimum at $r = 2^{1/6}\sigma$, with φ unrestricted. We add a third model, an RBF Neural Network receiving $K{=}12$ Gaussians centred across the training range and no wider, the SchNet and DimeNet featurization (Schütt et al., 2017; Gasteiger et al., 2020), before the same Neural Network.

The dissociation survives at this realism. On the symmetry axis, now moving through two angular coordinates jointly, the invariant model does not degrade, again identically, while the generic Neural

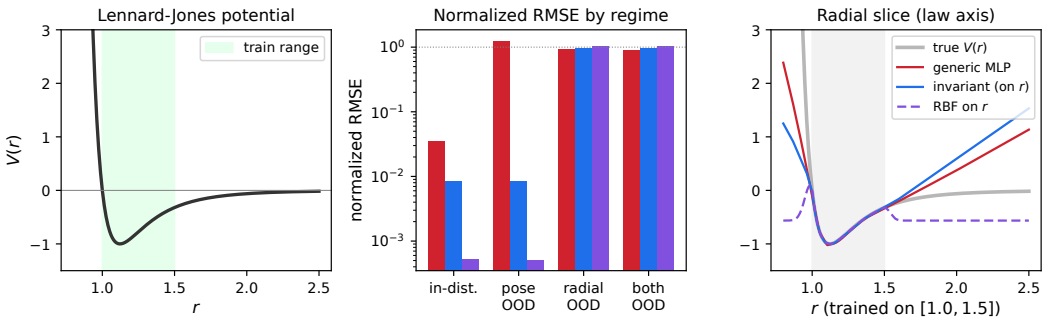


Figure 1: Instantiation III. Left: the Lennard-Jones potential, non-monotonic with an intrinsic scale σ; shaded is the training range. Center: normalized RMSE by regime on a log scale, with the dotted line marking error equal to signal. Right: a radial slice at $\theta = \pi/4$. The RBF model (dashed) is accurate inside its own centers and collapses to a near-constant outside them in both directions, while the raw-invariant and generic models at least continue a trend, wrong though it is.

Network degrades by more than an order of magnitude (Table 1). On the law axis the ordering inverts, and the inversion is the point: the RBF model is the most accurate of the three in-distribution in all twenty-four batches and the worst of the three on the law axis in all twenty-four. Outside $[\mu_1, \mu_K]$ every Gaussian is saturated near zero, so the Neural Network's input has collapsed to approximately a fixed vector regardless of how far r has moved (Figure 1). So a practitioner, based on validation error inside the training range, would choose precisely the model that fails the hardest outside of it.

4 Distance from the training data does not order the outcome

The results above are usually read through a distance-based lens: the model degrades because the test point is far from the training data, and a distance- or density-aware estimator would flag it as such (Ovadia et al., 2019). This section shows this reading is wrong for the Lennard-Jones setting of Instantiation III, and not marginally.

Design. We use the Lennard-Jones geometry with four particles at the vertices of a regular tetrahedron of edge s under a uniformly random rotation, so the target is exactly $V(s)$ while the input is a genuine rotated point set. Training covers $s \in [1.0, 1.5]$. Two directions leave it. **Compressing** ($s < 1.0$) drives the target up the repulsive wall as s^{-12}, from 0 to 1750, while moving the input only slightly, toward the centroid of the training cloud. **Expanding** ($s > 1.5$) moves the input a long way while $V \to 0$, so the law is nearly flat and the target stays inside its training range. Errors are reported over the sd of the training targets (§2). Four architectures, those of §5: an unnormalized ReLU Neural Network, a ReLU Neural Network with LayerNorm, a Transformer, and an unnormalized $\tanh$ Neural Network, five seeds each.

At matched distance, error differs by three orders of magnitude. Compressing to $s \in [0.70, 0.80)$ and expanding to $s \in [1.5, 1.7)$ sit at the same mean nearest-neighbour distance, the expanded regime slightly further from the training data, and their errors differ by $744\times$ to $6787\times$ across the four architectures (Table 2). These are not seed artefacts: computed within each seed, the ratio is never below $618\times$ for any model.

The ordering is inverted. The most distant regime we evaluate, expanding to $s \in [5, 8)$ at $17\times$ the in-distribution nearest-neighbour distance, incurs error 0.28 for the LayerNorm model; the nearest OOD regime, compressing to $s \in [0.60, 0.70)$, incurs 3679 for the same model, ten times closer to the training data and four orders of magnitude worse, in all twenty model-seed combinations (Table 2). We do not claim more than this: the rank correlation of error with distance over the twelve regimes carries nothing, its bootstrap intervals containing zero at $n = 12$ (Appendix B); the matched pair and the inversion carry the argument.

The same test in the model's own feature space. The claim predicts its own exception, and it is testable. Feature-space Mahalanobis (Lee et al., 2018), its background-corrected form (Ren et al., 2021) and feature-space nearest neighbours (Sun et al., 2022), on the pooled representation the output head reads, do what input distance cannot: the neighbour score orders the twelve regimes at

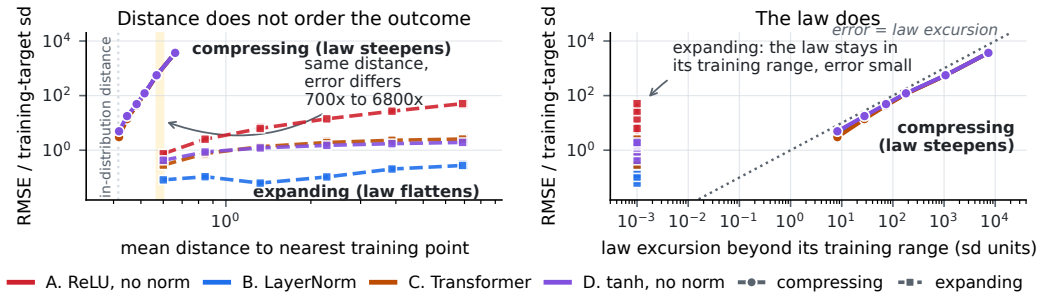


Figure 2: Left: error against mean distance to the nearest training point, the two directions separated and the matched pair shaded; the branches differ by three orders of magnitude at the same x. Right: the same errors against how far the law leaves its training range, in units of the training-target sd: the compressing branch lies on the diagonal, every expanding point at the left edge.

Table 2: RMSE over the training-target sd, mean of five seeds, with the ratio's seed range rounded outward. Matched pair: compressing $s \in [0.70, 0.80)$ and expanding $s \in [1.5, 1.7)$, at nearest-neighbour distances 0.568 and 0.600. Inversion: the nearest OOD regime, compressing $s \in [0.60, 0.70)$ at 0.661, against the most distant, expanding $s \in [5, 8)$ at 6.893 ($17\times$ the in-distribution 0.416).

	matched pair				inversion	
model	compress	expand	ratio	per seed	nearest OOD	most distant
ReLU, no norm	549.96	0.74	744×	[618, 863]	3671	50.78
LayerNorm	555.27	0.08	6787×	[5311, 10502]	3679	0.28
Transformer	549.43	0.28	1931×	[1374, 3017]	3673	2.56
tanh, no norm	557.79	0.43	1309×	[1185, 1481]	3682	1.93

$\rho \geq 0.92$ for the unnormalized, LayerNorm and Transformer models, where input distance is neither correlated nor anti-correlated with error (Table 7). On the tanh Neural Network, which §5 identifies as saturating the most, every feature score fails and the neighbour score signals the catastrophic compressed regime less OOD than the benign expanded one: a score defined on a representation inherits that representation's competence boundary.

And not only for this law. Swapping the law while holding everything else fixed, the inversion survives all four laws we tried, one of them not a power law, at $11\times$ to $6808\times$ over eight law–model combinations, the magnitude tracking how far the law travels (Appendix D).

Nor only in low dimension. Holding the law fixed, $V(r)$ on $r = \|x\|$ with $x \in \mathbb{R}^d$ and training radii $[1.0, 1.5]$ in every direction, we sweep d from 2 to 32 (Appendix E, Table 10). The geometry moves by more than an order of magnitude: the compressed shell of the pair is $16\times$ further from the training set than in-distribution samples are at $d = 2$ and nearer than they are from $d = 16$ on, while the expanded shells stay further at every d. The errors do not move, and the ratio stays above $573\times$ in every one of forty batches, so distance passes from a weak positive predictor of error to an anti-predictor.

Convex hull membership is no better. Every compressed configuration we checked lies inside the convex hull of the training set, against 30% of fresh in-distribution samples, so interpolation in the sense of Balestriero et al. (2021) incurs error equal to the signal.

What does order it. The right panel of Figure 2 plots the same errors against how far the law leaves its training range: the compressing branch lies along the diagonal, error $\approx$ excursion, and every expanding point collapses to the left edge where the excursion is zero. The error is set by how far the law travels, a property of the problem and the model class, and distance tracks competence only to the extent that the space in which it is measured already encodes the law.

