SYMMETRY, NOT DIMENSION: WHEN SPARSE-CODING ATOMS IN A FEATURE PLANE ARE IDENTIFIABLE

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse autoencoders trained from different seeds disagree, which concurrent work attributes to feature manifolds of dimension ≥ 2; in the sharpest analysis, from an isotropy assumption. For a feature plane, the sparse-coding objective under a rigid rotation of its atoms is a *circular cross-correlation* between the *angular density* and a configuration-dependent cost kernel, flat along the rotation iff the density's relevant Fourier coefficients vanish, else pinned with computable curvature. Under L_1, dimension ≥ 2 is necessary for that flatness, not sufficient. On densities from models' own rings, twelve-seed dictionaries meet predictions the density condition makes and symmetry cannot: agreement at an installed harmonic (strongest in 28 of 35 Gemma-2-2B tilts at $q \geq 2$ no symmetry explains; chance 8.8), and its loss when one Fourier coefficient is removed from a spread density, including the number circles of Llama-3.1-8B and Pythia-6.9B, but not from clustered rings. Only ReLU+L_1 seeds reliably sit near their landscape's minimum; TopK's in-plane optimum at $k \geq 2$ is flat on full-support densities. On released dictionaries and our own, ablating a dictionary's plane for a concept transfers across dictionaries in all 20 cells, its most selective latent in 7, failing where dictionaries disagree on its item. Making the data exactly symmetric in one plane removes the seeds' agreement on its atoms (matched atoms $2^\circ \to 16^\circ$ apart) and reinstalling one harmonic restores it; a plane found in natural text without probes transfers too. Claims should attach to the plane, not the atom. Code, records and some trained dictionaries are in the supplement; all will be public at `https://github.com/xxx/xxx` upon acceptance.

1 INTRODUCTION

Sparse autoencoders are the standard instrument for decomposing language-model activations into interpretable features (Cunningham et al., 2024; Bricken et al., 2023; Gao et al., 2025; Lieberum et al., 2024), and practice treats individual latents as *the* features of the model—catalogued, named, used for causal interventions (Marks et al., 2025)—on an identifiability assumption that does not hold in general: SAEs differing only in initialization can share as few as $\sim 30\%$ of their latents (Paulo & Belrose, 2026). Features can live on manifolds of dimension ≥ 2 (Engels et al., 2025), and concurrent work converges on a diagnosis: a multi-dimensional manifold admits many equally good bases (Gerasimov et al., 2026; Bhalla et al., 2026; Dalili & Mahdavi, 2026).

The diagnosis is incomplete, and the proofs show where. The sharpest version (Dalili & Mahdavi, 2026), on the population ℓ_1 risk, proves both of its results on activations whose direction is *uniform* in the subspace (their Eq. 4), and uses dimension only through the symmetry it creates; the synthetic demonstrations vary dimension, not the law (Bhalla et al., 2026; Gerasimov et al., 2026). Dimension and symmetry are confounded throughout, and predict differently (§4.1).

This paper replaces the dimension diagnosis with a criterion on the density. For a feature plane carrying m dictionary atoms, the population sparse-coding objective as a function of a rigid rotation φ of those atoms is a circular cross-correlation

$$J(\varphi) \;=\; (\rho \star \bar{c})(\varphi)$$

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.

between the *angular density* ρ of activations in the plane and a cost kernel $\bar{c}$ set by the atom configuration, so J is exactly flat iff $\hat{\rho}_k\,\hat{\bar{c}}_k = 0$ for every $k \neq 0$ (Proposition 3). Dimension ≥ 2 is necessary for the rotation to exist and, for a non-vanishing kernel, not sufficient for it to be flat; a plane whose angular density carries a harmonic its configuration couples to is not flat under rotation. The criterion is about one mode of the solution set, the rigid rotation the dimension diagnosis says is free: a flat rotation at an optimum is a continuum of optima, and with free atoms we found one only under isotropy (for L_1) and for TopK at $k \geq 2$, whose kernel vanishes (§4). Elsewhere what it supplies is the pinning strength, and that is what our tests measure. The criterion is quantitative (pinning strength is, generically, the curvature, $\approx 4\pi m^2 |\hat{\rho}_m||\hat{\bar{c}}_m|$ where the m-th harmonic dominates, Proposition 6), it has a graded form beyond the regular polygons the sharp version assumes (a configuration's *structure factor*, Proposition 4, exact under a sparsity hypothesis trained dictionaries do not meet), and it is a statement about *which* configurations a given density can pin, not about complexity: with the density fixed, a dictionary with *more* atoms can be *more* stable if its atom count lands on a harmonic the density carries. The mathematics is elementary—the contribution is the separation it makes and the tests that follow from it, not new theory.

What we test, and what could have failed. Because Proposition 3 is an identity, convolving a measured density with a measured kernel confirms it by construction; we separate illustrations from tests throughout, and the contributions are the tests:

- *Controlled.* At fixed dimension two, cross-seed jitter tracks the density's circular moment over an $8.6\times$ jitter range (Spearman -1.00 over six conditions), the predicted non-monotonicity in m replicates over 8 of 8 generators (Corollary 1's on-resonance case holds; off resonance trained seeds are graded), and a modular-arithmetic transformer with a known symmetry group has its objective deepen from flat to pinned as that symmetry is broken, in the representation ($5/5$ defect levels) and by dose ($3/3$; seeds unresolved).
- *Which coding rules inherit it.* The criterion constrains an objective, and only one rule's seeds reliably reach its pinned optimum: ReLU+L_1 seeds agree where it says in 20 of 21 cells and at every grid point, disagree on the flat control, and sit near their own landscape's minimum; TopK-1 and JumpReLU seeds sit where a random rotation would (§4.3).
- *A pretrained model's own geometry.* On Gemma-2-2B's weekday, month and clock-hour rings, whose density we measure rather than choose, twelve-seed dictionaries test four predictions fixed before training and one regularity (Table 2): the lock below the lattice (weaker at the order fixed in advance) and at an installed harmonic (28 of 35 with $q \geq 2$ where no symmetry explains it) hold, the lattice and λ rows weakly, and removing one Fourier coefficient unpins its own order on spread densities—two constructed, and the number circles of Llama-3.1-8B and Pythia-6.9B—but not on the clustered rings; the main contrast replicates in the full residual stream and a second model (§5.1). *Which* order locks is not predicted by the density alone.
- *A recommendation, at production scale*, which Proposition 1 motivates without the Fourier criterion and which we test empirically. Across twelve released GemmaScope SAEs, ablating the plane a dictionary assigns to weekdays reproduces across dictionaries ($+0.815$, above a matched control for 12 of 12) where ablating its most selective latent does not (-0.04). Over all 20 cells we ran—three layers and three concepts, the suite of Paulo & Belrose (2026), Llama Scope (He et al., 2024), Gemma-2-9B, and dictionaries we trained from six seeds on unprompted text—the plane beats its control in every one and the atom in 7 (§6).

Where a plane's rotation is left free, what can replicate is the plane and its atoms' relative geometry, not their phase: one way representation-level convergence (Kornblith et al., 2019; Roeder et al., 2021) coexists with partial feature-level agreement.

2 Related work, and where the prior indeterminacy comes from

SAE feature instability. Paulo & Belrose (2026) established the cross-seed instability phenomenon and its width/L_0 trends, Fel et al. (2025) report it for vision SAEs, Song et al. (2026) argue consistency should be a first-class metric, and Leask et al. (2025) show different widths do not yield canonical units. Three concurrent preprints attribute it to manifolds of dimension ≥ 2 (Gerasimov et al., 2026; Bhalla et al., 2026; Dalili & Mahdavi, 2026). Gerasimov et al. (2026) report the dissociation whose instability half Proposition 1 formalizes (unstable latents "concentrate in reproducible lower-rank subspaces"); *whether* a given plane's rotation is pinned, and how strongly, is what Propo-

sition 3 quantifies. Lan et al. (2024) find SAE feature spaces shared across models at the level of subspaces, and supervised subspaces match or beat SAE latents at isolating attributes (Chaudhary & Geiger, 2024) and steering (Wu et al., 2025); §6 is a symmetry-motivated case of both. Nelson et al. (2026a;b) give guarantees under sparsity and support-diversity assumptions outside our regime.

Where the symmetry enters. For the sharpest of the three, Dalili & Mahdavi (2026), the within-manifold law is a uniform slice (their Eq. 4), which forces a repeated eigenvalue: the eigenspace their Theorem 11 recovers is free to rotate exactly because the eigenvalue repeats, and their Theorem 8's constant is set by the uniform measure of the positive orthant. Neither result uses dimension except through the symmetry it makes room for (App. J); in a 2-plane, Proposition 3 says what decides it.

Identifiability theory. In ICA, Gaussian (spherically symmetric) sources are identifiable only up to rotation, and independent subspace analysis leaves the basis inside a subspace unidentified (Comon, 1994; Cardoso, 1998; Hyvärinen & Hoyer, 2000; Hyvärinen & Oja, 2000; Theis, 2006); varimax identifies leptokurtic factors (Rohe & Zeng, 2023), and disentanglement needs inductive bias (Locatello et al., 2019). Ours concerns the sparse-coding objective, not the likelihood, and replaces that all-or-nothing condition by a harmonic one (Corollary 1). Amplitude–phase sparse coding makes the phase the free coordinate (Cadieu & Olshausen, 2012); NMF and dictionary identifiability rest on support conditions, and Whittington et al. (2023) select bases for independent bounded sources.

Symmetry and degeneracy in interpretability. Regular polygon configurations in low-dimensional subspaces are familiar from the superposition literature (Elhage et al., 2022), where they arise from packing pressure, and categorical concepts are represented as polytopes (Park et al., 2025); our criterion says when one is pinned by the data. Hindupur et al. (2025) ask how each coding rule's bias meets a concept's geometry, Modell et al. (2025) where such manifolds come from; indeterminacy groups are derived from architecture (Godfrey et al., 2022) and parameter symmetries (Şimşek et al., 2021; Ziyin, 2024; Kunin et al., 2021), where we derive them from the data, and Higgins et al. (2018)'s symmetry-based disentanglement is the closest precedent (the phase as a group coordinate: Cohen & Welling, 2014; Sanborn et al., 2023). Non-uniqueness of mechanistic explanations is documented for circuits and causal abstractions (Méloux et al., 2025; Makelov et al., 2024; Sutter et al., 2025; Sun & Toneva, 2026); in modular arithmetic (Nanda et al., 2023; Chughtai et al., 2023) exact symmetry makes the SAE loss invariant over the group (§4.2). Feature splitting and absorption (Chanin et al., 2025; Bussmann et al., 2025) are the established account of the width and L_0 trends, which our criterion does not replace (§3.3); Michaud et al. (2025) model how SAEs tile manifolds, and critiques (Heap et al., 2025; Kantamneni et al., 2025) question what latents mean.

3 THE SYMMETRY CRITERION

3.1 SETUP: THE INDETERMINACY GROUP OF THE DATA

Activations $x \in \mathbb{R}^d$ follow a law P; a dictionary is $D = [d_1, \dots, d_M] \in \mathbb{R}^{d\times M}$ with unit columns, and non-negative sparse coding assigns the cost

$$c(x; D) \;=\; \min_{a\geq 0} \; \tfrac{1}{2}\|x - Da\|_2^2 + \lambda\|a\|_1, \qquad J(D) \;=\; \mathbb{E}_{x\sim P}\,[\,c(x; D)\,]. \tag{1}$$

Propositions 1 and 3 and Eq. 2 hold, with $\bar{c}$ built from that cost, for any per-sample cost that is *isometry-invariant*, $c(gx; gD) = c(x; D)$ for $g \in O(d)$—in particular for the JumpReLU (MSE + L_0) and TopK objectives of production SAEs (for TopK, with one active plane per sample; App. A); Proposition 4 uses the ℓ_1 kernel, and existence of an optimum is claimed for ℓ_1 with $\lambda > 0$ only.

Definition 1 (Indeterminacy group)**.** *$G(P) = \{g \in O(d) : g_{\#}P = P\}$, the isotropy group of P, for P with $\mathbb{E}\|x\|^2 < \infty$ (for symmetry about a center $\mu \neq 0$, see App. A).*

Proposition 1 (Orbit structure)**.** *The solution set $S = \arg\min_D J(D)$ is $G(P)$-invariant: $D \in S$ and $g \in G(P)$ imply $gD \in S$. Atoms are identifiable at best up to $G(P)$ (and permutation).*

Proposition 2 (The amortized SAE objective inherits the symmetry)**.** *For SAE parameters $\Theta = (W_e, b_e, W_d, b_d)$ define $g \cdot \Theta = (W_e g^\top,\, b_e,\, gW_d,\, gb_d)$. Then the encoder output $z_\Theta(x) = \sigma(W_e(x - b_d) + b_e)$ satisfies $z_{g\cdot\Theta}(gx) = z_\Theta(x)$ and the loss obeys $L(g \cdot \Theta) = L(\Theta)$ for $g \in G(P)$ (uncentered data: $gb_d \mapsto \mu + g(b_d - \mu)$).*

Proofs (App. A) are short consequences of equivariance. Proposition 1 moves "is this feature identifiable?" from the feature's dimension to $G(P)$, a property of the *data*: a continuous rotation in

$G(P)$ that moves an optimal atom leaves a circle of equally good optima. We call a plane's atoms identifiable when the optima restricted to it are isolated modulo $G(P)$; Proposition 3 tests this along the rigid-rotation mode.

3.2 Two-dimensional planes: flatness is a Fourier condition

Let a feature occupy a plane $V \cong \mathbb{R}^2$; write its activations in polar form $h = r\,u(\theta)$ with angular marginal ρ on $[0, 2\pi)$ and radial conditional independent of θ (App. A states when the dependent case reduces to this one, and App. H tests it). Let the plane carry m unit atoms at angles $\alpha = (\alpha_1, \ldots, \alpha_m)$ and let $\bar{c}(\psi) = \mathbb{E}_r\big[\min_{a\geq 0} \frac{1}{2}\|r\,u(\psi) - D(\alpha)a\|^2 + \lambda\|a\|_1\big]$ be the radius-averaged cost of a point at angle ψ. Rotating all atoms rigidly by φ shifts the kernel (Lemma 2, App. A), giving the objective restricted to the plane as a function of the rotation:

$$J(\varphi) \;=\; \int_0^{2\pi} \rho(\theta)\,\bar{c}(\theta - \varphi)\,d\theta \;=\; (\rho \star \bar{c})(\varphi) \;=\; 2\pi \sum_{k\in\mathbb{Z}} \overline{\hat{\rho}_k}\,\hat{\bar{c}}_k\,e^{-ik\varphi}. \tag{2}$$

($\hat{f}_k = \frac{1}{2\pi}\int f e^{-ik\theta}$; the bar on $\hat{\rho}_k$ is conjugation, the one in $\bar{c}$ part of its name.)

Equation 2 is the plane's own contribution to J, equal to the full objective up to a constant if the plane's atoms code only in-plane inputs (approximately so in §4.1). Otherwise a rotation also moves the cost of outside inputs that borrow the atoms, a cross-term that scales with their reconstruction share elsewhere and can dominate production read-outs (App. E).

Proposition 3 (Flatness criterion)**.** *With r independent of θ, J is constant on the circle iff $\hat{\rho}_k \hat{\bar{c}}_k = 0$ for every $k \neq 0$. In particular: (i) if ρ is uniform, J is flat for* every *atom configuration, so the rotation leaves the plane's atoms undetermined; (ii) if $\hat{\rho}_{k_0} \neq 0$ for some $k_0 \neq 0$, then J is not constant for any configuration with $\hat{\bar{c}}_{k_0} \neq 0$.*

The condition in (ii) is on the configuration, λ and the radial law, checkable by the exact solver; $\hat{\bar{c}}_{k_0}$ can vanish, e.g. for a regular m-gon, $m \nmid k_0$ (Corollary 1), or at isolated λ/r (Proposition 4).

Corollary 1 (Harmonic selectivity)**.** *If the atoms form a regular m-gon, $\hat{\bar{c}}_k = 0$ unless $m \mid k$, so only the harmonics $\hat{\rho}_m, \hat{\rho}_{2m}, \ldots$ can pin the rotation. For $\rho(\theta) = (1 + a\cos\ell\theta)/2\pi$ ($|a| \leq 1$), whose only non-constant harmonics are $k = \pm\ell$, the rotation is* exactly *flat unless $m \mid \ell$, however non-uniform ρ is. For ρ equally weighted on p equally spaced angles, $\hat{\rho}_k = 0$ unless $p \mid k$, so J has no harmonic below $\mathrm{lcm}(m, p)$: for $\gcd(m, p) = 1$ the first surviving harmonic is $k = mp$ and J stays within $4\pi\sum_{j\geq 1}|\hat{\rho}_{jmp}||\hat{\bar{c}}_{jmp}|$ of its mean, however concentrated the density is.*

No trained dictionary is exactly regular, so beyond one test of the binary rule (App. C) our experiments test the graded version.

Proposition 4 (Structure factor)**.** *For data at radius r, if $\lambda > 0$ is large enough that at most one atom is ever active (sufficient: $\lambda/r \geq \cos(\delta/2)$ for minimum gap δ), then for $q \neq 0$, $\hat{\bar{c}}_q = -\frac{m}{2}\hat{g}_q S_q e^{iq\psi_q}$, where $S_q e^{iq\psi_q} = \frac{1}{m}\sum_j e^{-iq\alpha_j}$ is the configuration's* structure factor *and $g(u) = \max(0, r\cos u - \lambda)^2$. The q-th harmonic of J then vanishes iff $\hat{\rho}_q = 0$, $\hat{g}_q = 0$, or $S_q = 0$; a regular m-gon has $S_q = \mathbf{1}[m \mid q]$, recovering Corollary 1 (App. A).*

S_q is computable from a decoder, given the plane, and replaces the binary $m \mid q$ test by a graded one: a trained dictionary, never regular, generically has $S_q > 0$ at every q and, on a density whose power lies off $m\mathbb{Z}$, is *more* strongly pinned than the idealization; App. I finds that ordering against regular polygons, though random angles do as well. The sparsity hypothesis is load-bearing: inside it the identity holds to quadrature accuracy (8×10^{-7} over 57 constructed configurations, App. B), but at most 5 of 155 GemmaScope configurations meet it at any λ we tried, outside it its *amplitude* is off by a median 42–129% (App. I), and on the ring dictionaries of §5.1 its *phase* survives to $4.9°$ where $S_q \geq 0.3$ (App. K): a graded ordering, not a formula.

Proposition 5 (Dimension is necessary, not sufficient)**.** *In a one-dimensional feature subspace no continuous indeterminacy exists ($O(1)$ is discrete, and for one-signed data with $P(r > \lambda) > 0$ an atom on the line strictly prefers the aligned sign). Together with Proposition 3: dimension ≥ 2 makes a continuous rigid rotation possible, and a* flat *one needs $\hat{\rho}_k \hat{\bar{c}}_k = 0$ for every $k \neq 0$, which fails whenever $\hat{\rho}_{k_0} \neq 0$ and $\hat{\bar{c}}_{k_0} \neq 0$ for some $k_0 \neq 0$; e.g. one atom has $\hat{\bar{c}}_1 \neq 0$ for every $\lambda < r$, so*

any ρ with $\hat{\rho}_1 \neq 0$ pins it (App. A). A kernel that vanishes, as for TopK at $k \geq 2$ on a regular m-gon, $m \geq 3$ (§4.3), is flat for every ρ. Both halves concern the in-plane rotation; we prove nothing about a one-parameter subgroup of $G(P)$ that does not rotate a single plane.

3.3 From flatness to a measurable quantity

Proposition 6 (Curvature formula)**.** *Differentiating Eq. 2 twice gives $J''(\varphi) = -2\pi \sum_{k\neq 0} k^2\, \overline{\hat{\rho}_k} \hat{\hat{c}}_k e^{-ik\varphi}$, which is ≥ 0 at any minimizer φ^*. Keeping only $k = \pm m$, $J''(\varphi^*) = 4\pi m^2 |\hat{\rho}_m||\hat{\hat{c}}_m| + E$ with $|E| \leq 4\pi \sum_{k\geq 1, k\neq m} k^2 |\hat{\rho}_k||\hat{\hat{c}}_k| + \eta_1^2\, 4\pi m^2 |\hat{\rho}_m||\hat{\hat{c}}_m|$, $\eta_1 = \sum_{k\geq 1, k\neq m} k|\hat{\rho}_k||\hat{\hat{c}}_k|/(m|\hat{\rho}_m||\hat{\hat{c}}_m|)$, at a minimizer within a quarter period of the m-term's, for $|\hat{\rho}_m \hat{\hat{c}}_m| > 0$ and a convergent sum (App. A), so the approximation $J''(\varphi^*) \approx 4\pi m^2 |\hat{\rho}_m||\hat{\hat{c}}_m|$ is controlled when η_1 and the k^2-weighted remainder are small. Measured densities from §4 on are reported as circular moments $R_k = 2\pi|\hat{\rho}_k| = |\mathbb{E}_\theta[e^{-ik\theta}]| \in [0, 1]$ (written $|\hat{\rho}_k|$ in results), in which the leading term reads $2m^2 R_m |\hat{\hat{c}}_m|$.*

Two modeling consequences of Eq. 2, each with its own assumptions, organize the experiments. A *fluctuation relation*: the learned phase spread decreases in J''; a Boltzmann model $p(\varphi) \propto e^{-J(\varphi)/T}$ over the *measured* landscape is the small-noise stationary law of constant-step SGD (Mandt et al., 2017) (0.87–1.05 of the predicted variance, nothing fitted) and phenomenological for annealed training, where one global T fits the synthetic regimes (App. C; not the ring's, App. K). A *tiling relation*: where $|\hat{\rho}_m||\hat{\hat{c}}_m|$ decays faster than m^{-2}, landscapes flatten as more atoms tile a manifold, so the predicted variable is m, the atoms *per manifold*, not L_0 or width. **That cuts against us:** raising L_0 *lowers* m on production SAEs (App. F), so the criterion predicts agreement *improving* with L_0, against the published trend, unresolved here (App. E).

A single-dictionary diagnostic follows: rotate a group of atoms rigidly in their decoder plane and record the loss—the *rotation landscape*, with its *depth* (range) and *curvature* at the minimum (App. C, given the true plane; it fails its baselines at production scale, App. E).

4 Controlled tests

In a 2-plane some optimal non-negative code has at most two active atoms, so Eq. 1 is solved in closed form and every static claim of §3 is checkable without learning: the identity holds to 4×10^{-12} relative, Corollary 1 is correct in $24/24$ cells with the off-resonance variation zero to machine precision (and, at the λ tested, pinning on resonance), and the curvature series matches finite differences on single-harmonic densities to 1.1×10^{-4} (App. B). One subtlety governs everything below: with *free* atom angles, exact ReLU+L_1 flatness survived only for the isotropic density in every case we tested; the off-resonance regular m-gons are not free optima (non-stationary unless $m \mid 2\ell$, saddles where stationary at ℓ=3), and the free optimum is irregular and sits at one phase modulo the data's symmetry (App. B). Away from exact symmetry the quantity to measure is curvature.

4.1 Dimension held fixed, identifiability swept

We construct data as sparse mixtures of six 2-planes in $\mathbb{R}^{48}$ whose population angular density we set, train ReLU+L_1 SAEs from 20 seeds per condition, and measure the cross-seed spread of their atom phases in each plane (App. C). The dimension is 2 in every phase-bearing condition, so none of the variation below is dimensional: (1) cross-seed jitter falls with the circular moment R_m (Spearman -1.00 over the 6 conditions, the independent units, whose means span $8.6\times$; -0.95 over their 36 planes; -1.00 again with each seed on its own data draw and frame; Fig. 2, App. C); (2) harmonic selectivity manifests, including the non-monotone signature in m that no scalar-complexity account produces (8 of 8 independent generators; App. C), but graded: with the phase read modulo the data's symmetry (Proposition 1), at m=4 the on-resonance ℓ=4 is pinned sharply (0.11) and ℓ=8 only partly (0.51), ℓ=3, 6 at chance (0.92, 0.95; this reading cannot settle the graded rule there) and ℓ=2 partly pinned (0.49), consistent with a free dictionary escaping a flat mode by *leaving* the polygon (App. C); (3) a one-temperature Boltzmann model on the measured landscape predicts the jitter ($R^2 = 0.58$ held out by condition, below 0.61 for R_m alone; 0.18 for m), and the single-SAE *curvature* ranks it as well as the exact oracle. A 1-D control has perfect cross-seed sign agreement, as Proposition 5's optimum forces, and each seed recovers the *plane* itself (App. C).

Table 1: **Only ReLU+L_1 seeds agree where the objective is pinned** (TopK-2's optimum is flat, App. P). Cross-seed agreement A of the configuration's phase at the density's harmonic (median over cells; 12 seeds; chance ≈ 0.26), over $M \in \{4, 6, 9\}$; "pinned > flat" is the pre-specified contrast (21 cells against 3 controls; exact one-sided Mann–Whitney, floor $p = 0.0005$). The last two columns are App. M's hyperparameter grid (hexagon, $M = 6$) against its matched control (descriptive, exact); TopK-2's and tied JumpReLU's grid p is carried by one of the grid's two installed harmonics. Kernel: $\max_{1 \le q \le 12} |\hat{\hat{c}}_q|$ at the rule's own seeds (median of 288; JumpReLU at $\theta_0 = 0.05$; App. P).

coding rule	$\max_q \lvert\hat{\hat{c}}_q\rvert$	A pinned	A uniform	cells sig.	pinned > flat	grid: A installed	pinned > flat, grid
ReLU+L_1	8.4×10^{-3}	**0.980**	0.223	**20/21**	$p = 0.0010$	**0.98** [0.97, 1.00]	$p = 7.8 \times 10^{-5}$
TopK, $k = 1$	1.3×10^{-1}	0.357	0.282	7/21	$p = 0.20$	0.21 [0.15, 0.70]	$p = 0.54$
TopK, $k = 2$ (non-neg.)	2.7×10^{-4}	0.389	0.470	10/21	$p = 0.47$	0.48 [0.08, 0.66]	$p = 0.011$
JumpReLU, tied θ	9.7×10^{-2}	0.450	0.296	10/21	$p = 0.25$	0.34 [0.05, 1.00]	$p = 2.2 \times 10^{-4}$
JumpReLU, per-atom θ	7.2×10^{-2}	0.355	0.343	5/21	$p = 0.25$	0.31 [0.04, 0.83]	$p = 1.7 \times 10^{-3}$

4.2 A TRAINED NETWORK'S REPRESENTATION, ITS SYMMETRY MADE EXACT

A one-layer transformer trained to 100% accuracy on $(a + b) \bmod 53$ represents the sum on a closed curve built from four Fourier planes (Nanda et al., 2023; Zhong et al., 2023). Circularizing its conditional-mean curve makes the input shift $s \mapsto s{+}1$ an orthogonal map of the curve, so by Proposition 2 the SAE loss on it is exactly $\mathbb{Z}_{53}$-invariant, a sanity check: the landscape over the 53 group elements is flat to a relative 2×10^{-7} (float32 rounding; 2×10^{-15} in float64). At the other end, as one knob interpolates to the network's own activations, the landscape deepens monotonically over all five defect levels, to 0.18 (median, 24 seeds; Fig. 3a, App. D). Reweighting the curve at a harmonic deepens it too (24 seeds per point), but most at the network's key frequencies, not where the dictionary's cost profile predicts (a failed test), while 24 seeds' orbit elements do not separate from uniformity at either end, an underpowered test (App. D).

4.3 WHICH CODING RULES INHERIT THE CRITERION

The criterion constrains an objective, so it transfers to any rotation-equivariant rule—but only through $\hat{\hat{c}}_k$, where rules differ sharply (Table 1). ReLU+L_1 seeds agree where the density pins the objective (on the six densities tilted at the first harmonic, at a phase the exact objective does not choose; App. P) and not on the uniform control, and over a hyperparameter grid with a matched control ReLU+L_1's weakest of 16 points has $A = 0.973$ (3 of 160 other-rule points reach it; App. M). The contrast survives a post hoc drop of every rule's $k{=}1, 3$ cells (low-order moments; 0.977 against 0.24–0.42), embedding each seed's plane in a random basis, which makes training rotation-equivariant in distribution (0.941 against 0.106 uniform), and a statistic with no firing weights or chosen harmonic (16 of 21 cells against 0–5). For TopK-1 and JumpReLU the free-atom optimum is pinned on these densities (21 of 21 cells, 19 for JumpReLU), yet on their own landscapes their seeds sit where a random rotation would (median percentile 0.52–0.63, against 0.05 for ReLU+L_1; App. P): their optimizers—or amortized encoders—stop short of it. TopK at $k \ge 2$ has no pinned optimum to find: two non-negative atoms reconstruct every direction they bracket, so its optimum reconstructs these densities exactly (21 of 21) and is flat, and 184 of 288 trained TopK-2 seeds sit flatter than any ReLU+L_1 seed; for them disagreement is the criterion's prediction.

5 A PRETRAINED MODEL'S OWN FEATURE MANIFOLD

Our controlled experiments build the density. A pretrained model supplies one built for its own reasons: Engels et al. (2025) report that language models place quantities such as weekdays and months on closed curves in a 2-plane, and in Gemma-2-2B (Gemma Team, 2024) the item means of the weekday ($p = 7$), month ($p = 12$) and clock-hour ($p = 12$) probes lie in cyclic order in their top-2 principal plane at $26/26$, $24/26$ and $23/26$ layers (App. I). In that plane the angles are the model's (item weights ours); we can measure the density, symmetrize it (each item cluster translated so its mean sits on a regular p-gon) and tilt it by reweighting contexts by $1 + 2r \cos q\theta_i$ ($r{=}0.35$) at their item's p-gon angle θ_i, *installing* harmonic q. A regular p-gon's harmonics form the lattice $p\mathbb{Z}$; the measured densities carry power below it (App. K).

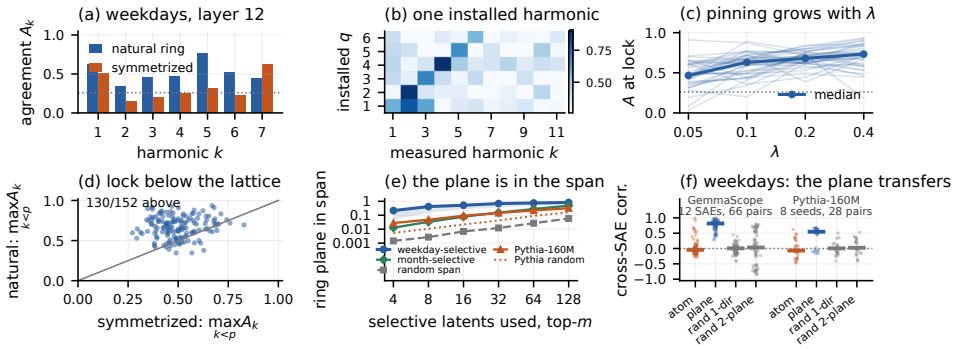

Figure 1: **(a)–(d): the criterion on a pretrained model's own ring** (§5.1); ReLU+L_1 dictionaries, 12 seeds, on Gemma-2-2B's in-plane probe density; A_k is cross-seed agreement at harmonic k, chance ≈ 0.26 at fixed k (dotted). (a) Natural weekday ring (layer 12, $M{=}p$, mean over λ): seeds lock below the lattice, most at $k{=}5$; symmetrized, most orders fall toward chance and $k = p$ rises. (b) An installed q locks $k = q$ (mean A_k over all rings' tilts; at $q{=}1$ the lock goes to $k{=}2$). (c) Agreement at each series' lock order against λ. (d) $\max_{k<p} A_k$ per setting (null median 0.43–0.48). **(e)–(f): the recommendation on dictionaries we did not train** (§6). (e) Fraction of the ring plane inside an SAE's top-m weekday-selective span, against the same rule on months and random spans. (f) Raw cross-SAE correlation of an ablation's effect, one point per pair, median bar.

Table 2: **Four predictions the criterion makes before training, one regularity, and what the seeds did.** Rows 1 and 2 set the $\mathbb{Z}_p$ case of Proposition 1 (symmetrized) against the non-symmetric density; rows 3–4 need the density condition and are where the criterion is on its own (row 3 at $\gcd(p, q) = 1$); row 5 is a regularity (App. T). A is cross-seed agreement *of one Fourier moment's phase*, not of the atom set (§5.1; chance $\approx$ 0.26; row 1 is $\max_{k<p} A_k$, null median 0.43–0.48). Rows 1–3 are 152 settings in 11 cells and 138 tilts, row 4 the projected, constructed 48-point clock density and the number circles of Llama-3.1-8B and Pythia-6.9B (App. K), row 5 within-(cell, M) trends. App. K gives the nulls, counts and controls.

	prediction	measured	check	verdict
1	below the lattice, the natural ring locks	0.67	0.49 symm.	130/152, 11/11 cells; 90 vs 18 lock
2	the lattice order $k = p$ survives symmetrizing	0.36	0.31 nat.	89/152, 8/11 cells, weakly
3	an installed q locks $k = q$	0.74	0.29 other k	strongest 28/35 at $\gcd{=}1$, $q{\geq}2$ (8.8); 113/138 signif.
	at $q = 1$	0.57	0.83 at $k{=}2$	fails (2/27 strongest)
4	removing $\hat{\rho}_{k^*}$ unpins k^*	0.76 → 0.41	sham 0.76 → 0.85	44/48, 36/36 in Qwen; 4/4, 3/3 layers
	number circles (fitted plane)	0.70 → 0.39	A_1 up in only 40/84	70/84, 7/7 layers (Llama)
		0.68 → 0.49	A_1 up in only 36/84	65/84, 7/7 layers (Pythia)
	clustered 7-, 12-point rings	0.30 → 0.44	subsets 7 → 4	fails (9/24)
5	*regularity*: A grows with λ	Spearman +0.80	symm. 20/38	33/38 series, weakly
	which order locks, density alone	13/90	chance 10	fails on clusters
	on the spread 48-point density	33/64	35/64 for $k{=}2$	fails

5.1 Trained seeds lock when the density says

We train ReLU+L_1 dictionaries from 12 seeds on the in-plane density of our probe set—the probes' ring-plane coordinates at layers 1, 3, 7 and 12, on the unit circle, the criterion's setting—with $M \in \{4, 6, k^*, p\}$ atoms, k^* the density's strongest harmonic below p, and $\lambda \in \{0.05, 0.1, 0.2, 0.4\}$, and record the cross-seed agreement A_k at every harmonic $k \leq p$. A_k concentrates the phase of the configuration's k-th *moment*, not the atom set (the seeds occupy a median 9.5 cluster subsets), so what these test is that a density pins the rotation of whatever configuration a seed reaches (atom by atom the effect is small; App. K). A *cell* is a (ring, layer) pair (11); a *setting* a (cell, M, λ) triple (152, since $k^* = 6$ coincides with $M = 6$ in six cells); a *tilt* one installed $q \leq p/2$ in one (cell, $M < p$) at $\lambda = 0.2$ (138); a setting *locks* when its $\max_{k<p} A_k$ clears the 95th percentile of its null. Table 2 and Fig. 1a–d give four predictions derivable before training and none built into the construction—rows 1–2 contrast the $\mathbb{Z}_p$ case of Proposition 1 (approximate: symmetrization leaves residual power) with the non-symmetric density, rows 3–4 what symmetry alone cannot supply—and a regularity, row 5, since $|\hat{\bar{c}}_k|$ rises with λ in only 74% of fixed configurations (App. T).

Row 1's symmetrized side sits above its null (18 settings lock against 90 natural, 8 expected), consistent with the residual harmonics symmetrization leaves; read at the one order fixed before training, k^*, row 1 is weaker (0.31 against 0.24, 9 of 11 cells). Row 3's installed order or its lattice alias is strongest below p in 93 of 138 tilts (chance 27), and in 28 of the 35 with $q \geq 2$ and $\gcd(p, q) = 1$, where no residual symmetry of the tilted density confines the seeds (chance 8.8; flat-optimum TopK-2: 9); at $q = 1$ it fails (2 of 27), and even TopK-2's seeds agree there (26 of 27). Row 4 keeps both sides non-symmetric and cuts one Fourier coefficient—the test a symmetrized control cannot be: on the 48 quarter-hours in the clock plane (spread: items overlap), bounded per-item weights take R_{k^*} from 0.23 to 0.02 and the agreement at k^* from 0.76 to 0.41 (44 of 48 settings) with the seeds still in a median 12 cluster subsets; an edit that also holds every other harmonic up to 8, which the first moves as a side effect, gives the same (0.40, 45 of 48; Qwen 35 of 36), and the same reweighting aimed at another harmonic leaves it (0.85, lower in 24 of 48), as does resampling alone. The box was chosen after an unbounded edit, which collapses the subsets, failed; the test fails on clustered rings (9 of 24). In the plane fitting Llama-3.1-8B's period-100 circle (Kantamneni & Tegmark, 2025), the same edit takes agreement at k^* from 0.70 to 0.39 (70 of 84 settings, 7 of 7 layers; Pythia-6.9B, 0.68 to 0.49), and in Llama rows 1 (at k^*), 3 and 5 hold (App. K).

Three replications, and the failure. *The full residual stream* (512-atom dictionaries, 60k Pile vectors plus the ring probes at 30% of the pool): 0.79 against 0.62 symmetrized (20 of 24 settings, 5 of 6 cells; 18 vs 12 lock), an installed harmonic still locking its own order or alias (0.63 against 0.32; App. L). *A second model*, Qwen2.5-1.5B (Qwen Team, 2024): 0.688 against 0.519 over 196 settings, 14 of 14 cells. *A spread density*, the 48 quarter-hours: 0.724 against 0.613 (50 of 64). Among the failures (App. S) are Table 2's last two rows: *which* order locks is not read off the density alone. On clustered rings its strongest harmonic picks the lock order in 13 of 90 settings (39 crediting the alias; chance 10 and 19.7), on the spread density in 33 of 64, no better than the constant $k = 2$ (35); weighting it by the trained kernel, as Proposition 3 does, adds little (35 and 40 of 90; App. K).

On GemmaScope's in-plane configurations the natural ℓ_1 landscape is deeper than the symmetrized in 155 of 155 cases, as for random angles, and the configuration sits at its 41st percentile, not its minimum (App. I).

6 THE RECOMMENDATION, AT PRODUCTION SCALE

Proposition 1 says only that a symmetry of the data leaves atoms free along its orbit; it does not say that subspaces replicate, and a language model's isotropy group is unknown. It motivates a recommendation, which this section tests empirically and without the Fourier criterion: attach a claim to the plane a dictionary assigns to a concept, not to one latent inside it, whose phase a symmetry, a flat objective or training that stops short (§4.3) leaves undetermined. In each SAE we rank concept-selective latents by an F-statistic across the item clusters, computed within *that SAE's own decoder*, and ablate in the residual stream either its single most selective latent (the *atom*) or the 2-plane its top-64 selective span assigns to the probes' cluster means (the *plane*). The statistic is the correlation between two SAEs' matrices of log-probability changes (63 weekday prompts $\times$ 7 days), each claim against a matched-rank control and tested by a leave-one-dictionary-out jackknife, which respects that pairs share dictionaries; we also *double-center* each matrix, removing effects that merely lower every item or every prompt (App. Q, R).

GemmaScope: the plane transfers, the most selective atom does not. At Gemma-2-2B layer 12 we take twelve released dictionaries (16k/32k/65k, L_0 from 21 to 445). The model's ring plane lies inside each one's top-64 selective span at a median 0.758 of its energy, against 0.028 for a random span of that rank and 0.279 for the same rule run on *months* (Fig. 1e). Ablating the plane reproduces across dictionaries at $+0.815$ over the 66 pairs (Fig. 1f); ablating the atom at -0.042, though its effect is an order of magnitude above its control's. The plane beats its control for 12 of 12 dictionaries and the atom for 7 of 12 (raw jackknife $p = 4 \times 10^{-4}$ and 0.10; one release, not independent training conditions). The atoms differ in *which* day they point at (all seven occur), and that is most of their failure: taking each dictionary's atom *for each day*, same-day atoms transfer at $+0.58$ against $+0.04$ for different days, above a random direction for 12 of 12 dictionaries—and below the plane for 12 of 12. A decoder-only rank-2 object does not transfer ($+0.16$), and a rank-1 direction built like the plane does ($+0.78$), less reliably. And the plane is specific: it moves its own concept's answers more than the other concepts' planes do, and more than it moves theirs, in all 32

dictionaries at three layers for each concept (10–32× for weekdays and months, 1.6–2.5× for clock hours; App. Q).

Three layers, three concepts, three releases. Across layers 5, 12 and 20 and weekdays, months and clock hours, double-centered, the plane beats its control in 9 of 9 (layer, concept) cells (jackknife $p \leq 5\times10^{-6}$) and the atom in 5 of 9 (Table 13), never at layer 5; with 100 control draws the weekday atom clears at 2 of 3 layers, weakly (App. R). Paulo & Belrose (2026) trained nine independently seeded 32k TopK dictionaries on Pythia-160M's layer-6 MLP output; on the eight released, the plane gives $+0.55$ against -0.07 for the atom (seed-level jackknife $p = 0.016$ and 0.55), though the day means form no ring there. In Llama Scope's dictionaries of Llama-3.1-8B (He et al., 2024; Grattafiori et al., 2024), from another laboratory, the plane beats its control at 7 of 7 layers for weekdays and for months, the atom at 3 and 1; in Gemma-2-9B, the plane at $+0.88$ and $+0.81$, the atom below control.

Without probes or per-item labels. Built from natural text alone—the latents whose activation falls mostly on tokens that *are* weekday names, and the top-2 principal subspace of their own reconstruction there, with no probe and no day label—the plane still transfers at layers 5, 12 and 20 ($+0.74$, $+0.66$, $+0.81$ double-centered, against $\leq +0.05$ for a firing-matched random latent group), though it is not the probe plane; the label-free latent for one weekday transfers weakly ($\leq +0.23$), and months repeat both (App. R).

Our own seeds, and the symmetry directly. Six seeds per coding rule, trained at GemmaScope's smallest width on 60M FineWeb tokens (Penedo et al., 2024), split the same way: the plane transfers across seeds in all six (rule, concept) cells, the atom in two. The in-plane configuration's phase agrees across ReLU+L_1 seeds (mean A_k 0.96 on weekdays), more than across TopK seeds (0.57; permutation $p = 0.001$; 0.95 against 0.78 read from 8 latents each, $p = 0.003$), and at that matched count in 7 of 9 (layer, concept) cells (weekdays at layer 20 go the other way, n.s.). Rotating every activation's component in the weekday plane about the ring's center by a random multiple of $2\pi/7$ during training—the data law is exactly $\mathbb{Z}_7$-symmetric there, all else unchanged—removes that agreement (0.40, chance 0.37; $p = 0.001$; also at layer 5) and the atoms' match (Hungarian-matched atoms 2.0° apart across seeds, 15.5° after; $p = 0.006$) while months and clock hours keep theirs; the same rotation of a variance-matched plane elsewhere removes nothing (0.95, $p = 0.21$), and $\mathbb{Z}_{12}$ in the month plane does the same for months (phase $0.70 \to 0.43$; atoms $7.8° \to 23.0°$ apart). Installing one harmonic back restores both ($A_2 = 1.00$; atoms 3.8° apart) with as many latents on the plane as the symmetrized seeds, some 400 (App. R).

Against supervised causal search. At layer 12, swapping a dictionary's plane between two weekday prompts moves the answer to the source's in a median 26% of pairs, its atom in 2%, the label-free plane in 7% and a random plane in none; the probe ring plane in 31%, and a rank-2 subspace trained by distributed alignment search (Geiger et al., 2024) in 42%. Supervised objects win, as on RAVEL (Huang et al., 2024), and no plane carries the whole day (DAS: 83% at rank 4; App. R).

7 Discussion and limitations

What could have come out otherwise. Proposition 3 is an identity, so we keep illustrations apart from tests: fifty-five, listed with outcomes in App. S, seventeen of them against us—nine in the controlled and ring settings, eight at production scale.

What to do about it. Attach claims to the subspace a dictionary assigns to a concept, which a symmetry leaves fixed, not to one latent's phase inside it (Proposition 1, §6); a claim about an atom should name its item, since item-matched atoms transfer, less reliably than the plane.

Limitations. The landscape is for a 2-plane (a 3-dimensional check, App. U; no landscape for $d > 2$); Proposition 4's sparsity hypothesis fails on trained dictionaries; the L_0 prediction has the wrong sign (App. E); removal fails on clustered rings; the modular dose response follows key frequencies, not the cost profile; the fluctuation relation is derived only at stationarity; the symmetrized control locks more often in the full stream; our frozen-encoder estimator fails its controls; the graded prediction fails on unprompted data; and the plane, unlike the atom, is anchored on the concept's cluster means (§6). We do not separate feature splitting and absorption from the rotational account.

AI USE STATEMENT

In this work, we used generative AI tools to assist in writing, editing, and debugging code, including code used to implement our methods and experiments. We did not use generative AI tools for any of the other tasks that require disclosure: in particular, we did not use them to generate synthetic data (the synthetic activations of §4.1 are sampled by our experiment code from a distribution we specified, stated in Appendix C; the AI assistance there was in writing that code), to develop theoretical models or conceptual frameworks, to formulate mathematical claims or assist in proofs, to propose or refine hypotheses, to design the research methodology or experiments, to assist with translation, to clean or reformat datasets, or to support data analysis or the interpretation of results. Additionally, we used generative AI tools to edit the manuscript to improve readability and clarity, and to obtain simulated peer-review-style feedback on drafts, which informed revisions to the writing and presentation. All AI-assisted code was reviewed and tested for correctness by the authors, and all AI-suggested edits and feedback were reviewed and adopted at the authors' discretion; the scientific content, claims, and results are entirely the authors' own. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Appendix A contains all proofs; every empirical claim traces to a supplied result file. The controlled experiments (§4) and the ring-seed runs (§5) run on CPU; nearly everything that runs a language model or trains a production-width dictionary uses a single RTX 4090, about a day in all, on released models and SAEs and public corpora (App. V). Code, per-feature records (61,870 rotation landscapes; 129,071 defined orthogonal-matching-pursuit retention rates among the 671,744 latents of nine SAEs) and per-run result records are in the supplementary material, and the consistency scan of Appendix V recomputes the numbers it guards from those records and checks that each appears in the PDF.

REFERENCES

Alekh Agarwal, Animashree Anandkumar, Prateek Jain, Praneeth Netrapalli, and Rashish Tandon. Learning sparsely used overcomplete dictionaries. In *Proceedings of the 27th Conference on Learning Theory (COLT)*, volume 35, pp. 123–137, 2014.

Sanjeev Arora, Rong Ge, and Ankur Moitra. New algorithms for learning incoherent and overcomplete dictionaries. In *Proceedings of the 27th Conference on Learning Theory (COLT)*, volume 35, pp. 779–806, 2014.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Usha Bhalla, Thomas Fel, Can Rager, Sheridan Feucht, et al. Do sparse autoencoders capture concept manifolds? *arXiv preprint arXiv:2604.28119*, 2026.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning (ICML)*, 2023.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L. Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features.

Bart Bussmann, Noa Nabeshima, Adam Karvonen, and Neel Nanda. Learning multi-level features with Matryoshka sparse autoencoders. In *International Conference on Machine Learning (ICML)*, 2025.

Charles F. Cadieu and Bruno A. Olshausen. Learning intermediate-level representations of form and motion from natural movies. *Neural Computation*, 24(4):827–866, 2012.

Jean-François Cardoso. Multidimensional independent component analysis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1998.

David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, Satvik Golechha, and Joseph Bloom. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2409.14507.

Pratik Chaudhari and Stefano Soatto. Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. In *International Conference on Learning Representations*, 2018.

Maheep Chaudhary and Atticus Geiger. Evaluating open-source sparse autoencoders on disentangling factual knowledge in GPT-2 small. *arXiv preprint arXiv:2409.04478*, 2024.

