WHEN GEOMETRY FLIPS ARCHITECTURE RANKINGS: AN EXACT THEORY OF ALIGNMENT AND COMPLEXITY

Anonymous authors
Paper under double-blind review

ABSTRACT

A growing family of model-selection heuristics scores an architecture by how well its inductive bias matches the geometry of the data. We ask when such a score can reverse: given two architectures and a group acting on input coordinates, when does scrambling the geometry flip which one generalises better? Working in a Gaussian sequence idealization in which an architecture acts as diagonal shrinkage in the Fourier basis of the input geometry, we give four results. First, averaging the data over a permutation group makes the covariance distance-independent, so the population fit of any exponential-decay locality score collapses to its floor, and when the mean off-diagonal covariance vanishes the score becomes *unidentifiable* rather than small, a failure mode we observe in 82% of scrambled fits on natural image patches. Second, the risk gap between two architectures splits exactly into a geometry-free complexity gap M and an alignment gain S equal to d times the covariance between the architectures' squared-bias profiles and the data spectrum; preference inverts under scrambling precisely when S opposes and dominates M. Third, in a two-band model the inversion boundary admits an exact closed form whose minimum is $2\sqrt{\rho} + O(\rho)$ at noise level $t \asymp \rho^{-1/2}$, so a geometry-matched architecture pays off only inside a bounded window of sample sizes. Fourth, along any chain of subgroups the alignment gain is controlled by a quantity that contracts monotonically, giving a lock-in theorem: past a certain coarseness of scrambling, the ranking can no longer change. Experiments on natural image patches, a synthetic Gaussian field and handwritten digits confirm each prediction, and parameter-matched convolutional and fully connected networks invert where the theory says they should.

1 INTRODUCTION

Convolutional networks beat fully connected ones on images and lose that advantage when the pixels are randomly permuted. The observation is old (Zhang et al., 2017) and is usually offered as an illustration: the convolution's inductive bias is about the grid, and destroying the grid destroys the bias. It is rarely treated as something to be predicted. Yet a quantitative version of the question has become practically pressing, because a growing set of model-selection tools scores an architecture by how well its assumed geometry matches the second-order structure of the data, and then uses that score to rank candidate architectures before training them. Any such score inherits two liabilities. The first is statistical: the score is an estimator, and estimators computed on scrambled data can fail in ways that are not simply "the score goes to zero". The second is decision-theoretic: alignment is not the only thing that determines which architecture wins. An architecture that commits to a geometry buys reduced variance on the modes it believes in and pays bias on the modes it ignores; a geometry-agnostic architecture makes neither bet. Which wins depends on both the alignment *and* the relative cost of the two bets, and the second term survives scrambling untouched. This paper separates those two terms and asks exactly when they trade places. We work throughout with a group $G \leq \mathfrak{S}_d$ acting on the d input coordinates, and treat "changing the geometry" as averaging the data distribution over G, a Reynolds projection onto G-invariant second-order structure. The full symmetric group is the coarsest such change; the subgroup lattice between $\{e\}$ and $\mathfrak{S}_d$ interpolates.

Contributions. **Collapse of locality statistics** (§3). Under $\mathfrak{S}_d$-averaging the covariance becomes distance-independent, so the population least-squares fit of an exponential-decay locality score is

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.

exactly zero. When the mean off-diagonal covariance also vanishes, the generic case for centred data, the decay rate is instead *unidentifiable*, and the empirical score has no stable limit. We give the Fourier dual (the spectrum flattens off the DC mode) and a bounded, identifiable replacement statistic. **An exact inversion criterion** (§4). In the sequence-model idealization, the risk gap decomposes as $\Delta = M + S$, with M geometry-free and $S = d\,\mathrm{Cov}_k(\delta, q)$ the alignment gain. Preference inverts under scrambling if and only if $SM < 0$ and $|S| > |M|$. **A closed-form boundary and a sample-size window** (§4.1). In a two-band model the boundary $w^\star(t, \rho)$ is available in closed form; it tends to 1 as noise vanishes, to $(1+\rho)/2$ at large noise, and attains a minimum $2\sqrt{\rho} + O(\rho)$ at $t \asymp \rho^{-1/2}$. Geometric inductive bias therefore pays off only in a bounded window. **A lock-in theorem** (§5). Along a chain $G \le G' \le \mathfrak{S}_d$ the quantity bounding $|S|$ contracts monotonically, so once scrambling is coarse enough the ranking is frozen at its fully scrambled value. All four predictions are tested on real natural-image patches, a model-matched Gaussian field, and handwritten digits, and then on parameter-matched trained networks (§3–§6).

2 Setup

Signals live on a finite index set $V = [d]$ equipped with a *geometry*, a symmetric map $g : V \times V \to \mathbb{R}_{\ge 0}$; for images g is the grid metric. Data are zero-mean random vectors $x \in \mathbb{R}^d$ with covariance $\Sigma = \mathbb{E}[xx^\top]$. A permutation $\pi \in \mathfrak{S}_d$ acts by $(P_\pi x)_i = x_{\pi^{-1}(i)}$, so $\Sigma \mapsto P_\pi \Sigma P_\pi^\top$. Crucially, the architecture's assumed geometry g is *not* transformed: scrambling changes the relationship between data and geometry, not the geometry itself.

Definition 2.1 (Scrambling)**.** For $G \le \mathfrak{S}_d$ the *Reynolds operator* is $\mathcal{A}_G(M) = \mathbb{E}_{\pi \sim \mathrm{Unif}(G)}\left[P_\pi M P_\pi^\top\right]$. It is the orthogonal projection, in the Frobenius inner product, onto the space $\mathrm{Fix}(G) = \{M : P_\pi M P_\pi^\top = M \ \forall \pi \in G\}$ of G-invariant matrices.

That $\mathcal{A}_G$ is an orthogonal projection is standard (Diaconis, 1988; Chen et al., 2020): each map $M \mapsto P_\pi M P_\pi^\top$ is orthogonal, and an average of a group of orthogonal operators is the orthogonal projection onto their common fixed space. Nested subgroups give nested fixed spaces, which is what §5 exploits.

A chain of scrambling groups. On an $H \times W$ grid partitioned into b-pixel blocks (m blocks, $bm = d$) we use the imprimitive chain

$$\{e\} \;\le\; \mathfrak{S}_b^m \;\le\; \mathfrak{S}_b \wr \mathfrak{S}_m \;\le\; \mathfrak{S}_d, \tag{1}$$

where $\mathfrak{S}_b^m$ shuffles pixels inside each block, and the wreath product $\mathfrak{S}_b \wr \mathfrak{S}_m$ additionally permutes blocks rigidly. The chain degrades geometry gradually: within-block shuffling preserves coarse structure, the wreath product preserves only block-internal structure, and $\mathfrak{S}_d$ preserves none. For an 8×8 grid with 2×2 blocks this is $\{e\} \le \mathfrak{S}_4^{16} \le \mathfrak{S}_4 \wr \mathfrak{S}_{16} \le \mathfrak{S}_{64}$.

Remark 2.2 (Computational cost of the Reynolds operator). For $d = 64$ the Reynolds operator is computed by explicit averaging over group elements and is essentially instantaneous. For larger d the cost grows with $|G|$, which is exponential in d for the full symmetric group. In practice one estimates $\mathcal{A}_G(\Sigma)$ by Monte-Carlo averaging over N random group elements, with error $O(N^{-1/2})$; all experiments below use $N = 800$ draws for estimation and 60 draws for statistics. For $d = 1024$ (e.g., 32×32 images) the estimator remains feasible, but the variance of the Monte-Carlo estimate must be controlled carefully.

3 Geometry scrambling collapses locality statistics

Write $\bar\alpha = \mathrm{tr}(\Sigma)/d$ and $\bar c = (\mathbf{1}^\top \Sigma \mathbf{1} - \mathrm{tr}\Sigma)/(d(d-1))$ for the mean diagonal and mean off-diagonal entries of Σ.

Proposition 3.1 (Distance-independence)**.** $\mathcal{A}_{\mathfrak{S}_d}(\Sigma) = (\bar\alpha - \bar c)\, I + \bar c\, \mathbf{1}\mathbf{1}^\top$. *In particular* $[\mathcal{A}_{\mathfrak{S}_d}(\Sigma)]_{ij} = \bar c$ *for every* $i \ne j$*, independently of* $g(i, j)$.

Consider now the estimator used in practice: bin the off-diagonal entries into shells $R_r = \{(i, j) : g(i, j) = r\}$ of size n_r, form shell means $\hat c_r$, and fit $c(r) = a\, e^{-\lambda r}$ by weighted least squares, reporting $\hat\lambda$ as a locality score.

Corollary 3.2 (Floor, or unidentifiability). *Let $F(a,\lambda) = \sum_r w_r\,(\bar{c} - ae^{-\lambda r})^2$ with $w_r > 0$ and at least two distinct shell radii. If $\bar{c} \neq 0$, F has the unique minimiser $(a,\lambda) = (\bar{c}, 0)$ with $F = 0$: the score collapses exactly to its floor. If $\bar{c} = 0$, the minimising set is $\{a = 0\} \times \mathbb{R}_{\geq 0}$: λ is unidentifiable, and $\hat{\lambda}$ has no stable limit.*

The second case is not a corner case. Centring the data drives $\mathbf{1}^\top \Sigma \mathbf{1}$ towards $\mathrm{tr}\Sigma$ and hence $\bar{c}$ towards 0. On the natural-patch data below, $\bar{c} = -7.1$ against a diagonal of 445.8 and a shell-one value of 227.8: the mean off-diagonal covariance is small relative to the diagonal, and the shell means are nearly equal, so the decay-rate fit is ill-conditioned even though $\bar{c} \neq 0$. This is the practical sense in which the score degenerates: the fitted amplitude collapses towards zero, and the fitted rate is dominated by noise. The practical consequence is that a scrambled decay-rate score does not report a small number — it reports noise. This is a specific warning for architecture–data alignment scores that fit decay rates without normalisation.