Making the law observed restores competence, and only there. The account is causal in form, so we test it that way, holding the law, the geometry, the architectures and the training-set size fixed and widening only the training range, so that the wider conditions have lower sample density

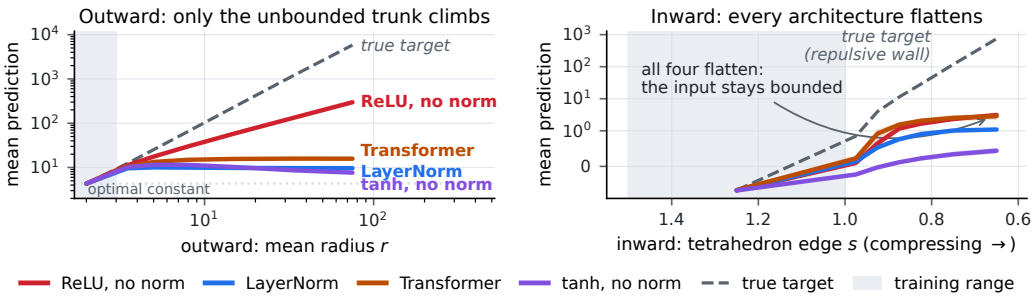


Figure 3: Four models, one task, one optimizer, one capacity budget, five seeds. Left: outward motion, where the target grows without bound; only the unnormalized ReLU trunk keeps climbing, and the three bounded trunks, the tanh one with no normalization anywhere, flatten. Mean predictions from in distribution to the furthest regime: $4.32 \rightarrow 297.35$ for ReLU without normalization, $4.32 \rightarrow 9.64$ with LayerNorm, $4.32 \rightarrow 15.79$ for the Transformer, and $4.33 \rightarrow 7.65$ for tanh, which peaks at 11.49 and declines toward the optimal constant of 4.34. Right: the Lennard-Jones compression of §4, where the input stays bounded and every architecture flattens.

and the improvement is observedness rather than more data. Each regime's error falls by an order of magnitude in the condition that first contains it, while $s \in [0.60, 0.70)$, which no condition observes, never falls below 0.60 (Appendix C). The gain is not free: the widest condition raises in-distribution error by more than thirty times at unchanged capacity.

5 When models fail, the shape follows representational boundedness

Two incompatible descriptions of OOD behaviour coexist in the literature. Because ReLU Neural Networks are continuous piecewise-affine functions, they prolong the affine map of an unbounded outer region and diverge (Hein et al., 2019; Xu et al., 2020). Neural Networks with high-dimensional inputs instead revert toward the constant that minimizes training loss without observing the input (Kang et al., 2024). Reported on different tasks and architectures, the two have never been compared under control.

Design. A set of eight particles in the plane with target the mean of $g(r) = r^2$, permutation invariant so a set architecture is the natural choice. Trained on $r \in [1, 3]$ and evaluated out to $r \in [50, 100]$, where the true mean reaches 5822. Four models differ only in whether the trunk is bounded and whether tokens mix: DeepSets with ReLU and no normalization, the same with LayerNorm, a pre-LN Transformer, and DeepSets with tanh activations and no normalization at all.

Result. From in-distribution to the furthest regime the unnormalized ReLU Neural Network's mean prediction grows by a factor of 69 and is still climbing over the final two steps, while the LayerNorm, Transformer and tanh Neural Networks all stay within a factor of four of their in-distribution value (Figure 3); the tanh Neural Network peaks and then declines toward the optimal constant. Seed variation never approaches the $69\times$ separation. The tanh arm is the control: with no normalization it saturates anyway, so the operative property is boundedness, not normalization. Smooth one-sided activations such as GELU do not escape this: a pair of GELU units of opposite input sign sums to $x \operatorname{erf}(x/\sqrt{2})$, which grows like $|x|$ on both tails, so a GELU Neural Network that uses units of both signs is unbounded on both sides (Rocha et al., 2026; Appendix J). Measured on the same task, a GELU DeepSets with no normalization climbs by $61\times$ and is still rising over the final two steps, as the ReLU trunk does and the bounded trunks do not.

Aggregate error conceals this entirely. In the same regime the four models report normalized RMSE between 0.95 and 1.00, indistinguishable, while one continues a trend with the wrong slope and three have stopped responding to their input.

And boundedness is not a defect. On the expanding direction of §4, where $V \rightarrow 0$, saturating is the correct behaviour: the LayerNorm model is the best of the four at every expanding regime and never exceeds 0.28, while the unnormalized Neural Network ramps away from the truth, to 50.78 at

Table 3: Calibration in Instantiation I's geometry, with the ensemble architecture of this section rather than that of Table 1; mean ± sd over thirty disjoint batches, each with its own training draw and its own ensemble seeds. The invariant ensemble stays calibrated on the axis its group acts on and loses calibration only off it. Ratio is ensemble disagreement over error; the disagreement column is omitted for space and is in `run_calib30.log` in the supplementary code.

		$g(r) = r^{-2}$		$g(r) = r^2$	
regime	model	error	ratio	error	ratio
in-distribution	generic	$0.008 \pm .001$	$1.51 \pm .29$	$0.002 \pm .001$	2.45 ± 1.24
	invariant	$0.045 \pm .028$	$1.58 \pm .34$	$0.002 \pm .001$	$2.04 \pm .60$
symmetry axis	generic	$2.415 \pm .465$	$0.40 \pm .13$	$0.423 \pm .062$	$0.62 \pm .17$
	invariant	$0.045 \pm .028$	$\mathbf{1.58 \pm .34}$	$0.002 \pm .001$	$\mathbf{2.04 \pm .60}$
law axis	generic	$2.416 \pm .155$	$\mathbf{0.18 \pm .05}$	$0.235 \pm .014$	$0.12 \pm .03$
	invariant	4.353 ± 1.26	$0.53 \pm .18$	$0.272 \pm .019$	$\mathbf{0.17 \pm .04}$

the most distant (Table 2). Each shape is right in one regime and wrong in the other, which is why the regime must be identified rather than averaged over.

6 Nor does the model's own uncertainty predict competence

Section 4 rules out an external signal; the remaining candidate is an internal one, the model reporting its own incompetence when the geometry does not. We train $K{=}6$ deep ensembles (Lakshminarayanan et al., 2017), seeds differing and everything else fixed, and compare error against ensemble disagreement, both normalized by the RMS of the true signal. Their ratio is the quantity of interest: near 1 the self-report is commensurate with how wrong the model is, well below 1 it is wrong without saying so.

The self-report has the same boundary as the prediction. On the symmetry axis the invariant ensemble's calibration is unchanged from in-distribution on both laws (Table 3); the row is identical to the in-distribution row, and identical by construction.

Proposition 1. *Let $f_\theta(\mathbf{x}) = h_\theta(\|\mathbf{x}\|)$ for any h_θ. Then $f_\theta(R\mathbf{x}) = f_\theta(\mathbf{x})$ for every rotation R, so along the group's orbits the model's error, its disagreement across seeds, and every quantity computed from its output are identical to their in-distribution values, whatever h_θ learned.*

The proposition is trivial, which is the point: orbit exactness owes nothing to training. Its measured witness is the two rows agreeing to a relative 4×10^{-16}. We report the row because the generic ensemble, which must learn the same symmetry from data, loses calibration on that axis. On the law axis, where nothing was imposed, the invariant ensemble is inaccurate and falsely confident, its ratio well below one on both laws. An architectural prior does not merely fail to help off its own orbits; it fails there in a way the model cannot see.

The same boundary in a standard metric. The disagreement ratio is our own construction, so we checked that nothing rests on it. We rebuilt the ensembles as Lakshminarayanan et al. (2017) prescribe, each member predicting a mean and a variance through a Gaussian negative log-likelihood with softplus positivity and combined as a K-component mixture, and scored them by prediction-interval coverage at a nominal 90% over eight disjoint batches. In-distribution both models cover 100%, over-dispersed, as ensembles this small usually are. Across the symmetry axis the invariant model keeps full coverage while the generic model's falls below a third; on the law axis both collapse to under 5% (Table 3; the r^2 task behaves the same way, Appendix H). So the decomposition is not an artefact of the ratio: it appears in a coverage metric the uncertainty literature already uses, and more sharply, since these intervals begin too wide and still miss the truth on more than nine tenths of law-axis points.

The worst case is the generic ensemble on the law axis of the inverse-square law, error more than twice the signal at a ratio of 0.18 (Table 3). This is consensus without competence: disagreement measures variance across seeds, and members that share the same piecewise-affine asymptotes extrapolate in the same direction for the same representational reason, so their agreement certifies nothing. Predictive uncertainty is known to degrade under shift (Ovadia et al., 2019); the decomposition adds which shift, and so the answer is not all of them.