Bilal Chughtai, Lawrence Chan, and Neel Nanda. A toy model of universality: Reverse engineering how networks learn group operations. In *International Conference on Machine Learning (ICML)*, 2023.

Taco Cohen and Max Welling. Learning the irreducible representations of commutative Lie groups. In *International Conference on Machine Learning*, 2014.

Pierre Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, 1994.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Seyed Arshan Dalili and Mehrdad Mahdavi. Subspace-aware sparse autoencoders for effective mechanistic interpretability. *arXiv preprint arXiv:2606.06333*, 2026.

David Donoho and Victoria Stodden. When does non-negative matrix factorization give a correct decomposition into parts? In *Advances in Neural Information Processing Systems (NIPS)*, 2003.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *Transformer Circuits Thread*, 2022. `https://transformer-circuits.pub/2022/toy_model/index.html`.

Joshua Engels, Eric J. Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2405.14860.

Thomas Fel, Ekdeep Singh Lubana, Jacob S. Prince, Matthew Kowal, Victor Boutin, Isabel Papadimitriou, Binxu Wang, Martin Wattenberg, Demba Ba, and Talia Konkle. Archetypal SAE: Adaptive and stable dictionary learning for concept extraction in large vision models. In *International Conference on Machine Learning (ICML)*, 2025.

Xiao Fu, Kejun Huang, and Nicholas D. Sidiropoulos. On identifiability of nonnegative matrix factorization. *IEEE Signal Processing Letters*, 25(3):328–332, 2018.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2406.04093.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Proceedings of the Third Conference on Causal Learning and Reasoning (CLeaR)*, volume 236 of *Proceedings of Machine Learning Research*, pp. 160–187, 2024.

Gemma Team. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Gleb Gerasimov, Timofei Rusalev, Nikita Balagansky, Daniil Laptev, Vadim Kurochkin, and Daniil Gavrilov. Unstable features, reproducible subspaces: Understanding seed dependence in sparse autoencoders. *arXiv preprint arXiv:2606.12138*, 2026.

Charles Godfrey, Davis Brown, Tegan Emerson, and Henry Kvinge. On the symmetries of deep learning models and their internal representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. arXiv:2205.14258.

Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Karol Gregor and Yann LeCun. Learning fast approximations of sparse coding. In *International Conference on Machine Learning (ICML)*, 2010.

Rémi Gribonval and Karin Schnass. Dictionary identification—sparse matrix-factorization via ℓ_1-minimization. *IEEE Transactions on Information Theory*, 56(7):3523–3539, 2010.

Zhengfu He, Wentao Shu, Xuyang Ge, Lingjie Chen, Junxuan Wang, Yunhua Zhou, Frances Liu, Qipeng Guo, Xuanjing Huang, Zuxuan Wu, Yu-Gang Jiang, and Xipeng Qiu. Llama Scope: Extracting millions of features from Llama-3.1-8B with sparse autoencoders. *arXiv preprint arXiv:2410.20526*, 2024.

Thomas Heap, Tim Lawson, Lucy Farnik, and Laurence Aitchison. Automated interpretability metrics do not distinguish trained and random transformers. *arXiv preprint arXiv:2501.17727*, 2025.

Irina Higgins, David Amos, David Pfau, Sébastien Racanière, Loïc Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018.

Christopher J. Hillar and Friedrich T. Sommer. When can dictionary learning uniquely recover sparse data from subsamples? *IEEE Transactions on Information Theory*, 61(11):6290–6297, 2015.

Sai Sumedh R. Hindupur, Ekdeep Singh Lubana, Thomas Fel, and Demba Ba. Projecting assumptions: The duality between sparse autoencoders and concept geometry. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Jing Huang, Zhengxuan Wu, Christopher Potts, Mor Geva, and Atticus Geiger. RAVEL: Evaluating interpretability methods on disentangling language model representations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Aapo Hyvärinen and Patrik Hoyer. Emergence of phase- and shift-invariant features by decomposition of natural images into independent feature subspaces. *Neural Computation*, 12(7):1705–1720, 2000.

Aapo Hyvärinen and Erkki Oja. Independent component analysis: Algorithms and applications. *Neural Networks*, 13(4–5):411–430, 2000.

Jean Kaddour. The MiniPile challenge for data-efficient language models. *arXiv preprint arXiv:2304.08442*, 2023.

Subhash Kantamneni and Max Tegmark. Language models use trigonometry to do addition. *arXiv preprint arXiv:2502.00873*, 2025.

Subhash Kantamneni, Joshua Engels, Senthooran Rajamanoharan, Max Tegmark, and Neel Nanda. Are sparse autoencoders useful? A case study in sparse probing. In *International Conference on Machine Learning (ICML)*, 2025. arXiv:2502.16681.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning (ICML)*, 2019.

Daniel Kunin, Javier Sagastuy-Brena, Surya Ganguli, Daniel L. K. Yamins, and Hidenori Tanaka. Neural mechanics: Symmetry and broken conservation laws in deep learning dynamics. In *International Conference on Learning Representations*, 2021.

Michael Lan, Philip Torr, Austin Meek, Ashkan Khakzar, David Krueger, and Fazl Barez. Sparse autoencoders reveal universal feature spaces across large language models. *arXiv preprint arXiv:2410.06981*, 2024.

Patrick Leask, Bart Bussmann, Michael Pearce, Joseph Bloom, Curt Tigges, Noura Al Moubayed, Lee Sharkey, and Neel Nanda. Sparse autoencoders do not find canonical units of analysis. In *International Conference on Learning Representations (ICLR)*, 2025.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma Scope: Open sparse autoencoders everywhere all at once on Gemma 2. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 278–300, 2024.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning (ICML)*, 2019.

Aleksandar Makelov, Georg Lange, and Neel Nanda. Is this the subspace you are looking for? An interpretability illusion for subspace activation patching. In *International Conference on Learning Representations (ICLR)*, 2024.

Stephan Mandt, Matthew D. Hoffman, and David M. Blei. Stochastic gradient descent as approximate Bayesian inference. *Journal of Machine Learning Research*, 18(134):1–35, 2017.

Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2403.19647.

Maxime Méloux, Silviu Maniu, François Portet, and Maxime Peyrard. Everything, everywhere, all at once: Is mechanistic interpretability identifiable? In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2502.20914.

Eric J. Michaud, Liv Gorton, and Tom McGrath. Understanding sparse autoencoder scaling in the presence of feature manifolds. *arXiv preprint arXiv:2509.02565*, 2025.

Alexander Modell, Patrick Rubin-Delanchy, and Nick Whiteley. The origins of representation manifolds in large language models. *arXiv preprint arXiv:2505.18235*, 2025.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2301.05217.

Walter Nelson, Marco Fumero, Theofanis Karaletsos, and Francesco Locatello. Statistical and structural identifiability in representation learning. In *International Conference on Learning Representations (ICLR)*, 2026a. arXiv:2603.11970.

Walter Nelson, Theofanis Karaletsos, and Francesco Locatello. Toward identifiable sparse autoencoders. In *International Conference on Machine Learning (ICML)*, 2026b. arXiv:2605.31245.

Charles O'Neill, Alim Gumran, and David Klindt. Compute optimal inference and provable amortisation gap in sparse autoencoders. *arXiv preprint arXiv:2411.13117*, 2024.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. In *International Conference on Learning Representations (ICLR)*, 2025.

Yagyensh C. Pati, Ramin Rezaiifar, and Perinkulam S. Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Asilomar Conference on Signals, Systems and Computers*, 1993.

Gonçalo Paulo and Nora Belrose. Sparse autoencoders trained on the same data learn different features. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2501.16615.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024. arXiv:2406.17557.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping ahead: Improving reconstruction fidelity with JumpReLU sparse autoencoders. *arXiv preprint arXiv:2407.14435*, 2024.

Geoffrey Roeder, Luke Metz, and Durk P. Kingma. On linear identifiability of learned representations. In *International Conference on Machine Learning (ICML)*, 2021.

Karl Rohe and Muzhe Zeng. Vintage factor analysis with Varimax performs statistical inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(4):1037–1060, 2023.

Sophia Sanborn, Christian Shewmake, Bruno Olshausen, and Christopher Hillar. Bispectral neural networks. In *International Conference on Learning Representations*, 2023.

Berfin Şimşek, François Ged, Arthur Jacot, Francesco Spadaro, Clément Hongler, Wulfram Gerstner, and Johanni Brea. Geometry of the loss landscape in overparameterized neural networks: Symmetries and invariances. In *International Conference on Machine Learning (ICML)*, pp. 9722–9732, 2021.

Xiangchen Song, Aashiq Muhamed, Yujia Zheng, Lingjing Kong, Zeyu Tang, Mona T. Diab, Virginia Smith, and Kun Zhang. Mechanistic interpretability should prioritize feature consistency in sparse autoencoders. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2026.

Daniel A. Spielman, Huan Wang, and John Wright. Exact recovery of sparsely-used dictionaries. In *Proceedings of the 25th Annual Conference on Learning Theory (COLT)*, volume 23, pp. 37.1–37.18, 2012.

Alan Sun and Mariya Toneva. Tracking equivalent mechanistic interpretations across neural networks. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2603.30002.

Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere I: Overview and the geometric picture. *IEEE Transactions on Information Theory*, 63(2):853–884, 2017.

Denis Sutter, Julian Minder, Thomas Hofmann, and Tiago Pimentel. The non-linear representation dilemma: Is causal abstraction enough for mechanistic interpretability? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2507.08802.

Fabian J. Theis. Towards a general independent subspace analysis. In *Advances in Neural Information Processing Systems (NIPS)*, volume 19, 2006.

James C. R. Whittington, Will Dorrell, Surya Ganguli, and Timothy E. J. Behrens. Disentanglement with biological constraints: A theory of functional cell types. In *International Conference on Learning Representations (ICLR)*, 2023.

Siqi Wu and Bin Yu. Local identifiability of ℓ_1-minimization dictionary learning: A sufficient and almost necessary condition. *Journal of Machine Learning Research*, 18(168):1–56, 2018.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. AxBench: Steering LLMs? Even simple baselines outperform sparse autoencoders. In *International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, 2025.

Sho Yaida. Fluctuation-dissipation relations for stochastic gradient descent. In *International Conference on Learning Representations*, 2019.

Ziqian Zhong, Ziming Liu, Max Tegmark, and Jacob Andreas. The clock and the pizza: Two stories in mechanistic explanation of neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2306.17844.

Liu Ziyin. Symmetry induces structure and constraint of learning. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2309.16932.

A PROOFS

Throughout, the discrete redundancy of permuting atoms is quotiented out without further comment (columns have unit norm, so there is no rescaling). J is the population objective and $\mathbb{E}\|x\|^2 < \infty$, so J is finite ($c \le \frac{1}{2}\|x\|^2$); for the ℓ_1 cost of Eq. 1 with $\lambda > 0$ it is also continuous on the compact set of unit-column dictionaries (an optimal code has $\|a^*\|_1 \le \|x\|^2/2\lambda$, a feasible set independent of D, so Berge's theorem applies, and dominated convergence, with $c \le \frac{1}{2}\|x\|^2$ integrable, carries continuity through the expectation), so the arg-min set is non-empty (for the L_0 and TopK costs, which are discontinuous in D, we claim no existence, and Proposition 1 is then a statement about whatever minimizers exist). $G(P)$ is defined for every P but contains linear isometries only, so it registers symmetries about the origin. A law symmetric about a center $\mu \neq 0$ carries its symmetry in the affine stabilizer $G_\mu(P) = \{g \in O(d) : x \mapsto \mu + g(x-\mu) \text{ preserves } P\}$: an SAE with a decoder bias inherits it as it stands (Proposition 2), whereas Eq. 1 has no bias, so Proposition 1 reaches it only after centering, which replaces the objective by the centered law's. The ring experiments center on the item-mean centroid, an estimate of μ (App. K); the synthetic planes (App. C) and the densities of App. U are symmetric about the origin and are not centered. Since the empirical law of a finite sample has a generically trivial isotropy group, the definition has content for the population law.

Lemma 1 (Equivariance of the coding cost)**.** *For every $g \in O(d)$, every x, and every D: $c(gx; gD) = c(x; D)$, for c as in Eq. 1 and, more generally, for any per-sample cost of the form $\inf_{a\in\mathcal{A}} F(\|x - Da\|, a)$ with a feasible set $\mathcal{A}$ that does not depend on (x, D) (signed codes included).*

Proof. g is an isometry, so $\|gx - gDa\| = \|x - Da\|$ for all a; the penalty depends only on a, and the feasible set $\{a \ge 0\}$ is untouched. The two minimization problems agree term by term. For the sparse-coding forms of the JumpReLU/TopK objectives the per-sample cost is a function of the reconstruction error norm and the code (its L_0 or support) over a feasible set that does not depend on (x, D), so the same argument applies; the amortized versions are covered by Proposition 2. □

Proof of Proposition 1. For $g \in G(P)$, $J(gD) = \mathbb{E}_{x\sim P}[c(x; gD)] = \mathbb{E}_{x\sim P}[c(g^{-1}x; D)] = \mathbb{E}_{y\sim g_\#^{-1}P}[c(y; D)] = \mathbb{E}_{y\sim P}[c(y; D)] = J(D)$, using Lemma 1 and $g^{-1} \in G(P)$. Hence J is constant on $G(P)$-orbits and the arg-min set is a union of orbits. A continuous rotation in $G(P)$ produces a circle of distinct optima only if some optimal atom has a component in the rotated plane; otherwise its orbit is a point. □

Proof of Proposition 2. For an encoder $z_\Theta(x) = \sigma(W_e(x - b_d) + b_e)$ with σ acting on code coordinates (thresholds and column norms are unchanged by g), $W_e g^\top(gx - gb_d) + b_e = W_e(x - b_d) + b_e$, so codes match: $z_{g\cdot\Theta}(gx) = z_\Theta(x)$. The reconstruction becomes $g(W_d z + b_d)$ and $\|gx - g\hat{x}\| = \|x - \hat{x}\|$; take expectations and use $g_\# P = P$. For a law symmetric about a mean $\mu \neq 0$, i.e. for $g \in G_\mu(P)$ (the map $x \mapsto \mu + g(x-\mu)$ preserves P), the same computation with $b_d \mapsto \mu + g(b_d - \mu)$ in place of gb_d gives, at the input $\mu + g(x - \mu)$, the same code $z_\Theta(x)$ and the reconstruction $\mu + g(\hat{x} - \mu)$, so an SAE with a decoder bias inherits the symmetry of the centered law without any centering. The proposition concerns the loss, not the optimizer: plain gradient descent from a rotation-invariant initialization is equivariant, but Adam's per-coordinate step sizes are not, so Adam-trained seeds can share a preference for the coordinate frame on a flat objective; App. M measures this by embedding each seed's plane in a random basis. □

Lemma 2 (Shift covariance and homogeneity)**.** *Write $c_r(\psi;\alpha)$ for the cost of the point $r\,u(\psi)$ against atoms at angles α. Then (i) $c_r(\psi;\,\alpha+\varphi\mathbf{1}) = c_r(\psi-\varphi;\,\alpha)$; (ii) for the ℓ_1 cost and $r>0$, $c_r(\psi;\alpha) = r^2\,c_1^{(\lambda/r)}(\psi;\alpha)$, where the superscript denotes the penalty used (a penalty homogeneous of degree s in a gives λ/r^{2-s} instead: λ/r^2 for L_0, and TopK, a constraint, has none).*

Proof. (i) is Lemma 1 with $g = R_\varphi \in SO(2)$, noting $R_{-\varphi}u(\psi) = u(\psi-\varphi)$. (ii) substitute $a = rb$ in Eq. 1 ($r>0$). Part (ii) is specific to the ℓ_1 penalty: under an L_0 penalty the effective penalty is λ/r^2, and TopK has none. □

Proof of Eq. 2 and Proposition 3. By Lemma 2(i), rotating all atoms rigidly by φ replaces $\bar c(\theta)$ by $\bar c(\theta-\varphi)$, so $J(\varphi) = \int \rho(\theta)\bar c(\theta-\varphi)\,d\theta$. With $\hat f_k = \frac{1}{2\pi}\int f(\theta)e^{-ik\theta}d\theta$, expanding $\bar c$ alone (for the ℓ_1 cost the per-sample cost is the Moreau envelope of a convex function of x, so its gradient, the residual, is 1-Lipschitz; $\bar c$ is therefore $C^{1,1}$ with a second derivative of bounded variation (piecewise smooth at a fixed radius), $\hat{\bar c}_k = O(k^{-3})$, and its series converges absolutely, which covers point-mass ρ too) and integrating against ρ by Fubini gives $J(\varphi) = 2\pi\sum_k \overline{\hat\rho_k}\hat{\bar c}_k e^{-ik\varphi}$ (using $\hat\rho_{-k} = \overline{\hat\rho_k}$, the complex conjugate, for real ρ). By uniqueness of Fourier coefficients of the continuous function J, it is constant iff every $k\neq 0$ coefficient vanishes, i.e. iff $\hat\rho_k\hat{\bar c}_k = 0$ for all $k\neq 0$. The sparse-coding L_0 and TopK costs are minima over supports of continuous functions of x (they are discontinuous in D, not in x), so their kernels are continuous, indeed Lipschitz, and J is continuous for every ρ. Only a cost computed through an encoder—a frozen JumpReLU gate, or TopK selection by correlation—can be discontinuous in x; then $\bar c$ is bounded but $\hat{\bar c}_k$ may decay only as $O(1/k)$: for $\rho\in L^1$, J is still continuous and its k-th coefficient is still $2\pi\overline{\hat\rho_k}\hat{\bar c}_k$, so the criterion holds as stated, and for $\rho\in L^2$ Eq. 2 holds pointwise with absolute convergence, since $\sum_k|\hat\rho_k||\hat{\bar c}_k| \le \|\hat\rho\|_2\|\hat{\bar c}\|_2 < \infty$; for a point-mass ρ, J can jump and "constant" means almost everywhere. Eq. 2 equals the full objective up to a constant when the plane's atoms code only in-plane inputs, in-plane inputs use no other atoms, and the per-sample cost is additive across orthogonal planes: true for ℓ_1 and L_0 penalties with atoms inside the planes, but not for TopK, whose budget $\|a\|_0 \le k$ is shared by every plane active in a sample, so for TopK this needs one active plane per sample (or a fixed per-plane budget). (i): uniform ρ has $\hat\rho_k = 0$ for all $k\neq 0$. (ii): if ρ is non-uniform, some $\hat\rho_{k_0}\neq 0$, and then the k_0 term of the series is nonzero exactly when $\hat{\bar c}_{k_0}\neq 0$, a condition on the configuration, the penalty and the radial law, not on the angular density. We deliberately do not assert that $\{\alpha : \hat{\bar c}_{k_0}(\alpha) = 0\}$ is measure-zero: $\bar c$ is defined through a minimum over active sets and is piecewise, though not globally, smooth in α, so the usual analytic argument does not apply as stated. What we do establish is computational: the exact 2-plane solver evaluates $\hat{\bar c}_k(\alpha)$ pointwise in closed form, and over the configurations of App. B, at the λ tested, the vanishing set coincides with the lattice condition $m\nmid k$ of Corollary 1. It need not in general: inside Proposition 4's regime $\hat{\bar c}_q$ vanishes wherever $\hat g_q$ does, e.g. for a regular triangle at $\lambda/r = \sqrt{0.3}$, where $\hat g_6 \propto \lambda(1-\lambda^2)^{5/2}(10\lambda^2-3) = 0$, and wherever $S_q = 0$ on an irregular configuration. If the radial law depends on θ, the reduction is not automatic and we state it precisely. Lemma 2(ii) gives $c_r(\psi;\alpha) = r^2 c_1^{(\lambda/r)}(\psi;\alpha)$: the radius rescales the kernel by r^2 *and* reshapes it through the effective penalty λ/r. When the shape is radius-free—for the ℓ_1 cost exactly only at $\lambda=0$ (TopK, which has no penalty, is always radius-free; L_0, with effective penalty λ/r^2, is not), approximately where λ/r varies little over the radial law—the r^2 factors can be absorbed into the density and every statement goes through with the reweighted marginal $\tilde\rho(\theta)\propto\mathbb{E}[r^2\mid\theta]\,\rho(\theta)$ in place of ρ. Otherwise each radius contributes its own kernel, $J = \sum_b w_b r_b^2\,(\rho_b \star c_1^{(\lambda/r_b)})$, and J is a *mixture* of such correlations rather than one; flatness off the configuration's own harmonics still holds term by term (for a regular m-gon every $c_1^{(\lambda/r)}$, as a function of the angle, has vanishing k-th coefficient unless $m\mid k$), but the relative strengths across $k\in m\mathbb{Z}$ need not be those of Proposition 6. This matters for App. H's objective-level computations, where the per-item mean radii span $2.1\times$ on the day ring at layer 7 before symmetrization, so that appendix reports the mixture computation as a control on the reduced one; the trained-seed runs of §5.1 use App. K's unit convention, every probe on the unit circle, where the reduction is exact. □

Proof of Corollary 1. A regular m-gon is invariant under rotation by $2\pi/m$, so $\bar c$ is $2\pi/m$-periodic and $\hat{\bar c}_k = 0$ unless $m\mid k$ (Lemma 2(i) applied to the symmetry of the configuration itself). For $\rho = (1+a\cos\ell\theta)/2\pi$ the only nonzero data harmonics are $k=\pm\ell$; the product $\hat\rho_k\hat{\bar c}_k$ vanishes for all $k\neq 0$ unless $m\mid\ell$. For the equally weighted $\rho = \frac{1}{p}\sum_j\delta_{2\pi j/p}$, $\hat\rho_k = \frac{1}{2\pi}\mathbf{1}[p\mid k]$ (generic

unequal weights put power at every k), and for a regular m-gon $\hat{\bar{c}}_k = 0$ unless $m \mid k$; the product is therefore supported on $m\mathbb{Z} \cap p\mathbb{Z} = \mathrm{lcm}(m,p)\mathbb{Z}$. Hence $J(\varphi) = J_0 + 4\pi \sum_{j\geq 1} \mathrm{Re}\big[\overline{\hat{\rho}_{jL}}\hat{\bar{c}}_{jL} e^{-ijL\varphi}\big]$ with $L = \mathrm{lcm}(m,p)$, so $|J - J_0| \leq 4\pi \sum_{j\geq 1} |\hat{\rho}_{jL}||\hat{\bar{c}}_{jL}|$, the range $\max J - \min J$ is at most twice that, and both vanish identically only if the sum does. For $\gcd(m,p) = 1$ (e.g. prime $p > m$), $L = mp$: the rotation is flat to *every* harmonic below mp, and residually pinned at order mp by an amount set by $|\hat{\bar{c}}_{mp}|$, which the exact 2-plane solver puts, for a 53-point comb in one plane, a proxy for the four-plane curve of §4.2, at $|\hat{\bar{c}}_{mp}|/\hat{\bar{c}}_0 = 3.0 \times 10^{-7}$, 1.3×10^{-6}, 3.4×10^{-7}, 6.0×10^{-7} for regular $m = 2, 3, 4, 5$ at $\lambda = 0.12$ (App. D); the landscape measured there over the group elements is exactly flat by Proposition 2 and does not resolve it. We therefore state the corollary as flatness up to the $\mathrm{lcm}(m,p)$ harmonic rather than as exact flatness; only the $\rho = (1 + a\cos\ell\theta)/2\pi$ case above, whose spectrum is finitely supported, is exactly flat off resonance. □

Proof of Proposition 5. In one dimension with non-negative codes and one-signed data $x = re$, $r \geq 0$, the candidate atoms are $\pm e$. With atom $+e$ and $r > \lambda$ the optimal code is $a = r - \lambda$ with cost $\frac{1}{2}r^2 - \frac{1}{2}(r-\lambda)^2 < \frac{1}{2}r^2$; with atom $-e$ the code is 0 and the cost $\frac{1}{2}r^2$. Taking expectations gives a strict gap whenever $P(r > \lambda) > 0$ (implied by $\lambda < \mathbb{E}[r]$); no continuous family of optima exists. (Atoms are restricted to the line; $\mathbb{E}r^2 < \infty$.) For the second sentence: by Proposition 3 a flat rigid rotation needs $\hat{\rho}_k \hat{\bar{c}}_k = 0$ for all $k \neq 0$, so it fails as soon as some $k_0 \neq 0$ has $\hat{\rho}_{k_0} \neq 0$ and $\hat{\bar{c}}_{k_0} \neq 0$. That this happens in two dimensions: a single atom ($m = 1$) is inside Proposition 4's regime trivially, so at radius $r = 1$, $\hat{\bar{c}}_1 = -\frac{1}{2}\hat{g}_1 e^{-i\alpha}$ with $\hat{g}_1 = \big(2\sqrt{1-\lambda^2} + \lambda^2\sqrt{1-\lambda^2} - 3\lambda \arccos\lambda\big)/3\pi > 0$ for $0 \leq \lambda < 1$ (the integrand $(\cos u - \lambda)^2 \cos u$ is positive on the arc $|u| < \arccos\lambda$), and any ρ with $\hat{\rho}_1 \neq 0$ pins it. Conversely, Corollary 1 shows that flatness can survive $\hat{\rho}_k \neq 0$, and a coding rule whose kernel vanishes identically (non-negative TopK at $k \geq 2$ on a regular m-gon, $m \geq 3$, App. M, and at its optimum on a full-support ρ, which reconstructs exactly, App. P) is flat for every ρ: the kernel, not the dimension, decides. □

Proof of Proposition 6. Term-by-term differentiation of Eq. 2 is legitimate whenever $\sum_k k^2 |\hat{\rho}_k||\hat{\bar{c}}_k| < \infty$, which for the ℓ_1 cost holds for square-integrable ρ (by Cauchy–Schwarz, since there $k^2|\hat{\bar{c}}_k| = O(1/k)$; for the other rules' kernels, whose coefficients can decay as slowly as $O(1/k)$, it needs a smoother ρ and is assumed, not shown); for a point-mass density J is a finite sum of translates of $\bar{c}$, twice differentiable except where an active set switches, and there the formula holds wherever J'' exists as a conditionally convergent series (symmetric partial sums), since $|\hat{\rho}_k|$ does not decay and $k^2|\hat{\bar{c}}_k|$ can decay only as $1/k$; the error bound is then finite only if its sum converges. It gives $J''(\varphi) = 2\pi \sum_{k\neq 0} (-ik)^2 \overline{\hat{\rho}_k}\hat{\bar{c}}_k e^{-ik\varphi} = -2\pi \sum_{k\neq 0} k^2 \overline{\hat{\rho}_k}\hat{\bar{c}}_k e^{-ik\varphi}$, and $J''(\varphi^*) \geq 0$ at a minimizer is the second-order condition, not a term-wise statement, since individual harmonics may contribute with either sign. Pairing k with $-k$ and writing $a_k = 4\pi|\hat{\rho}_k||\hat{\bar{c}}_k|$ and $\overline{\hat{\rho}_k}\hat{\bar{c}}_k = |\hat{\rho}_k||\hat{\bar{c}}_k|e^{i\delta_k}$, $J(\varphi) = J_0 + \sum_{k\geq 1} a_k \cos(k\varphi - \delta_k)$ and $J''(\varphi^*) = -\sum_{k\geq 1} k^2 a_k \cos(k\varphi^* - \delta_k)$. Assume $a_m > 0$ and write $m\varphi^* - \delta_m = \pi + \varepsilon$, so that $\varepsilon = 0$ at a minimizer of the m-th term alone; the proposition concerns the minimizer on the branch $|\varepsilon| \leq \pi/2$, i.e. within a quarter period of that term's minimum. Then $J''(\varphi^*) = m^2 a_m - m^2 a_m(1 - \cos\varepsilon) - \sum_{k\neq m} k^2 a_k \cos(k\varphi^* - \delta_k)$. The last sum is bounded in modulus by $\sum_{k\geq 1, k\neq m} k^2 a_k$. For the displacement, $J'(\varphi^*) = 0$ reads $m a_m \sin\varepsilon = \sum_{k\neq m} k a_k \sin(k\varphi^* - \delta_k)$, so $|\sin\varepsilon| \leq \eta_1 := \sum_{k\geq 1, k\neq m} k a_k/(m a_m)$, and on the branch $|\varepsilon| \leq \pi/2$, $1 - \cos\varepsilon \leq \sin^2\varepsilon \leq \eta_1^2$. Hence $|E| \leq \sum_{k\geq 1, k\neq m} k^2 a_k + \eta_1^2 m^2 a_m$, which is the statement, and the upper side is exact with no displacement term: $J''(\varphi^*) \leq m^2 a_m + \sum_{k\neq m} k^2 a_k$. Since $k \leq k^2$, $\eta_1 \leq m\eta$ with $\eta := \sum_{k\neq m} k^2 a_k/(m^2 a_m)$; when every other harmonic lies above m, $\eta_1 \leq \eta$. The displacement correction is therefore not uniform in m when lower harmonics are present, and at a local minimizer off that branch the approximation does not apply. For a regular m-gon the sum runs over multiples of m only, so the leading correction is the $2m$ harmonic. App. B's check ($\leq 1.1 \times 10^{-4}$) is on single-harmonic densities, where $E \equiv 0$, so it validates the series, not the size of E. Nothing above uses that m is the atom count: the bound holds with m replaced by any harmonic k_0 with $a_{k_0} > 0$, e.g. the dominant one of an irregular configuration. Nor is a pinned minimum always curved: one atom at $\lambda = 0.05$, $r = 1$, on $\rho \propto 1 + 0.5\cos\theta + b_2\cos 2\theta$ with $b_2 = -\hat{g}_1/(8\hat{g}_2) \approx -0.205$ has $J(\varphi) - J(0) = \frac{a_1}{2}(1 - \cos\varphi)^2$, a strict minimum with

$J''(0) = 0$; there the curvature understates the pinning, and a Boltzmann spread scales as $T^{1/4}$ rather than $T^{1/2}$. □

Free configurations. The statements above rotate the configuration rigidly. A trained SAE's atoms move freely, and the exact 2-plane numerics of App. B (the free-configuration paragraph: $\lambda = 0.3$, unit radius, $\rho \propto 1 + 0.8\cos\ell\theta$) show the graded picture for ReLU+L_1; except where derived below, these are numerical findings on the (m, ℓ) searched, not theorems. With uniform ρ the objective is flat for *every* configuration (a genuine continuum of optima); on resonance ($m \mid \ell$) the regular configuration is a strict local minimum (and, at that λ, the free optimum for $m \geq 3$—not at every λ: inside Proposition 4's regime J is a sum of per-atom terms, and at $\lambda = 0.9$ a triangle on $\rho \propto 1 + 0.9\cos 6\theta$ ties with irregular placements on the same wells; an antipodal pair, $m = 2$, is a local minimum that the free optimum leaves); off resonance the regular configuration has an exactly flat rigid mode but is not a free optimum—by the polygon's symmetry its per-atom gradients are $C\cos(\ell\alpha_j - \beta)$, with one amplitude C set by the ℓ-th coefficient of the per-atom derivative kernel and one phase β, so they vanish together only if $m \mid 2\ell$, unless the relevant kernel coefficient vanishes ($C = 0$, e.g. inside Proposition 4's regime wherever $\hat{g}_\ell = 0$), and where they do it is a saddle (free Hessian eigenvalues of both signs)—and the optimizer breaks regularity into a set of optima that, in every cell searched, is finite and equivalent under the data's $\mathbb{Z}_\ell$. Away from exact symmetry the quantity that orders instability is therefore the curvature/depth of the landscape, not a binary criterion, and that is what all our experiments measure.

Proposition 4 in full. Let the atoms be distinct unit vectors at angles $\alpha_1, \ldots, \alpha_m$ with minimum angular gap δ, let the data lie at radius r, and let $\lambda > 0$. Every optimal code has at most two active atoms: the optimal fit $y^* = Da^*$ is unique (the objective is strictly convex in the fit), and each active atom satisfies $d_j^\top(x - y^*) = \lambda > 0$, which at most two unit vectors in a plane do. (For any $\lambda \geq 0$, *some* optimal code has at most two: $\min\{\mathbf{1}^\top a : Da = y^*,\ a \geq 0\}$ has two equality constraints, so it has a basic optimal solution with at most two non-zeros.) If two atoms j, j' at angular separation $\gamma \geq \delta$ were both active, adding their stationarity conditions $d_j^\top(x - Da) = d_{j'}^\top(x - Da) = \lambda$ would give $(a_j + a_{j'})(1 + \cos\gamma) = r\cos(\theta - \alpha_j) + r\cos(\theta - \alpha_{j'}) - 2\lambda \leq 2r\cos(\gamma/2) - 2\lambda$. When $\lambda/r \geq \cos(\delta/2)$ the right side is ≤ 0; for $\gamma < \pi$ the left side is positive, and for $\gamma = \pi$ it is 0 while the right side is exactly $-2\lambda < 0$, so both cases are contradictions. So at most one atom is ever active (at $\lambda = 0$ an antipodal pair can carry two, though some optimal code uses one and the kernel below is unchanged), atom j exactly where $r\cos(\theta - \alpha_j) > \lambda$ (an arc of half-width $\arccos(\lambda/r)$; the arcs are disjoint), and the cost kernel is a *sum* of translates rather than a minimum over them: $\bar{c}(\theta) = r^2/2 - \frac{1}{2}\sum_j g(\theta - \alpha_j)$ with $g(u) = \max(0, r\cos u - \lambda)^2$. Hence

$$\hat{\bar{c}}_q \;=\; -\tfrac{m}{2}\,\hat{g}_q\,S_q\,e^{iq\psi_q}, \qquad \tfrac{1}{m}\textstyle\sum_j e^{-iq\alpha_j} \;=:\; S_q\,e^{iq\psi_q},\ S_q \geq 0,$$

for $q \neq 0$ (so $S_q = |Z_q|$ and $\psi_q = -\arg Z_q/q$ for the moment $Z_q = \frac{1}{m}\sum_j e^{iq\alpha_j}$ whose phase App. C and App. K read: the two phases have opposite signs), so the q-th harmonic of J vanishes iff $\hat{\rho}_q = 0$, or $\hat{g}_q = 0$, or $S_q = 0$. For a regular m-gon $S_q = \mathbf{1}[m \mid q]$, recovering Corollary 1. For a radial law the hypothesis must hold on its support, which must therefore be bounded, $r \leq \lambda/\cos(\delta/2)$ (the unbounded lognormal radius of App. C violates it whenever $\delta < \pi$), and $\hat{g}_q$ is averaged over r. Since $S_q \leq 1$ with equality iff all $q\alpha_j$ coincide, the regular m-gon maximizes S_q for $q \in m\mathbb{Z}$: irregularity raises pinning only through harmonics off $m\mathbb{Z}$.

B EXACT 2-PLANE NUMERICS

In a 2-plane, some optimal non-negative code has at most two active atoms, so the cost is computed exactly by enumerating all active sets of size ≤ 2 (closed-form least squares per set, feasibility check, minimum over sets); we verified this solver against long-run coordinate descent to 2×10^{-16}. On a 720-point angular grid: the convolution identity holds to a relative error $\leq 4 \times 10^{-12}$ across densities (harmonic, von Mises mixtures, near-discrete); at $\lambda = 0.3$ and unit radius, the rotation is pinned exactly when $m \mid \ell$ in all 24 cells for $m \in \{2, 3, 4, 6\}$, $\ell \in \{2, 3, 4, 6, 8, 12\}$, with off-resonance variation *exactly* zero to machine precision (Corollary 1 proves only the flat direction; the other can fail at isolated λ, App. A); the curvature series of Proposition 6 matches central finite differences (step $h = 2\pi/720$) on regular m-gons against $\rho = (1 + a\cos m\theta)/2\pi$, $m \in \{3, 4\}$, $a \in \{0.1, 0.2, 0.4, 0.8\}$, to relative error $\leq 1.1 \times 10^{-4}$ and is linear in a (with the density's only

harmonic at $k = m$, $E \equiv 0$, so the residual is the finite-difference step's own error, $(mh)^2/12 \approx 5.7 \times 10^{-5}$ and 1.0×10^{-4} at $m = 3, 4$, not the truncation error of the approximation). That check validates the series, not its error bound; on the multi-harmonic densities C1 and C2 of App. C, with regular m-gons at $m = 2, 3, 4$, where $E \neq 0$, the exact series puts $|E|$ at 4.5×10^{-4} to 6.6×10^{-2} of the leading term, inside Proposition 6's bound in all six cases, which the bound nearly attains on C1, whose harmonics are all in phase at the minimum.

Free configurations. With the atom angles free rather than rigidly rotated, at $\lambda = 0.3$ and unit radius on $\rho \propto 1 + 0.8\cos\ell\theta$, $m \in \{2, 3, 4, 6\}$, $\ell \in \{3, 4, 6\}$ (`results/e0b_free_configs.json`; the Hessian test of `results/e0_theory_check.json`): at uniform ρ the optima of 12 random restarts spread over 7–11 distinct phases, consistent with a continuum up to the optimizer's resolution; on resonance the regular configuration's free Hessian is positive definite (tested at $m = \ell = 3$ and 4) and the optimum stays regular (irregularity $\leq 3\times10^{-7}$ for $m \geq 3$); off resonance ($m = 2, 4, 6$ at $\ell = 3$) the regular configuration's free Hessian has a negative eigenvalue (-0.12, -0.030, -0.012) beside positive ones; at $m = 2$ and 6, where $m \mid 2\ell$, the configuration is stationary and so a saddle, while at $m = 4$ it is not stationary at all (its per-atom gradients cancel in sum but not singly, so it leaves the polygon at first order; on a density $1 + a\cos\ell\theta$ the polygon's symmetry makes atom j's gradient $\propto \cos(\ell\alpha_j + \chi)$ with one amplitude and one χ for all j, so at any ℓ, unless that amplitude vanishes, the gradients vanish together only if $m \mid 2\ell$, and then only at isolated rotations; that the stationary ones are saddles is tested at $\ell = 3$ only), and in all six off-resonance cells of the search ($\ell = 3, 4, 6$) the optimum leaves the polygon (irregularity 0.03–0.34) at a single phase modulo the data's $\mathbb{Z}_\ell$.

Proposition 4 inside its own hypothesis. The closed form of Proposition 4 holds when the atoms' active arcs are disjoint, which the GemmaScope configurations of App. I almost never satisfy (at most 5 of 155), and the twelve-seed ring dictionaries of §5.1 never do (0 of 17,184 trained (seed, q) pairs, App. K). The sweep here constructs configurations that do: over the 57 of them whose minimum angular gap clears the threshold at their λ, the closed form matches the exact solver to a maximum relative error of 8×10^{-7} and a median of 7×10^{-9}, the accuracy of a 720-point quadrature of a kernel whose second derivative jumps at the arc ends ($(2\pi/720)^3 \approx 7 \times 10^{-7}$) (`results/e8_structure_factor.json`, test `disjoint_regime_identity`). Outside the regime the amplitude degrades smoothly—6.6% at 0.9× the threshold and 23% at 0.75×—and on GemmaScope's trained configurations it is off by a median 94% at $\lambda = 0.15$ (App. I), while on the ring dictionaries of §5.1 the *phase* survives to 4.9° where $S_q \geq 0.3$ and to 10.1° over all pairs (App. K). That is the whole basis for using the structure factor as a graded ordering rather than as a formula.

C SYNTHETIC SWEEP DETAILS

Data. $n = 30{,}000$ samples in $\mathbb{R}^{48}$: six orthogonal 2-planes (the first twelve columns of a Gaussian matrix's QR factor), each active independently with probability $n_i/6$, where n_i, drawn per sample, is 1 (probability 0.7) or 2 (so a sample can carry no plane and is then noise alone); radius e^z, $z \sim \mathcal{N}(0, 0.25^2)$, independent of the angle; angles by inverse-CDF sampling from the condition's density on a 720-point grid; isotropic Gaussian noise $\sigma = 0.02$. One draw (data seed 12345, the same in every condition; neither centered nor rescaled) is shared by all 20 SAE seeds (0–19) of a condition, as is one embedding of the planes—unlike App. M's design, which resamples the data per seed; agreement here can therefore carry a finite-sample or frame preference, and does where the objective cannot choose: on the uniform density one plane in six is concentrated (Rayleigh $p = 8 \times 10^{-4}$), and at $\ell = 3$ the seeds share one of three equally good m-phases (below). Conditions: (A) $\rho \propto 1 + a\cos 4\theta$, $a \in \{0, .15, .3, .5, .7, .9\}$, 4 atoms per plane; (B) $a = 0.9$, $\ell \in \{2, 3, 4, 6, 8\}$, 4 atoms per plane; (C1) fixed single-mode density, 0.9 of a von Mises law ($\kappa = 2.5$) plus 0.1 uniform (circular moments $R_{1..4} = 0.69, 0.35, 0.13, 0.04$), atom budget $m \in \{2, 3, 4, 6, 8\}$; (C2) fixed unequal 3-mode density, 0.9 of a von Mises mixture (modes at $0, 2\pi/3, 4\pi/3$ with weights $0.5, 0.3, 0.2$, $\kappa = 5$) plus 0.1 uniform with harmonic profile peaked at k=3 ($R_{2..6} = 0.153, 0.341, 0.045, 0.019, 0.026$), $m \in \{2, 3, 4, 5, 6\}$; (D) controls: uniform density; 1-D features. **SAEs.** ReLU + L_1 ($\lambda = 0.3$), width = planes × atom budget, Adam (learning rate 3×10^{-3}, batch 512), cosine schedule, 25 epochs, 20 seeds per condition; decoder columns Gaussian at initialization and renormalized every step, the untied encoder initialized to the decoder's transpose, both biases at zero, the decoder bias subtracted before encoding, and no dead-latent resampling. **Measurement.** Atoms are assigned to planes by

in-plane norm fraction $\|P_V d_j\| \geq 0.75$ (within 41° of the plane); the plane's configuration phase is $\arg\left(\frac{1}{m}\sum_j e^{im\alpha_j}\right)/m$, defined modulo $2\pi/m$. That phase is invariant under the configuration's own $\mathbb{Z}_m$ but not under a symmetry of the data it does not contain: in condition (B) a $2\pi/\ell$ rotation maps an optimum to an equally good one (Proposition 1) and moves the m-phase unless $\ell \mid m$, so Table 3 also reads the phase at harmonic $\mathrm{lcm}(m, \ell)$, which is invariant under both. The atom count that enters every measurement is the count *recovered* in the plane (per plane, the median over the 20 seeds, at which every seed's phase is read, a seed with no atom in the plane being dropped; in Fig. 2b, the median over a condition's planes), not the budget: at budgets 6 and 8 in condition C1 the SAE puts 5 and 7 atoms in the plane, which is why Fig. 2b has abscissae $2, 3, 4, 5, 7$ for C1 and $2, 3, 4, 5, 6$ for C2. A seed can recover a count other than its plane's median (529 of the 2,640 (plane, seed) pairs: 37–60% at C1's budgets 3–8 and C2's 5 and 6, 3–24% elsewhere); its phase is still read at the median's harmonic, which can add jitter there. Cross-seed jitter is the circular standard deviation $\hat{\sigma} = \sqrt{-2\log \bar{R}}\,/\,w$ of the phase across the 20 seeds, where $w = 2\pi/(2\pi/m) = m$ maps the phase onto the unit circle and $\bar{R}$ is the resultant length; we report $\hat{\sigma}$ normalized by $(2\pi/m)/\sqrt{12}$, the standard deviation of the uniform law on one period. The two are different functionals: the circular estimator diverges as $\bar{R} \to 0$ whereas the normalizer is finite, so the normalized quantity is not exactly 1 for uniform phases; empirically, with 20 seeds it saturates just above 1 (and we clip it at 3), which is why the "uniform" reference line in Fig. 2a, drawn at 1, marks that level only approximately. Uncertainties quoted for a condition are ± 1 standard deviation over its 6 planes. For the non-monotonicity claim of Fig. 2b we also give tests over planes: the three sizes are trained on the same six planes and data, so the test is paired: exact one-sided Wilcoxon signed-rank gives $p = 0.047$ for m=3 against m=2 (lower in 5 of 6 planes) and $p = 0.016$ for m=3 against m=4 (6 of 6, the minimum attainable, $1/64$). Of the $23 \times 6 = 138$ feature subspaces, the 6 lines of the one-dimensional control have no phase, leaving $n = 132$ for every statistic that involves jitter, curvature or the Boltzmann fit. On those 6 the read-out is the sign of the atom instead, and cross-seed sign agreement is 1.000 on all of them, with the aligned atom strictly beating the flipped one (0.256 against 0.679 in loss): below dimension two there is no continuous indeterminacy to have, which is what Proposition 5 asserts and all this control can show—the non-negative code on one-signed data leaves the aligned atom no competitor. Oracle landscapes use the exact 2-plane solver on the plane's own activations; the deployable (single-SAE) landscape rotates the fitted SAE's atoms and evaluates the actual training loss, but it still takes the plane, and the atoms assigned to it, from the ground truth, so what it drops is the oracle's exact solver, not its knowledge of the plane. Both are computed for seed 0's configuration, the oracle one on the first 4,000 samples whose in-plane radius exceeds its median over the whole dataset, and profiled over one period $[-\pi/m, \pi/m]$ of the configuration's symmetry, at 49 rotations (oracle) and 25 (single-SAE); curvature is twice the leading coefficient of a least-squares quadratic through the 9 (oracle) or 7 (single-SAE) profile points around the minimum, and depth is $\max - \min$ in loss units, the convention of §3.3, distinct from the relative depth used on GemmaScope (App. E). For an irregular configuration that window need not span a period of J (2π on the C densities, π at $\ell = 2$), so depth is the range over the window; the minimum falls on its edge in 7 of the 132 oracle profiles and 3 single-SAE ones, all on densities with $R_m < 0.01$.

Condition (B): the binary rule on trained dictionaries. Table 3 reports the family that tests Corollary 1 directly, and it is the reason the body tests the graded criterion rather than the binary one. Corollary 1's hypothesis is a regular m-gon; a free dictionary is under no obligation to be one, and here it demonstrably is not: the atoms are least regular at $\ell = 2, 3$, two of the three rows where regularity would deny them the density's harmonic. The off-resonance ℓ=2 planes are therefore partly pinned rather than flat, through a configuration that has left the polygon, while at $\ell = 3, 6$ the quotiented phase is at chance. At ℓ=3 this reading cannot settle the graded rule: its predicted curvature equals ℓ=2's ($\ell^2|\hat{c}_\ell|$ 0.054 against 0.055) and its quotiented jitter is at chance (0.92 against 0.49), but chance modulo 30° is a physical spread of 8.7°, below the 12.7° of ℓ=2's seeds. Read at m, its seeds' circular spread is 15.8° and are concentrated in 6 of 6 planes (Rayleigh $p \leq 0.01$), on one of three m-phases the population objective ranks equally—a choice only the shared data draw or frame can make (Data, above). This is the same phenomenon as the saddle-breaking of §4, seen from the configuration side rather than the loss side, and it is a limitation of the corollary rather than of Proposition 3, which assumes nothing about the configuration.