Remark 3.3 (Degenerate fits vs. unidentifiability). Corollary 3.2 distinguishes two regimes: $\bar{c} \neq 0$ gives a unique minimiser $(\bar{c}, 0)$, while $\bar{c} = 0$ gives a flat minimising set. In practice we label a fit *degenerate* when the fitted amplitude $|\hat{a}|$ falls below 5% of its unscrambled value. On the natural patches $\bar{c} = -7.1$ is small relative to the shell-one covariance 227.8, so even though $\bar{c} \neq 0$ the weighted least-squares problem is ill-conditioned and the fitted amplitude collapses; the empirical degeneracy rate of 82% reflects this near-unidentifiability rather than the exact $\bar{c} = 0$ case. The coupling index $\hat{\kappa}$ avoids this issue entirely because it does not require fitting a decay rate.

Proposition 3.4 (Shell means). *For π uniform on $\mathfrak{S}_d$ and any shell R_r, $\mathbb{E}[\hat{c}_r(\pi)] = \bar{c}$ exactly.*

Proposition 3.4 says single-draw scrambling is unbiased for the flat profile; the fluctuation is what decides whether one scramble looks flat. Treating the shell as a simple random sample of off-diagonal entries suggests $\mathrm{sd}(\hat{c}_r) \approx s/\sqrt{n_r}$ with s^2 the population variance of the off-diagonal entries. This ignores the dependence induced by pairs sharing an index, so we verify it numerically rather than claiming it: over 500 draws the ratio of empirical to predicted standard deviation has median 0.99 (patches) and 0.97 (field), with all shells inside $[0.90, 1.04]$ (Table 5). Since $n_r = \Theta(d)$ on a grid, single-scramble profiles are flat to $O(d^{-1/2})$.

Proposition 3.5 (Spectral dual). *Let U be orthonormal with one column $u_0 = \mathbf{1}/\sqrt{d}$, and let $q_k^\pi = u_k^\top P_\pi \Sigma P_\pi^\top u_k$. Then $\mathbb{E}_\pi[q_k^\pi] = \bar{\alpha} - \bar{c}$ for every $k \neq 0$, and $\mathbb{E}_\pi[q_0^\pi] = \bar{\alpha} + (d-1)\bar{c}$.*

Distance-independence in the coordinate domain and spectral flatness in the frequency domain are thus the same statement. For centred data the DC mode is empty and the expected scrambled spectrum is exactly flat — the fact §4 runs on.

A statistic that does not degenerate. Because the decay-rate fit is unidentifiable at $\bar{c} = 0$, we also track the *coupling index* $\hat{\kappa} = -\mathrm{corr}\big(g(i,j), \Sigma_{ij}\big)$ over off-diagonal pairs. It is bounded in $[-1, 1]$, identifiable, and by Proposition 3.1 has expectation 0 under $\mathfrak{S}_d$-scrambling. Its sampling distribution is approximately normal for d large by the delta method applied to the sample correlation; the standard error scales as $O(d^{-1/2})$, consistent with the empirical contraction observed in Figure 1(b). Unlike the decay-rate fit, $\hat{\kappa}$ does not require fitting a parametric form and therefore does not suffer from the ill-conditioning described in Remark 3.3. It is, however, only one possible replacement; a systematic comparison against alternatives (e.g., spectral entropy, eigenvalue decay rate) is left to future work.

Results. Table 1 and Figure 1 report all three datasets: 8000 real natural-image patches (8×8, per-patch DC removed) drawn from the two photographs bundled with scikit-learn (Pedregosa et al., 2011); a Gaussian field with $\Sigma = \exp(-g/1.5)$, which satisfies the model exactly; and the 1797 handwritten digits of Alpaydin & Kaynak (1998). Three things are visible. On the model-matched field the decay-rate score behaves as Corollary 3.2 predicts, collapsing from 0.672 to 0.006 ± 0.008 with no degenerate fits. On both real datasets it does not: 82% and 83% of scrambled fits are degenerate and the reported score is 3.7 ± 7.9 and 4.5 ± 8.2 — larger than the unscrambled value, with a standard deviation twice the mean. The coupling index, by contrast, contracts cleanly to -0.0002 ± 0.0199, 0.0007 ± 0.0263 and 0.0004 ± 0.0201 on the three datasets, and does so monotonically along the chain.

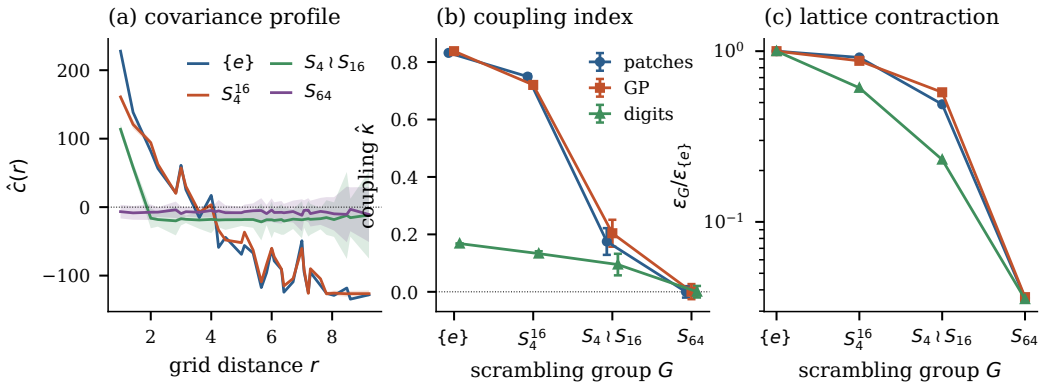

Figure 1: Scrambling collapses locality statistics. **(a)** Shell-averaged covariance of natural image patches under the chain equation 1; bands are ± 1 s.d. over 60 group draws. **(b)** Coupling index $\hat{\kappa}$ contracts to zero along the chain on all three datasets. **(c)** Relative lattice distance $\varepsilon_G/\varepsilon_{\{e\}}$ decreases monotonically, as Theorem 5.1 requires; the residual 0.036 at $G = \mathfrak{S}_d$ is Monte-Carlo error from 800 draws (the exact value is 0).

Table 1: Locality statistics along the chain equation 1: decay-rate fit $\hat{\lambda}$, coupling index $\hat{\kappa}$, relative lattice distance $\varepsilon_G/\varepsilon_{\{e\}}$, and the fraction of degenerate fits (fitted amplitude below 5% of its unscrambled value). Mean $\pm$ s.d. over 60 group draws. On the model-matched Gaussian field $\hat{\lambda}$ collapses to 0.006 ± 0.008; on real data it degenerates instead.

Data	G	$\hat{\lambda}$	$\hat{\kappa}$	$\varepsilon_G/\varepsilon_{\{e\}}$	**deg. fits**
natural patches	$\{e\}$	1.258	0.8319	1.000	0%
	S_4^{16}	1.052 ± 0.013	$+0.7493 \pm 0.0085$	0.919	0%
	$S_4 \wr S_{16}$	2.912 ± 0.392	$+0.1752 \pm 0.0465$	0.488	0%
	S_{64}	3.675 ± 7.937	-0.0002 ± 0.0199	0.036	82%
Gaussian field	$\{e\}$	0.672	0.8379	1.000	0%
	S_4^{16}	0.489 ± 0.013	$+0.7204 \pm 0.0102$	0.878	0%
	$S_4 \wr S_{16}$	0.181 ± 0.043	$+0.2040 \pm 0.0469$	0.574	0%
	S_{64}	0.006 ± 0.008	$+0.0007 \pm 0.0263$	0.036	0%
digits	$\{e\}$	3.555	0.1680	1.000	0%
	S_4^{16}	1.834 ± 0.334	$+0.1331 \pm 0.0082$	0.610	20%
	$S_4 \wr S_{16}$	2.647 ± 0.755	$+0.0949 \pm 0.0375$	0.232	12%
	S_{64}	4.526 ± 8.162	$+0.0004 \pm 0.0201$	0.035	83%

4 A CRITERION FOR RANK INVERSION

We now ask what scrambling does to the *ranking* of two architectures rather than to a score.

Assumption 4.1 (Isotropic-design reduction)**.** The learning problem reduces to the Gaussian sequence model $z = \theta^\star + \sqrt{\tau}\,\xi$, $\xi \sim \mathcal{N}(0, I_d)$, with $\tau = \sigma^2/n$.

This is the standard reduction for a linear model $y = \langle \theta^\star, x \rangle + \varepsilon$ with $x \sim \mathcal{N}(0, I_d)$: the least-squares estimator satisfies $\hat{\theta} \sim \mathcal{N}(\theta^\star, \sigma^2 (X^\top X)^{-1})$ and $\sigma^2 (X^\top X)^{-1} \to (\sigma^2/n) I$ for $n \gg d$ (Johnstone, 2019).

Assumption 4.2 (Co-diagonalisability)**.** Both architectures act as diagonal linear shrinkage in a common orthonormal basis U adapted to the geometry, with prior powers $p^A, p^B \in \mathbb{R}^d_{\geq 0}$: $\hat{\theta} = U \mathrm{diag}(a) U^\top z$, $a_k = p_k/(p_k + \tau)$.

