When Good Brain Predictors Make Bad Brain Controllers: A Regression Theory of Prediction–Control Dissociation

Anonymous authors
Paper under double-blind review

Abstract

Deep neural networks are increasingly used both as encoding models of visual cortex and as controllers that synthesize images to drive neural responses toward desired states. Recent *in vivo* experiments reveal a striking dissociation: models with nearly identical held-out accuracy can generate qualitatively different stimuli and differ sharply in control, and these differences track input-gradient structure better than predictive accuracy. We develop a geometric and analytical account of this phenomenon in a linear teacher–student model. The key is that control creates a self-generated distribution shift. Natural-image generalization, self-accentuation, and peer review evaluate the same fitted predictor under three different input distributions; correspondingly, their equal-error sets in weight space are ellipsoids, spheres, and hyperplanes. Using deterministic equivalents for ridge regression, we show that prediction-optimal and control-optimal regularization satisfy different spectral conditions, so cross-validation can leave large control error when responses are noisy and inputs are anisotropic. Extending the analysis to fixed linear feature maps, we construct a prediction-preserving gauge family with identical generalization and cross-validated regularization but arbitrarily different control. Control is most fragile when feature directions that survive regularization backpropagate into low-variance, high-spatial-frequency image directions. The resulting gradient-geometry factor extends to nonlinear networks as a Jacobian-based diagnostic and predicts across-model variation in *in vivo* control better than held-out prediction error. Thus predictive accuracy alone cannot certify control; the model's gradient geometry or generated stimuli must also be evaluated.

1 Introduction

A good predictive model of a neuron can, in principle, be used as a *model-based* controller: if $f(\mathbf{x})$ predicts the response of a neuron to an image $\mathbf{x}$, one can synthesize images that drive f to a target level and present them to the subject to modulate neuronal state. Aside from model-free generative search (Ponce et al., 2019; Xiao & Kreiman, 2020), model-based control of visual neurons has been successfully pursued across species (Bashivan et al., 2019; Walker et al., 2019). Recently, Prince, Wang, et al. (2026) scaled this paradigm into a benchmark of many deep-network encoding models and uncovered three puzzles: Encoding models that are nearly indistinguishable on predicting held-out natural images generate visibly different stimuli and differ several-fold in how well those stimuli modulate the neuron *in vivo*; models that fail to control can still recognize the successful stimuli of other models; and the best predictor of control success is not accuracy but the spectral structure of the model's input gradient.

We argue that these phenomena arise already in linear models, from a single fact: *control evaluates a model on a distribution the model generates itself.* Held-out prediction measures weight error in the metric of the data covariance Σ, which suppresses low-variance directions; accentuation moves along the fitted weight itself and therefore measures the same error in Euclidean geometry, where nothing is suppressed. For natural images, whose covariance spectrum decays over many orders of magnitude, this is the difference between not seeing an error and being dominated by it.

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.

Our contributions are: **(I)** a formulation of generalization, self-accentuation and peer review as evaluations of one predictor under three input distributions, with their equal-error sets (ellipsoid, sphere, hyperplane) exposing the dissociation geometrically (Sec. 3); **(ii)** deterministic equivalents for pixel-space ridge regression showing that prediction-optimal and control-optimal penalties satisfy different spectral conditions, so cross-validation in anisotropic data and teacher leaves a computable residual control error (Sec. 4); **(iii)** for ridge regression on a fixed linear feature map, a separation of the summary statistics that determine prediction from those that determine control, and a gauge family of feature maps with identical generalization but arbitrarily different control (Sec. 5); **(iv)** a response-independent gradient-geometry factor T_F that identifies fragile feature maps, extends to nonlinear networks as a diagnostic, and recovers model ordering in the neuronal control data. All statements are proved in the appendix, where we also restates every result in derivation order (App. D); numerical validation against Monte Carlo accompanies each section.

Related work. Encoding models built on deep networks now predict cortical responses so similarly that benchmarks no longer rank them (Yamins et al., 2014; Güçlü & van Gerven, 2015; Cadena et al., 2019; Schrimpf et al., 2018; Conwell et al., 2024). Read one way this is convergence (Huh et al., 2024; Kapoor et al., 2025); read another it is a weak test, answered by constraining the mapping (Prince et al., 2024) or by seeking stimuli that split models apart (Golan et al., 2020; Feather et al., 2023; Rajalingham et al., 2018; Geirhos et al., 2019). Closed-loop control is one limit of that strategy: the model writes the stimulus and the neuron grades it (Ponce et al., 2019; Xiao & Kreiman, 2020; Bashivan et al., 2019; Walker et al., 2019; Prince et al., 2026). Optimizing an input against a fitted model likewise underlies feature visualization (Olah et al., 2017; Hamblin et al., 2024; Fel et al., 2023) and adversarial attack (Nguyen et al., 2015; Goodfellow et al., 2015; Mądry et al., 2018), where robustness has been tied to alignment (Guo et al., 2022; Dapello et al., 2020; Taori et al., 2020). In each case the test distribution is created by the model under test, which we make analytical. Classic theory characterizes ridge risk under anisotropic covariance and signal–spectrum alignment (Dobriban & Wager, 2018; Hastie et al., 2022; Mei & Montanari, 2022; Mel & Ganguli, 2021), as well as generalization and optimal regularization under an externally specified train–test distribution shift (Canatar et al., 2021; Tripuraneni et al., 2021; Patil et al., 2024). We instead study a self-generated shift: the test distribution is supported on an optimization path created by the fitted predictor itself.

2 Prediction and control dissociate in neurons and in linear regression

Prince, Wang, et al. (2026) used the following closed-loop paradigm (Fig. 1, top): responses of visual cortical sites (V3/V4–IT, 25 sites in 5 macaques) to ∼1000 naturalistic images were recorded; for each site, ten encoding models were fit by ridge regression on the top 750 PCs of features from ten deep network backbones, with penalty and layer chosen by cross-validation per site; using *feature accentuation* (Hamblin et al., 2024), each model optimized an image using a variant of gradient ascent, and then synthesized images that drive its own prediction to eleven target levels from ten seed images (Fig. 1**A**); and the images were presented to the same sites in the *control session*. The control is successful if the proposed images evoked neural activations as predicted (data and models: Apps. B.6, B.7). Three observations motivate this paper.

Similar prediction, divergent control. The trained models were tightly matched on held-out natural images (Fig. 1**B.I**), yet the slope and correlation between predicted and measured responses to their own accentuated images varied widely across models, with far larger spread than on natural images. Many models reached their target *predicted* response by adding texture and high-spatial-frequency structure that left the neuron largely unmoved (Fig. 1**B.II**).

Recognition without generation. Presenting model A's images to model B ("peer review", Fig. 1**B.III**) shows an asymmetry: a model whose own images fail to modulate the neuron still predicts the neuronal responses to a successful model's images with good correlation and better R^2 than the generator itself. Models can recognize good control stimuli that they cannot produce.

Control tracks input-gradient structure. Across the 250 site–model pairs, control success is predicted by the spectral profile of the model's input gradient $\nabla_{\mathbf{x}} f$: models whose gradients con-

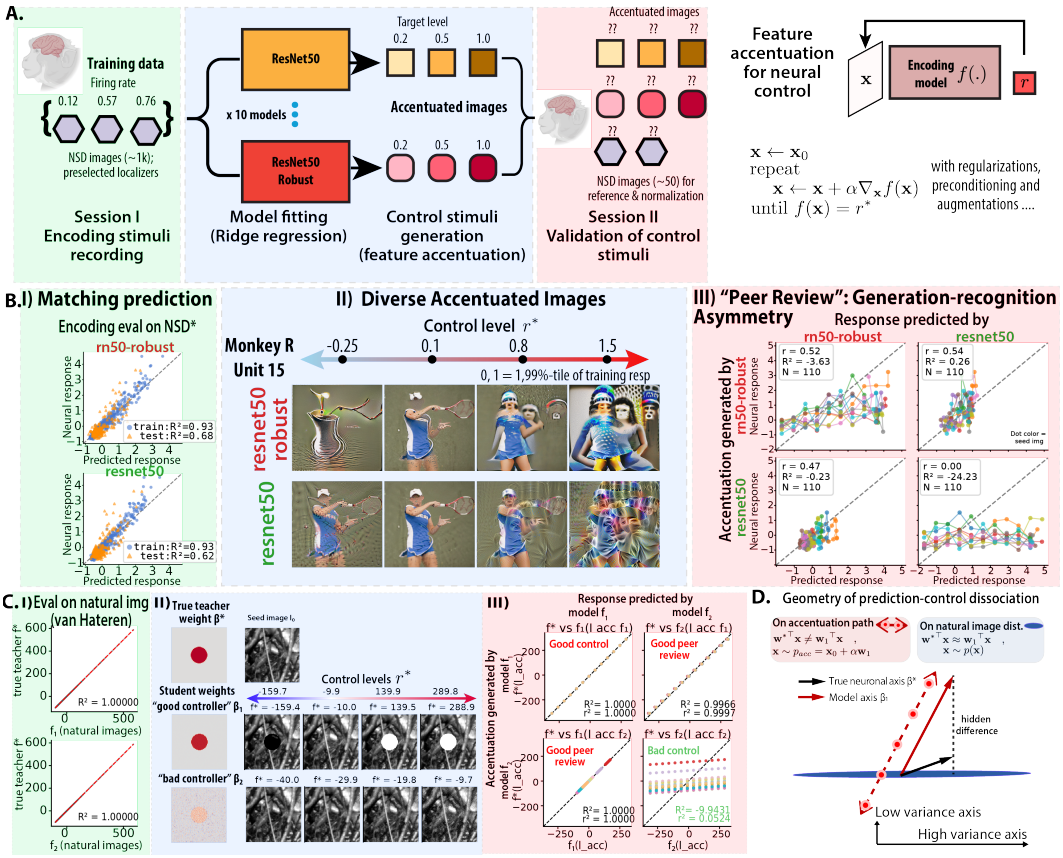

Figure 1: **The dissociation of prediction and control in neurons and in a linear teacher–student model. A.** Closed-loop experiment of Prince, Wang, et al. (2026). Encoding models are fit to responses to natural images, generate accentuated images for graded target responses, and are tested in a second session. **B. I** Two models predict held-out natural-image responses equally well. **II** Their accentuations from the same seed differ visibly. **III** When each model evaluated on the other's images (*peer review*), the weak controller (RN50) tracks neuronal responses on the strong controller's (RN50-robust) images but its own images barely modulate the neuron. **C.**: two linear students of a linear teacher (a disk-shaped weight on van Hateren images), weight differing along low-variance image directions, reproduce I-III qualitatively. **D.**: prediction on natural data hides weight error in small variance dim; the accentuation path exposes the hidden Euclidean difference. Methods see App. B.6, B.2

centrate power at low spatial frequencies, like natural images, control better. The two adversarially trained models are the strongest controllers, but the association persists without them (Prince, Wang, et al., 2026).

These phenomena are puzzling at first glance, but linear models on natural images reproduce all three observations (Fig. 1**C**). We construct the bad controller to deviate from the teacher along the lowest-variance, high-spatial-frequency directions of the image covariance, the good controller to deviate mildly along mid-band PCs; then matched prediction, diverse accentuations, unequal control and good peer review all follow. Dissecting this idealistic set up, the rest of the paper explains why *finite-sample* regression on *noisy responses* produces such errors, why *cross-validation does not remove them*, and how *feature maps* can amplify the dissociation.

3 Three evaluations, three geometries of one predictor

Teacher and student. We idealize a visual neuron as a noisy linear teacher, its encoding model as a linear student, and natural image inputs as $\mathbf{x} \in \mathbb{R}^d, \mathbf{x} \sim \mathcal{N}(0, \Sigma)$, a centered Gaussian distribution with structured covariance $\Sigma = \sum_k s_k \mathbf{u}_k \mathbf{u}_k^\top$. A linear teacher with weight $\beta^\star = \sum_k b_k \mathbf{u}_k$ produces noisy responses $y = \mathbf{x}^\top \beta^\star + \sigma\epsilon, \epsilon \sim \mathcal{N}(0, 1)$, and $S := \beta^{\star\top} \Sigma \beta^\star$ is the noiseless natural response

Evaluation	Covariance	MSE E	R^2	Slope	Equal-error set
Generalization	Σ	$\Delta^\top \Sigma \Delta$	$1 - \Delta^\top \Sigma \Delta / S$	$\hat{\beta}^\top \Sigma \beta^\star / \hat{\beta}^\top \Sigma \hat{\beta}$	ellipsoid
Accentuation	$\mathrm{Var}[\alpha]\, \hat{\beta}\hat{\beta}^\top$	$S\,(1 - N/D)^2$	$1 - (D/N - 1)^2$	N/D	sphere
Peer review	$\mathrm{Var}[\alpha']\, \beta' \beta'^\top$	$S\,(D_{peer} - N_{peer})^2 / \lVert\beta'\rVert^4$	$1 - (D_{peer}/N_{peer} - 1)^2$	N_{peer}/D_{peer}	hyperplane

Table 1: Evaluation metrics on three distributions, and their geometries. $\Delta := \hat{\beta} - \beta^\star$; on the student's own path the true and predicted gains are $N := \hat{\beta}^\top \beta^\star$, $D := \hat{\beta}^\top \hat{\beta}$, and on a peer's path $N_{peer} := \beta^{\star\top} \beta'$, $D_{peer} := \hat{\beta}^\top \beta'$; $z := D/N - 1$ in each row. Last column: level sets of each row's metrics in the space of student weights $\hat{\beta}$ (Prop. 7).

variance. The student is $f(\mathbf{x}) = \mathbf{x}^\top \hat{\beta}$, fitted from n pairs $(X, \mathbf{y})$; for linear feature-space students (Sec. 5) $\hat{\beta}$ is the effective pixel weight $F^\top \hat{w}$, which is also the gradient $\nabla_{\mathbf{x}} f$.

Three distributions. Generalization evaluates f on fresh natural inputs, $\mathbf{x} \sim \mathcal{N}(0, \Sigma)$. Accentuation evaluates f on inputs it generates itself: gradient ascent from a seed $\mathbf{x}_0$ moves along $\nabla_{\mathbf{x}} f = \hat{\beta}$, so the idealized path is $\mathbf{x} = \mathbf{x}_0 + \alpha \hat{\beta}$ with α set by the target level[1]. We study $\mathbf{x}_0 = 0$ and the target range is variance-matched, $\mathrm{Var}[\alpha](\hat{\beta}^\top \hat{\beta})^2 = S$, so the student's *prediction* has natural dynamic range along its path. Peer review evaluates f on the path $\mathbf{x}_0 + \alpha \beta'$ generated by another student β'. Writing $\Delta = \hat{\beta} - \beta^\star$ and applying the standard formulas for MSE, R^2 and the slope of true-on-predicted response to each distribution (Lem. 5) gives Table 1.

Prediction is covariance-weighted, control is Euclidean. Generalization error is $E_{gen} = \Delta^\top \Sigma \Delta$: a weight error Δ along $\mathbf{u}_k$ costs s_k times its squared length, so errors in low-variance directions can be nearly invisible. For accentuation, weight error is evaluated on a single path $\hat{\beta}$, and every metric is a function of the scalar z, or equivalently the ratio between quadratic forms D/N,

$$z_{acc} := \frac{D}{N} - 1 = \frac{\hat{\beta}^\top (\hat{\beta} - \beta^\star)}{\hat{\beta}^\top \beta^\star}, \qquad \frac{E_{acc}}{S} = \left(\frac{z}{1+z}\right)^2, \quad R^2_{acc} = 1 - z^2, \quad \mathrm{slope}_{acc} = \frac{1}{1+z}, \tag{1}$$

The three evaluations therefore have three geometries in the space of student weights (Table 1; Prop. 7; Fig. 2). **Prediction: ellipsoids.** The equal-E_{gen} sets are ellipsoids centered on $\beta^\star$ with semi-axis $\sqrt{E_{gen}/s_k}$ along $\mathbf{u}_k$, hence elongated along low-variance axes, where large weight errors are cheap. **Control: spheres.** The equal-z_{acc} sets are Euclidean spheres whose diameter joins 0 and $(1+z)\beta^\star$; the diameter encodes the error directly, and zero-error students live on the sphere through $\beta^\star$ at $z = 0$, not only at $\beta^\star$ itself. **Peer review: hyperplanes.** The equal-z_{peer} sets are hyperplanes normal to the generator β', so a peer reviewer is judged only on its component along the peer's path. A student can sit deep inside a small prediction ellipsoid, lie on a huge control sphere, and still lie on the zero-error hyperplane of a good generator: this is exactly the configuration of Fig. 1**C**, and it can be made arbitrarily extreme by adding weight along a direction with $s_k \to 0$ (Exp . 8).

The remaining question is therefore not whether the dissociation is possible but when *learning* produces it. We study three ingredients: finite-sample regression on noisy responses, the choice of regularization, and the feature map through which the gradient reaches the pixels.

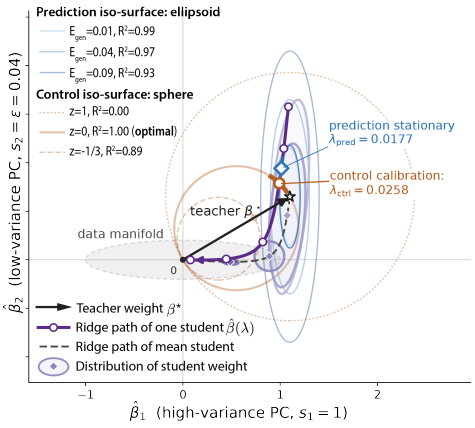

Figure 2: **Equal-error sets and the ridge path** in the plane of a high- and a low-variance PC. Prediction levels are ellipsoids around $\beta^\star$ (blue), control levels are the spheres with diameter joining 0 and $(1+z)\beta^\star$ (orange). The fitted weight traverses this plane as a function of λ (purple), scattering around the mean path (dashed black). The prediction-optimal penalty is the tangency of the path with an ellipsoid; perfect control is its crossing with the $z = 0$ sphere. Exact 2d construction: App. A.1.

[1] In practice, accentuation for DNN adds gradient preconditioning and augmentation (Hamblin et al., 2024; Fel et al., 2023), and some add image priors (Bashivan et al., 2019).

4 PIXEL-SPACE RIDGE: NOISE, SPECTRUM AND REGULARIZATION

Ridge regression in pixel space, $\hat{\beta} = \arg\min_\beta \frac{1}{n}\|X\beta - \mathbf{y}\|^2 + \lambda\|\beta\|^2$, yields

$$\hat{\beta} = \underbrace{(\hat{\Sigma} + \lambda I)^{-1}\hat{\Sigma}\beta^\star}_{\text{mean: signal shrinkage}} + \underbrace{\sigma(\hat{\Sigma} + \lambda I)^{-1}\tfrac{1}{n}X^\top \eta}_{\text{noise}}, \qquad \hat{\Sigma} = X^\top X/n, \quad \mathbf{y} = X\beta^\star + \sigma\eta \tag{2}$$

Since ridge penalizes weight in low-variance directions, one might expect it to remove these control-vulnerable components. We ask whether such control risk survive a cross-validated penalty.

The ridge path through the two landscapes. The question has a geometric reading (Fig. 2). For a given training set, Eq. 2 traces a path $\hat{\beta}(\lambda)$ in weight space from the least-squares or minimum-norm solution at $\lambda \to 0$ to the origin at $\lambda \to \infty$. Response noise and the random design scatter the fitted weight in a cloud around its mean. Ideally, cross-validation selects the point where the path is tangent to the smallest reachable prediction ellipsoid; whereas minimal control error asks where the path crosses the $z = 0$ sphere $D = N$; in general these two points do not coincide. The analysis below makes this geometric picture quantitative.

Toolkit. We use deterministic equivalents (DEs) in the proportional limit $d/n \to \gamma$ (Dobriban & Wager, 2018; Hastie et al., 2022). They replace functions of the empirical covariance $\hat{\Sigma}$ by deterministic functions of the population covariance Σ, e.g. $\hat{\Sigma}(\hat{\Sigma} + \lambda I)^{-1} \asymp \Sigma(\Sigma + \kappa I)^{-1}$, where the finite-sample randomness is absorbed into an effective penalty $\kappa(\lambda) \geq \lambda$, the unique positive root of $\kappa - \lambda = \frac{\kappa}{n}\operatorname{Tr}[\Sigma(\Sigma + \kappa I)^{-1}]$, one-to-one in λ (App. L). Applied to Eq. 2, the DE says that the signal term concentrates around a *mean fitted weight*, the teacher shrunk mode by mode,

$$\bar{\beta}_\kappa := \Sigma(\Sigma + \kappa I)^{-1}\beta^\star = \sum_k \frac{s_k}{s_k + \kappa} b_k \mathbf{u}_k, \tag{3}$$

while the fluctuation around the mean causes the *norm inflation* $V := \mathbb{E}\|\hat{\beta}\|^2 - \|\bar{\beta}\|^2 \geq 0$, which, as we will show, drives accentuation error. The mean weight $\bar{\beta}$ and inflation V also organize the feature-space case and the nonlinear diagnostic (Secs. 5, 6). Results use the teacher-weighted and unweighted spectral sums

$$B_{p,q}(\kappa) = \sum_k \frac{s_k^p}{(s_k + \kappa)^q} b_k^2, \qquad \mathrm{df}_{p,q}(\kappa) = \sum_k \frac{s_k^p}{(s_k + \kappa)^q}, \qquad \mathrm{df}_2 := \mathrm{df}_{2,2}, \tag{4}$$

so that, for instance, $\beta^{\star\top}\bar{\beta}_\kappa = B_{1,1}$ and $\|\bar{\beta}_\kappa\|^2 = B_{2,2}$. For the ratio N/D we use the leading approximation μ_N/μ_D from the DEs of N and D; second-order and distributional refinements matter only where the leading calibration error is already small (App. J).

Proposition 1 (Pixel ridge: prediction and control)**.** *With κ, B and* df *built from Σ,*

$$\begin{aligned} E_{gen} \asymp E_{gen,DE} &= \frac{n(\kappa^2 B_{1,2} + \sigma^2)}{n - \mathrm{df}_2} - \sigma^2, \qquad N \asymp \mu_N = \beta^{\star\top}\bar{\beta}_\kappa = B_{1,1}, \\ D \asymp \mu_D &= \|\bar{\beta}_\kappa\|^2 + V_{\mathrm{pix}}(\kappa), \qquad V_{\mathrm{pix}}(\kappa) = \frac{\mathrm{df}_{1,2}(\kappa)}{n}\big(E_{gen,DE} + \sigma^2\big). \end{aligned} \tag{5}$$

(Proofs and the per-mode decomposition behind these sums: App. F.)

E_{gen} recovers the classical in-distribution generalization error (Dobriban & Wager, 2018; Hastie et al., 2022). The two control moments have a plain reading. N is the alignment of the teacher with the mean fitted weight, and the first term of D is that weight's squared norm; both are set by the teacher and the shrinkage alone, and since each factor $s_k/(s_k + \kappa)$ lies in $[0, 1]$ they are ordered $\|\beta^\star\|^2 \geq B_{1,1} \geq B_{2,2}$ (Eq. 119), so the mean weight by itself *under*-shoots its own path ($D < N$, slope above one). The second term of D is the key issue. V_{pix} summarizes how limited data n, response noise σ^2 and prediction error E_{gen} jointly inflate the student norm, amplified by the spectral susceptibility $\mathrm{df}_{1,2} = \sum_k s_k/(s_k + \kappa)^2$, which counts the number of modes near the regularization threshold κ. Inflation enlarges D without changing N, so the student can easily *over*-predict: $\mathrm{slope}_{acc} = N/D$ drops below one, and $z = D/N - 1$ can grow until $R^2_{acc} = 1 - z^2$ turns negative. Note that V is not necessarily bad for control: at the special point $B_{1,1} = B_{2,2} + V_{\mathrm{pix}}$ the inflation exactly replenishes the norm lost to shrinkage, so $N = D$ and control is perfect at leading order (the valley in Fig. 3**B**). This is generally not where cross-validation lands, as the next proposition shows.

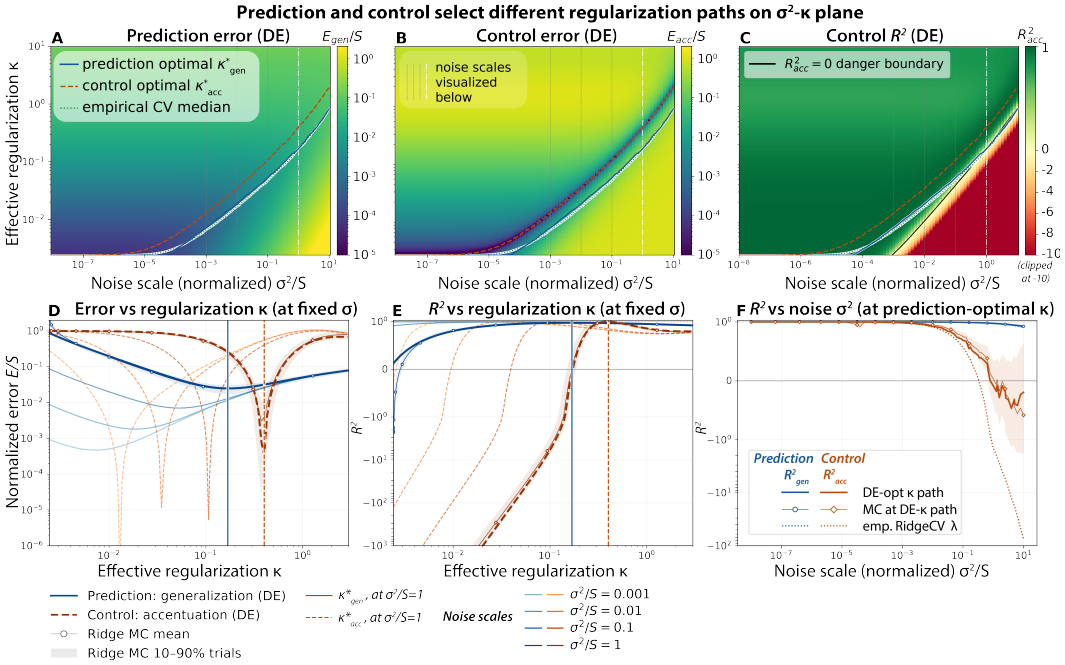

Figure 3: **Pixel-space ridge separates prediction- and control-optimal regularization on natural images.** Disk teacher on van Hateren images ($d = 10^4$, $n = 10^3$). **A–C** DE landscapes (Eq. 5) of E_{gen}, E_{acc} and R^2_{acc} over noise σ^2/S and effective penalty κ (Monte Carlo: Fig. 7). The prediction-optimal and median empirical RidgeCV paths track each other but not the control-optimal path; black contour: $R^2_{acc} = 0$. **D,E** Slices at four noise levels expose displaced prediction and control valleys (or peaks for R^2); markers and bands show the mean and 10–90% range over 100 natural-image trials at the highlighted noise $\sigma^2/S = 1$. **F** Along the prediction-optimal path, R^2_{gen} remains high while R^2_{acc} becomes strongly negative; empirical RidgeCV exhibits the same dissociation. Methods see App. B.4, A.2.

Proposition 2 (Prediction-optimal and control-optimal penalties differ)**.** *Among admissible values of κ, let $\kappa^\star_{\mathrm{gen}}$ be an interior differentiable minimizer of $E_{\mathrm{gen,DE}}$, and let $\kappa^\star_{\mathrm{acc}}$ denote a zero of the leading control error, equivalently satisfying $\mu_D = \mu_N$. Their defining conditions are*

$$E_{\mathrm{gen,DE}} + \sigma^2 = n\kappa \frac{B_{2,3}}{\mathrm{df}_{2,3}}, \qquad (\kappa = \kappa^\star_{\mathrm{gen}};\ \textit{prediction stationarity}), \tag{6}$$

$$E_{\mathrm{gen,DE}} + \sigma^2 = n\kappa \frac{B_{1,2}}{\mathrm{df}_{1,2}}, \qquad (\kappa = \kappa^\star_{\mathrm{acc}};\ \textit{zero leading control error}). \tag{7}$$

Further, at the prediction optimum $\kappa^\star_{\mathrm{gen}}$ the control error is

$$z^{(0)}_{acc,CV} = \frac{\kappa\, \mathrm{df}_{1,2}}{B_{1,1}} \left(\frac{B_{2,3}}{\mathrm{df}_{2,3}} - \frac{B_{1,2}}{\mathrm{df}_{1,2}} \right)_{\kappa = \kappa^\star_{gen}}. \tag{8}$$

(Proofs: App. G.)

Geometrically, along the ridge path (Fig. 2), $\kappa^\star_{\mathrm{gen}}$ lands at the tangency with the prediction ellipsoid and $\kappa^\star_{\mathrm{acc}}$ at its crossing of the control sphere. Both conditions equate the noisy prediction risk to $n\kappa$ times a spectral-band average of the teacher's per-mode power b_k^2, but the two bands sit at different places (Fig. 26). The control kernel $s_k/(s_k+\kappa)^2$ is a bump centered at $s_k = \kappa$, symmetric on a log axis; the prediction kernel $s_k^2/(s_k+\kappa)^3$ is the same bump multiplied by the shrinkage factor $s_k/(s_k+\kappa)$, which shifts it up to $s_k \approx 2\kappa$ and suppresses its low-variance side. Consequently $z^{(0)}_{acc,CV}$ compares the teacher power around 2κ with that around κ. It vanishes for isotropic data, or for a teacher whose per-mode power b_k^2 is constant across modes. For *an anisotropic spectrum and a structured teacher* the averages differ, and their difference is the control error left at the cross-validated penalty. When b_k^2 decreases toward the low-variance tail, as for a teacher aligned with high-variance image modes, $z^{(0)}_{acc,CV} > 0$ and the cross-validated student over-predicts ($\mathrm{slope}_{acc} < 1$).