Independent draws and frames. The 20 seeds of a condition share one data draw and one coordinate frame, so some of their agreement could be the sample's, or Adam's preference for the axes, rather than the population objective's. Rerunning the whole sweep with every seed after the

density harmonic ℓ (atoms fixed at $m = 4$)	Cor. 1 predicts	jitter at m (norm.)	jitter at $\mathrm{lcm}(m, \ell)$ (norm., quotiented)	gap CV of atoms	$\lvert\hat{\tilde{c}}_\ell\rvert$ of the recovered config.
2	flat	0.488 ± 0.142	0.488 ± 0.142	0.189	13.7×10^{-3}
3	flat	0.610 ± 0.040	0.921 ± 0.176	0.110	6.0×10^{-3}
4	**pinned**	0.106 ± 0.022	0.106 ± 0.022	0.058	12.0×10^{-3}
6	flat	0.881 ± 0.092	0.949 ± 0.166	0.052	0.5×10^{-3}
8	**pinned**	0.756 ± 0.205	0.506 ± 0.134	0.033	2.3×10^{-3}
regular 4-gon (reference)		—	—	0	$< 10^{-14}$ (ℓ=2, 3, 6)
uniform phase (chance)		≈ 1	≈ 1	—	—

Table 3: **Condition (B): the one test of the *binary* rule on trained SAEs, and why it comes out graded.** Four atoms per plane, 20 seeds, 6 planes per row (mean $\pm$ sd over planes); the density is $1 + 0.9\cos(\ell\theta)$, so Corollary 1 predicts a pinned configuration exactly when $4 \mid \ell$. The density is invariant under a $2\pi/\ell$ rotation, so by Proposition 1 an optimum and its rotation are equally good; at $\ell = 3, 6, 8$ they carry different m-phases (three, three and two antipodal values), and cross-seed agreement at harmonic m cannot register pinning there. The fourth column reads the phase at $\mathrm{lcm}(m, \ell)$, which quotients by both symmetries (`code/g54_condb_quotient.py`); at $\ell = 2, 4$ the two statistics coincide. Read that way the binary rule is not contradicted in 3 of the 5 rows—ℓ=4 pinned sharply (0.106), ℓ=3, 6 flat (0.92, 0.95), though only in the weak sense this reading allows: at harmonic 12 chance is a physical spread of 8.7°, so ℓ=2's 12.7° would read as flat too—and graded in the other two: ℓ=8, predicted pinned, is pinned only partly (0.506, and its m-phases split between the two antipodal optima: the resultant at $2m$ exceeds the one at m in five of six planes; quotienting removes that split, yet 0.506 is twice the 0.24 predicted from ℓ=4's jitter by the curvature ratio (spread $\propto 1/\sqrt{J''}$), a gap we do not explain), and ℓ=2, predicted flat, is partly pinned (0.488, itself 2.5 times the 0.20 the same ratio predicts from its measured kernel, so the magnitude is off by a factor of about two in both graded rows). Proposition 6's curvature $\propto \ell^2 |\hat{\rho}_\ell| |\hat{\tilde{c}}_\ell|$, evaluated on each row's own measured kernel (exact here, since the density has only the harmonics $0, \pm\ell$), ranks the rows at Spearman $+0.90$ against the quotiented jitter's negative (positive is the predicted direction) and $+0.80$ with the ℓ^2 weight dropped; against the m-phase jitter the order was reversed ($+0.70$ and $+0.90$), an artifact of the unquotiented statistic; in physical angle (each jitter divided by the harmonic it is read at, ℓ=3, 6 saturated) it is $+0.70$. The last two columns say why ℓ=2 is not flat: the recovered atoms are not a regular 4-gon, and they are least regular at $\ell = 2, 3$, where regularity would cost them a harmonic (gap CV 0.19, 0.11, 0.06, 0.05, 0.03 at $\ell = 2, 3, 4, 6, 8$; ℓ=6 is the exception). Computing the exact kernel of the *measured* configuration rather than of an idealized 4-gon recovers a non-zero $|\hat{\tilde{c}}_\ell|$ where the polygon has none (1.4×10^{-2} at ℓ=2 against 4×10^{-18}), so these planes are pinned by the criterion, through a configuration that left the polygon. The direction of causation is not resolved here: deformation could be the response to the harmonic or its cause. No plane is anywhere near the jitter clip of 3 (the observed maximum over all 132 planes is 1.42).

first on its own draw from the same population, embedded by its own Haar rotation of $\mathbb{R}^{48}$ and rotated back before measurement (`ANGSYM_RESEED=1`; `code/e1b_reseed_summary.py`), leaves what the body uses: the (A) family's jitter still falls with R_m at Spearman -1.00 (range 6.3×); $m = 3$ is still the minimum of the non-monotone family (C2: 0.14, against 0.32 at $m = 2$ and 1.08 at $m = 4$); condition (B) still pins $\ell = 4$ (0.16), partly pins $\ell = 2$ (0.56, concentrated in 5 of 6 planes) and $\ell = 8$ (0.48 quotiented), and leaves $\ell = 6$ at chance; and the pooled curvature correlation is -0.67 / -0.69 (oracle / single-SAE, against -0.77 shared). What changes is what should: the uniform control's jitter rises from 0.92 to 1.01, the spread of uniformly random phases, and at $\ell = 3$ the seeds' concentration on one of three equally good m-phases, in 6 of 6 planes with the shared draw, falls to 0 of 6 (jitter at m 0.61 $\rightarrow$ 1.00), so that choice was the shared draw's or frame's, as said above. The condition means of the two runs correlate at $+0.87$ (Spearman, 22 conditions).

In this setting the curvature is the slightly better predictor of jitter (Spearman -0.774 oracle / -0.772 single-SAE) and depth is close behind (-0.755 / -0.703); the two diverge only at production scale, where the profile is coarse (App. E). These pool all 132 planes with the m-phase jitter; reading the 18 condition-(B) planes at $\ell = 3, 6, 8$ modulo their symmetry instead gives -0.755 / -0.764 for the curvature and -0.696 / -0.656 for the depth, so the oracle and single-SAE curvatures stay within 0.01 of each other (the depths within 0.06). The Boltzmann prediction—the normalized circular

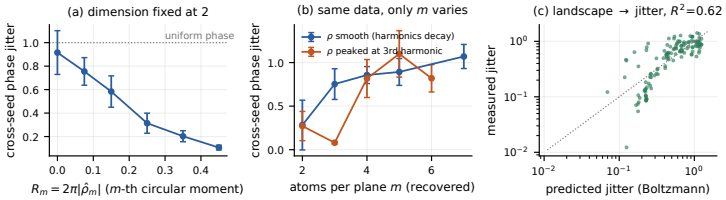

Figure 2: **Synthetic sweep with subspace dimension fixed at 2 everywhere.** (a) Cross-seed phase jitter of SAE atoms in a plane (20 seeds) against the m-th circular moment $R_m = 2\pi|\hat{\rho}_m|$ of the angular density, at constant dimension (Spearman -0.95 over the 36 planes; -1.00 over the 6 conditions, the independent units, exact one-sided $p = 0.0014$). Jitter is the circular spread in units of the uniform law's standard deviation, clipped at 3; uniform phases over 20 seeds give ≈ 1 (dotted; App. C). Points are condition means, bars ± 1 s.d. over their 6 planes. (b) Same data and dimension, only the atom count varies. With a smooth density (blue), more atoms means less stability, the width law. With a density peaked at the 3rd harmonic (orange), stability is *non-monotone* with a sharp optimum at $m = 3$ (0.080 ± 0.017 vs 0.272 ± 0.168 at $m = 2$): more atoms, *more* stable, exactly where $|\hat{\rho}_m|$ peaks. No account monotone in m produces this; one in which atoms match the density's modes does, and App. C gives the two results that separate it from the criterion. (c) A one-parameter Boltzmann model over the *measured* landscape fits the jitter (in sample) across all 132 planes of the 22 conditions (log-scale $R^2 = 0.62$) with a single global temperature (dotted: $y = x$). Planes within a condition share a data draw, so the 20 distinct conditions, not the planes, are the units, since (A) at $a = 0.9$ and at $a = 0$ repeat (B) at $\ell = 4$ and (D)'s uniform control; the (b) sweeps also reuse one density across m (App. C).

standard deviation of $p(\varphi) \propto e^{-J(\varphi)/T}$ over the plane's oracle profile, the functional applied to the seeds' phases—fits one global temperature T by least squares on log-spread over all planes; $T = 1.2 \times 10^{-3}$, log-scale $R^2 = 0.62$ in-sample (one fitted parameter), $n = 132$ planes. Fitted on 21 of the 22 conditions and scored on the held-out one, it gives $R^2 = 0.60$ (T within 1.09–1.27×10^{-3} across folds), against 0.15 for the same held-out scoring of a linear fit of log jitter on the atom count m and 0.67 for one on the moment R_m alone, so in this sweep the landscape adds no held-out accuracy over the density's moment (`code/g82_boltzmann_heldout.py`; R^2 over the 132 planes' held-out predictions). Two conditions repeat another's specification and runs—(A) at $a = 0.9$ is (B) at $\ell = 4$, and (A) at $a = 0$ is (D)'s uniform control—so 20 conditions are independent, and holding out one of a twin leaves its copy in training; counted once each (120 planes; `code/g87_recounts/dedup.py`), the held-out R^2 is 0.58 for the Boltzmann model, 0.61 for R_m and 0.18 for m, the values §4.1 quotes, and the by-condition curvature Spearman -0.880 / -0.881 (oracle / single-SAE). **The relation at stationarity.** The sweep's training is annealed (Adam, cosine schedule), so its Boltzmann temperature is fitted, not derived. Constant-step SGD has a stationary law, and in the small-noise limit the variance of a coordinate δ along which the objective has curvature J'' is $\eta_{\text{eff}}\sigma^2/(2BJ'')$, with η_{eff} the step size along δ, σ^2 the per-sample gradient variance along δ and B the batch size (Mandt et al., 2017). Away from that limit, and with anisotropic or state-dependent gradient noise, SGD's stationary law is in general not a Boltzmann law on the training loss (Chaudhari & Soatto, 2018; Yaida, 2019), so the relation is an approximation to be checked, not a theorem about our runs. We test it with nothing fitted (`code/g66_fluctuation.py`): four unit atoms in a plane, trained by plain SGD on their angles with the *exact* non-negative code, on data from $\rho \propto 1 + a\cos 4\theta$ ($\lambda = 0.3$, $B = 32$), with δ the rigid rotation, the mean of the four angles (so $\eta_{\text{eff}} = \eta/4$ for angle steps η); after a burn-in of 20 relaxation times we record δ for 200 more relaxation times, and compute J'' and σ^2 with the exact solver at the time-averaged configuration. Over $a \in \{0.1, 0.3, 0.6\}$ and $\eta \in \{0.02, 0.05, 0.1\}$ the observed variance spans $31\times$ and matches the prediction to a ratio of 0.87–1.05 (median 1.01; log–log slope 1.02); the one-parameter fit above is the annealed analogue of this derived relation, not a replacement for it. The dimension-style baseline (atom count m) correlates with jitter at $+0.54$ pooled ($R^2 = 0.15$ held out, above), and cannot reproduce the (a) sweep (constant m) or the (b) non-monotonicity; since m is constant over (A), (B) and (D), the fair comparison is on the C families, where it varies, and there too it trails the moment ($R^2 = 0.25$ held out against 0.47 for R_m, 10 conditions).

Figure 2(a)'s condition means span $8.6\times$ ($0.916 \to 0.106$) and individual planes $14\times$ ($1.21 \to 0.085$). In panel (b) the abscissa is the atom count *recovered* in the plane (median over seeds in each plane, then over the condition's planes), which can fall one or two short of the budget at the largest budgets. The principal-angle statistic (largest principal angle between a seed's in-plane span and the true plane) is $6.8^\circ \pm 1.3^\circ$ over the 132 planes (13.2° at worst) and $6.3^\circ \pm 0.9^\circ$ on the 72 planes whose density carries little power at the atom count ($R_m < 0.05$), against 85.6° for a random 2-plane in $\mathbb{R}^{48}$; it is conditional on the ≥ 0.75 in-plane assignment rule, so it says how well the recovered atoms span the plane, not how much of it the SAE captured, and an atom with three quarters of its norm in a plane is within 41° of it by construction, so the random-plane figure is the scale of the ambient space rather than a null. Without the rule (`code/e1b_plane_recovery.py`, which retrains the same seeds and reproduces the $6.8^\circ \pm 1.3^\circ$): over all 2,640 (plane, seed) pairs the SAE's reconstruction returns a median 0.953 of each plane's variance (5th percentile 0.80; the minimum, 0.28, is three seeds of the two-atom budget on the three-peaked density), every pair has at least one atom assigned and 97% have at least $m - 1$, and the two best-aligned atoms carry a median 0.99 of their norm in the plane. Over all 22 conditions the by-condition curvature comparison is -0.891 vs -0.893 (oracle vs single-SAE, the order used throughout); the depth summary gives -0.755 vs -0.703 by plane and -0.843 vs -0.769 by condition.

Replication over generators. Within the main sweep the non-monotonicity rests on six planes sharing one data draw (paired Wilcoxon $p = 0.047$ vs m=2, 0.016 vs m=4), so the independent-unit count is one. Rerunning the $m \in \{2, 3, 4\}$ family and the smooth-density width family with eight independent generators (data seeds $101, 202, \ldots, 808$; `code/e10_nonmono_replicates.py`) reproduces both: jitter at m=3 falls below both neighbors in 8 of 8 generators (the middle of three exchangeable values is the smallest with probability $1/3$, so $p = 3^{-8} = 1.5 \times 10^{-4}$; exact one-sided Wilcoxon $p = 0.0039$ against each neighbor), with medians 0.28 at m=2, 0.10 at m=3 and 0.73 at m=4, while on the smooth density the m=3 dictionaries sit at 0.68–0.90 and jitter rises with m in every generator (Spearman median $+0.95$, minimum $+0.80$). Table 2 of §5.1 remains the test on a density we did not choose. The regular-polygon hypothesis is load-bearing wherever we invoke it, and it bounds flatness by $\mathrm{lcm}(m, p)$ for a p-point density. The day ring's 4th harmonic is among the weakest it carries (weakest at layer 7, second weakest at 12 and 20; Table 8), which fits the m=4 non-monotone cases but not the others.

Against a simpler account (Fig. 2b). Seeds agreeing when atoms match modes is the clustered limit of the criterion, and two results lie outside it: the one-mode width family is pinned at m=2 (jitter 0.28) but not m=4 (jitter 0.86), and an installed q leaves its clusters' other harmonics unlocked (agreement 0.30 against 0.94 at q; App. K). A graded mode-matching account would also predict that jitter falls with the amplitude of four fixed modes ($0.76 \to 0.11$ over $a > 0$), so that result does not separate the two.

D MODULAR ADDITION DETAILS

One-layer transformer (width 64, 4 heads, MLP 256, no LayerNorm) trained on $(a + b) \bmod p$, $p = 53$, with an 85% train split of the 53^2 pairs (AdamW, learning rate 10^{-3}, $(\beta_1, \beta_2) = (0.9, 0.98)$, weight decay 0.5, batch 512, 9,000 steps, seed 0), to 100% train and test accuracy. The MLP output at the final position carries the sum: conditioning on $s = a+b \bmod p$ explains 77.4% of its variance, and four Fourier planes at key frequencies $\{1, 3, 9, 26\}$ explain 99.7% of that component. We circularize the sum curve by projecting to those planes and equalizing each plane's two axes, giving a curve C_s; the SAE data at level ε are $X = (1 - \varepsilon)C_s + \varepsilon(\bar{H}_s + r)$ over all 53^2 inputs, with $\bar{H}_s$ the conditional mean and r each input's residual, rescaled to unit mean norm ($\varepsilon \in \{0, 0.25, 0.5, 0.75, 1\}$; $\varepsilon = 1$ is the network's own activations), and the symmetry defect in Fig. 3a is the Frobenius ratio $\|[X_{s+1} - RX_s]_s\| / \|[X_s]_s\|$ over the 53 curve points, i.e. of the conditional means (the residual term εr, which also breaks the symmetry, is not counted). The generator R is the direct sum of in-plane rotations by $2\pi k/p$ over a single Gram–Schmidt orthonormal frame of the 8-dimensional key subspace (without the shared frame, R fails orthogonality at 3.7×10^{-2}; with it, 2.7×10^{-16}). SAEs: ReLU + L_1, width 16, $\lambda = 0.12$, Adam at 3×10^{-3} with a cosine schedule, batch 256, 150 epochs (75 for the replicate seeds of the reweighting family), 24 seeds, initialized as in App. C. Landscape evaluation applies the group element to $(W_d, W_e^\top, b_d)$ jointly, as Proposition 2 requires, and reports its range over its mean; omitting b_d makes the exactly-flat landscape appear spuriously non-flat, which we flag as an easy-to-make measurement error.

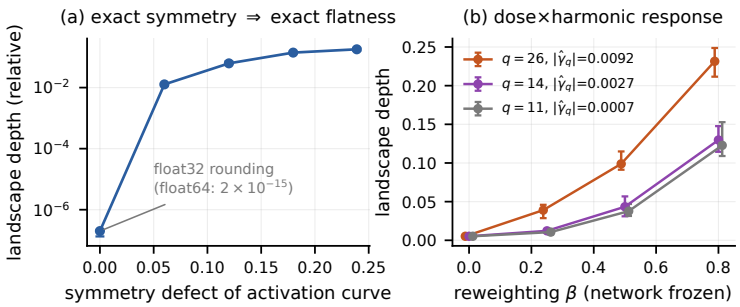

Figure 3: **Modular addition: an exactly-known indeterminacy group on a trained transformer.** (a) Rotation-landscape depth (range over mean loss) of SAEs on the network's activations, as the data are interpolated from the circularized curve (exact $\mathbb{Z}_{53}$ symmetry; the constructed map is orthogonal to 2.7×10^{-16} and satisfies $R^{53} = I$ to 7×10^{-15}) to the network's own activations; median and interquartile range over 24 replicate SAEs. Depth rises monotonically in the measured symmetry defect, from 2×10^{-7} to 0.18 (in 23 of the 24 seeds individually). The 2×10^{-7} is float32 rounding: re-evaluated in float64 the same landscapes are flat to 2×10^{-15}, as Proposition 2 requires over the 53 group elements; this panel does not measure the residual Corollary 1 predicts under a continuous rotation (App. D). (b) Causal check with the *network frozen*, on a resample of the circularized curve ($\varepsilon = 0$): reweighting it by $1 + \beta \cos(2\pi qs/p)$ breaks the symmetry of the data alone and deepens the landscape (relative depth, as in (a)) monotonically in β at all three probed harmonics q (median and interquartile range over 24 replicate SAEs per point). The harmonic the dictionary's cost profile feels most ($q = 26$, largest $|\hat{\gamma}_q|$) responds most (0.23 against 0.13 and 0.12 at $\beta = 0.8$, two-sided Mann–Whitney, normal approximation, $p = 3 \times 10^{-9}$ and 6×10^{-9}); the other two, whose $|\hat{\gamma}_q|$ differ by 3.7× in the one SAE that measures them, do not separate ($p = 0.81$). $q = 26$ is also a key frequency of the network, and a run that separates the two finds the response following the key frequencies, not $|\hat{\gamma}_q|$ (text).

How flat is "flat". The landscape of Fig. 3a is the objective at the 53 group elements R^t, $t = 0, \ldots, 52$, not along a continuous rotation. For exactly $\mathbb{Z}_{53}$-symmetric data Proposition 2 gives $L(R^t) = L(I)$ at every t with no residual, whatever configuration the dictionary reached: the harmonics in $53\mathbb{Z}$ that survive Corollary 1 under a *continuous* rotation are constant on that grid by construction (a 53-point comb has $|\hat{\rho}_{53j}| = |\hat{\rho}_0|$, and R^t samples the rotation exactly at the period of those harmonics). The measured 2×10^{-7} is a relative range, about two float32 units in the last place; re-evaluating the same 24 seeds' landscapes in float64 gives a median 1.9×10^{-15} (`code/e2b_modular_seeds.py`), so it is arithmetic noise, not a residual. A continuous rotation of the same dictionaries would carry a residual at the comb's harmonics—at the $\mathrm{lcm}(m, 53)$ harmonic for a regular m-gon, at 53 itself for an irregular configuration such as a trained one; for regular m-gons it is small: the exact 2-plane solver puts $|\hat{\tilde{c}}_{mp}|/\hat{\tilde{c}}_0$ at 3.0×10^{-7}, 1.3×10^{-6}, 3.4×10^{-7} and 6.0×10^{-7} for regular $m = 2, 3, 4, 5$ at unit radius and this experiment's $\lambda = 0.12$ (`code/e0c_lcm_residual.py`; evaluated on a 10,600-point grid, a multiple of 53 so the comb is exact and $k = 265$ is resolved, and unchanged to two digits from a 2,120-point grid), within a factor of five of one another, and a relative range of J ($4|\hat{\tilde{c}}_{mp}|/\hat{\tilde{c}}_0$, Corollary 1) of 1–5 $\times 10^{-6}$ under a continuous rotation, which the grid of group elements does not see by construction. These are regular m-gons in one plane, a proxy: the trained dictionaries hold 16 atoms across four planes and are not regular, so the proxy does not bound their residual. The dose experiment of panel (b), which breaks the symmetry of the data rather than rotating the atoms, is therefore the one that tests a prediction, an ordering by harmonic rather than a magnitude.

Orbit scatter and quotient. At exact symmetry the replicate seeds' recovered shifts (each seed's atoms located on the curve and cross-correlated with seed 0's) do not concentrate on any element of $\mathbb{Z}_{53}$ (Rayleigh $p = 0.54$, and 0.36 at $\varepsilon = 1$; a test with little power at 24 seeds, below), while raw cross-seed max-cosine is already 0.80; aligning seeds by the best group element raises it only to 0.85, so most of what separates two seeds is not an orbit move. We do *not* read that +0.05 as evidence of orbit structure. It is a maximum over 53 candidate alignments, and the same procedure yields +0.030 at $\varepsilon = 1$ (the network's own activations; $\varepsilon = 0$ is the circularized curve) and +0.086

at $\varepsilon = 0.5$, with no monotone relation to the symmetry defect. Without a null for the best-of-53 selection the statistic is uninterpretable, and we report it only for completeness; the orbit statement at $\varepsilon = 0$ is Proposition 2's, confirmed by the flat landscape; the shift distribution is consistent with it and does not test it.

Causal reweighting (partial result). With the network frozen we reweight the circularized ($\varepsilon = 0$) data by $w(s) \propto 1 + \beta\cos(2\pi qs/p)$ and re-train replicate SAEs. The reweighted data are 3×53^2 points resampled from the $\varepsilon = 0$ set with those weights. The dictionary can only feel harmonics present in its per-position cost profile $\gamma(s)$; we measure $|\hat{\gamma}_q|$ from a single SAE (seed 0, trained for 150 epochs on the unweighted set) and pick q at the top, middle, and bottom of its spectrum ($|\hat{\gamma}_q| = 0.0092, 0.0027, 0.0007$). Read on all 24 replicate SAEs per point (`code/e2b_modular_seeds.py`), the landscape depth responds to dose ($\beta \in \{0, 0.25, 0.5, 0.8\}$) at every harmonic: from a median 0.005 at $\beta = 0$ to 0.232 (q=26), 0.130 (q=14) and 0.123 (q=11) at $\beta = 0.8$, monotone in β in all three series (Fig. 3b), and the harmonic the dictionary feels most, which is also a key frequency of the network, responds most ($p = 3\times10^{-9}$ and 6×10^{-9} against the other two, two-sided Mann–Whitney over seeds, normal approximation; they do not separate, $p = 0.81$). The 24 seeds of a point share its one resample, so these p are over initializations, not over draws. At a fixed dictionary the reweighting would move the relative depth by $2\beta|\hat{\gamma}_q|/\bar{\gamma}$, 0.064, 0.019 and 0.005 at $\beta = 0.8$ for the probe SAE ($\bar{\gamma} = 0.229$); the retrained dictionaries go 3.6, 7 and 25 times deeper, so most of the response is not the probe's harmonic content, and which harmonic responds most follows the network's key frequencies instead (below). The two flatness baselines in this section are different measurements and differ by four orders of magnitude for a stated reason: the interpolation family is evaluated on the exact circularized curve, where the only residual is float32 rounding, while the reweighting family resamples a finite set, whose β=0 draw is not exactly symmetric and sets a depth floor of 5×10^{-3} (its SAEs also train for 75 epochs rather than 150). That breaking a symmetry deepens a landscape is what any account predicts; what is specific to the criterion is that the response follows the dictionary's own harmonic content, and it does at the harmonic that content singles out. That harmonic, $q = 26$, is also one of the network's four key frequencies and 14 and 11 are not, so this contrast does not separate the dictionary's harmonic content from the representation's own Fourier basis, and the run that separates them (below) finds the representation's basis. With the network, the dictionary size and the training protocol held fixed, changing only the angular symmetry of the data deepens the landscape by more than an order of magnitude. We compare no measured depth with Proposition 6's magnitude: the dictionaries span four planes, and the proposition is for one. Separately, with 24 seeds the cross-seed *concentration* statistic does not separate from uniformity (Rayleigh $p > 0.05$ at every dose and harmonic): the induced tilt is real in the objective, but at 24 seeds the Rayleigh test (critical resultant 0.35) has 80% power only for a population resultant of about 0.43, so this endpoint cannot distinguish a small effect from none. We report the dose effect as consistent with the mechanism (and with any account), the cross-seed endpoint as open, and the lead of the harmonic the dictionary feels most as explained by the key frequencies, not by the dictionary's content.

Key frequencies, not the cost profile (a failed test). To separate the two, a further run (`code/e2c_modular_keyfreq.py`; $\beta = 0.8$, the same interpolated representation, 24 replicate SAEs per cell, one fresh resample per cell) reweights at two harmonics that are *not* key frequencies but carry large $|\hat{\gamma}_q|$ ($q = 18, 24$: $|\hat{\gamma}_q| = 0.0060, 0.0057$, ranks 3 and 4 of 26) and at two key frequencies with small $|\hat{\gamma}_q|$ ($q = 9, 1$: $0.0014, 0.0018$, ranks 20 and 18). If the response followed the dictionary's harmonic content the first pair would deepen more; the reverse holds: medians 0.182 and 0.200 against 0.267 and 0.250, the key frequency deeper in all four pairings (one-sided Mann–Whitney $p \leq 7 \times 10^{-6}$, over initializations as above), with $q = 26$, key and largest, at 0.212 on its own new resample (below $q = 9$ and 1). The retrained dictionaries respond to where the network's own Fourier planes are, not to the probe dictionary's cost profile, so test (6) of App. S is a failure; the dose response itself, tests (4) and (5), is unaffected, and it is not specific to the criterion.

The generator's rotation frequencies are exactly the network's key frequencies $\{1, 3, 9, 26\}$, and over the group elements every $m < 53$ dictionary is flat (under a continuous rotation, up to a residual at the comb's harmonics, estimated only for regular m-gons through the 2-plane proxy). With the network frozen, reweighting the circularized curve deepens the landscape monotonically in dose at every harmonic probed, most at the network's key frequencies rather than where the dictionary's cost profile points. Replicate SAEs give cross-seed max-cosine 0.80 with recovered shifts spread

over $\mathbb{Z}_{53}$; the alignment gain from quotienting by the estimated orbit is a best-of-53 maximum and we do not read it as evidence.

E GEMMASCOPE DETAILS AND FULL TABLES

The consequences that fail at this scale. In the order Eq. 2 implies them. (i) The L_0 law: the criterion says pinning should follow the atoms *per manifold*, and on GemmaScope the measured trend with L_0 is a dilution artifact of the group's reconstruction energy rather than a statement about m (Table 4, and the unresolved tension §3.3 flags). (ii) The per-feature diagnostic: the group's reconstruction share predicts inference-time interchangeability (OMP retention) better than the landscape depth does, so at this scale the naive read-out measures the wrong thing (Table 5). Both are tests of a consequence that fail their control, not evidence that the identity is false, and the reason is the same: the quantity a frozen encoder exposes is the share, not the density. (iii) Manifold discovery, which would supply the planes these tests need, is blocked for a different reason: decoder geometry alone finds cross-seed planes for only 27 of $4{,}000$ candidate clusters, mostly atom pairs, on Pythia-160M's replicate seeds (App. N).

What §7 compresses. Reading the rotation landscape off a *frozen* encoder does not measure Eq. 1, which is why §5.1 re-solves the code. Over the 41 of 78 (ring, layer) cells at width 16k with two or more atoms in the plane ($\tau = 0.5$), a median 99.2% of the frozen-encoder landscape's power sits at $k = 1$ (minimum 85.8%) and at most 4×10^{-5} at $k \in p\mathbb{Z}$, natural and symmetrized differing by $1.06\times$; fourteen cells from five other item sets, four of which fail the cyclic-order test, reproduce it. Exact re-solving on the same residuals, with the plane centered, removes the bowl in six of the eleven cells (App. I's measurable ring, layer and width combinations at layers 7, 12 and 20, not the (ring, layer) cells of §5.1) and leaves it in the five whose residual offset exceeds half the ring radius, the natural ring deeper in 11 of 11 either way (median $2.1\times$): the bowl is mostly the residual's in-plane offset—without centering a 0.919 share stays at $k = 1$—and secondarily the amortization gap.

On GemmaScope itself—layers 5/12/20, widths 16k–262k, L_0 18–475 measured, 31 SAEs, $61{,}870$ features—the L_0 law—depth falling with L_0, which is what the tiling relation predicts if L_0 stood in for m, and the direction of the published trend—looks confirmed (mean Spearman -0.983) but the estimator rotates a fixed group of four atoms while L_0 grows 7–$21\times$ along a sweep ($27\times$ over all 31 SAEs), and normalizing the profile by the group's reconstruction energy weakens the trend to a mean -0.33 (four of six sweeps still negative). L_0 counts atoms co-active per *input* while the theory's m counts atoms tiling one *manifold*, and production SAEs make them anti-correlated, so the criterion at face value predicts agreement should *improve* with L_0, opposite to the published trend—a tension we do not resolve. The per-feature diagnostic looks confirmed too, but reconstruction share beats it, and controlling for it collapses every flatness statistic to between -0.02 and -0.12. We do not claim the diagnostic; a future estimator has to clear that baseline.

Concurrent work. A concurrent workshop submission, publicly posted in anonymous form and therefore not citable by name, reports that OMP re-solving the coding problem on a frozen GemmaScope dictionary attains slightly better reconstruction with a largely different support at a single width. It has priority on that observation; the dissociation reported here is reached independently across a five-point width sweep.

Setup. Gemma-2-2B residual-stream GemmaScope SAEs (Lieberum et al., 2024; Rajamanoharan et al., 2024); layers 5, 12, 20; widths 16k/32k/65k/131k/262k (layer 12) and 16k/65k (layers 5, 20); five released L_0 variants per (layer, width) sweep of Table 4 (three at layer-12 65k: tags 21, 72, 141) and, for the width law, one SAE per layer-12 width at matched L_0 (tags 82, 76, 72, 67, 67); 31 SAEs in all, overlapping App. Q's twelve only in part. Activations: a pool of 1.2M residual vectors at the SAE's hook point from a public Pile slice, sequence length 256, first token dropped; landscapes and firing statistics use the first 300k of the pool, and the OMP comparison a random 4,096 drawn from it (OMP retention is recorded for nine SAEs: Table 5's five layer-12 widths and, at layers 5 and 20, the 16k and 65k SAEs with L_0 tags 68, 53, 71 and 61). Hook verification: for every SAE the measured L_0 matches the advertised `average_l0` tag (measured 1–11% above the tag for all 31 SAEs, e.g. 84.6 vs 82, 68.0 vs 67) and FVU falls in $[0.093, 0.305]$ across the 31 SAEs, inside the expected band at both ends, two independent checks that we attached the SAE to the tensor it was trained on. **Landscapes.** The grouping rule differs between the two settings and we state it once:

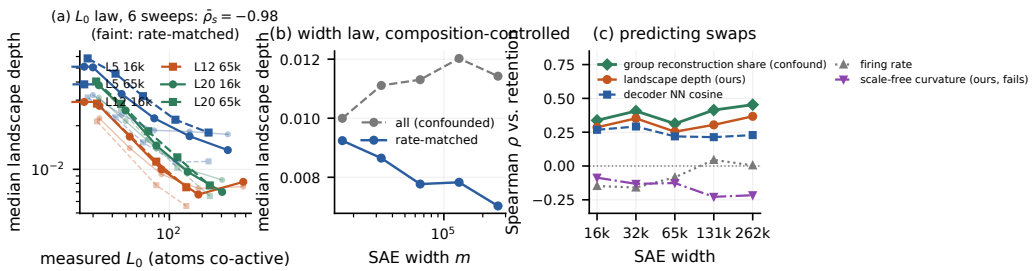

Figure 4: **GemmaScope, Gemma-2-2B: 31 production SAEs, all quantities from single SAEs.** (a) The co-activation (L_0) law: median rotation-landscape depth versus measured L_0, six sweeps (3 layers $\times$ 2 widths; not independent replicates, Table 4), all six decreasing and five of six strictly monotone (mean Spearman -0.98; depth falls 3.2–5.4$\times$ end to end). (b) The width law at layer 12 is invisible in the raw feature population (gray), which in fact trends the other way (Spearman $+0.90$), because wider SAEs hold rarer features, whose landscapes are deeper; restricted to a matched firing-rate band (blue) the predicted decrease appears (Spearman -0.90, exact permutation $p = 0.083$ two-sided at $n = 5$ widths, not significant at 0.05). The band $[10^{-3}, 10^{-2}]$ was chosen after seeing the raw comparison, so this panel is a confound diagnosis, not an independent test of the width law. (c) Per-feature prediction of inference-time interchangeability: Spearman of predictors against OMP retention at five widths. Landscape depth from a single SAE (orange) is above decoder nearest-neighbor cosine (blue) and firing rate (gray) at every width, individually significant at three of five though none survives a Holm correction over the five (paired bootstrap; the margin over cosine is $+0.02$ to $+0.14$ and not monotone, Table 5). But the group's reconstruction share (green, the top series) beats all three (Table 5) and absorbs depth entirely, and the scale-free curvature (purple) runs the wrong way, which is why this panel is a confound illustration and not a positive result.

the controlled experiments (App. C) take the plane from the ground truth and assign atoms to it by in-plane norm, because there the planes are known; the production experiments group by decoder cosine, because at 262k latents co-activation grouping needs codes for all pairs. For $\sim$2,000 sampled live features per SAE: group the feature with its 3 nearest decoder-cosine neighbors, restrict to the group's active samples (up to 1,024), rotate $(W_{\text{dec}}, W_{\text{enc}})$ of the group rigidly in the principal plane of the group's decoder vectors, and record MSE + penalty over 9 rotation angles in $[-\pi/4, \pi/4]$; depth is (max$-$min)/min. One full encode/decode pass per SAE provides residuals and firing statistics; everything else is restricted to sampled columns, keeping the largest SAE (262k) tractable on one GPU. The curvature reported in Table 5 is $2c_2/$energy from a quadratic $c_2t^2 + c_1t + c_0$ fit to the ≤ 5 profile points around the discrete minimum; when the minimum falls at the edge of the $[-\pi/4, \pi/4]$ window the fit extrapolates, which is our leading suspect for its sign reversal against the dense-profile result of App. C.

The dilution confound, in numbers. Every production landscape rotates a group of exactly four atoms. The group's share of the reconstruction on its own active samples, $E/(E + \text{MSE})$, falls monotonically along every L_0 sweep: layer 12 / 16k gives $0.041, 0.027, 0.016, 0.011, 0.010$ over $L_0 = 23 \ldots 475$, and the other five sweeps behave the same way (Spearman -1.000 in all six). Since the relative depth $(\max - \min)/\min$ is a ratio of a perturbation of that group's contribution to the total residual, it inherits the decline whether or not the angular argument holds. Renormalizing the profile range by E instead of by $\min$ gives a statistic that is nearly constant in L_0: the median $(\max - \min)/E$ sits between 0.58 and 0.72 at every L_0 level of every sweep, with sweep-level Spearman $+0.10, -0.40, +0.30, -0.50, -0.70, -0.80$ (mean -0.33). We cannot separate the two accounts with a fixed group size, and we say so here rather than reporting only the favorable normalization. The experiment that separates them is a group that grows with the manifold, for instance every atom within a decoder-cosine radius of the seed, so that the energy-normalized statistic is the one the theory speaks about; App. F reports that adaptive-group version, which shows the decline is dilution but, with m and L_0 moving together on the released grid, cannot test m on its own.

Width-law confound. Raw comparison across widths at matched L_0: median depth $0.0100 \rightarrow 0.0114$ (16k$\rightarrow$262k), Spearman $+0.900$, i.e. no flattening. Restricted to features with firing rate in $[10^{-3}, 10^{-2}]$: $0.00924, 0.00865, 0.00777, 0.00783, 0.00703$, i.e. $0.0092 \rightarrow 0.0070$ with one inver-

			all sampled features		firing rate $\in [10^{-3}, 10^{-2}]$		partial
layer	width	L_0 range (n)	median depth	Spearman	median depth	Spearman	$\rho_s(\text{depth}, L_0 \mid \log f)$
5	16k	20–343 (5)	$0.0505 \to 0.0136$	-1.000	$0.0325 \to 0.0174$	-1.000	-0.153
5	65k	18–229 (5)	$0.0582 \to 0.0179$	-1.000	$0.0313 \to 0.0113$	-0.900	-0.198
12	16k	23–475 (5)	$0.0274 \to 0.0082^{\dagger}$	-0.900	$0.0225 \to 0.0076$	-0.900	-0.295
12	65k	22–143 (3)	$0.0284 \to 0.0075$	-1.000	$0.0214 \to 0.0056$	-1.000	-0.369
20	16k	23–304 (5)	$0.0381 \to 0.0070$	-1.000	$0.0313 \to 0.0084$	-1.000	-0.471
20	65k	21–235 (5)	$0.0403 \to 0.0077$	-1.000	$0.0408 \to 0.0065$	-1.000	-0.473
			mean	$\mathbf{-0.983}$		$\mathbf{-0.967}$	$\mathbf{-0.327}$

Table 4: The L_0 law on production SAEs: six sweeps, all decreasing end to end, five of six strictly monotone (four of six in the firing-rate band), still decreasing when the firing-rate confound is controlled two ways, and not robust to the normalization: per unit of the group's own reconstruction energy the trend weakens (sweep-level Spearman mean -0.33, preceding paragraph) and with an adaptive group it reverses ($+1.000$ in both sweeps, Table 6). Depth columns give the sparsest and densest SAE of each sweep. The last column is the feature-level partial Spearman of landscape depth on the SAE's L_0 controlling for $\log$ firing rate ($n = 5{,}978$–$9{,}999$ features per sweep; all $p < 10^{-30}$ with features as units, descriptive; the unit-level test follows). Sweep-level Spearmans are over 3–5 L_0 levels; exact permutation p-values, two-sided, are 0.017 ($\rho = -1$, 5 levels), 0.083 ($\rho = -0.9$, 5 levels) and 0.333 ($\rho = -1$, 3 levels), and the two-sided sign test over the six sweeps gives $p = 0.031$. These are descriptive rather than inferential, since layers and widths of one base model are not independent replicates. The partial correlations are tested at the level of the independent unit (the SAE) by permuting L_0 labels within each sweep: $p = 1/120$ for four sweeps, 0.025 for layer 12 / 16k, and $1/6$ (the floor at three levels) for layer 12 / 65k. †The one non-monotone sweep: depth falls to 0.0067 at $L_0 = 184$ and rises to 0.0082 at $L_0 = 475$, the densest SAE GemmaScope ships at this width.

sion (131k above 65k by 6×10^{-5}), Spearman -0.900. With five widths the exact permutation p is 0.083 two-sided; the large-sample approximation reported by standard software ($p = 0.037$) is not valid at $n = 5$, and at sample sizes this small we quote the exact value. Wider SAEs allocate more features to rare concepts and rare features have deeper landscapes, so the population comparison mixes a composition shift with the within-feature effect. The same confound applies to the L_0 law, since at fixed width the mean firing rate is exactly L_0/m; Table 4 therefore reports the L_0 law both within the matched firing-rate band and as a feature-level partial correlation controlling for $\log$ firing rate, and the law survives both. The two controls differ in what they assume: the band comparison is nonparametric but discards 70–90% of the features, while the partial correlation uses all of them at the cost of a monotone-in-rank model of the firing-rate effect, so we report both rather than choosing.

The OMP comparison, in full. Freeze a dictionary and let orthogonal matching pursuit (Pati et al., 1993) re-solve the coding problem at the encoder's own sparsity: OMP reconstructs *better* than the amortized encoder ($0.90\times$ its error at 16k, $0.82\times$ at 262k) while keeping only 0.54 to 0.38 of the atoms it selected. That is the classical amortization gap (Gregor & LeCun, 2010; O'Neill et al., 2024); the increment is the *support-level* dissociation and its width trend (Table 5).

The amortization gap, in detail. Part of the amortization gap is coefficients, not support—a JumpReLU encoder is not least-squares optimal even on the atoms it chose, so we ran the control: refitting least squares on the *encoder's own* support recovers 4–7% of its error, and OMP still wins after it ($0.94\times \to 0.89\times$) and at matched support size ($0.97\times \to 0.93\times$), both broadly widening with width. The encoder's selection of atoms is therefore not privileged even for a fixed dictionary. Behind that: the group's reconstruction share falls from 0.04–0.08 to 0.01–0.03 across the L_0 sweep. The number of neighbors within cosine 0.25 of a feature falls from 21.7 to 0.9 (layer 12/16k) and $46.8 \to 1.8$ (layer 20/65k), and falls monotonically at cosine 0.15 and 0.35 too (Table 6 shows 0.25), and since a cosine threshold cannot see a wide-angle tiling at all (a regular m-gon with $m \leq 4$ has every pairwise cosine ≤ 0), we ran the rule that can: co-activation counts fall too, end to end, $0.52 \to 0.34$ and $0.82 \to 0.43$ (the second after a rise to 0.90). On curvature: in the controlled setting it is the *better* predictor (-0.774 vs -0.755), so we read the production reversal as a property of a 9-point profile whose minimum often sits at the window edge, without ruling out a genuine failure at scale.

F MANIFOLD-TRACKING GROUPS: THE CORRECTED PRODUCTION MEASUREMENT

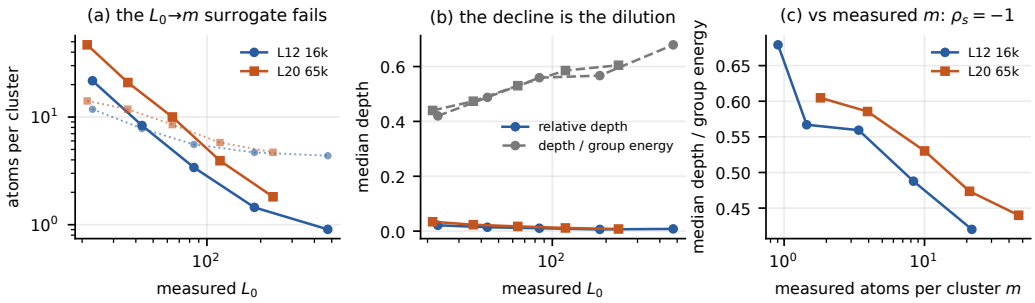

Figure 5: **The corrected measurement on 14 GemmaScope SAEs.** (a) The number of other atoms within decoder cosine 0.25 of a feature (solid) and the size of the adaptive rotated group (dotted; floor of four atoms) both *fall* as L_0 rises: L_0 is not a surrogate for atoms-per-manifold. (b) With the adaptive group, relative depth still falls with L_0 (colored) but depth per unit of the group's own reconstruction energy *rises* (gray): the decline is the dilution. (c) Against the measured atom count, the energy-normalized depth falls monotonically (Spearman -1.000 in both sweeps), the direction Proposition 6 predicts, though forced here, both quantities being monotone in L_0 (Table 6). The abscissa counts neighbors, the seed excluded. Three of the ten points have fewer than two on average, where the adaptive group sits at its floor of four atoms and coincides with the fixed group; restricted to the other seven the ordering is unchanged (-1.000 in each sweep, three and four points), which is a weaker statement than the panel suggests at a glance.

Same SAEs, hook and activation pool as App. E; 1,000 sampled live features per SAE; two L_0 sweeps (layer 12 / 16k and layer 20 / 65k, five levels each) plus the five layer-12 widths at matched L_0, whose 16k member (tag 82) is also in the first sweep: 14 SAEs. Each feature is measured twice: with the fixed four-atom group of App. E and with an adaptive group holding the seed and every atom within decoder cosine 0.25 of it, with at least 4 and at most 16 atoms (below three qualifying neighbors the floor makes it the fixed group). We store the profile range, the group's reconstruction energy and the residual minimum, so every normalization can be recomputed from the released records.

Neighbor counts below two. The mean number of neighbors within cosine 0.25, the seed excluded, falls under 2 at the densest SAEs (1.4 and 0.9 at layer 12/16k, 1.8 at layer 20/65k), so a cluster there holds two to three atoms, barely enough to tile a plane, and the adaptive group sits at its four-atom floor: at those SAEs it is the fixed group, and the reversal of the L_0 trend there is not a property of adaptive grouping. We keep them in the table because they are the measurement that falsifies the L_0 surrogate, and exclude them from any claim about the ordering.

What this does and does not license. The reordering in Fig. 5c uses the same ten SAEs (five per sweep) as Fig. 5b with a different abscissa; since cluster size and L_0 are perfectly anti-correlated here, the two panels contain the same ten points and the sign flip is a relabeling, not a second experiment. We report it because the relabeling is the theoretically correct one (m is the theory's variable and we measured it), but a real test needs a design in which m and L_0 are varied independently, which the released GemmaScope grid does not offer. Within a single SAE, where m varies across features at fixed L_0, the relation does not appear: restricting to features whose rotated group has the modal size and controlling for the group's reconstruction share and log firing rate, the partial Spearman of cluster size against energy-normalized depth ranges over $-0.109 \ldots + 0.139$ with inconsistent signs across the 14 SAEs (mean $+0.045$). We take the feature-level null seriously: it is equally consistent with the theory being wrong at this scale and with decoder cosine being a poor estimate of manifold membership for a single feature.

G CROSS-SEED PREDICTION: A NEGATIVE RESULT AT TWO SCALES

The landscape estimator does *not* beat decoder-geometry baselines for predicting *cross-seed* replication, and we report this at two scales. At $1.5\times$ overcompleteness (768 ReLU+L_1 latents, $\lambda = 0.15$, on $80{,}000$ centered, unit-mean-norm layer-3 residuals of Pythia-70M-deduped from the Pile slice of App. V; Adam at 3×10^{-3}, cosine schedule, batch $1{,}024$, 25 epochs; 8 replicate seeds, shared-latent fraction 0.32): best landscape variant AUC 0.62, against 0.82 for a decoder-geometry baseline (mean $|\cos|$ to the other atoms), 0.81 for logistic regression on the raw decoder vector and 0.72 for firing frequency. At $21\times$ overcompleteness (16,384 TopK latents, k=32, on Pythia-160M layer-6 residuals; 8 replicate seeds we trained on 50M MiniPile tokens in identical order, following the protocol of Paulo & Belrose (2026)—a suite distinct from their released 32k MLP-output dictionaries of §6; mean cross-seed max-cosine 0.715, mutual-NN shared fraction 0.562; 1,500 features audited): landscape depth correlates with cross-seed max-cosine at $+0.21$ (AUC 0.61) against $+0.56$ (AUC 0.78) for the decoder nearest-neighbor baseline.

The two shared-fraction numbers (0.32 and 0.562) are not comparable to each other or to the width trend of Paulo & Belrose (2026): they come from different base models (Pythia-70M vs 160M), different architectures (ReLU+L_1 vs TopK), different token budgets, and two different definitions (max-cosine ≥ 0.7 vs mutual nearest neighbor at ≥ 0.7). We report them to characterize each setting, not as a width comparison.

On this axis we can also ask the partial question we ask in Table 5: controlling for the decoder nearest-neighbor cosine, landscape depth correlates with cross-seed max-cosine at only $+0.05$, while the baseline controlling for depth stays at $+0.53$. The estimator adds essentially nothing to decoder geometry here.