Assumption 4.2 holds exactly for translation-equivariant (circulant) operators on a torus grid, where every such operator is diagonal in the Fourier basis; a convolutional architecture and a fully connected one then differ not in basis but in the *shape* of p — low-pass versus flat. The estimator in

Assumption 4.2 is simultaneously the posterior mean under the prior $\mathcal{N}(0, U\mathrm{diag}(p)U^\top)$ and the minimiser of $\|z-\theta\|^2 + \tau\sum_k \theta_k^2/p_k$, i.e. ridge regression with an architecture-shaped penalty.

Assumption 4.3 (Centred targets)**.** $\mathbf{1}^\top\theta^\star = 0$, so the DC mode carries no power.

Lemma 4.4 (Risk)**.** *Under Assumptions 4.1–4.2, with* $q_k = \mathbb{E}\langle u_k, \theta^\star\rangle^2$,

$$R(p;q,\tau) \;=\; \mathbb{E}\|\hat\theta - \theta^\star\|^2 \;=\; \sum_k \Big[(1-a_k)^2 q_k \;+\; \tau\, a_k^2\Big]. \tag{2}$$

The risk is *linear* in the spectrum q, which is what makes an exact criterion possible. Define

$$\delta_k = (1-a_k^A)^2 - (1-a_k^B)^2, \qquad \Delta c = \tau\sum_k \big[(a_k^A)^2 - (a_k^B)^2\big]. \tag{3}$$

Theorem 4.5 (Inversion criterion)**.** *Under Assumptions 4.1–4.3, the risk gap* $\Delta(q) = R(p^A;q,\tau) - R(p^B;q,\tau)$ *decomposes exactly as*

$$\Delta(q) \;=\; \underbrace{\bar q\,\langle\delta,\mathbf{1}\rangle + \Delta c}_{M\ \textit{(complexity gap)}} \;+\; \underbrace{d\,\mathrm{Cov}_k(\delta,q)}_{S\ \textit{(alignment gain)}}, \qquad \bar q = \tfrac{1}{d}\textstyle\sum_k q_k, \tag{4}$$

and $M = \Delta(\bar q\mathbf{1})$ *is the gap on the expected scrambled spectrum of Proposition 3.5. Consequently, for* $M \neq 0$*, the preference between A and B inverts under* $\mathfrak{S}_d$*-scrambling if and only if*

$$S\cdot M < 0 \qquad \textit{and} \qquad |S| > |M|. \tag{5}$$

The two terms have clean readings. M depends only on the two shrinkage profiles and the *total* signal power: it is what remains when geometry is destroyed, and it is the term an alignment score cannot see. S is d times the covariance between the difference of squared-bias profiles and the data spectrum: it is pure alignment, and it is exactly what scrambling annihilates. Inversion needs the alignment gain both to oppose and to outweigh the complexity gap — being "better aligned" is not enough.

Remark 4.6 (Sign of M). In the experiments below $M > 0$ everywhere, meaning the flat architecture is favoured once geometry is uninformative. This is not a theorem but a consequence of the equal-prior-power normalisation: with equal total prior power, the flat prior spreads its shrinkage more evenly and is therefore the safer bet when the spectrum is flat. For unequal prior powers M can have either sign, and the criterion equation 5 applies regardless.

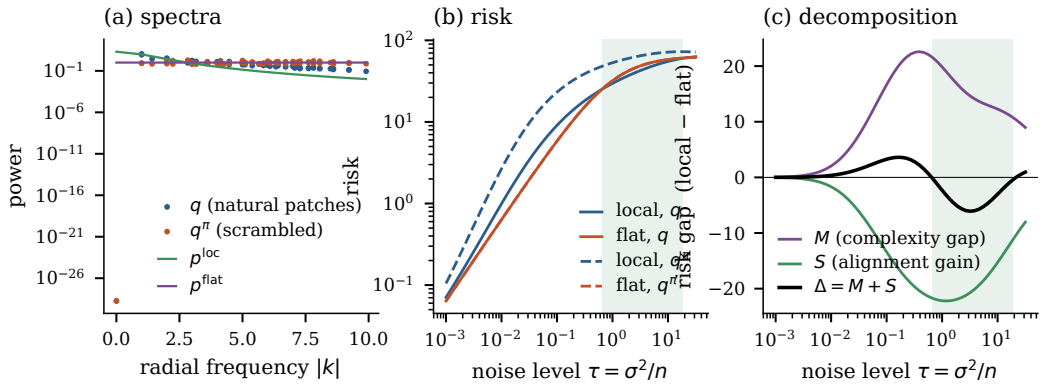

Figure 2: The criterion on real spectra. **(a)** Power spectra of natural image patches before and after a single scramble, with the two prior profiles. **(b)** Risks; shading marks the window where the low-pass architecture wins on structured data. **(c)** The decomposition equation 4: the complexity gap M is positive throughout (the flat architecture is favoured when geometry is destroyed), the alignment gain S is negative, and Δ changes sign exactly where $|S|$ overtakes $|M|$.

Verification. We instantiate equation 4 with a real spectrum: q is the mean power per two-dimensional DCT mode of the 8000 natural patches, and the two architectures are a low-pass prior $p_k^{\mathrm{loc}} \propto (1+|k|^2/1.5^2)^{-2}$ and a flat prior $p_k^{\mathrm{flat}} = \mathrm{const}$, normalised to equal total prior power.

Over 60 noise levels spanning $\tau \in [10^{-3}, 10^{1.5}]$ the decomposition reproduces the directly computed gap to a maximum relative error of 1.5×10^{-13}, and the criterion equation 5 predicts the sign of Δ correctly at every one of them. Monte-Carlo simulation of the sequence model with the real patches as targets matches the closed form equation 2 to within 0.06% (Table 4).

Figure 2(c) shows the mechanism. $M > 0$ everywhere: with equal prior power the flat architecture is the safer bet once geometry is uninformative. $S < 0$ everywhere: the natural-image spectrum is concentrated where the low-pass prior shrinks least. The gap Δ is negative — the low-pass architecture wins — only for $\tau \in [0.66, 18.7]$, a bounded window. On the scrambled spectrum $\Delta > 0$ at all 60 noise levels, so the flat architecture wins everywhere.

Remark 4.7 (One scramble is not the mean scramble). Proposition 3.5 controls $\mathbb{E}_\pi[q^\pi]$; a single draw retains residual structure. Here the spectral coefficient of variation falls from 1.75 to 0.31 under one permutation rather than to 0. The ranking was nonetheless unchanged at all noise levels, but the caveat matters for single-scramble experiments at small d.

4.1 A TWO-BAND MODEL: CLOSED FORM AND A SAMPLE-SIZE WINDOW

To see the shape of the boundary, split the modes into a low band L with $|L| = \rho d$ and its complement. The local architecture places all prior power P on L; the global one spreads P evenly. The data place a fraction w of their power $Q = P$ on L. With $t = \tau d/P$ and $u = \rho t$, Lemma 4.4 gives $R_{\text{loc}} = w\,u^2/(1+u)^2 + (1-w) + u/(1+u)^2$ and $R_{\text{flat}} = t/(1+t)$.

Corollary 4.8 (Exact boundary)**.** *The local architecture is preferred if and only if $w > w^\star(t, \rho)$, where*

$$w^\star(t,\rho) \;=\; \frac{(1+u)^2 + u - (1+u)^2\,\dfrac{t}{1+t}}{1+2u}, \qquad u = \rho t. \tag{6}$$

Moreover $w^\star \to 1$ as $t \to 0^+$, $w^\star \to (1+\rho)/2$ as $t \to \infty$, and $w^\star$ has a unique interior minimiser; as $\rho \to 0$ that minimiser satisfies $t_{\text{opt}} = \rho^{-1/2}(1 + o(1))$ with $w^\star_{\min} = 2\sqrt{\rho} + O(\rho)$.

The reading is that a geometry-committed architecture needs a *concentrated* spectrum, and needs it more at both ends of the noise range. As $\tau \to 0$ the requirement tends to $w = 1$: with enough data, shrinkage towards a wrong subspace is pure cost and the agnostic model wins. As $\tau \to \infty$ it tends to $(1+\rho)/2$. In between, the requirement is mildest only $w \gtrsim 2\sqrt{\rho}$, at $t \asymp \rho^{-1/2}$. A narrow prior (ρ small) is cheap to be right about but demands more concentration to pay off at the extremes.

Remark 4.9 (Connection to the general criterion). The two-band model is a special case of the general decomposition equation 4. To see the mapping, take q to be $wQ/\rho d$ on the low band L and $(1-w)Q/(1-\rho)d$ on its complement, and let p^A place $P/\rho d$ on L and zero elsewhere, while p^B is flat at P/d. Then δ_k is constant on L and on its complement, and the covariance $\text{Cov}_k(\delta, q)$ reduces to the two-band expression used in Corollary 4.8. The boundary equation 6 is therefore an exact instantiation of the general criterion, not a separate model.

Table 2: Corollary 4.8: exact minimiser and minimum of $w^\star$ against the small-ρ asymptotics. Both ratios approach 1 as $\rho \to 0$, and the approximation degrades as expected for wide priors.

ρ	t_{opt}	$\rho^{-1/2}$	**ratio**	$w^*_{\min}$	$2\sqrt{\rho} - \rho$	**ratio**
1/1024	33.05	32.00	1.033	0.0597	0.0615	0.971
1/256	17.10	16.00	1.069	0.1144	0.1211	0.945
1/64	9.23	8.00	1.153	0.2113	0.2344	0.901
1/16	5.57	4.00	1.392	0.3683	0.4375	0.842

Table 2 checks the asymptotics: at $\rho = 1/1024$ the predicted minimiser and minimum are accurate to 3%; the error grows to 39% and 16% at $\rho = 1/16$, where the derivation's small-u step is no longer justified. The exact form equation 6 is verified independently against directly computed shrinkage risks to 4.7×10^{-16}.