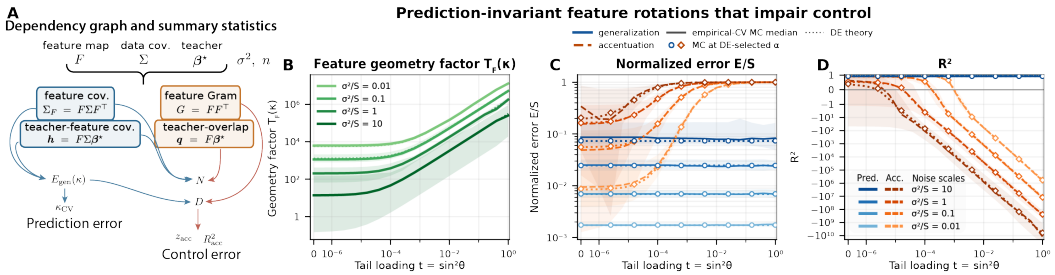

Figure 4: **Feature geometry decides control.** **A** Dependency graph: $(\Sigma_F, F\Sigma\beta^\star, S, \sigma^2, n)$ fix generalization and the cross-validated penalty; control additionally needs $G := FF^\top$ and $F\beta^\star$ (full graph: Fig. 27). **B-D** For van Hateren images, rotating the top-500-PC projector feature into the spectral tail: E_{gen} and R^2_{gen} are invariant while R^2_{acc} and the slope collapse; DE (lines) matches Gaussian-design Monte Carlo (markers). Construction: App. B.5; further rotations: App. A.3.

Over the noise–regularization plane for natural images (Fig. 3**A–C**), the DE tracks Monte Carlo except at very low noise or regularization (Fig. 7). Prediction- and control-optimal penalties form two distinct valleys, the E_{acc} valley a sharp cusp. At low noise, $\kappa^\star_{\text{gen}}$ is also benign to control; as noise grows it approaches and crosses the $R^2_{acc} = 0$ cliff, and control fails catastrophically.

This answers the question in the title for pixel regression: *control is bad, despite prediction-optimal regularization, when the response is noisy, the spectrum is anisotropic, and many modes sit near* κ, which is exactly the noisy neuron with natural-image regime.

Peer review: recognition without generation. Peer review evaluates $\hat\beta$ on the path of a peer β', so the roles of N and D are played by the true and predicted gains along the peer's path, $N_{peer} = \beta^{\star\top}\beta'$ and $D_{peer} = \hat\beta^\top\beta'$ of Table 1. Take the peer to be a ridge fit at the same κ on independent data. Both fits then concentrate around the same mean weight $\bar\beta_\kappa$ while their noise parts are uncorrelated, so the norm inflation drops out of the cross product: $D_{peer} \asymp \|\bar\beta_\kappa\|^2 = B_{2,2}$ and $N_{peer} \asymp \beta^{\star\top}\bar\beta_\kappa = B_{1,1}$ (Prop. 12). The peer slope and score are therefore

$$\text{slope}_{peer} = \frac{N_{peer}}{D_{peer}} \asymp \frac{B_{1,1}}{B_{2,2}} \in [1,\infty), \qquad R^{2,(0)}_{peer} = 1 - \left(1 - \frac{B_{2,2}}{B_{1,1}}\right)^2 \in (0,1], \tag{9}$$

using $B_{1,1} \geq B_{2,2}$ (Eq. 119). Note, on the leading order, the peer slope is never below one, i.e. the *student under-predicts the teacher on its peer's path*, and the peer R^2 is non-negative and degrades gracefully with noise through the selected κ. In contrast, on the self-generated path the norm inflation V_{pix} of Eq. 5 adds to D but not to N: so the student over-predicts the teacher ($\text{slope}_{acc} = N/D < 1$), and z_{acc} can grow without bound, driving R^2_{acc} negative. The sign differences in neural data confirm this leading order picture: in Prince, Wang, et al. (2026), peer-review scores between pairs of models have $R^2 = 0.04, [-0.06, 0.10]$, slope $= 1.35, [1.14, 1.59]$ (median, [IQR], $n = 90$), whereas self-evaluation scores have $R^2 = -4.17, [-4.59, -3.51]$, slope $= 0.38, [0.33, 0.42]$ ($n = 10$; App. A.4, analysis in App. B.9). Note, in simplistic pixel space peer review, the slope and R^2 are symmetric between models ($A \leftarrow B$ or $B \leftarrow A$). In practice the students are fitted to the same recordings through different feature maps; the review score then becomes asymmetric, and the evaluator can recognize its own noise in the generator's path only to the extent that the two maps route noise into the same pixel directions (Sec. 5; App. J.4).

5 Linear feature ridge: effect of gradient geometry

Pixel regression explains why one student fails; it cannot explain why ten equally predictive backbones fail differently. We idealize a backbone as a fixed linear map $F \in \mathbb{R}^{p\times d}$, fit $\hat w$ by ridge on features $F\mathbf{x}$, and accentuate along the pixel gradient $\hat\beta = F^\top\hat w$. Let $\Sigma_F := F\Sigma F^\top$ be the feature covariance and $G := FF^\top$ the Euclidean feature Gram. The best realizable feature weight and its irreducible error are $w^\star = \Sigma_F^\dagger F\Sigma\beta^\star$ and $S_\perp = S - w^{\star\top}\Sigma_F w^\star$ (App. 14).

Proposition 3 (Feature ridge: prediction and control)**.** *With* κ, B^F *and* df^F *built from* $(\Sigma_F, w^\star)$ *and aspect ratio* p/n,

$$E_{gen,DE} = \frac{n(\kappa^2 B^F_{1,2} + S_\perp + \sigma^2)}{n - \mathrm{df}^F_2} - \sigma^2, \qquad N_{DE} = \beta^{\star\top}\bar\beta_{F,\kappa}, \qquad D_{DE} = \|\bar\beta_{F,\kappa}\|^2 + V_F(\kappa). \tag{10}$$

where the mean fitted pixel weight $\bar{\beta}_{F,\kappa}$ is the realizable weight shrunk in feature space and back-projected, and the norm inflation carries a map-dependent gradient-geometry factor T_F*:*

$$\bar{\beta}_{F,\kappa} := F^\top(\Sigma_F + \kappa I)^{-1} F\Sigma\beta^\star, \quad V_F(\kappa) := \frac{1}{n}(E_{gen,DE} + \sigma^2)\, \underbrace{\mathrm{Tr}\left[G(\Sigma_F + \kappa I)^{-2}\Sigma_F\right]}_{T_F(\kappa):=}, \tag{11}$$

Setting $F = I$ gives $\bar{\beta}_{F,\kappa} = \bar{\beta}_\kappa$ and $T_F = \mathrm{df}_{1,2}$, recovering Prop. 1. (Proofs: App. H.)

The structure is the same as in pixel ridge, so the interpretation carries over: N is the teacher's alignment with the mean weight, D is that weight's norm plus an inflation, and the unrealizable signal $S_\perp$ simply acts as extra response noise. What changes is what determines the two objects. Prediction sees the feature map only through Σ_F and $F\Sigma\beta^\star$: the whole curve $E_{gen,DE}(\kappa)$, hence the cross-validated κ and R^2_{gen}, is a function of $(\Sigma_F, F\Sigma\beta^\star, S, \sigma^2, n)$. Control additionally depends on the Gram G and on the teacher overlap $F\beta^\star$, (Fig. 4**A**; Cor. 16). Thus the statistics governing control generically exceed those governing prediction. They coincide only in special cases, for instance white data $\Sigma = I$, where $G = \Sigma_F$ and $F\beta^\star = F\Sigma\beta^\star$ (App. I). This extra degree of freedom can be seen clearly through the group that transforms feature map without affecting prediction.

Proposition 4 (Prediction-preserving gauge)**.** *For $\Sigma \succ 0$ and any orthogonal Q with $Q\Sigma^{1/2}\beta^\star = \Sigma^{1/2}\beta^\star$, the map $F \mapsto F\Sigma^{1/2}Q\Sigma^{-1/2}$ preserves Σ_F and $F\Sigma\beta^\star$, hence the joint feature–response covariance, $E_{gen,DE}(\kappa)$ and the DE-based prediction-optimal penalty $\kappa^\star_{gen}$, while changing G and $F\beta^\star$ and therefore control (Prop. 18; construction in App. I).*

Note, the gauge acts nontrivially on control only when the input is anisotropic: in white data, control error is also invariant to this group. Variation of accentuation error grows with anisotropy. Rotating within it moves feature energy from high-variance to low-variance pixel directions at no cost to prediction (Fig. 4B): along a continuous path, E_{gen}, R^2_{gen} and the selected penalty are exactly constant, while $\mathrm{Tr}(FF^\top)$ grows by four orders of magnitude and R^2_{acc} falls from near one to -10^8.

What makes a map fragile. The response-independent part of V_F is the geometric factor T_F, which has the spectral form

$$T_F(\kappa) = \sum_{k:s_k>0} \left(\frac{s_k}{s_k+\kappa}\right)^2 \frac{\|g_k\|^2}{g_k^\top \Sigma g_k}, \qquad g_k := F^\top v_k, \tag{12}$$

with (s_k, v_k) the eigenpairs of Σ_F (Prop. 20). It averages, over the feature PCs that survive ridge shrinkage ($s_k^2/(s_k+\kappa)^2 \sim 1$), the ratio between the Euclidean energy of the backpropagated PC gradient $\|g_k\|^2$ and its energy under the natural-image covariance $g_k^\top \Sigma g_k$. Thus, a map is fragile when its *well-resolved feature directions backpropagate into low-variance, high-spatial-frequency pixel directions*. This is the linear counterpart of the third observation of Sec. 2, and it can be computed from the map and the image statistics alone.

Summary: When is control fragile although prediction is fine? Control fails when the norm inflation V_F overshoots the shrinkage deficit (Eqs. 5, 10). The inflation $V_F = (E_{gen} + \sigma^2)\, T_F(\kappa)/n$ is a product of three factors, **(i) response noise** $E_{gen} + \sigma^2$; **(ii) limited data** $1/n$; **(iii) gradient geometry** T_F, the ratio of Euclidean to covariance-weighted gradient energy, a property of the feature map; **(iv) mismatched regularization**: on anisotropic data, the cross-validated penalty is designed for prediction, but not control-optimal and can leave a large control error (Eq. 8). Held-out accuracy reflects only (i); (iii) and (iv) are invisible to it. Neuronal recordings are noisy and short (Goris et al., 2014; Churchland, 2010), supplying (i)–(ii); natural images make (iv) large for every model; and the feature map decides (iii), in which equally predictive maps differ without bound.

6 Nonlinear feature ridge: a generalized diagnostic for DNN

For linear feature, V_F is the main culprit of accentuation error. For nonlinear features ϕ, is there an analogous quantity that preludes control failure? Viewing a nonlinear feature map ϕ as locally linear, we use its Jacobian $J_\phi(x)$ to replace F in Eq. 12,

$$T_\phi(x,\kappa) = \sum_k \frac{s_k}{(s_k+\kappa)^2} \|J_\phi(x)^\top v_k\|^2, \qquad V_\phi := \frac{E_{val}}{n}\, \mathbb{E}_x\left[T_\phi(x,\kappa)\right], \tag{13}$$

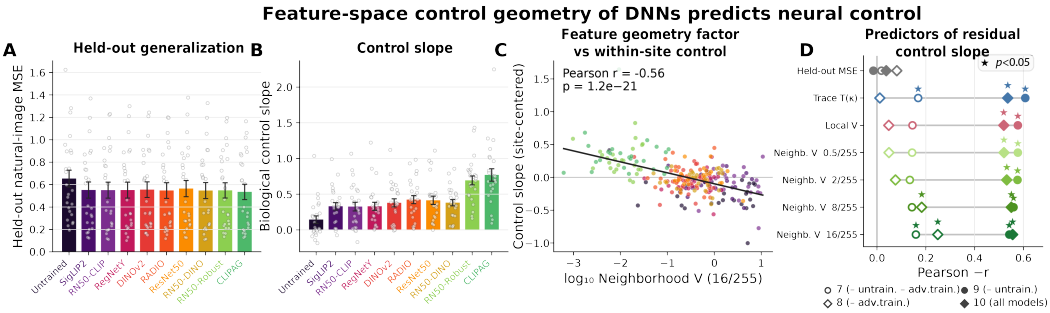

Figure 5: **Gradient-feature geometry in nonlinear model predicts biological control across encoding models.** Published preprocessed data of Prince, Wang, et al. (2026). **A** Held-out natural-image MSE in the control session ($E_{\text{val}} \approx E_{\text{gen}} + \sigma^2$). **B** Slope of observed-over-predicted responses to accentuated stimuli. **A,B**: mean over 25 sites (bars), sites (dots), ±1 SE. **C** Raw $\log_{10} V_\phi$ (16/255 neighborhood) against control slope with each site's ten-model mean removed (Pearson $r = -0.56$, 250 site–model pairs). **D** Pearson $-r$ between site-centered control slope and seven raw $\log_{10}$ predictors (held-out MSE, $T_\phi(\kappa)$, exact-local and neighborhood V_ϕ at four scales) for 10/9/8/7-model subsets (all / no untrained AlexNet / no robust models / neither); stars: raw two-sided $p < 0.05$. All estimators: App. A.5.

where (s_k, v_k) is the eigenpairs of feature covariance on natural-images, and $J_\phi(x)^\top v_k = \nabla_x [v_k^\top \phi(x)]$ is the pixel gradient of feature PC k, obtainable by one backward pass per PC, and V_ϕ mirrors the norm inflation V_F of Eq. 11 with $E_{val} \approx E_{gen} + \sigma^2$. Since an accentuation path travels far from the seed, we also evaluate smoothed versions in which the Jacobian is averaged, or the squared gradient is averaged, over a Gaussian neighborhood of the seed at pixel-noise scales from $0.5/255$ to $16/255$ (App. K).

We computed these quantities for all 250 site–model pairs (App. B.8), with κ set by each model's cross-validated penalty and feature spectrum, average over the ten seed images, and E_{val} the held-out error of the control session. Because sites differ greatly in controllability, we subtract each site's mean from the control outcome and correlate this residual with the predictors in log10 (Pearson r; Fig. 5**C,D**). Held-out error carries essentially no ordering information ($r = 0.02$ with control slope, -0.04 with control error, across the nine trained models), whereas exact-local V_ϕ reaches $r = -0.58$ and $+0.40$; the trace T_ϕ alone does as well (-0.61, $+0.43$), so the ordering mainly comes from gradient geometry. Among the seven conventionally trained models the slope association is weaker but same-signed ($r \approx -0.15$ to -0.17): the two adversarially trained models are the strongest controllers and have by far the smallest T_ϕ. Neighborhood smoothing ($16/255$) keeps more of the ordering without the robust models ($r = -0.25$ vs. -0.05 for exact-local V_ϕ over eight models). Thus, the gradient-geometry factor of linear feature maps carries over to deep networks, ranking equally predictive encoding models by *in vivo* control where held-out accuracy cannot.

7 Discussion

A single mechanism accounts for the three neuronal phenomena: prediction measures weight error in the covariance metric, control measures it along the model's own gradient. In the linear theory the control error at fixed prediction error is governed by the three factors: noise, spectrum, and map, of which only the first is visible from held-out accuracy. The practical implication is that selecting encoding models or penalties by held-out prediction cannot certify gradient-based control; validation must constrain the gradient, for instance through T_ϕ, or test generated stimuli directly. The high-frequency artifacts of failed accentuations are not a cosmetic defect but the expected signature of low-variance directions that natural images never constrained.

Limitations. The theory assumes Gaussian inputs, a linear teacher, a fixed linear feature map and a linear accentuation path from a centered seed, while some phenomena (e.g. non-trivial Pearson r) requires multiple non-centered seeds. The details of accentuation procedure (e.g. preconditioning and image prior) could be built in as natural extensions. The leading DEs predict typical fits and the location of the control optimum but not the fluctuation-dominated error at that optimum (App. J). Our theory did not prove the relation between the diagnostic and control error for nonlinear features; its biological validation is descriptive.

8 Reproducibility Statement

The theory assumptions, derivations, and deterministic equivalents are stated in the main text, with full derivations in the appendix. The code repository (will be released after camera ready) contains the core RMT library, CPU and GPU Monte Carlo scripts, notebooks recreating the figures, and saved outputs for the experiments shown here.

9 Ethics Statement

This work is theoretical and uses synthetic simulations, natural-image covariance spectra, and previously published neuronal recordings for validation. It does not introduce a deployed model, collect human-subject data, or release a dataset with personal information.

AI Use Statement

OpenAI ChatGPT, Codex and Anthropic Claude, including Claude Code, were used as general-purpose research, coding, and writing assistants throughout this work. The theory was developed iteratively between the authors and these systems. In some instances, the authors first derived a result and used an AI system to check the algebra, simplify the equations, or identify alternative interpretations; in others, an AI system proposed a derivation, simplification, or reading that the authors subsequently checked, corrected, and refined. The final conceptual synthesis emerged through this iterative process, while the authors determined the research questions, selected the scientific directions, and made all final decisions about the claims. The systems also assisted with designing, generating, debugging, profiling, executing, and documenting computational experiments and figure-generation code, as well as organizing, drafting, editing, and typesetting the manuscript. The authors reviewed all retained derivations and prose, inspected the code, validated numerical results against theory and direct computation, and checked the supporting references. The authors take full responsibility for the accuracy and integrity of the paper. No AI system is listed as an author.

References

Pouya Bashivan, Kohitij Kar, and James J. DiCarlo. Neural population control via deep image synthesis. *Science*, 364(6439):eaav9436, 2019. doi: 10.1126/science.aav9436.

Santiago A. Cadena, George H. Denfield, Edgar Y. Walker, Leon A. Gatys, Andreas S. Tolias, Matthias Bethge, and Alexander S. Ecker. Deep convolutional models improve predictions of macaque V1 responses to natural images. *PLoS Computational Biology*, 15(4):e1006897, 2019. doi: 10.1371/journal.pcbi.1006897.

Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Out-of-distribution generalization in kernel regression. *Advances in Neural Information Processing Systems*, 34:12600–12612, 2021.

M. M. Churchland. Stimulus onset quenches neural variability: a widespread cortical phenomenon. *Nat. Neurosci.*, 13, 2010. doi: 10.1038/nn.2501. URL https://doi.org/10.1038/nn.2501.

Colin Conwell, Jacob S. Prince, Kendrick N. Kay, George A. Alvarez, and Talia Konkle. A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nature Communications*, 15(1):9383, 2024. doi: 10.1038/s41467-024-53147-y.

Joel Dapello, Tiago Marques, Martin Schrimpf, Franziska Geiger, David D. Cox, and James J. DiCarlo. Simulating a primary visual cortex at the front of CNNs improves robustness to image perturbations. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018.

Jenelle Feather, Guillaume Leclerc, Aleksander Mądry, and Josh H. McDermott. Model metamers reveal divergent invariances between biological and artificial neural networks. *Nature Neuroscience*, 26(11):2017–2034, 2023. doi: 10.1038/s41593-023-01442-0.

Thomas Fel, Thibaut Boissin, Victor Boutin, Agustin Picard, Paul Novello, Julien Colin, Drew Linsley, Tom Rousseau, Rémi Cadène, Lore Goetschalckx, Laurent Gardes, and Thomas Serre. Unlocking feature visualization for deep network with MAgnitude constrained optimization. In *Advances in Neural Information Processing Systems*, volume 36, pp. 37813–37826, 2023.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019.

Tal Golan, Prashant C. Raju, and Nikolaus Kriegeskorte. Controversial stimuli: Pitting neural networks against each other as models of human cognition. *Proceedings of the National Academy of Sciences*, 117(47):29330–29337, 2020. doi: 10.1073/pnas.1912334117.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

R. L. T. Goris, J. A. Movshon, and E. P. Simoncelli. Partitioning neuronal variability. *Nat. Neurosci.*, 17, 2014. doi: 10.1038/nn.3711. URL https://doi.org/10.1038/nn.3711.

Umut Güçlü and Marcel A. J. van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27): 10005–10014, 2015. doi: 10.1523/JNEUROSCI.5023-14.2015.

Chong Guo, Michael Lee, Guillaume Leclerc, Joel Dapello, Yug Rao, Aleksander Mądry, and James DiCarlo. Adversarially trained neural representations may already be as robust as corresponding biological neural representations. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 8072–8081. PMLR, 2022.

Chris Hamblin, Thomas Fel, Srijani Saha, Talia Konkle, and George Alvarez. Feature accentuation: Revealing 'what' features respond to in natural images. *arXiv preprint arXiv:2402.10039*, 2024.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20617–20642, 2024.

Chaitanya Kapoor, Sudhanshu Srivastava, and Meenakshi Khosla. Bridging critical gaps in convergent learning: How representational alignment evolves across layers, training, and distribution shifts. In *Advances in Neural Information Processing Systems*, volume 38, pp. 155718–155732, 2025. doi: 10.52202/085713-4690.

Aleksander Mądry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022.

Gabriel Mel and Surya Ganguli. A theory of high dimensional regression with arbitrary correlations between input features and target functions: sample complexity, multiple descent curves and a hierarchy of phase transitions. In *International Conference on Machine Learning*, pp. 7578–7587. PMLR, 2021.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 427–436, 2015. doi: 10.1109/CVPR.2015.7298640.

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. In *Distill*, 2017. URL https://distill.pub/2017/feature-visualization/.

Pratik Patil, Jin-Hong Du, and Ryan J Tibshirani. Optimal ridge regularization for out-of-distribution prediction. *arXiv preprint arXiv:2404.01233*, 2024.

Carlos R. Ponce, Will Xiao, Peter F. Schade, Till S. Hartmann, Gabriel Kreiman, and Margaret S. Livingstone. Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. *Cell*, 177(4):999–1009.e10, 2019. doi: 10.1016/j.cell.2019.04.005.

Jacob S. Prince, Colin Conwell, George A. Alvarez, and Talia Konkle. A case for sparse positive alignment of neural systems. In *ICLR 2024 Workshop on Representational Alignment (Re-Align)*, 2024.

Jacob S. Prince, Binxu Wang, Thomas Fel, Akshay V. Jagadeesh, Parisa A. Vaziri, George A. Alvarez, Margaret S. Livingstone, and Talia Konkle. Parametric neural control differentiates top neural network models of primate visual cortex. *bioRxiv*, 2026. doi: 10.64898/2026.08.16.745063. URL https://www.biorxiv.org/content/10.64898/2026.08.16.745063v1. Preprint, posted August 20, 2026.

Rishi Rajalingham, Elias B. Issa, Pouya Bashivan, Kohitij Kar, Kailyn Schmidt, and James J. DiCarlo. Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience*, 38(33):7255–7269, 2018. doi: 10.1523/JNEUROSCI.0388-18.2018.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv*, pp. 407007, 2018. doi: 10.1101/407007.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. URL https://arxiv.org/abs/2007.00644.

Nilesh Tripuraneni, Ben Adlam, and Jeffrey Pennington. Covariate shift in high-dimensional random feature regression. *arXiv preprint arXiv:2111.08234*, 2021.

Edgar Y. Walker, Fabian H. Sinz, Erick Cobos, Taliah Muhammad, Emmanouil Froudarakis, Paul G. Fahey, Alexander S. Ecker, Jacob Reimer, Xaq Pitkow, and Andreas S. Tolias. Inception loops discover what excites neurons most using deep predictive models. *Nature Neuroscience*, 22(12):2060–2065, 2019. doi: 10.1038/s41593-019-0517-x.

Will Xiao and Gabriel Kreiman. XDream: Finding preferred stimuli for visual neurons using generative networks and gradient-free optimization. *PLoS computational biology*, 16(6):e1007973, 2020. doi: 10.1371/journal.pcbi.1007973. Publisher: Public Library of Science San Francisco, CA USA.

Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014. doi: 10.1073/pnas.1403112111.

CONTENTS

1 Introduction **1**

2 Prediction and control dissociate in neurons and in linear regression **2**

3 Three evaluations, three geometries of one predictor **3**

4 Pixel-space ridge: noise, spectrum and regularization **5**

5 Linear feature ridge: effect of gradient geometry **7**

6 Nonlinear feature ridge: a generalized diagnostic for DNN **8**

7 Discussion **9**

8 Reproducibility Statement **10**

9 Ethics Statement **10**

A Extended results **15**

A.1 Geometry of the ridge path 15

A.2 Pixel-space ridge on natural images 16

A.3 Linear feature ridge: the prediction-null family 20

A.4 Peer review in the neuronal experiment 22

A.5 Nonlinear gradient diagnostic: full comparison 24

B Detailed methods **31**

B.1 Natural-image data and covariance 31

B.2 Toy teacher–student example 31

B.3 Two-dimensional geometry illustration 32

B.4 Natural-image pixel-ridge landscape experiment 32

B.5 Linear feature maps and prediction-null rotations 33

B.6 Neuronal data 35

B.7 Encoding models 36

B.8 Computing the gradient-geometry diagnostic 36

B.9 Neuronal peer-review analysis 37

B.10 Statistical analysis 38

B.11 Software and compute 38

C Theory results roadmap **40**

D Main theory results in derivation order **40**

D.1 Setting, estimator and evaluation distributions 40

D.2 Metrics on an arbitrary evaluation distribution . . . 41
D.3 Deterministic-equivalent toolkit . . . 42
D.4 Pixel-space ridge regression . . . 42
D.5 Beyond ratio of means . . . 43
D.6 Linear feature-space ridge regression . . . 43
D.7 Nonlinear feature maps: a heuristic diagnostic . . . 45

E Evaluation geometry: setting, metrics and equal-error sets **45**
E.1 Setting . . . 45
E.2 Metrics on an arbitrary evaluation distribution . . . 46
E.3 Generalization, accentuation and peer review . . . 47
E.4 Geometry of equal-error sets . . . 48

F Pixel-space ridge: proofs **49**
F.1 Generalization risk . . . 49
F.2 The accentuation forms N and D . . . 50
F.3 The leading control error z_{acc} . . . 50

G Cross-validation versus control-optimal regularization: proofs **51**
G.1 Prediction-optimal stationarity . . . 51
G.2 Control-optimal condition and the risk at the control optimum . . . 52
G.3 Residual control error at the prediction optimum . . . 52

H Feature-space regression: proofs **53**
H.1 Setting and estimator . . . 53
H.2 Realizable decomposition . . . 54
H.3 Generalization risk and accentuation forms . . . 54
H.4 Tuning conditions in feature space . . . 56

I Gauge invariance and gradient fragility: proofs **56**
I.1 Distinct summary statistics for prediction and control . . . 56
I.2 The prediction-preserving gauge family . . . 56
I.3 Constructing the family . . . 57
I.4 A spectral diagnostic for fragile feature maps . . . 58

J Ratios of random quadratic forms, and peer review **58**
J.1 Second-order expansion of a ratio . . . 59
J.2 Gaussian distributional surrogate . . . 59
J.3 Peer review between independent fits . . . 59
J.4 Peer review on shared data across feature maps . . . 60

K Nonlinear gradient diagnostic **61**

K.1 Empirical validation against neuronal control 62

L Random-matrix toolkit **62**

L.1 Fixed point and deterministic equivalents 62

L.2 Spectral sums 63

A Extended results

The appendix has three parts. This section collects supplementary figures that extend the main-text figures. Section B describes the numerical experiments and the neuronal analyses in detail. The theory follows from Section C on: every result stated in derivation order, then proofs grouped by topic.

A.1 Geometry of the ridge path

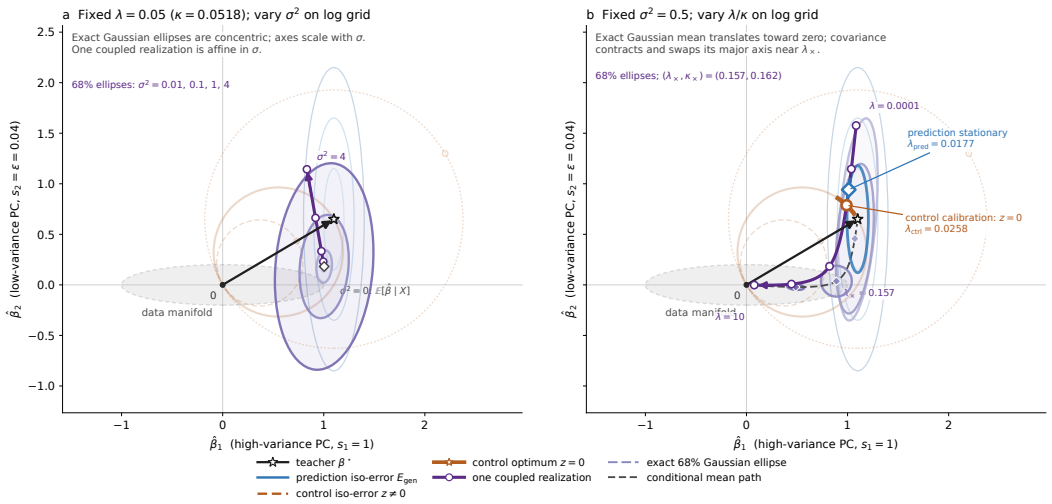

Figure 6: **Exact conditional geometry of ridge estimators in two dimensions**, in the plane of a high-variance ($s_1 = 1$) and a low-variance ($s_2 = 0.04$) PC for a fixed design X. **(a)** Fixed $\lambda = 0.05$ ($\kappa = 0.0518$), σ^2 on a log grid: the conditional law of $\hat{\beta} \mid X$ keeps its mean, its 68% ellipses grow with σ, and one coupled noise realization (purple) moves affinely in σ. **(b)** Fixed $\sigma^2 = 0.5$, λ on a log grid: the conditional mean path (dashed) shrinks toward the origin while the covariance contracts. Blue: prediction equal-error ellipses around $\beta^\star$; orange: control equal-error spheres (solid: $z = 0$). The prediction-stationary penalty ($\lambda_{\text{pred}} = 0.0177$) and the control-calibrated penalty ($\lambda_{\text{ctrl}} = 0.0258$) are distinct points on the path. Setting: App. B.3.