Contrast this with Table 5, where the marginal ranking is *reversed* on the inference-time axis at every width, though, as that table's last rows show, depth does not survive controlling for the group's reconstruction share. The two ground truths are only moderately correlated where we measured both (Spearman $+0.50$ at $1.5\times$, between each latent's OMP retention on the first seed's frozen dictionary—OMP at its median L_0 of 24, retention as in Table 5—and its mean cross-seed max-cosine; `code/e5_inference_identifiability.py`), and they evidently reward different signals: cross-seed disagreement at these training scales is dominated by effects that local decoder geometry captures (near-duplicate atoms and cluster membership) on top of optimization noise, whereas the interchangeability of atoms under a fixed dictionary is closer to what the theory describes, the flatness of the objective. "SAE feature instability" is therefore at least two distinct phenomena, and papers (including ours) should specify which axis a claim concerns. Once reconstruction share is controlled, our estimator has no positive claim on either axis.

H THE CYCLIC-FEATURE RING: SETUP, FULL TABLES, AND THE READ-OUTS THAT FAILED

Locating the ring. For each of p items (7 weekdays, 12 months) we build prompts from 38 (30) templates crossed with 12 sentence prefixes, and take the residual activation at the last token of the item word: $3{,}192$ probes for days, $4{,}320$ for months. That is the probe set for the density conditions and trained read-outs below; the ring's *geometry* we report on the set the rest of the paper uses (20 templates $\times$ 12 prefixes, Table 9), scanned over the embedding output and all 26 block outputs by `code/g45_ring_geometry.py`. The p item means lie in cyclic order in the plane of their top two principal components at all 27 of those hook points for weekdays and at 25 of 27 for months—it fails only at the last two blocks—where a random arrangement would pass with probability $2p/p! \approx 3 \times 10^{-3}$ ($p = 7$); restricted to the block outputs these are the $26/26$ and $24/26$ of Table 9, which the scan reproduces exactly. That plane carries 35–66% of the between-item variance, and the in-plane between-item radius is $7.6\times$ (days) and $10.3\times$ (months) the mean within-item scatter at layer 7, falling from $30\times$ and $67\times$ at the first block to $1.5\times$ and $1.2\times$ at the last; we use layers 7, 12 and 20, where both rings pass (§5.1 trains at layers 1, 3, 7 and 12, and §6 ablates at 5, 12 and 20).

The objective-level test. Write θ for the angle of a probe activation inside that plane. For a condition we choose weights w over the p items, form the weighted empirical angular density $\hat{\rho}$ on a 720-point grid, and convolve it with the exact cost kernel $\bar{c}$ of a regular m-gon computed by the solver

of App. B at $\lambda = 0.15$. The reported depth is $(\max J - \min J)/\overline{J}$. Two families of conditions (Table 7, Fig. 6): *(A)* each item cluster is translated in the plane so its mean sits on a regular p-gon at the ring's common mean radius (keeping the first item's angle), then moved back by ε of that translation (so $\varepsilon = 0$ is the regular ring and $\varepsilon = 1$ the model's own), preserving the model's own cyclic direction and all within-cluster structure; *(B)* at $\varepsilon = 0$, weights $w \propto 1 + 2r\cos(q\theta^{\text{ideal}})$ with $r = 0.35$, which for equally spaced angles sets $R_q = r$ and leaves every other harmonic below p at zero (its alias $p - q$ is the complex conjugate and is set too). One cell class is special: at $q = p/2$ the alias coincides with the tilt, so the installed moment doubles to $2r$ (measured $R_6 = 0.59$ on the month ring against 0.31–0.35 elsewhere). Those 15 cells receive a larger dose than the rest of the sweep; excluding them, the month-ring ordering is Spearman -0.73 against -0.78 with them, so the graded law does not rest on the doubled-dose cells.

Robustness of the selectivity result. Three controls, all in `code/g10_radius_control.py` and `results/gpu/g10_radius_control.json.gz`. *(i) The radial reduction.* App. A shows the reduction to a single kernel is exact only when the kernel shape is radius-free; here it is not (λ/r varies). At the symmetric point the construction has already placed every cluster at the ring's common mean radius, so the residual spread is within-cluster scatter (CV 0.075–0.167), and recomputing every cell as the radius mixture $\sum_b w_b r_b^2 (\rho_b \star c_1^{(\lambda/r_b)})$ over 12 radius bins leaves the ordering intact: Spearman(κ, depth ratio) moves from -0.899 to -0.795 (days) and from -0.744 to -0.768 (months); the cell-by-cell depth *ratios* of the two arms agree at Spearman 0.84 (days) and 0.75 (months); and the 51 flat month cells stay flat ($1.06\times \to 1.04\times$). The before-column here is this script's own single-kernel arm, not Fig. 6's: the two pipelines share the density and the solver but bin and normalize differently, so the same statistic reads $-0.899/-0.744$ here against $-0.895/-0.783$ there. The comparison that matters is within this table, between arms computed the same way; the day discrepancy is 0.004 and the month one 0.039, and on months the mixture moves the correlation the *other* way once measured against Fig. 6's baseline. We give both rather than pick. We note for completeness that the reduction is *not* harmless if the between-item radius spread is retained. Rotating each cluster to the ideal angle at its own radius, a construction we do not use, drops the day ordering to -0.13, which is what makes the control worth running rather than assuming. *(ii) The penalty.* $\lambda = 0.15$ is arbitrary; sweeping it gives Spearman $-0.825, -0.874, -0.895, -0.901, -0.884$ (days) and $-0.822, -0.832, -0.783, -0.740, -0.789$ (months) at $\lambda = 0.05, 0.10, 0.15, 0.25, 0.40$, so nothing here turns on the choice. *(iii) What the density is, and is not.* The angular law is the model's; the *item weighting* is ours, uniform at $q = 0$ and tilted by construction otherwise, and natural text does not use weekdays and months uniformly. A uniform-over-items weighting is itself a symmetry assumption, of the same family as the one we fault in §2, the difference being that ours is stated and swept (*q is* the sweep) rather than hidden in a hypothesis. What the experiment tests is the criterion's arithmetic on a real model's angular geometry, not the frequency with which that geometry is visited.

What we could not measure: a trained read-out on this manifold. We also tried the obvious experiment, training replicate SAEs on these activations under each condition and measuring cross-seed phase jitter, in the full residual stream, the top-16 principal subspace and the ring plane itself, 24–32 seeds per cell. We report it as a negative result with a diagnosis rather than as evidence: in the full stream the apparent effect on $m = 2$ is collinear with how much of each atom lies in the plane (Spearman -0.92 between in-plane fraction and jitter across conditions), which is a property of the read-out, not of flatness; inside the plane the dictionaries are far from a regular tiling of the circle (mean angular-gap coefficient of variation 0.36–0.55, where a regular m-gon gives 0), so there is no rotational configuration whose phase could be pinned, and every condition returns jitter ≈ 1. The reason is structural: a p-point cyclic feature coded by $m < p$ atoms does not present the optimizer with a rotation but with a choice of *which* m of the p clusters to claim. Seeds then differ combinatorially, and the phase at the dictionary's own order $k = m$—the only order this read-out examined—is the wrong summary: the configurations are not regular m-gons, so their m-th harmonic carries no lock.

The same runs, read harmonic by harmonic. The stored per-seed atom angles let us recompute the statistic of §5.1 on these runs without retraining anything (`code/g26_fullstream_harmonics.py`). Each atom enters weighted by its squared in-plane norm fraction, the weight the cost kernel gives it, and the null of $\max_{k<p} A_k$ is simulated per cell with that cell's own weights. Over the twelve (ring, layer, m) cells—days at layers 7 and 12 and months at layer 7, $m = 2$–5; the in-plane runs cover the day ring only, eight cells—the

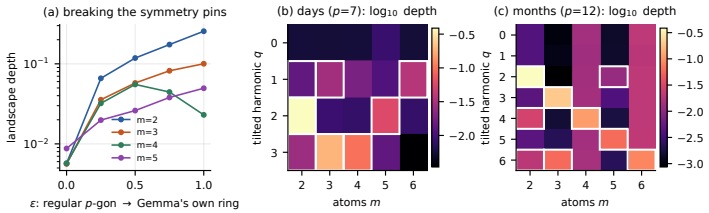

Figure 6: **The criterion, evaluated exactly on Gemma-2-2B's own cyclic features.** The manifold is the model's representation of the days of the week ($p = 7$) and of the months ($p = 12$). Nothing is trained: the angular density is the model's own, measured inside the plane; the cost kernel is the exact 2-plane solver of App. B; $J(\varphi)$ is their circular cross-correlation. (a) Symmetrizing the ring onto a regular p-gon flattens the landscape and interpolating back deepens it, by $45\times$ for the layer-7 m=2 day curve (the top one of the four plotted, m = 2–5) and a median $7.4\times$ over all 30 curves. (b,c) At the symmetric point we tilt exactly one harmonic q. $\kappa = \min\{jm : jm \equiv \pm q \bmod p\}$ is the first order the configuration can feel; white boxes mark cells where m divides q or $p - q$, so that the configuration feels the installed harmonic or its alias directly. Depth falls off with κ (Spearman -0.90 days, -0.78 months). Table 7 gives layer 7 cell by cell; the full grid is in the released records.

natural ring is above the symmetrized one in 12 of 12 in the full residual stream (median 0.871 against 0.598), 11 of 12 in the top-16 principal subspace (0.674 against 0.366) and 6 of 8 in the plane (0.481 against 0.313). What separates the three spaces is the symmetrized side. In the full stream it is not at chance—10 of 12 symmetrized cells still exceed their own null, and not through the first harmonic (0.573 with $k = 1$ excluded)—because these dictionaries hold two to five atoms for the whole residual stream and put a median of only 0.33 of an atom's norm in the plane, so the in-plane angles inherit the orientation of shared off-plane structure. Removing that contamination removes the effect: 5 of 12 symmetrized cells exceed their null in the 16-dimensional subspace and 1 of 8 in the plane. The gap between natural and symmetrized grows with how much of the atoms lies in the plane (Spearman $+0.48$ over the twelve cells). Selectivity, on the other hand, does not appear at all: over the 48 (cell, q) tilts the agreement at the installed order beats the other orders in 23 and is the strongest order in 13, both at chance. That is the coding rule, not the read-out—these dictionaries are TopK with $k = 1$, which §4.3 finds does not inherit the criterion at an installed low harmonic—and it is why §5.1 trains ReLU+L_1. Proposition 1 still constrains this regime: the symmetrized density is (approximately) $\mathbb{Z}_p$-invariant, so the solution set is a union of discrete $\mathbb{Z}_p$-orbits and the "which clusters" indeterminacy is indeterminacy *along* that orbit; what is lost below p is the continuous criterion, not the orbit structure. The continuous picture our synthetic sweep validates (App. C) requires an angular density that is spread rather than concentrated on p atoms, which is exactly the regime the objective-level test above isolates.

Further detail. On the 12-point ring the 51 cells with no feelable order at all sit at $1.05\times$ their untilted baseline, and outside the disjointness regime of Proposition 4 the closed form degrades to 6.6% median relative error at λ a tenth below the threshold and 23% at a quarter below (App. B).

Detail from the interpolation curves. Twenty-nine of the 30 curves end higher than they start and 21 rise monotonically; of the nine that turn over, four do so by more than 10%, concentrated at $m \in \{4, 5, 6\}$ on the day ring, whose 4th harmonic is among the weakest it carries (Table 8)—which fits the m=4 cases and not the others. The single curve that ends lower does so by 0.7%. The $m = 2$ layer-7 endpoints are $5.7 \times 10^{-3} \to 2.6 \times 10^{-1}$ (days) and $4.3 \times 10^{-3} \to 7.5 \times 10^{-2}$ (months).

The objective-level grid, in detail. The month grid's flat side has IQR 0.99–1.30; the $4.85\times$ tail is the m=3 row, whose baseline is the grid's smallest. On the day ring depth falls from $58\times$ baseline at κ=2 through $7.3\times$ at κ=4 to $0.84\times$ at κ=9, with a median $0.97\times$ over all cells with $\kappa \geq 9$. The truncated harmonic series of Eq. 2, $4\sum_{j\geq 1} R_{jm}|\hat{c}_{jm}|$ (Corollary 1's range bound), reproduces the solver's depth to a median ratio of 1.09 over the 90 tilted month cells (1.14 over the 45 day cells), rank-ordering them at Spearman $+0.993$ ($+0.997$). On contingency: a model whose ring density were rotation-invariant, as the isotropy assumptions of prior analyses would have it, would show *no* deepening, so the measured deepening (a median $7.4\times$, up to $45\times$) is a real-model case where an isotropy-based account predicts flat and the criterion predicts pinned; the model's harmonic content

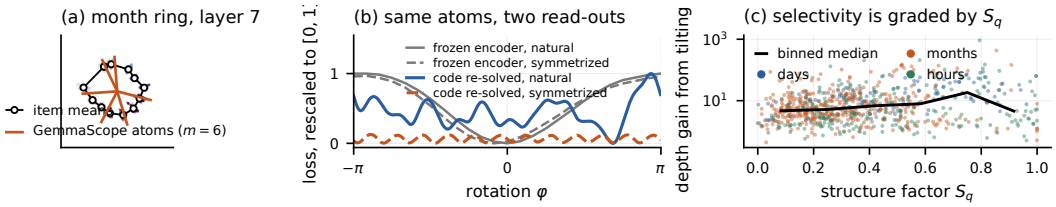

Figure 7: **The criterion evaluated on GemmaScope's own configuration.** (a) Gemma-2-2B's month ring in its top-2 principal plane at layer 7, with the six width-16k GemmaScope atoms lying in that plane ($\tau = 0.5$). The plane is located from probes alone; the atoms are whatever the released SAE learned. (b) Those atoms rotated rigidly in that plane, read two ways, each pair rescaled by the natural curve's range. The SAE's own loss (gray) is a first-harmonic bowl that barely separates the real ring from a symmetrized one; re-solving the code at every rotation gives an irregular profile (blue) that symmetrizing flattens to a small 12-fold ripple (orange), a $7.9\times$ depth ratio here. (c) Over all 149 distinct configurations (no arc filter) and 762 (cell, q) pairs, the depth gain from tilting harmonic q rises with the structure factor S_q; partial Spearman $+0.43$ given q and cell.

(Table 8) is likewise measured, not chosen. The selectivity grid, by contrast, computes the criterion on densities we construct, and the series comparison is a tightness check, not a test the model could fail: both sides come from the same density and kernel, so their agreement certifies that the leading harmonics carry the landscape, not that Gemma has a particular property.

I THE CRITERION ON GEMMASCOPE'S OWN CONFIGURATIONS

What §5.1 summarizes. Evaluated with no training on GemmaScope's in-plane configurations (ReLU+L_1 code re-solved at every rotation, $\lambda = 0.15$), the natural ring is deeper than the symmetrized one in 155 of 155 cases (median $23.0\times$), but random angles at the same m give $22.7\times$ ($p = 0.55$ against the trained ones) and item sets carrying no ring deepen too, so this confirms the identity rather than tests it. Three bounds follow below: the structure factor orders selectivity only weakly ($+0.43$), GemmaScope's configuration sits only at the 41st percentile of its landscape (the ℓ_1 landscape we compute, not GemmaScope's own loss, which barely separates the two rings, $1.06\times$), and its atoms are only weakly locked to the item angles—what carries over is the identity, not the optimum.

Item sets and plane location. Probes are 20 sentence templates $\times$ 12 prefixes $\times$ p items, read at the last token of the item, so every item weighs the same; weighting each item by its count in a FineWeb shard instead (2,186–3,070 per weekday, 2,363–4,618 per month) leaves the strongest harmonic below p unchanged in 8 of 8 weekday and month cells, including the four where it leads the runner-up by under 0.06, and moves no $|\hat{\rho}_k|$ by more than 0.10 (`code/g71_frequency_weights.py`; clock hours, 1 to 13 per hour in the shard, are too rare to weight); the ring plane is the top-2 principal plane of the p item means, which involves no SAE. Table 9 lists what we tried. Weekdays, months and clock hours put their item means in cyclic order in that plane at essentially every layer and are the three used in §5.1; the unordered control (`animals`) fails the order test everywhere, as it must. `Seasons` ($p = 4$), `compass` and `zodiac` fail it too—with $p = 4$ a pass would be weak evidence (chance $1/3$) and seasons fail at all three layers, and the zodiac plane carries little variance ($\mathrm{ev}_2 = 0.26$)—so we do not use them; `numbers` passes, consistent with Engels et al. (2025) reporting circular structure for numeric quantities. We also report the largest angular gap between item means in units of the uniform gap $2\pi/p$, since an *arc* passes a cyclic-order test (the angular order along an arc is the linear order) while a ring does not leave one side of the circle empty.

In-plane atom counts. An atom counts as in-plane when $\|P_V w_j\|/\|w_j\| \geq \tau$. Over the 26 layers at width 16k, the median count is 6 at $\tau = 0.3$ (range 0–18) and 2 at $\tau = 0.5$ (range 0–8); 70 of 78 cells have at least two atoms at $\tau = 0.3$. On the six (ring, layer) cells where both widths were measured, the count rises with width at every one: $3 \to 4$, $2 \to 3$, $0 \to 3$ (weekdays, layers 7/12/20) and $6 \to 10$, $4 \to 7$, $5 \to 9$ (months), $20 \to 36$ in total at $\tau = 0.5$.

Objective-level landscape. The density is the empirical angular density of the probes in the plane; the cost kernel is the exact 2-plane solver of App. B evaluated at the *measured* atom angles; J is

their circular convolution, so the code is re-solved at every rotation by construction. Symmetrization translates each item cluster so its mean sits on a regular p-gon at the common mean radius, as in §5. Three thresholds $\tau \in \{0.3, 0.4, 0.5\}$ over 70 (ring, layer) cells give 166 configurations with $m \geq 2$, of which 149 are distinct atom sets (two thresholds often select the same atoms; the structure-factor analysis of Fig. 7c merges those, every other count below keeps them as separate SAE settings). Restricting to cells whose plane passes the cyclic-order test *and* has a maximum item gap below 2.5 uniform gaps—the arc filter—leaves 155 of the 166 (m from 2 to 18, median 4; angular-gap coefficient of variation median 0.62, nowhere near regular). **Every number of §5's GemmaScope paragraph and of this appendix is computed on those 155 configurations at $\lambda = 0.15$ unless a different set or λ is named** (the 41st percentile, for one, uses the 70 cells at $\tau = 0.3$). GemmaScope puts more atoms on the ring with width, in 6 of 6 cells ($20 \to 36$ atoms from 16k to 65k). Per ring the median natural-to-symmetrized depth ratios are 34.9 / 18.7 / 25.3 (weekdays / months / hours), reported descriptively since layers of one model are not independent; for the same reason every p-value in this appendix, which treats configurations or cells as units, is descriptive. The natural-to-symmetrized depth ratio exceeds 1 in $155/155$ (so in all 138 distinct atom sets among them; random angles at the same m exceed 1 in all 155 as well) at every λ we tried, with medians 25.7, 23.0, 22.6, 16.2 and 11.0 at $\lambda = 0.05, 0.15, 0.25, 0.40, 0.60$.

Controls on that comparison. Three, run on the same cells with the same solver (`code/g19_controls.py`). *(i) Configuration.* At $\lambda = 0.15$, replacing the trained angles by the regular m-gon of the same m gives a median $8.1\times$; replacing them by *random* angles at the same m gives $22.7\times$, and by random angles matched to the trained configuration's angular-gap dispersion, $22.7\times$. Paired against the trained configuration these are $2.08\times$ (Wilcoxon $p = 3 \times 10^{-13}$ over the 155 configurations, which share cells and atom sets, so descriptive) and $1.08\times$ ($p = 0.55$). So the trained-versus-regular gap is real and runs in the direction Proposition 4's structure-factor heuristic suggests, and the trained-versus-random gap is absent: the experiment establishes that the criterion behaves as stated on a configuration nobody chose, not that GemmaScope's configuration is special. *(ii) Negative control on the manifold.* The same computation on the item sets that fail the cyclic-order test gives a median $14.3\times$ over 50 configurations (four item sets at three layers, two widths and three thresholds, where $m \geq 2$), 48 of them above 1: seasons 7.0, zodiac 12.0, animals 15.8, compass 18.1. Symmetrization removes angular structure wherever there is any, and these densities are not p-fold symmetric, so the criterion predicts them pinned too; what the control rules out is reading the ring result as evidence that *rings* are special. *(iii) Sparsity.* On the same 155 configurations the trained-to-regular ratio of medians falls monotonically with λ: 4.53, 2.83, 1.64, 1.48, 1.16 at $\lambda = 0.05, 0.15, 0.25, 0.40, 0.60$. Proposition 4 is exact in the disjoint-support regime $\lambda/r \geq \cos(\delta/2)$, the large-$\lambda$ end, and the sweep approaches that regime without reaching it: the trained configurations have minimum gaps δ of a median 27° (IQR 11°–62°; median $m = 4$, range 2–18), so the threshold is a median 0.973, and 0, 1, 1, 1 and 5 of the 155 configurations satisfy it at the five λ (those with $m \leq 3$). Against the exact kernel the closed form of the proposition is off by a median 129%, 94%, 79%, 62% and 42% at those λ (`code/e8b_prop3_on_trained.py`). So on trained dictionaries the proposition is the structural statement—an irregular configuration couples to every harmonic—and S_q is a heuristic ordering rather than a computable kernel; the structure-factor test of Fig. 7c is a test of that heuristic at $\lambda = 0.15$, where 154 of 155 configurations are outside the regime, and its $+0.43$ should be read as such. Restricting that test to the regime is not possible at any λ we ran (≤ 5 cells). The identity itself (Proposition 3) uses the exact kernel throughout and does not depend on the proposition: the product-spectrum check at the end of this appendix is computed with it. We report the shrinking margin as a limitation of the effect, not as support for it.

Frozen-encoder landscapes. For the same cells we also rotated the atom group and read the SAE's own loss, as a production audit naturally would. Over the 41 of the 78 (ring, layer) cells at width 16k whose plane holds at least two atoms at $\tau = 0.5$, the landscape's power is at $k = 1$ (median 0.992, minimum 0.858) and essentially absent at $k \in p\mathbb{Z}$ (maximum 4×10^{-5}), and the natural and symmetrized rings differ by a median $1.06\times$ (range 0.99–1.26). The same holds on the 14 measurable cells of the other five item sets (median 98.0% at $k = 1$, maximum 8×10^{-4} at $k \in p\mathbb{Z}$), four of which fail the cyclic-order test—seasons, compass points and the unordered animal control are not rings at all, and the zodiac's plane carries too little variance to order. That is the point: the $k = 1$ bowl does not depend on whether the plane holds a ring. Whether it comes from the encoder or from the residual's in-plane offset is settled by the re-solved comparison below: exact re-solving on the uncentered residual keeps a 0.919 share at $k = 1$, so the offset accounts for most of it and the amortization gap for the rest.

Negative controls. Whether GemmaScope's configuration sits at the minimum of the ℓ_1 landscape we compute for it is a separate question from whether that landscape is pinned, and the answer is no: over the 70 cells at $\tau = 0.3$ the as-trained rotation sits at a median 41st percentile of its own landscape (chance 50; Wilcoxon $p = 0.02$), and on the day ring at the 63rd, the wrong side. Likewise the configuration's phase at harmonic p relative to the item lattice is only weakly non-uniform ($R = 0.26$ over 70 cells, Rayleigh $p = 0.009$) and not significantly so at $\tau = 0.4$ or 0.5. GemmaScope optimizes a different objective at a different sparsity on a different corpus, and its atoms serve many manifolds at once; we would not expect it to sit at our minimum, and it does not.

Re-solving on the residuals themselves. The objective-level result above discards radii: it convolves the measured angular density with the exact kernel, which is the setting the theory is stated in. As a check that nothing depends on that, we also re-solved the non-negative code exactly on the residuals themselves, at $\tau = 0.4$ over the 11 measurable (ring, layer, width) cells, rotating GemmaScope's own atoms inside the plane and enumerating the at-most-two-atom optimal supports in closed form (the enumeration is exact in two dimensions; it agrees with L-BFGS-B restarts to 9×10^{-16}). Read raw, the landscape is still a first-harmonic bowl (median $k = 1$ share 0.919 at $\lambda = 0.15$, natural/symmetrized $1.11\times$), because the in-plane projection of the residual carries a large offset and a radial spread far wider than the ring. Centering the plane on the manifold removes it in six of the eleven cells and not in the other five: the $k = 1$ share is ≤ 0.001 in six (median over all eleven 0.001, a median that hides the split) and 0.82–0.96 in five, which are exactly the five whose residual in-plane offset exceeds half the ring radius (0.51–0.87 against 0.06–0.40 for the other six). Where the bowl goes, power moves to $k \in p\mathbb{Z}$ (median 0.021, up to 0.173 on the 19-atom month group at layer 7; 0.000 in the five bowl cells), and the natural ring is deeper than the symmetrized one in 11 of 11 cells either way (median $2.10\times$; $1.64\times$ at $\lambda = 0.05$, $2.13\times$ at $\lambda = 0.30$). The depths are far smaller than the angle-only computation's because the ring is a modest modulation on top of that radial spread; the sign and the harmonic content are what the criterion predicts.

A numerical consistency check on the identity. `code/g20_falsifiable.py` recomputes, for each of the 155 configurations, the product spectrum $\sum_k |\hat{\rho}_k||\hat{\tilde{c}}_k|$ of that cell's own density and its own configuration, and correlates it with the measured landscape depth: Spearman $+0.975$ over the 155 configurations, $+0.961$ controlling for m and ring. This is a check that the numerics obey Proposition 3, not a test of it—the triangle inequality already bounds depth by twice that sum. The one thing it does show is that the density alone is not enough: the density's total non-uniformity correlates with depth at only $+0.116$ (two-sided $p = 0.15$), so which cells are pinned cannot be read off the ring without the atoms.

J PLACEMENT AGAINST CLASSICAL IDENTIFIABILITY RESULTS

The receipts, in full. For the sharpest of the three concurrent preprints, Dalili & Mahdavi (2026), the within-manifold law is the uniform slice $h = t\,u$, u uniform on the unit sphere of V_i (their Eq. 4). That law forces $\Sigma_i = (\mathbb{E}t^2/d_i)P_{V_i}$, a d_i-fold repeated eigenvalue (their §A.3.1), so any basis within the eigenspace their Theorem 11 recovers is free to rotate, exactly because the eigenvalue repeats. Uniformity is sufficient for the repeat, not necessary: any law on the slice with isotropic second moment repeats it (in a 2-plane, any ρ with $\hat{\rho}_2 = 0$), and in a 2-plane any anisotropic one has distinct eigenvalues (in higher dimension a repeat needs a symmetry of the second moment, not of the law)—the harmonic reading of Proposition 3 with a kernel supported at $k = \pm 2$. Their Theorem 8, that the true basis is not an ℓ_1 minimizer, is powered by the same assumption: its constant $2^{-(d_i+1)}$ is half the uniform measure 2^{-d_i} of the positive orthant. Neither result uses the dimension except through the symmetry it makes room for. Our Proposition 3 is the statement of what, in that room, decides the matter.

Against the classical results. NMF uniqueness rests on conic-hull conditions such as sufficient scattering (Donoho & Stodden, 2003; Fu et al., 2018), which are blind to how mass is distributed inside the cone: two laws with identical support can sit on opposite sides of our criterion. Classical dictionary-learning identifiability—complete (Spielman et al., 2012; Sun et al., 2017; Gribonval & Schnass, 2010; Wu & Yu, 2018) and overcomplete (Arora et al., 2014; Agarwal et al., 2014; Hillar & Sommer, 2015)—obtains uniqueness up to permutation and sign under sparse-support and incoherence conditions; those conditions implicitly break rotational symmetry, and we conjecture that breaking it is the load-bearing ingredient; we do not re-derive their theorems. Roeder et al.

(2021) prove *linear* identifiability of learned representations, where the information is preserved while coordinates vary, an analogue, at the level of representations and for a model-class rather than a data symmetry, of Proposition 1. Whittington et al. (2023) select a basis through non-negativity and activity constraints when the sources are independent and bounded, so that their joint support is a product set—itself a support condition, which an anisotropic density on another support need not meet; within a plane, Proposition 3(ii) asks instead for a harmonic the configuration couples to. Our theory also assumes exact symmetries and rigid rotations, which trained dictionaries do not satisfy: they deform away from the polygon where regularity would deny them a harmonic (App. C; the direction of causation is open), which is why we test the graded criterion.

K TRAINED SEEDS ON THE MODEL'S OWN RING

Design. For each (ring, layer) with a cyclic plane at layers 1, 3, 7 and 12—eleven cells; the clock-hour ring is not cyclic at layer 1—the in-plane coordinates of the probes of App. I (20 templates $\times$ 12 prefixes $\times$ p items, the set behind Table 9) are centered on the item-mean centroid and scaled to mean item radius 1. The *unit* convention puts every probe on the unit circle (the criterion's setting); the *raw* convention keeps each probe's own radius. Symmetrization translates each item cluster so its mean sits on the regular p-gon of radius 1 that keeps the first item's angle and the model's cyclic order (in the unit convention every probe is then put back on the circle); a tilt reweights the symmetrized contexts by $1 + 2r\cos(q\theta_i)$, θ_i the item's p-gon angle, with $r = 0.35$ (at $q = p/2$, where the tilt and its alias coincide, the installed moment is $2r|\cos(p\theta_0/2)|$ for point masses, θ_0 the first item's angle, not App. H's $2r$: 0.06–0.69 on the seven 12-point cells' weighted densities, against 0.23–0.35 at every other q; those 19 tilts, none in the $\gcd(p, q) = 1$ count, are the weakest with $q \geq 2$, the installed order strongest in 9), and resamples as many contexts as there are probes, with replacement. Dictionaries are ReLU+L_1—chosen after TopK-1 dictionaries on this manifold showed no selectivity (App. H), for the reason §4.3 gives—with $M \in \{4, 6, k^*, p\}$ atoms, where k^* is the natural density's strongest harmonic below p (6 for the weekday ring at every layer; 3, 9, 2 for the clock-hour ring at layers 3, 7, 12; 11, 11, 6, 6 for the month ring; $|\hat{\rho}_{k^*}|$ exceeds the runner-up by 1.02–2.8$\times$ and, where $k^* = p - 1$ (six cells), $|\hat{\rho}_1|$ by 1.3–7.2$\times$), trained from 12 seeds (5000–5011) for 2,500 steps at $\lambda \in \{0.05, 0.1, 0.2, 0.4\}$ (tilts at $\lambda = 0.2$) with the training code of App. M: the plane itself ($d = 2$), Adam at a constant learning rate 3×10^{-3}, batches of 1,024 drawn with replacement from one fixed probe sample (1,680 probes for weekdays, 2,880 for months and clock hours) shared by every seed (so the symmetrized arm, which shares it, is the reference for any finite-sample fluctuation), initialization as in Table 14, and atoms firing on under 10^{-4} of the first 4,096 points, checked every 500 steps, re-drawn in random directions during the first half. The statistic is A_k, the circular concentration across seeds of the phase $\arg Z_k$ of the configuration's order parameter at harmonic k (Z_k as in App. M: atoms firing on over 1% of the points, weighted by firing frequency), for every $k \leq p$; with 12 seeds chance is ≈ 0.26 and the Rayleigh test is calibrated to 0.049. Code: `code/g21_real_ring_seeds.py`. The layer sets differ across analyses (1, 3, 7, 12 here; 7, 12, 20 in App. H; 3, 7, 12, 20 for the 48 quarter-hours; 7, 12 in App. L; 3, 7, 14, 21, 25 in Qwen and 7, 14, 21 for its removal); no common rule chose them, and we have not rerun the analyses on one set. Other choices also followed earlier runs: the coding rule, the read-out over every $k < p$ (App. H's first read-out examined $k = m$ only), the spread index and the removal's box (below).

Results, unit convention (152 (cell, M, λ) settings, 138 tilts). Symmetrization translates each cluster without reshaping it in the plane (its angular width scales roughly by the ratio of its old to its new radius), so the density's power below p falls from a median 0.30 ($\max_{k<p} |\hat{\rho}_k|$, one value per cell) to 0.04 (0.001–0.10), not to zero. Below the lattice: $\max_{k<p} A_k$ is a median 0.667 on the natural ring against 0.490 symmetrized, natural higher in 130/152 (Wilcoxon $p = 1.2 \times 10^{-19}$, settings as units and so descriptive; cell-level tests under *Reading Table 2* below). Read instead at the one order fixed before training, the density's strongest harmonic k^*, the contrast is weaker but holds: 0.314 against 0.237, natural higher in 100 of 152 (one-sided Wilcoxon $p = 1.7 \times 10^{-7}$, descriptive) and in 9 of 11 cells (sign test $p = 0.033$). *Where the maximum falls, and whether it carries the result.* $\max_{k<p}$ invites the worry that the comparison is driven by $k = 1$, the order at which the tilt analysis below finds a cluster-claiming mechanism rather than a rotational lock, and which symmetrization removes by construction. It is not: over the 152 settings the natural ring's $\arg\max_{k<p} A_k$ is spread over eleven orders, at $k = 1$ in 35 of them (23%) and at $k \geq 2$ in the rest

($k = 4$ in 31, $k = 2$ in 24, $k = 5$ in 20, $k = 3$ in 13, $k = 10$ in 10, and $k \in \{6, \ldots, 9, 11\}$ in the remaining 19). Recomputing the comparison with the suspect orders excluded leaves it intact: excluding $k = 1$ gives 0.629 against 0.461, natural higher in $126/152$ ($p = 1.4 \times 10^{-17}$), and excluding $k \in \{1, 2\}$ gives 0.607 against 0.451, higher in $118/152$ ($p = 1.7 \times 10^{-15}$). The margin narrows, as it must when the maximum is taken over fewer orders, and the conclusion does not depend on the two orders the mechanism could be confused with; the symmetrized ring's significant fraction over all 1,432 (setting, $k < p$) pairs is 0.071. At the lattice: A_p is 0.310 natural against 0.361 symmetrized, symmetrized higher in $89/152$ (one-sided Wilcoxon signed-rank $p = 3.0 \times 10^{-3}$; a sign test gives 0.02). Tilts: A at the installed order is a median 0.739, significant in $113/138$, against 0.293 at other orders excluding the alias $p - q$. Within each (cell, M), Spearman of λ against A at the series' lock order (the $k < p$ with the highest agreement averaged over the four λ, so no single λ chooses it) is a median $+0.80$, positive in $33/38$. Choosing each series' order on the first seed batch and reading the trend on the second (the replicate batch below) removes that selection: median $+0.90$, positive in 29 of 38 (sign test $p = 8 \times 10^{-4}$). *The data-level control.* λ does not only scale the kernel: it changes how many atoms survive and how regular the reached configuration is, so a trend in λ could be generic optimizer behavior. The same series read on the *symmetrized* density—same seeds, same λ, same statistic, a density stripped of the harmonic that does the pinning—rises in only 20 of 38 (median Spearman $+0.20$; 17.4 expected by chance, as four points give a positive Spearman with probability $11/24$) against the natural side's 33 of 38 ($+0.80$); on $\max_{k<p} A_k$ it is $22/38$ ($+0.20$) against $37/38$ ($+0.80$). That comparison is only half a control, because below the lattice the symmetrized density has little power by construction (a median 0.04), so a weak trend there is expected. Where the symmetrized side *can* express a generic λ effect it does, and it is not small: at the lattice with $M = p$ (44 combinations) its A_p runs 0.28, 0.25, 0.38, 0.80 over the four λ—the atoms settling onto the p-gon only once the penalty is large, a rise twice row 5's ($+0.52$ against $+0.26$), at a different order. So λ has a generic effect on where seeds converge, and the question is whether row 5 is that effect wearing the density's harmonic. Two confounds a reader would name are measurable on the stored configurations. Atom death: in 36 of the 38 series every atom is live at every λ (the two exceptions differ by a fraction of an atom), so the trend is not fewer surviving atoms. Regularity: the reached configurations do become slightly more regular with λ (median gap CV $0.465 \to 0.408$, falling in 27 of 38), and partialling that out leaves the trend where it was—A at the lock order rises with λ in 35 of 38 series (median partial Spearman $+0.84$, descriptive at four points per series) against 33 of 38 ($+0.80$) raw. What we cannot rule out with these runs is a generic convergence that happens to act at the density's own order. The control that isolates the objective's own λ-dependence is to freeze the configurations reached at one λ and read them under the others: taking the twelve shapes reached at $\lambda = 0.05$ in each series, placing each at its own exact optimal rotation under $\lambda \in \{0.05, 0.1, 0.2, 0.4\}$, and reading A at the lock order across them (`code/g52_frozen_lambda.py`) leaves the agreement where it was—$0.64 \to 0.63$, Spearman $+0.20$, positive in 20 of 38 series—against the trained series' $0.47 \to 0.73$. So on these shapes the penalty barely moves the optimum (it reshapes the kernel through λ/r, and for an irregular shape it can move the minimizer, but here the agreement at the optimal rotation does not change), and the trained rise is the seeds concentrating around optima that stay put. That is the form the fluctuation relation predicts and also the form a generic convergence would take; the symmetrized series above, which rise in 20 of 38 against 33 of 38, are the half of the separation these runs afford, and we leave the attribution there. *Null of the maximum.* Because $\max_{k<p} A_k$ is a maximum over $p - 1$ harmonics, we simulated its null (12 seeds of M uniformly random atoms each; `code/g21_null.py`): median 0.43 and 95th percentile 0.62 at $p = 7$, 0.48 and 0.65 at $p = 12$, within 0.01 across M. A combination *locks* when its natural $\max_{k<p} A_k$ exceeds that 95th percentile: 90 of 152 do, against 18 symmetrized (7.6 expected under the i.i.d. null, binomial $p = 6 \times 10^{-4}$; the symmetrized arm, which shares the pipeline's sample and seeds, is the stricter reference). Lock order: among those 90, the observed order (the setting's own $\arg\max_{k<p} A_k$, not row 5's series order) equals the density's strongest harmonic in 13 and equals $\arg\max_k |\hat{\rho}_k||\hat{\tilde{c}}_k|$ with the trained configuration's kernel in 35 (chance ≈ 10); the trained kernel alone picks it in 28 and the configurations' median $|Z_k|$ (here unweighted, $Z_k = m^{-1} \sum_j e^{ik\alpha_j}$) alone in 17, so most of what the product adds is the trained configuration's, and none of these rules is fixed before training. The density alone fails at $M = p$ as well as below it (6 of 32 and 7 of 58), so the clusters an $m < p$ dictionary chooses are not the whole explanation; crediting the lattice alias $p - k$, as row 3 does, gives 39 of 90 for the density and 40 for the product, against 19.7 by chance—the density alone at twice chance, the kernel adding one (`code/g81_lockorder_split.py`); the best constant

rule, which reads no density, picks 20 ($k = 1$), or 25 with the alias credited ($k = 2$). Row 2 is weak at every penalty and absent at $\lambda = 0.1$, where the natural median is the higher: the symmetrized A_p is the higher in 21, 19, 27 and 22 of the 38 settings at $\lambda = 0.05, 0.1, 0.2, 0.4$ (medians 0.29, 0.27, 0.40, 0.50 against 0.26, 0.30, 0.30, 0.47 natural), and the symmetrized A_p clears its single-order null (Rayleigh $p < 0.05$) in 47 of 152 settings against 32 natural. *Raw convention* (same 152 combinations and 138 tilts): 0.760 vs 0.506 below the lattice, $137/152$, $p = 2.2 \times 10^{-23}$; lattice 0.328 vs 0.354, $88/152$, $p = 2.3 \times 10^{-3}$; tilts 0.790 at the installed order ($111/138$ significant) vs 0.319 elsewhere; λ trend median $+0.90$, positive in $32/38$; 116 combinations lock above the null; lock order matched by the product in $45/116$ and by the density alone in $12/116$.

Why the agreement is harmonic-resolved. Proposition 3 pins the rotation of *one* configuration; if the twelve seeds reached the same configuration up to rotation, a pinned optimum would make them agree at every harmonic that configuration carries. They do not: on the natural ring the twelve seeds of a combination occupy a median 9.5 distinct cluster subsets (a seed's subset: the items whose mean probe angle is the nearest to one of its atoms) (4 to 12; every one of the 152 combinations has more than one), and seed pairs with the same live-atom count are $11.6°$ apart after the best rigid rotation (IQR $8.6°$–$15.0°$; $22.3°$ for random configurations; `code/g21_subsets.py`). Read atom by atom with no rotation allowed (each atom to the nearest atom of the other seed, symmetrized; `code/g70_atom_level.py`), the natural density brings seeds closer than its symmetrization does, but only slightly: a median $16.3°$ against $17.0°$, closer in 97 of 152 settings and in 10 of 11 cells (cell-level t-test $p = 0.01$). So the identity acts seed by seed, on the rotation, and leaves the cluster subset free. Nor is row 1 seeds sharing subsets (`code/g80_subset_sharing.py`): the symmetrized seeds occupy as many subsets as the natural ones (a median 10 against 9.5; 4.0% against 6.4% of seed pairs share one), and over only the pairs that claim different subsets the natural ring still agrees more below the lattice (the mean over pairs of cos of their k-th moment phases, maximized over $k < p$: 0.40 against 0.16, natural higher in 126 of 152 settings and 11 of 11 cells; all pairs, 0.44 against 0.17; atoms unweighted, since the firing weights were not stored). Nor is row 1 only the symmetrized moments being smaller: the symmetrized configurations are somewhat more regular (median gap CV 0.33 against 0.40, lower in 113 of 152 settings) and their median $|Z_k|$ below p is 0.16 against 0.21, but weighting each seed's phase by $|Z_k|$ leaves the natural ring higher ($\max_{k<p}$ 0.75 against 0.62, 110 of 152 settings, 11 of 11 cells), and over the 711 (setting, k) pairs whose two arms' median $|Z_k|$ lie within $1.25\times$ of each other the natural ring agrees more in 61% (0.30 against 0.24; atoms unweighted). For a density whose only power below p is at $\pm q$ (the symmetrized ring with one installed harmonic, up to the residual), each seed's objective is $J_s(\varphi) = 2|\hat{\rho}_q||\hat{\bar{c}}_q^{(s)}|\cos(q\varphi + \arg\hat{\rho}_q - \arg\hat{\bar{c}}_q^{(s)})$ plus lattice terms, and by Proposition 4—whose phase statement holds on the trained configurations to a median $4.9°$ where the configuration couples to q, below—$\arg\hat{\bar{c}}_q^{(s)}$ is, modulo π, minus the phase of the configuration's q-th moment Z_q (below): the minimum aligns that moment with the density's harmonic whatever the configuration, while every other moment's phase is set by which clusters the seed claimed. Agreement at q (and its alias $p - q$) and chance elsewhere is therefore what the identity predicts for seeds that differ, and on the natural density, where several $|\hat{\rho}_k||S_k^{(s)}|$ compete, the order that wins depends on the configurations reached—the two-factor limit reported above.

Predicting which order locks. The comparisons above are qualitative. The identity also makes a graded prediction over every (cell, M, λ, $k < p$) at once: Eq. 2 gives the objective's k-th Fourier amplitude as $2|\hat{\rho}_k||\hat{\bar{c}}_k|$ (measured-moment convention), so the agreement at order k should follow that product. Over the 1,432 pairs on the natural ring the rank correlation is $+0.289$ ($p = 5 \times 10^{-29}$, two-sided with the pairs, which share settings, as units, so descriptive; `code/g27_predict_Ak.py`), and it is the product rather than either factor that carries it: the density alone gives $+0.128$ and the kernel alone $+0.263$, with partial rank correlations of $+0.140$ for the density given the kernel and $+0.269$ for the kernel given the density. Within a single (cell, M, λ), where no between-cell scale can help, the median is $+0.314$, positive in 127 of 152. The product separates the pairs that actually lock ($A_k \geq 0.6$) from the rest with AUC 0.679. On the symmetrized ring, where the product is near zero at every order below p, it predicts nothing and measures nothing: $+0.023$ ($p = 0.38$). The prediction is a ranking, not yet a calibration. Reading it through the Boltzmann model of §3.3 with one global temperature fitted on these data reproduces the ordering of the binned medians ($0.29 \to 0.39$ predicted against $0.28 \to 0.47$ measured, over quintiles of the product) but underpredicts the strongest bin, so its RMSE (0.208) does not beat predicting the mean (0.202). We report the rank test as the result and the calibration as an open end.

The same controls that App. O applies. A rank correlation between two sequences indexed by k can be manufactured by a shared trend in k, and App. O shows the probe-free version of this statistic is exactly that. The same two controls are owed here and pass (`code/g40_appK_controls.py`). The trend is weaker to begin with: on a clustered ring the density's harmonics do not decay—$|\hat{\rho}_k|$ *rises* toward the lattice order (Spearman with k $+0.139$)—while the kernel's fall (-0.738) and the agreement barely moves (-0.192). With the rank of k partialled out of both sides the pooled correlation is $+0.221$ (from $+0.289$) and the within-setting median $+0.258$, positive in 102 of 149 (from $+0.314$; three settings drop out, in which partialling k leaves one side without variance). And a setting's predictor fits its own agreement spectrum better than other settings' at the same p: $+0.258$ against $+0.057$ with k partialled, its own higher in 90 of 149 (Wilcoxon $p = 4.3 \times 10^{-3}$; $+0.314$ against $+0.191$, 95 of 152, $p = 1.5 \times 10^{-4}$ unpartialled). What survives is smaller than the raw number, and it is the number we stand behind.

The phase statement of Proposition 4 on the trained configurations. The account above routes through Proposition 4, whose disjoint-arc hypothesis none of the 17,184 trained (seed, q) pairs here satisfies, and whose magnitude is only a heuristic outside it (App. I). What the mechanism uses is only its *phase*: $\hat{\bar{c}}_q = -\frac{m}{2}\hat{g}_q\overline{Z_q}$ with $\hat{g}_q$ real, so $\arg\hat{\bar{c}}_q + \arg Z_q \equiv 0 \pmod{\pi}$ where $Z_q = m^{-1}\sum_j e^{iq\alpha_j}$, atoms unweighted. Computing $\hat{\bar{c}}_q$ with the exact 2-plane solver for every trained seed, that relation holds to a median 10.1° over all pairs (resultant of $e^{2i(\arg\hat{\bar{c}}_q + \arg Z_q)}$ 0.651), and to a median 4.9° (resultant 0.838) over the 6,344 pairs whose configuration couples to q at all ($S_q \geq 0.3$)—the pairs where pinning can happen. Where $S_q < 0.3$ the deviation is 15.3°, and the phase of a moment that small is not what the objective aligns. The closed form's amplitude is a heuristic on these configurations; its phase is not.

Replication over seeds. An independent batch of 12 seeds (5100–5111; `code/run_g21r.sh`) on the same 152 combinations (unit convention, no tilts) reproduces the comparisons: $\max_{k<p} A_k$ natural 0.642 against 0.481 symmetrized, natural higher in 122/152 (Wilcoxon $p = 3 \times 10^{-17}$); at the lattice, symmetrized higher in 84/152 (one-sided Wilcoxon $p = 3 \times 10^{-3}$); symmetrized significant fraction below p 0.063. The lock order is only partly a property of the density: of the 90 combinations locked in the first batch, the second locks above the null in 58, is significant at the first batch's order in 59 (Rayleigh $p < 0.05$; chance ≈ 4.5) and has the same strongest order in 33 (chance ≈ 10). At the first batch's order it locks at the same phase: pooling the 24 configurations leaves the concentration at the first batch's order at 0.64 against a per-batch 0.65, within $0.8\times$ in 87 of 90. Code: `code/g21_replicate.py`.