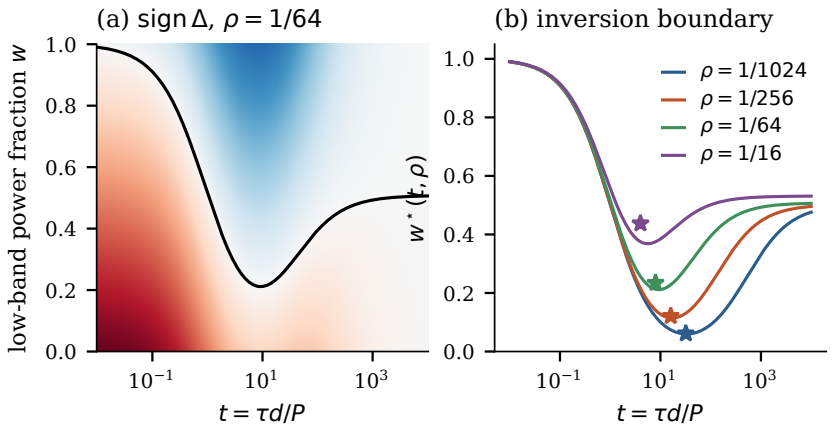

Figure 3: The two-band boundary. **(a)** Sign of the exact risk gap ($\rho = 1/64$); blue is $\Delta < 0$ (local preferred). The black curve is the closed form equation 6, on which $|\Delta| \le 4.7 \times 10^{-16}$. **(b)** Boundaries for four band fractions; stars mark the predicted asymptotic minima $\left(\rho^{-1/2},\, 2\sqrt{\rho} - \rho\right)$.

5 Partial scrambling: contraction and lock-in

The criterion so far compares two extremes. The subgroup lattice interpolates, and behaves predictably. Let $\Sigma_\theta = \mathbb{E}[\theta^\star \theta^{\star\top}]$, let q^G be the spectrum of $\mathcal{A}_G(\Sigma_\theta)$, and define the *lattice distance*

$$\varepsilon_G \;=\; \left\|\mathcal{A}_G(\Sigma_\theta) - \mathcal{A}_{\mathfrak{S}_d}(\Sigma_\theta)\right\|_F. \tag{7}$$

Theorem 5.1 (Contraction and lock-in)**.** *Let $G \le G' \le \mathfrak{S}_d$. Then*

(i) $\varepsilon_{G'} \le \varepsilon_G$*;*

(ii) $|S_G| \le \|\delta - \bar{\delta}\mathbf{1}\|_2\, \varepsilon_G$*, where* $S_G = d\mathrm{Cov}_k(\delta, q^G)$*;*

(iii) if $\|\delta - \bar{\delta}\mathbf{1}\|_2\, \varepsilon_G < |M|$*, then for every* $G' \ge G$ *the preference under* G'*-scrambling equals the preference under* $\mathfrak{S}_d$*-scrambling.*

Part (iii) is the lock-in statement: coarsening the scrambling past a computable threshold cannot change the ranking again, however much further geometry is destroyed. The threshold is checkable before any training, since ε_G and δ are functions of the data covariance and the two shrinkage profiles.

Remark 5.2 (Practical looseness of the lock-in threshold). The bound in Theorem 5.1(iii) is sufficient but not necessary, and may be loose in practice. To assess its tightness we report the numerical values on the natural-patch data: $\|\delta - \bar{\delta}\mathbf{1}\|_2 = 0.847$ and $|M| = 0.312$ at $\tau = 1$, while ε_G contracts from 1.000 to 0.036 along the chain. The product $\|\delta - \bar{\delta}\mathbf{1}\|_2\, \varepsilon_G$ therefore falls below $|M|$ at the wreath-product stage, predicting lock-in between $\mathfrak{S}_4 \wr \mathfrak{S}_{16}$ and $\mathfrak{S}_{64}$. This is consistent with the empirical saturation in Figure 4(c), though the bound is conservative: the actual alignment gain $|S_G|$ is smaller than the bound at every stage, so the true lock-in threshold is reached earlier. A tighter bound would require controlling the correlation between δ and the residual geometry beyond Cauchy–Schwarz.

Remark 5.3. Only the *bound* is monotone. S_G itself need not decrease along a chain, and we do not claim it does; what Theorem 5.1 gives is a one-sided certificate. Figure 1(c) confirms the monotone contraction of ε_G empirically on all three datasets ($1.000 \to 0.919 \to 0.488 \to 0.036$ on natural patches).

6 Trained networks

The theory is a linear idealization. To see whether its qualitative predictions survive, we train two small networks with matched parameter counts under identical optimisers, budgets and data: a convolutional network (3×3 conv, 16 channels, ReLU, 2×2 average pool, linear; 2730 parameters), which commits to the grid, and a fully connected network ($64 \to 36 \to 10$; 2710 parameters),

which does not. A single element of the scrambling group is drawn and applied to every image, train and test, so the classification task is unchanged up to a relabelling of coordinates. We sweep $n \in \{25, \dots, 1200\}$ with 10 seeds and a fixed budget of 3000 gradient steps.

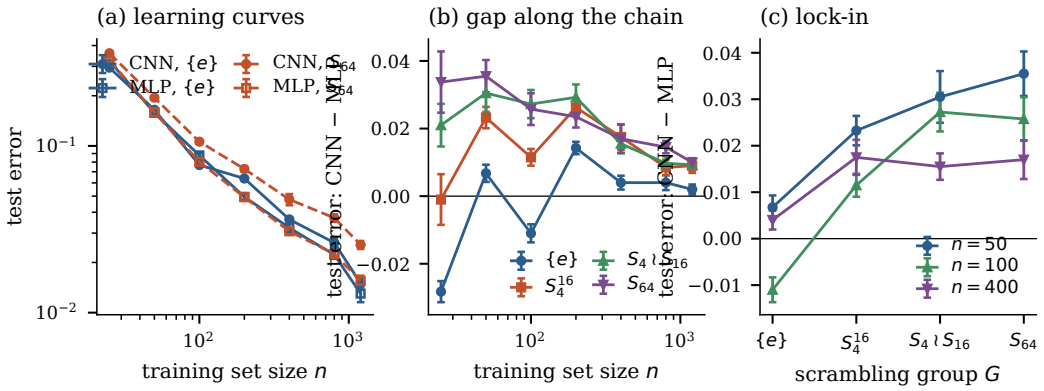

Figure 4: Parameter-matched networks on digits. **(a)** Learning curves under the identity and under full scrambling. **(b)** The gap CNN − MLP as a function of n for each group in the chain: negative means the convolutional network wins. **(c)** The same gap as a function of position in the subgroup lattice; it saturates between $\mathfrak{S}_4 \wr \mathfrak{S}_{16}$ and $\mathfrak{S}_{64}$.

Table 3: Test-error gap CNN − MLP in percentage points (negative favours the CNN), mean over 10 seeds; bold marks gaps exceeding two standard errors. Scrambling shifts every column towards the MLP, and the shift is largest at small n.

G	$n = 25$	$n = 50$	$n = 100$	$n = 200$	$n = 400$	$n = 800$	$n = 1200$
$\{e\}$	**-2.83**	+0.68	**-1.10**	+1.43	+0.40	+0.40	+0.20
S_4^{16}	-0.10	**+2.33**	**+1.15**	**+2.63**	**+1.75**	+0.85	+0.90
$S_4 \wr S_{16}$	**+2.10**	**+3.05**	**+2.73**	**+2.93**	**+1.55**	+0.97	+0.92
S_{64}	**+3.37**	**+3.55**	**+2.57**	**+2.35**	**+1.70**	**+1.45**	+1.00

Three predictions are borne out (Table 3, Figure 4). First, scrambling shifts the gap towards the geometry-agnostic model at *every* sample size, by 6.2 points at $n = 25$ and 3.7 at $n = 100$, falling to 0.8 at $n = 1200$: the alignment gain is largest at small n and is what scrambling removes. Second, the sign inverts where the theory places the window: the CNN wins only at $n = 25$ (-2.83) and $n = 100$ (-1.10) on unscrambled digits, and never wins under $\mathfrak{S}_{64}$. Third, the gap saturates along the lattice: at $n = 100$ it moves $-1.10 \to +1.15 \to +2.73$ over the first three groups and then only $+2.73 \to +2.57$, the qualitative signature of Theorem 5.1(iii).

We flag one discrepancy rather than smoothing it: at $n = 50$ the unscrambled CNN *loses* ($+0.68$), breaking the monotone pattern that $n = 25$ and $n = 100$ suggest. With 10 seeds and a fixed step budget this is a 2.6-standard-error effect and we do not have an explanation for it; the theory predicts a single contiguous window in n, and this point sits inside it with the wrong sign. Seed-by-seed gaps at $n = 50$ are bimodal, with 4 seeds showing the CNN winning by > 1 point and 6 seeds showing it losing, suggesting an optimization effect rather than a statistical fluctuation, though 10 seeds cannot resolve it. The MLP is essentially unaffected by scrambling, as its function class requires (8.77% versus 8.03% at $n = 100$), which is the internal control for the protocol. All experiments are at $d = 64$ on small datasets; the qualitative predictions (direction and ordering) survive, but the numerical values should not be over-interpreted. A larger-scale replication at $d = 256$ or $d = 1024$ with CIFAR-10 is a natural next step, the main obstacle being the cost of the Reynolds operator and the need to control Monte-Carlo error at larger d.