A.2 PIXEL-SPACE RIDGE ON NATURAL IMAGES

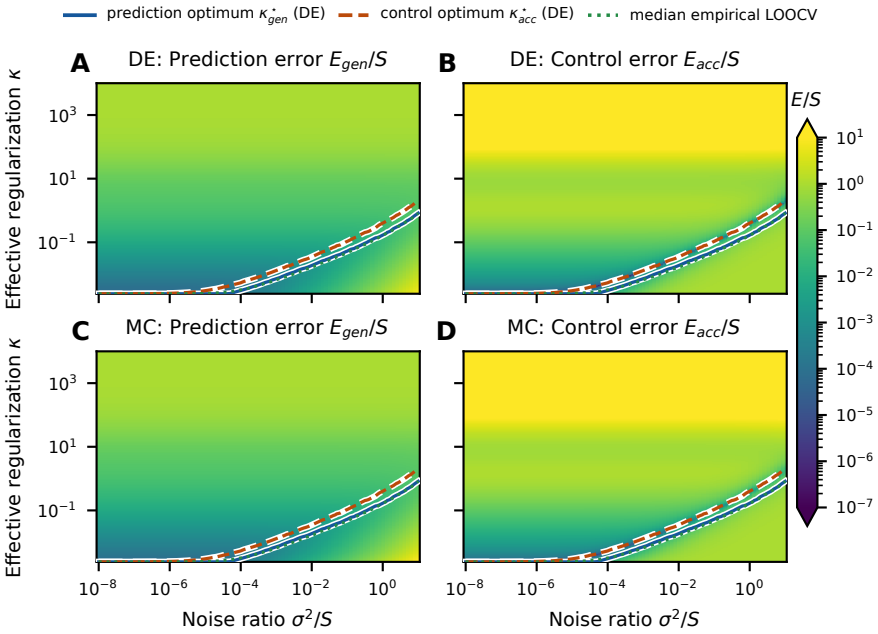

Figure 7: **Error landscapes: deterministic equivalent versus natural-image Monte Carlo** over the full grid of Fig. 3 (114 noise levels $\times$ 478 penalties, κ up to 10^4). **A, B** Leading DE of the normalized prediction error E_{gen}/S and control error E_{acc}/S (Eq. 5). **C, D** Mean over 100 natural-image Monte Carlo trials on the same grid. Lines (same in all panels): prediction optimum $\kappa^{\star}_{gen}$ (blue), control optimum $\kappa^{\star}_{acc}$ (orange dashed), both from the DE, and median empirical LOOCV penalty (green dotted). For $\kappa \geq 10^{-2}$ the Monte Carlo mean and the DE differ by a median of 2% for E_{gen} and 10% for E_{acc}; larger deviations are confined to the near-ridgeless floor $\kappa \approx 2.4 \times 10^{-3}$ and to the narrow E_{acc} valley, where the leading DE vanishes. Methods: App. B.4.

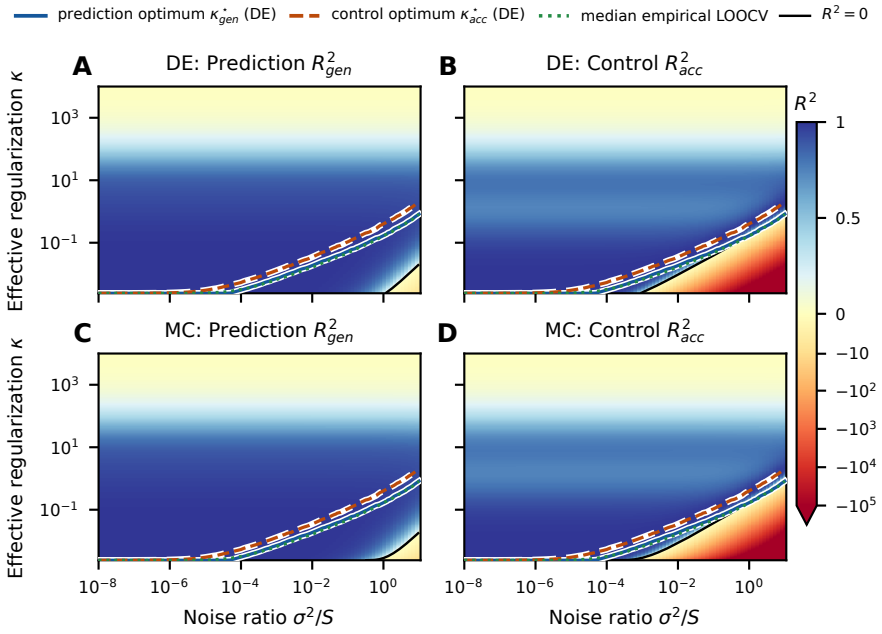

Figure 8: R^2 **landscapes**, layout as in Fig. 7: prediction $R^2_{gen} = 1 - E_{gen}/S$ (left) and control $R^2_{acc} = 1 - z^2$ (right). Color is linear on $[0, 1]$ and logarithmic in $|R^2|$ below 0; black contour: $R^2 = 0$. Prediction stays above zero except at the highest noise with the weakest penalty, while control falls below zero in a wedge under the prediction-optimal path, reaching $R^2_{acc} < -10^4$ at weak regularization; the prediction-optimal path meets the wedge at $\sigma^2/S \approx 1$ (Fig. 3).

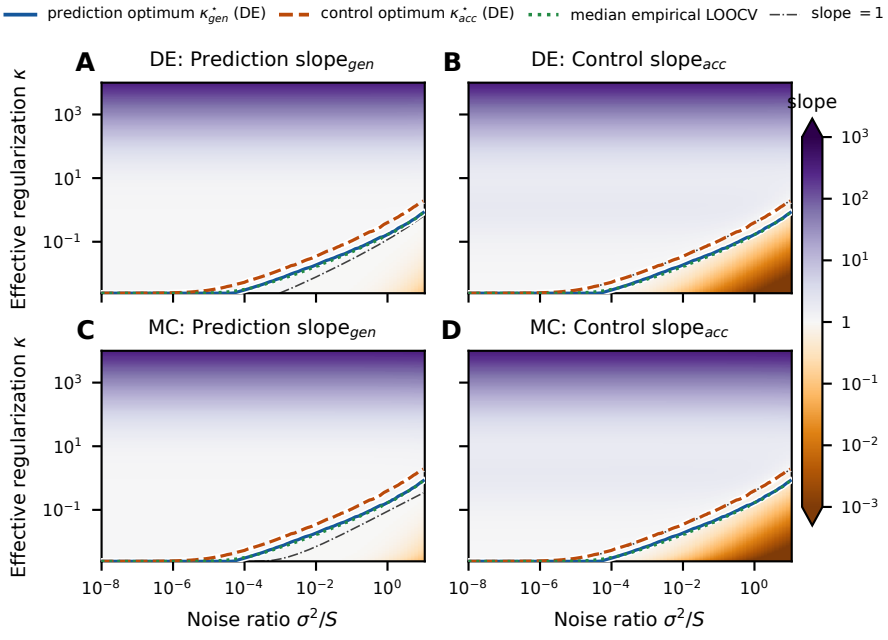

Figure 9: **Slope landscapes**, layout as in Fig. 7: true-on-fitted slope for generalization (left) and $\text{slope}_{acc} = N/D$ for accentuation (right); color is logarithmic and centered on the calibrated value 1 (dash-dotted contour). Strong shrinkage makes both slopes exceed one. Below the prediction-optimal path the generalization slope falls by at most about one order of magnitude, whereas the accentuation slope falls by up to four: the accentuated images move the teacher far less than the student predicts.

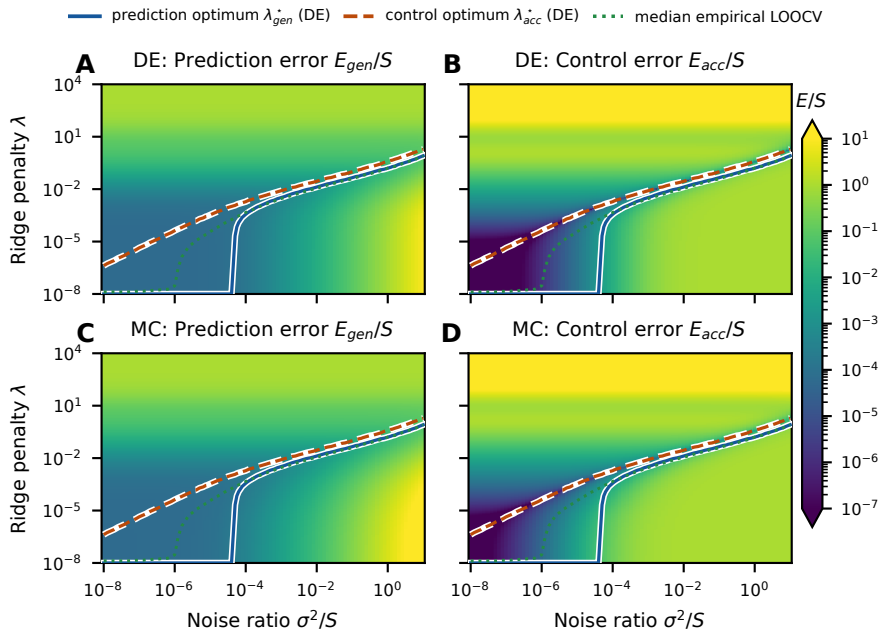

Figure 10: **Error landscapes against the ridge penalty** λ ($\alpha = n\lambda$) instead of the effective penalty κ; data, panels and lines as in Fig. 7. Because $d > n$, all $\lambda \lesssim 10^{-4}$ map onto the near-ridgeless floor of κ, so this view resolves the low-noise corner that the κ view compresses. There the prediction optimum is ridgeless ($\lambda^{\star}_{gen}$ at the grid floor for $\sigma^2/S \lesssim 5 \times 10^{-5}$), whereas the control optimum $\lambda^{\star}_{acc}$ shrinks in proportion to the noise along a deep valley of E_{acc}: at $\sigma^2/S = 10^{-6}$ the two optima differ 4000-fold in λ but only by 3% in κ. For $\sigma^2/S \geq 10^{-3}$, $\lambda^{\star}_{acc} \approx 2$–$2.5\,\lambda^{\star}_{gen}$.

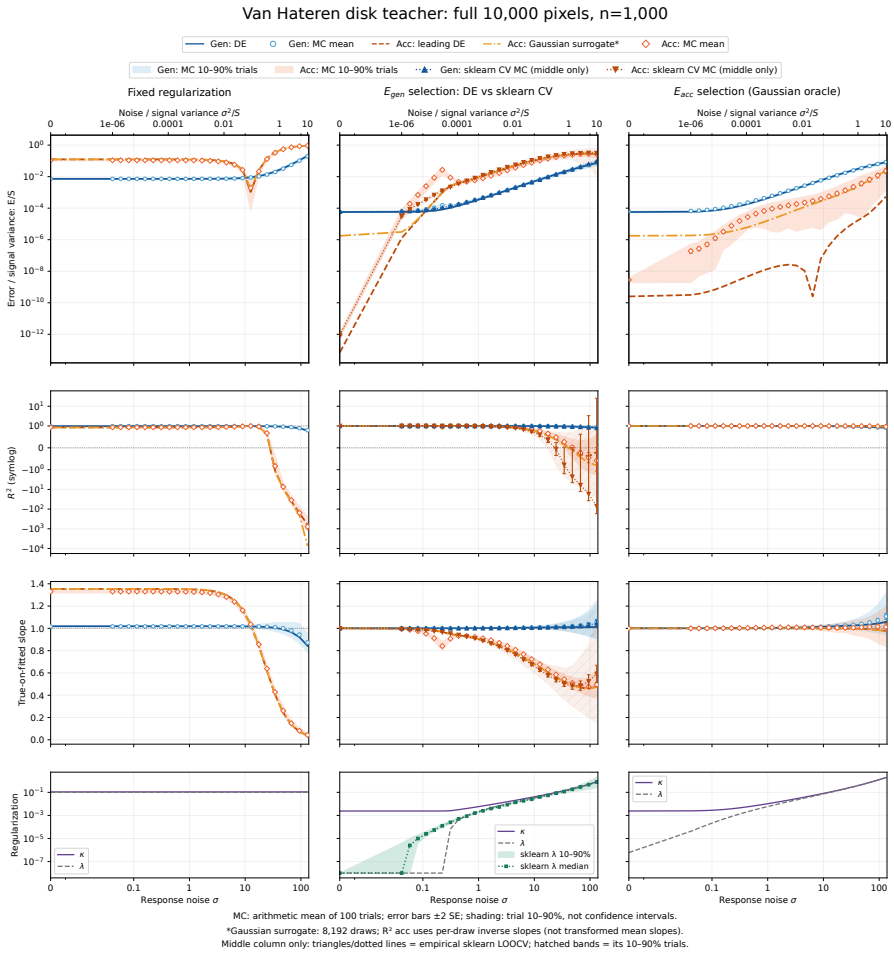

Figure 11: **Fixed, prediction-selected and control-selected penalties along the noise axis** (van Hateren disk teacher, $d = 10^4$, $n = 10^3$). Columns: fixed regularization; penalty minimizing the DE of E_{gen} (with empirical LOOCV in triangles); penalty minimizing E_{acc} (Gaussian oracle). Rows: normalized error, R^2 (symlog), true-on-fitted slope, and the selected κ and λ. Blue: generalization; orange: accentuation, leading DE (dashed), Gaussian surrogate (App. J.2; dash-dot) and Monte Carlo (markers: mean of 100 trials; bands: 10–90% of trials). At the prediction-selected penalty the accentuation R^2 turns negative at high noise; at the control-selected penalty the leading DE underestimates the accentuation error by orders of magnitude, and the Gaussian surrogate is needed.

A.3 LINEAR FEATURE RIDGE: THE PREDICTION-NULL FAMILY

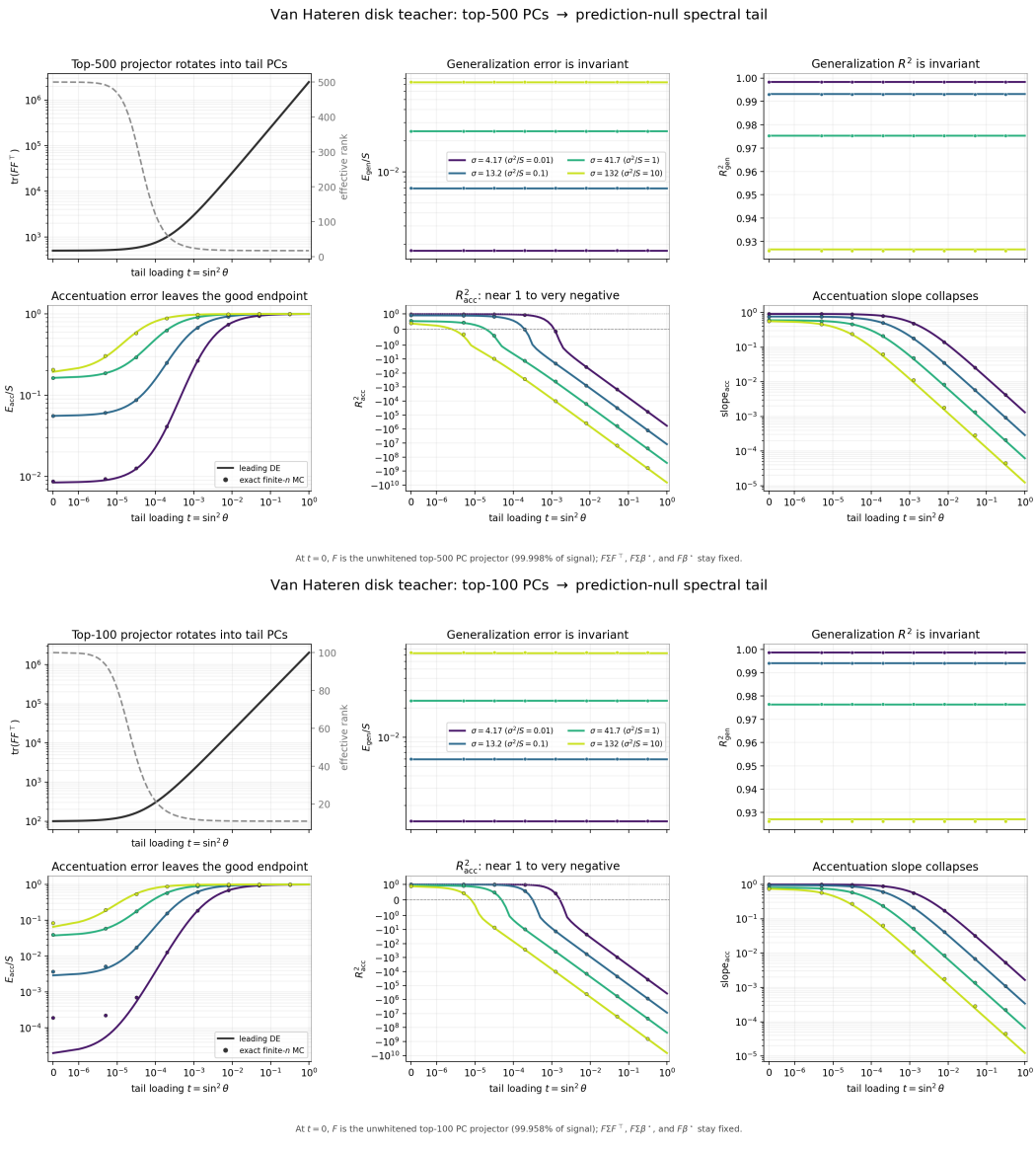

Figure 12: **Continuous prediction-null rotation on van Hateren images** (Corollary 19), starting from the top-500 (top) and top-100 (bottom) PC projectors and rotating their null part into the spectral tail. Generalization error, R^2_{gen} and the selected penalty are exactly constant in the tail loading t, while $\mathrm{Tr}(FF^\top)$, accentuation error, R^2_{acc} and slope change by orders of magnitude. Lines: leading DE; markers: finite-n Monte Carlo.

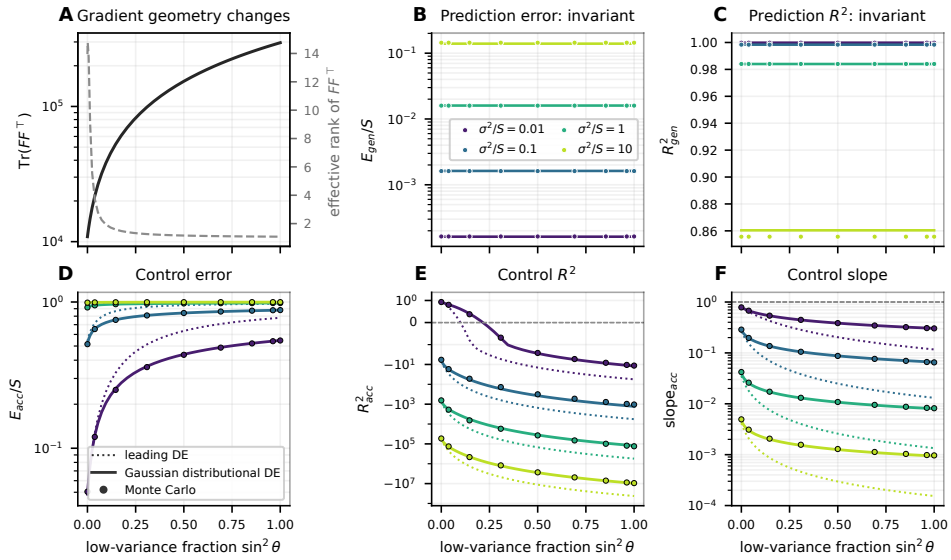

Figure 13: **Rotating a single prediction-null row** ($p = 16$)**.** One feature row turns from a high-variance direction (PC rank 21) to a low-variance one (PC rank 9500; input-variance ratio 3×10^4), all other rows fixed, so that Σ_F, $F\Sigma\beta^\star$ and $F\beta^\star$ stay constant. Colors: $\sigma^2/S = 0.01, 0.1, 1, 10$. **A** $\mathrm{Tr}(FF^\top)$ grows 27-fold while the effective rank of $FF^\top$ falls from 14 to 1. **B, C** E_{gen}/S and R^2_{gen} are exactly invariant. **D–F** E_{acc}/S, R^2_{acc} and slope$_{acc}$ degrade monotonically; at $\sigma^2/S = 0.01$, R^2_{acc} falls from 0.93 to -11. Because $FF^\top$ is dominated by one row, the control ratio does not concentrate: the leading DE (dotted) overstates the loss, whereas the Gaussian distributional DE (solid; App. J.2) tracks Monte Carlo (markers: 500 trials; mean, median for R^2_{acc}). Penalty fixed at the DE-selected prediction optimum.

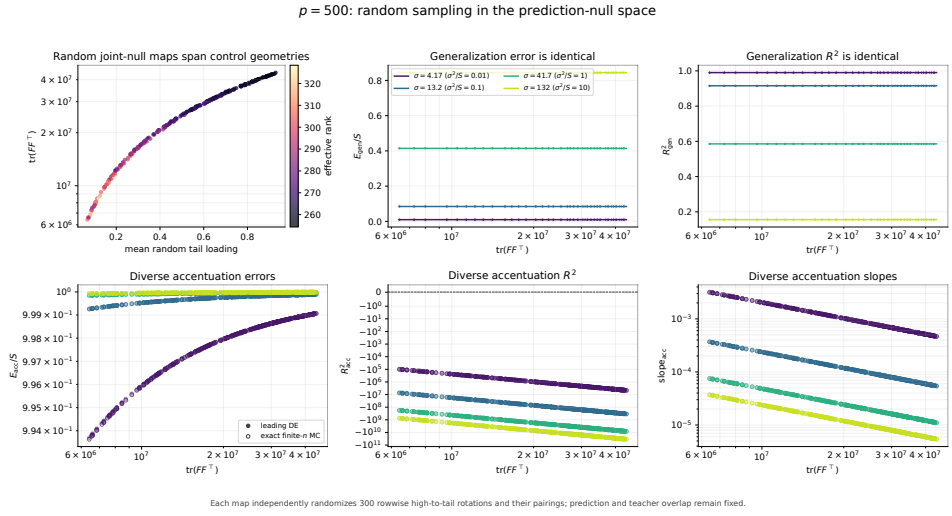

Figure 14: **Random maps from the prediction-null family** ($p = 500$)**.** Each map independently randomizes 300 row-wise high-to-tail rotations and their pairings; Σ_F, $F\Sigma\beta^\star$ and the teacher overlap stay fixed. Top: $\mathrm{Tr}(FF^\top)$ against the mean tail loading (color: effective rank), and E_{gen}/S and R^2_{gen} at four noise levels, identical for all maps. Bottom: E_{acc}/S, R^2_{acc} and slope$_{acc}$ spread over orders of magnitude and degrade with $\mathrm{Tr}(FF^\top)$. Filled: leading DE; open: exact finite-n Monte Carlo.

A.4 PEER REVIEW IN THE NEURONAL EXPERIMENT

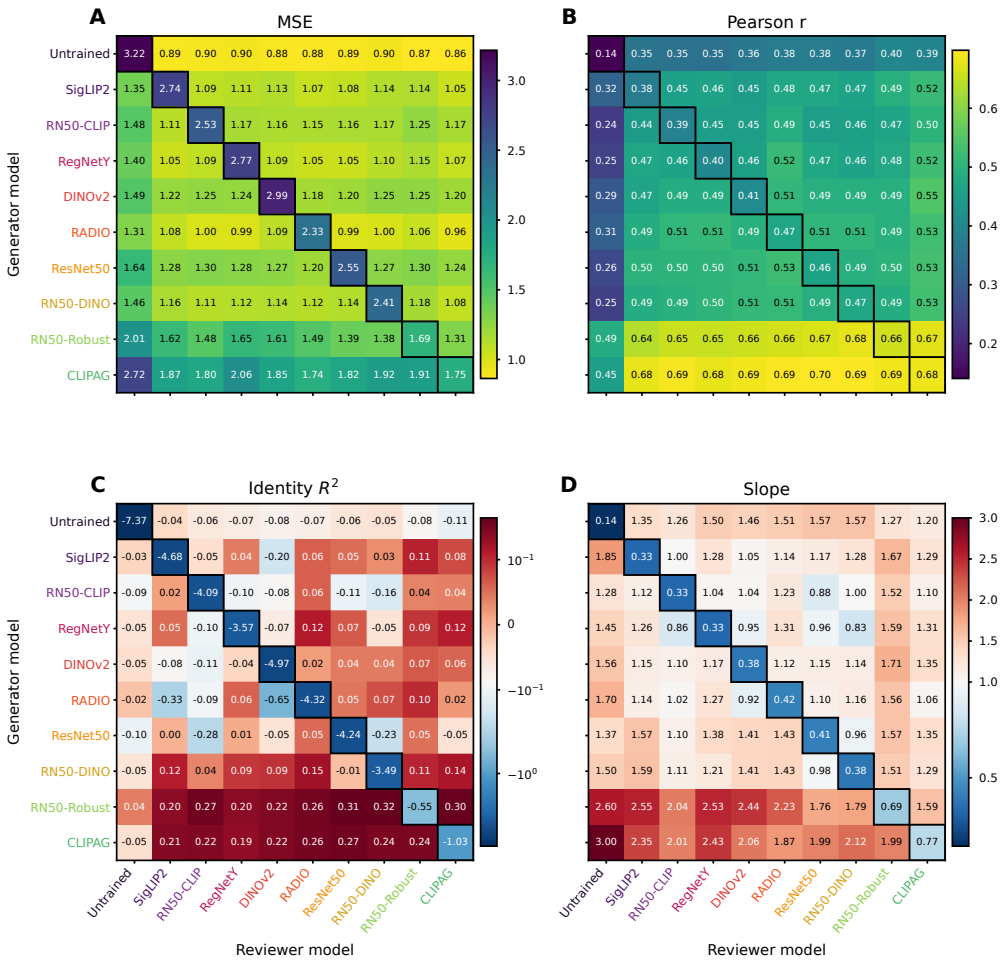

Figure 15: **Neuronal peer review.** Rows: model that generated the accentuated images; columns: model whose predictions are compared with the measured responses; boxed diagonal: self-evaluation. Each cell is the equal-weight mean over 25 sites of **A** MSE, **B** Pearson r, **C** identity R^2 (symlog color scale) and **D** the OLS slope of measured on predicted response (color centered at one), with predictions mapped to the control session by the frozen site affine map. Self-evaluation has slope below one and strongly negative R^2 for every model, whereas peer slopes are mostly above one (Sec. 4). Methods: App. B.9.

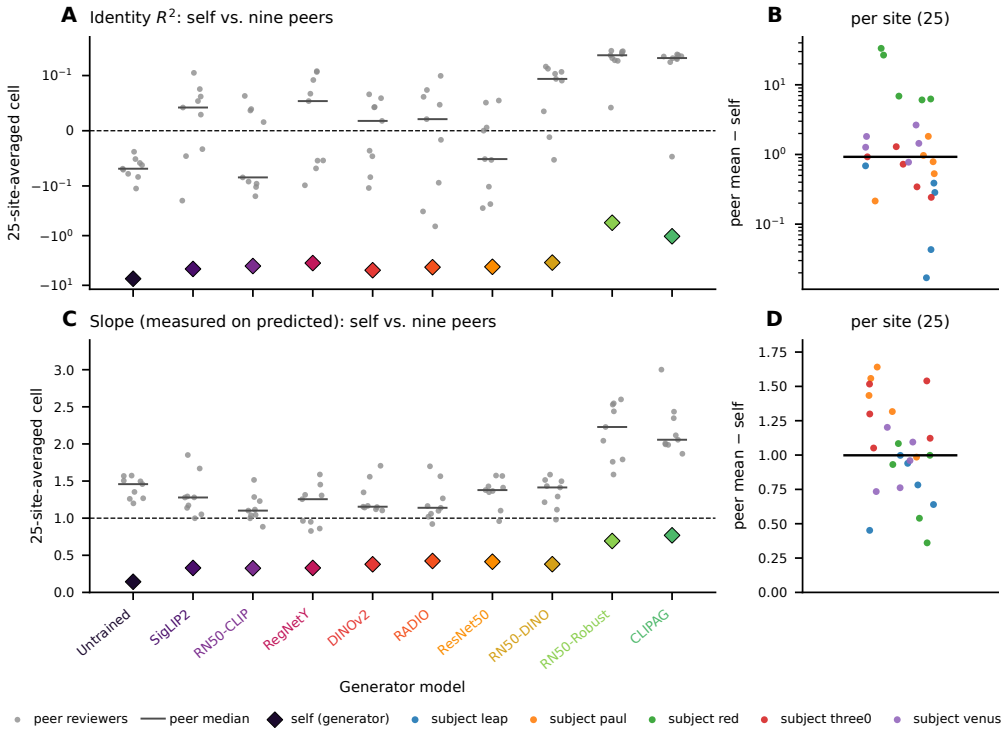

Figure 16: **Self-evaluation versus peer review.** Top row: identity R^2; bottom row: slope. **A, C** For each generator model, the self cell (diamond) and the nine peer reviewers (gray dots; bar: median), from the 25-site-averaged matrices of Fig. 15C,D (R^2 on a symlog axis). Every self slope is below one and every self R^2 is negative; 80 of 90 peer slopes exceed one. Over matrix cells, peer versus self: R^2 0.04 $[-0.06, 0.10]$ versus -4.17 $[-4.59, -3.51]$, slope 1.35 $[1.14, 1.59]$ versus 0.38 $[0.33, 0.42]$ (median [IQR]). **B, D** Paired site contrasts for the same metric (mean over peer cells minus mean over self cells; log axis in B, linear from zero in D; bar: median), colored by subject: all 25 sites favor peer review on both metrics. Reviewer-label permutation $p < 5 \times 10^{-6}$; with five subjects the exact sign-flip test attains its minimum $p = 0.0625$. Methods: App. B.9.