Cluster choice and the tilt. A tilt by $1 + 2r\cos(q\theta)$ puts extra mass on the q clusters nearest the peaks of $\cos(q\theta)$, so a sceptic may propose that the seeds merely claim those clusters. The two accounts differ in where agreement should appear. The tilted density has power only at $\pm q$ (and at the alias $p - q$, since it is sampled on p points), so the criterion pins the seeds' q-th moment and nothing else below p; a claimed subset of q equally spaced clusters has power at every multiple of q, so claiming it would make the seeds agree at $2q, 3q, \ldots$ as well. On the two 12-point rings with $q = 2$ the accounts disagree at $k \in \{4, 6, 8\}$: over the 19 tilt runs, A is a median 0.94 at q, 0.41 at the alias, 0.30 at those subset harmonics and 0.26 at the remaining orders (chance ≈ 0.26). With $q = 3$ the free order is $k = 6$ (9 is the alias): 0.64 at q, 0.41 at $k = 6$, 0.25 elsewhere—a trace of the subset, and a clear gap to q. On the 7-point ring, where no $q \in \{2, 3\}$ equally spaced clusters exist, A at the multiples of q below p is 0.19 and 0.18 ($k = 4, 6$; significant in none of the 8 tilts) for $q = 2$ and 0.50 ($k = 6$; 4 of 8) for $q = 3$, against 0.94 and 0.75 at q. Where the subset account does win is $q = 1$: a tilt by $1 + 2r\cos\theta$ loads one side of the ring, every seed claims that side, and a one-sided subset's second moment is fixed by its first, so the seeds agree at $k = 2$ (0.83, significant in 25 of 27 tilts) more than at the installed $k = 1$ (0.57)—visible as the dark $(1, 2)$ entry of Fig. 1b. The criterion pins one moment of each configuration and says nothing about the rest; when the density itself selects the subset, the rest come along. A rule whose in-plane optimum is flat shows the same reading is not a pin: TopK-2 seeds on these tilts agree at the installed $k = 1$ in 26 of 27 ($A_1 = 0.68$), while at $q \geq 2$ they show no selectivity (9 of 35 at $\gcd(p, q) = 1$, chance 8.8; App. M). Read as each other's controls, the tilts are only partly what a symmetry argument cannot supply: a tilt at q keeps the symmetry $\mathbb{Z}_g$, $g = \gcd(p, q)$, of its item weights (up to the symmetrization residual), so Proposition 1 alone confines agreement to multiples of g—at $q \in \{4, 6\}$ on the 12-point rings, already q and its alias—and Proposition 3 alone says where only when $g = 1$. Over the 138 tilts the installed order or its alias is the strongest order below p in 93 (chance 27 under any null blind to the

tilt, since each (cell, M)'s installed orders and their aliases partition the orders below p; 91 of the 111 with $q \geq 2$), against a median 0.37 at the orders the other tilts would have installed (significant in 137 of 618, 22% against a calibrated 5%; 58 of them in the $q = 1$ tilts above, and without those and the multiples of g a tilt's symmetry allows, 68 of 450, 15%; the strongest order in 45 of the 618, 18 of the 450) (`code/g34_tilt_subset.py`). Over the 62 tilts with $g = 1$, where only Proposition 3 says where, the installed order or its alias is the strongest in 30 (chance 14.9), and in 28 of the 35 of those with $q \geq 2$ (chance 8.8 with every order equally likely, 3.5 under the untilted symmetrized runs' own distribution of the strongest order; significant at q in 34).

Reading Table 2. The null median of $\max_{k<p}$ is 0.43–0.48 and its 95th percentile 0.62–0.65. Counts in rows 1–3 are over settings, which share cells, so the cell-level counts are the conservative ones (11 of 11 for row 1; 8 of 11 for row 2, one-sided sign test $p = 0.11$, hence "weakly"; Table 10); setting-level Wilcoxon p-values in this appendix treat settings as independent and are descriptive, and so, less severely, are the cell- and layer-level tests below, since cells and layers of one model are not independent either; the two models are the only independent units. Per cell (Table 10) the symmetrized side sits above the null median on the weekday ring (0.48–0.52 against 0.43) and straddles it on the 12-point rings (0.44–0.61 against 0.48, four cells below), with clock hours at layer 12 (0.61) near the 95th percentile (0.65) and also the full-stream exception (App. L). With the cell as the unit—one mean difference per cell, a one-sided t-test over the 11 cells (`code/g64_cell_level.py`)—row 1 gives $+0.16$ ($t = 11.2$, $p = 3 \times 10^{-7}$) and row 3, the installed order against the others below p, $+0.37$ ($t = 22.3$, $p = 4 \times 10^{-10}$), positive in 11 of 11 cells each; a random-intercept model with cells as groups gives 0.16 and 0.38. Row 4 is the 48-point clock density, whose items are spread over the circle (every seed claims a different subset of them, before and after the edit); the edit is fixed per layer, so the density-level count is by layer: lower at all four Gemma layers (11, 11, 12 and 10 of 12 settings at layers 3, 7, 12, 20) and at all three Qwen2.5-1.5B layers (12 of 12 each), 7 of 7 layers (one-sided sign test $p = 1/128$ with the layers of two models as units; $1/16$ and $1/8$ within Gemma and within Qwen); it replicates on Qwen with only the model changed (36 of 36); the lock is weakened rather than relocated ($\max_{k<p} A_k$ 0.93 $\to$ 0.74 and 0.94 $\to$ 0.78, still above the null's 95th percentile in 12 of 24 and 24 of 36 settings, below); and it does *not* hold on the clustered rings, a failure counted against the criterion (App. S, (17)). Row 5's mechanism check is the kernel coefficient the curvature contains, rising with λ on configurations chosen without reference to any training run (74% of series) and on the trained ones (84%, App. T), and its control the symmetrized series above; row 3's $q = 1$ exception is discussed above.

Editing one coefficient of a non-symmetric density. Every comparison above replaces the natural density by a *symmetric* one, and a symmetric density is what Proposition 1 already speaks about: the orbit structure alone predicts that the seeds stop agreeing below the lattice, with no reference to which harmonic carries the power. The test that only Proposition 3 can pass keeps every condition non-symmetric and changes one Fourier coefficient (`code/g36_harmonic_surgery.py`), on the 48-point ring at layers 12 and 20, $M \in \{4, 6, 8, 12\}$, $\lambda \in \{0.1, 0.2, 0.4\}$, twelve seeds.

Removal. The items stay where they are; only their relative mass changes. For item j let u_j be the mean of $e^{-ik^*\theta}$ over that item's contexts, and take per-item weights with $\sum_j w_j = 1$ inside the box $w_j \in [\frac{1}{2p}, \frac{2}{p}]$ that minimize $|\sum_j w_j u_j|$ (400 rounds, from uniform weights, of projecting onto the affine set $\{\sum_j w_j u_j = 0, \sum_j w_j = 1\}$, clipping to the box and renormalizing; the box keeps the edit interpretable and makes the removal partial); contexts are then resampled with those weights, as many as there are probes and with replacement, as the tilt does. The natural arm is the original sample, not resampled; the two controls that separate the edit from the reweighting and the resampling are reported below (*Sham*). Removing the density's own strongest harmonic k^* takes $|\hat{\rho}_{k^*}|$ from a median 0.305 to 0.075 and takes the agreement at k^* from 0.93 to 0.40, lower in 22 of the 24 settings (Wilcoxon $p = 2.0 \times 10^{-6}$). The same design at the ring's other two cyclic layers, 3 and 7 ($k^* = 3$ at both; `--layers 3` and 7), gives 0.63 $\to$ 0.40 and 0.64 $\to$ 0.44, lower in 11 of 12 at each ($p = 1.7 \times 10^{-3}$, 4.6×10^{-3}); over the four layers, 0.76 $\to$ 0.41, 44 of 48 ($p = 1.1 \times 10^{-10}$), with $|\hat{\rho}_{k^*}|$ 0.23 $\to$ 0.02. Per layer the dose and the drop go together: $|\hat{\rho}_{k^*}|$ goes 0.260 $\to$ 0.012, 0.135 $\to$ 0.008, 0.193 $\to$ 0.022 and 0.417 $\to$ 0.128 at layers 3, 7, 12 and 20, and the agreement 0.63 $\to$ 0.40, 0.64 $\to$ 0.44, 0.87 $\to$ 0.26 and 0.96 $\to$ 0.79, so layer 20, which keeps the most of its harmonic, loses the least. At layers 12 and 20 the removed $k^* = 2$ is largely the projection's ellipticity (below), and Qwen's plane is elliptical too (axis ratio 1.29–1.61, $|\hat{\rho}_2| = 0.26$–0.34); at layers 3 and 7, where $k^* = 3$, the drop holds in 22 of 24. The seeds occupy a

median 12 distinct cluster subsets before and after, and no condition falls below four, so the drop is not the shared-subset mechanism of the $q = 1$ tilt above.

Sham. Two arms, run afresh alongside a new natural arm and a new removal at all four layers (`g36_harmonic_surgery.py --sham`; `code/g89_removal_sham.py`), separate the edit from what it rides on: the same boxed reweighting aimed at another harmonic q_{sham} (the strongest low order other than k^*, its neighbors and $2k^*$: 10, 7, 10, 5 at layers 3, 7, 12, 20), whose weights span a comparable range (0.57–1.57× uniform against the removal's 0.50–1.84×, medians) and which leaves $|\hat{\rho}_{k^*}|$ in place (median 0.26 after), and the natural contexts resampled with uniform weights. The removal replicates ($0.76 \to 0.42$ at k^*, lower in 45 of 48 settings and at 4 of 4 layers); the sham does not lower the agreement ($0.76 \to 0.85$, lower in 24 of 48, one-sided Wilcoxon $p = 0.81$) and neither does the resampling (0.78, lower in 17 of 48); the sham sits above the removal in 44 of 48. One caveat: at layer 12 the sham's target is weak ($|\hat{\rho}_{10}| = 0.045$), so there the sham tests the reweighting more than the removal of a comparable coefficient; at the other three layers it removes a coefficient of 0.06–0.18.

A clean edit. The boxed removal moves other coefficients as a side effect ($|\hat{\rho}_1|$ rises at every layer, below), so a reader can ask whether the seeds lose agreement at k^* because k^* went or because the spectrum changed. A second edit (`g36 --removal clean`; `code/g92_clean_removal.py`), in the same box, minimizes $|\hat{\rho}_{k^*}|$ with every other harmonic up to 8 held at its natural complex value by a penalty: in the weights the other coefficients then move by at most 0.002 at layers 7 and 12, 0.022 at layer 3 and 0.045 at layer 20, the two layers where the box allows only a partial removal. On the resampled densities the seeds train on, $|\hat{\rho}_{k^*}|$ goes $0.260 \to 0.087$, $0.135 \to 0.031$, $0.193 \to 0.003$ and $0.417 \to 0.241$ at layers 3, 7, 12 and 20, and no other $|\hat{\rho}_j|$, $j \le 8$, moves by more than 0.035, 0.018, 0.027 and 0.070 (resampling noise included). The agreement at k^* falls as under the boxed edit: $0.76 \to 0.40$, lower in 45 of 48 settings and at 4 of 4 layers (one-sided Wilcoxon $p = 1.2 \times 10^{-9}$, descriptive; $0.63 \to 0.43$, $0.64 \to 0.38$, $0.87 \to 0.17$ and $0.96 \to 0.85$), again least where least of the harmonic goes. On Qwen2.5-1.5B's own 48-point density (layers 7, 14, 21, where $k^* = 2$ throughout), the same clean edit, which in the weights moves no other coefficient by more than 0.018 while taking $|\hat{\rho}_2|$ to 0.04–0.08, takes the agreement at k^* from 0.93 to 0.52, lower in 35 of 36 settings and at 3 of 3 layers ($p = 2 \times 10^{-10}$, descriptive). On the plane fitting Llama-3.1-8B's number circle the box lets only about half the coefficient go ($|\hat{\rho}_{k^*}|$ a median $0.23 \to 0.10$ in the weights, other coefficients within 0.042), and the drop is correspondingly smaller: $0.70 \to 0.57$, lower in 73 of 84 settings and at 6 of 7 layers ($p = 1.5 \times 10^{-9}$, descriptive), against $0.70 \to 0.39$ for the boxed edit, which removes more; on Pythia-6.9B's, with the same half removal (a median $0.23 \to 0.10$; others within 0.031), only weakly: $0.68 \to 0.64$, lower in 61 of 84 and at 5 of 7 layers ($p = 8 \times 10^{-4}$, descriptive), against $0.68 \to 0.49$ boxed. So the unpinning follows the one coefficient, and its size, not the rest of the spectrum.

The removal in a second model, with the design fixed by the above. The strongest objection to this arm is that it is one surgery on one density in one model, with the criterion's distinctive content resting on it. So we ran it again, changing only the model and keeping the design, the reading rule and the diagnostic: Qwen2.5-1.5B, its own 48 quarter-hour probes projected into its own clock-hour plane (`code/g28_continuous_rings.py` with `--model`); layers 7, 14 and 21 (cyclic at all three, maximum item gap 5.9–7.4 uniform gaps); the same $M \in \{4, 6, 8, 12\}$, $\lambda \in \{0.1, 0.2, 0.4\}$, twelve seeds, 36 settings. The harmonic to remove is again the density's own strongest, here $k^* = 2$ at $|\hat{\rho}_2| = 0.314$.

Removal takes $|\hat{\rho}_{k^*}|$ to 0.031 and the agreement at k^* from a median 0.933 to 0.479, **lower in** 36 **of** 36 **settings** (Wilcoxon $p = 1.5 \times 10^{-11}$), with the subset count a median 12 before and 12 after—so the drop is again the density's and not the shared subset's. This is the prediction Proposition 1 cannot make, holding on a model the design never saw. What did not vary is the construction: the same item set, the same projection into a coarser ring's plane, so this is a replication across models of one design and not a second kind of density.

The competition arm is the weaker half here and says why in the criterion's own terms. The harmonic the rule installs is the weakest of the orders 2–8 more than one away from k^*, which for this density is $q = 8$ at all three layers; installing it at $r = 0.5$ raises the agreement at q from a median 0.324 to 0.468, higher in 28 of 36 ($p = 8.0 \times 10^{-5}$), but never makes q the *argmax*: $k^* = 2$ keeps the lock in all 36. The graded movement is m-ordered as the structure factor suggests—higher in 4 of

9 settings at $M = 4$, 7 of 9 at $M = 6$, 9 of 9 at $M = 8$ and 8 of 9 at $M = 12$—because $|\hat{\tilde{c}}_8|$ is small for a four-atom dictionary however much density power sits at 8. On Gemma the installed harmonic was $q = 4$ and the lock order moved to it in 9 of 24 settings (below). We report both: the removal prediction replicates outright in a second model, and the competition prediction replicates in its graded form, bounded by the kernel in the way Corollary 1 and Proposition 4 suggest.

The box matters, and came second. Our first run had none: the same construction drives $|\hat{\rho}_{k^*}|$ to 0.011 but needs weights spanning $27\times$, and the resulting density is concentrated enough that every seed claims the same few heavy items: the subset count falls to 7 and the seeds then agree at k^* anyway (0.76), through the subset rather than the density. We report the bounded version, which removes less of the harmonic and leaves the mechanism interpretable, and the reader should weigh that it was chosen after the unbounded run.

The clustered rings. The same design on the clustered 7- and 12-point rings (days and months, layers 7 and 12, $M \in \{4, 6\}$, 24 settings) does *not* reproduce it, and the diagnostic says why. Cancelling k^* there takes $|\hat{\rho}_{k^*}|$ from 0.232 to 0.018 but leaves the agreement at k^* where it was (0.30 natural against 0.44 removed, lower in 9 of 24, $p = 0.86$)—and the distinct-subset count falls from 7 to 4, below our collapse threshold (fewer than four distinct subsets among the twelve seeds) in 12 of the 24 conditions. We read two things into it. First, on a clustered ring the density's strongest harmonic is not where the seeds agree in the first place (a median 0.30 at k^*, against chance ≈ 0.26; the same fact as the 13-of-90 lock-order result of §5.1), so there is nothing there for the surgery to remove. Second, with $m < p$ the seeds claim clusters, and the reweighting makes them claim the same ones. The surgery bites on the spread density, where the subset count stays at 12, and not where the configuration controls the lock. We count this arm as a failed prediction (App. S, (17)); the reading above comes after the fact and does not rescue it.

Competition. Reweighting the natural contexts by $1 + 2r\cos(q\theta)$ at $q = 4$ at both layers—chosen not by the rule of the later runs, which would have picked 5 and 7; $|\hat{\rho}_4|$ is 0.05 at layer 12 but 0.12 at layer 20—raises the agreement at q from a median 0.58 to 0.82 at $r = 0.2$ and 0.93 at $r = 0.5$, higher in 22 and 24 of 24 settings ($p = 1.1 \times 10^{-6}$ and 6.0×10^{-8}), with the subset count unchanged at 12. The *lock order*—the argmax over 47 harmonics—moves to q in only 9 of the 24, because the bounded reweighting at $q = 4$ also raises the density's own $k^* = 2$ ($|\hat{\rho}_{k^*}|$ from a median 0.31 to 0.45 at $r = 0.5$, while $|\hat{\rho}_q|$ rises to 0.55): both moments are pinned at once, which is what the identity says. The installed harmonic has to be a low order for this to work at all—$|\hat{\tilde{c}}_q|$ decays with q for a small dictionary, so installing $q = 36$ moves nothing however much density power it carries, which the criterion itself predicts.

Amplitude at fixed mode count. Installing $q \in \{2, 3\}$ at amplitudes $r \in \{0.1, 0.2, 0.35, 0.45\}$—every cluster keeps positive mass, so the mode count is p throughout and only the harmonic's amplitude moves—with $M = 4$ over the eleven cells (`code/run_g21t.sh`): A_q rises with r in 19 of 22 (cell, q) series (Spearman median $+0.80$), from a median 0.62 at $r = 0.1$ to 0.81 at $r = 0.45$. At $q = 2$ the medians are 0.73, 0.76, 0.84, 0.91, significant in $11/11$ cells at every r; at $q = 3$, 0.49, 0.66, 0.63, 0.63, significant in 5, 9, 11, 11. Code: `code/g21_rsweep.py`.

A second model for rows 1, 3 and 5. Apart from the removal above, everything above is Gemma-2-2B. Qwen2.5-1.5B places the same three item sets on closed curves in their top-2 principal plane—the cyclic-order test passes at 14 of the 15 (ring, layer) cells we probed, at layers $3, 7, 14, 21, 25$ of its 28 (`code/g30_second_model.py`)—so the test runs there unchanged. Over 196 (cell, M, λ) settings in 14 cells the natural ring is a median 0.688 against 0.519 symmetrized, higher in 156 of 196 (Wilcoxon $p = 4.3 \times 10^{-23}$) and in 14 of 14 cells; 126 natural settings clear the calibrated null against 43 symmetrized; an installed harmonic locks its own order at a median 0.721 against 0.330 elsewhere and is the strongest order in 98 of 180 tilts; and agreement rises with λ in 45 of 49 (cell, M) series (median Spearman $+0.80$). Every comparison comes out as in Gemma, in a different model of a different size by a different laboratory; two rates differ, the symmetrized side clearing its null more often (43 of 196 against 18 of 152) and the installed order being strongest less often (98 of 180 against 93 of 138).

Which order the seeds lock at, on the spread density. §5.1 reports that the density's strongest harmonic picks the observed lock order in 13 of the 90 locked combinations of the 7- and 12-point rings—chance. Those are clustered densities, where an $m < p$ dictionary also faces a choice of which clusters to claim, a degree of freedom that Proposition 3, a statement about rotations under

any angular density, does not describe, and which a density spread over the circle does not pose. The 48-point ring is that density, and the prediction was never read on it. Reading it on the runs already reported (`code/g35_lockorder_spread.py`; no retraining), with the null of $\max_{k<p} A_k$ at $p = 48$ simulated per M as `g21_null.py` does at $p = 7, 12$ (95th percentile 0.71–0.72 across M): the observed $\arg\max_{k<p} A_k$ equals $\arg\max_{k<p} |\hat{\rho}_k|$ in 33 of 64 settings and in 24 of the 36 locked ones, against a uniform chance of $1/47$—but lock orders pile up at low k, and the constant rule "$k = 2$", which reads no density, is right in 35 of 64 (26 of the 36 locked); the density's top three contain it in 43 of 64; folding in the trained kernel, $\arg\max |\hat{\rho}_k||\hat{c}_k|$ picks it in 39 of 64 (27 of the 36 locked). Over the locked settings only, the denominator of the clustered count, the density (24) trails the constant rule (26) and the product ties it. The symmetrized side is the control: symmetrization flattens the density below the lattice, only 3 settings lock there at all, and none of the three is predicted.

Per layer, and where the power comes from. The prediction holds where the density has a dominant harmonic and fails where its spectrum is flat, which is the theory's own contrast, though four layers make it descriptive rather than a test (Spearman $+0.95$ between a layer's dominance ratio and its hit rate). Layer 20: strongest $k = 2$ at $|\hat{\rho}_2| = 0.417$ against 0.142 for the runner-up, lock order matched in 15 of 16. Layer 12: $k = 2$ at 0.193 against 0.103, matched in 12 of 16. Layer 3: $k = 3$ at 0.260 against 0.177, matched in 3 of 16. Layer 7: $k = 3$ at 0.135 with $k = 2$ at 0.129 a 5% behind—the prediction has no resolution there—matched in 3 of 16, with the seeds taking $k = 2$. *The $k = 2$ power is the projection's ellipticity*: the unit convention removes radial ellipticity, but an elliptical ring crowds its points in angle, and that crowding is a second harmonic. Across the four layers the plane's axis ratio (1.164, 1.139, 1.224, 1.462) tracks $|\hat{\rho}_2|$ (0.120, 0.129, 0.193, 0.417) at Spearman $+0.80$. The removal test above is agnostic to where a harmonic comes from, and there $k = 2$ is simply the density's strongest; the lock-order reading is not, because an elliptical projection predicts $k = 2$ without reading the density. Over the 32 settings whose predicted order is *not* $k = 2$, the lock order is picked exactly in 6 of 32 and within the top three in 15 of 32—nine and seven times a uniform chance, but no better than the constant rules "$k = 3$" (6 of 32) or "$k = 2$" (8 of 32). Read against a low-order baseline, the density alone does not predict the lock order on the spread density either; Table 2 and App. S count it as a failure.

A density spread over 48 points. On a 7- or 12-point ring a dictionary of $m < p$ atoms faces a choice of which clusters to claim, which is not the continuous rotation Proposition 3 is stated for. We call a density *spread* when neighboring items overlap in angle: the overlap ratio ω, the median within-item circular standard deviation over the item spacing $2\pi/p$, is 0.02–0.41 (median 0.12) on the eleven clustered cells and 0.49–6.1 on every spread density used here (medians 1.5 and 1.9 for the 48 quarter-hours in Gemma and Qwen, 5.2 for the compass bearings, 1.8 and 1.4 for the Llama and Pythia number circles; `code/g83_spread_index.py`). The ratio was defined after these runs and no cut-off was fixed in advance, but the two groups do not overlap, so any cut-off between 0.41 and 0.49 sorts them alike. In degrees, the largest item gap of every spread density (35–72°) is within the 75° that App. I's arc filter admits at $p = 12$, while one of the eleven clustered cells, clock hours at layer 3 (2.98 uniform gaps, 89°), would fail that filter; without it rows 1–3 read 114 of 136 settings (10 of 10 cells), 83 of 136 and 26 of 32 ($q \geq 2$, $\gcd(p, q) = 1$). Gemma places 48 quarter-hour clock times in cyclic order around the *clock ring's own* plane—circular-circular correlation between item index and angle -0.82, $+0.97$, $+0.94$, $+0.92$ at layers 3, 7, 12, 20 (the sign is the ring's orientation), with the items spread between the hour marks rather than piled on them (`code/g28_continuous_rings.py`, `--plane-ring hours`). With $M \in \{4, 6, 8, 12\}$ atoms, so $M \ll p = 48$, the natural density is a median 0.724 against 0.613 symmetrized over 64 settings (four cells $\times$ four M $\times$ four λ), higher in 50 of 64 ($p = 3.3 \times 10^{-8}$) and in 4 of 4 cells, with 36 natural settings above the null against 3 symmetrized; an installed harmonic locks at 0.549 against 0.251 elsewhere; and the λ trend is $+1.00$ in 16 of 16 series. Read in their own top-2 plane, rather than the clock's, the quarter-hours do not form a ring at all (no layer passes the cyclic-order test, maximum gaps of 5–18 uniform gaps), and neither do weekly calendar dates or compass bearings in degrees; we report that as the reason the main text uses the item sets it does.

What that costs the surgery arm, stated plainly. The removal prediction is the one only Proposition 3 makes, and the two densities on which it holds in Gemma and Qwen are both of this constructed kind—a fine item set projected into a coarser one's plane—while it fails on the weekday and month rings, which are the manifolds the paper opens with and the ones prior work studies.

Nothing in Proposition 3 asks for a ring: it is a condition on an angular density, and a density spread over the circle in *any* plane is a legitimate test bed. In Gemma the only spread density we have found in a plane the model supplies on its own is the compass density below, and it is confounded; Llama-3.1-8B supplies one, and the removal prediction holds there (next paragraph). The test that would settle it is the same design on a continuous quantity read in its own top-2 plane. Of the item sets available to us (dates, compass bearings, numbers) none forms a ring in its own plane—their item means are not in cyclic order—but Proposition 3 does not ask for a ring, only for an angular density with power below the lattice, and the 36 compass bearings at layers 12 and 20 are spread enough in their own plane to carry the design (maximum item gap 5.4 and 4.5 uniform gaps). So we ran it there, with the cyclic-order gate lifted and everything else as above (`--require-cyclic 0`; 24 settings). Two things differ from the constructed densities and both are reported. The bounded removal takes out less of the harmonic here—$|\hat{\rho}_{k^*}|$ falls from 0.209 to 0.069 at layer 12 and from 0.346 to 0.185 at layer 20, against $0.305 \to 0.075$ and $0.314 \to 0.031$ on the two clock densities—and the drop it produces is correspondingly smaller: the agreement at k^* goes from a median 0.774 to 0.607, lower in 18 of 24 settings (Wilcoxon $p = 0.010$; 8 of 12 at layer 12, 10 of 12 at layer 20), with the subset count $12 \to 11$, so it is not the shared-subset mechanism. The competition arm does not move the argmax in any of the 24 and moves the graded agreement at q only at layer 20 (11 of 12 at $r = 0.2$). A partial removal that produces a partial, significant unpinning with the confound controlled is what the criterion predicts for a partial removal, and it is the first time in Gemma the removal prediction has held on a density in a plane the model supplied on its own; we report it at that strength—weaker than the 36 of 36 above, on one item set, and with the lock handed to the $k = 1$ the edit adds (below)—rather than as a fourth clean replication.

A spread density the model supplies: Llama-3.1-8B's number circle. Llama-3.1-8B (Grattafiori et al., 2024) tokenizes every integer below 1000 as one token and represents integers on a helix with Fourier periods 2, 5, 10 and 100 (Kantamneni & Tegmark, 2025); its period-100 component places $0, \ldots, 99$ around a circle, a spread density that we did not construct. We read 6,000 probes (ten sentence templates, six prefixes, the hundred integers, at the integer's token) at layers 0, 2, 4, 8, 12, 16 and 20 and take the least-squares plane of the period-100 component (`code/g55_llama_harvest.py`). The hypothesis fixes the period; the model supplies the plane, which is essentially the item means' own top-2 plane (it carries 0.21–0.28 of their variance, against 0.22–0.31 for the top-2 principal plane), the item means follow $2\pi n/100$ with phase coherence $|\frac{1}{100}\sum_n e^{i(\theta_n \mp 2\pi n/100)}|$ (the better orientation) of 0.96–0.97, and the density covers the circle but for a gap of 47–71° at the seam between 99 and 0. Its strongest harmonic below p is $k^* = 3$ at six of the seven layers and 2 at layer 20. The design above runs unchanged ($M \in \{4, 6, 8, 12\}$, $\lambda \in \{0.1, 0.2, 0.4\}$, twelve seeds, the same box, 84 settings at those seven layers; `code/g65_numbers_summary.py`). The bounded reweighting removes the harmonic almost entirely ($|\hat{\rho}_{k^*}|$ from 0.20–0.29 to at most 0.018, weights within 0.50–1.87× uniform), and the agreement at k^* falls from a median 0.70 to 0.39, lower in 70 of 84 settings (Wilcoxon $p = 4 \times 10^{-11}$) and at 7 of 7 layers (sign test $p = 1/128$); in the 66 settings where the natural density had locked the seeds at k^* it falls in 64. Unlike on the clock density, agreement does not disappear: at the edited density's new strongest harmonic it rises ($0.82 \to 0.91$, higher in 72 of 84), which is what Proposition 3 predicts when a second harmonic remains; but the seeds already agreed more there than at k^* before the edit (0.82 against 0.70; higher in 56 of 84), so the lock strengthens where it was rather than moves, and the density's strongest harmonic was the seeds' lock order in only 21 of 84 settings to begin with. The edit raises $|\hat{\rho}_1|$ at all seven layers (0.07–0.13 to 0.14–0.21), and A_1 does not follow ($0.45 \to 0.38$, higher in 40 of 84). Installing a harmonic ($r = 0.35$) raises agreement at it in 64 of 84. Of the 14 settings that do not fall, 12 are four-atom dictionaries and 12 had no lock at k^* to lose. The removal prediction therefore holds on a density a model supplies on its own, in a third model, at every layer tested. The other rows replicate there too (`code/g21_real_ring_seeds.py` unchanged, $M \in \{3, 4, 6, 8\}$, four penalties, the six layers 0, 4, 8, 12, 16, 20; `code/g68_numbers_table2.py`). Read at k^*, fixed before training (with $p = 100$ the maximum over 99 orders is chance-dominated), the natural density locks the seeds more than its symmetrization onto the regular 100-gon (0.49 against 0.27, higher in 81 of 96 settings and at 6 of 6 layers); an installed harmonic is locked at its own order (0.74 against 0.31 at the other orders up to 12; significant in 98 of 144 tilts, 6 of 6 layers); and agreement at k^* rises with λ (median Spearman +0.80, positive in 23 of 24 series).

A fourth model: Pythia-6.9B. The same harvest on Pythia-6.9B (Biderman et al., 2023), another tokenizer and another training corpus (`code/g55_llama_harvest.py --model`

pythia-6.9b), gives a period-100 plane of the same kind: it carries 0.19–0.28 of the item means' variance against 0.20–0.30 for their top-2 principal plane, phase coherence is 0.95–0.98, the seam gap is 41–69°, and $k^* = 3$ at six of the seven layers and 2 at layer 20. The removal runs unchanged (84 settings at the same seven layers; `code/g65_numbers_summary.py pythia`): $|\hat{\rho}_{k^*}|$ falls from 0.16–0.24 to at most 0.015 (weights within 0.50–1.81× uniform), and the agreement at k^* from a median 0.68 to 0.49, lower in 65 of 84 settings (Wilcoxon $p = 2 \times 10^{-8}$) and at 7 of 7 layers (sign test $p = 1/128$); in the 67 settings where the seeds had locked at k^* it falls in 60. Installing a harmonic raises agreement at it in 61 of 84. Agreement at the edited density's new strongest harmonic rises less cleanly than in Llama ($0.81 \to 0.90$, higher in 58 of 84, $p = 0.07$, one-sided Wilcoxon; a sign test gives 3×10^{-4}), and here too the seeds already agreed more there than at k^* before the edit (0.81 against 0.68, in 59 of 84); the edit raises $|\hat{\rho}_1|$ (0.07–0.13 to 0.10–0.25 over the layers) and A_1 does not follow ($0.37 \to 0.40$, higher in 36 of 84).

Where the agreement goes, and what the edit does to the rest of the spectrum. Reading only A at k^* leaves two questions open: whether the seeds stop agreeing or merely agree somewhere else, and whether the bounded reweighting edits one coefficient or several. Both are answerable from the stored spectra (`code/g50_surgery_spectrum.py`). The edit is not clean. Weights confined to $[\frac{1}{2}, 2]\times$ uniform cannot cancel k^* without loading one side of the circle, so every removal raises $|\hat{\rho}_1|$: from 0.095 to 0.136 on the Gemma clock density (the mean of its two layers, $0.048 \to 0.101$ and $0.142 \to 0.171$; the edit is fixed per layer, so these are two values, not 24), from 0.117 to 0.223 on Qwen's (its middle layer; 0.10–0.16 to 0.17–0.24 over the three), where it becomes the strongest coefficient below the lattice at two of the three layers, and to 0.155 and 0.211 on the compass density, where it is the strongest at both. The seeds' first moment follows that side effect on two of the five edited densities: A_1 rises in 30 of 36 Qwen settings ($0.456 \to 0.541$, $p = 1.8 \times 10^{-5}$; this paragraph's p-values are two-sided Wilcoxon) and in 20 of 24 compass settings ($0.293 \to 0.484$, $p = 2.0 \times 10^{-5}$; 12 of 12 at layer 20, where $|\hat{\rho}_1|$ goes from 0.044 to 0.211), and not on Gemma (13 of 24) or on the number circles (40 and 36 of 84, above), where $|\hat{\rho}_1|$ doubles as it does on Qwen's. Where it follows, that is the criterion acting on the coefficient the edit added; where it does not, we have no account of it and count the $k = 1$ reading for neither side. Either way it has to be read alongside the removal. Read on the whole spectrum, the two constructed densities lose the lock and do not relocate it: $\max_{k<p} A_k$ falls from 0.934 to 0.741 on Gemma (lower in 22 of 24, $p = 1.8 \times 10^{-4}$; settings above the null of the maximum, its 95th percentile 0.72 at $p = 48$, $19 \to 12$) and from 0.935 to 0.782 on Qwen (lower in 31 of 36, $p = 2.2 \times 10^{-8}$; $36 \to 24$), the strongest order other than k^* and its alias does not rise ($0.710 \to 0.706$ on Gemma, $0.791 \to 0.758$ on Qwen), and the argmax lands on the density's new strongest order in only 4 of 24 and 14 of 36. On the compass density the lock relocates rather than dissolves: $\max_{k<p} A_k$ is 0.907 before and 0.890 after (lower in 18 of 24, $p = 0.07$; $24 \to 20$ above the null, 0.70 at $p = 36$), and what the surgery takes from $k^* = 3$ it hands to $k = 1$, the order at which the tilt analysis above found a cluster-claiming mechanism rather than a rotational lock. So the own-plane result is an unpinning of the order, not of the dictionary, and we count it at that strength.

Atoms that must serve competing directions. The same test in a 16-dimensional subspace—the ring plane plus the top 14 principal directions of the rest of the residual stream, which carry 66–84% of the off-plane variance—with $M \in \{24, 32\}$ atoms at $\lambda \in \{0.1, 0.2, 0.3\}$, six cells (`code/g23_export_pca.py`, `g24_subspace_seeds.py`). Between 7.6 and 20 atoms land in the plane on average per (cell, M, λ). Below the lattice the natural ring locks almost perfectly ($\max_{k<p} A_k$ median 0.99) and the symmetrized one far less (0.67, at the null's 95th percentile; natural higher in $34/36$, $p = 2.8 \times 10^{-10}$); at the lattice the symmetrized ring locks perfectly (1.00 against 0.97, higher in $35/36$). With many atoms the dictionary reproduces the cluster pattern, so every harmonic of an irregular pattern locks and only the p-th harmonic of a regular one does.

What is and is not tested here. The comparisons are between trained dictionaries, so nothing in them is entailed by the convolution identity; each direction could have come out the other way. The clustered rings count in both directions: rows 1 and 3 hold on them, and the removal and lock-order predictions that fail on them are counted against the criterion (App. S, (17) and (21)), not set outside its scope. What the experiment does not do is predict the lock order from the density alone, and we report that failure rather than rescue it: the order is set jointly by the density and by the configuration the optimizer reaches.

L THE TRAINED-SEED TEST ON THE FULL RESIDUAL STREAM

App. K trains inside the ring plane, and inside a 16-dimensional subspace containing it, because that is where the proposition's setting is closest. This section runs the same test closer to where a real SAE lives: on Gemma-2-2B's whole 2304-dimensional residual stream, with the ring probes only a minority of the training data (`code/g25_fullstream_seeds.py`).

Design. The pool is $60{,}000$ residual vectors from the same public 10k-document Pile slice used in App. E, taken at the layer's own hook point, plus the ring probes of App. I repeated until they are 30% of the pool—without that upsampling a dictionary of this size gives the ring almost no atoms, and we say so rather than hide it. Activations are rescaled so the mean norm is $\sqrt{d}$, the convention production SAEs are trained under. Dictionaries are ReLU+L_1 with M atoms trained from 12 seeds by Adam for $6{,}000$ steps at batch 256 (learning rate 3×10^{-3}, cosine schedule, initialization as in Table 14), with atoms that fired on no sample in the last 500 steps re-set, during the first half, to randomly drawn pool vectors. Only the ring probes' in-plane component differs between conditions: *natural* is the model's own, *symmetrized* rotates each item cluster about the ring's center until its circular-mean angle sits on the regular p-gon and rescales it about that center to the ring's average mean radius (so each cluster keeps its angular profile, which App. K's translation does not), and *tilt* q resamples the symmetrized probes with weight $1 + 2r\cos(q\theta_i)$ at the item's p-gon angle θ_i, $r = 0.35$. Every off-plane coordinate, and all $60{,}000$ general-text vectors, are identical across conditions.

The statistic. Each atom's decoder column is projected into the ring plane and enters weighted by its squared in-plane norm fraction $\|P_V w\|^2/\|w\|^2$, the weight the cost kernel gives it; the reported A_k uses the six atoms with the largest such fraction; the soft variant over all atoms, weighted the same way, gives 0.845 against 0.660 below the lattice (natural higher in 21 of 24), 0.677 against 0.328 for the tilts, and the ring plane above the random planes in 18 of 18 (`g25_analysis.py --key A`). The effective number of atoms the dictionary puts in the plane is the participation ratio of those weights. The null of $\max_{k<p} A_k$ is simulated per setting with that setting's own weights.

Results ($M = 512$, $\lambda \in \{0.03, 0.05, 0.1, 0.2\}$, six (ring, layer) cells at layers 7 and 12; 84 runs: 24 natural, 24 symmetrized and 24 tilts ($q \in \{2, 3\}$ at $\lambda \in \{0.03, 0.1\}$), and 12 re-runs of the natural and symmetrized $\lambda = 0.1$ settings, because the control below was added after the $\lambda = 0.03$ and 0.1 runs; the $\lambda = 0.05$ and 0.2 runs store it directly). Below the lattice the natural ring is a median 0.789 against 0.620 symmetrized, higher in 20 of 24 (cell, λ) settings (Wilcoxon $p = 8.9 \times 10^{-4}$) and in 5 of the 6 cells; the exception is the clock-hour ring at layer 12. An installed harmonic locks its own order or its alias, 0.631 against 0.323 at the other orders below p, and it is the strongest order in 12 of 24 tilts (chance 5.6). Agreement at the lock order rises with λ (median Spearman $+0.90$, positive in 4 of 6 series). The effective number of in-plane atoms is a median 15.0 (8.5–26.1) over all runs.

Two controls. The first is the symmetrized ring, as in App. K. The second asks whether the ring plane is special at all: the same trained dictionaries, read in three random 2-planes orthogonal to it, give a median 0.524 against 0.803 in the ring plane over the 18 settings the control was run on ($\lambda \geq 0.05$; the $\lambda = 0.03$ runs predate it), with the ring plane higher in 16 of them. That 0.52 is the background this setting cannot remove, and it is our explanation, not a measurement, of why the symmetrized side sits at 0.62 rather than at the 0.43–0.48 of a uniform null: in the plane we can make the *whole* angular density p-fold symmetric, but in the residual stream we can only do that to the ring's own probes, and the $60{,}000$ general-text vectors keep whatever angular structure they have in that plane. The criterion reads the density that is actually there, and after symmetrization that density is not uniform. 18 of the 24 natural settings clear their own null against 12 symmetrized. The consequence is a smaller margin, which is what we measure.

M WHICH CODING RULES INHERIT THE CRITERION

Propositions 1–3 hold for any isometry-invariant cost, while Proposition 4 and the error bound of Proposition 6 use the ℓ_1 kernel (§3). Production SAEs use TopK or JumpReLU, so what the criterion says about them is a fair question, and the answer has two parts.

The objective transfers, through one factor only. Flatness of J is a symmetry statement: if the coding rule commutes with rotations of the dictionary, the whole orbit has the same value whenever

the density is invariant. Every rule here is rotation-equivariant in that sense, so the rule enters only through $\hat{\bar{c}}_k$. Table 11 evaluates $\bar{c}$ for the same configurations under each rule, and one principle organizes it: *a kernel is flat exactly where the code reconstructs at no angular cost*. ReLU+L_1's kernel is never flat (0.115–0.320 on irregular configurations, 0.024–0.063 on regular ones), because the penalty is paid even where reconstruction is exact—the ℓ_1 term is what gives the kernel its angular structure. JumpReLU's collapses to 7×10^{-16} on a regular hexagon, whose live atoms reconstruct exactly, but is strongly structured (0.31–1.85) on irregular configurations. TopK is decided by its sign constraint. The selected atoms reconstruct a point exactly when they *bracket* it; with signed coefficients any two independent atoms span the plane, the residual is identically zero, and the kernel measures 10^{-29} for every configuration, so no density can pin it. Deployed TopK SAEs gate through a ReLU (Gao et al., 2025), and a non-negative pair spans only the cone between the two atoms: on a regular m-gon with $m \geq 3$ the two largest correlations always bracket, so the kernel is again machine-zero, but on generic irregular configurations (the random ones of Table 11), points fall outside the selected cone and the kernel (0.155–0.500) is comparable to ReLU+L_1's; at TopK-2's own trained seeds it is nearly flat (App. P). We flag this because it is easy to get backwards: TopK's angular structure exists only away from the polygon, and only because of the sign constraint.

Only one rule's seeds agree where the objective is pinned. We then trained SAEs under each rule on data whose angular density we control: eight densities (a uniform control; the regular hexagon and 12-gon; the hexagon tilted at $q \in \{1, 2, 3\}$ and the 12-gon at $q \in \{1, 2\}$, amplitude $r = 0.35$; items are wrapped Gaussians of width 0.05), three dictionary sizes $M \in \{4, 6, 9\}$ spanning $M < p$, $M = p$ and $M > p$ on the hexagon (all $M < p$ on the 12-gon), and 12 seeds each with the data resampled per seed, so agreement across seeds reflects the population objective rather than a shared finite-sample fluctuation. Data are $32{,}768$ points of the ring embedded in $\mathbb{R}^8$ with isotropic noise 0.03; training is Adam at 3×10^{-3}, batch $1{,}024$, $4{,}000$ steps, dead atoms redrawn in random directions during the first half, with $\lambda = 0.2$ for ReLU+L_1, $\lambda_0 = 0.005$ and a learned threshold (initialized at 0.02, sigmoid straight-through bandwidth 0.1) for JumpReLU (`code/g13_coding_rules.py`). Identifiability is measured without reference to the theory: each seed's configuration has a complex order parameter $Z_k = m^{-1} \sum_j w_j e^{ik\alpha_j}$ (atoms weighted by firing frequency) at k_{test}, the density's harmonic (q for a tilted ring, p for a regular one, 1 for the uniform control), whose phase lives on a circle of circumference $2\pi/k$ and is therefore already quotiented by the configuration's own symmetry; A is the circular concentration of that phase across seeds (the median over cells is reported), with a Rayleigh test. A is a resultant length, so App. C's normalized jitter is $\sqrt{-2 \log A}\,\sqrt{12}/2\pi$: $A = 0.98$ is a jitter of 0.11, and chance at 12 seeds, $A \approx 0.26$, about 0.9 (that sweep trains ReLU+L_1 at $\lambda = 0.3$, this one at 0.2).

ReLU+L_1 behaves exactly as the criterion says: $A = 0.980$ on densities carrying a pinning harmonic against 0.223 on the uniform control, significant in 20 of 21 cells and 0 of 3 controls (Mann–Whitney $p = 0.0010$). It holds in the under-complete regime too—at $M = 4$ against $p = 6$, where seeds must choose *which* clusters to claim, $A = 0.999$ and all four cells are significant, and on the 12-gon, all nine of whose cells are under-complete, A is a median 0.90 (8 of 9 significant; the plain 12-gon at $M = 9$ is the one miss of the 21)—and the median over each dictionary size, pooling $p = 6$ and 12, is 0.980 at $M = 6$ and 0.970 at $M = 9$. The other four rules do not: A between 0.355 and 0.450 on pinned densities, no rule separating pinned from flat at better than $p = 0.20$. Three of the four—TopK-1 and both JumpReLU rules—have structured kernels on generic irregular configurations (Table 11; TopK-1 also at every one of its own trained seeds, App. P) and pinned free-atom optima, so the objective is pinned and training does not settle where it says: for those rules harmonic selectivity is training-mediated, through the optimizer or the amortized encoder, which these runs do not separate (App. P). TopK-2 is the different case: evaluated at its own trained seeds, its kernel is a median 5×10^{-4} of the zero code and its largest Fourier coefficient is below the smallest any ReLU+L_1 seed reaches at 184 of the 288 seeds, so those seeds have reached the exactly reconstructing optimum of App. P and their disagreement is the flatness the criterion predicts (the other 104, whose kernel is in ReLU+L_1's range, are reported there). (On a regular m-gon the TopK-2 kernel is flat, and the JumpReLU kernels are flat on the regular hexagon, Table 11; what the experiment shows is that the configurations TopK-1 and JumpReLU reach are neither regular nor flat—JumpReLU's seeds have median gap CV 0.34, regular in 5 and 9 of 288 (tied, per-atom), and a largest kernel harmonic of 0.097 and 0.072 at $\theta_0 = 0.05$, against ReLU+L_1's 8.4×10^{-3}—and that their seeds do not agree there.) One caveat on the per-cell counts. Between 5 and 10 of 21 pinned cells are individually significant for these rules, and for TopK-2 one of the three *uniform*-density controls is too—a

rate above the 0.05 the test is calibrated at (we verified the implementation returns 0.049 on $20{,}000$ draws of 12 uniform phases). Nine of the eleven significant TopK-2 cells (ten pinned, one control) sit at $k_{\text{test}} \in \{1, 3\}$, where the statistic is a low-order moment of the atom cloud (at $k = 1$, its centroid) and a tilted density moves it without pinning the configuration. We therefore read the pinned-versus-flat contrast, which is the pre-specified comparison, and report the per-cell counts descriptively. Two re-readings apply the same scrutiny to every rule (`code/g79_rule_reanalysis.py`). Dropping every rule's $k_{\text{test}} \in \{1, 3\}$ cells (9 of 21) leaves ReLU+L_1 at 0.977 (11 of 12 significant) and the others at 0.236–0.417 (2 to 5 of 12). The three controls, read at $k = 1$, are kept; against them ReLU+L_1's contrast is exact $p = 0.0044$ and the others' $p \geq 0.18$. The set $\{1, 3\}$ was chosen after seeing where TopK-2's significant cells fall, so this re-reading is post hoc; the third below chooses no harmonic. Other drops read the same way: without $k = 1$ (the six $q = 1$ tilts) ReLU+L_1 is at 0.986 (14 of 15; exact $p = 0.0025$) and the others at 0.27–0.36 ($p \geq 0.24$), without $k \in \{1, 2\}$ at 0.986 (8 of 9; $p = 0.009$) against 0.23–0.40 ($p \geq 0.10$); without $\{1, 2, 3\}$ only the six untilted polygon cells remain, and there TopK-1, which locks the plain hexagon's lattice (below), matches ReLU+L_1 (0.70 against 0.57, both $p = 0.024$). And since Adam's per-coordinate steps are not rotation-equivariant while every seed's plane was embedded along the same two coordinate axes, we reran all 120 cells with each seed's plane embedded along a fresh Haar-random 2-plane of $\mathbb{R}^8$ (`--rand-basis`), which makes the pipeline rotation-equivariant in distribution: ReLU+L_1 gives 0.941 against 0.106 on the uniform control (18 of 21 significant; exact $p = 0.0010$) and the other rules 0.329–0.493 against 0.204–0.354 (exact $p \geq 0.06$), TopK-2's uniform-density agreement falling from 0.470 to 0.335. Read at every harmonic the pinned cells use ($k \in \{1, 2, 3, 6, 12\}$, with the firing weights these runs store), ReLU+L_1's three controls stay at 0.07–0.49 (median 0.22), so its contrast does not rest on reading them at $k = 1$; read the same way, the other rules' contrasts clear $p < 0.05$ at one of the five harmonics at most (TopK-2 at none; smallest $p = 0.008$, TopK-1 at $k = 12$), ReLU+L_1's at all five ($p \leq 0.008$; exact, 21 cells against 3 controls). The contrast belongs to the objective, not to the coordinate frame. A third re-reading removes both of the statistic's own choices (`code/g84_rule_alignment.py`): the firing weights, which could carry the density's harmonic into the phase (atoms on heavy clusters fire more), and the harmonic k_{test} itself. For each pair of seeds we find the rotation that best aligns their *unit*-weight atom-angle histograms (von Mises kernel, $\kappa = 12$) and score $c(0)/\max_\delta c(\delta)$, how nearly the best alignment is no rotation at all, averaged over pairs and compared with 200 draws that rotate each seed uniformly. ReLU+L_1 seeds align at no rotation in 16 of 21 pinned cells and 0 of 3 controls (median score 0.51 against -0.04; exact $p = 0.015$), the other rules in 0 to 5 (exact $p \geq 0.15$); with random bases, 16 of 21 against 0 to 3 (exact $p = 0.005$ against $p \geq 0.30$). The pooled contrast of the hyperparameter grid (defined below, Table 12) merges its two installed harmonics, $q = 2$ and $q = 3$, and a split by harmonic is only fair if it is applied to every rule, so here it is for all five:

	$q = 2$			$q = 3$			pooled
rule	pinned	uniform	p	pinned	uniform	p	p
ReLU+L_1	0.976	0.208	0.014	0.998	0.310	0.014	7.8×10^{-5}
TopK-1	0.196	0.140	0.010	0.229	0.431	0.99	0.54
TopK-2	0.294	0.345	0.56	0.562	0.245	3×10^{-4}	0.011
JumpReLU, tied	0.304	0.255	0.098	0.384	0.222	2.5×10^{-4}	2.2×10^{-4}
JumpReLU, per-atom	0.305	0.220	0.005	0.294	0.222	0.045	1.7×10^{-3}

(One-sided exact Mann–Whitney over grid points, ReLU+L_1's 8 distinct runs entered once each; grid points are settings rather than samples, so these are descriptive.) Only ReLU+L_1 separates pinned from flat at both installed harmonics with a margin: its pinned agreement is ≥ 0.976 at both, against ≤ 0.562 for every other rule. Per-atom JumpReLU is also below 0.05 at both halves, but at 0.305 and 0.294 against controls of 0.220 and 0.222; TopK-2's and tied JumpReLU's pooled p is carried by $q = 3$, and TopK-1, not significant pooled, separates only at $q = 2$ (and it runs the wrong way at $q = 3$). For TopK-1 and tied JumpReLU that is what a rule whose training does not find the pinned optimum looks like; TopK-2's optimum is flat (App. P), so its $q = 3$ separation is the low-order-moment effect of the remark above—the weight-free statistic aligns TopK-2's seeds in none of the three $q = 3$ cells of Table 1's design and ReLU+L_1's in all three—and this is the uniform statement of that remark rather than a split applied where it was convenient. TopK-2's $p = 0.011$ in Table 1 is therefore not evidence against the degeneracy of §4.3; App. P gives the reason its seeds cannot lock.