Table 4: Radial chirp: novel-versus-determined AUC of six standard metrics, oriented as a detector uses them (0.5 is chance; below 0.5 ranks the answerable queries as the less answerable). Mean $\pm$ sd over eight disjoint batches.

score	AUC
input-space kNN distance	$0.301 \pm .018$
feature-space Mahalanobis	$0.312 \pm .015$
feature-space kNN	$0.311 \pm .015$
ensemble disagreement	$0.314 \pm .011$
Gaussian-likelihood head, sd	$0.305 \pm .018$
MC dropout, sd	$0.319 \pm .015$

Table 5: Transformer on $y = \max(x) - \min(x)$ over sorted length-8 inputs: permutation-shift degradation by positional scheme, ratio of batch-mean errors over five seed batches; the ordering was predicted before the run.

scheme	in-dist.	permuted	degradation
NoPE	0.0418	0.0416	$1.0\times$
ALiBi	0.0379	0.0501	$1.3\times$
Sinusoidal	0.0052	0.4353	$83.7\times$
learned APE	0.0032	0.5789	$180.9\times$

There is also a class of error visible without any uncertainty machinery. The law $g(r) = r^{-2}$ is strictly positive, yet on the law axis the generic ensemble's mean prediction is negative for 69% of test points and the invariant ensemble's for 83%. A system with any representation of its own commitments could have caught this; none of the models here can. The strongest form of that gap follows.

Distant but determined. Take a target whose recovered law entails a constraint the data do not state: the radial chirp $f(x, y) = \sin(x^2 + y^2)$, trained on the annular wedge $r \in [1, 2.2]$, $\theta \in [0, \pi/2]$, and queried radially at $r \in [2.2, 4]$. The enumerative search of §2 recovers the symbolic expression from raw coordinates, and it entails periodicity in $u = x^2 + y^2$ with period 2π. When each query's u is folded back into the training support, the radial queries partition into those the training data already fix relative to that constraint and a novel remainder. The determined share is 56% by the geometry of the fold and $56\% \pm 1.4$ measured over the eight batches of Table 4, and the partition alternates in bands of radius rather than growing with distance. We make the notion precise.

Definition 1 (Determination relative to a class)**.** *A query x is* determined *by data D under hypothesis class $\mathcal{H}$ when every $h \in \mathcal{H}$ consistent with D takes the same value at x.*

Determination is a property of the pair $(D, \mathcal{H})$, not of x's position: the same D leaves a query open under a class of universal approximators and fixes it under the class of 2π-periodic functions of u. On the determined subset the recovered expression is exact to round-off, while the ensemble of this section is no better there than on the novel subset (4.4 against 3.9), since neither subset lies in its class. What matters for abstention is that every standard score ranks the partition backwards (Table 4): the determined queries lie farther from the training data and draw more ensemble disagreement than the novel ones (Appendix F), and the same holds for Mahalanobis and k-nearest-neighbour distance in the model's own feature space, for the predictive variance of a Gaussian-likelihood head and for Monte-Carlo dropout (Lee et al., 2018; Sun et al., 2022; Lakshminarayanan et al., 2017; Gal & Ghahramani, 2016). Oriented as a detector uses them, all six score AUC 0.30 to 0.32 for telling novel from determined, below chance in every batch, so a detector built from any of them abstains on the queries whose answers the data already fix, because the bands are placed by the law, to which none of the scores has access. Appendix F gives the setup, a Bayesian control, and a sine-activation sweep showing training loss carries the same blindness.

7 The same decomposition with no hand-built structure

The featurizer of §2 is hand-built, which invites the reading that the dissociation is an artefact of that construction; a sequence task with a Transformer, the axis fixed by a routine architectural choice, gives the same answer. The target is $y = \max(x) - \min(x)$ over length-8 sequences whose training inputs are sorted, so that $y = x_7 - x_0$ is a positional shortcut; the target is permutation-invariant, so shuffling the input is a covariate shift with the label unchanged, and whether the shortcut is representable is decided by the positional encoding alone. With no positional encoding and mean pooling the model is permutation-invariant by construction; ALiBi (Press et al., 2022) exposes relative position, sinusoidal encoding (Vaswani et al., 2017) absolute position, and a learned absolute encoding exposes it fully. Table 5 shows degradation ordered exactly by how much position the encoding exposes, from $1.0\times$, unchanged to float precision, to $180.9\times$, while length and magnitude shifts

degrade every scheme. Along the axis the featurization does not expose the model is exact at any distance; along the axes it does, it inherits. Appendix A gives the design, the batch-level inversion of the two middle rows, and a prediction of ours that failed.

8 Discussion

The results have one reading: a model is competent in a region when it computes a representation structurally equivalent to the generating law there, in the sense of Rocha et al. (2026), a correspondence between the model's operations and the mechanism's relations that holds over the whole domain by construction. This happens when the structure was imposed, as along a group's orbits, or when the law lies in the model's class and the region was observed. Every quantity that is a function of the geometry between test point and data fails to track that equivalence, in §4 by three orders of magnitude at matched distance and, as the dimension grows, with the wrong sign. Every quantity that is a function of the law and the model class does track it: whether the shift lies along an imposed group's orbits (§3, §7), how far the law leaves its training range (§4), whether the representation can continue the law's growth (§5), and whether the data determine the query under the class (§6). Exactness also survives composition: a chain of exact stages is exact, and a fitted stage contributes an error bounded above by its residual times the downstream Lipschitz constants, an upper bound that compensation between stages can beat but that the composite's class enforces, since compositions of piecewise-affine stages are piecewise-affine (Appendices K and I).

Symmetry buys exactness only along the group's orbits, so a single OOD number for an equivariant model averages a direction where it is exact with one where it is not, a property of the test distribution rather than of the model; Instantiation III shows the cost, since selecting on in-distribution validation error picks the RBF model, the worst choice outside the training range. Flagging unfamiliar inputs converts a confident wrong answer into an abstention, a gain, but §6 shows every standard flag raised in the wrong place.

9 Related work

Hein et al. (2019) and Xu et al. (2020) establish divergence along rays and Kang et al. (2024) reversion toward a constant; §5 identifies boundedness rather than dimension as what separates the two. Softmax confidence (Hendrycks & Gimpel, 2017), Mahalanobis distance (Lee et al., 2018; Ren et al., 2021), free energy (Liu et al., 2020) and feature-space neighbours (Sun et al., 2022) share the premise §4 examines, Nalisnick et al. (2019) show it can invert, and Balestriero et al. (2021) argue test points almost always fall outside the training hull, of which §4 reports the complementary case. Geometric deep learning builds symmetry into an architecture (Cohen & Welling, 2016; Bronstein et al., 2021; Batzner et al., 2022; Villar et al., 2021), and Delétang et al. (2023) find across 20,910 models that architecture class forecasts OOD generalization while data and training time do not. Appendix G extends this survey.

10 Limitations

The settings are synthetic and of modest dimension, at most 32, chosen so the competence boundary is known exactly; the Lennard-Jones instantiations use a real potential but data from its closed form. All Neural Networks are small: the Transformers of §5 and §7 are single blocks and the RBF model is the SchNet featurization rather than SchNet. The symbolic comparison is favourable by construction where the law lies in the grammar, so it is a reference floor and not a method. §4 refutes input-space distance only; in a representation that still carries the law, distance recovers the ordering.

11 Conclusion

OOD failure has no single direction. Separated, the directions show that an imposed symmetry buys exactness along its own orbits and nothing off them, that the shape of failure follows representational boundedness, that distance and dispersion can order the outcome backwards, and that abstention needs determination relative to the hypothesis class.

Reproducibility Statement

Every experiment is implemented directly over NumPy with a small hand-written reverse-mode autodiff, gradient-checked against central differences to a maximum relative error of 7×10^{-9}, so the study runs on one CPU with no deep-learning framework and no GPU. The supplementary material contains the code for every experiment and the raw per-batch arrays each table is computed from; its `README.md` maps each table to the script and saved array that produce it, and records the two conventions that matter for comparison: correlations taken over the twelve OOD regimes, and RMSE reported over the training-target sd where the signal decays toward zero. Appendix H gives architectures, optimizer settings, batch counts and seeds in full. Appendix A reports one entry that is sensitive at the second significant figure to the floating-point behaviour of the machine, with the measurement that establishes it.

The symbolic column of Table 1 is quoted as a bound for a measured reason. In Instantiation I the search returns $\mathrm{inv}(\mathrm{add}(\mathrm{sq}(x), \mathrm{sq}(y)))$, computing $1/(x^2 + y^2)$, while the target is generated as $\|\mathbf{x}\|^{-2}$; the two float paths differ on 201 of 400 points at a maximum relative difference of 3.2×10^{-16}. The least-squares residual is 3.6×10^{-32} in mean square against 2.0×10^{-33} for the closed-form coefficients, so seventeen eighteenths of what a run reports is the solver rather than the fit. In Instantiation II the ordering reverses, round-off along the candidate's own path dominating, and the two agree to within a factor of 2.5. Reproduced by `code/symbolic_floor.py`.