A.5 NONLINEAR GRADIENT DIAGNOSTIC: FULL COMPARISON

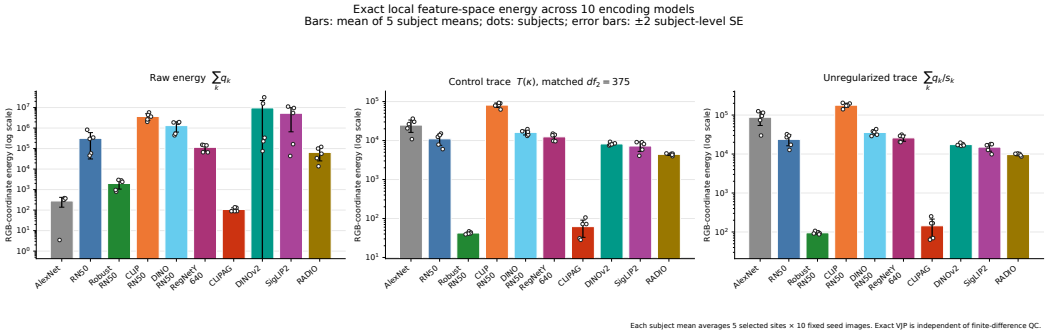

Figure 17: **Exact local gradient energy of the ten encoding models**: raw energy $\sum_k q_k$, control trace $T(\kappa)$ at matched $\mathrm{df}_2 = 375$, and unregularized trace $\sum_k q_k/s_k$, with $q_k = \|J_\phi^\top v_k\|^2$ in RGB $[0, 1]$ coordinates. Bars: mean of five subject means (each over 5 sites $\times$ 10 seed images); dots: subjects; error bars: ± 2 subject-level SE. The two adversarially trained models (robust RN50, CLIPAG) have a control trace two to three orders of magnitude below the others.

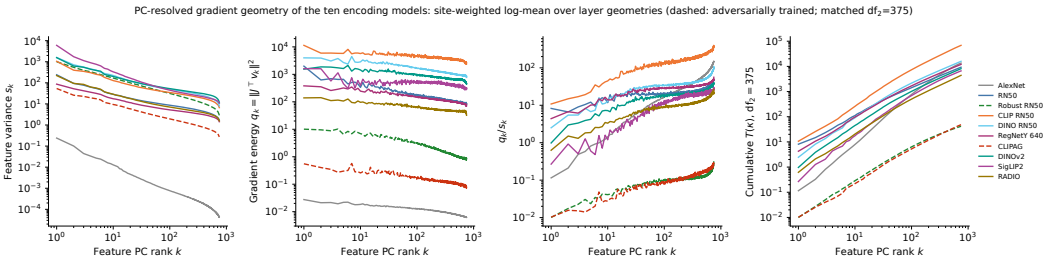

Figure 18: **PC-resolved gradient geometry of the ten encoding models.** From left: training feature variance s_k, exact local gradient energy $q_k = \|J_\phi^\top v_k\|^2$ (mean over the ten seed images), their ratio q_k/s_k, and the cumulative control trace $\sum_{j \le k} s_j q_j/(s_j + \kappa)^2$ at the penalty giving $\mathrm{df}_2 = 375$, against feature PC rank. Each curve is the site-weighted log-mean over the layer geometries a model uses across the 25 sites (84 geometries in total). The adversarially trained models (dashed) have q_k/s_k two to three orders of magnitude below the others at every rank, so feature-space estimation noise back-propagates to little pixel-gradient energy; for the other models the ratio grows toward the tail and the trace accumulates over hundreds of PCs.

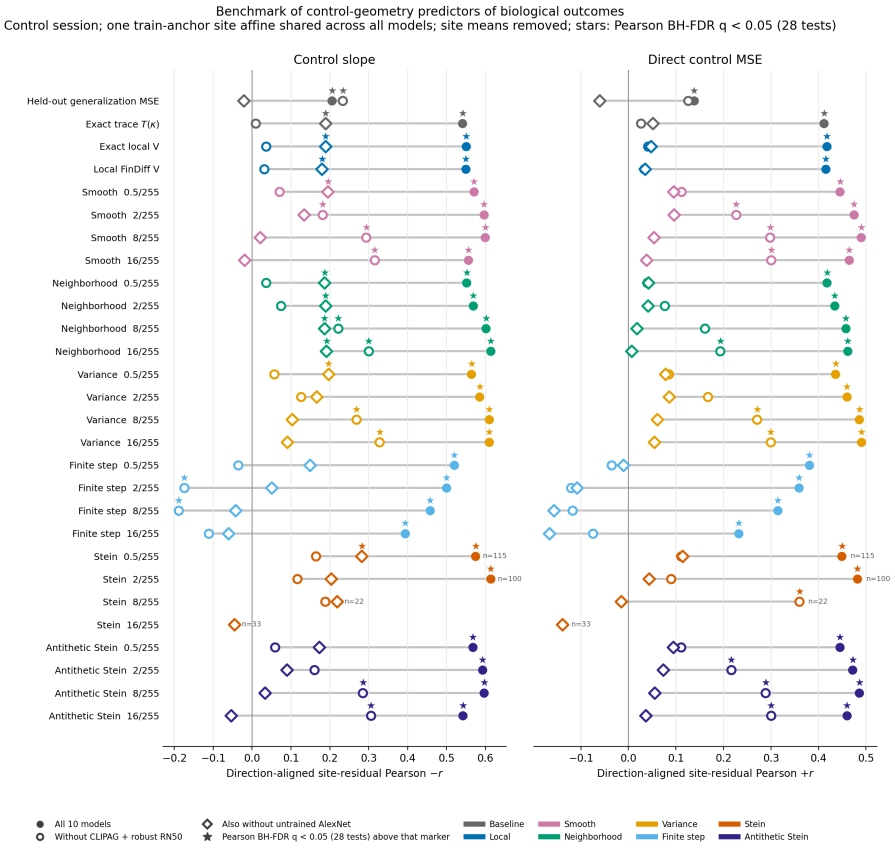

Figure 19: **All candidate predictors of biological control.** For each $\log_{10}$ predictor, the site-residual Pearson correlation with control slope (left, sign flipped) and with direct control MSE (right), so that positive values are in the predicted direction. Estimators (App. K): held-out MSE, exact trace, exact and finite-difference local V_ϕ, and smoothed-Jacobian, neighborhood, response-variance, finite-step, one-sided Stein and antithetic Stein versions at neighborhood scales 0.5–$16/255$. Filled: all 10 models ($n = 250$); open circles: without CLIPAG and robust RN50; diamonds: also without the untrained AlexNet. Stars: Benjamini–Hochberg $q < 0.05$ over the 28 tests; n is annotated where fewer rows give a valid estimate. Statistics: App. B.10.

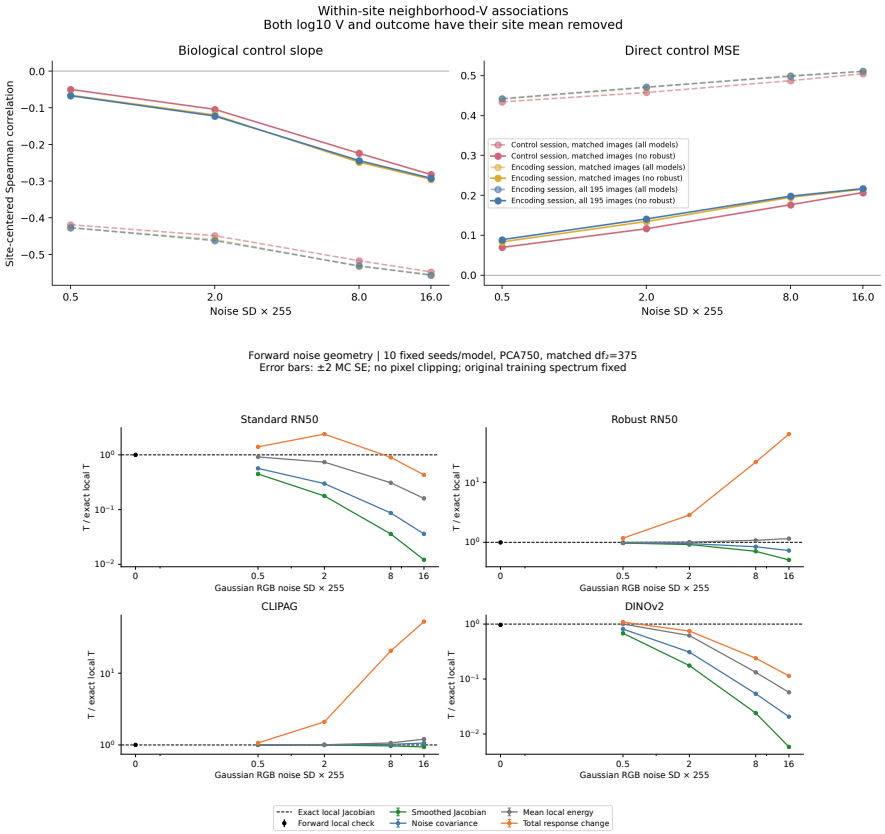

Figure 20: **Dependence on the neighborhood scale τ.** Top: site-centered Spearman correlation of neighborhood V_ϕ with control slope and with direct control MSE against τ, for three endpoint definitions (control session with matched images; encoding session with matched or all 195 held-out images), with all models (dashed) and without the robust models (solid). Bottom: neighborhood estimators of the trace relative to the exact local trace for four models at matched $\mathrm{df}_2 = 375$ (10 seeds; error bars ± 2 MC SE). For standard networks smoothing removes most of the local gradient energy, whereas for the robust networks it barely changes it.

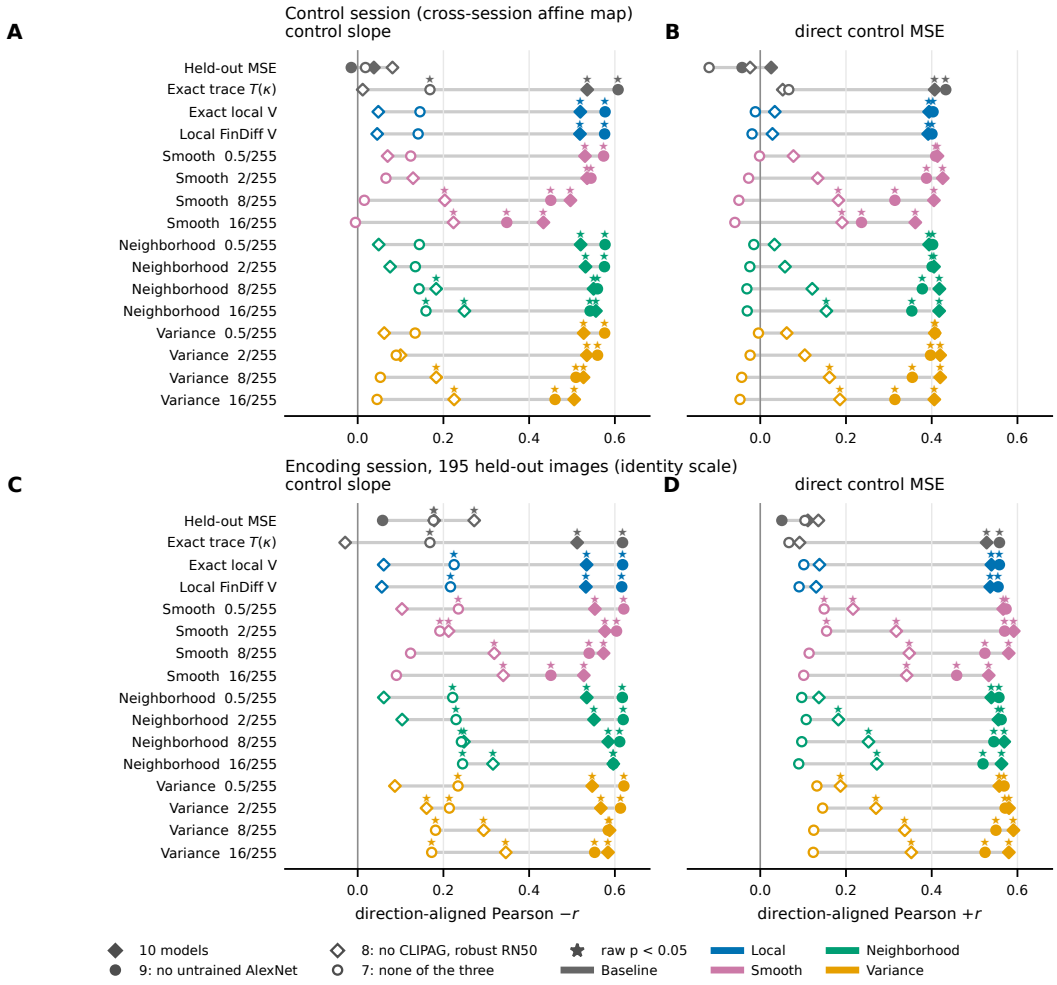

Figure 21: **Robustness of the model ranking to the endpoint.** Same convention as Fig. 5D: raw $\log_{10}$ predictor against the outcome minus its fixed all-10-model site mean; direction-aligned Pearson correlation (sign flipped for control slope, so positive is the predicted direction). **A, B** Control session with the cross-session affine map (Eq. 23); A extends Fig. 5D to more estimators, B uses the direct control MSE. **C, D** Encoding-session endpoint: held-out MSE and V_ϕ evaluated on the 195 held-out encoding-session images, outcomes on the upstream (identity) scale without the affine map. Markers: all 10 models (filled diamond), without the untrained AlexNet (filled circle), without CLIPAG and robust RN50 (open diamond), without all three (open circle); stars: raw $p < 0.05$. For 9–10 models, the exact trace, exact local V_ϕ and neighborhood V_ϕ $(16/255)$ reach $r = 0.35$–0.62 at both endpoints and for both outcomes, while held-out MSE stays at or below 0.18. Spearman and predictor-residualized versions: Figs. 22 and 23.

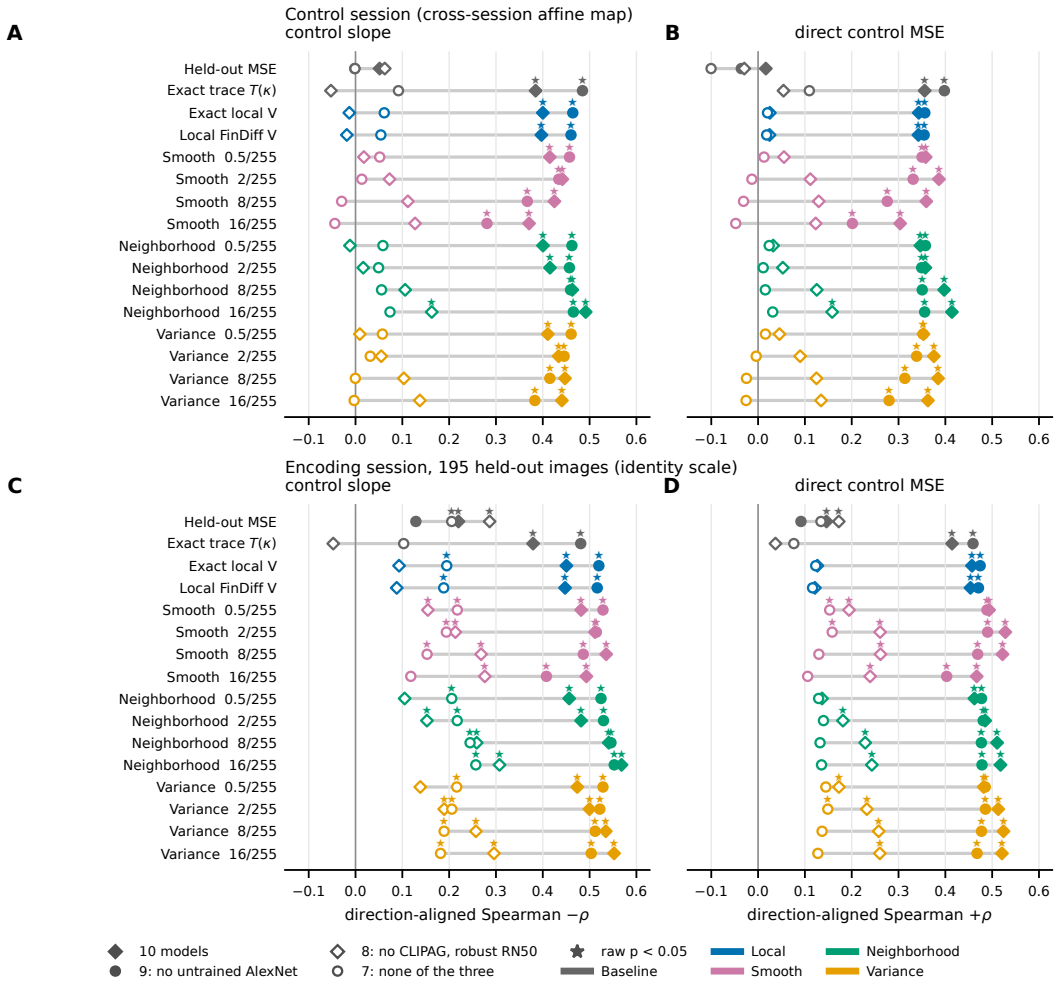

Figure 22: **Rank-based version** of Fig. 21: identical, but with Spearman ρ instead of Pearson r (stars: raw Spearman $p < 0.05$). The geometry–control association is not driven by a few extreme values: for 9–10 models the same three predictors reach $\rho = 0.34$–0.57, while held-out MSE stays at or below 0.22.

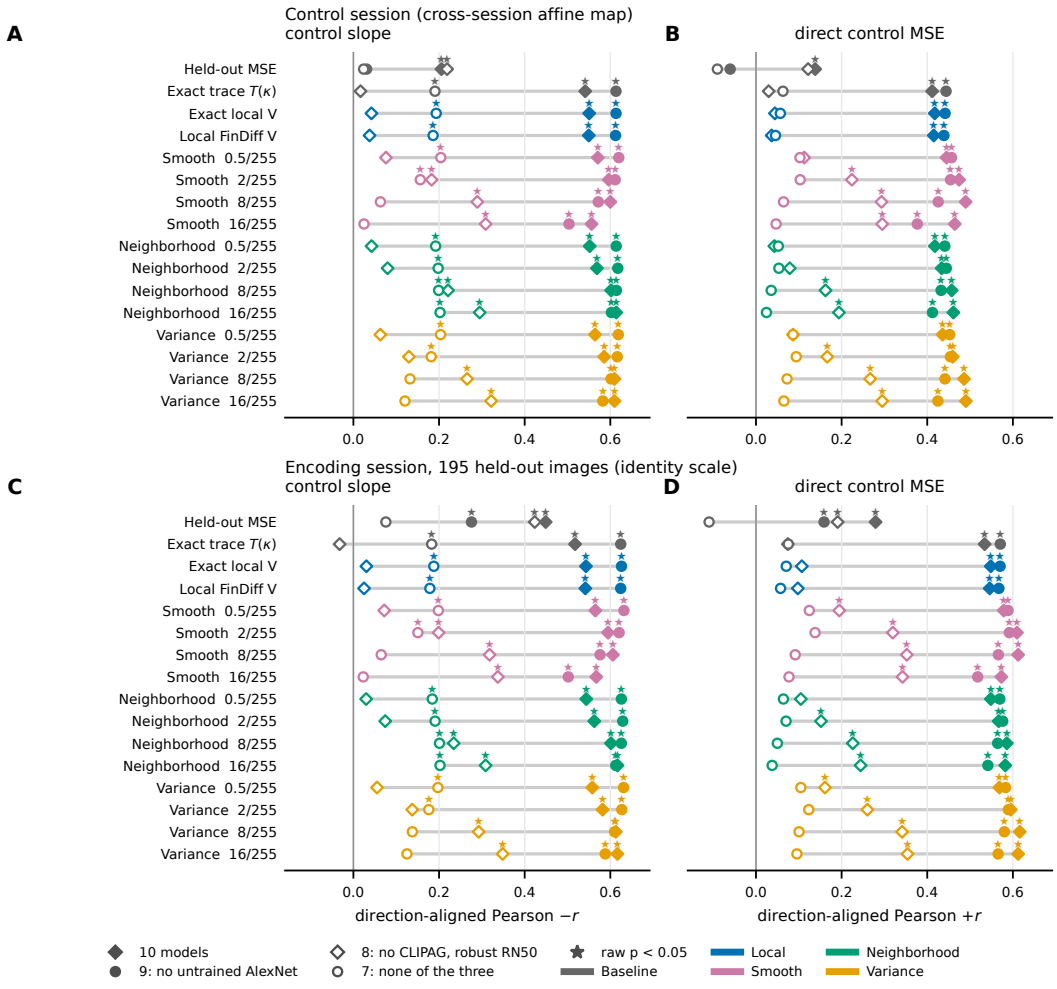

Figure 23: **Predictor-residualized version** of Fig. 21: identical, but the $\log_{10}$ predictor is also centered on its fixed all-10-model site mean, so only within-site differences between models enter on both axes. The geometry predictors change little ($r = 0.41$–0.63 for 9–10 models). Held-out MSE gains a positive association here, up to $r = 0.21$ in the control session and 0.45 at the encoding endpoint, largest with the untrained AlexNet included; it vanishes for the seven conventional models ($r \leq 0.08$).

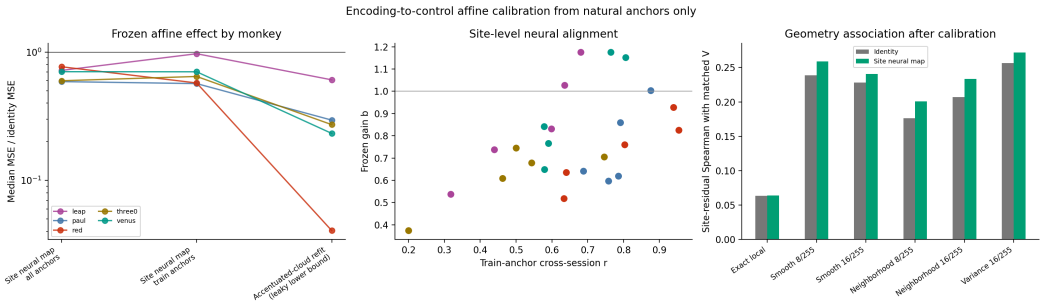

Figure 24: **Encoding-to-control affine calibration from natural anchors only.** Left: median MSE relative to the identity map, by monkey, for a site map fitted on all anchors, on training anchors only (used throughout), and a leaky refit on the accentuated responses (lower bound). Middle: frozen gain b of each site against its train-anchor cross-session correlation. Right: site-residual Spearman correlation of matched V_ϕ with the control outcome under the identity and the site map. Methods: App. B.6.

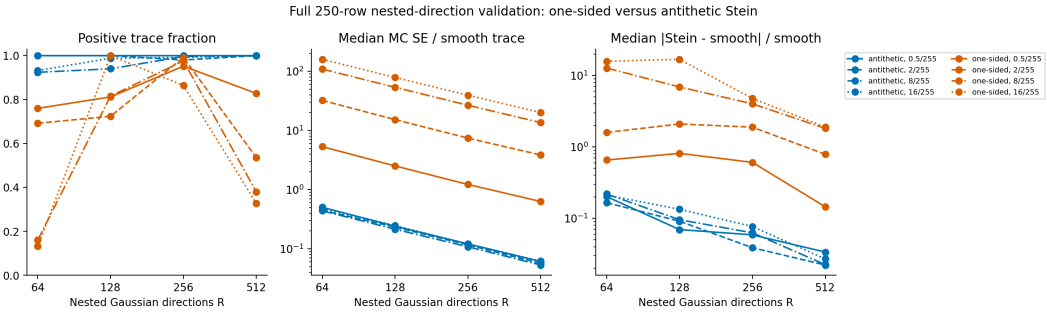

Figure 25: **Stein estimators of the smoothed trace**, nested over $R = 64$–512 Gaussian directions on all 250 site–model pairs: fraction of positive trace estimates, median Monte Carlo SE relative to the smoothed trace, and median relative deviation from the smoothed-Jacobian trace, for the one-sided (orange) and antithetic (blue) estimators (Eq. 107) at four neighborhood scales. Methods: App. B.8.

B Detailed methods

B.1 Natural-image data and covariance

All pixel-space and linear-feature experiments on natural images (Figs. 3 and 4; Apps. A.2, A.3) share one image set, covariance and teacher. The toy example of Fig. 1 uses a separate, simpler pipeline (App. B.2).

Images and patches. We used the uncalibrated van Hateren natural-image database (`.iml` files, 1024×1536 pixels, 16-bit big-endian luminance L). Each image was mapped to $\mathrm{round}\big(255 \log(1+L)/\log 65536\big)/255 \in [0,1]$, i.e. log luminance quantized to eight bits. From 4,000 randomly chosen source images we drew 100×100 patches at uniformly random positions: a *training pool* of 2,000 patches (one from each of 2,000 images) and a *population pool* of 20,000 patches (ten from each of the other 2,000 images). The two pools share no source image, so fitted models are always evaluated on images from different photographs. One crop per training image keeps training samples independent: with 25 crops per image, correlated sibling crops leaked through leave-one-out cross-validation, which then chose the interpolating boundary of the penalty grid in 26 of 100 trials already at $\sigma^2/S = 0.03$. Patches are vectorized in $\mathbb{R}^d$, $d = 10^4$.

Population covariance. The population pool was centered once, and its empirical covariance $\Sigma = \frac{1}{20{,}000} P^\top P$ (P: centered patch matrix) was diagonalized on a GPU in single precision, $\Sigma = \sum_k s_k \mathbf{u}_k \mathbf{u}_k^\top$. With 20,000 samples in $d = 10^4$ dimensions all eigenvalues are positive; they span 7.2 decades ($s_1 = 30.7$, $s_d = 2.0 \times 10^{-6}$, $\mathrm{Tr}\,\Sigma = 46.9$). The leading PC, essentially mean luminance, carries 65% of the total variance, the top 100 PCs 89%. This Σ serves both as the covariance entering every deterministic equivalent and as the evaluation distribution for E_{gen}.

Disk teacher. The teacher $\beta^\star$ is the indicator of a centered disk of radius 0.3 on the $[-1,1]^2$ patch grid (radius $\approx$ 15 pixels; 692 pixels, so $\|\beta^\star\|^2 = 692$), and $b_k = \mathbf{u}_k^\top \beta^\star$. Its natural response variance is $S = \sum_k s_k b_k^2 = 1736.34$. The teacher is strongly structured relative to the spectrum: the first PC carries 90.7% of S but only 7.4% of $\|\beta^\star\|^2$, and PCs beyond the 100th carry 13.8% of $\|\beta^\star\|^2$ while contributing 0.04% of S. It is this Euclidean teacher energy in low-variance directions that prediction barely weights and accentuation exposes.

B.2 Toy teacher–student example

The linear example is a construction rather than a fit: it isolates the geometry of Sec. 3 with two students that differ from the teacher only along low-variance directions.

Images and covariance. We drew 12,000 random 100×100 patches from 500 van Hateren images (image and patch positions fixed by seed 42). Each source image was min–max normalized to $[0,1]$ and a constant 0.15 was subtracted; unlike App. B.1, no log transform was applied. The eigenvectors $\mathbf{u}_k$ (ordered by decreasing variance) were obtained from the SVD of the centered patch matrix.

Teacher and students. The teacher $\beta^\star$ is the same radius-0.3 disk as in App. B.1. For an index slice K of eigenvectors let $u_K \propto \sum_{k \in K} (\mathbf{u}_k^\top \beta^\star)\mathbf{u}_k$ be the unit-norm projection of the teacher onto their span. The two students are

$$w_1 = \beta^\star + \alpha_1 u_{K_1}, \qquad w_2 = \beta^\star + \alpha_2 u_{K_2},$$

with K_1 the 400 eigenvectors ranked $d - 499$ to $d - 100$ (moderately low variance; $d = 10^4$), $\alpha_1 = 1$, and K_2 the 100 lowest-variance eigenvectors, $\alpha_2 = 100$. Since $\|\beta^\star\| \approx 26$, w_1 is visually indistinguishable from the teacher while w_2 carries a large high-spatial-frequency component (Fig. 1C.II). Both perturbations are nearly invisible to prediction: on the natural patches $1 - \mathrm{Var}(f_i - f^\star)/\mathrm{Var}(f^\star)$ equals 1.00000 to five decimals for both students (Fig. 1C.I), although $\|w_2\| \approx 104$ is four times $\|\beta^\star\|$.

Accentuation. For a linear model $f(\mathbf{x}) = w^\top \mathbf{x}$, gradient ascent to a target t from a seed $\mathbf{x}_0$ has the closed-form endpoint $\mathbf{x} = \mathbf{x}_0 + \lambda w$ with $\lambda = (t - w^\top \mathbf{x}_0)/\|w\|^2$, so each student reaches every

target exactly on its own path. Targets were 13 evenly spaced levels between $\mu \pm 3\sigma$ of the teacher response on natural patches (-309.6 to 289.8), and seeds were 20 natural patches (seed 7); unlike the analysis of Sec. 3, seeds are natural images rather than the null seed. Figure 1C.II shows every third target for one seed ($f^\star(\mathbf{x}_0) = -31.4$).