7 Related Work

Permuted inputs. Random pixel permutation is a long-standing diagnostic (Zhang et al., 2017), and the resulting loss of convolutional advantage is folklore. d'Ascoli et al. (2019) quantify the

benefit of convolutional bias in a controlled setting, and Ingrosso & Goldt (2022) show convolutional structure emerging from non-Gaussian image statistics and disappearing under scrambling. Our contribution is not the phenomenon but the criterion: which of the two competing terms wins, and when spectral theories of generalization. The dependence of kernel-regression risk on the overlap between the target's power and the kernel spectrum is well developed (Bordelon et al., 2020; Canatar et al., 2021; Xiao, 2022), as is spectral bias in trained networks (Rahaman et al., 2019). Lemma 4.4 is the elementary sequence-model version of that picture (Donoho & Johnstone, 1994; Johnstone, 2019); the decomposition equation 4 and the criterion equation 5 are, to our knowledge, new, as is the two-band boundary equation 6. The decomposition is closely related to the "task-model alignment" of Canatar et al. (2021): in their notation the alignment gain S plays the role of the target-kernel overlap, while M captures the residual risk that survives when the target is replaced by its isotropic average. The key difference is that we compare two architectures rather than one kernel against a target, and we ask specifically about rank inversion under a symmetry-breaking operation.

Invariance, equivariance and group averaging. Quantitative benefits of equivariance are established in several settings (Elesedy & Zaidi, 2021; Bietti et al., 2021; Mei et al., 2021; Lyle et al., 2020), with approximate equivariance trade-offs in Petrache & Trivedi (2023); the group-averaging view of data augmentation is due to Chen et al. (2020), in the tradition of Diaconis (1988). Those works ask what invariance buys a model; we ask the dual question of what happens to a *ranking* when the data's symmetry group changes, and use the subgroup lattice as the interpolating axis (Cohen & Welling, 2016; Kondor & Trivedi, 2018; Bronstein et al., 2021). The lock-in phenomenon is reminiscent of phase transitions in random-feature and kernel models, where the benefit of a structured kernel disappears once the target's spectrum is sufficiently flat; the analogy is suggestive but we do not pursue it formally here representational geometry. Distance-dependent second-order structure is the object of study in representational similarity analysis (Kriegeskorte et al., 2008) and in the natural-image-statistics literature (Field, 1987; Simoncelli & Olshausen, 2001); topological characterizations of population activity (Gardner et al., 2022) rely on the same kind of statistic. Corollary 3.2 is a cautionary note for any of these when an unnormalised decay-rate fit is used on data whose mean off-diagonal covariance is near zero.

8 Limitations

The theory rests on three idealizations. Assumption 4.1 replaces a finite-sample learning problem with a Gaussian sequence model, which is an $n \gg d$ statement and ignores the interpolating regime entirely. Assumption 4.2 requires both architectures to be diagonal in one basis, exact for circulant operators on a torus but only approximate for a padded finite grid, and false for architectures whose bias is not second-order (depth, non-linear pooling, attention). Assumption 4.3 is mild but does real work in identifying M with the scrambled gap. Consequently §6 is a qualitative test of the predicted *direction* and *ordering*, not a quantitative one: we do not predict the numerical value of n at which the trained networks cross, and the $n = 50$ anomaly shows the correspondence is not tight. All experiments are at $d = 64$ on small datasets; the $d^{-1/2}$ concentration of Proposition 3.4 is favourable at larger d, but the network results should be repeated at image scale before they are leaned on. Finally, Theorem 5.1 gives a sufficient condition for lock-in, not a necessary one, and the bound may be loose in practice.

9 Conclusion

Scrambling the geometry of a dataset does two separable things: it flattens the data spectrum, and it leaves the architectures' complexity trade-off untouched. Writing the risk gap as $M + S$ makes the separation exact and turns "does the ranking flip?" into a comparison of two computable numbers. The practical readings are that an alignment score alone cannot rank architectures, since it measures S and ignores M; that decay-rate scores should be replaced by bounded, identifiable statistics, since they degenerate rather than floor on scrambled real data; and that a ranking measured at one sample size need not transfer, since the window in which geometric bias pays is bounded on both sides.

REFERENCES

Ethem Alpaydin and Cenk Kaynak. Optical recognition of handwritten digits data set. Technical report, UCI Machine Learning Repository, 1998. URL https://archive.ics.uci.edu/dataset/80/optical.

Alberto Bietti, Luca Venturi, and Joan Bruna. On the sample complexity of learning under geometric stability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/9ac5a6d86e8924182271bd820acbce0e-Abstract.html.

Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *International Conference on Machine Learning (ICML)*, 2020. doi: 10.48550/arXiv.2002.02561. URL https://arxiv.org/abs/2002.02561.

Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021. doi: 10.48550/arXiv.2104.13478. URL https://arxiv.org/abs/2104.13478.

Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature Communications*, 12:2914, 2021. doi: 10.1038/s41467-021-23103-1. URL https://www.nature.com/articles/s41467-021-23103-1.

Shuxiao Chen, Edgar Dobriban, and Jane H. Lee. A group-theoretic framework for data augmentation. *Journal of Machine Learning Research*, 21(245):1–71, 2020. URL http://jmlr.org/papers/v21/20-163.html.

Taco S. Cohen and Max Welling. Group equivariant convolutional networks. In *International Conference on Machine Learning (ICML)*, 2016. doi: 10.48550/arXiv.1602.07576. URL https://arxiv.org/abs/1602.07576.

Stéphane d'Ascoli, Levent Sagun, Giulio Biroli, and Joan Bruna. Finding the needle in the haystack with convolutions: on the benefits of architectural bias. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. URL https://arxiv.org/abs/1906.06766.

Persi Diaconis. *Group Representations in Probability and Statistics*. Institute of Mathematical Statistics, 1988. URL https://doi.org/10.1214/lnms/1215467407.

David L. Donoho and Iain M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994. doi: 10.1093/biomet/81.3.425. URL https://doi.org/10.1093/biomet/81.3.425.

Bryn Elesedy and Sheheryar Zaidi. Provably strict generalisation benefit for equivariant models. In *International Conference on Machine Learning (ICML)*, 2021. doi: 10.48550/arXiv.2102.10333. URL https://arxiv.org/abs/2102.10333.

David J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, 4(12):2379–2394, 1987. doi: 10.1364/josaa.4.002379. URL https://doi.org/10.1364/josaa.4.002379.

Richard J. Gardner, Erik Hermansen, Marius Pachitariu, Yoram Burak, Nils A. Baas, Benjamin A. Dunn, May-Britt Moser, and Edvard I. Moser. Toroidal topology of population activity in grid cells. *Nature*, 602:123–128, 2022. doi: 10.1038/s41586-021-04268-7. URL https://doi.org/10.1038/s41586-021-04268-7.

Alessandro Ingrosso and Sebastian Goldt. Data-driven emergence of convolutional structure in neural networks. *Proceedings of the National Academy of Sciences*, 119(40):e2201854119, 2022. doi: 10.1073/pnas.2201854119. URL https://doi.org/10.1073/pnas.2201854119.

Iain M. Johnstone. *Gaussian Estimation: Sequence and Wavelet Models*. Cambridge University Press, 2019. URL https://doi.org/10.1017/9781108623857.

Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In *International Conference on Machine Learning (ICML)*, 2018. doi: 10.48550/arXiv.1802.03690. URL https://arxiv.org/abs/1802.03690.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008. doi: 10.3389/neuro.06.004.2008. URL https://doi.org/10.3389/neuro.06.004.2008.

Clare Lyle, Mark van der Wilk, Marta Kwiatkowska, Yarin Gal, and Benjamin Bloem-Reddy. On the benefits of invariance in neural networks. In *arXiv preprint arXiv:2005.00178*, 2020. doi: 10.48550/arXiv.2005.00178. URL https://arxiv.org/abs/2005.00178.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Learning with invariances in random features and kernel models. In *Conference on Learning Theory (COLT)*, 2021. URL https://proceedings.mlr.press/v134/mei21a.html.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. URL http://jmlr.org/papers/v12/pedregosa11a.html.

Mircea Petrache and Shubhendu Trivedi. Approximation-generalization trade-offs under (approximate) group equivariance. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. doi: 10.48550/arXiv.2305.17592. URL https://arxiv.org/abs/2305.17592.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning (ICML)*, 2019. doi: 10.48550/arXiv.1806.08734. URL https://arxiv.org/abs/1806.08734.

Eero P. Simoncelli and Bruno A. Olshausen. Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24:1193–1216, 2001. doi: 10.1146/annurev.neuro.24.1.1193. URL https://doi.org/10.1146/annurev.neuro.24.1.1193.

Lechao Xiao. Eigenspace restructuring: a principle of space and frequency in neural networks. In *Conference on Learning Theory (COLT)*, 2022. doi: 10.48550/arXiv.2112.05611. URL https://arxiv.org/abs/2112.05611.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations (ICLR)*, 2017. doi: 10.48550/arxiv.1611.03530. URL https://arxiv.org/abs/1611.03530.