The calibration figures in Table 3 are means over thirty disjoint batches, each with its own training draw and ensemble seeds. Single-draw values sit up to a factor of two from the batch mean, and the ratio §6 rests on is stable where its constituents are not, $1.61 \pm .30$ at five batches against $1.58 \pm .34$ at thirty. Five batches were not enough. At five, the invariant model's in-distribution error looked two-lobed; over thirty it does not, ranging across $[0.008, 0.112]$. Within that range there is a concentration in one narrow band whose mechanism we decline to name, consistent with a collapse to a fixed function but unconfirmed at the level of hidden units: eight of the thirty batches fall in $[0.0769, 0.0843]$, a band whose coefficient of variation is 2.8% against 64% for the sample as a whole. Tested on the log scale, the appropriate one since an absolute-width criterion is satisfied too easily near the bottom of a right-skewed range, against a lognormal null, the concentration has $p = 0.004$ at the stated window size and $p = 0.003$ once the choice of window size is charged for. It is consistent with a collapse to a fixed function but unconfirmed at the level of hidden units.

Ethics Statement

This work uses only synthetic data generated from closed-form physical laws. It involves no human subjects, no personal or scraped data, and no released model or dataset. Its practical implication, that distance- and uncertainty-based competence signals can be silent exactly where a model is most wrong, argues for more caution in safety-critical deployment rather than less, and we have tried to state the boundaries of the claim precisely enough that it is not read more broadly than the evidence supports.

Use of Generative AI

The research question, the two-axis protocol, the choice of instantiations, the conclusions drawn and every claim made in this paper are the authors' own, and the authors take full responsibility for its contents. Within that, we used generative AI tools (Anthropic's Claude) throughout, for tasks the ICLR policy requires to be disclosed: implementing the experimental code and running the experiments of Sections 3 to 7 and the appendices, including the batch replication, calibration, coverage and detector studies; deriving the determined fraction reported in Appendix F and drafting the composition argument of Appendix K; assisting in the interpretation of results; and drafting sections of the text and polishing the writing. The datasets are samples of the closed-form laws stated in the paper, generated by code the tools wrote. We also used the tools for figure generation, literature retrieval and formatting. No number in the paper rests on a tool's output alone: every reported quantity was recomputed from the released code.

REFERENCES

Randall Balestriero, Jerome Pesenti, and Yann LeCun. Learning in high dimension always amounts to extrapolation, 2021. URL https://arxiv.org/abs/2110.09485.

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. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://arxiv.org/abs/2206.07697.

Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P. Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E. Smidt, and Boris Kozinsky. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications*, 13(1):2453, 2022. URL https://arxiv.org/abs/2101.03164.

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

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pp. 2990–2999, 2016.

Grégoire Delétang, Anian Ruoss, Jordi Grau-Moya, Tim Genewein, Li Kevin Wenliang, Elliot Catt, Chris Cundy, Marcus Hutter, Shane Legg, Joel Veness, and Pedro A. Ortega. Neural networks and the Chomsky hierarchy. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2207.02098.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, 2016.

Johannes Gasteiger, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs. *International Conference on Learning Representations (ICLR)*, 2020. URL https://arxiv.org/abs/2003.03123.

Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why ReLU networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 41–50, 2019. doi: 10.1109/CVPR.2019.00013.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations (ICLR)*, 2017. URL https://arxiv.org/abs/1610.02136.

John Edward Jones. On the determination of molecular fields. II. From the equation of state of a gas. *Proceedings of the Royal Society of London. Series A*, 106(738):463–477, 1924. doi: 10.1098/rspa.1924.0082.

Katie Kang, Amrith Setlur, Claire Tomlin, and Sergey Levine. Deep neural networks tend to extrapolate predictably, 2024. URL https://arxiv.org/abs/2310.00873.

Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/abs/2305.19466.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. URL https://arxiv.org/abs/1612.01474.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. URL https://arxiv.org/abs/1807.03888.

Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. URL https://arxiv.org/abs/2010.03759.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? In *International Conference on Learning Representations (ICLR)*, 2019. URL https://arxiv.org/abs/1810.09136.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. URL https://arxiv.org/abs/1906.02530.

Mircea Petrache and Shubhendu Trivedi. Approximation-generalization trade-offs under (approximate) group equivariance. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/abs/2305.17592.

Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations (ICLR)*, 2022.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006.

Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A simple fix to mahalanobis distance for improving near-OOD detection, 2021. URL https://arxiv.org/abs/2106.09022.

Filipe Marinho Rocha, Inês Dutra, Vítor Santos Costa, and Luís Paulo Reis. Exactness at inference: A representational criterion for out-of-distribution generalization. arXiv:2609.24942, 2026. URL https://arxiv.org/abs/2609.24942.

Kristof T. Schütt, Pieter-Jan Kindermans, Huziel E. Sauceda, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. URL https://arxiv.org/abs/1706.08566.

Ivan Sosnovik, Michał Szmaja, and Arnold Smeulders. Scale-equivariant steerable networks. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://arxiv.org/abs/1910.11093.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning (ICML)*, 2022. URL https://arxiv.org/abs/2204.06507.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, pp. 5998–6008, 2017.

Soledad Villar, David W. Hogg, Kate Storey-Fisher, Weichi Yao, and Ben Blum-Smith. Scalars are universal: Equivariant machine learning, structured like classical physics. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. URL https://arxiv.org/abs/2106.06610.

Keyulu Xu, Jingling Li, Mozhi Zhang, Simon S. Du, Ken-ichi Kawarabayashi, and Stefanie Jegelka. How neural networks extrapolate: From feedforward to graph neural networks. In *International Conference on Learning Representations (ICLR)*, 2020.

A A TRANSFORMER INSTANCE OF THE TWO-AXIS DECOMPOSITION

The protocol of §2 imposes structure with a hand-built invariant featurizer in two dimensions, which invites the reading that the dissociation is an artefact of that construction. This appendix runs the same measurement protocol, $K{=}6$ ensembles, normalized RMSE against normalized ensemble spread, ratio = disagreement over error, on a sequence task with a Transformer, where the axis is fixed by a routine architectural choice rather than by a symmetry we imposed.

Design. The target is $y = \max(x) - \min(x)$ over length-8 sequences of scalars, with training inputs sorted ascending, which makes $y = x_{L-1} - x_0$ exactly: a positional shortcut requiring no comparison logic. The target is permutation-invariant, so shuffling the input is pure covariate shift with the label unchanged. Whether the shortcut is representable is decided entirely by the positional encoding, giving a ladder of four: NoPE exposes nothing, so with mean pooling the model is exactly permutation-invariant and the shortcut is not representable even in principle; ALiBi exposes relative distance only; sinusoidal encoding exposes fixed absolute position; and a learned absolute encoding (APE) exposes it fully. The ordering was predicted before the experiment was run. Models are $d_{\text{model}} = 32$, 4000 iterations, five disjoint seed batches, implemented in the same NumPy autodiff as the rest of the paper and gradient-checked to 3×10^{-9}; the NoPE permutation-invariance is verified to 1.8×10^{-15}, i.e. float noise.

Table 6: Permutation-OOD degradation by positional scheme, five seed batches, ratio of batch-mean errors. Inheritance is axis-conditional and ordered by how much absolute position the encoding exposes. NoPE is unchanged by construction, not by luck.

scheme	in-distribution	permutation OOD	degradation
NoPE	0.0418	0.0416	1.0×
ALiBi	0.0379	0.0501	1.3×
Sinusoidal	0.0052	0.4353	83.7×
APE	0.0032	0.5789	180.9×

Inheritance is axis-conditional. Table 6, which repeats Table 5 of the main text, reproduces the pattern of §3 in a setting with no hand-built invariance: along the axis the featurization does not expose, the model is exact at any distance; along the axes it does, it degrades by two orders of magnitude. Length and magnitude shifts degrade every scheme, so no featurization here collapses those axes. The ordering holds on the five-batch mean and inverts in one batch of five, where sinusoidal encoding degrades by $98\times$ against APE's $83\times$; we report the mean and note the inversion rather than claiming the ordering holds uniformly. As an external check the ladder reproduces the NoPE-beats-APE length ordering of Kazemnejad et al. (2023), a known result arrived at here inside an unrelated calibration protocol.

A prediction of ours that failed. We expected the positional schemes to fail silently, every seed taking the same shortcut, so that the ensemble agrees on the same wrong answer and the ratio falls well below 1. It does not: APE's permutation-OOD ratio is 1.20, which is loud. What governs the ratio is not whether the failure is severe but whether it is shared: ALiBi's member errors correlate at 0.86 and it is the scheme that goes quiet, while APE's learned position vectors differ per seed, so its failure is idiosyncratic and the spread exposes it. Across all 160 cells the rank correlation between member-error correlation and the calibration ratio is -0.82, against -0.60 for error magnitude as a confound. This does not rescue disagreement as a competence signal; it identifies when disagreement happens to work, which is when different seeds fail differently, a property of the training procedure rather than of the model's relation to the law.