Peer review. Each student's accentuated images were evaluated by the teacher and by both students, giving the 2×2 generator–evaluator panels of Fig. 1C.III (all 20 seeds $\times$ 13 targets; colors mark seeds). Throughout Fig. 1C we report $R^2 = 1 - \mathrm{Var}(f^\star - f_e)/\mathrm{Var}(f^\star)$, where $f^\star$ is the teacher and f_e the evaluating model on the same images, i.e. the paper's convention of normalizing by the variance of the true response (Table 1), together with the squared Pearson correlation r^2. Each student predicts the teacher's response to the other's accentuated images almost perfectly ($R^2 = 1.0000$ and 0.9966), and w_1 also controls through its own images ($R^2 = 1.0000$), whereas w_2's own images barely move the teacher ($R^2 = -9.94$, $r^2 = 0.05$). These values come from a deterministic reproduction of the original run.

B.3 Two-dimensional geometry illustration

Setting. Figures 2 and 6 use $d = 2$ with $\Sigma = \mathrm{diag}(1, \varepsilon)$, $\varepsilon = 0.04$, so the second coordinate is a low-variance direction, and teacher $\beta^\star = (1.1, 0.65)^\top$, giving $S = \beta^{\star\top}\Sigma\beta^\star = 1.2269$. One Gaussian design $X \in \mathbb{R}^{40\times 2}$ with rows $\mathcal{N}(0, \Sigma)$ is drawn once (seed 20260918) and held fixed. With $n = 40 \gg d$ the effective penalty is close to the nominal one ($\kappa = 0.0518$ at $\lambda = 0.05$); the figures therefore label paths by λ in the convention of Eq. 2.

Exact conditional law. For fixed X and $\mathbf{y} = X\beta^\star + \sigma\eta$, the ridge estimator is exactly Gaussian,

$$\hat\beta \mid X \sim \mathcal{N}\big(A_\lambda X\beta^\star,\ \sigma^2 A_\lambda A_\lambda^\top\big), \qquad A_\lambda = (X^\top X + n\lambda I)^{-1}X^\top.$$

Ellipses are its 68% probability regions (Mahalanobis radius $\sqrt{-2\log 0.32}$); the dashed path is the conditional mean $A_\lambda X\beta^\star$ over λ, i.e. the noise average for this design, not the deterministic equivalent $\bar\beta_\kappa$ of Eq. 3. The purple path is one noise realization η, kept fixed along both sweeps so that the path is continuous; it was chosen among four draws for visible motion along the low-variance axis. 250 further draws verify the ellipses. Panel (a) fixes $\lambda = 0.05$ and varies σ^2 on 81 log-spaced values in $[10^{-4}, 4]$ (clouds at $\sigma^2 = 0.01, 0.1, 1, 4$); panel (b) fixes $\sigma^2 = 0.5$ and varies λ on 101 log-spaced values in $[10^{-4}, 10]$ (clouds at $\lambda = 10^{-4}, 10^{-2}, 1, 10$ and at the value $\lambda_\times = \sqrt{\det X^\top X}/n$ where the covariance ellipse swaps its major axis).

Equal-error sets and marked penalties. Prediction contours are the ellipses $\Delta^\top\Sigma\Delta = E$ for $E = 0.01, 0.04, 0.09$ ($R^2_{gen} = 0.99, 0.97, 0.93$); control contours are the circles $\hat\beta^\top\hat\beta - (1+z)\hat\beta^\top\beta^\star = 0$ through the origin and $(1+z)\beta^\star$ for $z = -1/3, 0, 1$ ($R^2_{acc} = 1 - z^2 = 0.89, 1, 0$) (Prop. 7). For the plotted realization, the prediction-stationary penalty $\lambda_{\mathrm{pred}} = 0.0177$ minimizes $E_{gen}(\hat\beta(\lambda))$ (golden-section search in $\log\lambda$), and the control-calibrated penalty $\lambda_{\mathrm{ctrl}} = 0.0258$ is the root of $\hat\beta^\top(\hat\beta - \beta^\star) = 0$ ($z = 0$) closest to λ_{pred} (bisection in $\log\lambda$). Both are oracle quantities for this draw, illustrating what cross-validation targets on average rather than a cross-validated selection.

B.4 Natural-image pixel-ridge landscape experiment

This subsection describes the experiment underlying Fig. 3 and Figs. 7–11. Its Monte Carlo component uses actual natural-image patches, rather than Gaussian designs having the same covariance.

Images, teacher, and population reference. Images, pools, covariance and disk teacher are those of App. B.1: a 2,000-patch training pool, a disjoint 20,000-patch population pool that defines Σ and is used to evaluate every fitted model, and a disk teacher with $S = \beta^{\star\top}\Sigma\beta^\star = 1736.34$.

Noise–regularization grid and deterministic equivalents. We evaluated 114 response-noise levels, comprising zero and a logarithmic grid of $\sigma^2/S \in [10^{-8}, 10]$, and 478 ridge penalties. We use the scikit-learn convention

$$\widehat\beta_{\alpha,\sigma} = X^\top(XX^\top + \alpha I_n)^{-1}\boldsymbol{y}_\sigma, \qquad \alpha = n\lambda, \tag{14}$$

with $\alpha \in [10^{-5}, 10^{7}]$ ($\lambda \in [10^{-8}, 10^{4}]$). The effective penalty κ shown in Fig. 3 is the positive solution of $n\lambda = \kappa\{n - \mathrm{df}_1(\kappa)\}$; Fig. 10 shows the error landscapes against λ itself. At each of the 54,492 grid points we evaluated the deterministic equivalents using the eigenvalues of the empirical population covariance and the disk teacher's coordinates in that eigenbasis. The prediction-optimal and control-optimal paths minimize, respectively, the DE prediction error and the leading-DE accentuation error over the finite penalty grid at each noise level. All heatmaps and slices are direct evaluations on this grid; no smoothing or interpolation was used.

Natural-image Monte Carlo. We ran 100 paired trials. On trial t, using seed $20260918 + t$, we sampled $n = 1,000$ patches without replacement from the 2,000-patch training pool, centered every pixel using that sample, and drew a centered $\epsilon \sim \mathcal{N}(0, I_n)$. For every noise level we used the response

$$\boldsymbol{y}_\sigma = X\beta^\star + \sigma\epsilon. \tag{15}$$

Thus the design and unit-noise direction were shared across the entire noise axis within a trial. This pairing reduces differences caused by resampling and makes each trial a coherent path as σ varies.

Because $n < d$, we diagonalized the $n \times n$ kernel matrix $XX^\top$ once per trial. For every penalty the fitted weight is exactly affine in the noise scale,

$$\widehat{\beta}_{\alpha,\sigma} = X^\top(XX^\top + \alpha I_n)^{-1}X\beta^\star + \sigma X^\top(XX^\top + \alpha I_n)^{-1}\epsilon = w_{0,\alpha} + \sigma w_{1,\alpha}. \tag{16}$$

We therefore cached the linear or quadratic coefficients in σ of all required quadratic forms and reconstructed the complete noise axis exactly. This avoids refitting 54,492 models per trial and is not an interpolation.

Let $P \in \mathbb{R}^{20,000\times d}$ be the centered population-patch matrix, $N = \beta^{\star\top}\widehat{\beta}$, $D = \|\widehat{\beta}\|^2$, $Q = \|P\widehat{\beta}\|^2/20{,}000$, and $C = (P\beta^\star)^\top P\widehat{\beta}/20{,}000$. We report

$$\frac{E_{gen}}{S} = \frac{\|P(\widehat{\beta} - \beta^\star)\|^2}{20{,}000S}, \qquad R^2_{gen} = 1 - \frac{E_{gen}}{S}, \qquad \mathrm{slope}_{gen} = \frac{C}{Q}, \tag{17}$$

$$\frac{E_{acc}}{S} = \left(1 - \frac{N}{D}\right)^2, \qquad R^2_{acc} = 1 - \left(\frac{D}{N} - 1\right)^2, \qquad \mathrm{slope}_{acc} = \frac{N}{D}. \tag{18}$$

The fitted intercept is accounted for by sample centering and by the cross-validation leverage below, but it is omitted from these population weight-response metrics.

Empirical RidgeCV. For each trial and noise level we selected α by exact leave-one-out cross-validation over the same 478-point grid. If $XX^\top = U\,\mathrm{diag}(q_j)U^\top$, the diagonal of the fitted-intercept ridge smoother is

$$h_i(\alpha) = \sum_j U_{ij}^2 \frac{q_j}{q_j + \alpha} + \frac{1}{n}, \tag{19}$$

and the LOOCV loss is the mean of $[(y_i - \widehat{y}_i)/(1 - h_i)]^2$. As in Eq. 16, this loss is quadratic in σ, so its full landscape was reconstructed from three cached coefficients. The green path in Fig. 3 is the median selected κ across trials. Metrics labeled empirical RidgeCV were instead evaluated at each trial's own selected penalty before aggregation; they are not evaluations at the median penalty. We verified the selected penalties and minimum losses against `sklearn.linear_model.RidgeCV` at three noise levels.

Aggregation and checks. At every grid point we saved all 100 trial values and computed their arithmetic mean, standard error, median, and 10th and 90th percentiles. Bands in Fig. 3 show the 10–90% trial range, not a confidence interval; markers show the arithmetic mean. We also checked the cached noise-polynomial evaluation against an independent direct ridge solve for all reported metrics. Computation used float64 linear algebra on a GPU, and plot-ready summaries were cached separately from the rendered figure.

B.5 Linear feature maps and prediction-null rotations

All feature-map experiments used the population covariance and disk-teacher alignment estimated in Section B.1. The finite-sample designs were Gaussian with these measured second moments: if $\Sigma =$

$U\,\mathrm{diag}(s_1,\ldots,s_d)U^\top$, with eigenvalues in decreasing order, we sampled independent population-PC scores and generated $y = {\beta^\star}^\top x + \sigma\epsilon$. Thus natural images determined the spectrum and teacher alignment, whereas the Monte Carlo experiments tested the Gaussian theory under controlled finite-sample sampling. Unless stated otherwise, $d = 10{,}000$, $n = 1000$, and $\sigma^2/S \in \{0.01, 0.1, 1, 10\}$, where $S = {\beta^\star}^\top \Sigma \beta^\star$.

Top-PC feature map and continuous null rotation. The reference map $F(0)$ returned the first p population-PC coordinates, with $p = 500$ in the main experiment and $p = 100$ in the companion experiment. Its feature covariance and feature–teacher covariance were

$$\Sigma_F = F(0)\Sigma F(0)^\top = \mathrm{diag}(s_1,\ldots,s_p),$$
$$F(0)\Sigma\beta^\star = (s_1\beta_1^\star,\ldots,s_p\beta_p^\star)^\top.$$

The unrepresented teacher power was treated as the independent residual variance $S_\perp$. We constructed prediction-equivalent maps in whitened coordinates. Writing $A = F\Sigma^{1/2}$, $b = \Sigma^{1/2}\beta^\star$, and $c = \Sigma^{-1/2}\beta^\star$, we decomposed the top-PC map as $A(0) = P + R$, where P is its projection onto $\mathrm{span}(b, c)$ and R is orthogonal to both constraints. We then formed a row frame V supported on the bottom 40% of the image spectrum, also orthogonal to b and c, and normalized it so that $VV^\top = RR^\top$ and all cross-Gram terms vanish. The continuous path was

$$A(t) = P + \sqrt{1-t}\,R + \sqrt{t}\,V, \qquad F(t) = A(t)\Sigma^{-1/2}, \qquad t = \sin^2\theta \in [0,1]. \tag{20}$$

It preserves, to numerical precision,

$$F(t)\Sigma F(t)^\top = \mathrm{diag}(s_1,\ldots,s_p), \quad F(t)\Sigma\beta^\star = F(0)\Sigma\beta^\star, \quad F(t)\beta^\star = F(0)\beta^\star, \tag{21}$$

while changing $G(t) = F(t)F(t)^\top$. The first two equalities fix the entire Gaussian prediction problem; the third additionally fixes the control numerator, isolating the effect of gradient geometry in the control denominator. We evaluated $t = 0$ and 61 logarithmically spaced values from 10^{-6} to 1. The tail frames and all Monte Carlo draws were generated from fixed seeds.

Prediction-selected penalty and Monte Carlo. For each $(p, \sigma^2/S)$, we evaluated 281 unnormalized ridge penalties $\alpha \in [10^{-4}, 10^5]$ on a logarithmic grid and selected the minimizer of the deterministic-equivalent leave-one-out prediction risk. The normalized penalty is $\lambda = \alpha/n$, and κ was obtained from the fixed-point equation on the invariant feature spectrum. This selection uses the population spectrum and teacher alignment and is therefore a theoretical prediction-optimal benchmark, not empirical cross-validation. The selected penalty was fixed across rotations and Monte Carlo trials.

Finite-n ridge fits were sampled exactly from the Gaussian feature design. Because Eq. 21 makes the joint law of $(F(t)x, y)$ identical for every t, a trial used one ridge estimate for all rotations and evaluated it under each $G(t)$. This pairing removes irrelevant Monte Carlo variation when comparing control geometries. We used 300 trials for $p = 500$ and 1000 trials for $p = 100$. Figure markers report trial means except for R^2_{acc}, whose heavy lower tail motivated the trial median; curves are leading deterministic equivalents. Population metrics were computed from the known Gaussian moments using the definitions in Section 3.

For every rotation and selected κ, we evaluated the control-variance geometry directly as

$$T_{F(t)}(\kappa) = \mathrm{Tr}\big[G(t)(\Sigma_F + \kappa I)^{-2}\Sigma_F\big]. \tag{22}$$

Because Σ_F is fixed along the path, changes in this trace arise only from the pixel-gradient Gram $G(t)$.

Held-out empirical-CV robustness experiment. To verify that the result was not an artifact of DE-based penalty selection, we repeated the $p = 500$ experiment with an explicit validation set. Each of 30 paired trials contained 1000 training and 500 validation examples. The underlying Gaussian inputs, teacher responses, response noise, and split were shared across rotations, but applying $F(t)$ produced a different realized feature matrix at every rotation. We therefore selected α separately for every trial, rotation, and noise level by minimizing held-out prediction MSE over 71 logarithmically spaced values from 10^{-2} to 10^5. No teacher coefficient, population risk, or control

outcome entered this selection. Only afterward did we use the known population moments to evaluate E_{gen}, E_{acc}, R^2_{gen}, and R^2_{acc}. We report trial medians and 10–90% intervals. The selected α values were also mapped through the fixed-point equation to trial-specific κ values for Eq. 22. The paired design isolates the rotation effect, while separate validation selection allows finite-sample Gram matrices and CV choices to differ across otherwise prediction-equivalent maps.

Random maps. The random-family experiment used a canonicalized $p = 500$ problem with $\Sigma_F = I_p$ and the realizable teacher aligned to the first feature axis. Of the remaining 499 joint-null rows, 300 were rotated and 199 were fixed. Each rotatable high-variance row was a unit direction supported on three PCs drawn from ranks $0.002d$–$0.15d$ and orthogonal to both b and c; its tail counterpart was constructed analogously from ranks $0.60d$–d. Fixed rows drew one coordinate from each intermediate band $0.15d$–$0.30d$, $0.30d$–$0.45d$, and $0.45d$–$0.60d$. For each of 300 random maps we sampled a global tail propensity $u \sim \text{Unif}(0, 1)$, independent row loadings $t_j \sim \text{Beta}(1 + 12u, 1 + 12(1 - u))$, and a random high–tail pairing. Every draw retained the same $(\Sigma_F, F\Sigma\beta^\star, F\beta^\star)$ but produced a different diagonal spectrum of G. We evaluated the maps using 2000 shared exact Gaussian ridge trials at each of the four noise levels.

Single-row rotation. Fig. 13 uses the smallest version of the construction, $p = 16$ with $\Sigma_F = I_p$. The first row is the teacher axis $b/\sqrt{S}$ in whitened coordinates. The second row rotates in one plane, $\cos\theta\, u_{\text{high}} + \sin\theta\, u_{\text{low}}$, where u_{high} and u_{low} are unit directions orthogonal to b and c, supported on the PC-rank bands 0.2–3% and 60–95% of the spectrum, and chosen to minimize and maximize the inverse-variance Rayleigh quotient within their bands (dominant PC ranks 21 and 9500). The remaining 14 rows are random orthonormal directions in the band 3–15%, also orthogonal to b and c, and fixed along the path. Every angle therefore satisfies Eq. 21, and only the second row's contribution to G changes. We used 41 angles in $[0, \pi/2]$, a penalty selected on a 281-point grid $\alpha \in [10^{-6}, 10^{8}]$ by the DE leave-one-out prediction risk, 500 shared exact ridge trials, and a 200,000-draw Gaussian distributional surrogate (App. J.2) for the control ratios.

B.6 Neuronal data

Recordings and stimulus sets. We reanalyzed the recordings and encoding-model outputs of Prince, Wang, et al. (2026) using the authors' released preprocessing code (commit `9b6fd95228d1b77146f8b279c7c228914ca34bb8`). The analysis contains five selected sites from each of five macaques (25 sites total): anterior IT in Monkey Red, central IT in Monkey Paul, three V3 and two V4 sites in Monkey Venus, and superior temporal sulcus sites in Monkeys Leap and Three0. Red and Paul were recorded with chronic 64-channel arrays; Venus, Leap, and Three0 were recorded with Neuropixels probes. Site selection followed the reliability screening of the original study. Area and subject are consequently partly confounded and we do not interpret area differences in the present analysis.

The encoding session contained 969 natural images: 649 images from the NSD shared-image set, 259 segmented object and animal images, and 61 fLoc images. A common fixed split assigned 774 images to model fitting and 195 to held-out evaluation. Control sessions were collected on 4–12 May 2025. For every site and encoding model, stimuli were synthesized from ten held-out NSD seed images toward as many as 11 evenly spaced target responses. Targets spanned the measured encoding-response range, from the first percentile minus one quarter bin width to the 99th percentile plus one half bin width. Depending on recording coverage, 36–110 accentuated images were available per site–model pair (median 102). The control sessions also repeated natural images shared with the encoding session. The held-out subset used below contained 50 images for Red, Paul, and Venus, 24 for Leap, and 22 for Three0; the complementary fitting-anchor counts were 50, 50, 50, 25, and 24.

Neural preprocessing. We used the response windows and exclusion rules of Prince, Wang, et al. (2026). Responses were averaged over 100–400 ms after image onset for Red and Paul and over 80–250 ms for Venus, Leap, and Three0. Within each control day, trials with robust deviation $|y - \text{median}(y)|/(1.4826\,\text{MAD}) > 15$ were removed. The remaining responses were standardized using that day's natural-image anchors; a day with fewer than five anchor trials used all of that day's trials. Responses were then averaged over stimulus repetitions. Near-duplicate Leap stimuli were merged

using the released pixel-correlation criterion ($r \geq 0.90$). At comparison time, predicted responses were lower bounded by the fitted neuron's firing floor, defined by the original pipeline.

Cross-session response alignment. The preceding day-wise anchor standardization is part of the upstream pipeline, but it does not place the encoding and control recordings on an identical affine scale. We therefore added one downstream map per site, shared by all ten models. On natural anchors belonging to the 774-image fitting split, we fitted

$$y_{is}^{\mathrm{ctrl}} = a_s + b_s y_{is}^{\mathrm{enc}} + \epsilon_{is}. \tag{23}$$

The fitted (a_s, b_s) was then frozen. A model prediction $\hat{y}_{ism}$ was mapped as $\hat{y}_{ism}^{\mathrm{ctrl}} = a_s + b_s \max\{\hat{y}_{ism}, f_s\}$, where f_s is the site's firing floor. No accentuated response entered this fit, and the same map was used for every model at a site. Held-out natural-image MSE, E_{val}, was evaluated on the disjoint control-session anchors listed above. The control MSE was the mean squared difference between mapped predictions and measured accentuated responses, without outcome refitting. Control slope was the ordinary-least-squares coefficient in the regression of measured on mapped predicted responses, with an intercept, over all available accentuated stimuli for that site–model pair. We also retained Pearson correlation and identity $R^2 = 1 - \sum_i (y_i - \hat{y}_i)^2 / \sum_i (y_i - \bar{y})^2$.

Stimulus synthesis. Accentuation used the released Fourier-parameterized feature-visualization procedure (Hamblin et al., 2024): a 1024-pixel image parameterization with inverse-frequency preconditioning, random crops and injected noise was optimized with NAdam (learning rate 12) for at most 6000 steps. The iterate whose post-hoc model prediction came closest to the requested target was retained. A prespecified ladder of augmentation-noise and Fourier-decay settings was searched independently for each site–model axis, as described by Prince, Wang, et al. (2026).

B.7 Encoding models

For each backbone and candidate layer, we flattened the layer activation and fitted a 750-component PCA transform on the 774 encoding-training images. PCA scores were centered but not whitened. A separate linear readout was fitted for every neural site using `RidgeCV` with an intercept and $\alpha \in \{10^{-4}, 10^{-3}, \ldots, 10^{9}\}$. The implementation used efficient leave-one-out ridge cross-validation to select α independently for each response channel. Candidate layers were ResNet and RegNet bottleneck blocks, AlexNet convolution/ReLU/max-pool stages, and transformer blocks (including the SigLIP2 attention pool). For each site and backbone, the layer with the largest R^2 on the fixed 195-image split was exported. Thus this split was used both for layer selection and for the reported encoding accuracy; its MSE is a post-selection validation measure rather than an untouched final test set.

The ten backbones and the number of distinct selected layer–PCA geometries are listed in Table 2. Sites that selected the same model, layer, PCA mean, and PCA axes shared one geometry calculation. Hashing these objects reduced 250 site–model readouts to 84 unique geometries, while all readout-specific regularization and biological outcomes remained at the 250-row level. Model loading, preprocessing, and layer names were taken from the released experimental utilities and reproduced with the same checkpoints.

B.8 Computing the gradient-geometry diagnostic

For a model–layer geometry, let s_k be the variance of PCA score k on the 774 fitting images and let α_{sm} be the selected scikit-learn ridge penalty for site s and model m. Because scikit-learn minimizes an unnormalized residual sum of squares, we set $\lambda_{sm} = \alpha_{sm}/n$ with $n = 774$ and solved the positive fixed-point equation

$$\kappa_{sm}\left[1 - \frac{1}{n}\sum_{k=1}^{750}\frac{s_k}{s_k + \kappa_{sm}}\right] = \lambda_{sm}. \tag{24}$$

The 750-dimensional training spectrum is used here as a plug-in approximation to the population spectrum. Across the 250 readouts, the median solved κ was 3.48 and the median $\mathrm{df}_2 = \sum_k [s_k/(s_k + \kappa)]^2$ was 231.5.

Model	Architecture	Training	# geometries
AlexNet (untrained)	AlexNet (TF-slim variant)	released untrained-control checkpoint[†]	11
ResNet-50	ResNet-50	ImageNet supervised	9
Robust ResNet-50	ResNet-50	ImageNet adversarial, ℓ_∞, $\epsilon = 8/255$	6
CLIP RN50	ResNet-50	CLIP contrastive	9
DINO RN50	ResNet-50	DINO self-supervised	6
RegNetY-640	RegNetY-640	SEER self-supervised	9
CLIPAG ViT-B/32	ViT-B/32	adversarially fine-tuned CLIP	7
DINOv2 ViT-B/14-reg	ViT-B/14 + registers	DINOv2 self-supervised	9
SigLIP2 ViT-B/16	ViT-B/16	SigLIP2	10
RADIO v2.5-B	ViT-B	RADIO distillation	8

Table 2: **Encoding models.** The final column counts distinct selected layer–PCA feature maps across the 25 sites (84 in total). [†]The released analysis designates the Brain-Score checkpoint `AlexNet_training_seed_01` as the untrained control. The robust ResNet-50 used checkpoint `imagenet_linf_8_pure.pt`; CLIPAG used `CLIPAG_ViTB32.pt`. Exact checkpoint identifiers and preprocessing are inherited from the released model-loading utilities.

We defined the differentiable input at the 224×224 RGB grid, before model-specific channel normalization, so $d = 150{,}528$ and derivatives are in RGB units on $[0, 1]$. The upstream resize from the synthesized image was not differentiated. For each of the ten fixed seed images and every PCA coordinate k, reverse-mode automatic differentiation gave

$$q_k(x) = \|J_\phi(x)^\top u_k\|_2^2, \qquad T_\phi(x, \kappa) = \sum_{k=1}^{750} q_k(x) \frac{s_k}{(s_k + \kappa)^2}. \tag{25}$$

The 750 vector–Jacobian products were evaluated in bounded blocks (up to 32 PCA directions per backward pass) and averaged over seeds. The primary theory-motivated proxy was $V_{\phi,sm} = E_{\mathrm{val},sm}\, \mathbb{E}_x[T_\phi(x, \kappa_{sm})]/n$, using the held-out control-session anchor MSE from Eq. 23.

We additionally evaluated finite-neighborhood versions at per-channel Gaussian noise scales $\tau \in \{0.5, 2, 8, 16\}/255$. Perturbed RGB values were not clipped. Each seed used $R = 64$ Gaussian outer directions and $L = 8$ independent neighborhood centers. The smoothed-Jacobian estimator targeted $\|\mathbb{E}_z J_\phi(x + \tau z)^\top u_k\|^2$ and removed finite-L square bias with the off-diagonal U-statistic of Eq. 106. The neighborhood estimator targeted $\mathbb{E}_z \|J_\phi(x + \tau z)^\top u_k\|^2$. We also retained the centered response covariance, total finite response change, its odd/even decomposition, and a local symmetric finite-difference check. Monte Carlo summaries were formed at the outer-direction level before summing over PCs, so cross-PC covariance was retained. As a separate forward-only validation, an antithetic Stein estimator used $R = 512$ common-prefix directions and replaced $f(x + \tau z) - f(x)$ by $[f(x + \tau z) - f(x - \tau z)]/2$ to cancel even-response variance.

All spectral vectors were stored before κ weighting, allowing traces to be recomputed without rerunning a network. The production run covered 84/84 unique geometries and 840 geometry–seed archives. Exact PCA-cache reproduction, serial-versus-batched VJPs, and the identities "variance plus drift equals total" and "odd plus even equals total" were checked; the largest relative decomposition residual was 5.96×10^{-8}. Nine CLIP-RN50 geometries had marginal finite-difference diagnostics, so their step-dependent smooth and neighborhood estimates were retained as sensitivities; exact VJPs and finite-response covariance do not depend on that step size. Total production compute was estimated at approximately 1.4 H100 GPU-hours.

B.9 Neuronal peer-review analysis

For every site s and generator model g, we selected the personalized accentuated stimuli generated for that site's recorded unit. Each of the ten reviewer models r predicted the identical stimulus set. Reviewer predictions were firing-floor clamped and transformed using the same frozen site-level map (a_s, b_s) as the self prediction. This yielded $25 \times 10 \times 10 = 2500$ site–generator–reviewer cells; rows of the displayed matrices are generators, columns are reviewers, and diagonal cells are

self-evaluations. Within each cell we computed MSE, Pearson r, identity R^2, and the OLS slope of measured on reviewer-predicted responses. Metrics were first computed within site and then averaged with equal weight over the 25 sites; stimuli were never pooled across sites. Available stimulus counts, total trials, and mean repeats were retained for every cell.

For the self-versus-peer comparison, the 10 diagonal and 90 off-diagonal cells of the 25-site-averaged matrix were summarized by medians and interquartile ranges. These cell counts are descriptive because rows and columns share models. Dependence-aware inference used three complementary contrasts. First, within each generator row we compared its diagonal value with the mean of its nine peer reviewers and applied a paired Wilcoxon test over ten generators. Second, we permuted the global reviewer labels 200,000 times while preserving the complete matrix rows and columns, testing whether the named self diagonal was unusually separated from a randomly matched diagonal. Third, we formed a peer-minus-self contrast within each site, averaged the five sites within each subject, and used 100,000 equal-subject bootstrap resamples for confidence intervals. With only five subjects, we also report the exact two-sided subject sign-flip test, whose minimum attainable p value when all subjects agree is 0.0625. For slope, signed deviation from one and $|\text{slope} - 1|$ were kept separate: a change from slope below one to above one is not by itself improved calibration.

B.10 Statistical analysis

The biological synopsis contains one row per site and model ($25 \times 10 = 250$). Positive geometry and error predictors were analyzed on a base-10 logarithmic scale. For Fig. 5, the biological control slope was centered within site using the mean of all ten models,

$$y^{\circ}_{sm} = y_{sm} - \frac{1}{10} \sum_{m'=1}^{10} y_{sm'}, \tag{26}$$

whereas the displayed predictors remained at their raw log scale. Subset comparisons used all ten models; the nine trained models; eight models after removing RN50-Robust and CLIPAG; and seven models after additionally removing the untrained AlexNet. The site means in Eq. 26 were estimated once from all ten models and held fixed before subset filtering. Direction-aligned correlations were reported as $-r$ for control slope, because the theory predicts that larger control-error proxies accompany smaller slopes.

The figure reports ordinary two-sided Pearson correlations and marks raw $p < 0.05$ as exploratory. The 250 site–model rows are repeated measurements, so these pointwise p values are not treated as independent-sample confirmatory tests. Companion tables report Spearman correlations and linear regressions with site-clustered standard errors. A sensitivity analysis also removed each site's mean from the log predictor; another recomputed site means within each 10/9/8/7-model subset. For the complete 28-predictor benchmark, Benjamini–Hochberg correction was applied within each biological endpoint and model subset; the paper-facing seven-predictor panel displays the unadjusted threshold and the full FDR view is reported separately.

Finite-image uncertainty in E_{val} was assessed with 2000 bootstrap resamples. Held-out image identities were resampled jointly for every site–model row within a monkey, preserving the shared-stimulus dependence; this procedure quantifies image-set uncertainty but not uncertainty across animals, neural trials, or Jacobian probes. As an out-of-sample sensitivity, fixed log-linear predictors of control slope were evaluated by leave-one-model- out and leave-one-subject-out prediction. Predictor standardization and all regression coefficients were fitted within each training fold, with no hyperparameter search on held-out outcomes. Throughout, we emphasize effect sizes, model-subset sensitivity, and held-out comparisons because only five subjects contribute independent animal-level replication.