The same three rules on the model's own ring. (Tied JumpReLU was not trained on the ring; TopK-2, the flat-objective control, is below.) The comparison above is on densities we build. Repeating App. K's test with the coding rule swapped—same eleven cells, same dictionary sizes, same twelve seeds, ReLU+L_1 replaced by TopK-1 or JumpReLU with per-atom thresholds at $\lambda_0 \in \{0.005, 0.01, 0.02, 0.05\}$ (`code/g29_analysis.py`)—gives the same ordering on Gemma's own geometry. ReLU+L_1: natural 0.667 against 0.490 symmetrized, 130 of 152 settings, 90 above the null against 18, tilts 0.739 against 0.293. JumpReLU: 0.487 against 0.470, 88 of 152 (Wilcoxon $p = 5 \times 10^{-3}$), only 11 settings above the null against 6, tilts 0.485 against 0.248—a detectable trace of selectivity and no usable separation, and its agreement does not grow with the penalty (median Spearman -0.30). TopK-1, which has no penalty and so one setting per (cell, M): 0.499 against 0.492, 20 of 38 (Wilcoxon $p = 0.25$), 6 of 38 above the null against 2 (a rate of 16%, above JumpReLU's 11 of 152, but with no natural-against-symmetrized margin), tilts 0.295 against 0.266—no selectivity. TopK-2 is the control this design lacked: its in-plane optimum is flat (App. P), so the criterion predicts no lock and no selectivity, and whatever agreement it shows is the statistic's, not the objective's (same cells, sizes and seeds, `g21 --rule topk2`, one setting per (cell, M); `code/g29_analysis.py`). Below the lattice it gives 0.471 against 0.449 symmetrized, natural higher in 27 of 38 (Wilcoxon $p = 0.02$, descriptive) but only 2 settings above the null against 1 (ReLU+L_1: 90 against 18). At an installed $q \geq 2$ it shows no selectivity: the installed order or its alias is the strongest below p in 9 of the 35 tilts with $\gcd(p, q) = 1$ (chance 8.8; ReLU+L_1 28) and 14 of the 76 others (ReLU+L_1 63), significant at q in 1 and 5. At $q = 1$, though, its seeds agree too: the installed $k = 1$ is the strongest order in 26 of 27 tilts ($A_1 = 0.68$, significant in 25), where ReLU+L_1's lock goes to $k = 2$. A first-harmonic tilt loads one side of the ring, the atoms follow the mass, and the configuration's first moment—its centroid—points at the heavy side whatever the objective's pinning; agreement there is the low-order-moment effect App. M flags on the constructed densities, not a pin. So the $q \geq 2$ selectivity of Table 2's row 3 is the objective's, and the $q = 1$ row, which fails for ReLU+L_1, is not a reading of the criterion at all. The rule a production SAE uses decides whether the criterion is visible in its dictionary, on a density nobody designed.

The objective under each production rule. §4.3 and Table 1 report where twelve *seeds* end up. That leaves two readings a reader cannot separate: either the criterion is a property of the ReLU+L_1 objective alone—in which case it says nothing about deployed SAEs, which use the other rules—or it holds for every rule's objective and training does not reach the pinned optimum, because the optimizer stops short or because the amortized encoder's objective is flatter than the exact code's. Proposition 3 is about an objective, so the question is answerable with no training: swap $\bar{c}$ for the rule's own cost kernel and convolve. `code/g37_rule_landscape.py` does this on App. I's 155 GemmaScope configurations and their natural and symmetrized densities, with each rule's kernel computed exactly (TopK gates through a ReLU, so its coefficients are non-negative and its kernel is angle-dependent; `code/e9_kernel_by_rule.py`). At $\lambda = 0.15$ and the thresholds shown, the natural-over-symmetrized depth ratio is a median $23.1\times$ for ReLU+L_1 ($155/155$ cells above 1; App. I's $23.0\times$ is the same comparison through the direct re-solve rather than the rule's kernel, which differs in the third digit), $13.9\times$, $7.3\times$ and $5.4\times$ for JumpReLU at $\theta = 0.2, 0.4, 0.6$ ($155/155$ each), $12.1\times$ for TopK-1 ($155/155$), $6.4\times$ for TopK-2 ($154/155$) and $2.4\times$ for TopK-3 (111 of the 119 configurations with at least three atoms). In every case a median ≥ 0.95 of the landscape's power below $4p$ sits below the lattice, where the symmetrized density has none.

As in §5.1 this is an *illustration*, not a test: projecting a density onto a regular p-gon zeroes the pinning harmonics, so a ratio above 1 follows from the identity, and the regular and random controls exceed 1 too ($8.1\times$ and $24.1\times$ for ReLU+L_1, the random angles an independent draw from App. I's $22.7\times$; $5.8\times$ and $11.1\times$ for TopK-1). What it establishes is the scope of the criterion, which is what §4.3's negative results leave open: the convolution and its flatness condition are statements about every objective here, not about one of them. Set beside the trained-seed counts: TopK-1's objective is pinned on the natural ring in 155 of 155 cells ($12.1\times$, though random angles give $11.1\times$ too), and its seeds clear the calibrated null in 6 of 38 settings. The two numbers come from different configurations (GemmaScope's, and the seeds' own), so the gap between them is indicative only; read as an amortization gap, it is where a production SAE would lose the identifiability its own objective could give it.

Is the contrast a hyperparameter accident? The pre-specified contrast has three uniform-density controls, so its floor is $p = 0.0005$ (complete separation); ReLU+L_1 scores 0.0010, one rank short of it, and the other rules 0.20–0.47. A contrast with three controls cannot rank the rules

finely. We therefore reran the hexagon with an installed harmonic $q \in \{2, 3\}$ ($r = 0.35$, $M = 6$, $d = 8$, 12 seeds, data resampled per seed), the plain hexagon (read at its lattice order $k = 6$) and a uniform density over a hyperparameter grid per rule (`code/g22_rule_hparams.py`): learning rate $\in \{3 \times 10^{-3}, 10^{-2}\}$, steps $\in \{4{,}000, 12{,}000\}$, dead-atom re-initialization at random or at the worst-reconstructed data points, and for JumpReLU the L_0 penalty $\lambda_0 \in \{0.002, 0.01\}$ and the straight-through bandwidth $\in \{0.1, 0.3\}$ (ReLU+L_1's λ is not varied, 0.2 throughout, nor TopK's k, so JumpReLU has four times the recipes in which to lock)—16 grid points per TopK and ReLU+L_1 rule and 64 per JumpReLU variant, pooled over the two installed harmonics (8 and 32 at each; no ReLU+L_1 atom dies, so its two re-initialization settings give identical runs and its 16 points are 8 distinct runs, each entered once in the tests), and the same grid trained once on the uniform density and read at both $k = 2$ and $k = 3$ as a control of matched size (so the pooled test counts each control run twice; the split above reads each harmonic's control once). Table 12 gives the result and Table 1 carries its last two columns. ReLU+L_1 locks at the installed harmonic at every grid point (0.973–0.999, all significant) and never on the uniform density (0.11–0.47); TopK-1 reaches $A \geq 0.8$ at none (0.15–0.70, one significant) and does not separate pinned from flat ($p = 0.54$), although on the plain hexagon it does lock the lattice (A at $k = 6$, 0.70–1.00) with 3–5 live atoms; TopK-2 separates them only through $q = 3$ (7 of 8 significant, against 1 of 8 at $q = 2$; pooled $p = 0.011$). Both JumpReLU variants separate pinned from flat ($p = 2.2 \times 10^{-4}$ tied, 1.7×10^{-3} per-atom) but reach $A \geq 0.8$ at only 6 of 64 and 1 of 64 grid points, all at $\lambda_0 = 0.01$ (median 0.39 and 0.32 there against 0.30 and 0.27 at $\lambda_0 = 0.002$), with 2–5 of the 6 atoms alive. Agreement is read from each seed's live atoms, so seeds that keep different numbers alive can disagree for that reason alone; on Table 1's cells the per-seed percentile of App. P, which compares no two seeds, puts ReLU+L_1 first by the same wide margin (0.05 against 0.40–0.63), though it orders the other four differently. Three of the 160 other-rule grid points reach ReLU+L_1's worst point (0.973). The reading of §4.3 stands with one refinement: JumpReLU's training can reach the pinned configuration, and on this grid does so more often at the larger of its two penalties—the direction in which the kernel grows with the penalty (App. T)—though on the model's ring its agreement does not grow with the penalty over four values (median Spearman -0.30, above), and at no recipe we tried does it do so reliably. Code: `code/g22_analysis.py`.

N Cross-seed consensus planes on a real LM

Proposition 1 suggests that the orbit, i.e. the subspace, is what can replicate across seeds where the coordinates inside it do not, which suggests discovering feature planes by cross-seed consensus rather than by decoder geometry alone. We tried this on the eight replicate TopK SAEs of App. G (16,384 latents, $k = 32$, Pythia-160M layer 6, 300k held-out activations). For 4,000 sampled live features we formed the decoder-cosine cluster (≥ 0.30, up to 8 atoms), kept the clusters that are planar (top-2 principal directions carrying $\geq 70\%$ of the cluster's decoder energy, every atom at least 60% in-plane; any two-atom cluster meets both), and asked how many of the other seven seeds span the same plane to within 15° of mean principal angle with an atom count within one of the cluster's; a plane survives if at least four do.

Twenty-seven clusters out of 4,000 survive. Their median cross-seed subspace error is 11.5° and their configuration phases are very stable (median jitter 0.022), but they are mostly atom pairs, not tilings: 19 of 27 have $m = 2$ (median nearest-neighbor decoder cosine 0.35), and the median fraction of the cluster's active samples on which two of its atoms fire together is 0.046. The angular density inside them is strongly concentrated ($|\hat{\rho}_1| = 0.82$ on average), which is the pinned regime, so the criterion predicts exactly the stability observed, but with no spread in the predictor there is no test. We report this as the reason production-scale validation of the criterion is currently blocked on manifold discovery, and as the motivation for using an externally identified manifold in §5.

O Orbit claims against atom claims, and the criterion on unprompted data

Two experiments on dictionaries trained on Gemma-2-2B's residual stream with nothing constructed in the pool: no ring probes, no symmetrization, no 2-plane restriction. One is a negative and one is a positive with a sharp boundary; we report both.

Shared setup. Twelve ReLU+L_1 dictionaries of $M = 512$ atoms are trained by Adam for 8,000 steps at batch 512 on 120,000 Pile residual vectors—general-text vectors of the pools harvested for App. L, which trains on 60,000 of them, with the probes *dropped*, scaled and trained as there (Table 14)—at each of layers 7 and 12 and each $\lambda \in \{0.1, 0.2\}$, 48 in all (`code/g38_probefree.py`, stage `train`).

O.1 The graded prediction on unprompted data: a negative

Question. Every positive real-model result in the paper either trains on a pool containing the ring probes or works inside a 2-plane. Does the graded form of the criterion—does $|\hat{\rho}_k||\hat{\tilde{c}}_k|$ order where the seeds agree—survive on data the model produces unprompted?

Design. In each of 76 (plane, layer, λ) cells—19 planes at two layers and two λ—we compute the data's angular density in the plane, each point weighted by its in-plane norm (a proxy: the $\lambda = 0$ reduction of App. A weights by its square), the exact 2-plane cost kernel of each seed's top-16 in-plane atoms, and the cross-seed concentration A_k of the configuration's k-th phase (every atom, weighted by its squared in-plane norm fraction), for $k = 1 \ldots 24$. The planes are the three ring planes the model supplies (days, months, clock hours), eight random planes of the full 2304-dimensional space, and eight random planes inside the data's top-32 principal subspace, each read at two layers and two λ. The criterion is not a statement about rings, so it should hold in all of them. The dictionary is not ignoring these planes: the participation ratio of the atoms' in-plane norm fractions is a median 172 of 512 atoms (range 101–266).

Result, and why the raw number is the null. Spearman$(|\hat{\rho}_k||\hat{\tilde{c}}_k|, A_k)$ is a median $+0.512$, positive in 74 of 76 cells. That is not evidence; that is the decay both share: both sides fall with k—the density's harmonics decay, the kernel's decay, cross-seed agreement decays—so a positive rank correlation is what two decaying sequences produce by construction. Two controls decide it (`code/g38_analysis.py`).

Partial. With the rank of k partialled out of both sides the correlation is a median $+0.054$, positive in 44 of 76 cells ($+0.045$ for $k \geq 2$).

Cross-plane. If the prediction is about a plane's own density and its own trained configuration, a plane's predictor must fit its own agreement spectrum better than it fits another plane's. This control holds the decay fixed by construction. A plane's predictor scores $+0.054$ against its own A and $+0.030$ against other planes' at the same layer and λ; its own is higher in 39 of 76, a coin flip (Wilcoxon $p = 0.46$).

Neither factor alone rescues it: the density alone gives $+0.092$ (median over the 72 cells where it is defined) and the kernel alone -0.009, both partialled on k. The same two controls applied to the clustered-ring statistic of App. K leave it standing ($+0.258$ partialled, own setting over others 90 of 149, Wilcoxon $p = 4.3 \times 10^{-3}$), which is what separates a prediction from a shared trend.

What this does and does not say. The prediction is not measurable here, and we do not know which of two readings is right: the criterion may be false at this scale, or its ingredients as we estimate them—a density read at one layer in one plane, a kernel from 16 atoms out of 512—may be too noisy to carry the ordering. The experiment does not separate them, so the main text reports the failure rather than a diagnosis. It is the same blocker as App. N: without an identified manifold there is nothing to read the density on.

O.2 Intervening on an atom against intervening on its orbit: a positive with a boundary

Question. Proposition 1 motivates attaching interpretability claims to orbit invariants rather than to atom identities. Here we test that on our own dictionaries (`code/g39_orbit_intervention.py`).

Design. For each seed (layer 12, $\lambda = 0.1$), the weekday-selective atoms are found from that seed's own decoder directions by an F-statistic across the seven day clusters of the probe set; nothing shared between seeds enters, so neither intervention is defined by a common plane and the comparison is not circular. Two interventions project the residual stream at layer 12 off a subspace, at every position: *atom* removes the seed's single most selective atom direction (an atom claim), *orbit*

removes the 2-plane of the top-2 principal directions of its top 8 selective atoms' decoder vectors (an orbit claim; a decoder-only construction, unlike App. Q's probe-derived plane, which released dictionaries need: there this decoder-only object transfers at only $+0.161$). Two controls remove a random 1-direction and a random 2-plane of the same rank. The read-out is the change in the log-probabilities of the seven weekday tokens at the final position of 63 prompts whose natural continuation is a weekday ("`Monday, Tuesday, Wednesday,`" → "`Thursday`"; seven days × nine templates). The model's top weekday is the correct one in 100% of them unablated. Each seed gives a 63×7 effect matrix and the statistic is the correlation between two seeds' matrices, over the 66 pairs.

Where the seeds share a subspace. Run on dictionaries trained on the same pool with the ring probes mixed back in at the fraction §5 uses (1,680 probes repeated 15 times into 60,000 general vectors, 29.6% of the pool), the two claims separate cleanly:

	atom	orbit	random 1-dir	random 2-plane
cross-seed correlation (median)	-0.102	$+0.692$	$+0.010$	-0.018
IQR	$[-0.39, +0.33]$	$[+0.46, +0.79]$	$[-0.12, +0.13]$	$[-0.15, +0.10]$
pairs above 0.9 / below 0.5	5 / 55	9 / 21	0 / 66	0 / 66
mean $\lvert\Delta \log p\rvert$	1.807	1.718	0.048	0.049

The two interventions are the same size—38× and 35× their matched random control—so the atom claim is not failing because it does nothing. It is failing because what it does differs from seed to seed. And the seeds do agree on a subspace here: the largest principal angle between two seeds' selective planes is a median 23.1°, and each plane sits 38.7° from the model's own ring plane, while the effects of their most selective atoms do not agree (the atom column above). That is the contrast Proposition 1 motivates, in one line: the subspace replicates, the coordinates in it do not. These dictionaries have seen the probes, so this arm shows the atom/orbit contrast under a shared subspace; it is not evidence that a shared subspace arises unprompted.

Where they do not. On the probe-free dictionaries of the previous subsection the same procedure gives orbit $+0.069$ and atom $+0.016$, neither distinguishable from the random controls, with effects of 0.192 and 0.199—four times the controls but a tenth of the arm above. The reason is visible in the diagnostic: the seeds' selective planes are a median 87.9° apart and 87.3° from the ring plane, where two random 2-planes in 2304 dimensions have a largest principal angle of 89.5° (median over 3,000 pairs, 2.5th percentile 88.2°). The selector finds almost no shared subspace (the median pair sits just inside the random 2.5th percentile), so there is little for the orbit claim to be about, and it reproduces no better than the atom claim.

Reading. The recommendation is conditional, and the condition is measurable in advance from the seeds themselves. Where the seeds put a common subspace on a concept, ablating that subspace reproduces across seeds and ablating an atom does not; where they do not, neither claim reproduces. What fails is manifold discovery by cross-seed consensus—the blocker of App. N; App. R finds planes from token identity instead—not the statement about what a claim should attach to once a manifold is found.

P WHICH CONFIGURATIONS EACH RULE'S OPTIMUM ACTUALLY IS

§4.3 reports where twelve seeds of each rule end up. A sharper question is what each rule's optimum is (the best of 26 starts; the search does not certify a global one), because a regular m-gon couples only to harmonics in $m\mathbb{Z}$ and, for the rules whose kernel vanishes on the polygon (Table 11), to none at all. The objective restricted to a plane is computable, so the question is a search rather than an experiment (`code/g41_free_atoms.py`).

Design. Over 21 (density, m) cells—the weekday, month and clock-hour densities measured on Gemma at layer 7, the 48-point clock density at layers 12 and 20, two single-harmonic controls, and $m \in \{3, 4, 6\}$—we minimize J over all m atom angles (Nelder–Mead with at most 6,000 iterations, $\mathrm{xatol} = 10^{-5}$ and $\mathrm{fatol} = 10^{-14}$, from the regular m-gon, a copy with Gaussian noise of 0.05 rad added and 24 random starts, the first angle fixed). Every candidate is scored at its *own* best rotation, so the search is about the configuration's shape and a rigid rotation cannot flatter one candidate over

another, and the regular m-gon is always among the starts, so a reported optimum is never worse than the configuration in question. At the minimizer we read the gap coefficient of variation (is it regular?) and the depth of the rotation landscape in units of the zero code's cost (is it pinned?).

rule	gap CV	regular	reconstructs exactly	pinned at the optimum	landscape depth
ReLU+L_1 ($\lambda = 0.2$)	0.118	$4/21$	$0/21$	$\mathbf{21/21}$	3.1×10^{-2}
TopK, $k = 1$	0.229	$2/21$	$0/21$	$\mathbf{21/21}$	1.1×10^{-1}
TopK, $k = 2$ (non-neg.)	0.005	$13/21$	$\mathbf{21/21}$	$\mathbf{0/21}$	5.5×10^{-8}
JumpReLU ($\theta_0 = 0.05$)	0.028	$10/21$	$0/21$	$19/21$	3.6×10^{-2}
JumpReLU ($\theta_0 = 0.30$)	0.063	$8/21$	$0/21$	$\mathbf{21/21}$	4.8×10^{-2}

Reading. Four of the five rules have a pinned optimum, and it is mostly not the polygon: the minimizer's gaps are uneven (median CV 0.03–0.23; regular, CV < 0.02, in 2–4 of 21 cells for ReLU+L_1 and TopK-1 and 8–10 for JumpReLU) and it beats the best-rotated regular m-gon by a median 0.7–11% in objective. For those rules the polygon's flatness cannot explain the disagreement, and §4.3's reading stands: a pinned optimum exists and training (the optimizer or the amortized encoder) usually fails to reach it.

TopK at $k = 2$ is the exception, and the criterion predicts it. Its optimum reconstructs the density's support exactly ($J^* \sim 10^{-17}$, at the numerical floor in 21 of 21 cells), so the landscape there is flat to 5.5×10^{-8} of the zero code—nearly six orders below every other rule; the residual is mass that a rotation moves onto angles outside the support, which the optimal pair need not bracket. Its seeds therefore have no objective-level reason to agree, and their disagreement in Table 1 is the criterion's own prediction rather than a failure of it. This is the same mechanism as the regular m-gon of Corollary 1, reached by the rule's own optimum instead of by construction: a code that pays no angular cost leaves the rotation free.

On Table 1's own densities. The search above uses the models' densities; the trained SAEs of Table 1 saw the eight constructed ones, at $M \in \{4, 6, 9\}$. Re-run on the seven pinning densities of that design (`g41_free_atoms.py --density-set table1`), the optimum is pinned in 21 of 21 cells for ReLU+L_1 and TopK-1 and 19 of 21 for JumpReLU at either reference threshold ($\theta_0 = 0.05$ and 0.30), the two exceptions the hexagon and 12-gon densities at $M = 6$, whose optimum is the regular hexagon, where the JumpReLU kernel collapses (Table 11); TopK-2's optimum reconstructs exactly in 21 of 21, flat there to 4×10^{-17}, since these densities have full support. So on the very densities its seeds were trained on, each production rule's objective has a pinned optimum (TopK-2 excepted) that its seeds do not agree on. Whether the gap is the optimizer's or the amortized encoder's—the search minimizes the exact code's objective, training the encoder's—these runs do not separate.

Where the trained seeds sit, and the two statements reconciled. Two things this paper says about TopK-2 can be heard as a conflict: App. M reports the objective pinned under the TopK-2 kernel (6.4× on GemmaScope's irregular configurations, and 1.5×10^{-1} on random ones, Table 11), and this appendix reports the rule's optimum flat. Both are statements about configurations nobody trained with TopK-2, where the non-negative pair code does leave a residual, and the rule's own minimizer, where it does not. What decides which mechanism the *seeds'* disagreement reflects is where TopK-2's own trained seeds sit, and the stored configurations of Table 1's cells answer it (`code/g51_trained_topk2.py`; 288 seeds in 24 cells, each rule's kernel evaluated at its own seeds' angles). The TopK-2 seeds mostly reach the degenerate optimum: at the reached configuration the in-plane kernel is a median 5×10^{-4} of the zero code and its largest Fourier coefficient over $1 \le q \le 12$ a median 2.7×10^{-4} (absolute, at the design's unit radius, where the zero code costs $1/2$; Table 1's kernel column uses the same units), thirty times below the ReLU+L_1 seeds' 8.4×10^{-3} in the same cells, and 184 of 288 seeds sit below the *smallest* value any ReLU+L_1 seed reaches (1.8×10^{-3}; 175 below 10^{-3}, where no ReLU+L_1 seed is). (A fixed threshold of 10^{-2} is not a separation: 159 of the 288 ReLU+L_1 seeds fall below it too.) TopK-1's seeds sit at configurations with a structured kernel ($\max_q |\hat{\tilde{c}}_q| \ge 10^{-3}$), 288 of 288. For those 184 seeds the disagreement is the flatness. The rest, 104 seeds (32 at $M = 4$, 23 at $M = 6$, 49 at $M = 9$, seven of them on the uniform control), are irregular where the others are not (median gap CV 0.38 against 0.09 over all TopK-2 seeds), and their kernel is in ReLU+L_1's range (median $\max_q |\hat{\tilde{c}}_q| = 8.6 \times 10^{-3}$), so their disagreement a training account (optimizer or encoder) has to carry. TopK-1 is the opposite case:

its kernel is structured at every one of its 288 trained seeds (median $\max_q |\hat{c}_q| = 0.133$, minimum 0.027), so on the pinning densities its disagreement is training that does not reach the exact code's pinned optimum, which is what §4.3 says of it.

Do the seeds sit at their own landscape's minimum? Cross-seed agreement infers this; it can be measured directly (`code/g85_seeds_vs_optimum.py`). For each stored seed of Table 1's 21 pinned cells, the rule's exact in-plane kernel is built at that seed's own live in-plane atoms' angles and correlated with the cell's density, which gives J over rigid rotations of the configuration the seed reached; we record the fraction of rotations with lower J than the one it was trained into (0: at the landscape's minimum; 0.5: no better than a random rotation). JumpReLU's kernel is taken at its reference threshold 0.05, since trained thresholds were not stored.

	ReLU+L_1	TopK-1	TopK-2	JumpReLU, tied	JumpReLU, per-atom
median percentile (252 seeds)	**0.05**	0.63	0.40	0.52	0.56
seeds in the bottom 5%	**50%**	14%	15%	16%	10%
cells whose median seed is in the bottom 5%	**12/21**	2/21	0/21	1/21	0/21
landscape depth (zero code)	0.046	0.12	0.0036	0.056	0.037

ReLU+L_1's seeds are at or next to their landscape's minimum; the other pinned rules' seeds are where a random rotation would put them (TopK-1's two cells at the minimum are the plain hexagon at $M = 4, 6$), on landscapes of comparable depth (0.8–2.6 times ReLU+L_1's); TopK-2's landscape is thirteen times shallower than ReLU+L_1's. This is the direct form of Table 1's reading. ReLU+L_1's exceptions are nine cells: eight far from the minimum, the six $q = 1$ tilts and the 12-gon at $M \in \{6, 9\}$, and one just outside the bottom 5%, the 12-gon tilted at $q = 2$ with $M = 4$ (median percentile 0.065). The eight are informative: at the $q = 1$ tilts its seeds agree ($A = 0.63$–0.999) on covering the ring's *light* side, a rotation the exact code's kernel ranks among its worst (median percentile 0.68–0.99). The exact code at $\lambda = 0.2$ is the objective the free-atom search minimizes, not the trained encoder's, and these are the cells where the two disagree—a trained ReLU+L_1 dictionary can be pinned, and agree across seeds, at a phase the exact objective does not choose. Without those six cells Table 1's contrast stands (0.986 against 0.27–0.36; App. M).

And it is not special to $k = 2$. Re-running the search at $k \in \{2, 3, 4, 8\}$ over 42 (density, m) cells each—m from 3 to 12 (`code/g41_free_atoms.py --tag _topk_sweep`, which uses the exact non-negative code over the selected set: in a plane that code is a projection onto a convex cone, so it is the minimum over the pairs of the selected atoms):

	TopK-2	TopK-3	TopK-4	TopK-8
optimum reconstructs exactly	42/42	42/42	42/42	42/42
optimum is pinned	0/42	0/42	0/42	0/42

The reason is elementary once stated: the top-k selected set contains the top-2, so if two of the selected atoms bracket a direction the code reconstructs it, and a larger k can only make that easier. Every value of $k \geq 2$ we can search has an exactly reconstructing optimum with nothing left to pin on these densities—not on every ρ: the kernel vanishes on the whole circle only when the selected atoms bracket every direction (for $k = 2$, with selection by correlation as here, only on a regular m-gon, $m \geq 3$; any irregularity leaves a residual), and on a density without full support an exactly reconstructing optimum can still be pinned, by rotations that carry mass outside the cones it brackets. That suggests reading the seed disagreement of the $k = 32$ suite in §6 as the criterion's own prediction, with the caveat that in a $k = 32$ SAE most selected latents lie outside the plane, so the in-plane budget is smaller than k; it does not license the same reading for TopK-1, whose optimum *is* pinned (21/21).

Q THE RECOMMENDATION ON RELEASED DICTIONARIES

§6 tests the recommendation Proposition 1 motivates on dictionaries released by other groups. This section gives the design, the controls, and the numbers. The twelve share no latent index even at a shared width, so the comparison is geometric. Since the plane uses cluster means every dictionary shares, three controls ask what the dictionaries contribute: the *fixed* probe plane's effect correlates

only $+0.131$ with an assigned plane's, the dictionary-only rank-2 object transfers at $+0.161$, and the rank-1 shadow in the same span at $+0.784$ but unreliably (21 of 66 pairs below 0.5, against 9; details below).

Design. For each released SAE, the weekday-selective latents are ranked by an F-statistic across the seven day clusters of the probe set of App. I, computed on *that SAE's own decoder* projections (between- over within-cluster sum of squares of the probes projected on the latent's unit decoder direction). Nothing shared between SAEs enters the ranking except the probes, which is why the controls below are rank-matched constructions on *random* spans: shared probes projected onto random subspaces do not agree, so the probes alone, seen through a random span, which holds only about 0.03 of the ring at rank 64, do not make SAEs agree; the fixed probe plane below is the control for the ring itself. The same probes rank the latents and build the plane; we did not split them. Each SAE's random span is its own, so these controls share no input across SAEs; the fixed probe plane below is the control that does. Two objects are then formed per SAE—the atom from its decoder alone, the plane from the shared cluster means as seen through its decoder span, so the plane is not a decoder-only object and the controls below ask what the decoder contributes to it:

- the *atom*: the single most selective latent's decoder direction (rank 1);
- the *plane*: the top-2 principal directions of the day-cluster means as seen through the span of that SAE's top-m selective latents, $m = 64$ (rank 2). This is the 2-plane the SAE assigns to the concept. The top-2 principal directions of the selected *decoder vectors* are not that plane and do not replicate (74–88° apart); we report the distinction because it is easy to get wrong.

Each is ablated by projecting the residual stream off it at every position, and each is read against a control of the same rank: a random direction for the atom, and the same plane construction applied to a random rank-m span for the plane. The two are not alike: the plane's control shares the cluster means and is the harder one, while the atom's asks only whether its effect is consistent at all; no control selects a latent the way the atom is selected. The read-out is the change in the seven weekday tokens' log-probabilities at the last position of the 63 prompts of App. O; per SAE that is a 63×7 matrix, and the statistic is the correlation between two SAEs' matrices. Code: `code/g43_gemmascope_atoms.py` and `code/g42_public_saes.py`.

GemmaScope (Gemma-2-2B, layer 12). Twelve released dictionaries: widths 16k ($L_0 = 22, 41, 82, 176, 445$), 32k ($22, 40, 76, 155$) and 65k ($21, 38, 72$). The model's top weekday is the correct one in 100% of the prompts unablated.

top-m	4	8	32	64
ring plane inside the selective span, median	0.216	0.419	0.691	0.758
minimum over the twelve	0.048	0.143	0.359	0.467
random span of the same rank	0.0014	0.0027	0.0122	0.0267
ratio	151×	156×	57×	28×

	atom	plane	random rank-1	random rank-2
cross-SAE correlation (median)	-0.042	$\mathbf{+0.815}$	$+0.005$	$+0.034$
IQR	$[-0.13, +0.11]$	$[+0.67, +0.88]$	$[-0.12, +0.08]$	$[-0.54, +0.55]$
pairs below 0.5	59/66	9/66	66/66	46/66
mean $\lvert\Delta \log p\rvert$	0.421	1.310	0.036	0.142
beats its matched control	35/66	**58/66**	—	—
Wilcoxon p	0.27	2×10^{-11}	—	—

The assigned planes sit a median 42.1° from each other (43.9° within a width, 40.4° across widths) and 33.2° from the model's own ring plane, against 89° for the rank-matched random construction. Reported in float16; a bfloat16 replication of the nine 16k and 32k dictionaries gives $+0.843$ for the plane against $+0.841$, -0.032 for the atom against -0.021, and the plane above its control in 36 of 36 pairs at both precisions, so the precision is not carrying the result.

What the dictionaries contribute, and what the probes do. The assigned plane is built from the day-cluster means, which every dictionary shares, so a reader is right to ask whether the agreement is the probes' doing. Three measurements say it is not the probes alone. First, the *same* cluster means pushed through a random rank-64 span give planes 89° apart whose effects correlate $+0.034$, though

such a span holds only about 0.03 of the ring. Second, ablating the *fixed* probe-derived ring plane—one object, identical for all twelve dictionaries—produces an effect that correlates only $+0.131$ (minimum -0.093) with an assigned plane's, while assigned planes correlate $+0.815$ with each other and the fixed plane's own effect is larger (2.06 against 1.31): what the dictionaries agree on is not the object the probes handed them. Third, in the other direction, the purely dictionary-derived rank-2 object—the top-2 principal directions of the selected decoder vectors, with no probe quantity in it at all—transfers at $+0.161$ and is above its control for only 7 of 12 dictionaries ($p = 0.39$), consistent with those directions lying 74–88° apart, although its effect (0.61) is larger than the atom's (0.42). Neither ingredient alone carries the result. One more control separates rank from construction: the rank-1 principal direction of the cluster means seen through the same top-64 span, with the same object built through a random span as its matched control. The atom and the plane differ in rank, in whether a probe quantity enters, and in how many latents feed them, and this cell holds the last two fixed while varying the first. The rank-1 shadow does transfer—median $+0.784$, with an effect (0.84) between the atom's and the plane's, above its control for 52 of 66 pairs and 10 of 12 dictionaries (sign test $p = 0.019$)—but far less reliably than the plane: its interquartile range is $+0.001$ to $+0.904$ against the plane's $+0.666$ to $+0.880$, a third of its pairs (21 of 66) fall below 0.5 against 9 of 66 for the plane, and its matched control is already at $+0.346$, so its margin over control (median paired difference) is $+0.376$ against the plane's $+0.700$; both controls are bimodal (interquartile ranges -0.56 to $+0.64$ and -0.54 to $+0.55$), so these margins move with the draw, and with the fresh controls of the three-concept run below ($+0.03$ and $+0.20$) they are $+0.58$ and $+0.54$. On the Pythia suite the same cell gives $+0.264$ against the plane's $+0.547$ (8 of 8 seeds above control, margin $+0.311$ against $+0.417$). So both readings hold in part: a good part of what transfers is the concept's shadow in the dictionary's span, present already at rank 1, and on GemmaScope the second dimension is what makes it *stable*—the 2-plane stays above 0.5 in 57 of 66 pairs where the rank-1 shadow does in 45; on Pythia, at the seed level, the rank-1 shadow is at least as reliable (8 of 8 against 7 of 8). Read as a 2×2 of rank (1 or 2) by construction (cluster means through the span, or decoder vectors alone), GemmaScope gives $+0.784$ and $+0.815$ against -0.042 (the atom) and $+0.161$: the cluster means carry most of the transfer, the second dimension its stability.

The angles, and why they do not predict the transfer. The $33°$ and $42°$ above are *largest* principal angles, the worst-case measure (`code/g46_selection_null.py` records both; the smallest are $26.5°$ to the ring plane and $27.4°$ between assigned planes). On either measure the assigned planes are closer to the ring plane than to each other, which invites the inference that ablating the ring plane should transfer at least as well—and it does not, by a factor of six. The inference is worth taking seriously, because within the family of assigned planes the effect correlation *does* track the angle: over the 66 pairs, Spearman -0.78 against the largest angle and -0.78 against the smallest ($p < 10^{-13}$ with the 66 pairs, which share dictionaries, as units, so descriptive; angles span 19°–68° and correlations 0.28–0.99). A linear fit of that relation predicts $+0.845$ at $33.2°$, and the pairs actually lying between $29°$ and $37°$ have a median correlation of 0.845. The fixed ring plane at the same angle gives $+0.131$. So orientation is not what is missing. What separates them is that an assigned plane is the cluster means seen *through that dictionary's selective span*—a projection of the ring into the subspace the dictionary's own features occupy, holding a median 0.758 of the ring plane—while the fixed plane is the ring itself, with a 57% larger effect magnitude (2.06 against 1.31). Two planes at the same angle can be different interventions, and here they are. Across the twelve, an SAE's angle to the ring plane and its correlation with the fixed plane's effect move in the same direction (Spearman -0.41), but with 12 units that is not significant (two-sided $p = 0.19$) and we do not lean on it.

Three nulls for the containment statistic. A random rank-m span counts dimensions and nothing else, and the latents are ranked by an F-statistic across the very clusters whose top-2 plane is being contained, so two further nulls ask whether the *selection rule* is what does the work (`code/g46_selection_null.py`, 20 label permutations per dictionary):

rank m	weekday-selective	random span	permuted labels	month-selective	real > month
4	0.216	0.0015	0.0014	0.013	12/12
8	0.419	0.0031	0.0034	0.033	12/12
16	0.521	0.0069	0.0071	0.079	12/12
32	0.691	0.0141	0.0158	0.155	12/12
64	0.758	0.0283	0.0368	0.279	12/12
128	0.820	0.0540	0.0688	0.488	12/12

The random-span column here, the draw §6 quotes, is independent of the one in the first table above (0.0031 against 0.0027 at top-8, 0.028 against 0.027 at top-64); the two differ by at most 16% of the smaller (6% at top-64), and each median is below the analytic mean m/d ($d = 2304$) or at most 2% above it, as expected of the right-skewed law of a random span's containment, which is the check that the draw is not doing any work. Permuting the day labels reduces the selective span to about that random span (0.0034 against 0.0031 at top-8, 0.0368 against 0.0283 at top-64): the rule contributes little without the labels, so the dimension-counting null was not flattering the result after all. The hard null is the month-selective span, a real concept selected at full strength by the same rule on the same dictionary; the weekday plane sits inside it at 0.279 at top-64, so the weekday selection's margin over *that* is $2.7\times$ rather than $27\times$, and it holds for every dictionary at every rank. The margin shrinks with m because a wider span contains more of everything—at top-128 the month span already holds half the weekday plane—which is the reason to read the statistic at the rank the intervention uses and not at the rank that flatters it. §6 quotes top-64 against the month null.

Pairs are not independent units. The 66 pairs come from 12 dictionaries, each appearing in 11 of them, so a pair-level Wilcoxon overstates the evidence and we do not rest on it. At the dictionary level—one number per dictionary, its median advantage over the matched control across its 11 pairs—the plane is positive for 12 of 12 (exact sign test $p = 2.4 \times 10^{-4}$, the smallest value 12 units can give), the atom for 7 of 12 ($p = 0.39$; a count that moves with the control draw, App. R) and the decoder-only object for 7 of 12. The pair-level values are 58 of 66 ($p = 2 \times 10^{-11}$) and 35 of 66 ($p = 0.27$). §6 quotes the counts from this sign test and the p-values from the leave-one-dictionary-out jackknife (`code/g53_dyadic_jackknife.py`: the mean pair advantage over its leave-one-dictionary-out standard error $\left(\frac{n-1}{n}\sum_s(\bar{d}_{(s)} - \bar{d}_{(\cdot)})^2\right)^{1/2}$, referred one-sided to Student's t with $n-1$ degrees of freedom), which also respects the shared dictionaries: 4×10^{-4} for the plane and 0.10 for the atom. It is an approximation, not an exact test, so the unit count does not floor it; where it falls below the 2^{-n} floor of a sign test over its n units, as in several cells of Table 13, it reads as a large t, not as an exact tail probability.

The multi-seed suite of Paulo & Belrose (2026) (Pythia-160M, layer-6 MLP). Eight of the nine independently seeded 32k TopK dictionaries, $k = 32$, trained by that work on its own data in its own protocol (the eight in its released checkpoint set, `k32-sae-mlp-32k-seed2` to `-seed9`, read and ablated at the output of block 6's MLP). Over the 28 pairs the plane reproduces at $+0.547$ (IQR $+0.34$–$+0.62$, 24 of 28 beating the matched control) and the atom at -0.067 (14 of 28). The assigned planes here sit a median $62°$ from each other and $70°$ from the cluster plane (largest principal angle, against GemmaScope's $42°$ and $33°$): the eight seeds agree on a plane far less closely than the twelve GemmaScope dictionaries do, which is the geometric face of the weaker manifold. **What the seeds contribute here.** The two controls that make the GemmaScope result interpretable were run on this suite too, and this is the suite where the seeds are actually independent. Ablating the fixed probe-derived plane, identical for all eight, gives an effect correlating $+0.359$ (minimum -0.256) with an assigned plane's, against $+0.547$ between assigned planes: the shared probes' effect is more aligned with the transfer here than on GemmaScope ($+0.359$ against $+0.131$; correlations are not additive shares), which fits the weaker manifold and the wider angles above. The rank-1 shadow through the same span transfers at $+0.264$ (8 of 8 seeds above its matched control, $p = 0.004$), half the plane's median, where on GemmaScope the two medians are close. So on the independently seeded suite the plane transfers, the atom does not, the dictionaries contribute beyond the probes, and the margin of all three is smaller than on Gemma; we state it at that strength. The independence argument above applies here too, and we apply it: at the *seed* level the plane beats its matched control for 7 of the 8 seeds (exact sign test $p = 0.035$; median advantage $+0.49$) and the atom for 5 of 8 ($p = 0.36$). With eight units the exact sign test cannot go below $p = 3.9 \times 10^{-3}$, so 0.035 is what it can show and the pair-level 5.5×10^{-6} is not; §6 quotes the seed-level jackknife ($p = 0.016$ for the plane, 0.55 for the atom; 2.5×10^{-4} and 0.82 double-centered, App. R), a t-

statistic that no unit count floors. Both suites' unit-level tests are recomputed from the released per-pair correlations by `code/g47_unit_level_tests.py`, which also re-derives the dictionary-level values above and asserts they agree. The plane's effect (0.054) is *smaller* than its own control's (0.059), so its reproducibility is not an artifact of size. The model's top weekday is correct in 65% of the prompts unablated, which is why we read a correlation of effects and not an accuracy.

Why the two differ, and what that costs us. The day-cluster means at Pythia's layer-6 MLP output are not in cyclic order (largest gap 2.4 uniform gaps, top-2 variance 0.71) and the cluster plane sits inside the selective span at 0.048 against 0.011 random at top-8—$4.5\times$, against $156\times$ for Gemma at that rank; at the top-64 the intervention uses, 0.220 against 0.089, $2.5\times$ against Gemma's 27–$28\times$ (the two random draws). (Every containment ratio in this appendix names its rank, because the statistic rises with m for every construction and a ratio quoted at one rank does not transfer to another.) A scan of layers 3–9 at both the MLP output and the residual stream finds no cyclic weekday or month arrangement in that model, so this is the manifold being weak rather than the SAEs being at fault, and it is the boundary App. O already draws with our own dictionaries. We report both numbers rather than the stronger one alone: the recommendation is conditional, and its condition—how much of the concept's plane a dictionary's selective span contains—is measurable in advance, without any intervention.

The same contrast on three concepts, and what it does and does not decide. Twelve dictionaries varying in width and sparsity are also what feature splitting predicts: a concept's mass lands on different latents while its subspace survives. That account predicts the same pattern for every concept. The criterion does not: where the in-plane density has power at a harmonic the configuration couples to, the atoms are identifiable and should transfer, so the orbit-over-atom gap should depend on the density. We therefore ran the identical pipeline—selection by the dictionary's own decoder, the assigned plane, the ablation read-out, the matched controls—on the three concepts with cached probe pools at layer 12, changing only the prompts and the read-out tokens (`code/g49_concepts.py`; 108 prompts for the 12-item concepts, baseline top-1 99% for months and 79% for clock hours). This is an independent run with fresh random controls, so its weekday row differs from the tables above in the control-dependent counts (the atom beats its control in 0 of 12 dictionaries here against 7 of 12 there, the plane in 65 of 66 pairs against 58). "Span holds" is the ring plane's containment at top-64; "atom $>$ ctrl" counts dictionaries whose median advantage over the matched control is positive:

concept	$\max_{k<p}\lvert\hat{\rho}_k\rvert$	at k	span holds	atom	plane	rank-1	plane$-$atom	atom $>$ ctrl
weekdays	0.255	6	0.758	-0.042	$+0.815$	$+0.784$	$+0.86$	$0/12$
months	0.209	6	0.876	$+0.678$	$+0.893$	$+0.861$	$+0.22$	$10/12$
clock hours	0.156	2	0.471	$+0.834$	$+0.837$	$+0.786$	$+0.00$	$9/12$

Three things follow, and the third is against us. First, the plane transfers on all three—above its matched control for 12 of 12 dictionaries on each concept, 65, 65 and 66 of 66 pairs—so the recommendation is not a weekday fact. Second, the atom's transfer *is* concept-dependent: null for weekdays, $+0.678$ for months and $+0.834$ for clock hours, where it reproduces as well as the plane does. That dependence is not a width effect: split into the 19 same-width and 47 different-width pairs, the weekday atom fails at both (-0.091 and -0.025) and the month and hour atoms transfer at both ($+0.72$ and $+0.67$; $+0.89$ and $+0.83$). So the weekday result is not feature splitting across widths wearing a different name (splitting across sparsities at one width, and absorption, this split does not test), and "the atom does not transfer" is a statement about weekdays, not about atoms. Third, the dependence does not run the criterion's way. The density's strongest sub-lattice harmonic is largest for weekdays and smallest for clock hours, and the gap orders the same way, which is the reverse of "a more concentrated density makes the atom more identifiable". The criterion's atom-level statement is made for a fixed configuration, and these twelve dictionaries hold different numbers of atoms in these planes, so it does not speak to cross-dictionary atom transfer directly—but that is a limit of what it says, not a reason to read the ordering as support. On released dictionaries the concept-dependence of atom transfer is real, is not width-driven, and is not predicted by the density.