One entry is environment-sensitive, and we can say why. An independent re-run on different hardware reproduced NoPE, ALiBi and APE to under 1% and the sinusoidal permutation-OOD entry to 7.5%. The pipeline is seed-determined: the same cell is bit-identical across processes and across BLAS thread counts at every checkpoint from 1 to 4000 iterations. The gap is amplification, not nondeterminism. Perturbing one initial weight by a single unit in the last place (2.8×10^{-17}) and training as normal moves the sinusoidal member's permutation-OOD prediction by 5.6×10^{-2} and its error by 1.54%, while the identical perturbation moves NoPE by 3.2×10^{-13}, APE by 2.0×10^{-8} and ALiBi by 1.8×10^{-15}, none of the three changing the reported error in the fourth decimal. Under sinusoidal encoding this task sits in a basin where last-bit arithmetic differences between two environments, a different BLAS kernel is enough, reach the second significant figure.

The ordering, the refutation and both rank correlations are unaffected; the sinusoidal entry should be read to one significant figure. This is the appendix-scale version of the residual-floor argument of the Reproducibility Statement: part of what a run reports is a property of the machine that ran it, and the remedy is to measure how much rather than to assume it is negligible.

Scope. This is one task family and a small model; it answers whether the axis-conditional pattern survives a change of architecture and featurization, not whether it holds at scale. The sorted-input design plants the shortcut deliberately, which is what makes the ladder interpretable and which a reader may reasonably read as constructing the result; the defence is that the ordering was predicted in advance and the silent-failure prediction was refuted, which is not the signature of a rigged setup.

B BOOTSTRAP INTERVALS FOR THE COMPETENCE-SCORE CORRELATIONS

A method measuring distance in the raw input, or in a representation not known to encode the law, would on the matched pair of §4 have flagged the benign direction and passed the catastrophic one. That is the sense in which the scores below fail: not that they are uninformative, but that the ordering they produce is the wrong one where it matters.

Rank correlations over the twelve OOD regimes of §4, with 95% intervals from 4000 bootstrap resamples of the regimes and the fraction of draws preserving the point estimate's sign. Scores are averaged over five seeds before the correlation is taken.

Table 7: Point estimate, bootstrap interval, and sign stability. The input-distance column is the one §4 declines to rest anything on: every interval contains zero. The two feature-space columns that work do not have that problem, and the $\tanh$ row shows the failure is not noise-limited but representational.

model	input NN	Mahalanobis	rel. Mahalanobis	feature kNN
ReLU, no norm	$+0.03\ [-0.45, 0.61]\ 55\%$	$0.58\ [0.05, 0.86]\ 98\%$	$0.83\ [0.43, 0.97]\ 100\%$	$0.98\ [0.87, 1.00]\ 100\%$
LayerNorm	$-0.48\ [-0.66, 0.11]\ 96\%$	$0.97\ [0.78, 1.00]\ 100\%$	$-0.66\ [-0.86, -0.18]\ 99\%$	$0.92\ [0.63, 1.00]\ 100\%$
Transformer	$-0.43\ [-0.56, 0.14]\ 94\%$	$0.82\ [0.34, 1.00]\ 100\%$	$-0.40\ [-0.82, 0.22]\ 89\%$	$0.96\ [0.74, 1.00]\ 100\%$
$\tanh$, no norm	$-0.43\ [-0.56, 0.14]\ 94\%$	$-0.23\ [-0.55, 0.43]\ 77\%$	$-0.44\ [-0.59, 0.13]\ 94\%$	$-0.51\ [-0.57, 0.13]\ 96\%$

Two details that did not fit in §4. First, the background-corrected Mahalanobis score of Ren et al. (2021) does not transfer to this setting: it ranges from $+0.83$ to -0.66 across the four architectures and is the only score whose sign depends on which model it is attached to, with seed-to-seed spread up to ± 0.38 against ± 0.04 for the uncorrected version. It was designed for near-OOD detection in classification, where the background fit removes a nuisance direction; here it removes signal. Second, the input-distance column reproduces the values §4 reports to three decimals, because it is computed on the same trained models; it is a consistency check on the harness rather than an independent replication.

C THE OBSERVEDNESS ABLATION IN FULL

Normalized RMSE by training range, five seeds, $N_{\text{train}} = 3000$ in every condition so that the widest range has the lowest sample density. Bracketed entries are regimes inside that condition's training range. The diagonal is the result: each regime drops by an order of magnitude in the condition that first observes it. The first column is the control: no condition observes $s \in [0.60, 0.70)$, and it never reaches the level the observed regimes do.

Two honest qualifications. The never-observed column does improve as the range widens, from 0.997 to 0.595, because $[0.60, 0.70)$ is adjacent to a region that becomes observed and the law is smooth across the boundary; the claim is that it does not approach the order-of-magnitude drop the observed regimes show, not that it is unaffected. And the rightmost two columns move the wrong way, sharply: a training range whose targets span 0 to ~ 255 forces a fixed target normalization and a fixed capacity to serve the repulsive wall, and the nearly flat region is what gets sacrificed. Observedness is not free, and the trade-off is itself a statement about capacity allocation rather than about distance.

Table 8: Widening the training range, at fixed N. Bracketed = inside the training range.

condition	$[0.60, 0.70)$	$[0.70, 0.80)$	$[0.85, 0.90)$	$[1.00, 1.50)$	expand $[1.5, 1.7)$
DeepSets, ReLU, no norm					
A train $[1.00, 1.5]$	0.997	0.984	0.901	[0.049]	0.742
B train $[0.85, 1.5]$	0.943	0.753	[0.068]	[0.166]	1.080
C train $[0.70, 1.5]$	0.595	[0.056]	[0.139]	[1.645]	10.756
DeepSets + LayerNorm					
A train $[1.00, 1.5]$	0.999	0.993	0.936	[0.042]	0.082
B train $[0.85, 1.5]$	0.968	0.810	[0.093]	[0.139]	0.269
C train $[0.70, 1.5]$	0.654	[0.057]	[0.142]	[1.199]	3.377

D THE MATCHED PAIR ACROSS LAW FAMILIES

The geometry, the four-particle tetrahedron, the architectures, the optimizer and the training range $s \in [1.0, 1.5]$ are all held fixed; only the law changes. Each law shares the structure the argument depends on, steep just below the training range and flat far above it, and nothing else. Born–Mayer is included because it is not a power law, so the result cannot be attributed to an exponent. Errors are RMSE over the sd of the training targets, the fixed scale §4 uses, since normalized RMSE measures its own denominator wherever the signal decays toward zero. Five seeds each.

Table 9: The matched pair of §4 under four laws. Compressing to $s \in [0.70, 0.80)$ sits at nearest-neighbour distance 0.568; expanding to $s \in [1.5, 1.7)$ sits at 0.600, slightly further from the training data. The nearer set is the worse one in every row.

law	model	err. compress	err. expand	ratio	ρ(dist, log err)
Lennard-Jones $4(s^{-12} - s^{-6})$	ReLU, no norm	562.4	0.759	741×	+0.22
(control, re-run here)	LayerNorm	557.6	0.082	6808×	−0.36
Coulomb $1/s$	ReLU, no norm	1.52	0.136	11×	+0.89
	LayerNorm	1.64	0.066	25×	+0.68
Inverse cube $1/s^3$	ReLU, no norm	4.38	0.179	24×	+0.71
	LayerNorm	4.82	0.032	149×	−0.01
Born–Mayer $e^{-(s-1)/0.15}$	ReLU, no norm	12.9	0.146	88×	+0.40
	LayerNorm	14.9	0.089	168×	+0.07

Two things separate cleanly. The sign of the effect is a property of the geometry and holds in all eight rows: the compressed set is nearer and worse. The magnitude is a property of the law, spanning $11\times$ to $6808\times$ and ordered by how steeply the target climbs below the training range, which is what §4 claims sets the error, now varied deliberately rather than observed once.

The rank correlation behaves differently, and the difference is instructive. It is strongly negative for Lennard-Jones, near zero for the intermediate laws, and strongly positive for Coulomb, where the mild $1/s$ climb leaves so little room to be wrong on the compressing side that the expanding regimes dominate the ranking. A single number summarising twelve experimenter-chosen conditions is therefore sensitive to a choice the matched comparison is immune to, which is why §4 declines to rest on it. The Lennard-Jones row also reproduces the $744\times$ and $6787\times$ of §4 to within rounding, from a separate script, which is the consistency check that makes the other rows readable.