B.11 Software and compute

Experiments were implemented in Python 3.10 using PyTorch 2.4, NumPy 2.0, SciPy 1.14, pandas 2.2, scikit-learn 1.6, and statsmodels 0.14; figures and progress monitoring used Matplotlib 3.8 and `tqdm`. Linear-feature and null-rotation simulations used the measured $d = 10{,}000$ natural-image spectrum, $n = 1{,}000$ samples, and feature dimension up to $p = 1{,}000$ ($p = 500$ in the main construction). These simulations ran on CPU and required less than five CPU-minutes in aggregate;

the full-dimensional PCA/whitening check required less than one minute on a single NVIDIA H100 80-GB GPU.

The largest pixel-space landscape comprised 100 paired natural-image trials, 114 noise levels, and 478 ridge values. Exact reuse of ridge-path coefficients across noise levels reduced this to one factorization per trial; including Gaussian-surrogate evaluation, the landscape required approximately one H100 GPU-minute. The nonlinear gradient-geometry analysis evaluated 10 pretrained encoders at 84 unique layer geometries (250 site–model pairs), using 750 feature PCs and 10 seed images. Batched vector–Jacobian products and finite-neighborhood diagnostics required approximately two recorded H100 GPU-hours, or a few GPU-hours including validation and reruns. Intermediate arrays and summaries were cached separately so that figures could be regenerated without recomputing model gradients; the principal cached outputs occupied approximately 6 GB, excluding source datasets and pretrained model weights.

C THEORY RESULTS ROADMAP

The theory part of the appendix is organized in two layers (the extended results and the empirical methods precede it, in Sections A and B). Section D states every theoretical result of the paper once, in the order in which it is derived, so that it can be read as a self-contained summary. The sections that follow contain the proofs, arranged by topic; each begins with a short motivation, restates the result it proves, gives the proof in numbered steps, and closes with remarks on interpretation. Readers who want the statements only can stop at Section D; readers who want a specific derivation can jump to it by the pointers given there. Unless stated otherwise, every result concerns the linear teacher–student model with Gaussian inputs; the nonlinear diagnostic of Section K is the only exception, and it is a diagnostic rather than a theorem.

1. **Evaluation geometry** (Section E). Generalization, self-accentuation and peer review are evaluations of one fitted weight under three input covariances. Their equal-error sets are covariance ellipsoids, Euclidean spheres and peer-normal hyperplanes.
2. **Pixel-space ridge** (Section F). Deterministic equivalents for the generalization risk, for the two accentuation forms $N = \hat{\beta}^\top \beta^\star$ and $D = \hat{\beta}^\top \hat{\beta}$, and for the per-eigenmode weight error. The Euclidean estimation-noise term in D is what prediction does not see.
3. **Prediction- versus control-optimal tuning** (Section G). The two optima equate the noisy prediction risk to different teacher-weighted spectral averages; prediction-optimal cross-validation leaves a computable residual control error.
4. **Feature-space regression** (Section H). Ridge on a fixed linear feature map inherits the pixel-space prediction theory with effective noise $\sigma^2 + S_\perp$, while pixel-space control additionally depends on $FF^\top$ and $F\beta^\star$.
5. **Gauge invariance and gradient fragility** (Section I). Transformations preserving $F\Sigma F^\top$ and $F\Sigma\beta^\star$ preserve prediction and its cross-validated optimum but not control; the fragility factor T_F is a ridge-weighted ratio of Euclidean to natural gradient energy.
6. **Ratios of random quadratic forms and peer review** (Section J). Second-order and distributional refinements of the ratio-of-DE approximation, the deterministic equivalent of peer review between independent fits, and the exact conditional form of peer review between models fitted to the same data through different feature maps.
7. **Nonlinear diagnostic** (Section K). Replacing F by a local Jacobian, with its estimators and its empirical association with neuronal control.
8. **Random-matrix toolkit** (Section L). The fixed-point equation, the one- and two-resolvent deterministic equivalents, and the identities for the spectral sums $B_{p,q}$ and $\mathrm{df}_{p,q}$ used throughout.

D MAIN THEORY RESULTS IN DERIVATION ORDER

This section collects every theoretical statement used in the paper, in the order in which it is derived, so that a reader who prefers a linear exposition can read it as a self-contained summary before the proofs. Each statement names the section that proves it. All results concern the linear teacher–student model of Section D.1; only Section D.7 concerns nonlinear feature maps, and it is a diagnostic rather than a theorem.

D.1 SETTING, ESTIMATOR AND EVALUATION DISTRIBUTIONS

Data and teacher. Inputs are $\mathbf{x} \sim \mathcal{N}(0, \Sigma)$ with $\Sigma \succeq 0$ diagonalized as $\Sigma = \sum_{k=1}^{d} s_k \mathbf{u}_k \mathbf{u}_k^\top$. The teacher is linear with weight $\beta^\star = \sum_k b_k \mathbf{u}_k$ and the observed response is

$$y = \mathbf{x}^\top \beta^\star + \sigma\epsilon, \qquad \epsilon \sim \mathcal{N}(0,1), \qquad S := \beta^{\star\top} \Sigma \beta^\star = \sum_k s_k b_k^2. \tag{27}$$

Given n samples stacked as $X \in \mathbb{R}^{n\times d}$, $\mathbf{y} = X\beta^\star + \sigma\eta$, write $\hat{\Sigma} = X^\top X/n$ and $\gamma = d/n$.

Students. A pixel-space student is $f(\mathbf{x}) = \mathbf{x}^\top \hat{\beta}$. A feature-space student uses a fixed linear map $F \in \mathbb{R}^{p\times d}$ and reads out $f(\mathbf{x}) = \hat{w}^\top F\mathbf{x}$; its effective pixel weight is $\hat{\beta} := F^\top \hat{w}$, which is also the input gradient $\nabla_{\mathbf{x}} f$. Ridge regression with penalty λ gives

$$\hat{\beta} = (\hat{\Sigma}+\lambda I)^{-1}\hat{\Sigma}\beta^\star + \sigma(\hat{\Sigma}+\lambda I)^{-1}\tfrac{1}{n}X^\top\eta \quad \text{(pixel)}, \qquad \hat{w} = (\hat{\Sigma}_F+\lambda I)^{-1}\tfrac{1}{n}(FX^\top)\mathbf{y} \quad \text{(feature)}, \tag{28}$$

with $\hat{\Sigma}_F = FX^\top XF^\top/n$. Software that minimizes $\|\mathbf{y} - X\beta\|^2 + \alpha\|\beta\|^2$ uses $\alpha = n\lambda$.

Three evaluation distributions. The same fitted weight is evaluated on three input distributions.

1. *Generalization*: $\mathbf{x} \sim \mathcal{N}(0, \Sigma)$.
2. *Self-accentuation*: $\mathbf{x} = \mathbf{x}_0 + \alpha\hat{\beta}$, $\alpha \sim p(\alpha)$, the idealized gradient-ascent path of the student. The main text takes $\mathbf{x}_0 = 0$, $\mathbb{E}[\alpha] = 0$, and the variance-matched range $\mathrm{Var}[\alpha]\,(\hat{\beta}^\top\hat{\beta})^2 = S$, so that the student-predicted response variance along the path equals the natural response variance.
3. *Peer review*: $\mathbf{x} = \mathbf{x}_0 + \alpha\beta'$ for a path generated by another student β', with the analogous variance matching.

Nonzero and multiple seeds are treated in Section E.3.

D.2 Metrics on an arbitrary evaluation distribution

Lemma 5 (Metrics of a linear predictor against a noiseless linear teacher)**.** *Let the evaluation inputs have mean μ_p and covariance Σ_p, and let $\Delta := \hat{\beta} - \beta^\star$. Then*

$$\mathrm{MSE} = \Delta^\top(\mu_p\mu_p^\top + \Sigma_p)\Delta, \qquad R^2 = 1 - \frac{\Delta^\top(\mu_p\mu_p^\top + \Sigma_p)\Delta}{\beta^{\star\top}\Sigma_p\beta^\star}, \tag{29}$$

$$\mathrm{slope} = \frac{\hat{\beta}^\top\Sigma_p\beta^\star}{\hat{\beta}^\top\Sigma_p\hat{\beta}}, \qquad r = \frac{\hat{\beta}^\top\Sigma_p\beta^\star}{\sqrt{(\hat{\beta}^\top\Sigma_p\hat{\beta})(\beta^{\star\top}\Sigma_p\beta^\star)}}, \tag{30}$$

where slope is the regression coefficient of the true response on the predicted response and R^2 is the direct (uncalibrated) coefficient of determination. Proof: Lemmas 23–26.

Corollary 6 (The three evaluations)**.** *Substituting the three covariances Σ, $\mathrm{Var}[\alpha]\hat{\beta}\hat{\beta}^\top$, and $\mathrm{Var}[\alpha]\beta'\beta'^\top$ into Lemma 5 gives Table 3. In particular, with $N := \hat{\beta}^\top\beta^\star$, $D := \hat{\beta}^\top\hat{\beta}$ and variance matching,*

$$\frac{E_{gen}}{S} = \frac{\Delta^\top\Sigma\Delta}{S}, \qquad \frac{E_{acc}}{S} = \Big(1-\frac{N}{D}\Big)^2, \qquad R^2_{acc} = 1-\Big(\frac{D}{N}-1\Big)^2, \qquad \mathrm{slope}_{acc} = \frac{N}{D}. \tag{31}$$

Pearson r is identically 1 (or undefined) on a noiseless rank-one path and is therefore not informative for accentuation. All accentuation metrics are functions of the single scalar $z_{acc} := D/N - 1$. Proof: Section E.3.

Proposition 7 (Geometry of equal-error sets)**.** *Fix $\beta^\star$ and $\Sigma \succ 0$. In the space of student weights $\hat{\beta}$:*

1. *the equal-E_{gen} sets are ellipsoids $\Delta^\top\Sigma\Delta = c$ centered at $\beta^\star$, with semiaxis $\sqrt{c/s_k}$ along $\mathbf{u}_k$;*
2. *the equal-z_{acc} sets (hence equal E_{acc}, R^2_{acc}, slope) are Euclidean spheres with diameter from 0 to $(1+z)\beta^\star$: $\|\hat{\beta} - \frac{1+z}{2}\beta^\star\|^2 = \frac{(1+z)^2}{4}\|\beta^\star\|^2$;*
3. *for a fixed peer β', the equal-peer-error sets are hyperplanes normal to β' through $a\beta^\star$ with $a = \beta'^\top\hat{\beta}/\beta'^\top\beta^\star$.*

Proof: Section E.4.

Example 8 (Exact dissociation without learning)**.** Let $\Sigma = \mathrm{diag}(1, \varepsilon)$, $\beta^\star = \hat{\beta}_1 = (1, 0)^\top$ and $\hat{\beta}_2 = (1, \varepsilon^{-q})^\top$ with $0 < q < 1/2$. As $\varepsilon \to 0$, $E_{gen}(\hat{\beta}_2) = \varepsilon^{1-2q} \to 0$ while $\mathrm{slope}_{acc}(\hat{\beta}_2) \to 0$ and $R^2_{acc}(\hat{\beta}_2) \to -\infty$; both peer-review scores $R^2_{peer}(\hat{\beta}_2 \leftarrow \hat{\beta}_1)$ and $R^2_{peer}(\hat{\beta}_1 \leftarrow \hat{\beta}_2)$ equal 1. Thus indistinguishable prediction, arbitrarily poor control, and perfect mutual peer review coexist.

D.3 Deterministic-equivalent toolkit

Effective regularization and spectral sums. For a population spectrum $\{s_k\}$ and aspect ratio $\gamma = d/n$, let $\kappa = \kappa(\lambda)$ be the unique positive solution of

$$\kappa - \lambda = \frac{\kappa}{n} \operatorname{Tr}\big[\Sigma(\Sigma + \kappa I)^{-1}\big] = \frac{\kappa}{n} \operatorname{df}_1(\kappa), \tag{32}$$

and define, for $p, q \geq 0$,

$$B_{p,q}(\kappa) = \sum_k \frac{s_k^p}{(s_k + \kappa)^q} b_k^2, \qquad \operatorname{df}_{p,q}(\kappa) = \sum_k \frac{s_k^p}{(s_k + \kappa)^q}, \qquad \operatorname{df}_1 := \operatorname{df}_{1,1},\ \operatorname{df}_2 := \operatorname{df}_{2,2}. \tag{33}$$

The relation $n\lambda = \kappa\,(n - \operatorname{df}_1)$ and the derivative $d\kappa/d\lambda = n/(n - \operatorname{df}_2)$ follow from Eq. 32. The notation $A \asymp B$ means that $A - B \to 0$ almost surely in the proportional limit $n, d \to \infty$, $d/n \to \gamma$; the one- and two-resolvent deterministic equivalents used throughout are listed in Section L.

D.4 Pixel-space ridge regression

Proposition 9 (Per-eigenmode weight error)**.** *For pixel ridge under Gaussian design,*

$$\mathbb{E}\big[(\mathbf{u}_k^\top(\hat\beta - \beta^\star))^2\big] \asymp \underbrace{\frac{\kappa^2}{(s_k + \kappa)^2} b_k^2}_{\textit{shrinkage bias}} + \underbrace{\frac{s_k}{(s_k + \kappa)^2} \frac{\kappa^2 B_{1,2}(\kappa) + \sigma^2}{n - \operatorname{df}_2(\kappa)}}_{\textit{finite-sample signal and noise variance}}. \tag{34}$$

The variance term is largest for modes with $s_k \approx \kappa$ and, relative to the mode's own variance s_k, is largest for low-variance modes. Proof: Section F.

Proposition 10 (Pixel-ridge deterministic equivalents for prediction and control)**.** *With κ from Eq. 32 on the spectrum of Σ,*

$$E_{gen} \asymp E_{gen,DE} := \frac{n\big(\kappa^2 B_{1,2} + \sigma^2\big)}{n - \operatorname{df}_2} - \sigma^2, \tag{35}$$

$$N \asymp \mu_N := B_{1,1}, \tag{36}$$

$$D \asymp \mu_D := B_{2,2} + \underbrace{\frac{\operatorname{df}_{1,2}}{n}\big(E_{gen,DE} + \sigma^2\big)}_{V_{\mathrm{pix}}(\kappa)} = \|\bar\beta_\kappa\|^2 + V_{\mathrm{pix}}(\kappa), \qquad \bar\beta_\kappa := \Sigma(\Sigma + \kappa I)^{-1}\beta^\star. \tag{37}$$

Equation 35 is the sum of Eq. 34 weighted by s_k; Eq. 37 is its unweighted sum plus the squared mean. The leading (ratio-of-DE) accentuation metrics follow from Eq. 31 with $N/D \approx \mu_N/\mu_D$, i.e.

$$z_{acc}^{(0)} = \frac{\mu_D}{\mu_N} - 1 = \frac{B_{2,2} - B_{1,1}}{B_{1,1}} + \frac{\operatorname{df}_{1,2}\,(E_{gen,DE} + \sigma^2)}{n\,B_{1,1}}. \tag{38}$$

The first term is a shrinkage bias ($B_{2,2} \leq B_{1,1}$, so it is negative: ridge shrinkage makes the student under-shoot along its own path); the second term is Euclidean estimation variance, which prediction weights by s_k but control counts in full. Proof: Section F.

Proposition 11 (Prediction-optimal and control-optimal regularization)**.**

1. *If $E_{gen,DE}(\kappa)$ has an interior differentiable minimizer $\kappa^\star_{gen}$ on the admissible $\lambda \geq 0$ path, then*

$$E_{gen,DE} + \sigma^2 = n\kappa \frac{B_{2,3}}{\operatorname{df}_{2,3}} \quad \textit{at } \kappa = \kappa^\star_{gen}. \tag{39}$$

2. *At a point with $\mu_N > 0$, the leading accentuation error vanishes ($\mu_D = \mu_N$, i.e. $z_{acc}^{(0)} = 0$, $R_{acc}^{2,(0)} = 1$, $\operatorname{slope}_{acc}^{(0)} = 1$) if and only if*

$$E_{gen,DE} + \sigma^2 = n\kappa \frac{B_{1,2}}{\operatorname{df}_{1,2}} \qquad \Longleftrightarrow \qquad \sigma^2 = n\lambda \frac{B_{1,2}}{\operatorname{df}_{1,2}}. \tag{40}$$

At such a root $\kappa^\star_{acc}$ the prediction risk is $E_{gen,DE}(\kappa^\star_{acc}) = \sigma^2 \operatorname{df}_1/(n - \operatorname{df}_1)$.

3. *At an interior prediction optimum the residual leading control error is*

$$z^{(0)}_{acc,CV} = \frac{\kappa\,\mathrm{df}_{1,2}}{B_{1,1}}\left(\frac{B_{2,3}}{\mathrm{df}_{2,3}} - \frac{B_{1,2}}{\mathrm{df}_{1,2}}\right)_{\kappa=\kappa^\star_{gen}}, \tag{41}$$

which is zero only when the two teacher-weighted spectral averages coincide, e.g. for isotropic Σ, but not in general for anisotropic spectra.

The zero-error statement concerns the leading ratio-of-DE approximation; the finite-sample expectation retains $\mathrm{Var}(N/D)$. Existence, uniqueness, and boundary optima are discussed in Section G. Proof: Section G.

Proposition 12 (Peer review between independent ridge fits)**.** *Let $\hat\beta$ and β' be ridge fits from independent samples with the same κ. Then $\hat\beta^\top\beta' \asymp B_{2,2}$ and $\beta^{\star\top}\beta' \asymp B_{1,1}$, so the leading peer gain and score are*

$$G_{peer} := \frac{\hat\beta^\top\beta'}{\beta^{\star\top}\beta'} \asymp \frac{B_{2,2}}{B_{1,1}}, \qquad R^{2,(0)}_{peer} = 1 - \left(1 - \frac{B_{2,2}}{B_{1,1}}\right)^2. \tag{42}$$

Equivalently, with $z_{peer} := \hat\beta^\top\beta'/\beta^{\star\top}\beta' - 1 \asymp B_{2,2}/B_{1,1} - 1 \in (-1, 0]$, the self-review error of Proposition 10 decomposes as $z^{(0)}_{acc} = z_{peer} + V_{\mathrm{pix}}/B_{1,1}$: self-review is peer review plus the norm inflation. The peer path removes the Euclidean variance term of Eq. 37 because independent noise realizations are uncorrelated; peer review therefore measures shared shrinkage bias only. Proof: Section J, peer-review subsection.

D.5 Beyond ratio of means

Proposition 13 (Second-order ratio correction)**.** *Write $N = \mu_N + \delta_N$, $D = \mu_D + \delta_D$. To second order,*

$$\mathbb{E}\Big[\frac{N}{D}\Big] \approx \frac{\mu_N}{\mu_D}\Big(1 - \frac{\mathrm{Cov}(N,D)}{\mu_N\mu_D} + \frac{\mathrm{Var}(D)}{\mu_D^2}\Big), \qquad \mathbb{E}\Big[\Big(1 - \frac{N}{D}\Big)^2\Big] = \Big(1 - \mathbb{E}\Big[\frac{N}{D}\Big]\Big)^2 + \mathrm{Var}\Big(\frac{N}{D}\Big). \tag{43}$$

With $\hat\beta = g + h$ (g the shrunken signal, h the noise part), the response-noise contribution is

$$\mathrm{Var}\Big(\frac{N}{D}\Big) \approx \frac{1}{\mu_D^2}\Big[\frac{\sigma^2}{n}\sum_k \frac{s_k b_k^2}{(s_k+\kappa)^2}\Big(1 - \frac{2\mu_N}{\mu_D}\frac{s_k}{s_k+\kappa}\Big)^2 + \frac{2\mu_N^2\sigma^4}{\mu_D^2 n^2}\sum_k \frac{s_k^2}{(s_k+\kappa)^4}\Big], \tag{44}$$

which omits random-design fluctuations and therefore under-estimates the variance near ridgeless operating points. The variance term dominates E_{acc} precisely when the mean calibration error is small (e.g. near $\kappa^\star_{acc}$), which is why leading DEs predict the location of the control optimum but not the value of E_{acc} there. Proof: Section J.1.

Remark 1 (Gaussian distributional surrogate). Approximating $\hat\beta$ by a Gaussian with the DE mean $\Sigma(\Sigma+\kappa I)^{-1}\beta^\star$ and diagonal per-mode variance from Eq. 34, and pushing it through N and D, yields medians, quantiles and the probability of sign changes of N. It is a surrogate, not a proved distributional deterministic equivalent: it does not enforce the exact ridgeless identity $N = D$, and it inherits the missing random-design covariance. Section J.2.

D.6 Linear feature-space ridge regression

Lemma 14 (Two population-optimal feature weights)**.** *Let $\Sigma_F := F\Sigma F^\top$ and $G := FF^\top$. The feature weight minimizing population prediction error and its irreducible error are*

$$w^\star = \Sigma_F^\dagger F\Sigma\beta^\star, \qquad S_\perp := \mathbb{E}_{\mathbf{x}}\big[(\mathbf{x}^\top\beta^\star - \mathbf{x}^\top F^\top w^\star)^2\big] = S - w^{\star\top}\Sigma_F w^\star, \tag{45}$$

whereas the feature weight whose pixel image best reconstructs the teacher is $w^P = G^\dagger F\beta^\star$. They coincide when $\Sigma \propto I$ on the row space of F and differ otherwise. Proof: Section H.

Proposition 15 (Feature-ridge deterministic equivalents)**.** *Let κ solve Eq. 32 on the spectrum of Σ_F with aspect ratio p/n, and define the feature-space spectral sums*

$$B^F_{a,b}(\kappa) := w^{\star\top}\Sigma_F^a(\Sigma_F+\kappa I)^{-b}w^\star, \qquad \mathrm{df}^F_{a,b}(\kappa) := \mathrm{Tr}\big[\Sigma_F^a(\Sigma_F+\kappa I)^{-b}\big]. \tag{46}$$

Then

$$E_{gen,DE} = \frac{n}{n - \mathrm{df}_2^F}\left(\kappa^2 B_{1,2}^F + S_\perp + \sigma^2\right) - \sigma^2, \tag{47}$$

$$N_{DE} = {\beta^\star}^\top \bar{\beta}_{F,\kappa}, \qquad \bar{\beta}_{F,\kappa} := F^\top(\Sigma_F + \kappa I)^{-1} F\Sigma\beta^\star = F^\top \Sigma_F (\Sigma_F + \kappa I)^{-1} w^\star, \tag{48}$$

$$D_{DE} = \|\bar{\beta}_{F,\kappa}\|^2 + \underbrace{\frac{E_{gen,DE} + \sigma^2}{n} T_F(\kappa)}_{V_F(\kappa)}, \qquad T_F(\kappa) := \mathrm{Tr}\big[G(\Sigma_F + \kappa I)^{-2}\Sigma_F\big]. \tag{49}$$

Setting $F = I$ recovers Proposition 10. The effective training noise for learning $w^\star$ is $\sigma^2 + S_\perp$. Proof: Section H.

Corollary 16 (Distinct summary statistics for prediction and control)**.** *At leading order, $E_{gen,DE}(\kappa)$ for every κ, hence the DE-predicted CV choice κ_{CV} and R^2_{gen}, are functions of*

$$\mathcal{P} := \big(\Sigma_F,\ F\Sigma\beta^\star,\ S,\ \sigma^2,\ n\big) \tag{50}$$

only, whereas N_{DE}, D_{DE} and every accentuation metric additionally depend on the Euclidean geometry $G := FF^\top$ (through T_F and $\|\bar{\beta}_{F,\kappa}\|^2$) and on the teacher image $F\beta^\star$ (through N_{DE}). Fixing $\mathcal{P}$ fixes prediction; it does not fix control (Figure 27).

Proposition 17 (Prediction- and control-optimal feature ridge)**.** *An interior minimizer of Eq. 47 satisfies $E_{gen,DE} + \sigma^2 = n\kappa B^F_{2,3}/\mathrm{df}^F_{2,3}$, the same form as Eq. 39 with $\sigma^2 \mapsto \sigma^2 + S_\perp$ inside $E_{gen,DE}$. The leading accentuation error vanishes exactly when*

$$E_{gen,DE} + \sigma^2 = n\,\frac{{\beta^\star}^\top \bar{\beta}_{F,\kappa} - \|\bar{\beta}_{F,\kappa}\|^2}{T_F(\kappa)} = n\,\frac{(w^P - w^\star)^\top G\bar{w}_\kappa + \kappa\, {w^\star}^\top(\Sigma_F + \kappa I)^{-1} G\bar{w}_\kappa}{T_F(\kappa)}, \tag{51}$$

with $\bar{w}_\kappa := \Sigma_F(\Sigma_F + \kappa I)^{-1} w^\star$. The first condition depends on $\mathcal{P}$ alone; the second depends on G and $F\beta^\star$. An admissible root is not guaranteed for an arbitrary F; in particular, when $w^P \neq w^\star$ the numerator can be negative for all κ, in which case no ridge level calibrates control. Proof: Section H.4.

Proposition 18 (Prediction-preserving gauge family)**.** *Assume $\Sigma \succ 0$ and let $\mathcal{G}_b := \{Q \in O(d) : Q\Sigma^{1/2}\beta^\star = \Sigma^{1/2}\beta^\star\}$. For $Q \in \mathcal{G}_b$ the map $F \mapsto F_Q := F\Sigma^{1/2}Q\Sigma^{-1/2}$ preserves Σ_F and $F\Sigma\beta^\star$, hence every element of $\mathcal{P}$, the Gaussian training law of $(F\mathbf{x}, y)$, the full curve $E_{gen,DE}(\kappa)$ and the DE-based CV choice; it changes G and $F\beta^\star$ in general, and therefore changes N_{DE}, D_{DE} and all accentuation metrics. The subgroup $\mathcal{G}_{b,c} := \{Q \in \mathcal{G}_b : Q\Sigma^{-1/2}\beta^\star = \Sigma^{-1/2}\beta^\star\}$ additionally fixes $F\beta^\star$, isolating G as the only control-relevant statistic that varies. When $\Sigma = I$ the family acts trivially on both prediction and control. Members of the family need not give identical CV choices for the same finite realization of X. Proof: Section I.*

Corollary 19 (Continuous prediction-null rotation)**.** *Write $A := F\Sigma^{1/2}$ and let P be the part of A needed to fix $A\Sigma^{1/2}\beta^\star$ (and optionally $A\Sigma^{-1/2}\beta^\star$). For any two row frames R, V in the orthogonal complement with $RR^\top = VV^\top$ and $RV^\top = 0$, the path $A(t) = P + \sqrt{1-t}\,R + \sqrt{t}\,V$, $F(t) = A(t)\Sigma^{-1/2}$, $t \in [0,1]$, lies in the gauge family. Choosing R in high-variance and V in low-variance pixel directions moves feature energy continuously from the spectral head to the tail: E_{gen}, R^2_{gen} and κ_{CV} are constant in t while $\mathrm{Tr}(F(t)F(t)^\top)$, T_F and z_{acc} grow without bound. Proof: Section I, construction subsection.*

Proposition 20 (Spectral form of the fragility factor)**.** *Let $\Sigma_F = \sum_k s_k v_k v_k^\top$ and $g_k := F^\top v_k$ be the pixel gradient of feature PC k. Then*

$$T_F(\kappa) = \sum_{k:s_k>0} \left(\frac{s_k}{s_k + \kappa}\right)^2 \frac{\|g_k\|^2}{g_k^\top \Sigma g_k}, \qquad s_k = g_k^\top \Sigma g_k. \tag{52}$$

T_F is a ridge-weighted average, over feature PCs that survive shrinkage, of the ratio between the Euclidean and the natural-covariance energy of the backpropagated gradient. It is response-independent and large when feature PCs backpropagate into low-variance (high spatial frequency) input directions. Proof: Section I.4.

Proposition 21 (Peer review on shared data across feature maps)**.** *Let two students with maps F_1, F_2 and penalties λ_1, λ_2 be fitted to the same (X, η), and write $M_i := F_i^\top(F_i\hat{\Sigma}F_i^\top + \lambda_i I)^{-1}F_i$, so that*

$\hat\beta_i = M_i\hat\Sigma\beta^\star + \frac{\sigma}{n}M_iX^\top\eta =: g_i + \varepsilon_i$. *Conditionally on* X,

$$\mathbb{E}_\eta\big[\hat\beta_1^\top\hat\beta_2\big] = g_1^\top g_2 + \Omega_{12}, \qquad \Omega_{12} := \frac{\sigma^2}{n}\,\mathrm{Tr}\,\big[M_1\hat\Sigma M_2\big], \qquad |\Omega_{12}| \le \sqrt{\Omega_{11}\Omega_{22}}, \tag{53}$$

while $\mathbb{E}_\eta[\beta^{\star\top}\hat\beta_2] = \beta^{\star\top}g_2$. *The peer gain of student 1 on student 2's path is therefore, at the level of conditional means,* $G_{1\leftarrow 2} \approx (g_1^\top g_2 + \Omega_{12})/\beta^{\star\top}g_2$, *to be compared with the self gain* $G_{2\leftarrow 2} \approx (\|g_2\|^2 + \Omega_{22})/\beta^{\star\top}g_2$. *The shared-noise overlap* Ω_{12} *is the inner product of the two models' noise-driven pixel weights: it equals the self noise energy* Ω_{22} *for identical students, vanishes for fits on independent data (recovering Proposition 12, generalized to different maps by* $g_i \asymp \bar\beta_i$*), and is small whenever the two maps route estimation noise into different pixel directions. All terms are explicit functions of the empirical design; their population deterministic equivalents for two different feature maps on shared data are left open. Proof: Section J.4.*

Remark 2 (When is control fragile although prediction is fine). Combining Propositions 15, 17 and 20, the leading control error at fixed prediction error grows with three separable factors: (i) the estimation-noise level $(E_{gen,DE} + \sigma^2)/n$; (ii) the spectral mismatch of Eq. 41, i.e. how far the CV-selected κ sits from the control-calibrated κ, which vanishes for isotropic data and grows with anisotropy and with teacher energy in low-variance modes; (iii) the gradient-geometry factor T_F, which vanishes only if surviving feature gradients live in high-variance input directions. Only factor (i) is visible from held-out prediction.