What predicts it pair by pair: where the atom points. Inside a ring, the most selective latent also picks a direction, i.e. an item. `code/g77_atom_item.py` records, for every dictionary of

the nine (layer, concept) cells of App. R, the item whose centered mean projects most on the atom, and splits the atom's pairs by it. Pairs whose atoms point at the same item correlate at a median $+0.85$ double-centered and other pairs at $+0.07$, higher in 8 of 8 cells that have both kinds (sign test $p = 1/256$; in the ninth, clock hours at layer 20, all ten atoms point at one hour). The weekday atoms spread over all seven days at layer 12 and five at layers 5 and 20; at layer 20 eight of ten month atoms and all ten hour atoms point at one item, and the atom transfers there (Table 13). Item agreement does not sort the other seven cells: their commonest item holds 3 to 5 of the ten or twelve atoms, and the atom transfers in three of them (months and clock hours at layer 12, weekdays at layer 20) and not in four; at layer 12, clock-hour atoms pointing at different hours still correlate at $+0.84$. So pairs that agree on the atom's phase—which item it picks out—transfer and pairs that disagree mostly do not, but how often a cell's atoms agree does not by itself decide the cell.

Atoms matched by item. The argmax atom compares, in most pairs, a latent for one item with a latent for another. `code/g86_item_atoms.py` removes the selection step: in each dictionary and for every item j, the *item-j atom* is the latent, among the dictionary's top-64 selective ones (the set the plane is built from), whose unit decoder direction projects most on item j's centered cluster mean; each is ablated as the argmax atom is, and same-item atoms are correlated across dictionaries (double-centered; float32 on CPU, with the argmax atom, the plane and a random direction recomputed in the same run). At layer 12 on weekdays the twelve dictionaries pick seven distinct latents for the seven days (a median), and same-day atoms transfer at a median $+0.58$ (interquartile range $+0.35$ to $+0.83$; 58% of pairs above 0.5), against $+0.04$ for different-day atoms, $+0.10$ for the argmax atom, -0.02 for a random direction and $+0.85$ for the plane (94% of pairs above 0.5). Every day's atoms transfer (per-day medians 0.36–0.77); item atoms beat the random direction for 12 of 12 dictionaries (sign test $p = 2.4 \times 10^{-4}$), and the plane beats them for 12 of 12. On months the same holds (a median 10.5 distinct latents for the twelve months): same-month atoms transfer at $+0.71$ (interquartile range $+0.40$ to $+0.93$; 65% of pairs above 0.5) against $+0.02$ for different months, $+0.50$ for the argmax atom, $+0.02$ for a random direction and $+0.95$ for the plane, above the random direction for 12 of 12 dictionaries and below the plane for 12 of 12. On clock hours item atoms transfer less ($+0.39$, interquartile range $+0.19$ to $+0.85$; $+0.04$ for different hours), still above the random direction for 12 of 12 and below the plane ($+0.91$) for 12 of 12, while the argmax atom does well ($+0.85$): there the dictionaries' most selective latent is the same kind of object in each, so selection does not hurt it. At layers 5 and 20 (ten dictionaries each, 16k and 65k) item atoms transfer as well: at layer 5 at $+0.74$, $+0.77$ and $+0.51$ for weekdays, months and clock hours, where the argmax atom gives $+0.06$, $+0.02$ and $+0.23$, and at layer 20 at $+0.59$ for weekdays (argmax $+0.14$), above the random direction and below the plane in every dictionary. Most of the argmax atom's failure is therefore selection—the dictionaries disagree on which day is most selective—and an atom claim indexed by its item survives across dictionaries, less reliably than the plane. The item atom is a latent's own decoder direction, which the cluster means choose but do not build, so its transfer is not the plane's anchoring.

What the plane is about: a specificity matrix. The transfer statistic is a correlation, and a correlation says that an object is found reproducibly, not what it is about (App. R). The size of an ablation's effect says the latter, so `code/g91_specificity.py` ablates each of the twelve dictionaries' assigned planes for weekdays, months and clock hours on every concept's prompts (mean $|\Delta \log p|$ over that concept's read-out tokens) and on unrelated text (next-token KL over 32 Pile documents, every position), with a random 2-plane drawn inside the same top-64 selective span as the within-span control. Each plane moves its own concept most, in 12 of 12 dictionaries for all three concepts, and against both comparisons: its effect on its own prompts over its effect on the other concepts' prompts is a median 12.6 for weekdays, 13.0 for months and 1.6 for clock hours, and over the other concepts' planes' effect on its prompts 4.8, 7.0 and 3.8 (1.31 against 0.09 and 0.21 for weekdays). The within-span random plane moves a concept's prompts by 0.14, 0.16 and 0.06. Ablating the weekday plane drops top-1 accuracy on the weekday prompts from 1.00 to 0.74 while the month and hour planes leave it at 1.00 (months $0.99 \to 0.86$; clock hours $0.79 \to 0.73$), and on unrelated text every plane disturbs the next-token distribution by about 10^{-3} nats per token, no more than the within-span random plane (0.0009 against 0.0012 for weekdays). At layers 5 and 20 (ten dictionaries each, 16k and 65k) the same holds in every dictionary for every concept, with ratios over the plane's effect on the other concepts' prompts of 13.9, 14.1 and 2.5 at layer 5 and 10.1, 31.9 and 2.0 at layer 20 (weekdays, months, clock hours), and on unrelated text again no more than the within-span random plane (`--layer 5, 20`). So the plane is not generic damage that happens to be reproducible: it is specific to its concept, strongly for weekdays and months and weakly for clock

hours, whose read-out tokens are number words the other concepts' planes can also reach. Which of the two occurs the criterion does not predict, and we do not claim the atom-level prediction on production dictionaries. What the criterion is entitled to at this scale is what §7 already says: the orbit recommendation, which Proposition 1 motivates without the Fourier criterion, and which this table extends from one concept to three.

R THE RECOMMENDATION ACROSS LAYERS, MODELS AND RELEASES

App. Q runs the transfer test at one layer of Gemma-2-2B and on one Pythia hook. This section runs it, unchanged, at three layers and three concepts of Gemma-2-2B, in Gemma-2-9B, and in a third release trained by a different laboratory on a different model family, and re-reads every cell with the shared component of the effect removed.

Removing the shared component. The statistic of §6 correlates two dictionaries' matrices of item log-probability changes (prompts $\times$ items). An ablation that lowers every item, or lowers whichever item is correct, correlates across any two dictionaries without their agreeing on anything about the concept. We therefore also *double-center* each matrix (subtract its prompt means and its item means) before correlating, and, most strictly, drop the correct item's entry first, so that what is compared is *which wrong items* gain or lose (`code/g61_effects_analysis.py`, from the saved effect matrices). Table 13 gives the raw and double-centered medians; the off-target reading gives the same verdict on the plane and on the atom against their controls in every cell (plane medians $+0.43$ to $+0.99$, jackknife $p \leq 0.009$ in the 18 cells that use it and 7 of 7 layers for each Llama Scope concept; the atom clears its control in the same seven cells), and is in the results file.

Gemma-2-2B, three layers and three concepts. GemmaScope's residual dictionaries at layers 5 and 20 (16k and 65k widths, all five released sparsities each, 10 dictionaries per layer; `average_l0` $18, 34, 68, 143, 309$ and $17, 29, 53, 105, 211$ at layer 5, $22, 38, 71, 139, 294$ and $20, 34, 61, 114, 221$ at layer 20) join the twelve of layer 12; probes are harvested per layer as in App. I, and the rings are in cyclic order at all three (`code/g49_concepts.py`, `code/g43_gemmascope_atoms.py`). In all nine (layer, concept) cells the plane beats its matched-rank control (double-centered jackknife $p \leq 5 \times 10^{-6}$), and it beats the atom in seven ($p < 0.05$; the two exceptions are months and clock hours at layer 20, where both transfer); the atom beats its own control in five of nine at $p < 0.05$, and never at layer 5. The concept-dependence of App. Q is therefore also a layer-dependence: at layer 5 the atom fails for months and clock hours too (-0.04 and $+0.06$ raw), while at layer 20 months' atom transfers ($+0.95$ raw) and so does the clock-hour atom, which matches the plane there ($+0.64$ against $+0.69$ double-centered). The plane transfers everywhere; the atom transfers where it happens to.

Many draws of each control. Each test above sets a claim against *one* random draw of its matched control per dictionary, and a fresh draw can move the atom's dictionary-level count (App. Q: 7 of 12 in one run, 0 of 12 in another). `code/g69_control_draws.py` gives every weekday dictionary at layers 5, 12 and 20 100 independent draws of each control and compares each pair's claim with the mean over draws. Double-centered, the plane stays far above its averaged control ($+0.91$, $+0.85$, $+0.84$ against $+0.12$, $+0.09$, $+0.17$; 10 of 10, 12 of 12 and 10 of 10 dictionaries, jackknife $p \leq 8 \times 10^{-8}$), and every one of the 45, 66 and 45 pairs beats 95% of its 100 draws. The atom's own correlation is small everywhere ($+0.06$, $+0.10$, $+0.14$ against $+0.08$, $+0.05$, $+0.09$ for the averaged control; the jackknife tests the mean paired difference, $+0.10$, $+0.10$ and $+0.15$, positive at layer 5 too although the medians order the other way); it clears the jackknife at layers 12 and 20 ($p = 0.03$ and 7×10^{-4}) but not at 5 (0.09), and only 8, 10 and 11 pairs beat 95% of their draws; its median advantage over the averaged control is positive for 3 of 10, 10 of 12 and 7 of 10 dictionaries (0, 2 and 5 raw, where single draws gave 7 and 0 of 12 at layer 12). So the weekday atom is not strictly null, as a single draw can make it look: it transfers weakly at two of three layers, far below the plane. At layer 12 the same 100 draws for months and clock hours (`--concept months,hours`) leave both objects standing: the plane at $+0.95$ and $+0.91$ double-centered against averaged controls of $+0.12$ and $+0.09$ (12 of 12 dictionaries, all 66 pairs above 95% of their draws, jackknife $p \leq 2 \times 10^{-15}$), and the atom at $+0.50$ and $+0.85$ against $+0.05$ and $+0.09$ (10 of 12 dictionaries, 39 of 66 pairs each; $p = 0.02$ and 0.03), so the months and clock-hour atoms' transfer in Table 13 is not a single draw's.

A third release: Llama Scope on Llama-3.1-8B. He et al. (2024) released residual-stream dictionaries of Llama-3.1-8B (Grattafiori et al., 2024) at every layer in two widths (32,768 and 131,072 latents; TopK-trained, JumpReLU at inference), from a different laboratory, architecture and model family (`code/g56_llamascope.py`, probes from `code/g55_llama_harvest.py`; weekday and month rings in cyclic order at all seven layers tested, $4, 8, \ldots, 28$; checkpoints `Llama3_1-8B-Base-L20R-8x` and `-32x` at layer 20, likewise at the others). With two widths per layer the pair is one per layer, so the unit is the layer. Double-centered, the plane beats its control at 7 of 7 layers for weekdays and for months (sign test $p = 0.008$ each), the atom at 3 of 7 and 1 of 7, and the plane beats the atom at 6 of 7 and 7 of 7. As in Gemma, deep layers are where weekday atoms transfer too (layers 20–28).

A larger model: GemmaScope on Gemma-2-9B. We run ten of the dictionaries DeepMind released for Gemma-2-9B's residual stream after block 20 (seven 16k sparsities, three 131k; `average_l0` $11, 20, 36, 58, 138, 310, 408$ and $19, 53, 114$) through the same test with probes harvested in the 9B model (`code/g63_gemma9b.py`; weekday and month rings in cyclic order). The ring plane sits inside the selective span at a median 0.59 (weekdays) and 0.80 (months); double-centered, the plane transfers at $+0.88$ and $+0.81$ and beats its control at jackknife $p = 2 \times 10^{-6}$ and 0.003, while the atom transfers at $+0.08$ and $+0.22$, *below* its random-direction control in both (ablating a random direction has a consistent effect in this model, $+0.40$), and the plane beats the atom in both ($p \leq 7 \times 10^{-4}$).

The Pythia suite, double-centered. The eight seeds of App. Q's second suite give $+0.42$ for the plane and -0.06 for the atom once the shared component is removed; the plane now beats its control at jackknife $p = 2.5 \times 10^{-4}$ (27 of 28 pairs, against $p = 0.016$ raw), the atom does not ($p = 0.82$).

Without probes or item labels. Every plane above is built from the probes' cluster means. `code/g57_unsup_plane.py` builds one from natural text alone, for each released dictionary at layers 5, 12 and 20: over 3.6M tokens of a public Pile slice (1,692 of them weekday names), the latents whose activation mass falls at least half on tokens that *are* weekday names (a median 6.5–11 per dictionary; the token's identity says "a weekday", never which), and the plane is the top-2 principal subspace of those latents' own reconstruction over the weekday-name positions. No probe and no day label enters; the probe ring plane is used only to evaluate it afterwards. The label-free plane transfers at all three layers—double-centered $+0.74$, $+0.66$ and $+0.81$ at layers 5, 12 and 20, against $+0.03$, $+0.04$ and $+0.05$ for the same construction on a random latent group matched in firing rate (jackknife $p = 5 \times 10^{-6}$, 1.5×10^{-6} and 2×10^{-9}). It is not the probe plane: it holds a median 0.67, 0.36 and 0.03 of it, so what the dictionaries agree on is the subspace their weekday-name latents use, which at layer 20 is almost disjoint from the ring the probes find. Its effect is of the probe plane's size (1.04 against 1.31) at layer 12, and 0.89 against 1.35 and 0.38 against 0.48 at layers 5 and 20. Two label-free atoms separate what the symmetry does. The latent most specific to one weekday name (share ≥ 0.99 in the median; Sunday in 7 of the 12 layer-12 dictionaries, a different day in the rest) breaks the rotation, and it transfers weakly at every layer ($+0.14$, $+0.23$, $+0.10$ double-centered). The latent with the largest weekday mass fraction spreads that mass over several days (no day above 40%) in 9 of 10, 10 of 12 and 5 of 10 dictionaries; at layer 12 it lies mostly off the ring plane (in-plane norm 0.24) and along the weekday centroid (cosine 0.63), encoding the ring's rotation-invariant center, and it transfers where it is of that kind ($+0.43$ and $+0.73$ at layers 5 and 12) and not at layer 20 ($+0.09$). What is invariant under the ring's rotation transfers; the atom that picks a phase does not—the distinction Proposition 1 draws, now without labels. Two limits. The decoder-only object (top-2 principal directions of the group's unweighted decoder vectors) transfers little ($+0.12$, $+0.11$, $+0.23$), so the plane needs the group's *usage*, not just its span; and a plane found the same way for months has an effect on the weekday prompts that is consistent across dictionaries too ($+0.95$, $+0.94$, $+0.59$) at a third of the magnitude or less (0.32, 0.34, 0.07): cross-dictionary agreement measures that an object is found reproducibly, and its effect's size, not its correlation, says what the object is about. *Not the residual stream's own plane.* Different dictionaries reconstruct the same residuals, so their label-free planes could agree merely by each lying near the stream's own top-2 principal plane at the weekday-name positions, an object no dictionary enters. They do not (`code/g88_resid_plane.py`; the same text and positions): at layers 5, 12 and 20 they sit at principal angles of a median 37°–45° and 70°–89° from it, sharing 0.25–0.37 of its energy, and ablating it has an effect that correlates with theirs at -0.03, $+0.13$ and $+0.04$ double-centered, against $+0.74$, $+0.66$ and $+0.81$ among the dictionaries themselves. What the dictionaries agree on is a subspace of their own, not the stream's dominant variance at those

tokens. The same construction for months (the nine month names that are not also common words; `--concept months`) gives a label-free plane that beats its random group at all three layers but more weakly at layers 5 and 12 ($+0.30$, $+0.33$ and $+0.77$ double-centered against $+0.08$, $+0.04$ and $+0.09$; jackknife $p = 0.035$, 0.003 and 4×10^{-7}), that holds little of the month probe plane (0.10, 0.17, 0.00), and whose phase-picking atom again transfers weakly at every layer ($\leq +0.10$); the weekday-name plane, found the same way, moves the month prompts consistently too ($+0.53$, $+0.41$, $+0.80$) at 0.26, 0.25 and 0.60 of the month plane's effect.

Steering, not only ablation. Ablation removes; a practitioner also wants to move the model along the concept, and the plane's own symmetry says how (`code/g60_plane_steering.py`). At the weekday token of 35 next-day prompts we rotate the activation's component in a plane about the ring's center by $s \cdot 2\pi/7$ ($s = 1, 2, 3$), or patch it to the target day's centroid (Engels et al., 2025), and count how often the answer moves by s days; center, orientation and centroids come from the probes for every plane alike. Averaged over s and the two interventions, each released dictionary's plane moves the answer in a median 24% and 21% of prompts at layers 5 and 12 (range 12–29%), against 30% and 24% for the probe plane itself and none for a random plane; the label-free planes, which hold less of the ring, 13% and 4%. At layer 20 nothing moves it ($\leq 1\%$, the probe plane included): the day is no longer read from this plane at that token. Steering at one layer and one position is weak for every plane, so we claim the plane for ablation; what steering adds is that the dictionaries' planes match the probe plane's own, weak, effect.

Against supervised causal search. Distributed alignment search (DAS; Geiger et al., 2024) learns a rank-r subspace such that swapping its content from a source prompt into a base prompt makes the model answer as the source would; the share of swaps that do so, the interchange-intervention accuracy (IIA), is the standard causal read-out, and RAVEL (Huang et al., 2024) uses it to compare SAEs with DAS. `code/g73_das_baseline.py` runs it on 42 next-weekday prompts (six templates, seven days; the model answers 90% of them correctly), over the 216 base–source pairs of one template that the model answers correctly, swapping the subspace's component at the weekday token after block 12; DAS is trained on three templates and scored on the other three, both ways (400 Adam steps at learning rate 2×10^{-3} on batches of up to 32 pairs, minimizing the cross-entropy of the source item's successor over the item tokens, the subspace being the orthonormal QR factor of a Gaussian-initialized $d \times r$ matrix; three initializations at rank 2). A dictionary's plane moves the answer to the one the source implies in a median 26% of pairs (13–31% over the twelve), its atom in 2%, the label-free plane in 7% and a random plane in none; the probe ring plane does so in 31% and DAS's own rank-2 plane in 42% (42–44% over initializations). Against rank-matched ceilings—DAS reaches 12% at rank 1 and 42% at rank 2—the atom reaches a sixth of what a rank-1 object can and the plane about three fifths of what a rank-2 one can, so part of the plane's advantage is its rank. These ceilings are DAS as run here: on its own training templates the rank-2 subspace reaches 51% and the rank-4 one 95%, and trained five times longer (2,000 steps) from five initializations (`--steps 2000 --das-seeds 5`) it reaches 42–44% on the held-out templates at rank 2 (49–52% on its own) and 85% at rank 4, so the rank-2 ceiling is not a short search's and the three fifths stands (0.26 against 0.44); a different search objective could still raise it. The IIA values are shares of pairs that reuse the same 42 prompts and are given without confidence intervals; the spreads over dictionaries and over DAS initializations above are the variability we measured. The dictionaries' planes share a mean squared principal cosine of 0.49 with the DAS plane (a random plane 0.001, the ring plane 0.59, DAS with itself across initializations 0.77). No plane holds the whole variable: DAS reaches 12% at rank 1 and 83% at rank 4, and the rank-6 span of the item means—the subspace a linear concept eraser such as LEACE (Belrose et al., 2023) removes, up to whitening—swaps it in 100%. Swapping at every position from the weekday token on changes none of this (27%, 2% and 44%). The same design at layer 5 orders the objects the same way (plane 30%, 25–32% over the two widths' dictionaries; atom 5%; ring plane 36%; DAS 44% at rank 2 and 84% at rank 4), and the dictionary's own selective span carries the rest as its rank grows (70% at rank 4, 100% at rank 6); so does months at layer 12 (72 prompts, 680 pairs; plane 20%, atom 2%, DAS 22% and 60%, selective span 47% at rank 4 and 98% at rank 11), and months at layer 5 (plane 20%, 19–22% over the dictionaries; atom 1%; DAS 23% and 57%). At layer 20 nothing swapped at the weekday token moves the answer (every object, the item-means span included, $\leq 1\%$) while DAS trained on every position from that token on reaches 94% at rank 4 (months the same: at most 4% at the month token, DAS at rank 4 included, and 73% at rank 4 from it on): by then the day has left the day token, and a plane read there is not where the model uses it, though its ablation still transfers ($+0.84$ double-centered, Table 13), so transfer there

marks an object found reproducibly, not where the day is read (`g73_das_baseline_{days_L5, days_L20, months_L5, months_L12, months_L20}.json`). Supervised causal search thus orders the objects as §6 does—the plane carries the day, the atom almost none of it—and adds that at layers 5 and 12 the day occupies more than a plane: the recommendation is to attach claims to the subspace a dictionary assigns to the concept, of whatever rank the concept needs.

Our own dictionaries, trained on unprompted activations. App. O's probe-free dictionaries were trained on 120,000 general vectors, too few for a dictionary to spend atoms on weekdays. `code/g58_natural_seeds.py` trains at GemmaScope's smallest width instead: 16,384 latents on Gemma-2-2B's residual stream after block 12, 60M tokens of FineWeb (Penedo et al., 2024) in packed 512-token windows, no probes, no upsampling, six seeds per coding rule that differ only in initialization and see the same activations in the same order (full recipe in Table 14 and below it). ReLU+L_1 (unit-norm decoder, L_1 coefficient ramped in over the first 1,000 steps and then set by a slow controller, one value for all six seeds, toward a mean L_0 of 64 over the first 30% of training and frozen) ends at $L_0 = 53$ and a fraction of variance unexplained of 0.31 with at most one dead latent per seed (silent for the last 20M tokens); TopK ($k = 64$, AuxK on dead latents (Gao et al., 2025)) at 0.27 with 41 dead per seed on average. `code/g59_natural_analysis.py` then asks §6's questions of these seeds. On weekdays the ring plane lies in each seed's top-64 selective span at a median 0.82 of its energy (ReLU+L_1) and 0.72 (TopK): a dictionary trained on text that merely mentions weekdays spends atoms on their ring. Ablating the plane each seed assigns transfers across the six seeds of a rule at +0.98 and +0.97 double-centered (jackknife $p = 0.002$ and 2×10^{-4} against a random-span plane) and between seeds of different rules at +0.59 raw; the most selective atom transfers at +0.05 and −0.01, not above its own random control. Months split the same way for TopK, while for ReLU+L_1 the atom beats its control, weakly (+0.12, $p = 0.02$). Clock hours, whose ring the spans hold at only 0.40 and 0.46, are the exception: there the atom correlates about as highly as the plane (+0.91 and +0.74 double-centered), clearing its control for ReLU+L_1 ($p = 0.048$) and not for TopK ($p = 0.06$). Over the six (rule, concept) cells the plane beats its control in six and the atom in two (Table 13, last six rows). The coding-rule contrast of §4.3 reappears at this width in the form the criterion predicts for TopK at $k \geq 2$, whose in-plane optimum is flat when at least two in-plane latents fire together (App. P's caveat for large k; we did not count them per token, and the two rules also differ in L_0, 53 against 64). Each seed's selective atoms with at least 0.3 of their norm in the ring plane ($\|P_V w\|/\|w\| \geq 0.3$) form a configuration whose k-th moment has a phase, and ReLU+L_1 seeds agree on it more than TopK seeds do: mean A_k over $k < p$ of 0.94 against 0.62 for weekdays, 0.74 against 0.52 for months and 0.66 against 0.47 for clock hours (exact permutation $p = 0.003$, 0.010 and 0.044 over the 924 ways to split the twelve seeds in two; `code/g72_rule_phase_test.py`), with more atoms in the plane for ReLU+L_1 (4–6 per seed against 3–4 for weekdays). Selecting atoms by their selectivity on the probes can imprint the probes' own angles on the configuration, so we also read it from *every* latent with at least 0.3 of its norm in the plane, with no selection (`code/g78_inplane_phase.py`): 0.96 against 0.57 for weekdays ($p = 0.001$), 0.70 against 0.58 for months ($p = 0.08$) and 0.58 against 0.50 for clock hours ($p = 0.19$)—the weekday contrast is the clear one, the other two lean the same way. Part of it is the number of latents read: at 0.5 of the norm, where both rules keep one or two weekday latents per seed, it is 0.94 against 0.81 ($p = 0.22$), while months sharpen (0.76 against 0.50, $p = 0.014$). Read instead from each seed's N latents with the largest in-plane fraction, so both rules contribute the same number (`g78 --topn`), the weekday contrast holds at every $N \geq 4$: 0.98, 0.95, 0.99 against 0.68, 0.78, 0.84 at $N = 4, 8, 16$ (permutation $p = 0.002$, 0.003, 0.001; 0.996 against 0.79 at $N = 2$, $p = 0.09$); months lean the same way ($p = 0.008$–0.15) and clock hours only at $N = 4$.

Two more layers. The same design at layer 5 (`g58 --layer 5`; the same token stream, recipe and seed indices; ReLU+L_1 ends at $L_0 = 55$, fraction of variance unexplained 0.21, no dead latents; TopK at 0.20 with 66–93 dead) gives the rule contrast for all three concepts. Read from every latent with at least 0.3 of its norm in the plane, ReLU+L_1 seeds agree at 0.87, 0.83 and 0.90 against TopK's 0.38, 0.61 and 0.39 for weekdays, months and clock hours (permutation $p = 0.001$, 0.011, 0.001; `g78 --layer 5`); from a matched $N = 8$ latents, 0.85, 0.80, 0.88 against 0.32, 0.60, 0.44 ($p = 0.001$, 0.017, 0.001); and atom by atom (Hungarian, $N = 8$, as below), ReLU+L_1's atoms sit 5.6°, 2.0° and 10.3° from their partners against TopK's 21.2°, 13.3° and 19.4° ($p = 0.006$, 0.011, 0.08; `g90 --layer 5`). The plane transfers across the seeds of each rule for every concept (raw median +0.69 to +0.99, jackknife $p \leq 0.007$ against a random-span plane), and at this layer ReLU+L_1's most selective weekday and clock-hour latents transfer as well (+0.99 and +0.94 raw,

$p = 0.022$ and 0.001 against a random direction) where TopK's do not ($+0.01$ and $+0.01$, $p = 0.09$ and 0.23); months' atom transfers for neither ($+0.04$, -0.06). Where the pinned rule's seeds agree on the in-plane atoms they can also agree on the single most selective one; the flat rule's seeds do neither. At layer 20 (ReLU+L_1 at $L_0 = 58$, fraction of variance unexplained 0.27, no dead latents; TopK at 0.23 with 55–82 dead) the contrast holds for months and clock hours but not for weekdays. Months: 0.89 against 0.72 from every latent at 0.3 of the norm ($p = 0.002$), 0.84 against 0.69 at $N = 8$ ($p = 0.004$). Clock hours: 0.73 against 0.34 at $N = 8$ and 0.81 against 0.47 at $N = 16$ ($p = 0.006$, 0.002; from every latent 0.54 against 0.44, $p = 0.23$, with one or two hour latents per seed in the plane), and atom by atom $4.7°$ against $18.0°$ ($p = 0.006$). Weekdays go the other way, not significantly: 0.68 against 0.82 from every latent and 0.66 against 0.70 at $N = 8$ ($p = 0.83$, 0.73), atoms $8.6°$ against $7.1°$, with some three weekday latents per seed in a plane the seeds' spans hold at only 0.54–0.61. The plane transfers across the seeds of each rule for every concept ($+0.70$ to $+0.98$ raw, jackknife $p \leq 0.04$), and ReLU+L_1's most selective latent does for all three concepts ($+0.86$, $+1.00$, $+1.00$), TopK's only for months ($+0.99$; weekdays $+0.14$, $p = 0.22$). Over the three layers and three concepts, then, ReLU+L_1 seeds agree more than TopK seeds at a matched count of 8 latents in 7 of the 9 (layer, concept) cells ($p < 0.05$ each; weekdays at layer 20 and clock hours at layer 12 are the exceptions, neither significant the other way), from every in-plane latent in 5, and atom by atom in 4.

The symmetry, made exact. The criterion's first prediction at this width is Proposition 1's: make the data symmetric under a rotation of one plane, and the seeds are free to lose their agreement on that plane's phase, and have no reason to lose it elsewhere. `code/g58_natural_seeds.py --sym-pool` retrains the six ReLU+L_1 seeds on the same token stream in the same order, with every activation's component in the weekday ring plane rotated about the ring's center by $2\pi j/7$, j drawn uniformly from $\{0, \ldots, 6\}$ per token. The data distribution—the law the tokens are drawn from, not the one-pass sample, in which each token is rotated once—is then exactly invariant under $\mathbb{Z}_7$ in that plane, every in-plane harmonic below 7 vanishes, and everything off the plane, the order and the initializations are those of the natural run. The rotation acts on every token's component in the plane, so the whole stream is spread around the ring's center and the seeds spend some 400 latents on the plane (5–8 in the natural run; about 130 at 0.5 of the norm). Read from every in-plane latent, the weekday phase agreement falls from 0.96 to 0.40, the level six uniformly random phases give (0.37; permutation $p = 0.001$, the smallest the 924 splits allow; 0.94 to 0.40 at 0.5 of the norm), while months (0.70 against 0.66, $p = 0.32$) and clock hours (0.58 against 0.60, $p = 0.56$), whose planes are nearly orthogonal to it (squared-cosine overlap ≤ 0.07), keep theirs. Since the symmetrized seeds are read from some 400 latents and the natural ones from 5–8, the two are also compared at a matched count, each seed read from its N latents with the largest in-plane fraction (`g78 --topn`): 0.95 against 0.42 at $N = 8$ and 0.99 against 0.39 at $N = 16$ (permutation $p = 0.001$, 0.008), so the fall is not the larger pool's. A sham arm applies the identical random $\mathbb{Z}_7$ rotation to another 2-plane of the stream instead (`--sham --sham-matched`): the adjacent eigenpair of the stream's covariance whose variance matches the ring plane's (37.45 against 37.53), so the rotation displaces a comparable share of each activation (0.0084 of $|h|^2$ against the ring plane's 0.0102; a random orthogonal plane, also run, displaces only 0.0016). The weekday agreement survives it (0.96 against 0.95, permutation $p = 0.21$; the random-plane sham 0.96 against 0.92, $p = 0.32$), and so do months (0.70 against 0.69) and clock hours (0.58 against 0.75): what removes the weekday phase is symmetry in the weekday plane, not a perturbation of that size elsewhere. The L_1 controller, one coefficient shared by an arm's six seeds, ends at 3.74 in the natural run, 3.86 symmetrized, 3.86 in the variance-matched sham and 3.75 in the random-plane sham (mean L_0 53.4, 53.9, 53.9, 53.8), so the sham that keeps the agreement also matches the symmetrized arm's penalty. The selected reading moves the same way less cleanly (0.94 to 0.65, $p = 0.025$; months unchanged, $p = 0.28$), as the imprint argument says it must. And the plane still transfers across the symmetrized seeds ($+0.99$ double-centered, jackknife $p = 2 \times 10^{-4}$; their spans hold 0.98 of the ring, so each seed's plane is nearly the ring plane and this transfer is close to guaranteed) while the atom does not ($+0.10$, $p = 0.40$; nor did it in the natural run, $+0.05$): the symmetry removes what the seeds' atoms agreed on and leaves what their planes did. Phase agreement and atom transfer are different read-outs: A_k is the phase of the in-plane configuration's k-th moment, while the ablated atom is the single most selective latent, which need not lie in the plane and can point at a different day in each seed (we did not record which). Installing one harmonic back tests the other half at this width (`--tilt-q 2`): the same symmetrized run with each point then moved along its circle about c by $(2r/q)\sin q\theta$, $q = 2$, $r = 0.35$, which installs harmonic 2 and, more weakly, its multiples. Read

from every in-plane latent, the seeds agree again at $k = 2$ ($A_2 = 1.00$ against 0.56 symmetrized), at its multiples 4 and 6 (1.00, 0.97) and at the alias $7 - 2 = 5$ (0.95), while $k = 1$ and 3 stay at chance (0.31, 0.24); over $k < 7$, 0.75 against 0.40 (permutation $p = 0.016$; at matched counts $N = 8$ and 16, 0.76 and 0.88 against 0.42 and 0.39, $p = 0.004$ and 0.001), and months and clock hours unchanged ($p \geq 0.5$). The symmetry removes the phase; one Fourier coefficient puts it back where that coefficient says. The tilted run spends as many latents on the plane as the symmetrized one (399–414 per seed at 0.3 of the norm, against 395–409; 5–8 in the natural run), so the loss of agreement is not the reallocation of some 400 latents to the plane, which the tiling relation of §3.3 could otherwise explain: with the same allocation, one harmonic restores it. The rotation also largely decouples a token's position in the plane from its day, so the plane says little about which day it is; the tilted run keeps that decoupling and its seeds agree again, so the loss follows the harmonic rather than the lost day information (a label-preserving symmetrization, each day's cluster moved onto a regular heptagon, we did not run).

Atom by atom. A_k summarizes a seed's atoms by the phase of one moment, so seeds could agree on it while their atoms do not. Matching the atoms themselves (`code/g90_atom_match.py`): each seed's N latents with the largest weekday-plane fraction are paired one-to-one with another seed's by the Hungarian algorithm on circular angle distance in the plane, a pair of seeds is scored by the median matched distance, the score is averaged over an arm's 15 pairs, and arms are compared by the exact permutation test over the 924 splits; each arm also has its own rotation null, the same pairs with one seed turned by a uniform random angle (200 draws). Natural ReLU+L_1 seeds' eight atoms sit a median $2.0°$ from their partners, against $27.7°$ for the rotation null; symmetrized, $15.5°$ against $19.7°$ (ratio to the null 0.07 against 0.79; $p = 0.006$; at $N = 7$ and 16, $3.2°$ against $16.1°$ and $2.2°$ against $13.5°$, $p = 0.011$ and 0.004). The variance-matched sham leaves them ($2.7°$, $p = 0.24$); installing harmonic 2 back re-aligns them ($3.8°$, null $43°$; $p = 0.001$ against the symmetrized arm); and TopK seeds, whose in-plane optimum is flat, match more loosely than ReLU+L_1's ($9.7°$, $p = 0.024$; $7.1°$ and $6.2°$ at $N = 7$ and 16, $p = 0.09$ and 0.06). The months and clock-hour planes, which were not symmetrized, move less and not reliably ($7.8° \to 13.9°$ and $6.9° \to 9.4°$ at $N = 8$, $p = 0.22$ and 0.15; months at $N = 7$, $5.4° \to 14.6°$, $p = 0.03$). (Reading every latent with at least 0.5 of its norm in the plane does not work here: the symmetrized seeds have some 130 such latents spread around the circle, so any two of them match closely at random, a null of $2.7°$.) So what the symmetry removes is the atoms' positions in the plane, not only a summary of them; this is a different object from the ablation transfer of each dictionary's single most selective latent above, which fails in both arms.

A second concept. The same intervention on the month plane (`g58 --sym-pool g25_pool_months_L12.pt`: every activation's component in it rotated about the ring's center by $2\pi j/12$, j uniform per token; same stream, order and initializations) replicates it for months. The months phase agreement falls from 0.70 to 0.43 read from every in-plane latent (permutation $p = 0.010$), 0.82 to 0.45 at $N = 8$ and 0.77 to 0.36 at $N = 16$ ($p = 0.007$, 0.001), and atom by atom the match goes from $7.8°$ to $23.0°$ at $N = 8$, which is its rotation null ($22.5°$; $p = 0.002$; $3.9° \to 12.9°$ at $N = 16$, null $13.3°$). Weekdays keep both (0.96 against 0.92, $p = 0.09$; atoms $2.0°$ against $1.7°$). Clock hours keep their phase (0.58 against 0.65 from every latent; 0.68 against 0.60 at $N = 8$, $p = 0.09$) but their atoms match more loosely ($6.9° \to 14.1°$ at $N = 8$, $p = 0.004$): the symmetrized seeds put some 126 latents on the month plane, which shares 0.07 of its energy with the hour plane, while the hour plane holds one latent or none per seed at 0.3 of the norm, so a top-N read of the hour plane can pick up rotated month latents; we have not separated that from a real spillover. Installing harmonic 5 back ($\gcd(12, 5) = 1$; `--tilt-q 5`, $r = 0.35$) restores the agreement where that coefficient says: $A_5 = 0.99$ and, at its multiple, $A_{10} = 0.99$ (0.15 and 0.38 symmetrized), with $k = 1$–4 still at chance (0.18–0.44), so the mean over $k < 12$ does not separate (0.51 against 0.43, $p = 0.26$); and it re-aligns the atoms ($5.8°$ against $23.0°$ at $N = 8$, $p = 0.004$; $2.7°$ against $21.1°$ at $N = 7$), with as many latents on the plane (117–137 per seed against 121–128). As for weekdays, one symmetry removes what the seeds agreed on in its own plane, atoms included, and one Fourier coefficient puts it back.

And a second layer. The weekday intervention at layer 5 (`g58 --layer 5 --sym-pool g25_pool_days_L5.pt`; same stream, order and initializations as the layer-5 seeds above) gives the same. The weekday phase agreement falls from 0.87 to 0.35 read from every in-plane latent ($p = 0.003$) and from 0.85 to 0.40 at $N = 8$ ($p = 0.003$), and atom by atom the match goes from $5.6°$ to $21.9°$ at $N = 8$, near its rotation null ($24.7°$; $p = 0.001$; $4.0° \to 23.8°$ at $N = 7$). Months

keep both (0.83 against 0.82; atoms 2.0° against 4.1°, $p = 0.24$); clock hours keep their atoms (10.3° against 11.1°) and their phase at matched counts (0.88 against 0.81 at $N = 8$, $p = 0.07$), though the every-latent read dips (0.90 against 0.77, $p = 0.006$). The symmetrized seeds again put many latents on the plane (132–136 at 0.3 of the norm).

S EVERY TEST, AND HOW IT CAME OUT

§7 separates illustrations from tests and gives the count; this is the list. *Illustrations*, which could not fail: the flat end of §4.2, the depth comparisons on GemmaScope's configurations (§5.1, App. I, M) and the numerics of App. B. *Tests*, which could, numbered one per prediction, with their outcomes; those whose predicted outcome did not occur at the stated level are marked †, and "weakly" marks an outcome in the predicted direction that clears its test only at the setting level, is small, or would follow from a generic account as well. *Controlled* (§4; App. C, D, M, P, U): (1) whether jitter tracks $R_m = 2\pi|\hat{\rho}_m|$ at fixed dimension (yes); (2) whether selectivity is non-monotone in m across generators (yes, 8 of 8); (3)† whether the binary rule holds on trained dictionaries (in 3 of 5 rows once the phase is read modulo the data's symmetry; ℓ=8 and ℓ=2 partly pinned); (4) whether breaking the modular transformer's exact symmetry in the representation deepens the landscape (yes, 5 of 5), (5) and by dose (yes, 3 of 3), (6)† in the order the dictionary's harmonic content predicts (no: the harmonic it singles out responds most, but it is a key frequency of the network, and key frequencies with small $|\hat{\gamma}_q|$ respond more than non-key harmonics with large, 0.25–0.27 against 0.18–0.20; App. D), (7)† and moves the seeds (not resolved at 24 seeds); (8)† whether production rules' seeds agree where the criterion says (TopK-1 no, JumpReLU rarely: against those rules' training, whose exact-code optimum is pinned); (9) whether each rule's own optimum is pinned (yes for four of five; no for TopK at $k \geq 2$, as the criterion says); (10) whether, on a 3-dimensional feature, same-shaped seeds differ by exactly the density's symmetry group (yes on the strongly anisotropic densities), (11)† and on a weak, smooth anisotropy (no: seeds scatter beyond the group). *On a pretrained model's own rings* (§5; App. K, T): Table 2's rows, (12) row 1, the natural ring locks below the lattice more than its symmetrization (yes, 130 of 152, 11 of 11 cells); (13) row 2, the lattice order survives symmetrizing (weakly); (14) row 3, an installed $q \geq 2$ locks at $k = q$ (yes, 91 of 111 strongest at q or its alias; 28 of 35 where $\gcd(p, q) = 1$, which symmetry does not constrain), (15)† and at $q = 1$ (no: the seeds lock at $k = 2$); (16) row 4, removing $\hat{\rho}_{k^*}$ unpins k^* on the spread 48-point densities (yes, 44 of 48 at four layers; 36 of 36 in Qwen), (17)† on the clustered weekday and month rings (no: lower in 9 of 24, $p = 0.86$; a failure on the manifolds the paper opens with), (18) and on the number circles of Llama-3.1-8B and Pythia-6.9B (yes, 70 and 65 of 84 settings, 7 of 7 layers each); (19) row 5, agreement grows with λ (weakly, since a generic convergence would take the same form); (20) rows 1, 3 and 5 on Llama-3.1-8B's number circle (yes, 6 of 6 layers for rows 1 and 3, 23 of 24 series for row 5); (21)† whether the density alone predicts the lock order on clustered rings (no), (22)† and on the spread density (no: 33 of 64, against 35 for the constant rule $k = 2$); (23) whether S_q orders selectivity on configurations we did not choose (weakly; App. I); (24) whether the fluctuation relation's variance holds at stationarity with nothing fitted (yes, 0.87–1.05 of the prediction; App. C). *The recommendation, at production scale* (§6; App. Q, R): (25) whether dictionaries agree on the plane and not on the atom (yes for the plane, which beats its control in 20 of 20 cells of Table 13, 17 at a Bonferroni threshold by the jackknife's t approximation, at most 11 by exact sign tests over dictionaries; the atom does in 7, so "not on the atom" holds in 13; with the 100-draw control of (28) the atom clears in 8, and by exact sign tests over each cell's units in 6), (26) whether the atom's failure is its phase, i.e. whether atoms pointing at the same item transfer (yes, higher than other pairs in 8 of 8 cells; posed after (25), and item agreement does not sort the cells, App. Q), (27)† and whether the atom's failure is ordered by the density across concepts (no); (28) whether the weekday atom's weak transfer survives averaging its control over 100 draws (yes: weak at two of three layers, far below the plane, and the months and clock-hour atoms at layer 12 hold too; a check posed after the single-draw result); (29) whether a plane built without probes or item labels transfers (yes, at three layers: $+0.66$ to $+0.81$ for weekdays against at most $+0.05$, and $+0.30$ to $+0.77$ for months, more weakly; a months plane found the same way also moves weekday prompts consistently, $+0.59$ to $+0.95$, at a third of the effect or less, so this tests that the plane is found reproducibly, not what it is about), (30) and whether the label-free latent that picks a phase transfers only weakly (yes, $\leq +0.23$ for weekdays and $\leq +0.10$ for months at every layer); (31) whether a dictionary's plane steers the answer as the probe plane does (weakly: both move it in a fifth to a third of prompts at layers 5 and 12, neither at layer 20); (32) whether

a dictionary's plane carries the variable in interchange interventions where its atom does not (yes: 26% against 2%, supervised search 42%); (33) whether seeds we trained on unprompted text agree on the plane and not the atom (yes, 6 of 6 cells against 2), (34) whether their ReLU+L_1 seeds agree on the in-plane phase more than TopK seeds, whose in-plane optimum is flat (at a matched latent count in 7 of the 9 (layer, concept) cells, none significantly the other way: all three at layer 5, weekdays and months at layer 12, months and clock hours at layer 20; weekdays at layer 20 go the other way, not significantly), (35) whether making the data exactly $\mathbb{Z}_7$-symmetric in the weekday plane removes the seeds' phase agreement there and only there (yes, at layer 12 and again at layer 5; that the plane still transfers is close to guaranteed, each seed's span holding 0.98 of the ring, and is not counted). *Production diagnostics* (App. I, E, F, G, O; (37)–(39) through the frozen-encoder estimator): (36)† whether GemmaScope sits at its predicted minimum (no; App. I); (37)† whether the L_0 law survives normalizing by the group's reconstruction energy (no: flat, and reversed with an adaptive group); (38)† whether the per-feature flatness diagnostic beats the group's reconstruction share at predicting inference-time interchangeability (no); (39)† whether the width law appears in the raw feature population (no; only in a firing-rate band chosen after the fact, and not significantly); (40)† whether landscape depth predicts cross-seed replication better than decoder geometry (no, at two scales); (41)† whether the graded prediction orders where seeds agree on unprompted data (no). *Follow-up tests* (App. M, P, Q, R, O, C, K): (42) whether Table 1's rule contrast survives a statistic with no firing weights and no chosen harmonic (yes: 16 of 21 cells for ReLU+L_1, 0 to 5 for the others); (43) whether each rule's trained seeds sit at their own landscape's minimum as Table 1 implies (yes: ReLU+L_1 at a median percentile 0.05, the other pinned rules at 0.52–0.63; nine ReLU+L_1 cells miss the bottom 5%, eight of them by far, App. P); (44) whether installing one harmonic back into the exactly symmetrized weekday plane restores agreement at it and at the multiples its displacement also installs (yes: $A_2 = 1.00$ against 0.56, with $k = 1, 3$ at chance); (45) whether the symmetrization's effect survives reading both arms from matched latent counts (yes: 0.95 against 0.42 at $N = 8$); (46) whether the same rotation of a variance-matched plane elsewhere in the stream leaves the weekday agreement intact (yes: 0.95 against 0.96, $p = 0.21$); (47) whether atoms matched by the item they point at transfer across GemmaScope dictionaries (yes: +0.58 on weekdays, +0.71 on months and +0.39 on clock hours at layer 12, above a random direction for 12 of 12 dictionaries and below the plane for 12 of 12 in each, and likewise at layers 5 and 20; posed after (26)); (48) whether exactly symmetrizing the weekday plane removes the seeds' agreement atom by atom, not only in one moment's phase (yes: Hungarian-matched atoms 2.0° apart become 15.5°; the sham leaves them, and one installed harmonic re-aligns them with the same some 400 latents on the plane; App. R); (49)† whether dictionaries trained without the probes, on 120,000 general vectors, agree on the weekday plane and not the atom (no: plane +0.069 and atom +0.016, both at their controls, the seeds' selective planes at random angles; too few vectors for a dictionary to spend atoms on weekdays; App. O); (50) whether the synthetic sweep's results survive giving every seed its own data draw and coordinate frame (yes: Spearman −1.00, the $m = 3$ minimum, and condition (B)'s pinned and partly pinned rows; the concentration at $\ell = 3$ that only a shared draw could produce disappears; App. C); (51) whether a rule whose in-plane optimum is flat, TopK-2, shows the ring's installed-harmonic selectivity (no, as predicted: the installed order is strongest in 9 of 35 tilts with $\gcd(p, q) = 1$, $q \geq 2$, chance 8.8, against ReLU+L_1's 28; but its seeds agree at $q = 1$, 26 of 27, so agreement at a first-harmonic tilt is not a pinning read-out; App. M); (52) whether the label-free planes the dictionaries agree on differ from the residual stream's own top-2 plane at the weekday-name tokens (yes: 37°–89° away, effect correlation $\leq +0.13$ against +0.66 to +0.81 among the dictionaries; App. R); (53) whether the plane a dictionary assigns to a concept is specific to it, i.e. moves its own concept's prompts more than the other concepts' planes do and more than it moves theirs (yes, in all 32 dictionaries at three layers for each of three concepts; 10–32× for weekdays and months, 1.6–2.5× for clock hours; App. Q); (54) whether removing $\hat{\rho}_{k^*}$ with every other harmonic up to 8 held at its natural value unpins k^* as the boxed edit does (yes: 0.76 → 0.40, 45 of 48 settings, 4 of 4 layers; in Qwen 0.93 → 0.52, 35 of 36; on the number circles, where the box removes only half, weakly: Llama 0.70 → 0.57, 73 of 84, Pythia 0.68 → 0.64, 61 of 84; App. K); (55) whether making the data exactly $\mathbb{Z}_{12}$-symmetric in the month plane does for months what $\mathbb{Z}_7$ does for weekdays (yes: phase 0.70 → 0.43, matched atoms 7.8° → 23.0°, at their rotation null; weekdays unchanged; clock hours keep their phase but their atoms loosen, 6.9° → 14.1°; installing harmonic 5 back restores A_5, 0.15 → 0.99, and the atoms, 23.0° → 5.8°; App. R). One further test could not be run: discovering feature planes by cross-seed consensus (App. N) returned mostly atom pairs that rarely fire together (19 of 27 survivors), with no spread in the predictor to test against. Fifty-five tests, seventeen against us: nine in the controlled and ring settings (3, 6, 7, 8, 11, 15, 17,

21, 22; 8 is the production rules' training, tested in the 2-plane) and eight at production scale (27, 36–41, 49). Of the thirty-eight others, four came out only weakly (13, 19, 23, 31) and one, 34, in 7 of 9 (layer, concept) cells. Six of the passes could fail but would not discriminate the criterion from a generic account—(4) and (5), since breaking a symmetry deepens a landscape on any account, (12) and (13), which Proposition 1 predicts without the Fourier criterion (App. K), (9), a numerical optimization of the objective, and (24), a simulation of the stationary-SGD approximation—and several share runs ((12), (13) and (19) use the same 152 settings, (14) and (15) the same 138 tilts), so the count is a ledger, not a measure of independent support. Nor is it corrected for multiplicity: we apply no family-wise or false-discovery correction across the fifty-five, and some analysis choices were made after earlier results had been seen, each flagged where it occurs—the ring-seed layer sets and coding rule, the bounded removal, the spread index and the competition order (App. K), the control added to App. L's runs, App. E's firing-rate band, App. M's $k \in \{1, 3\}$ re-reading, and tests (26), (28), (47), (48) and (50)–(55), posed after the results they follow up. Each outcome is to be read as one prediction at its own unit-level test, not as a member of a corrected family. Two passes, (25) and (33), hold in full only for the plane: "not on the atom" holds in 13 of 20 cells and in 4 of 6. Fifteen tests were posed after other results: (26), (28), (42)–(48) and (50)–(55).