E THE MATCHED PAIR ACROSS INPUT DIMENSIONS

§4 measures the matched pair on the four-particle tetrahedron, a twelve-dimensional input whose training manifold has four intrinsic degrees of freedom, the edge s and a rotation. Nearest-neighbour distances concentrate as the dimension grows, so a result about distance on such a manifold could be read as specific to it. This appendix holds the law fixed and sweeps the dimension. The target is the Lennard-Jones potential $V(r) = 4(r^{-12} - r^{-6})$ on the norm $r = \|x\|$ of a single point $x \in \mathbb{R}^d$, $d \in \{2, 4, 8, 16, 32\}$, with directions uniform on the sphere and training radii $r \in [1.0, 1.5]$, so the symmetry group is $O(d)$, the training set covers every orbit, and the only way out of it is radial: the

two directions of §4, in the same twelve shells, with nothing else varying. Three Neural Networks are trained: the generic ReLU Neural Network on x, a $\tanh$ Neural Network on x, and the invariant Neural Network on r, whose problem is the same one-dimensional regression at every d and serves as the control. All three are the two-hidden-layer Neural Network of §2, with 64 units per layer. The training set has $2000d$ points, linear in the dimension, so that the generic Neural Networks are not starved of data as d grows, and training is minibatch Adam, 256 points per step, 8000 steps, cosine schedule from 3×10^{-3}, so that its cost does not. Inputs to the generic Neural Networks are standardised per coordinate on training statistics; distances are computed on the raw inputs, whose radii are the same at every d, against the full training set, and are reported relative to the in-distribution baseline, the same quantity for 1200 fresh in-range samples. Eight disjoint batches per d, each with its own training draw, model seeds and test draws; errors are RMSE over the sd of the training targets, as in §4.

Table 10: The two directions across input dimension: $V(r)$ on $r = \|x\|$, $x \in \mathbb{R}^d$, training $r \in [1.0, 1.5]$ in every direction, eight batches per d (the summary of Tables 11 and 12). Distance is the mean nearest-neighbour distance to the training set relative to that of fresh in-distribution samples, so 1 is as near as the training data is to itself. Errors are RMSE over the training-target standard deviation for the generic ReLU Neural Network on x; the invariant Neural Network on r has in-distribution error 0.065 to 0.076 at every d. The ratio is of batch means for the pair of §4; ρ is Spearman's correlation of distance with log error over the twelve shells.

	in-dist.	compress $[0.70, 0.80)$		expand $[1.5, 1.7)$			compress $[0.60, 0.70)$		expand $[5, 8)$		
d	error	dist.	error	dist.	error	ratio	dist.	error	dist.	error	ρ
2	0.075	16.2	545	7.3	0.65	844×	22.5	3671	316	49	+0.23
4	0.209	2.22	548	1.76	0.74	740×	2.85	3671	36	46	+0.11
8	0.420	1.09	553	1.33	0.81	682×	1.22	3691	12	53	−0.06
16	0.520	0.88	558	1.23	0.69	807×	0.91	3643	7.4	47	−0.32
32	0.564	0.82	556	1.21	0.63	888×	0.81	3670	5.7	43	−0.55

Table 11: Every shell at every dimension. Left: mean nearest-neighbour distance to the training set relative to the in-distribution baseline (raw baselines 0.016, 0.139, 0.418, 0.735 and 0.992 for $d = 2$ to 32). Right: error of the generic ReLU Neural Network on x. Batch sds are below 10% of the mean for every compressed entry and below 23% for every expanded one.

	distance / baseline					error, ReLU on x				
regime	d=2	4	8	16	32	d=2	4	8	16	32
in-distribution	1.00	0.99	1.00	1.00	1.00	0.075	0.209	0.420	0.520	0.564
compress $r \in [0.95, 1.0)$	2.58	1.07	0.92	0.88	0.87	2.95	3.70	4.87	5.26	5.45
compress $r \in [0.90, 0.95)$	5.47	1.27	0.94	0.87	0.86	13.7	15.1	16.9	17.7	18.0
compress $r \in [0.85, 0.90)$	8.49	1.51	0.97	0.87	0.84	42.9	44.9	47.7	48.2	49.0
compress $r \in [0.80, 0.85)$	11.6	1.78	1.01	0.87	0.83	115	117	120	121	122
compress $r \in [0.70, 0.80)$	16.2	2.22	1.09	0.88	0.82	545	548	553	558	556
compress $r \in [0.60, 0.70)$	22.5	2.85	1.22	0.91	0.81	3671	3671	3691	3643	3670
expand $r \in [1.5, 1.7)$	7.27	1.76	1.33	1.23	1.21	0.645	0.740	0.811	0.691	0.626
expand $r \in [1.7, 2.0)$	22.6	3.25	1.74	1.46	1.39	2.36	2.40	2.64	2.28	2.03
expand $r \in [2.0, 2.5)$	47.8	5.93	2.52	1.88	1.71	5.98	5.75	6.43	5.65	5.13
expand $r \in [2.5, 3.5)$	94.9	11.2	4.19	2.78	2.36	13.5	12.7	14.4	12.8	11.6
expand $r \in [3.5, 5.0)$	174	20.2	7.07	4.38	3.53	26.1	24.4	27.7	24.7	22.7
expand $r \in [5.0, 8.0)$	316	36.3	12.4	7.38	5.73	49.2	45.6	52.7	46.9	43.0

Four things are visible in Tables 11 and 12. First, the distance scale collapses. The raw in-distribution baseline grows from 0.016 at $d = 2$ to 0.992 at $d = 32$, and relative to it the span between the nearest and the furthest shell shrinks from a factor of about 120 to a factor of 7. The compressed shells cross the baseline between $d = 8$, where three of the six are already nearer to the training set than in-distribution samples are, and $d = 16$, where all six are, the catastrophic $r \in [0.60, 0.70)$ included; the expanded shells remain further at every d. In the geometry a distance-based detector sees, the direction along which the target climbs by three orders of magnitude is, from $d = 16$ on, the most in-distribution place in the test set. Second, the compressed errors are

Table 12: The three Neural Networks at every dimension. The pair is that of §4, compress $r \in [0.70, 0.80)$ against expand $r \in [1.5, 1.7)$; the ratio is of batch means, followed by the within-batch range; nearest and furthest are the shells $r \in [0.60, 0.70)$ and $r \in [5, 8)$; ρ is Spearman's correlation of distance with log error over the twelve shells, mean over batches.

Neural Network	d	in-dist.	compress	expand	ratio	range	nearest	furthest	ρ
ReLU on x	2	0.075 ± 0.015	545	0.65	844×	[574, 1151]	3671	49.2	+0.23
	4	0.209 ± 0.017	548	0.74	740×	[652, 891]	3671	45.6	+0.11
	8	0.420 ± 0.020	553	0.81	682×	[606, 739]	3691	52.7	−0.06
	16	0.520 ± 0.019	558	0.69	807×	[773, 839]	3643	46.9	−0.32
	32	0.564 ± 0.021	556	0.63	888×	[840, 918]	3670	43.0	−0.55
$\tanh$ on x	2	0.145 ± 0.014	553	0.37	1475×	[1326, 1779]	3683	1.7	−0.23
	4	0.512 ± 0.030	557	0.67	831×	[743, 885]	3683	3.0	−0.32
	8	0.771 ± 0.015	561	0.47	1188×	[1006, 1246]	3700	7.2	−0.48
	16	0.833 ± 0.019	565	0.70	810×	[754, 890]	3652	4.8	−0.61
	32	0.865 ± 0.022	563	0.81	695×	[667, 721]	3678	3.7	−0.77
invariant on r	2	0.070 ± 0.013	552	0.88	629×	[464, 2626]	3681	48.8	+0.22
	4	0.076 ± 0.014	554	1.02	542×	[343, 1388]	3680	56.5	+0.15
	8	0.065 ± 0.005	554	0.72	773×	[449, 4968]	3692	44.2	−0.10
	16	0.070 ± 0.007	557	0.85	657×	[431, 8064]	3643	49.3	−0.27
	32	0.068 ± 0.007	553	0.73	756×	[507, 10431]	3668	43.5	−0.53

the same for all three Neural Networks to within a few percent, 3643 to 3700 on the nearest shell, because none of them has any information about the repulsive wall and each returns something of the order of its training range where the target lies 1100 to 7400 training standard deviations above it: the error there is the law's excursion, whatever the model. Third, the expanded branch separates the Neural Networks the way §5 predicts. The ReLU Neural Network on x and the invariant Neural Network on r continue a trend, reaching 43 to 56 on the furthest shell, and the $\tanh$ Neural Network saturates, at 1.7 to 7.2. Fourth, the sign never changes. For the pair of §4 the compressed shell is the worse one in all 120 Neural Network–dimension–batch combinations, and from $d = 8$ on it is also the nearer one, the pair's relative distances moving from 16.2 against 7.3 at $d = 2$ to 0.82 against 1.21 at $d = 32$ while the ReLU Neural Network's ratio stays between 682× and 888×. Where that pair is not matched in distance, at $d = 2$ and $d = 4$, a pair chosen on distance alone is: compress $r \in [0.60, 0.70)$ against expand $r \in [1.7, 2.0)$ at $d = 2$, relative distances 22.5 and 22.6, ratio 1553×, and compress $r \in [0.85, 0.90)$ against expand $r \in [1.5, 1.7)$ at $d = 4$, 1.51 and 1.76, ratio 61×, with the nearer shell the worse one in every batch.