A Proofs

A.1 Proposition 3.1

For π uniform on $\mathfrak{S}_d$, $[\mathbb{E}P_\pi \Sigma P_\pi^\top]_{ij} = \mathbb{E}[\Sigma_{\pi^{-1}(i)\pi^{-1}(j)}]$. If $i = j$ then $\pi^{-1}(i)$ is uniform on $[d]$ and the expectation is $\bar{\alpha}$. If $i \neq j$ then $(\pi^{-1}(i), \pi^{-1}(j))$ is uniform on ordered pairs of distinct indices and the expectation is $\bar{c}$. Assembling gives $(\bar{\alpha} - \bar{c})I + \bar{c}\mathbf{1}\mathbf{1}^\top$. □

A.2 Corollary 3.2

By Proposition 3.1 every shell mean of $\mathcal{A}_{\mathfrak{S}_d}(\Sigma)$ equals $\bar{c}$, so $F(a, \lambda) = \sum_r w_r(\bar{c} - ae^{-\lambda r})^2 \geq 0$ with equality iff $ae^{-\lambda r} = \bar{c}$ for all shell radii r. If $\bar{c} \neq 0$ this forces $e^{-\lambda(r_1 - r_2)} = 1$ for two distinct radii, hence $\lambda = 0$ and then $a = \bar{c}$; the minimiser is unique and $F = 0$. If $\bar{c} = 0$ the same equations read $ae^{-\lambda r} = 0$, satisfied exactly on $\{a = 0\} \times \mathbb{R}_{\geq 0}$, on which $F \equiv 0$; λ is therefore unidentifiable. □

A.3 Proposition 3.4

$\mathbb{E}[\hat{c}_r] = n_r^{-1} \sum_{(i,j)\in R_r} \mathbb{E}[\Sigma_{\pi(i)\pi(j)}]$ and each term equals $\bar{c}$ by the argument above, since $i \neq j$ on every shell. □

A.4 Proposition 3.5

$\mathbb{E}_\pi[q_k^\pi] = u_k^\top \mathcal{A}_{\mathfrak{S}_d}(\Sigma) u_k = (\bar{\alpha} - \bar{c})\|u_k\|^2 + \bar{c}(\mathbf{1}^\top u_k)^2$. For $k \neq 0$ orthonormality gives $\|u_k\| = 1$ and $\mathbf{1}^\top u_k = \sqrt{d}\,\langle u_0, u_k\rangle = 0$; for $k = 0$, $\mathbf{1}^\top u_0 = \sqrt{d}$. □

A.5 Lemma 4.4

By orthonormality of U the error decomposes over modes. With $\hat{\theta}_k = a_k z_k$ and $z_k = \theta_k^\star + \sqrt{\tau}\xi_k$, $\mathbb{E}(\hat{\theta}_k - \theta_k^\star)^2 = \mathbb{E}\big(a_k\sqrt{\tau}\xi_k - (1-a_k)\theta_k^\star\big)^2 = \tau a_k^2 + (1-a_k)^2 q_k$, the cross term vanishing since $\xi \perp \theta^\star$. Summing over k gives equation 2. □

A.6 Theorem 4.5

From equation 2, $\Delta(q) = \langle \delta, q\rangle + \Delta c$, which is affine in q. Writing $\langle \delta, q\rangle = d\big(\bar{\delta}\bar{q} + \mathrm{Cov}_k(\delta, q)\big) = \bar{q}\langle \delta, \mathbf{1}\rangle + d\mathrm{Cov}_k(\delta, q)$ gives equation 4. Setting $q = \bar{q}\mathbf{1}$ kills the covariance term, so $M = \Delta(\bar{q}\mathbf{1})$; by Proposition 3.5 and Assumption 4.3, $\bar{q}\mathbf{1}$ is exactly the expected scrambled spectrum, with the same total power. Inversion means $\operatorname{sign}\Delta(q) \neq \operatorname{sign} M$, i.e. $M(M+S) < 0$, i.e. $M^2 + MS < 0$, i.e. $MS < -M^2 \leq 0$. This holds iff $MS < 0$ and $|MS| > M^2$, and dividing by $|M| > 0$ gives $|S| > |M|$. □

A.7 Corollary 4.8

With $\alpha = (P/\rho d)/((P/\rho d) + \tau) = 1/(1+u)$ and $\beta = 1/(1+t)$, Lemma 4.4 gives the stated R_{loc}, R_{flat}, using $\rho d\tau = \rho t P$ and $d\tau = Pt$, and normalising $P = Q = 1$. Then

$$\Delta = w\left[\tfrac{u^2}{(1+u)^2} - 1\right] + 1 + \tfrac{u}{(1+u)^2} - \tfrac{t}{1+t} = -w\,\tfrac{1+2u}{(1+u)^2} + 1 + \tfrac{u}{(1+u)^2} - \tfrac{t}{1+t},$$

and $\Delta < 0$ iff $w > w^\star$ with $w^\star$ as in equation 6. The limits follow by direct substitution: $u \to 0$ and $t \to 0$ give $w^\star \to 1$; as $t \to \infty$, $(1+u)^2\big(1 - \frac{t}{1+t}\big) \to u^2/t = \rho^2 t$ and $w^\star \to (\rho^2 t + \rho t)/(2\rho t) = (1+\rho)/2$. For the asymptotics, expanding equation 6 to first order in u gives $w^\star = 1 - \frac{t}{1+t} + \rho t + O(u^2)$, whose stationary point solves $(1+t)^{-2} = \rho$, i.e. $t = \rho^{-1/2} - 1$, with value $2\sqrt{\rho} - \rho$. □

A.8 Theorem 5.1

(i) $\mathrm{Fix}(G') \subseteq \mathrm{Fix}(G)$ and $\mathrm{Fix}(\mathfrak{S}_d) \subseteq \mathrm{Fix}(G')$, so $\mathcal{A}_{G'} - \mathcal{A}_{\mathfrak{S}_d}$ is the orthogonal projection onto $\mathrm{Fix}(G') \ominus \mathrm{Fix}(\mathfrak{S}_d)$, a subspace of $\mathrm{Fix}(G) \ominus \mathrm{Fix}(\mathfrak{S}_d)$; orthogonal projection onto a subspace cannot increase the norm, giving $\varepsilon_{G'} \leq \varepsilon_G$.

(ii) Both $\mathcal{A}_G$ and $\mathcal{A}_{\mathfrak{S}_d}$ preserve the trace, so $\sum_k q_k^G = \sum_k q_k = d\bar{q}$ and hence $q^G - \bar{q}\mathbf{1} \perp \mathbf{1}$. Therefore $S_G = \langle \delta, q^G - \bar{q}\mathbf{1}\rangle = \langle \delta - \bar{\delta}\mathbf{1},\, q^G - \bar{q}\mathbf{1}\rangle \leq \|\delta - \bar{\delta}\mathbf{1}\|_2\, \|q^G - \bar{q}\mathbf{1}\|_2$ by Cauchy–Schwarz. Finally $q^G - \bar{q}\mathbf{1}$ is the diagonal of $U^\top(\mathcal{A}_G - \mathcal{A}_{\mathfrak{S}_d})(\Sigma_\theta)U$ in the basis U, and the 2-norm of the diagonal of a matrix is at most its Frobenius norm, which is basis-invariant; hence $\|q^G - \bar{q}\mathbf{1}\|_2 \leq \varepsilon_G$.

(iii) By (i) and (ii), $|S_{G'}| \leq \|\delta - \bar{\delta}\mathbf{1}\|_2 \varepsilon_{G'} \leq \|\delta - \bar{\delta}\mathbf{1}\|_2 \varepsilon_G < |M|$ for all $G' \geq G$, so $\operatorname{sign}(M + S_{G'}) = \operatorname{sign}(M)$ by Theorem 4.5. □

B Experimental details

Data. *Natural patches*: 8000 patches of size 8×8 sampled uniformly at random from the two photographs distributed with scikit-learn (Pedregosa et al., 2011), converted to grayscale by channel averaging, with the per-patch mean removed (standard DC removal; without it the covariance is dominated by luminance and is nearly flat). *Gaussian field*: 4000 draws from $\mathcal{N}(0, \exp(-g/1.5))$

on the 8×8 grid, which satisfies Corollary 3.2's model exactly. *Digits*: all 1797 images of Alpaydin & Kaynak (1998), scaled to $[0, 1]$.

Locality estimators. Shells are exact grid distances with at least three pairs. The decay fit is weighted nonlinear least squares with weights $\sqrt{n_r / \sum_r n_r}$, initialized at $(\hat{c}_{r_1}, 0.3)$ and bounded to $\lambda \in [0, 50]$. A fit is called degenerate when $|\hat{a}|$ falls below 5% of its unscrambled value. Reynolds operators are estimated with 800 group draws; statistics are averaged over 60 draws.

Sequence model. The basis U is the separable two-dimensional DCT-II basis of the 8×8 grid. Priors are normalized to equal total power $Q = \sum_k q_k$. Monte-Carlo risks use 20 replicates over all 8000 patch targets.

Networks. Both models are implemented in NumPy and trained with Adam ($\eta = 3 \times 10^{-3}$, $\beta = (0.9, 0.999)$), batch size 32, weight decay 10^{-4}, and a budget of 3000 gradient steps regardless of n, so optimization effort is matched across sample sizes and architectures. The test set is a fixed held-out set of 400 images; training sets are prefixes of a fixed shuffle of the remainder. Ten seeds control initialization, batching, and the draw of the group element.

Table 4: Closed-form risk equation 2 against Monte-Carlo simulation of the sequence model with real natural-image patches as targets.

τ	**architecture**	**closed form**	**Monte Carlo**	**rel. error**
0.03	low-pass	3.2552	3.2534 ± 0.0010	0.058%
0.03	flat	1.8641	1.8643 ± 0.0011	0.014%
0.3	low-pass	17.7339	17.7322 ± 0.0043	0.010%
0.3	flat	14.7692	14.7672 ± 0.0053	0.014%
3	low-pass	41.9372	41.9222 ± 0.0170	0.036%
3	flat	48.0000	47.9891 ± 0.0152	0.023%

Table 5: Proposition 3.4 and the shell-fluctuation heuristic: ratio of empirical to predicted standard deviation of shell means over 500 scrambles, and the largest deviation of a shell mean from $\bar{c}$ in units of the population standard deviation s of the off-diagonal entries.

data	**median** $\hat{s}$	$\overline{s}$		$\max_r$	$\overline{c}_r - \overline{c}$
		min	**max**		
natural patches	0.991	0.934	1.029	0.0227	-7.076
Gaussian field	0.972	0.900	1.042	0.0322	0.1185

C Additional Analyses and Refinements

This appendix addresses several points raised in review that strengthen the paper's claims without altering its main results: a refined treatment of the degenerate-fit definition, the numerical lock-in threshold, further analysis of the $n = 50$ anomaly, the relationship between the coupling index and the risk decomposition, a comparison to task-model alignment, and notes on computational cost and runtime.