The gradient, setting by setting. §7 says the symmetrized control locks more often in the full stream. The natural-over-symmetrized rate is flat (130 of 152 in the 2-plane, 20 of 24 in the full residual stream, 156 of 196 in Qwen, median margins 0.18, 0.17, 0.17); what grows is the symmetrized side's own lock rate (18 of 152 settings clear the null in the 2-plane, 43 of 196 in Qwen, 12 of 24 in the full stream), and on unprompted data the graded prediction fails outright (App. O). The production depth law that seemed to follow the published L_0 trend does not survive energy normalization (App. E), and whether what is left is limited by the criterion or by manifold discovery is open.

T THE PENALTY AND THE KERNEL: A REGULARITY

Table 2's fifth row, a regularity, is that agreement grows with λ, and Proposition 6 is where to look for it. The proposition gives the curvature as $4\pi m^2|\hat{\rho}_m||\hat{\tilde{c}}_m| + E$; it fixes the dependence on m and on the density, but λ enters only through $|\hat{\tilde{c}}_m|$, the sign of whose λ-dependence the proposition does not establish. The exact 2-plane solver settles it without further training (`code/g44_lambda_kernel.py`): for every configuration the twelve seeds reached, we re-solve the code at each λ and read the kernel at that series' lock order k, whose term $4\pi k^2|\hat{\rho}_k||\hat{\tilde{c}}_k|$ of the full sum takes the place of the $k = m$ term the proposition keeps.

Over the 38 (cell, M) series, $|\hat{\tilde{c}}_k|$ *rises* with λ in 32 of them (median Spearman $+1.00$, a median $3.57\times$ from $\lambda = 0.05$ to 0.4), so the curvature Proposition 6 names does grow with the penalty, and the measured agreement rises in 33 of the 38, the same direction. Two quantities that both move with λ will co-trend, so this is consistency rather than attribution; App. K's symmetrized and frozen-configuration controls are what separation these runs afford.

But that computation reads the trained configurations and each series' measured lock order, so on its own it cannot support calling this a prediction the identity fixes *before* a dictionary is trained. Both dependencies come out (`code/g48_lambda_fixedconfigs.py`): we fix the configurations in advance—the regular m-gon at each $m \in \{2, \ldots, 12\}$ and 20 random configurations at each m, drawn with seed 0; unit radius, $\lambda \in \{0.05, 0.1, 0.2, 0.4\}$—and sweep every harmonic $k \leq 12$ rather than the one the seeds picked out. Nothing in it reads a density, a trained dictionary or a result file. Over the 2,663 (configuration, k) series that carry the harmonic at all (relative to $\hat{\tilde{c}}_0$ above 10^{-5}, which drops the quadrature residue a regular 7- or 11-gon leaves at harmonics Corollary 1 zeroes), $|\hat{\tilde{c}}_k|$ rises with λ in 1,963 (median Spearman $+0.80$, a median $2.41\times$), and restricted to $k = m$—the dictionary's own order—in 163 of 220. The relative kernel falls in 80% of them, as it does on the trained configurations. The sign is therefore not read off the outcome, but it is not universal either: it holds in 74% of series chosen without reference to any training run and in 84% of the series the trained configurations give, while inside Proposition 4's regime a regular 2-gon's kernel $|\hat{\tilde{c}}_2| = |\hat{g}_2|$ *falls* with λ (0.114 to 0.047 over the same range), as every kernel must as $\lambda/r \to 1$. It is the outcome-independent 74%, for this ensemble and range of λ, that entitles the row in Table 2—as a regularity, not a consequence of Proposition 6. The regular polygons carry only $k \in m\mathbb{Z}$ (23 series) and behave like the rest (16 of 23, median $+0.80$).

The *relative* kernel $|\hat{\hat{c}}_k|/\hat{\hat{c}}_0$ moves the other way, falling in 37 of 38 series (median Spearman -1.00, $0.59\times$ end to end), because the zeroth harmonic grows faster. That is not a contradiction and it is worth stating plainly, because the two quantities answer different questions: the curvature at the minimum is absolute and is what the fluctuation relation of §3.3 turns into a phase spread, while a landscape's depth measured relative to its own mean is the ratio. That reading assumes an effective temperature that does not change with λ; under Adam, whose steps are nearly invariant to the loss's scale, the relative kernel may be what sets the phase spread, and then this row would be expected to run the other way. We have not measured the gradient-noise scale per λ that would decide it. App. I's natural-over-symmetrized depth *ratios* also fall with λ (25.7 to 11.0 across $\lambda = 0.05$ to 0.6, on GemmaScope's configurations rather than the seeds'), for a different reason: the mean $\hat{\hat{c}}_0$ cancels in that ratio, so its fall says only that the symmetrized landscape's range grows faster with λ than the natural one's, not that the natural curvature falls.

U THREE DIMENSIONS: THE ORBIT STATEMENT BEYOND THE PLANE

Proposition 3 and the curvature law are statements about a 2-plane; Proposition 1 is not, and it makes a prediction a 3-dimensional feature can test without new theory: the solution set is invariant under the group that leaves the density invariant, whatever that group is, so seeds are free to differ by it; if, further, the optimizer reaches the optimum and the density pins what the group does not move—an assumption, not part of the proposition—they should differ by no more. Densities of the same dimension but different symmetry should then leave seeds different amounts of freedom. `code/g75_sphere_seeds.py` puts 4,000 activations on the unit sphere of $\mathbb{R}^3$, with six densities: *isotropic* ($G(P) = O(3)$); *axial band*, a polar cap holding 40% of the mass and a band at latitude $-20°$, uniform in longitude ($G(P)$: rotations about the axis and reflections in planes containing it); *axial weak*, $\rho \propto 1 + 0.6\cos\vartheta + 0.5P_2(\cos\vartheta)$, the same group with a weak, smooth anisotropy; *octahedral*, six equal von Mises–Fisher clusters ($\kappa = 40$) at $\pm e_1, \pm e_2, \pm e_3$ (the 48 isometries of the cube); *perturbed*, the same clusters moved by 15–25° and reweighted (trivial group); and *generic*, eight clusters at random points (trivial group). ReLU+L_1 autoencoders (unit-norm decoder, ReLU encoder with its own bias, no decoder bias; every group above fixes the origin, so Proposition 2 applies to the uncentered law as it stands) are trained from 12 seeds at $M \in \{4, 5, 8\}$ and $\lambda \in \{0.1, 0.2\}$ (3,000 Adam steps at learning rate 3×10^{-3}, batches of 512). Two seeds are compared by the symmetric Chamfer angle between their atom sets as trained, after the best isometry in the density's own group, and after the best isometry in $O(3)$ (nearest-neighbor Kabsch from 65 starts, on the set and on its mirror image). Seeds can also reach different configurations, and a best isometry brings any two sets closer, so the main read-out is conditional: among pairs that reach the same configuration up to an isometry (within 5°), the fraction that also agree as trained, and after the density's own group (within 5° each).

density	group	same shape (pairs)	agree as trained	agree after the group	as trained, all pairs
isotropic	$O(3)$	369/396	0.5%	100% (by definition)	0.5%
axial band	axial	278/396	27%	99%	19%
octahedral	O_h (48)	199/396	23%	100%	12%
perturbed	trivial	189/396	76%	76%	36%
generic	trivial	225/396	99%	99%	56%
axial weak	axial	337/396	7%	28%	6%

On the two strongly anisotropic symmetric densities, same-shaped seeds differ by the density's group and by nothing more (99% and 100% agree once it is allowed for, against 27% and 23% as trained); on the two without symmetry they agree as trained (99% generic; 76% perturbed, most of the rest in one setting that splits, below); the isotropic row holds by construction. All twelve seeds share one sample of 4,000 points, whose finite-sample irregularity breaks a continuous group slightly; that, or atoms on the axis, which the axial group does not move, may be why 27% of same-shaped axial-band pairs agree as trained, and we have not run the per-seed resampling that would separate the two; nor do Propositions 1–2, which concern the population law, strictly cover an objective computed on one sample, whose own group is generically trivial. Empirically, then, dimension 3 leaves the orientation undetermined as far as the density's group does, and, where the density pins the rest strongly, no further (Proposition 1 gives only the first half). Two exceptions: on

the perturbed density one setting ($M = 4$, $\lambda = 0.1$) splits between two orientations, and the weak smooth anisotropy pins the tilt so weakly that same-shaped seeds scatter beyond the group as well—consistent with the 2-plane picture, in which a small Fourier coefficient pins weakly, though in $d = 3$ we have no closed-form landscape to make that a prediction. Unconditionally (last column), fewer pairs agree, because seeds also land on different configurations, which no group describes.

V ASSETS AND COMPUTE

Models: Gemma-2-2B and Gemma-2-9B (Gemma Team, 2024) (Gemma Terms of Use) with their GemmaScope SAEs (Lieberum et al., 2024) (CC-BY-4.0); Llama-3.1-8B (Grattafiori et al., 2024) (Llama 3.1 Community License) with the Llama Scope SAEs (He et al., 2024) (Apache-2.0, as listed on the release's collection page); Pythia-70M, Pythia-160M and Pythia-6.9B (Biderman et al., 2023) (Apache-2.0) and the replicate SAEs of Paulo & Belrose (2026) (released without a stated license); Qwen2.5-1.5B (Qwen Team, 2024) (Apache-2.0). Data: FineWeb (Penedo et al., 2024) (ODC-By; the shard `000_00000.parquet` of its `sample/10BT` subset), MiniPile (Kaddour, 2023) (dataset card: "other") and a 10k-document slice of the Pile (Gao et al., 2020) (`NeelNanda/pile-10k`; dataset card: BigScience BLOOM RAIL 1.0). Each is used under the terms of its public release; the exact checkpoints, SAE variants and file lists are in the supplementary code, which also defines every item list, template, prefix and prompt (probes: `g11`, `g14`, `g28`, `g55`; the 38 (30) density-probe templates of App. H: `g6c`; intervention, steering and DAS prompts: `g43`, `g49`, `g60`, `g73`). The controlled experiments of §4, the ring-seed runs of §5 and the 3-D check ran on CPU (two to fourteen cores); everything else that runs a language model or trains a production-width dictionary, except the item-matched atoms of App. Q (on CPU), used a single 24 GB RTX 4090 (about a day of GPU time in all, of which the six-seed dictionaries of §6 took 65–80 minutes per run, a run training one coding rule's six seeds together in one batched pass). Recounts over stored records that the appendices quote—held-out fits by condition, per-plane Rayleigh tests, own-seed kernels, tilts split by $\gcd(p, q)$, lock-order rules with and without alias credit, the number circles' side effects—are reproduced by `code/g87_recounts.py` into `results/g87_recounts.json`. Unless a p-value is marked two-sided it is one-sided, in the direction the criterion predicts; two-sided values are marked, and are those where no direction was predicted in advance or, as for the L_0 and width laws of App. E, where we tested conservatively. Layers of one model share weights and data, so where layers or (ring, layer) cells are the units—Llama Scope, the number circles, the cell-level tests of App. K—a test measures consistency across depth rather than independent replication; tests over settings, pairs, configurations or features are descriptive, and the unit-level test is given beside them. The supplementary material includes `scripts/consistency_scan.py`, which recomputes the paper's reported quantities from the released result files and asserts that each appears in the compiled PDF; it carries one or more checks per guarded number and exits non-zero on any mismatch. It guards the numbers as they are computed, not every restatement of them elsewhere in the text. It loads every number it recomputes from a named result file, so it is also the map from claims to files.

Recipes. Every dictionary of Table 14 has unit-norm decoder columns, Gaussian at initialization and renormalized after each step, an untied encoder initialized to the decoder's transpose (0.1 times it at production width), a zero encoder bias, and a decoder bias subtracted before encoding (zero at initialization except at production width; the 3-D check has none); the table gives the rest (`code/angsym/sae.py`, `g3`, `g13`, `g25`, `g58`, `g75`). Two setups are outside the table: the 3-D check states its recipe in App. U (no decoder bias), and the Pythia-160M TopK suite of App. G ($k = 32$, Adam at 3×10^{-4}, batches of 2,048) follows the protocol of Paulo & Belrose (2026) (`g3`). At production width (`code/g58_natural_seeds.py`) the decoder bias starts at the mean of the first 65,536 scaled activations and each decoder column's gradient is projected off the column before the step; the L_1 coefficient starts at 2.2, is ramped in over the first 1,000 steps and, until 30% of training, is multiplied each step by $\exp(0.002\,\mathrm{clip}((\bar{L}_0 - 64)/64, -1, 1))$, $\bar{L}_0$ the mean over the six seeds; a ReLU+L_1 latent silent for 1,000 steps is re-initialized, every 1,000 steps, toward batch inputs drawn in proportion to their squared error; AuxK reconstructs the residual from the top 512 latents silent for 2,000 steps, on a quarter of the batch, with weight $1/32$ on its error relative to the residual's; activations pass through a 2^{18}-token shuffle buffer. That run trains under bfloat16 autocast with float32 parameters, the table's other rows in float32 (App. L and O and the Pythia-160M run store their inputs in float16). Language models run in bfloat16, except the Gemma-2-2B ablation, steering and control-draw runs and the Pythia-6.9B harvest (float16) and the Pythia-70M run, the

Pythia-160M ablations and the item-matched atoms of App. Q (float32). The modular transformer of App. D reads the tokens $(a, b, =)$ through a learned embedding of $p + 1$ symbols plus a learned positional embedding initialized from $\mathcal{N}(0, 0.02^2)$; it has no biases, and every other weight takes PyTorch's default initialization under seed 0.

Definitions and hooks. A_k is always the resultant length across seeds of $e^{i \arg Z_k}$, $Z_k = \sum_j w_j e^{ik\alpha_j}$ over a seed's in-plane atoms, a seed with none being dropped (no stored seed of App. K or M has none, so there every A_k is over 12 seeds). The atoms and weights are: atoms firing on over 1% of the points and at least 0.5 in the plane, weighted by firing frequency (App. K, M); every atom, weighted by its squared in-plane norm fraction (App. H, L—whose headline keeps the six largest—and O); latents with at least 0.3 of their norm in the plane, weighted by that fraction (App. R); unit weights where App. K says so. App. C's phase uses atoms at least 0.75 in the plane, unweighted. Depth is a landscape's range: in loss units in App. C; over its mean in App. D, H, I and M; over its minimum in App. E and F; in units of the zero code's cost in App. P; depths are therefore compared within an appendix, never across. The exact-objective landscapes $J(\varphi)$ are evaluated on 720 rotations in App. H, I and P and on 2,048 in App. M's rule landscapes; App. P's percentile counts a rotation as lower only if its J is below the trained rotation's by more than 10^{-12}, so ties count as not lower. Every Rayleigh p is $e^{-Z}\big(1 + (2Z - Z^2)/4n\big)$, $Z = nA^2$, over the n seeds' unit phasors, clipped to $[0, 1]$. App. K's kernel-weighted lock-order rules take each seed's exact kernel at that setting's own λ, averaged over the seeds with at least two atoms. A series' λ trend (App. K, T) is a Spearman correlation over four points and descriptive on its own (exact one-sided $p = 1/24$ at $+1$, $4/24$ at $+0.8$); the evidence is the count of series. The transfer statistic of §6 is the Pearson correlation between two dictionaries' flattened prompts $\times$ items matrices of log-probability changes, over all entries (63×7 for weekdays, 108×12 for months and clock hours), double-centered where so marked and with each prompt's correct item left out for the off-target reading. In App. E a feature is *live* if it fires at least 30 times in the 300k-vector landscape pool; up to 2,000 live features per SAE are drawn uniformly (generator seed 0), one being skipped if its group has fewer than 64 active samples; a latent's OMP retention rate is the share of the encoder's selections of it, over the 4,096 OMP activations, that OMP also makes, *defined* when the encoder selects it at least 5 times. The reproducibility statement's 129,071 are the defined rates over all 671,744 latents of the nine SAEs; Table 5 counts those of the sampled features. The arc filter's cut at 2.5 uniform gaps separates a ring, whose largest gap is near 1, from an arc, whose one large gap spans the part of the circle it does not cover (`g15`, `g16`). Every read and ablation in Gemma-2, Llama-3.1, Qwen2.5 and Pythia-6.9B is at the residual stream after block L (`hidden_states[`L`+1]`), the tensor GemmaScope's and Llama Scope's residual dictionaries read; GemmaScope dictionaries are named by their `average_l0` tags where they are used (§6, App. Q, R), and Pythia-70M is the deduplicated model; for Pythia-70M and Pythia-160M a layer-L residual is `hidden_states[`L`]`.

Probes, prompts and tokens. A probe is the residual at the last token of the item inside a template sentence behind one of 12 prefixes (none, "Note:", "Remember,", "For the record,", "As discussed,", "By the way,", "In short,", "To be clear,", "Honestly,", "Apparently,", "In fact,", "Briefly,"), tokenized with BOS; a probe whose item tokenizes differently in context is dropped. The set of App. I, K and §6 has 20 templates per ring (`g11` for weekdays and months; `g14` for clock hours, ten hour sentences and ten generic ones), so a clock hour is read at the final token of "o'clock", which its twelve items share. The ablation prompts of §6 (`g43`, `g49`) take, for an item a followed by b and c, five list templates answered by the item after c ("a, b, c,", "The days go a, b, c,", "In order: a, b, c,", "a, then b, then c, then", "After a comes b, after b comes c, after c comes") and four answered by b ("The day after a is", "Yesterday was a, so today is", "If today is a, tomorrow is", "a is followed by"), with month and hour analogues ($7 \times 9 = 63$ and $12 \times 9 = 108$ prompts); no probe template is among them. The read-out is the log-softmax over the whole vocabulary at the last position, at the first token of the item word with a leading space (for clock hours, the number word). DAS (`g73`) adds "The next day after a is" and "Tomorrow comes after a, so it is" to the four next-item templates (42 prompts; 72 for months); steering (`g60`) uses the same six and drops the seven prompts that open with the day, whose name then has no leading-space token (35). The label-free planes (`g57`) count as a weekday name every vocabulary entry that decodes, stripped and lower-cased, to one of the seven names, and as a month name for the months plane one of the nine other than May, March and August, which are also common words.

Data sampling. The general-text pool of App. L and O (`g25`) takes `pile-10k` documents in a random order (seed 0), skips those under 200 characters or 16 tokens, truncates each to 128 tokens

and keeps every position but the first (BOS); App. L trains on 60,000 of its vectors at an even stride, App. O on the first 120,000. App. E's 1.2M pool (`g1`) and App. G's Pythia-70M pool (`e3`) take the documents in dataset order that reach 256 and 128 tokens, without the first position; the Pythia-160M run (`g3`) reads the MiniPile shards in file order in the same way at 256 tokens. The label-free planes (`g57`) read all 10,000 `pile-10k` documents in a random order (seed 0), truncated to 512 tokens, BOS excluded: the 3.6M tokens of App. R. The production-width run (`g58`) streams the FineWeb shard in file order, documents joined by EOS into 512-token windows that each start with BOS, whose position is dropped. App. I's item counts are over 39,982 FineWeb documents (`results/g71_item_counts.json`).

Checkpoints. Every language model is the base (pretrained) checkpoint, none instruction-tuned. GemmaScope dictionaries are the release's `params.npz` files, e.g. `layer_12/width_16k/average_l0_82`; Llama Scope's are `Llama3_1-8B-Base-L12R-8x` and `-32x` and their analogues at the other layers; the suite of Paulo & Belrose (2026) is the `EleutherAI/sae_overlap` release.

Seeds. Initializations are seeded by seed index, data draws separately. App. C: SAE seeds 0–19; one data draw, plane embedding included, from generator seed 12345. App. D: transformer, split and batches seed 0; SAE seeds 0–23; the single SAE whose $|\hat{\gamma}_q|$ picks the reweighted harmonics is seed 0, and the reweighting resamples use generator 11. App. G: seeds 0–7 at both scales. App. H's trained read-outs: seeds from 0 (24 or 32), batches from generator 4242. App. K: seeds 5000–5011 (replicate 5100–5111), tilt resampling generator 7, and the null of $\max_{k<p} A_k$ from 4,000 simulations per (M, p) (seed 0). App. L and O: seeds 5000–5011, batches from generator 4242. App. M: seeds 1000–1011, each on its own data draw (4400–4411; random bases 88000–88011). App. P: random starts from generator 0. App. Q and R: random controls from generator 0, drawn in turn for each dictionary (Llama Scope: the layer index; the 100-draw controls: 1234). Our own dictionaries: seeds 7000–7005 in each rule's run, so the two rules start from the same six initializations; symmetrizing rotations from generator 4321, the random sham plane from 97531; the 2^{18}-token buffer is shuffled by PyTorch's default generator, which separate runs consume differently, so the natural, symmetrized and sham runs share the token stream and its order into the buffer but not their minibatches. DAS: seed $100r + 10i + f$ for rank r, initialization i and fold f; its random planes and the steering's from generators 11 and 1. App. U: seeds 1000–1011 on one data draw (seed 12345).

Script map (scripts in `code/`, records in `results/` or `results/gpu/`, named without extension). Fig. 1: `paper_figures_main.py` from `g21_real_ring_seeds_*_L*`, `g43_gemmascope_atoms_all`, `g42_public_saes`, `g46_selection_null_L12`. Figs. 2, 3, 4: `paper_figures.py` from `e1_synthetic_sweep`; `e2_modular`, `e2b_modular_seeds`; `e6_L*`, `e7_L12`. Fig. 5, Table 6: `g4_manifold.py`, `g4_analysis.py` (`g4_manifold`). Fig. 6, Table 7: `g8_ring_objective.py` from `ring_projections` (`g8_ring_objective`); Table 8 is the circular moments of the probe angles in `ring_projections`. Fig. 7: `paper_figures_trained.py` from `g12_projections`, `g15_objective_trained`, `g11_ring_gemmascope`. Table 1: `g13_coding_rules.py` (`g13_coding_rules`), `g22_rule_hparams.py` and `g22_analysis.py` (`g22_summary`), the kernel column from `g87_recounts`. Table 2: `g21_analysis.py`, `g36_analysis.py`, `g89_removal_sham.py`, `g65_numbers_summary.py`, `g68_numbers_table2.py`, `g87_recounts.py`. Table 3: `g54_condb_quotient.py`. Tables 4 and 5: `g1_gemmascope.py` (`e6_L*`, `e7_L12`), `g2_analysis.py`, `g9_ls_refit.py` (`g9_lsrefit_L12`). Table 9: `g14_rings_multi.py`, `g15_all_layers.py`, checked by `g45_ring_geometry.py`. Table 10: `g21_real_ring_seeds.py`, `g21_analysis.py` (`g21_summary`). Table 11: `e9_kernel_by_rule.py`. Table 12: `g22_analysis.py`. Table 13: `g61_effects_analysis.py`, `g62_transfer_table.py` (`g62_transfer_table`).

	16k	32k	65k	131k	262k
overcompleteness	$7\times$	$14\times$	$28\times$	$57\times$	$114\times$
encoder–OMP Jaccard	0.400	0.369	0.334	0.306	0.278
OMP / encoder recon. error	0.903	0.872	0.851	0.827	0.821
vs. encoder support + LS refit	0.936	0.924	0.897	0.884	0.887
... and at matched support size	0.970	0.959	0.937	0.926	0.929
LS refit / encoder recon. error	0.959	0.951	0.942	0.936	0.932
mean retention	0.544	0.510	0.458	0.414	0.381
features with a defined retention rate	1,469	1,111	774	481	285
Spearman vs. per-feature retention:					
landscape depth (ours)	$+0.286$	$+0.352$	$+0.254$	$+0.304$	$+0.367$
decoder NN cosine	$+0.268$	$+0.292$	$+0.219$	$+0.213$	$+0.229$
firing rate	-0.148	-0.159	-0.084	$+0.045$	$+0.005$
scale-free curvature	-0.086	-0.132	-0.126	-0.228	-0.216
group reconstruction share	$\mathbf{+0.338}$	$\mathbf{+0.406}$	$\mathbf{+0.313}$	$\mathbf{+0.414}$	$\mathbf{+0.453}$
margin (depth − NN cosine)	$+0.019$	$+0.060$	$+0.035$	$+0.091$	$+0.139$
bootstrap 95% CI	$[-.03, .07]$	$[.00, .11]$	$[-.03, .10]$	$[.01, .16]$	$[.04, .25]$
bootstrap p	0.485	**0.038**	0.327	**0.016**	**0.013**
partial: depth | NN cosine	$+0.182$	$+0.241$	$+0.161$	$+0.223$	$+0.296$
partial: NN cosine | depth	$+0.149$	$+0.129$	$+0.093$	$+0.019$	$+0.024$
partial: depth | share	-0.035	-0.022	-0.041	-0.119	-0.067
partial: depth/energy | share	-0.018	-0.032	-0.068	-0.120	-0.095
partial: share | depth	$+0.190$	$+0.218$	$+0.194$	$+0.316$	$+0.293$

Table 5: Inference-time identifiability at layer 12 (matched-L_0 SAEs; OMP run on 4,096 activations with k set to the encoder's median L_0). Correlations are over the sampled features that OMP ever had the chance to keep, whose count is given in the table. Because depth and the decoder-cosine baseline are dependent, the rotated group being the feature plus its 3 nearest decoder-cosine neighbors, the margin is tested by a 2,000-fold paired bootstrap over features (95% CIs shown) rather than by comparing marginal p-values. The decisive rows are the reconstruction-share row and the last three. Two caveats on the margin: the five bootstrap tests are not corrected for multiplicity (three are individually below 0.05), and the feature population is selected by an OMP criterion whose stringency changes with width, so the count of features with a defined retention rate falls from $1{,}469$ to 285 across the sweep—the margin's rise, which is not monotone, is therefore measured on a shrinking, differentially selected sample. "Group reconstruction share" is $E/(E + \mathrm{MSE})$, where E is the mean squared norm of the group's contribution to the reconstruction on its active samples: a quantity with no connection to flatness, which outpredicts every flatness statistic at every width and, once partialled out, leaves depth (in either normalization) at zero or slightly negative. We report this rather than the favorable comparison. The three added rows are the control the comparison needs: the encoder's raw coefficients are amortized and thresholded, so the headline row confounds *which* atoms are chosen with *what coefficients* they get. Refitting least squares on the encoder's own support removes the second factor (row 6: it recovers 4–7% of the error, more at larger width), and restricting OMP to each sample's own $|S_{\mathrm{enc}}|$ removes any cardinality advantage. OMP still reconstructs better in both controls, and the margin still broadly grows with overcompleteness, so the dissociation is a property of the support rather than of the coefficients. Rerun as `code/g9_ls_refit.py` on the same activations, sample and dictionaries; the 262k column is a separate invocation with an identical draw sequence, and its encoder error differs from the headline run in the third significant figure (2605.9 vs 2614.7) because the median L_0 rounds to 65 rather than 66. That median, and so OMP's k, differs between the two runs at four of the five widths ($80, 74, 73, 68, 65$ in the rerun against $79, 75, 72, 68, 66$), so the headline OMP row and the three added rows come from different runs and need not multiply exactly (the rerun's own OMP/encoder ratios are $0.898, 0.878, 0.845, 0.827, 0.826$). Jaccard is per activation, $|S_{\mathrm{enc}} \cap S_{\mathrm{OMP}}|/|S_{\mathrm{enc}} \cup S_{\mathrm{OMP}}|$ averaged over the $4{,}096$; retention is per feature, the fraction of the activations on which the encoder fires it that OMP also selects (features the encoder fires on at least 5 of them), averaged over features.

sweep	quantity	value at each L_0 level (sparsest → densest)				
L12/16k	measured L_0	23	43	85	184	475
	neighbors within $\cos \geq 0.25$	21.7	8.3	3.4	1.4	0.9
	co-active partners ($J \geq 0.2$)	0.52	0.48	0.37	0.35	0.34
	median relative depth	0.0212	0.0142	0.0105	0.0063	0.0081
	median depth / group energy	0.420	0.488	0.559	0.567	0.679
L20/65k	measured L_0	21	36	64	118	235
	neighbors within $\cos \geq 0.25$	46.8	20.9	10.0	3.9	1.8
	co-active partners ($J \geq 0.2$)	0.82	0.90	0.76	0.55	0.43
	median relative depth	0.0341	0.0233	0.0172	0.0115	0.0077
	median depth / group energy	0.440	0.474	0.530	0.586	0.605

Table 6: Adaptive-group measurements. "Co-active partners" is the mean number of other atoms whose firing pattern overlaps the feature's at Jaccard ≥ 0.2 on the shared activation pool, the grouping rule a cosine threshold cannot mimic; it falls with L_0 in both sweeps ($0.52 \to 0.34$, Spearman -1.000; $0.82 \to 0.43$, -0.900), which is the co-activation check that App. E's decoder-cosine grouping lacks. Spearman against L_0: relative depth $-0.900/-1.000$; cluster size $-1.000/-1.000$; depth per group energy $+1.000/+1.000$. Spearman of depth per group energy against the *measured* atom count: -1.000 in both sweeps (exact permutation $p = 0.017$, two-sided), but both quantities are monotone in L_0 over the same five points, so that -1 holds by construction and is not an independent test of Proposition 6; a real test needs m and L_0 to vary independently, which this grid does not provide.

		(A) depth vs. ε (regular → real)					(B) depth at $\varepsilon = 0$		
ring	m	0	0.25	0.5	0.75	1	κ present	finite κ	no κ
days (p=7)	2	5.7e-03	6.6e-02	1.2e-01	1.7e-01	2.6e-01	2, 4, 6	3.93× (9)	— (0)
	3	5.6e-03	3.6e-02	5.8e-02	8.2e-02	1.0e-01	3, 6, 9	3.23× (9)	— (0)
	4	5.7e-03	3.2e-02	5.5e-02	4.4e-02	2.3e-02	4, 8, 12	1.58× (9)	— (0)
	5	8.8e-03	2.0e-02	2.6e-02	3.8e-02	5.0e-02	5, 10, 15	1.28× (9)	— (0)
	6	—	—	—	—	—	6, 12, 18	1.48× (9)	— (0)
months (p=12)	2	4.3e-03	2.4e-02	4.4e-02	6.1e-02	7.5e-02	2, 4, 6	5.59× (9)	1.09× (9)
	3	1.0e-03	1.2e-02	2.4e-02	3.4e-02	4.4e-02	3, 6	105× (6)	2.00× (12)
	4	1.4e-02	2.0e-02	2.6e-02	3.0e-02	3.5e-02	4	9.67× (3)	1.04× (15)
	5	1.4e-03	8.4e-03	1.7e-02	2.3e-02	2.9e-02	5, 10, …, 30	1.62× (18)	— (0)
	6	—	—	—	—	—	6	5.03× (3)	1.01× (15)

Table 7: The criterion on Gemma-2-2B's own cyclic features, (A) at layer 7 (layers 12 and 20 are in the released records and agree; the full $5 \times 5 \times 3 \times 2$ grid is in `results/gpu/G8_RESULTS.md`). (A) symmetrizing the ring flattens the objective and interpolating back deepens it; (A) columns are relative depths $(\max J - \min J)/\overline{J}$ at $q = 0$, and the m=6 row is omitted there only for width. (B) columns are *ratios to each row's own untilted depth*, pooled over the three layers, partitioned by whether the configuration has a finite $\kappa = \min\{jm : jm \equiv \pm q \bmod p\}$, the first aliased order it can feel (not the k^* of §5.1, which is the density's strongest harmonic), which exists iff $\gcd(m, p) \mid q$. Across the 12-point ring that is 39 cells at a median $5.0\times$ against 51 at $1.05\times$; the m=3 row is where the flat side is least clean ($2.00\times$), and it is also the row with the smallest untilted baseline in the grid (1.0×10^{-3}), so a fixed absolute leakage buys it the largest ratio. Within the finite-κ cells depth falls off with κ on both rings. The truncated series of Eq. 2 predicts the individual depths with Spearman $+0.997$ (days, 45 cells) and $+0.993$ (months, 90 cells), median predicted/measured 1.14 and 1.09.

ring	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8
days, layer 7	0.212	0.222	0.158	**0.049**	0.261	0.306	0.397	0.176
days, layer 12	0.187	0.168	**0.097**	0.117	0.224	0.252	0.276	0.177
days, layer 20	0.157	0.171	0.146	0.087	0.240	0.383	0.287	**0.083**
months, layer 7	**0.034**	0.062	0.058	0.118	0.163	0.201	0.117	0.065
months, layer 12	**0.048**	0.075	0.126	0.125	0.163	0.192	0.094	0.090
months, layer 20	0.078	0.071	0.258	0.092	0.167	0.075	0.096	**0.070**

Table 8: Circular moments R_k of the model's own (unsymmetrized, $\varepsilon = 1$) ring density, on this appendix's probe set (App. K uses that of Table 9); the weakest harmonic of each row is bold. On the day ring the 4th is the weakest at layer 7 and the second weakest at layers 12 and 20, which is consistent with the m=4 interpolation curves turning over: the configuration gains least from de-symmetrization where the model's density carries least. It does not explain the m=5 and m=6 turnovers, whose harmonics are among the ring's strongest, and we leave those unexplained rather than fit a story to them.

Table 9: Item sets. Cells are (layer) $\times$ (item set); the three rings used in §5.1 are measured at all 26 layers—the 26 block outputs; App. H scans the embedding output as well—and the rest at layers 7/12/20. The clock-hour ring has no measurable embedding row, since its twelve items share their final token. ev_2 is the fraction of the item means' variance in the top-2 plane; the gap column is the largest angular gap between consecutive item means, in units of $2\pi/p$.

item set	p	cells	cyclic order	ev_2 (med.)	max gap (med., range)
weekdays	7	26	26/26	0.58	1.64 (1.27–1.80)
months	12	26	24/26	0.42	1.86 (1.39–2.64)
clock hours	12	26	23/26	0.69	1.92 (1.54–3.45)
numbers one–ten	10	3	3/3	0.55	2.41 (2.36–2.72)
seasons	4	3	0/3	0.73	1.18 (1.15–1.30)
compass points	8	3	0/3	0.55	2.28 (2.19–2.80)
zodiac signs	12	3	0/3	0.26	2.72 (2.57–3.19)
farm animals (control)	8	3	0/3	0.40	2.12 (2.11–2.13)

Table 10: Per-cell medians over M and λ, unit convention. "below" is $\max_{k<p} A_k$; "lattice" is A_p.

ring	layer	p	k^*	below, natural	below, symm.	lattice, natural	lattice, symm.
weekdays	1	7	6	0.63	0.48	0.48	0.41
weekdays	3	7	6	0.64	0.52	0.50	0.29
weekdays	7	7	6	0.66	0.50	0.33	0.58
weekdays	12	7	6	0.71	0.48	0.37	0.52
clock hours	3	12	3	0.74	0.46	0.35	0.40
clock hours	7	12	9	0.71	0.51	0.29	0.30
clock hours	12	12	2	0.74	0.61	0.36	0.29
months	1	12	11	0.63	0.50	0.28	0.33
months	3	12	11	0.64	0.47	0.26	0.28
months	7	12	6	0.63	0.44	0.25	0.32
months	12	12	6	0.64	0.46	0.19	0.35

Table 11: Cost-kernel structure by coding rule, for unit-radius data in a 2-plane, over two regular polygons ($m \in \{4, 6\}$) and six random configurations (three each at $m = 4$ and 6). "Range" is $\max_\theta \bar{c} - \min_\theta \bar{c}$; the last column is the largest $\max_{1 \le q \le 12} |\hat{\bar{c}}_q|$ over the random configurations. JumpReLU is evaluated at two fixed reference thresholds θ_0, as the reconstruction cost of the gated projection $a_j = \langle x, d_j \rangle \, \mathbf{1}[\langle x, d_j \rangle > \theta_0]$ (encoder tied to the decoder, no L_0 term), and TopK as the non-negative least-squares code on the k atoms most correlated with x; the trained SAEs learn their thresholds. A rule with no angular structure in its kernel cannot be pinned by any density.

coding rule	range, random cfgs	range, regular cfgs	$\max_q \lvert\hat{\bar{c}}_q\rvert$
ReLU+L_1 ($\lambda = 0.2$)	0.115–0.320	0.024–0.063	9.4×10^{-2}
TopK, $k = 1$	0.345–0.500	0.125–0.250	1.4×10^{-1}
TopK, $k = 2$ (non-negative)	0.155–0.500	$< \mathbf{2 \times 10^{-31}}$	1.5×10^{-1}
TopK, $k = 2$ (signed)	$< \mathbf{2 \times 10^{-29}}$	$< \mathbf{2 \times 10^{-31}}$	$\mathbf{2 \times 10^{-30}}$
JumpReLU ($\theta_0 = 0.05$)	0.334–1.811	7×10^{-16}–1.2×10^{-3}	3.7×10^{-1}
JumpReLU ($\theta_0 = 0.30$)	0.313–1.852	7×10^{-16}–4.5×10^{-2}	3.6×10^{-1}

Table 12: Hyperparameter grid at the installed harmonic ($q \in \{2, 3\}$ pooled) and on the uniform control (read at $k \in \{2, 3\}$). A is the cross-seed agreement at the installed order; "≥ 0.8" counts grid points above a fixed threshold; ReLU+L_1's weakest point is 0.973, which three of the 160 other-rule points reach. The last column is a one-sided exact Mann–Whitney test; ReLU+L_1's two re-initialization settings coincide, so its 16 points are 8 distinct runs, each entered once.

coding rule	grid points	A installed, median [min, max]	sig.	≥ 0.8	A uniform, median [max]	pinned > flat
ReLU+L_1	16 (8 distinct)	0.984 [0.973, 0.999]	16/16	16	0.289 [0.472]	$p = 7.8 \times 10^{-5}$
TopK, $k = 1$	16	0.213 [0.152, 0.700]	1/16	0	0.267 [0.632]	$p = 0.54$
TopK, $k = 2$ (non-neg.)	16	0.475 [0.080, 0.663]	8/16	0	0.262 [0.525]	$p = 0.011$
JumpReLU, tied θ	64	0.335 [0.045, 0.995]	17/64	6	0.233 [0.610]	$p = 2.2 \times 10^{-4}$
JumpReLU, per-atom θ	64	0.305 [0.037, 0.832]	11/64	1	0.220 [0.642]	$p = 1.7 \times 10^{-3}$

Table 13: **The transfer test in every cell run.** Median cross-dictionary correlation of an ablation's effect, raw / double-centered; p is the one-sided leave-one-dictionary-out jackknife (t_{n-1} approximation, App. Q) of the double-centered difference against the matched-rank control (plane against a random-span plane, atom against a random direction; "ctrl" is that control's double-centered median, from one draw per dictionary, and in Gemma-2-9B and Llama Scope a random direction's effect already correlates at about $+0.4$), and the last column the plane against the atom. Llama Scope has one pair per layer, so its entries are medians over seven layers and counts of layers. The last six rows are our own dictionaries, six seeds per coding rule trained on unprompted text (`code/g58_natural_seeds.py`); their pairs are seeds of one rule. The p-values are per cell and uncorrected; at a Bonferroni threshold of $0.05/20$ the plane still beats its control in 17 of the 20 cells (not Gemma-2-9B months, $p = 0.003$, nor either Llama Scope cell, whose seven-layer sign test cannot go below $1/128$) and the atom in 1. The jackknife p is a t approximation that no unit count floors; held instead to an exact sign test over each cell's dictionaries, seeds or layers, the plane is above its control for every unit of every cell ($p \leq 1/64$), which clears $0.05/20$ only in the 11 cells with at least nine units, and the atom clears 0.05 in 6 cells (not Gemma-2-2B layer-20 weekdays, 8 of 10, nor our ReLU+L_1 months, 4 of 6, but our TopK clock hours, 6 of 6) and $0.05/20$ in 1. The cells are not independent replications: nine share one model and release, and our six share data and token order. The atom's verdicts also move with the control draw: against the mean of 100 draws the layer-12 weekday atom clears ($p = 0.03$; App. R).

release	model	layer	concept	plane	ctrl	p	atom	ctrl	p	plane > atom
GemmaScope	Gemma-2-2B	L5	days	$+0.84$ / $+0.91$	$+0.12$	2×10^{-7}	-0.03 / $+0.06$	$+0.08$	0.12	1×10^{-6}
GemmaScope	Gemma-2-2B	L12	days	$+0.81$ / $+0.85$	$+0.18$	4×10^{-6}	-0.04 / $+0.10$	$+0.05$	0.07	1×10^{-7}
GemmaScope	Gemma-2-2B	L20	days	$+0.77$ / $+0.84$	$+0.26$	7×10^{-7}	$+0.07$ / $+0.14$	$+0.07$	0.005	1×10^{-7}
GemmaScope	Gemma-2-2B	L5	hours	$+0.95$ / $+0.97$	$+0.05$	9×10^{-10}	$+0.06$ / $+0.23$	$+0.10$	0.22	7×10^{-5}
GemmaScope	Gemma-2-2B	L12	hours	$+0.84$ / $+0.91$	$+0.08$	3×10^{-9}	$+0.83$ / $+0.85$	$+0.10$	0.03	0.02
GemmaScope	Gemma-2-2B	L20	hours	$+0.50$ / $+0.69$	$+0.10$	2×10^{-6}	$+0.67$ / $+0.64$	$+0.09$	1×10^{-5}	0.58
GemmaScope	Gemma-2-2B	L5	months	$+0.96$ / $+0.99$	$+0.20$	1×10^{-9}	-0.04 / $+0.02$	$+0.08$	0.27	1×10^{-7}
GemmaScope	Gemma-2-2B	L12	months	$+0.89$ / $+0.95$	0.00	3×10^{-12}	$+0.68$ / $+0.50$	$+0.07$	0.02	0.003
GemmaScope	Gemma-2-2B	L20	months	$+0.97$ / $+0.99$	$+0.06$	6×10^{-9}	$+0.95$ / $+0.97$	$+0.07$	0.010	0.053
GemmaScope	Gemma-2-9B	L20	days	$+0.88$ / $+0.88$	$+0.32$	2×10^{-6}	$+0.04$ / $+0.08$	$+0.40$	1.00	7×10^{-7}
GemmaScope	Gemma-2-9B	L20	months	$+0.72$ / $+0.81$	$+0.38$	0.003	$+0.13$ / $+0.22$	$+0.41$	0.97	7×10^{-4}
Paulo & Belrose	Pythia-160M	L6 (MLP)	days	$+0.55$ / $+0.42$	$+0.10$	2×10^{-4}	-0.07 / -0.06	$+0.04$	0.82	7×10^{-6}
Llama Scope	Llama-3.1-8B	7 layers	days	$+0.68$ / $+0.76$	$+0.35$	7/7	$+0.30$ / $+0.31$	$+0.43$	3/7	6/7
Llama Scope	Llama-3.1-8B	7 layers	months	$+0.68$ / $+0.82$	$+0.29$	7/7	$+0.12$ / $+0.11$	$+0.39$	1/7	7/7
ours, ReLU+L_1	Gemma-2-2B	L12	days	$+0.95$ / $+0.98$	$+0.21$	0.002	$+0.03$ / $+0.05$	$+0.11$	0.66	2×10^{-4}
ours, ReLU+L_1	Gemma-2-2B	L12	hours	$+0.92$ / $+0.89$	-0.08	9×10^{-4}	$+0.92$ / $+0.91$	$+0.19$	0.048	0.26
ours, ReLU+L_1	Gemma-2-2B	L12	months	$+0.89$ / $+0.95$	-0.08	6×10^{-5}	$+0.07$ / $+0.12$	$+0.06$	0.02	0.008
ours, TopK	Gemma-2-2B	L12	days	$+0.94$ / $+0.97$	$+0.16$	2×10^{-4}	-0.02 / -0.01	-0.02	0.20	0.008
ours, TopK	Gemma-2-2B	L12	hours	$+0.87$ / $+0.96$	$+0.13$	2×10^{-4}	$+0.89$ / $+0.74$	$+0.14$	0.06	0.14
ours, TopK	Gemma-2-2B	L12	months	$+0.68$ / $+0.94$	-0.02	1×10^{-4}	-0.11 / -0.02	$+0.09$	0.38	4×10^{-4}

Table 14: Training recipes, all with Adam; every re-set acts only in the first half of training. [a]Latents firing on under 10^{-4} of the first 4,096 points, checked every 500 steps. [b]Latents silent for 500 steps. [c]Linear warm-up over 1,000 steps, linear decay over the last 20%. [d]ReLU+L_1 latents silent for 1,000 steps, checked every 1,000 steps; TopK instead uses AuxK from step 2,000 on.

setting	learning rate	schedule	batch × steps (or epochs)	dead latents re-set to	inputs
synthetic planes (App. C)	3×10^{-3}	cosine	512×25 epochs	not re-set	as generated
modular addition (App. D)	3×10^{-3}	cosine	256×150 epochs	not re-set	unit mean norm
Pythia-70M (App. G)	3×10^{-3}	cosine	$1{,}024 \times 25$ epochs	not re-set	centered, unit mean norm
Pythia-160M, TopK (App. G)	3×10^{-4}	constant	$2{,}048 \times 50$M tokens	not re-set	centered, unit mean norm
planes and rings (App. K, M)	3×10^{-3}	constant	$1{,}024 \times 2{,}500$ or 4,000	random directions[a]	as built there
3-D check (App. U)	3×10^{-3}	constant	$512 \times 3{,}000$	not re-set	unit vectors
residual stream (App. L, O)	3×10^{-3}	cosine	$256 \times 6{,}000$ or $512 \times 8{,}000$	random pool vectors[b]	mean norm $\sqrt{d}$
production width (App. R)	3×10^{-4}	warm-up, decay[c]	4,096 tokens $\times$ 14,648	high-error inputs[d]	mean squared norm d