Two costs are reported rather than hidden. The generic Neural Networks' in-distribution error rises with d, from 0.075 to 0.564 for the ReLU Neural Network and from 0.145 to 0.865 for the $\tanh$ Neural Network, on training sets that grow linearly with d; the invariant Neural Network's stays between 0.065 and 0.076. The norm has to be learned from samples and that gets harder with the dimension, while the invariant Neural Network's problem does not change. Larger training sets and a wider Neural Network reduce the generic error at $d = 32$ but do not remove it, 0.44 in one check at 256,000 points and 128 units, and we kept the protocol uniform across d rather than tune it per dimension. The $\tanh$ Neural Network fits the training range least well under the optimizer settings shared by all three, which is why its in-distribution column is the weakest; its compressed errors are indistinguishable from the others and its expanded errors are the saturating ones, so we report it for the shape of its failure rather than as a competitive baseline. Finally, the rank correlation of distance with log error, which §4 declines to rest on, turns from +0.23 to −0.55 over the sweep for the ReLU Neural Network: as the dimension grows, distance stops predicting the outcome and begins to predict its opposite.

F Distant but determined: a radial chirp

Sections 4 and 6 are negative results: an external geometric signal and an internal disagreement signal each fail to order the error. This appendix constructs the strongest form of the failure, a

setting in which a subset of distant queries is provably answerable from the training data alone, and both signals rank that subset backwards.

Setup. The target is a radial chirp $f(x, y) = \sin(x^2 + y^2)$, trained on 1200 points from the annular wedge $r \in [1, 2.2]$, $\theta \in [0, \pi/2]$, and probed on 2000 fresh points per regime (in-distribution; angular, $\theta \in [\pi/2, 2\pi]$ at trained radii; radial, $r \in [2.2, 4]$ on the trained wedge), the two-axis geometry of §2 with a law that oscillates. The same bottom-up enumeration as §2, given only the raw coordinates, selects $\sin(\mathrm{add}(\mathrm{sq}(x), \mathrm{sq}(y)))$ with coefficient one and intercept zero from 6.1×10^5 candidates, at residual 8.8×10^{-33}. Two global constraints follow from the recovered expression, neither supplied by us: the range bound $|f| \leq 1$, and periodicity in $u = x^2 + y^2$ with period 2π.

The partition. Training covers $u \in [1, 4.84]$, which is 0.61 of a period by deliberate choice. Folding a query's u back by $u \mapsto (u - 1) \bmod 2\pi + 1$ lands 56.3% of the radial extrapolation region inside the training support (Figure 4): on that subset the training data already fixes the answer relative to the recovered periodicity constraint, and evaluating the recovered expression there gives error 1.3×10^{-16}. The remaining 43.7% is genuinely novel, in the sense that training data plus the constraint do not determine it. Membership is fixed by radius, but it alternates in bands, so no distance threshold recovers the partition.

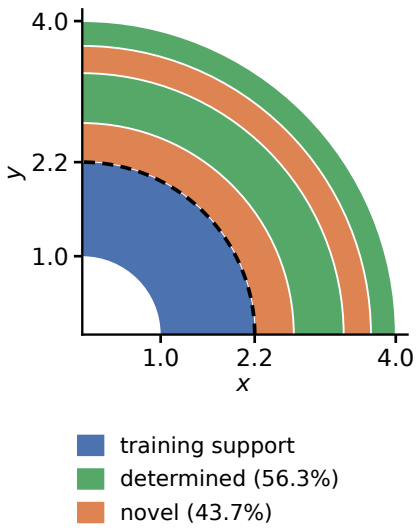


Figure 4: Training annulus ($r \in [1, 2.2]$) and radial query region ($r \in [2.2, 4]$) on the same wedge. Periodicity in $u = x^2 + y^2$ folds alternating query bands back inside the training u-support (determined, 56.3% in aggregate) or leaves them outside it (novel, 43.7%). The partition is not monotone in distance from the training boundary (dashed).

Table 13: Radial queries by periodicity partition. The ensemble of §6's architecture fails equally on both subsets and is falsely confident on both; a nearest-neighbour distance detector abstains on both, which is a false abstention on the determined 56.3%. "N/A" marks what training data alone can certify, not where the recovered expression could be evaluated.

radial subset	ensemble err.	ratio	symbolic err.	distance detector
determined (56.3%)	4.43	0.23	1.3×10^{-16}	abstains (falsely)
novel (43.7%)	4.21	0.24	N/A	abstains (correctly)

Both signals point the wrong way. Table 4 in the main text replicates the finding over eight disjoint batches and extends it to feature-space Mahalanobis and kNN distance, a Gaussian-likelihood head and Monte-Carlo dropout, all of which score AUC 0.30 to 0.32; the single draw reported here is the one the rest of this appendix is computed on. The determined queries sit farther from the training data on average (mean nearest-neighbour distance 1.16 against 0.82) and draw more ensemble

disagreement (mean 0.78 against 0.57): as classifiers of determinedness, distance and disagreement score AUC 0.69 and 0.67 in the wrong direction, equivalently 0.31 and 0.33 in the direction a detector would use them. These are moderate effect sizes pointed backwards, not chance, because the periodic bands are placed by the law, to which neither signal has access. A distance-aware detector therefore abstains precisely on queries whose answers the data already fixes.

What kind of model carries the signal. The Neural Networks have no internal access to any of this. A sine-activation sweep (three seeds, two initialization scales) improves training loss by one to two orders of magnitude over ReLU (2.5–46×10^{-6} against 2.9×10^{-4}) with no extrapolation gain, radial error 1.10–1.31 against 2.33: a low loss is a good fit whether or not the law was recovered, and certifies nothing. The enumerative search separates the two cases from the inside, on training data alone. When $\sin$ is in the operator set, the winner separates from the best behaviorally distinct runner-up by roughly thirty orders of magnitude in residual; removing $\sin$, an ablation of a capability the answer needs, leaves the top candidates within 4% of one another at residual 2.2×10^{-2}, returned with no warning. The separation is a property of commitment, not of discreteness: a continuous three-parameter family $A \sin(\omega u + \varphi)$, scored by the same description-length criterion against the second-best mode of its likelihood surface, earns a gap on the same scale as the discrete search's (up to 71.0 nats per point when the grid contains $\omega = 1$, against the search's 70), and the gap is stable under grid refinement, while a naive nearest-candidate rule collapses with resolution because in a continuum the nearest candidate is an artifact of the sampling. A model that blends hypotheses by unconstrained gradient descent obtains no gap under either rule.

A Bayesian control. A Gaussian process with a periodic kernel on $u = x^2 + y^2$, its period fit by marginal likelihood rather than fixed, selects 6.2832, which is 2π to five decimals, in four of nine restarts (the other five hit a spurious boundary optimum), and given the period its error sits at the numerical floor in every regime, the determined and novel subsets included (Rasmussen & Williams, 2006). The control worth reporting is the 2×2 separating the kernel from the coordinate: a Matérn kernel on u alone captures 81% of the total marginal-likelihood gap over a generic GP on raw (x, y), yet fails radially at error 1.439, indistinguishable from the generic model's 1.441; only the periodic kernel buys the law axis. Evidence and competence decompose differently, and the internal evidence signal is the less discriminating of the two. The form of the structure, periodicity in u, was supplied by us; only its parameter was found, whereas the enumerative search recovers the constraint itself from raw coordinates.

Table 14: Kernel-versus-coordinate controls on the chirp, identical 400-point training subsample; LML is log marginal likelihood, errors are normalized RMSE.

kernel	coordinate	LML (nats)	in-dist.	angular	radial
Matérn	raw (x, y)	1323.0	0.0043	0.8296	1.4412
Matérn	$u = x^2 + y^2$	2897.4	0.0001	0.0000	1.4393
ExpSineSquared	raw (x, y)	463.9	0.0171	1.7334	1.1242
ExpSineSquared	$u = x^2 + y^2$	3261.1	0.0000	0.0000	0.0000

G Extended related work

Extrapolation behaviour. Hein et al. (2019) show ReLU Neural Networks give arbitrarily confident predictions far from data, and Xu et al. (2020) prove convergence to linear functions along any ray, with the conditional that linear targets are learned given a diverse training distribution. Kang et al. (2024) report reversion toward the optimal constant for high-dimensional inputs. We are not aware of a controlled comparison of the two on a common task; §5 identifies representational boundedness, not input dimension, as the separator.

Interpolation and the convex hull. Balestriero et al. (2021) argue that in high dimension test points almost always fall outside the training hull, so learning is essentially always extrapolation. §4 reports the complementary case, which their framing predicts should be benign and is not: points 100% inside the hull with error equal to the signal.

Post-hoc competence scores. Maximum softmax probability (Hendrycks & Gimpel, 2017), feature-space Mahalanobis distance (Lee et al., 2018) and its background-corrected form (Ren et al., 2021),

logit free energy (Liu et al., 2020) and feature-space nearest neighbours (Sun et al., 2022) all share the premise §4 examines, that competence is recoverable from where a test point sits relative to the training distribution. Nalisnick et al. (2019) show it can invert outright: generative models assign higher likelihood to OOD inputs than to their own training data. We add a regression setting where the inversion has an identified cause, and where the feature-space members recover the ordering precisely when the representation they are defined on still carries the law.