D.7 Nonlinear feature maps: a heuristic diagnostic

Local diagnostic. For a differentiable feature map $\phi : \mathbb{R}^d \to \mathbb{R}^p$ with Jacobian $J_\phi(x)$ and feature covariance $\Sigma_\phi = \sum_k s_k v_k v_k^\top$ on natural inputs, define

$$T_\phi(x,\kappa) := \sum_k \frac{s_k}{(s_k+\kappa)^2}\|J_\phi(x)^\top v_k\|^2, \qquad V_\phi := \frac{E_{val}}{n}\,\mathbb{E}_x[T_\phi(x,\kappa)], \tag{54}$$

where E_{val} is the held-out prediction error and κ is computed from the empirical feature spectrum and the CV-selected penalty. For $\phi(x) = Fx$ this equals T_F and the variance term of Eq. 49. For nonlinear ϕ the identity $s_k = g_k^\top\Sigma g_k$ fails, so Eq. 54 is a diagnostic motivated by the linear theory; it is evaluated in exact-local, finite-difference, smoothed-Jacobian, neighborhood and finite-step versions (Section K).

Empirical claim. On the neuronal control data (25 sites, 10 encoding models), with each site's mean control outcome removed, V_ϕ and T_ϕ are associated with control slope and with direct control error across the nine trained models, whereas held-out prediction error is not; the association is weaker but same-signed among the seven conventionally trained models. These are descriptive correlations with repeated models within sites, not independent-sample tests (Sec. 6, Fig. 5).

E Evaluation geometry: setting, metrics and equal-error sets

This section fixes the model, derives the closed forms of the evaluation metrics on an arbitrary input distribution, specializes them to the three distributions that matter for prediction and control, and proves the geometric description of their equal-error sets. Nothing here involves random-matrix theory: all statements are exact for a fixed fitted weight $\hat\beta$.

E.1 Setting

Teacher, data and student. Inputs are drawn from a centered Gaussian with a structured covariance, $\mathbf{x} \sim \mathcal{N}(0,\Sigma)$, and responses come from a noisy linear teacher, the idealization of a neuron with ground-truth weight $\beta^\star$,

$$y = \beta^{\star\top}\mathbf{x} + \sigma\epsilon, \qquad \epsilon \sim \mathcal{N}(0,1). \tag{55}$$

The student, the idealization of a neuron predictive model, is the bias-free linear predictor $\hat y = \hat\beta^\top\mathbf{x}$. The n training inputs are stacked as rows of $X \in \mathbb{R}^{n\times d}$ and the realized unit-variance noise as $\eta \in \mathbb{R}^n$, so that $\mathbf{y} = X\beta^\star + \sigma\eta$ and $\hat\Sigma = X^\top X/n$. Data are centered and biases are ignored throughout; we study the effect of the sample size n and the noise level σ.

Lemma 22 (Ridge estimator as shrunken signal plus noise)**.** *The ridge estimator with penalty λ, $\hat{\beta} = \arg\min_\beta \frac{1}{n}\|X\beta - \mathbf{y}\|^2 + \lambda\|\beta\|^2$, equals*

$$\begin{aligned} \hat{\beta} &= \underbrace{(\hat{\Sigma} + \lambda I_d)^{-1}\hat{\Sigma}\beta^\star}_{\textit{signal}} + \underbrace{\sigma(\hat{\Sigma} + \lambda I_d)^{-1}\tfrac{1}{n}X^\top\eta}_{\textit{noise}}, \\ \hat{\beta} - \beta^\star &= -\lambda(\hat{\Sigma} + \lambda I_d)^{-1}\beta^\star + \sigma(\hat{\Sigma} + \lambda I_d)^{-1}\tfrac{1}{n}X^\top\eta. \end{aligned} \tag{56}$$

Proof. The normal equations give $\hat{\beta} = (X^\top X + n\lambda I_d)^{-1}X^\top\mathbf{y}$. Substituting $\mathbf{y} = X\beta^\star + \sigma\eta$ and dividing numerator and denominator by n gives the first identity. The second follows from $(\hat{\Sigma} + \lambda I)^{-1}\hat{\Sigma} - I = -\lambda(\hat{\Sigma} + \lambda I)^{-1}$. □

Idealized feature accentuation. Feature accentuation moves a seed image $\mathbf{x}_0$ along the input gradient of the predictive model until the prediction reaches a target level $y^\star$. For a linear student the gradient is constant, $\nabla_\mathbf{x} f = \hat{\beta}$, so the iteration $\mathbf{x} \leftarrow \mathbf{x} + \alpha\nabla_\mathbf{x} f$ stays on a line and the target is reached at

$$\mathbf{x}^\star = \mathbf{x}_0 + \frac{y^\star - \hat{\beta}^\top\mathbf{x}_0}{\|\hat{\beta}\|^2}\,\hat{\beta}, \tag{57}$$

the minimum-norm change of $\mathbf{x}_0$ that attains $f(\mathbf{x}^\star) = y^\star$. We therefore write accentuated inputs as $\mathbf{x}^\star = \mathbf{x}_0 + \alpha\hat{\beta}$, where the distribution of the step α encodes the experimenter's choice of target levels. Two features of this idealization are used repeatedly: the path direction is the *estimated* weight, not the true one; and for a null seed, $\mathbf{x}_0 = 0$, the path is the one-dimensional ray $\alpha\hat{\beta}$.

E.2 Metrics on an arbitrary evaluation distribution

Prediction and control differ in *where* the same predictor is evaluated, and some of the evaluation distributions are generated by the predictor itself. It is therefore convenient to have the metrics for an arbitrary evaluation distribution p with mean μ_p and covariance Σ_p, against the noiseless teacher response $y^\star = \beta^{\star\top}\mathbf{x}$.

Lemma 23 (Mean squared error)**.** $\mathrm{MSE}[p] = \mathbb{E}_{\mathbf{x}\sim p}\big[(\hat{y} - y^\star)^2\big] = (\hat{\beta} - \beta^\star)^\top(\mu_p\mu_p^\top + \Sigma_p)(\hat{\beta} - \beta^\star)$.

Lemma 24 (Direct R^2)**.** $R^2[p] = 1 - \dfrac{\mathbb{E}[(\hat{y} - y^\star)^2]}{\mathrm{Var}[y^\star]} = 1 - \dfrac{(\hat{\beta} - \beta^\star)^\top(\mu_p\mu_p^\top + \Sigma_p)(\hat{\beta} - \beta^\star)}{\beta^{\star\top}\Sigma_p\beta^\star} = 1 - \dfrac{\mathrm{MSE}[p]}{\beta^{\star\top}\Sigma_p\beta^\star}$.

Lemma 25 (Pearson correlation)**.** $r[p] = \dfrac{\mathrm{Cov}[\hat{y}, y^\star]}{\sqrt{\mathrm{Var}[\hat{y}]\mathrm{Var}[y^\star]}} = \dfrac{\hat{\beta}^\top\Sigma_p\beta^\star}{\sqrt{(\hat{\beta}^\top\Sigma_p\hat{\beta})(\beta^{\star\top}\Sigma_p\beta^\star)}}$.

Lemma 26 (Slope of true on predicted)**.** *The slope of the least-squares line (with intercept) of $y^\star$ on $\hat{y}$ is* $\mathrm{slope}_{y^\star\leftarrow\hat{y}}[p] = \dfrac{\mathrm{Cov}[\hat{y}, y^\star]}{\mathrm{Var}[\hat{y}]} = \dfrac{\hat{\beta}^\top\Sigma_p\beta^\star}{\hat{\beta}^\top\Sigma_p\hat{\beta}}$.

Proof. All four follow from $\hat{y} - y^\star = (\hat{\beta} - \beta^\star)^\top\mathbf{x}$, from $\mathbb{E}_p[\mathbf{x}\mathbf{x}^\top] = \mu_p\mu_p^\top + \Sigma_p$, and from $\mathrm{Cov}[a^\top\mathbf{x}, b^\top\mathbf{x}] = a^\top\Sigma_p b$ for fixed vectors a, b. □

Remark 3 (Which metrics see the center, and the rank-one degeneracy). MSE and direct R^2 depend on the center μ_p of the evaluation distribution; Pearson r and the slope do not. When Σ_p has rank one, $\Sigma_p = \mathbf{v}\mathbf{v}^\top$, the Pearson correlation equals ± 1 whenever $\hat{\beta}^\top\mathbf{v} \neq 0$ and $\beta^{\star\top}\mathbf{v} \neq 0$, and is undefined otherwise. On a single noiseless accentuation path the correlation therefore carries no calibration information; it becomes informative only with observation noise or with several seed images.

	μ_p	Σ_p	MSE	R^2	Slope
General	μ_p	Σ_p	$\Delta^\top(\mu_p\mu_p^\top+\Sigma_p)\Delta$	$1-\frac{\Delta^\top(\mu_p\mu_p^\top+\Sigma_p)\Delta}{\beta^{\star\top}\Sigma_p\beta^\star}$	$\frac{\hat\beta^\top\Sigma_p\beta^\star}{\hat\beta^\top\Sigma_p\hat\beta}$
Generalization	0	Σ	$\Delta^\top\Sigma\Delta$	$1-\frac{\Delta^\top\Sigma\Delta}{S}$	$\frac{\hat\beta^\top\Sigma\beta^\star}{\hat\beta^\top\Sigma\hat\beta}$
Accentuation	0	$\mathrm{Var}[\alpha]\hat\beta\hat\beta^\top$	$\mathrm{Var}[\alpha]\big(\hat\beta^\top\Delta\big)^2$	$1-\Big(\frac{\hat\beta^\top\hat\beta}{\hat\beta^\top\beta^\star}-1\Big)^2$	$\frac{\hat\beta^\top\beta^\star}{\hat\beta^\top\hat\beta}$
Peer review	0	$\mathrm{Var}[\alpha']\beta'\beta'^\top$	$\mathrm{Var}[\alpha']\big(\beta'^\top\Delta\big)^2$	$1-\Big(\frac{\beta'^\top\hat\beta}{\beta'^\top\beta^\star}-1\Big)^2$	$\frac{\beta'^\top\beta^\star}{\beta'^\top\hat\beta}$

Table 3: **Evaluation metrics for generalization, accentuation and peer review** with null seed $\mathbf{x}_0=0$ and $\Delta:=\hat\beta-\beta^\star$. The Pearson correlation is omitted: on the rank-one accentuation and peer paths it equals ±1 (undefined when $\hat\beta^\top\beta^\star=0$, respectively $\hat\beta^\top\beta'=0$ or $\beta'^\top\beta^\star=0$).

E.3 Generalization, accentuation and peer review

Three evaluation distributions are of interest. The key difference is who chooses the inputs: nature, the model itself, or another model.

1. *Generalization*: fresh samples from the natural distribution, $\mu_p=0$, $\Sigma_p=\Sigma$, giving the classical population risk $E_{gen}=(\hat\beta-\beta^\star)^\top\Sigma(\hat\beta-\beta^\star)$.
2. *Self-accentuation*: inputs $\mathbf{x}_0+\alpha\hat\beta$ with $\mathbb{E}[\alpha]=0$, so $\mu_p=\mathbf{x}_0$ and $\Sigma_p=\mathrm{Var}[\alpha]\,\hat\beta\hat\beta^\top$, giving $E_{acc}=(\hat\beta-\beta^\star)^\top(\mathbf{x}_0\mathbf{x}_0^\top+\mathrm{Var}[\alpha]\hat\beta\hat\beta^\top)(\hat\beta-\beta^\star)$.
3. *Peer review*: inputs $\mathbf{x}_0+\alpha'\beta'$ generated by another model β', giving $E_{peer}=(\hat\beta-\beta^\star)^\top(\mathbf{x}_0\mathbf{x}_0^\top+\mathrm{Var}[\alpha']\beta'\beta'^\top)(\hat\beta-\beta^\star)$.

Null seed. In the main text the seed is null, $\mathbf{x}_0=0$, corresponding to a gray initialization. The three errors then read

$$E_{gen}=(\hat\beta-\beta^\star)^\top\Sigma(\hat\beta-\beta^\star),\qquad E_{acc}=\mathrm{Var}[\alpha]\big(\hat\beta^\top(\hat\beta-\beta^\star)\big)^2,\qquad E_{peer}=\mathrm{Var}[\alpha']\big(\beta'^\top(\hat\beta-\beta^\star)\big)^2. \tag{58}$$

For a natural seed $\mathbf{x}_0\neq0$ the MSE and R^2 acquire the additional rank-one term $(\mathbf{x}_0^\top(\hat\beta-\beta^\star))^2$, which is the prediction error at the seed itself, while the slope and Pearson r are unchanged (Remark 3); averaging over natural seeds $\mathbf{x}_0\sim\mathcal{N}(0,\Sigma)$ adds E_{gen} to the path error. Multiple seeds also make the Pearson correlation informative. We do not pursue these extensions here.

Variance-matched control range. The accentuation MSE depends on how far the target is driven, through $\mathrm{Var}[\alpha]$, whereas R^2, slope and r do not. To make errors comparable across models we fix the target levels so that the *predicted* response variance along the path equals the natural response variance of the teacher:

$$\mathrm{Var}_{\mathbf{x}=\mathbf{x}_0+\alpha\hat\beta}[\hat\beta^\top\mathbf{x}]=\mathrm{Var}_{\mathbf{x}\sim\mathcal{N}(0,\Sigma)}[\beta^{\star\top}\mathbf{x}]\quad\Longleftrightarrow\quad\mathrm{Var}[\alpha]\,(\hat\beta^\top\hat\beta)^2=S:=\beta^{\star\top}\Sigma\beta^\star, \tag{59}$$

and analogously $\mathrm{Var}[\alpha'](\beta'^\top\beta')^2=S$ for the peer.

Corollary 27 (Accentuation and peer-review metrics; restatement of Corollary 6)**.** *With the null seed and variance matching, and $N:=\hat\beta^\top\beta^\star$, $D:=\hat\beta^\top\hat\beta$, the metrics of Table 3 reduce to*

$$\frac{E_{acc}}{S}=\Big(1-\frac{N}{D}\Big)^2,\qquad R^2_{acc}=1-\Big(\frac{D}{N}-1\Big)^2,\qquad \mathrm{slope}_{acc}=\frac{N}{D}, \tag{60}$$

and, for a peer path,

$$\frac{E_{peer}}{S}=\Big(\frac{\beta'^\top\hat\beta}{\beta'^\top\beta'}-\frac{\beta'^\top\beta^\star}{\beta'^\top\beta'}\Big)^2,\qquad R^2_{peer}=1-\Big(1-\frac{\beta'^\top\hat\beta}{\beta'^\top\beta^\star}\Big)^2,\qquad \mathrm{slope}_{peer}=\frac{\beta'^\top\beta^\star}{\beta'^\top\hat\beta}. \tag{61}$$

In particular every accentuation metric is a function of the single scalar $z_{acc} := D/N - 1$*:*

$$R^2_{acc} = 1 - z^2_{acc}, \qquad \text{slope}_{acc} = \frac{1}{1 + z_{acc}}, \qquad \frac{E_{acc}}{S} = \Big(\frac{z_{acc}}{1 + z_{acc}}\Big)^2. \tag{62}$$

Proof. Accentuation. Substituting Eq. 59 into Eq. 58,

$$E_{acc} = \frac{S}{(\hat\beta^\top \hat\beta)^2}\big(\hat\beta^\top(\hat\beta - \beta^\star)\big)^2 = S\Big(1 - \frac{\hat\beta^\top \beta^\star}{\hat\beta^\top \hat\beta}\Big)^2.$$

The true response variance along the path is $\mathrm{Var}[\beta^{\star\top}(\alpha\hat\beta)] = (\beta^{\star\top}\hat\beta)^2\mathrm{Var}[\alpha]$, so

$$R^2_{acc} = 1 - \frac{\mathrm{Var}[\alpha](\hat\beta^\top(\hat\beta - \beta^\star))^2}{\mathrm{Var}[\alpha](\beta^{\star\top}\hat\beta)^2} = 1 - \Big(\frac{\hat\beta^\top\hat\beta}{\beta^{\star\top}\hat\beta} - 1\Big)^2,$$

and the slope of $\beta^{\star\top}(\alpha\hat\beta)$ on $\hat\beta^\top(\alpha\hat\beta)$ is $\beta^{\star\top}\hat\beta/\hat\beta^\top\hat\beta$ because both are proportional to α. *Peer review.* The same three steps with $\hat\beta$ replaced by β' in the path and $\mathrm{Var}[\alpha'] = S/(\beta'^\top\beta')^2$ give Eq. 61. *Reduction to* z. Write $N/D = 1/(1+z)$ and $D/N = 1 + z$ in Eq. 60. □

Remark 4 (The two singularities of accentuation). The normalized error E_{acc}/S diverges only as $D \to 0$, i.e. when the student has no weight at all; the score R^2_{acc} instead becomes very negative when the *true* gain $N \to 0$ while D stays finite. Very negative accentuation R^2 therefore need not mean unbounded raw error: as $N \to 0$, $R^2_{acc} \to -\infty$ while $E_{acc}/S \to 1$. This is why R^2 is the more sensitive diagnostic of control failure.

E.4 Geometry of equal-error sets

The three evaluations induce different geometries in the space of student weights. Fixing the level of each metric and asking which $\hat\beta$ attain it gives the following picture.

Proposition 28 (Restatement of Proposition 7)**.** *Let* $\Sigma \succ 0$ *and fix* $\beta^\star$*.*

1. *The equal-*E_{gen} *sets are the ellipsoids*
$$(\hat\beta - \beta^\star)^\top \Sigma(\hat\beta - \beta^\star) = c, \tag{63}$$
centered at $\beta^\star$ *with semiaxis* $\sqrt{c/s_k}$ *along the population eigenvector* $\mathbf{u}_k$*.*

2. *The equal-*z_{acc} *sets are the spheres*
$$\Big\|\hat\beta - \frac{a}{2}\beta^\star\Big\|^2 = \frac{a^2}{4}\|\beta^\star\|^2, \qquad a := 1 + z_{acc}, \tag{64}$$
whose diameter joins 0 *and* $a\beta^\star$*; in particular perfect calibration* $z_{acc} = 0$ *is the sphere with diameter* $[0, \beta^\star]$*, not the point* $\beta^\star$*.*

3. *For a fixed peer* β'*, the equal-peer-error sets are the hyperplanes*
$$\beta'^\top(\hat\beta - a_{peer}\beta^\star) = 0, \qquad a_{peer} := \frac{\beta'^\top\hat\beta}{\beta'^\top\beta^\star}, \tag{65}$$
normal to β'*.*

Proof. (i) is the definition of E_{gen} in the eigenbasis of Σ. (ii) For a finite level z with $\hat\beta^\top\beta^\star \neq 0$, the condition $D/N = a$ reads $\hat\beta^\top\hat\beta = a\,\hat\beta^\top\beta^\star$; completing the square gives Eq. 64. Since R^2_{acc}, slope_{acc} and E_{acc}/S are functions of z_{acc} alone (Eq. 62), their level sets are unions of such spheres. (iii) Each peer metric in Eq. 61 is a function of $\beta'^\top\hat\beta/\beta'^\top\beta^\star$, which is constant on affine hyperplanes normal to β'. □

Remark 5 (Why prediction hides what control exposes). The ellipsoid tolerates Euclidean weight errors of size $\sqrt{c/s_k}$ along the kth eigendirection, so errors along low-variance directions are almost free for prediction. The sphere and the hyperplane are Euclidean objects and weight all directions equally. Consequently a student may sit deep inside a small prediction ellipsoid, far from the calibration sphere, while lying exactly on the peer hyperplane of a good generator. No scalar prediction score determines the other two geometries.

Example 29 (Proof of Example 8)**.** With $\Sigma = \mathrm{diag}(1, \varepsilon)$, $\beta^\star = \hat\beta_1 = (1, 0)^\top$ and $\hat\beta_2 = (1, a_\varepsilon)^\top$, $a_\varepsilon = \varepsilon^{-q}$, $0 < q < \frac{1}{2}$: $E_{gen}(\hat\beta_2) = \varepsilon a_\varepsilon^2 = \varepsilon^{1-2q} \to 0$ and $R^2_{gen}(\hat\beta_2) \to 1$. On its own path $N_2 = 1$ and $D_2 = 1 + a_\varepsilon^2$, so $\mathrm{slope}_{acc}(\hat\beta_2) = (1 + a_\varepsilon^2)^{-1} \to 0$ and $R^2_{acc}(\hat\beta_2) = 1 - a_\varepsilon^4 \to -\infty$. For peer review, $\hat\beta_1^\top \hat\beta_2 / \hat\beta_1^\top \beta^\star = 1$ and $\hat\beta_2^\top \hat\beta_1 / \hat\beta_2^\top \beta^\star = 1$, so both peer scores equal 1.

F PIXEL-SPACE RIDGE: PROOFS

The geometry of the previous section says where a student must *not* be for control to succeed. This section computes where ridge regression on noisy data actually puts it. We work in the proportional limit $n, d \to \infty$, $d/n \to \gamma$, with Gaussian design, and use the deterministic equivalents (DEs) collected in Section L: the effective penalty $\kappa(\lambda)$ of Eq. 108, the one-point equivalent Eq. 110, the two-point equivalent Eq. 112, and the derivative $d\kappa/d\lambda = n/(n - \mathrm{df}_2)$ of Eq. 117. Throughout, $\Sigma = \sum_k s_k \mathbf{u}_k \mathbf{u}_k^\top$, $\beta^\star = \sum_k b_k \mathbf{u}_k$, and $B_{p,q}$, $\mathrm{df}_{p,q}$ are the spectral sums of Eq. 33.

Proposition 30 (Restatement of Propositions 9 and 10)**.** *For pixel ridge under Gaussian design,*

$$E_{gen} \asymp E_{gen,DE} := \frac{n(\kappa^2 B_{1,2} + \sigma^2)}{n - \mathrm{df}_2} - \sigma^2, \tag{66}$$

$$\mu_N := \mathbb{E}[N] \asymp B_{1,1}, \tag{67}$$

$$\mu_D := \mathbb{E}[D] \asymp B_{2,2} + \underbrace{\frac{\mathrm{df}_{1,2}}{n}\big(E_{gen,DE} + \sigma^2\big)}_{V_{\mathrm{pix}}(\kappa)}$$

$$= \|\bar\beta_\kappa\|^2 + V_{\mathrm{pix}}(\kappa), \qquad \bar\beta_\kappa := \Sigma(\Sigma + \kappa I)^{-1}\beta^\star, \tag{68}$$

$$\mathbb{E}\big[(\mathbf{u}_k^\top(\hat\beta - \beta^\star))^2\big] \asymp \frac{\kappa^2}{(s_k + \kappa)^2} b_k^2 + \frac{s_k}{(s_k + \kappa)^2} \frac{\kappa^2 B_{1,2} + \sigma^2}{n - \mathrm{df}_2}. \tag{69}$$

F.1 GENERALIZATION RISK

Proof of Eq. 66. Step 1: conditional expectation over the noise. Insert the error decomposition Eq. 56 into $E_{gen} = (\hat\beta - \beta^\star)^\top \Sigma (\hat\beta - \beta^\star)$. The cross term is linear in η and averages to zero; the quadratic term uses $\mathbb{E}_\eta[\eta\eta^\top] = I_n$:

$$\mathbb{E}_\eta[E_{gen}] = \lambda^2 \beta^{\star\top}(\hat\Sigma + \lambda I)^{-1}\Sigma(\hat\Sigma + \lambda I)^{-1}\beta^\star + \frac{\sigma^2}{n}\,\mathrm{Tr}\left[(\hat\Sigma + \lambda I)^{-1}\hat\Sigma(\hat\Sigma + \lambda I)^{-1}\Sigma\right]. \tag{70}$$

Step 2: noise term. Writing the trace as a λ-derivative and using the one-point DE,

$$\mathrm{Tr}\left[\Sigma(\hat\Sigma + \lambda I)^{-2}\hat\Sigma\right] = -\frac{d}{d\lambda}\,\mathrm{Tr}\left[\Sigma(\hat\Sigma + \lambda I)^{-1}\hat\Sigma\right] \asymp -\frac{d}{d\lambda}\,\mathrm{Tr}\left[\Sigma(\Sigma + \kappa(\lambda) I)^{-1}\Sigma\right] = \frac{d\kappa}{d\lambda}\,\mathrm{df}_2(\kappa),$$

so that with Eq. 117 the noise term equals $\sigma^2 \mathrm{df}_2/(n - \mathrm{df}_2)$.

Step 3: bias term. Apply the two-point DE Eq. 112 with $A = \beta^\star\beta^{\star\top}$ and $B = \Sigma$:

$$\lambda^2\,\mathrm{Tr}\left[\beta^\star\beta^{\star\top}(\hat\Sigma + \lambda I)^{-1}\Sigma(\hat\Sigma + \lambda I)^{-1}\right] \asymp \kappa^2 \beta^{\star\top}(\Sigma + \kappa I)^{-1}\Sigma(\Sigma + \kappa I)^{-1}\beta^\star$$
$$+ \kappa^2 \frac{\beta^{\star\top}(\Sigma + \kappa I)^{-2}\Sigma\beta^\star\ \ \mathrm{Tr}[\Sigma(\Sigma + \kappa I)^{-2}\Sigma]}{n - \mathrm{df}_2}$$
$$= \frac{n}{n - \mathrm{df}_2}\,\kappa^2 B_{1,2}(\kappa).$$

Step 4: assemble. Adding the two terms,

$$\mathbb{E}_\eta[E_{gen}] \asymp \frac{n}{n - \mathrm{df}_2}\kappa^2 B_{1,2} + \sigma^2 \frac{\mathrm{df}_2}{n - \mathrm{df}_2} = \frac{n(\kappa^2 B_{1,2} + \sigma^2)}{n - \mathrm{df}_2} - \sigma^2, \tag{71}$$

which is Eq. 66 and agrees with the classical anisotropic ridge risk (Dobriban & Wager, 2018; Hastie et al., 2022). □

F.2 The accentuation forms N and D

Proof of Eqs. 67 and 68. Numerator. From Eq. 56, $N = \beta^{\star\top}(\hat{\Sigma}+\lambda I)^{-1}\hat{\Sigma}\beta^\star + \sigma\beta^{\star\top}(\hat{\Sigma}+\lambda I)^{-1}\frac{1}{n}X^\top\eta$. The noise part has zero mean, and the one-point DE gives

$$\mu_N = \mathbb{E}_{X,\eta}[N] \asymp \beta^{\star\top}(\Sigma+\kappa I)^{-1}\Sigma\beta^\star = B_{1,1}(\kappa). \tag{72}$$

Denominator. Expanding $D = \hat{\beta}^\top\hat{\beta}$ and averaging over η kills the cross term:

$$\mathbb{E}_\eta[D] = \beta^{\star\top}\hat{\Sigma}^2(\hat{\Sigma}+\lambda I)^{-2}\beta^\star + \frac{\sigma^2}{n}\operatorname{Tr}\left[(\hat{\Sigma}+\lambda I)^{-2}\hat{\Sigma}\right].$$

The noise trace is handled as in Step 2 above, $\operatorname{Tr}[(\hat{\Sigma}+\lambda I)^{-2}\hat{\Sigma}] \asymp \frac{d\kappa}{d\lambda}\operatorname{Tr}[\Sigma(\Sigma+\kappa I)^{-2}] = \frac{n}{n-\mathrm{df}_2}\mathrm{df}_{1,2}$. The signal term is the two-point DE with $A = \beta^\star\beta^{\star\top}$, $B = I$:

$$\beta^{\star\top}\hat{\Sigma}^2(\hat{\Sigma}+\lambda I)^{-2}\beta^\star \asymp \beta^{\star\top}(\Sigma+\kappa I)^{-2}\Sigma^2\beta^\star + \frac{\kappa^2\,\mathrm{df}_{1,2}}{n-\mathrm{df}_2}\,\beta^{\star\top}(\Sigma+\kappa I)^{-2}\Sigma\beta^\star.$$

Collecting,

$$\mu_D \asymp B_{2,2} + \frac{\mathrm{df}_{1,2}}{n-\mathrm{df}_2}\left(\sigma^2+\kappa^2 B_{1,2}\right) = B_{2,2} + \frac{\mathrm{df}_{1,2}}{n}\left(E_{gen,DE}+\sigma^2\right), \tag{73}$$

where the last equality uses Eq. 66. □

Proof of Eq. 69. Project Eq. 56 onto $\mathbf{u}_k$ and average over η:

$$\mathbb{E}_\eta\left[(\mathbf{u}_k^\top(\hat{\beta}-\beta^\star))^2\right] = \lambda^2\operatorname{Tr}\left[\mathbf{u}_k\mathbf{u}_k^\top(\hat{\Sigma}+\lambda I)^{-1}\beta^\star\beta^{\star\top}(\hat{\Sigma}+\lambda I)^{-1}\right] + \frac{\sigma^2}{n}\operatorname{Tr}\left[\mathbf{u}_k\mathbf{u}_k^\top(\hat{\Sigma}+\lambda I)^{-1}\hat{\Sigma}(\hat{\Sigma}+\lambda I)^{-1}\right].$$

The two-point DE with $A = \mathbf{u}_k\mathbf{u}_k^\top$, $B = \beta^\star\beta^{\star\top}$ gives for the first term $\kappa^2\left[b_k^2/(s_k+\kappa)^2 + \frac{s_k}{(s_k+\kappa)^2}\frac{B_{1,2}}{n-\mathrm{df}_2}\right]$, and the derivative argument of Step 2 gives for the second $\frac{\sigma^2}{n}\frac{n}{n-\mathrm{df}_2}\frac{s_k}{(s_k+\kappa)^2}$. Summing yields Eq. 69. Weighting by s_k and summing over k recovers Eq. 66; summing without weights and adding the squared mean $\sum_k(s_k b_k/(s_k+\kappa))^2 = B_{2,2}$ recovers Eq. 68. □

Remark 6 (Reading the three quantities).