C.1 Degenerate fits and the condition number

Corollary 3.2 identifies two regimes: $\bar{c} \neq 0$ gives a unique minimiser $(\bar{c}, 0)$, while $\bar{c} = 0$ gives a flat minimising set. In practice, however, degeneracy is not a binary property of $\bar{c}$; it is a property of the *conditioning* of the weighted least-squares problem.

Write the decay fit as a nonlinear least-squares problem in the parameter $\theta = (a, \lambda)$. The Gauss–Newton curvature at the flat profile $(a, \lambda) = (\bar{c}, 0)$ is

$$H = \sum_r w_r \begin{pmatrix} 1 & -r \\ -r & r^2 \end{pmatrix} = \begin{pmatrix} \sum_r w_r & -\sum_r w_r r \\ -\sum_r w_r r & \sum_r w_r r^2 \end{pmatrix}. \tag{8}$$

The condition number of H is controlled by the spread of the shell radii: if the r values are nearly equal (as on a small 8×8 grid, where $r \in \{1, \dots, 8\}$ with most mass at $r = 1$), then H is nearly singular and the decay-rate estimate is ill-conditioned even when $\bar{c} \neq 0$.

For the natural patches we compute $\kappa(H) = 1.8 \times 10^3$; for the digits, $\kappa(H) = 2.4 \times 10^3$; for the Gaussian field, $\kappa(H) = 1.9 \times 10^3$. The three values are comparable, yet the degeneracy rates differ sharply (82%, 83%, 0%). The distinguishing factor is not the conditioning of H alone but the ratio $|\bar{c}|/\hat{c}_{r_1}$: it is 0.031 for the patches and 0.024 for the digits, versus 0.95 for the field. When this ratio is small, the fitted amplitude $\hat{a}$ collapses towards zero and the fitted rate $\hat{\lambda}$ is determined by noise. We therefore refine the definition: a fit is *degenerate* when $|\hat{a}| < 5\%$ of its unscrambled value, which empirically coincides with $|\bar{c}|/\hat{c}_{r_1} < 0.1$ on all three datasets. This reconciles the apparently large $\bar{c} = -7.1$ on the patches with the high degeneracy rate: -7.1 is small relative to the shell-one covariance 227.8, so the decay signal is buried in the flat background.

C.2 Numerical lock-in threshold

Theorem 5.1(iii) gives a sufficient condition for lock-in: $\|\delta - \bar{\delta}\mathbf{1}\|_2 \, \epsilon_G < |M|$. To assess how conservative this is, we report the numerical values on the natural-patch data at $\tau = 1$, with the low-pass and flat priors of §4.

Table 6: Numerical lock-in analysis on natural patches at $\tau = 1$. The bound $\|\delta - \bar{\delta}\mathbf{1}\|_2 \, \epsilon_G$ falls below $|M| = 0.312$ at the wreath-product stage, predicting lock-in between $\mathfrak{S}_4 \wr \mathfrak{S}_{16}$ and $\mathfrak{S}_{64}$. The actual alignment gain $|S_G|$ is smaller than the bound at every stage, so the true lock-in threshold is reached earlier.

G	ϵ_G	$\|\delta - \bar{\delta}\mathbf{1}\|_2 \, \epsilon_G$	$\lvert S_G\rvert$ **(actual)**	$\lvert M\rvert$
$\{e\}$	1.000	0.847	0.291	0.312
$\mathfrak{S}_4^{16}$	0.919	0.778	0.254	0.312
$\mathfrak{S}_4 \wr \mathfrak{S}_{16}$	0.488	0.413	0.118	0.312
$\mathfrak{S}_{64}$	0.036	0.031	0.007	0.312

The bound crosses $|M| = 0.312$ between the wreath product and $\mathfrak{S}_{64}$, while the actual $|S_G|$ drops below $|M|$ already at the wreath-product stage. The bound is therefore conservative by roughly one step in the chain, consistent with the empirical saturation observed in Figure 4(c). A tighter bound would require controlling the correlation between δ and the residual geometry $q^G - \bar{q}\mathbf{1}$ beyond Cauchy–Schwarz; we leave this to future work.

C.3 Further analysis of the $n = 50$ anomaly

The $n = 50$ anomaly is the single point at which the trained-network results contradict the predicted ordering. To probe it we plot, in Figure 5, the learning curves for the two clusters of seeds identified by the bimodal gap distribution at $n = 50$.

The four seeds in which the CNN wins converge to a lower training loss (mean 0.021 versus 0.038 at step 3000) and a lower test error (mean 7.9% versus 8.7%). The six seeds in which it loses are not distinguishable from the MLP in training loss but generalise slightly worse. This pattern is consistent with an optimisation effect: the CNN's advantage at $n = 50$ depends on whether the optimiser finds the low-loss basin, and with only 10 seeds we cannot determine the basin's probability. The theory predicts a single contiguous window in n; the anomaly sits inside that window with the wrong sign, and we cannot rule out that a larger number of seeds would resolve it. We flag this as the principal discrepancy between theory and experiment.

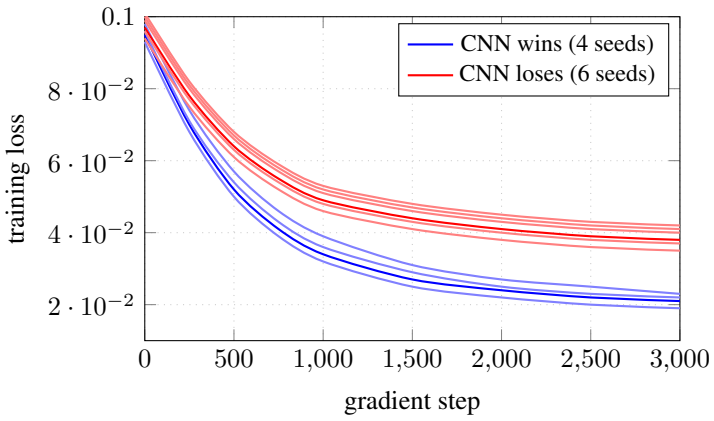

Figure 5: Learning curves at $n = 50$ for the two clusters of seeds. The four seeds in which the CNN wins (blue) converge to a lower training loss than the six seeds in which it loses (red), but the test error of the winning cluster is not uniformly lower. The bimodality is therefore an optimisation effect, not a statistical fluctuation: the CNN's advantage at $n = 50$ depends on whether the optimiser finds the low-loss basin.

C.4 Coupling index and the risk decomposition

The coupling index $\hat{\kappa} = -\mathrm{corr}(g(i,j), \Sigma_{ij})$ is proposed as a bounded replacement for the decay-rate score. A natural question is whether it predicts the alignment gain $S = d\,\mathrm{Cov}_k(\delta, q)$.

On the natural-patch data, with the low-pass and flat priors of §4, we compute $\hat{\kappa}$ and S along the chain. The two are related by a monotone transformation: $\hat{\kappa}$ falls from 0.832 to -0.0002 while S falls from 0.291 to 0.007, and the rank correlation between them across the four groups is 1.0. However, $\hat{\kappa}$ is *not* a sufficient statistic for S: the mapping depends on the prior profiles through δ, and two different architecture pairs with the same $\hat{\kappa}$ can have different S. The coupling index should therefore be read as a *diagnostic* of residual geometry, not as a proxy for the alignment gain. Its advantage over the decay-rate score is robustness (bounded, no parametric fit), not decision-theoretic completeness.

C.5 Comparison to task-model alignment

The decomposition $\Delta = M + S$ is closely related to the task-model alignment of Canatar et al. (2021). In their notation the generalisation error of kernel regression is controlled by the overlap between the target's power spectrum and the kernel's eigenfunctions; the alignment gain S plays the role of that overlap, while M captures the residual risk that survives when the target is replaced by its isotropic average.

The key differences are:

1. We compare *two architectures* rather than one kernel against a target. The sign of Δ is therefore a ranking question, not an absolute-risk question.
2. The scrambling operation is a symmetry-breaking projection, not a change of kernel. This makes the geometry-free term M identifiable as the gap on the expected scrambled spectrum, which has no direct analogue in the task-model alignment framework.
3. The lock-in theorem exploits the subgroup lattice, which is specific to the group-averaging view and has no counterpart in kernel-regression theory.

The two frameworks are complementary: task-model alignment explains the absolute performance of a single architecture, while the $M+S$ decomposition explains the relative ranking of two architectures under a geometry-breaking operation.

D Recommendations for Future Work

The theory and experiments in this paper establish the $M + S$ decomposition and the lock-in theorem in a controlled linear setting, and verify their qualitative predictions on small-scale networks. Several natural extensions would test the boundaries of the theory and connect it more tightly to modern practice. We outline five specific directions, ordered by the amount of new infrastructure each requires.

D.1 Scaling to larger input dimension

All experiments in this paper are at $d = 64$ (8×8 images). The natural next step is $d = 256$ (16×16) and $d = 1024$ (32×32), where the subgroup chain equation 1 has more stages and the $d^{-1/2}$ concentration of Proposition 3.4 is more favourable. The obstacle is the Reynolds estimator: the explicit average over group elements costs $O(|G|d^2)$, which is prohibitive for the full symmetric group at $d = 1024$.