Equivariance and its limits. Geometric deep learning gives a principled account of building symmetry into an architecture (Cohen & Welling, 2016; Bronstein et al., 2021) with strong empirical gains (Batzner et al., 2022). Villar et al. (2021) prove the scalar construction universal, and Petrache & Trivedi (2023) analyse approximate and misspecified symmetry. Our contribution is not that equivariance has limits but a protocol that localizes them to directions, surviving at physical realism.

Architecture class and generalization. Delétang et al. (2023) show across 20,910 models that architecture class forecasts OOD generalization while data and training time do not; we add the directional decomposition their setting does not require.

H Reproducibility details

All experiments are implemented directly over NumPy with a small reverse-mode autodiff, gradient-checked against central differences to a maximum relative error of 7×10^{-9} through attention, softmax and LayerNorm, so the whole study runs without a deep learning framework. Neural Networks in §5 and §4 use $d_{\text{model}} = 64$, a feed-forward width of 128, four attention heads where applicable, full-batch Adam at $\eta = 2 \times 10^{-3}$ with gradients clipped at ± 5, batch 256 and 1500 steps, on 3000 training configurations, with five seeds per model and 1200 to 1500 fresh samples per regime. Inputs are scaled by the training maximum and targets standardized on training statistics only. Instantiations I to III follow the architecture and optimizer of §2 with two hidden layers of 64 units, on 1200 training points for I and II and 1800 for III; §6 uses two hidden layers of 48 units, $K{=}6$ members and 1200 training points. The dimension sweep of Table 10 trains the same two-hidden-layer Neural Network of 64 units on $2000d$ points with minibatch Adam, 256 points per step, 8000 steps and a cosine schedule from 3×10^{-3}, with eight batches per dimension (Appendix E).

Every table in this paper reports mean $\pm$ sd over disjoint batches, each with its own training draw and its own model seeds: thirty batches for Table 3, twenty-four for Table 1, eight for the coverage study, eight per dimension for Table 10, and five seeds per model for Table 2 and §§4–5, where the reported contrasts are also given as within-seed ranges. No number in the paper comes from a single run. The prediction-interval check of §6 (Gaussian-likelihood ensembles, nominal 90%, eight batches, $g(r) = r^{-2}$) covers $100.0\% \pm 0.0$ for both models in distribution, $100.0\% \pm 0.0$ (invariant) against $31.5\% \pm 11.6$ (generic) across the symmetry axis, and $4.0\% \pm 0.6$ (generic) and $4.9\% \pm 3.6$ (invariant) on the law axis; on $g(r) = r^2$ the corresponding values are $100.0\% \pm 0.0$, $100.0\% \pm 0.0$ against $75.7\% \pm 12.4$, and $6.9\% \pm 1.3$ against $9.1\% \pm 2.3$ (`run_nll.log`).

We went to thirty batches in §6 after five proved able to mislead. The invariant model's in-distribution error looked two-lobed at five; an independent reimplementation of the same protocol saw the same appearance with the lobes at different levels and the membership reversed, which is how a five-point sample of a broad distribution behaves. Thirty batches rule out two lobes and establish something narrower in their place, reported in §6. Two negative results on our own method are worth recording, since both initially pointed the wrong way. The 1-versus-2-component BIC comparison we first used is close to blind to a tight cluster embedded in a diffuse background, because a contiguous two-way split cannot represent a spike with spread on both sides of it; and a window-width test on the raw scale is drawn to the bottom of a right-skewed range, where at $k = 8$ it returns a band of 20% coefficient of variation in preference to one of 2.8% at identical absolute width. Only the log-scale version of that test sees the structure. The reimplementation supplies two external checks: on the five-batch setting it agrees with ours on all eighteen reported quantities within the sum of the two sds, and on twenty-four batches of its own it finds the same concentration in the same place, $[0.0767, 0.0795]$ against our $[0.0769, 0.0843]$.

The symbolic column of Table 1 is a bound rather than a value. The Reproducibility Statement gives the measurement behind that choice, and `symbolic_floor.py` in the supplement reproduces it in a few seconds: the residual of an exactly recovered expression is float64 round-off whose

magnitude is set partly by the solver and partly by which floating-point path the target was generated along, so its digits are not a property of the model.

The supplementary material contains the code for every experiment and the raw per-batch outputs each table is computed from, so every reported number can be re-derived without retraining. Its `README.md` maps each table to the script and the saved array that produce it.

I THE CHEBYSHEV FLOOR OF A PIECEWISE-AFFINE FIT

A fitted piecewise-affine surrogate leaves a residual even inside the training range when the target is curved (Rocha et al., 2026). Let f be twice continuously differentiable and let a cell of the partition have width h. By the Chebyshev equioscillation theorem, the best uniform affine approximation of f on the cell is the chord displaced by half its maximum deviation, and its error equioscillates at three points. If f'' does not vanish on the cell, that error lies between $h^2 \min|f''|/16$ and $h^2 \max|f''|/16$, the extremes taken over the cell, with equality when f'' is constant. A surrogate that interpolates the training samples exactly therefore still carries a curved residual between them. Refining the partition lowers the floor as h^2 but never to zero, so a piecewise-affine surrogate is not structurally equivalent to a curved target; it suppresses the residual inside the training range, and outside it the divergence of §5 takes over.

J DIVERGENCE OF GELU NEURAL NETWORKS ON BOTH TAILS

Let $\mathrm{GELU}(x) = x\Phi(x)$, with Φ the standard normal cumulative distribution function. A single unit tends to 0 as $x \to -\infty$ and grows like x as $x \to +\infty$, so it is bounded on one tail. A Neural Network built from such units need not be (Rocha et al., 2026). For two units with input weights $+1$ and -1 and output weights 1,

$$\mathrm{GELU}(x) + \mathrm{GELU}(-x) = x\Phi(x) - x\Phi(-x) = x\,(2\Phi(x) - 1) = x\,\mathrm{erf}\Big(x/\sqrt{2}\Big)\,,$$

and since $\mathrm{erf}(z) \to \pm 1$ as $z \to \pm\infty$ the sum grows like $|x|$ on both tails. A fit to an even target, or to any target that rises on both sides of the training range, typically uses units of both input signs, and the resulting Neural Network is then unbounded on both tails. Boundedness is a property of the trained representation, not of the activation alone.

Measured on the task of §5, with the DeepSets trunk of the ReLU model, every ReLU replaced by GELU, no normalization, and the same data, optimizer, capacity budget and five seeds, the mean prediction from in distribution to the furthest regime goes from 4.33 to 264.50, a factor of 61 (per seed 53 to 66), still climbing by $4.9\times$ over the final two steps; the ReLU trunk rerun alongside it goes from 4.32 to 297.24, a factor of 69, reproducing §5. On the Lennard-Jones compression, where the input stays bounded, the two behave alike again, reaching 6.35 and 3.12 against a true mean of 773 and both rising by $1.9\times$ over the final two steps. The supplementary code reproduces both runs (`regime_gelu.py`); note that the test sets of §5 are seeded through Python's string hash, so `PYTHONHASHSEED` must be fixed for a bit-identical rerun.

K THE INFERENCE-PATH PROPAGATION BOUND

Let $g_1, \ldots, g_n$ be the exact stages of a pipeline and $\hat{g}_1, \ldots, \hat{g}_n$ the fitted stages standing in for them, with $\varepsilon_k \geq 0$ and $L_k \geq 0$ such that

$$\big|\hat{g}_k(u) - g_k(v)\big| \;\leq\; \varepsilon_k + L_k\,|u - v| \qquad \text{for all } u, v.$$

Writing $x_k = g_k(x_{k-1})$ and $\hat{x}_k = \hat{g}_k(\hat{x}_{k-1})$ from a common input $x_0 = \hat{x}_0$, the error $e_k = |\hat{x}_k - x_k|$ satisfies $e_k \leq \varepsilon_k + L_k\, e_{k-1}$, which unrolls to

$$e_n \;\leq\; \sum_{k \leq n} \varepsilon_k \prod_{j > k} L_j$$

(Rocha et al., 2026). If every stage on the path of a direction is exact, as along the symmetry axis where the invariant projection removes the orbit coordinate, every ε_k is zero and so is the bound,

whatever the L_k. If a fitted Neural Network lies on the path, as along the law axis, it contributes a term $\varepsilon_k > 0$ to the bound, and when the downstream constants satisfy $L_j \geq c > 1$ that term grows geometrically with the number of stages that follow. The bound is an upper bound and not a lower one: two fitted stages can cancel each other's errors, as $\hat{g}_1(x) = x + \varepsilon$ followed by $\hat{g}_2(z) = z - \varepsilon$ does, and a downstream stage that is insensitive to a discrepancy annihilates it, so a nonzero ε_k does not by itself force a nonzero e_n. What forces it in the settings of this paper is the class of the composite: a composition of continuous piecewise-affine stages is continuous piecewise-affine, so no arrangement of fitted ReLU stages realizes a curved law exactly on an interval, whatever compensation training finds among them, and the same closure argument applies to any class closed under composition that excludes the target.