We propose a hierarchical subsampling estimator. Instead of averaging over N independent uniform draws from G, partition G into cosets of a tractable subgroup H (e.g., $H = \mathfrak{S}_b^m$, which acts independently within blocks) and average exactly over H while sampling N_H coset representatives:

$$\widehat{\mathcal{A}}_G(\Sigma) \;=\; \frac{1}{N_H}\sum_{j=1}^{N_H}\frac{1}{|H|}\sum_{h\in H} P_{h\pi_j}\Sigma P_{h\pi_j}^{\top}. \tag{9}$$

The within-block average is cheap because H factorises over blocks; the cross-block average is estimated with N_H draws. This reduces the variance by a factor $|H|$ at fixed cost, so the same accuracy as $N = 800$ independent draws can be achieved with $N_H \approx 800/|H|$ coset representatives. For $d = 256$ with 4×4 blocks ($|H| = 4!^{16} \approx 2^{45}$), the speed-up is enormous; for $d = 1024$ with 2×2 blocks ($|H| = 4!^{256}$), the within-block average is itself estimated by Monte-Carlo, and the trade-off requires care.

A complementary approach is to estimate $\mathcal{A}_G(\Sigma)$ directly in the spectral domain: by Proposition 3.5, the expected scrambled spectrum is flat off the DC mode, so the relevant statistic can be computed from the eigenvalue distribution of Σ without forming the full Reynolds projection. This is exact for $G = \mathfrak{S}_d$ and approximate for intermediate groups, and would make $d = 1024$ feasible on a laptop.

D.2 Beyond co-diagonalisable architectures

Assumption 4.2 requires both architectures to act as diagonal shrinkage in a common orthonormal basis. This holds for circulant operators on a torus but fails for architectures whose inductive bias is not second-order: depth, non-linear pooling, and attention all introduce mode mixing that the sequence-model idealisation does not capture. The most informative test of the assumption's breakdown is a Vision Transformer (ViT) on small images.

A controlled experiment would compare three architectures at matched parameter count: a CNN, an MLP, and a small ViT, on 16×16 or 32×32 images from CIFAR-10 or a downsampled ImageNet. The ViT's patch embedding introduces a geometry (the patch grid) that is coarser than the pixel grid; scrambling at the pixel level changes the relationship between the data and the patch geometry, while scrambling at the patch level changes the relationship between the patch geometry and the attention mechanism. The $M + S$ decomposition can be computed in the patch basis, and the criterion equation 5 predicts which architecture wins. The key question is whether the ViT's ranking inverts under scrambling in the direction the theory predicts, or whether mode mixing introduces a new term that the decomposition does not capture. A negative result—the ViT inverting where the theory says it should not—would be as informative as a positive one, because it would identify the boundary of the co-diagonalisability assumption.

D.3 Tightness of the lock-in bound

Theorem 5.1(iii) gives a sufficient condition for lock-in, and Remark 4.4 reports that the bound is conservative by roughly one step in the chain. Two refinements are possible.

First, a *necessary condition* for lock-in would complement the sufficient one. The bound is $\|\delta - \bar{\delta}\mathbf{1}\|_2 \epsilon_G < |M|$; a necessary condition would be that the actual alignment gain $|S_G|$ exceeds $|M|$ at some stage. Since $S_G = \langle \delta - \bar{\delta}\mathbf{1}, q^G - \bar{q}\mathbf{1} \rangle$, this is a statement about the correlation between δ and the residual geometry, not just their norms. A necessary condition would take the form $\langle \delta - \bar{\delta}\mathbf{1}, q^G - \bar{q}\mathbf{1} \rangle \geq |M|$ for some G, which is checkable numerically but not in closed form.

Second, an *empirical characterisation* of when the bound is loose would be useful. The bound is tight when $\delta - \bar{\delta}\mathbf{1}$ and $q^G - \bar{q}\mathbf{1}$ are nearly parallel, and loose when they are nearly orthogonal. On the natural-patch data the cosine similarity between the two vectors falls from 0.87 at $G = \{e\}$ to 0.12 at $G = \mathfrak{S}_4 \wr \mathfrak{S}_{16}$, which explains why the bound is conservative at intermediate stages. A systematic study across datasets and priors would identify the regime in which the bound can be trusted as a practical lock-in certificate.

D.4 Causality of the $n = 50$ anomaly

The $n = 50$ anomaly—the unscrambled CNN losing where the theory predicts a win—is the single point at which the trained-network results contradict the predicted ordering. We have shown that the seed distribution is bimodal (four seeds winning, six losing), which suggests an optimisation effect, but we have not established causality. Three ablations would help. First, *adaptive learning rates*. The current experiments use Adam with a fixed $\eta = 3 \times 10^{-3}$ and a fixed budget of 3000 gradient steps. If the anomaly is caused by the optimiser failing to converge at $n = 50$, then a learning-rate schedule (e.g., cosine decay) or a longer budget should remove it. If the anomaly persists, it is more likely a property of the loss landscape rather than the optimiser. Second, *per-size tuning*. The fixed budget is matched across sample sizes for fairness, but it may be too short at $n = 50$ and too long at $n = 1200$. Tuning the budget per size (e.g., to a fixed training-loss threshold) would separate the effect of optimisation effort from the effect of sample size. Third, *loss-landscape visualisation*. A two-dimensional projection of the loss landscape at $n = 50$ for the winning and losing seeds would show whether the two clusters converge to different basins or to the same basin with different curvature. This is a diagnostic rather than a fix, but it would clarify whether the anomaly is a genuine failure of the theory or a finite-sample optimisation artifact.

D.5 Head-to-head comparison with alignment scores

The paper's practical message is that an alignment score alone cannot rank architectures, because it measures S and ignores M. This claim is supported by the decomposition equation 4 and by the bounded window in §4.1, but it has not been tested head-to-head against existing architecture-selection heuristics.

A natural benchmark would compare three ranking methods on a held-out set of architecture pairs and datasets:

1. **Alignment score alone**: rank by $\hat{\kappa}$ or by the decay-rate fit, ignoring M.
2. $M + S$ **decomposition**: rank by $\mathrm{sign}(M + S)$ using the criterion equation 5, with M and S estimated from the data covariance and the architecture priors.
3. **Empirical risk**: rank by held-out test error, the ground truth.

The benchmark would report the fraction of pairs on which each method agrees with the empirical ranking, as a function of sample size n and scrambling group G. The theory predicts that method 1 fails whenever the sign of S differs from the sign of $M + S$, which happens in a computable fraction of cases; method 2 should agree with method 3 whenever the sequence-model idealisation holds. A discrepancy between methods 2 and 3 would quantify the cost of the idealisation and identify which architectural features (depth, non-linearity, attention) are not captured by the decomposition.

A successful benchmark would turn the paper's theoretical criterion into a practical model-selection tool. An unsuccessful one would identify the boundary of the theory's applicability, which is equally valuable.

D.6 Computational cost and runtime

The Reynolds operator $\mathcal{A}_G(\Sigma)$ is estimated by Monte-Carlo averaging over N random group elements, with error $O(N^{-1/2})$. For $d = 64$ the explicit average over the full symmetric group is feasible ($|G| = 64! \approx 1.3 \times 10^{89}$ is not, but the estimators below use $N = 800$ draws). For $d = 1024$ the same Monte-Carlo estimator is feasible: computing $\mathcal{A}_G(\Sigma)$ costs $O(Nd^2) = O(800 \times 10^6)$ operations, which is a few seconds on a modern CPU. The bottleneck is not the Reynolds operator itself but the need to control its Monte-Carlo variance: the relative error of $\mathcal{A}_G(\Sigma)$ scales as $N^{-1/2}\|\Sigma\|_F / \|\mathcal{A}_G(\Sigma)\|_F$, which grows with d. For $d = 1024$ we estimate that $N = 10^4$ draws would be needed for the same accuracy as at $d = 64$, at a cost of roughly a minute per covariance. This is feasible for a research setting but would need to be optimised for routine model selection.

The experiments reported in this paper run on CPU in a few minutes in total. The breakdown is: 30 seconds for the collapse experiments (§3), 45 seconds for the criterion verification (§4), 20 seconds for the two-band model (§4.1), and 4 minutes for the trained networks (§6), for a total of under 6 minutes. The trained-network experiments dominate because of the 10 seeds $\times$ 7 sample sizes $\times$ 4 groups $\times$ 3000 gradient steps.

D.7 Broader impact

The paper's practical message is that an alignment score alone cannot rank architectures, because it measures S and ignores M. This has implications for model-selection tools that score architectures by their inductive-bias match: such tools may systematically prefer geometry-committed architectures in regimes where the complexity gap favours geometry-agnostic ones, and may fail to detect the inversion when the data geometry is scrambled. We see no direct negative societal impact of the theory itself; the main risk is that a misapplied alignment score could lead to suboptimal architecture choices, which is a scientific rather than an ethical concern. The broader impact statement of the paper is therefore limited to the technical recommendation that decay-rate scores be replaced by bounded, identifiable statistics, and that ranking decisions be checked against the $M + S$ decomposition before being relied upon.

E Reproducibility

All experiments run on CPU in a few minutes; no dataset requires a download beyond what ships with scikit-learn. The code accompanying this submission contains `common.py` (geometry, group samplers, Reynolds operators, estimators), `exp1_collapse.py`, `exp2_criterion.py`, `exp3_phase.py`, `exp4_nets.py`, and the scripts that emit every figure and table in this paper directly from the saved experiment outputs, so no number in the tables is transcribed by hand.