1. $\mu_N = \sum_k \frac{s_k}{s_k+\kappa}b_k^2 \geq 0$ is the teacher power in the modes that ridge resolves, i.e. those with $s_k \gtrsim \kappa$; it vanishes only if $\Sigma\beta^\star = 0$.
2. μ_D is the squared norm of the shrunken mean weight, $B_{2,2}$, plus the *Euclidean* energy of estimation noise: $(E_{gen,DE}+\sigma^2)/n$ per effective sample, multiplied by $\mathrm{df}_{1,2} = \sum_k s_k/(s_k+\kappa)^2$, a count of modes in a bell-shaped band around $s_k \approx \kappa$ (the weight $4\kappa s_k/(s_k+\kappa)^2$ peaks at 1 for $s_k = \kappa$ with constant width on a logarithmic axis). Prediction weights this same noise by s_k; control does not.
3. The per-mode error Eq. 69 is largest, relative to the mode's own variance s_k, for low-variance modes: the variance term divided by s_k is $(s_k+\kappa)^{-2}$ times a mode-independent factor.

F.3 The leading control error z_{acc}

Substituting $N \approx \mu_N$, $D \approx \mu_D$ into Eq. 62 gives the leading (ratio-of-DE) accentuation metrics. The central quantity admits a compact form.

Lemma 31 (Leading z_{acc} in closed form)**.**

$$z_{acc}^{(0)} := \frac{\mu_D}{\mu_N} - 1 = \frac{\sigma^2\,\mathrm{df}_{1,2}(\kappa) - n\lambda\,B_{1,2}(\kappa)}{(n-\mathrm{df}_2(\kappa))\,B_{1,1}(\kappa)} = \frac{B_{2,2}-B_{1,1}}{B_{1,1}} + \frac{\mathrm{df}_{1,2}(E_{gen,DE}+\sigma^2)}{n\,B_{1,1}}. \tag{74}$$

Proof. From Eqs. 72 and 73,

$$\mu_D - \mu_N = B_{2,2} - B_{1,1} + \frac{\mathrm{df}_{1,2}}{n - \mathrm{df}_2}(\sigma^2 + \kappa^2 B_{1,2}) = -\kappa B_{1,2} + \frac{\mathrm{df}_{1,2}}{n - \mathrm{df}_2}(\sigma^2 + \kappa^2 B_{1,2}),$$

using the recursion $B_{1,1} - B_{2,2} = \kappa B_{1,2}$ of Eq. 120. The coefficient of $B_{1,2}$ is

$$\frac{\kappa^2 \mathrm{df}_{1,2}}{n - \mathrm{df}_2} - \kappa = \frac{\kappa\,(\kappa \mathrm{df}_{1,2} + \mathrm{df}_2 - n)}{n - \mathrm{df}_2} = \frac{\kappa\,(\mathrm{df}_1 - n)}{n - \mathrm{df}_2} = -\frac{n\lambda}{n - \mathrm{df}_2},$$

by $\mathrm{df}_1 = \mathrm{df}_2 + \kappa \mathrm{df}_{1,2}$ and the fixed-point relation $n\lambda = \kappa(n - \mathrm{df}_1)$. Dividing by $\mu_N = B_{1,1}$ gives the first form; the second is Eq. 68 divided by $B_{1,1}$. □

Remark 7 (Two competing terms, and when R^2_{acc} turns negative). The two forms of Eq. 74 expose the competition. The ridge term $-n\lambda B_{1,2}$ (equivalently $B_{2,2} - B_{1,1} < 0$) is shrinkage: the student under-shoots along its own path, $z < 0$, slope > 1. The noise term $\sigma^2 \mathrm{df}_{1,2}$ is Euclidean noise energy: the student over-shoots, $z > 0$, slope < 1. Since $\mu_N, \mu_D > 0$, always $z > -1$, and $R^2_{acc} = 1 - z^2$ crosses zero at $z = 1$. Writing the condition $z \geq 1$ as

$$\sigma^2\,\mathrm{df}_{1,2} \;\geq\; n\lambda\,B_{1,2} + (n - \mathrm{df}_2)\,B_{1,1},$$

one sees that at fixed λ (hence fixed κ) the right-hand side is bounded, $B_{1,1} \leq \|\beta^\star\|^2$, while the left-hand side grows without bound in σ^2: for any fixed regularization, enough response noise drives R^2_{acc} to any negative value. Whether tuning λ with the noise can prevent this is the subject of the next section.

G Cross-validation versus control-optimal regularization: proofs

Along the ridge path the fitted weight moves from the least-squares or minimum-norm solution ($\lambda \to 0$) to the origin ($\lambda \to \infty$). Cross-validation selects the point of the path that is tangent to the smallest reachable prediction ellipsoid; control calibration asks where the path crosses the sphere $D = N$. This section shows that the two selections obey different spectral conditions and computes the control error left over by prediction-optimal tuning. Since $d\kappa/d\lambda > 0$ (Eq. 117), selecting λ is equivalent to selecting κ, and all conditions are written in κ.

Proposition 32 (Restatement of Proposition 11)**.** *1. If $E_{gen,DE}(\kappa)$ has an interior differentiable minimum at $\kappa^\star_{gen}$ on the admissible path $\lambda \geq 0$, then there*

$$E_{gen,DE} + \sigma^2 = n\kappa \frac{B_{2,3}}{\mathrm{df}_{2,3}}. \tag{75}$$

2. At a point with $\mu_N > 0$, the leading control error vanishes, $z^{(0)}_{acc} = 0$, if and only if

$$E_{gen,DE} + \sigma^2 = n\kappa \frac{B_{1,2}}{\mathrm{df}_{1,2}} \quad \Longleftrightarrow \quad \sigma^2 = n\lambda \frac{B_{1,2}}{\mathrm{df}_{1,2}}, \tag{76}$$

and at such a root $E_{gen,DE} = \sigma^2\,\mathrm{df}_1/(n - \mathrm{df}_1)$.

3. At an interior prediction optimum,

$$z^{(0)}_{acc,CV} = \frac{\kappa\,\mathrm{df}_{1,2}}{B_{1,1}} \left(\frac{B_{2,3}}{\mathrm{df}_{2,3}} - \frac{B_{1,2}}{\mathrm{df}_{1,2}} \right)_{\kappa = \kappa^\star_{gen}}. \tag{77}$$

G.1 Prediction-optimal stationarity

Proof of Eq. 75. Write the noisy prediction risk of Eq. 66 as

$$E_{gen,DE} + \sigma^2 = n\,\frac{H(\kappa)}{q(\kappa)}, \qquad H(\kappa) := \kappa^2 B_{1,2}(\kappa) + \sigma^2, \qquad q(\kappa) := n - \mathrm{df}_2(\kappa).$$

By the derivative identities of Eq. 121,

$$H'(\kappa) = \frac{d}{d\kappa}\sum_k \frac{\kappa^2 s_k}{(s_k+\kappa)^2} b_k^2 = 2\sum_k \frac{\kappa s_k^2}{(s_k+\kappa)^3} b_k^2 = 2\kappa B_{2,3}, \qquad q'(\kappa) = -\frac{d}{d\kappa}\mathrm{df}_2 = 2\,\mathrm{df}_{2,3}.$$

An interior stationary point of H/q satisfies $qH' = Hq'$, i.e.

$$(n - \mathrm{df}_2)\,\kappa\, B_{2,3} = \mathrm{df}_{2,3}\big(\kappa^2 B_{1,2} + \sigma^2\big), \tag{78}$$

which is Eq. 75 after multiplying by $n/(q\,\mathrm{df}_{2,3})$. The condition is necessary; a stationary point need not be the global minimum, and the minimum may sit at the boundary $\lambda = 0$ (see Remark 8). □

G.2 Control-optimal condition and the risk at the control optimum

Proof of Eq. 76. From Lemma 31,

$$\mu_D - \mu_N = -\kappa B_{1,2} + \frac{\mathrm{df}_{1,2}}{n}\big(E_{gen,DE} + \sigma^2\big). \tag{79}$$

With $\mu_N > 0$, $z_{acc}^{(0)} = 0$ iff this vanishes, which is the first form of Eq. 76. Substituting $E_{gen,DE} + \sigma^2 = nH/q$ and clearing denominators,

$$\begin{aligned} & \mathrm{df}_{1,2}(\kappa^2 B_{1,2} + \sigma^2) = (n - \mathrm{df}_2)\kappa B_{1,2} \\ \iff \quad & \sigma^2 \mathrm{df}_{1,2} = \kappa\big(n - \mathrm{df}_2 - \kappa \mathrm{df}_{1,2}\big) B_{1,2} = \kappa(n - \mathrm{df}_1) B_{1,2} = n\lambda B_{1,2}, \end{aligned}$$

using $\mathrm{df}_1 = \mathrm{df}_2 + \kappa \mathrm{df}_{1,2}$ and the fixed point $n\lambda = \kappa(n - \mathrm{df}_1)$. At a root with $\lambda, \sigma^2 > 0$ the two forms divide to $(E_{gen,DE} + \sigma^2)/\sigma^2 = \kappa/\lambda = n/(n - \mathrm{df}_1)$, hence

$$E_{gen,DE}(\kappa^\star_{acc}) = \sigma^2 \frac{\mathrm{df}_1(\kappa^\star_{acc})}{n - \mathrm{df}_1(\kappa^\star_{acc})}. \tag{80}$$

The teacher enters only through the location of the root, so this is not a teacher-independent value of the risk. □

G.3 Residual control error at the prediction optimum

Proof of Eq. 77. Insert Eq. 75 into Eq. 79:

$$\mu_D - \mu_N = -\kappa B_{1,2} + \kappa\,\mathrm{df}_{1,2}\frac{B_{2,3}}{\mathrm{df}_{2,3}} = \kappa\,\mathrm{df}_{1,2}\Big(\frac{B_{2,3}}{\mathrm{df}_{2,3}} - \frac{B_{1,2}}{\mathrm{df}_{1,2}}\Big),$$

and divide by $\mu_N = B_{1,1}$. □

Remark 8 (Reading the residual).

1. *Two spectral averages.* Define the probability weights over modes $w_k^{(1)} := s_k^2(s_k+\kappa)^{-3}/\mathrm{df}_{2,3}$ and $w_k^{(2)} := s_k(s_k+\kappa)^{-2}/\mathrm{df}_{1,2}$. Then $B_{2,3}/\mathrm{df}_{2,3} - B_{1,2}/\mathrm{df}_{1,2} = \sum_k (w_k^{(1)} - w_k^{(2)}) b_k^2$, and $w_k^{(1)}/w_k^{(2)} \propto s_k/(s_k+\kappa)$ is increasing in s_k. On a logarithmic axis in s, $w^{(2)}$ is a symmetric bump centered at $s = \kappa$ (half-maximum between 0.17κ and 5.8κ), while $w^{(1)}$ peaks at $s = 2\kappa$ with its low-variance side suppressed by the extra factor (Figure 26); the two kernels are the κ-derivatives of the linear and of the squared shrinkage factor, $s/(s+\kappa)^2 = -\partial_\kappa[s/(s+\kappa)]$ and $s^2/(s+\kappa)^3 = -\frac{1}{2}\partial_\kappa[s/(s+\kappa)]^2$. The residual is therefore positive (over-shoot, slope below one) when the teacher's power b_k^2 sits preferentially on higher-variance modes than the noise-counting weight $w^{(2)}$ does, which is the typical situation for natural signals, and negative in the opposite case.
2. *Isotropic data.* If $s_k \equiv s$, both averages equal $\|\beta^\star\|^2/d$ and the residual vanishes: prediction-optimal tuning calibrates control. The dissociation is a consequence of anisotropy.
3. *Magnitude.* The prefactor $\kappa\,\mathrm{df}_{1,2}$ counts the modes in the band $s_k \approx \kappa$ (Section F); for a power-law spectrum this count grows as κ moves into the bulk, so the residual is amplified precisely in the noisy regime where cross-validation selects a larger penalty.

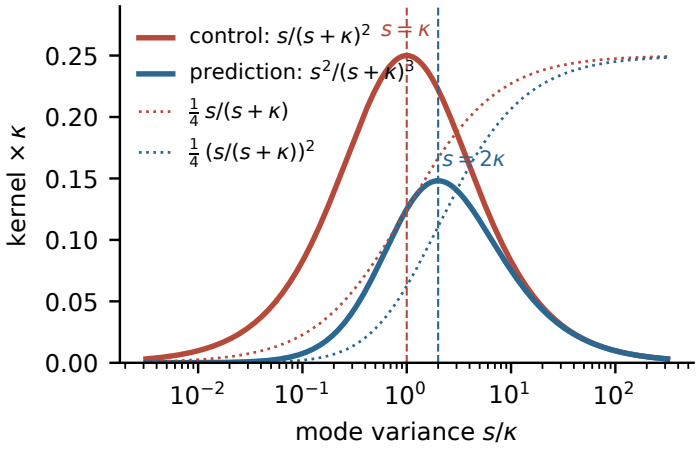

Figure 26: **The two spectral kernels of Proposition 32.** Solid: the control kernel $s/(s+\kappa)^2$ (peak at $s=\kappa$, symmetric on the log axis) and the prediction kernel $s^2/(s+\kappa)^3$ (peak at $s=2\kappa$), both scaled by κ. Dotted: the linear and squared shrinkage factors whose κ-derivatives they are. The residual control error at the prediction optimum, Eq. 77, is the difference between the teacher power averaged over the blue band and over the red band.

4. *Boundary optima.* At low noise the DE prediction risk can be minimized at the boundary $\lambda \to 0$ of the admissible path; there the stationarity formula does not apply and the ridgeless identity $N = D$ makes the leading control error vanish. In the van Hateren experiment the interior optimum appears above a noise threshold, and Eq. 77 was checked against the leading DE evaluated at the numerically selected optimum and against empirical RidgeCV.

5. *Existence of a control root.* Eq. 76 need not have an admissible solution for every $(\Sigma, \beta^\star, \sigma^2, n)$; it is a condition, not a guarantee. Where a root exists, the leading metrics are perfectly calibrated ($z_{acc}^{(0)} = 0$), but the finite-sample expectation retains the fluctuation term $\mathrm{Var}(N/D)$, see Section J.

H Feature-space regression: proofs

Pixel regression explains why a single student fails. To explain why equally predictive backbones fail differently, we let the student read out from a fixed linear feature map and accentuate through it. The structure of the results is that prediction only sees the feature map through $\Sigma_F = F\Sigma F^\top$ and $F\Sigma\beta^\star$, while control also sees $G := FF^\top$ and $F\beta^\star$.

H.1 Setting and estimator

The teacher is unchanged, $y = \beta^{\star\top}\mathbf{x} + \sigma\epsilon$. The student reads out linearly from features $F\mathbf{x}$, $F \in \mathbb{R}^{p\times d}$ fixed: $f(\mathbf{x}) = w^\top F\mathbf{x}$, so its effective pixel weight is $\hat\beta = F^\top\hat w$ and must lie in the row space of F. We condition on F throughout; it is a structured map, not a random sketch. The accentuation path $\mathbf{x}_0 + \alpha\hat\beta = \mathbf{x}_0 + \alpha F^\top\hat w$ also lies in the row space of F.

Lemma 33 (Feature ridge estimator)**.** *With $Z := XF^\top \in \mathbb{R}^{n\times p}$ and $\hat\Sigma_F := Z^\top Z/n = F\hat\Sigma F^\top$,*

$$\hat w = (\hat\Sigma_F + \lambda I_p)^{-1}F\hat\Sigma\beta^\star + \frac{\sigma}{n}(\hat\Sigma_F + \lambda I_p)^{-1}FX^\top\eta, \qquad \hat\beta = F^\top\hat w. \tag{81}$$

Proof. Apply Lemma 22 to the design Z and note $Z^\top\mathbf{y} = FX^\top(X\beta^\star + \sigma\eta)$. □

H.2 Realizable decomposition

Since $p < d$ in general, the teacher is usually not representable in the feature space. Two notions of "best representable weight" arise, and the difference between them is the first sign of the dissociation.

Lemma 34 (Restatement of Lemma 14)**.** *The population prediction risk $\mathbb{E}_{\mathbf{x}}[(\beta^{\star\top}\mathbf{x} - w^\top F\mathbf{x})^2]$ is minimized by*

$$w^\star = (F\Sigma F^\top)^\dagger F\Sigma\beta^\star = \Sigma_F^\dagger F\Sigma\beta^\star, \tag{82}$$

characterized by the orthogonality condition

$$F\Sigma(\beta^\star - F^\top w^\star) = 0. \tag{83}$$

The Euclidean reconstruction error $\|\beta^\star - F^\top w\|^2$ is instead minimized by the projection $w^P = (FF^\top)^\dagger F\beta^\star$, and $F^\top w^P$ is the orthogonal projection of $\beta^\star$ onto the row space of F. The two coincide when Σ acts as a multiple of the identity on the row space of F, in particular for white data, and differ otherwise.

Proof. Expanding, the risk equals $S - 2w^\top F\Sigma\beta^\star + w^\top\Sigma_F w$, whose gradient vanishes at $\Sigma_F w = F\Sigma\beta^\star$; the minimum-norm solution is Eq. 82, and $\Sigma_F w^\star = F\Sigma\beta^\star$ is Eq. 83 since $F\Sigma\beta^\star$ lies in the range of Σ_F. The reconstruction problem is the same computation with Σ replaced by I. □

Lemma 35 (The unrealizable signal acts as noise)**.** *Let $r := X(\beta^\star - F^\top w^\star) \in \mathbb{R}^n$ be the per-sample residual of the realizable teacher. Then*

$$\hat{w} = \underbrace{(\hat{\Sigma}_F + \lambda I_p)^{-1}\hat{\Sigma}_F w^\star}_{\text{shrinkage of the realizable weight}} + \underbrace{\frac{1}{n}(\hat{\Sigma}_F + \lambda I_p)^{-1}Z^\top(r + \sigma\eta)}_{\text{noise and unrealizable residual}}, \tag{84}$$

and under Gaussian inputs the residuals r_i are independent of the features Z, mutually uncorrelated, with variance

$$S_\perp := \mathrm{Var}[r_i] = (\beta^\star - F^\top w^\star)^\top\Sigma(\beta^\star - F^\top w^\star) = S - w^{\star\top}\Sigma_F w^\star = \Sigma_{yy} - \Sigma_{Fy}^\top\Sigma_F^\dagger\Sigma_{Fy}, \tag{85}$$

where $\Sigma_{Fy} = F\Sigma\beta^\star = \mathrm{Cov}(F\mathbf{x}, \beta^{\star\top}\mathbf{x})$ and $\Sigma_{yy} = S$.

Proof. Writing $X\beta^\star = Zw^\star + r$ gives $F\hat{\Sigma}\beta^\star = \hat{\Sigma}_F w^\star + \frac{1}{n}Z^\top r$, and substituting into Eq. 81 gives Eq. 84. For a single sample, $\mathbf{z}_i = F\mathbf{x}_i$ and $r_i = \mathbf{x}_i^\top(\beta^\star - F^\top w^\star)$ are jointly Gaussian with $\mathbb{E}[\mathbf{z}_i r_i] = F\Sigma(\beta^\star - F^\top w^\star) = 0$ by Eq. 83, hence independent; different samples are independent by construction. The variance follows by expanding the quadratic form and using the orthogonality condition once more to drop the cross term. □

The decomposition Eq. 84 is the key structural fact of feature-space regression: the fit of $w^\star$ from the design Z sees the unrealizable part of the teacher exactly as an additional response noise of variance $S_\perp$. Note that $S_\perp$ is a fixed geometric quantity of $(F, \Sigma, \beta^\star)$, not a random variable and not a function of σ, λ or n.

H.3 Generalization risk and accentuation forms

Proposition 36 (Restatement of Proposition 15)**.** *Let κ solve the fixed-point equation on the spectrum of Σ_F with aspect ratio p/n, and let $B_{a,b}^F$, $\mathrm{df}_{a,b}^F$ be the spectral sums built from $(\Sigma_F, w^\star)$. Then*

$$E_{gen} \asymp E_{gen,DE} = \frac{n}{n - \mathrm{df}_2^F}\left(\kappa^2 B_{1,2}^F + S_\perp + \sigma^2\right) - \sigma^2, \tag{86}$$

$$\mu_N \asymp \beta^{\star\top}\bar{\beta}_{F,\kappa}, \tag{87}$$

$$\bar{\beta}_{F,\kappa} := F^\top(\Sigma_F + \kappa I_p)^{-1}F\Sigma\beta^\star = F^\top\Sigma_F(\Sigma_F + \kappa I_p)^{-1}w^\star = (F^\top F\Sigma + \kappa I_d)^{-1}F^\top F\Sigma\beta^\star, \tag{88}$$

$$\mu_D \asymp \|\bar{\beta}_{F,\kappa}\|^2 + \underbrace{\frac{E_{gen,DE} + \sigma^2}{n}T_F(\kappa)}_{V_F(\kappa)}, \qquad T_F(\kappa) := \mathrm{Tr}\left[G(\Sigma_F + \kappa I_p)^{-2}\Sigma_F\right]. \tag{89}$$

Proof of Eq. 86. Step 1: split off the unrealizable error. With $\Delta_w := \hat{w} - w^\star$,

$$E_{gen} = (F^\top \hat{w} - \beta^\star)^\top \Sigma (F^\top \hat{w} - \beta^\star) = \Delta_w^\top \Sigma_F \Delta_w + S_\perp + 2\Delta_w^\top F\Sigma(F^\top w^\star - \beta^\star) = \Delta_w^\top \Sigma_F \Delta_w + S_\perp,$$

the cross term vanishing by Eq. 83.

Step 2: reduce to the pixel computation. By Lemma 35, $\Delta_w = -\lambda(\hat{\Sigma}_F + \lambda I)^{-1} w^\star + \frac{1}{n}(\hat{\Sigma}_F + \lambda I)^{-1} Z^\top (r + \sigma\eta)$, where conditionally on Z the vector $r + \sigma\eta$ has mean zero and covariance $(S_\perp + \sigma^2) I_n$. This is Eq. 56 with $(X, \Sigma, \beta^\star, \sigma^2)$ replaced by $(Z, \Sigma_F, w^\star, S_\perp + \sigma^2)$, so the proof of Eq. 66 applies verbatim:

$$\mathbb{E}\big[\Delta_w^\top \Sigma_F \Delta_w\big] \asymp \frac{n}{n - \mathrm{df}_2^F} \kappa^2 B_{1,2}^F + (S_\perp + \sigma^2) \frac{\mathrm{df}_2^F}{n - \mathrm{df}_2^F}.$$

Adding $S_\perp$ and rearranging gives Eq. 86. □

Proof of Eqs. 88 and 89. Numerator. $N = \beta^{\star\top} F^\top \hat{w}$. The noise part of Eq. 84 has zero conditional mean given Z (since $\mathbb{E}[r \mid Z] = 0$ and $\mathbb{E}[\eta] = 0$), and the one-point DE $(\hat{\Sigma}_F + \lambda I)^{-1} \hat{\Sigma}_F \asymp (\Sigma_F + \kappa I)^{-1} \Sigma_F$ gives

$$\mu_N \asymp \beta^{\star\top} F^\top (\Sigma_F + \kappa I)^{-1} \Sigma_F w^\star = \beta^{\star\top} F^\top (\Sigma_F + \kappa I)^{-1} F \Sigma \beta^\star = \beta^{\star\top} \bar{\beta}_{F,\kappa}, \tag{90}$$

where $\Sigma_F w^\star = \Sigma_F \Sigma_F^\dagger F \Sigma \beta^\star = F\Sigma\beta^\star$ because $F\Sigma\beta^\star$ lies in the range of Σ_F. The third form of $\bar{\beta}_{F,\kappa}$ is the push-through identity $F^\top (F M F^\top + \kappa I)^{-1} F = (F^\top F M + \kappa I)^{-1} F^\top F$ with $M = \Sigma$.

Denominator. $D = \hat{w}^\top G \hat{w}$. Averaging over (r, η) kills the cross term and leaves

$$\mathbb{E}_{r,\eta}[D] = w^{\star\top} \hat{\Sigma}_F (\hat{\Sigma}_F + \lambda I)^{-1} G (\hat{\Sigma}_F + \lambda I)^{-1} \hat{\Sigma}_F w^\star + (S_\perp + \sigma^2) \frac{1}{n} \operatorname{Tr}\big[\hat{\Sigma}_F (\hat{\Sigma}_F + \lambda I)^{-2} G\big].$$

The noise trace is the derivative argument of Section F on the design Z: $\frac{1}{n}\operatorname{Tr}[\hat{\Sigma}_F(\hat{\Sigma}_F + \lambda I)^{-2} G] \asymp T_F(\kappa)/(n - \mathrm{df}_2^F)$. The signal term is the two-point DE Eq. 112 on Z with $A = w^\star w^{\star\top}$ and $B = G$:

$$w^{\star\top} \Sigma_F (\Sigma_F + \kappa I)^{-1} G (\Sigma_F + \kappa I)^{-1} \Sigma_F w^\star + \frac{\kappa^2 T_F(\kappa)}{n - \mathrm{df}_2^F} w^{\star\top} (\Sigma_F + \kappa I)^{-2} \Sigma_F w^\star.$$

The first piece is $\|F^\top (\Sigma_F + \kappa I)^{-1} \Sigma_F w^\star\|^2 = \|\bar{\beta}_{F,\kappa}\|^2$, and the second combines with the noise trace into

$$\mu_D \asymp \|\bar{\beta}_{F,\kappa}\|^2 + \frac{T_F(\kappa)}{n - \mathrm{df}_2^F} \Big(\kappa^2 B_{1,2}^F + S_\perp + \sigma^2\Big) = \|\bar{\beta}_{F,\kappa}\|^2 + \frac{E_{gen,DE} + \sigma^2}{n} T_F(\kappa), \tag{91}$$

by Eq. 86. □

Remark 9 (Reading the feature-space forms).

1. $\bar{\beta}_{F,\kappa}$ is the realizable weight $w^\star$, shrunk by the usual ridge factor $\Sigma_F(\Sigma_F + \kappa I)^{-1}$ in feature space, and then mapped to pixels by $F^\top$. Setting $F = I_d$ gives $\Sigma_F = \Sigma$, $w^\star = \beta^\star$, $S_\perp = 0$, $G = I$, $T_F = \mathrm{df}_{1,2}$, and Proposition 36 reduces to Proposition 30.

2. Prediction and control read the teacher through two different vectors: $h := F\Sigma\beta^\star = \mathrm{Cov}(F\mathbf{x}, \beta^{\star\top}\mathbf{x})$ enters $w^\star$, $S_\perp$ and E_{gen}, and measures how the teacher covaries with the features on the natural distribution; $q := F\beta^\star$ enters $N = \hat{w}^\top q$ and measures how the teacher responds when inputs move along the back-propagated gradient $F^\top \hat{w}$. Likewise Σ_F is the covariance the fit sees, while G is the Euclidean length that back-propagation assigns to feature directions. This is the origin of Corollary 16.

3. Holding E_{gen} fixed, the leading control error $z_{acc} = \mu_D/\mu_N - 1$ can still change through σ^2 (at fixed E_{gen}, i.e. through n), through $\|\bar{\beta}_{F,\kappa}\|^2$ and $\beta^{\star\top}\bar{\beta}_{F,\kappa}$, and through T_F.

H.4 Tuning conditions in feature space

Proof of Proposition 17. Prediction optimum. Eq. 86 has the form $E_{gen,DE} + \sigma^2 = nH_F/q_F$ with $H_F = \kappa^2 B_{1,2}^F + \tau^2$, $\tau^2 := S_\perp + \sigma^2$, and $q_F = n - \mathrm{df}_2^F$. The computation of Section G applies with the spectrum of Σ_F and the teacher $w^\star$, giving at an interior optimum

$$q_F\,\kappa\,B_{2,3}^F = H_F\,\mathrm{df}_{2,3}^F \quad\Longleftrightarrow\quad E_{gen,DE} + \sigma^2 = n\kappa\frac{B_{2,3}^F}{\mathrm{df}_{2,3}^F}. \tag{92}$$

$S_\perp$ enters the location of the optimum through H_F even though it is absorbed into $E_{gen,DE} + \sigma^2$ on the right.

Control optimum. From Proposition 36,

$$\mu_D - \mu_N = \|\bar\beta_{F,\kappa}\|^2 - \beta^{\star\top}\bar\beta_{F,\kappa} + \frac{E_{gen,DE} + \sigma^2}{n}\,T_F(\kappa), \tag{93}$$

which vanishes iff $E_{gen,DE} + \sigma^2 = n\,(\beta^{\star\top}\bar\beta_{F,\kappa} - \|\bar\beta_{F,\kappa}\|^2)/T_F$ when $T_F > 0$. To expose the geometry of the numerator, write $\bar w_\kappa := \Sigma_F(\Sigma_F + \kappa I)^{-1}w^\star = w^\star - \kappa(\Sigma_F + \kappa I)^{-1}w^\star$ so that $\bar\beta_{F,\kappa} = F^\top\bar w_\kappa$, and use $F\beta^\star = Gw^P$ (as $F\beta^\star$ lies in the range of G):

$$\begin{aligned}\beta^{\star\top}\bar\beta_{F,\kappa} - \|\bar\beta_{F,\kappa}\|^2 &= (\beta^\star - F^\top\bar w_\kappa)^\top F^\top\bar w_\kappa = (F\beta^\star - G\bar w_\kappa)^\top\bar w_\kappa\\ &= \big(Gw^P - Gw^\star + \kappa G(\Sigma_F + \kappa I)^{-1}w^\star\big)^\top\bar w_\kappa\\ &= (w^P - w^\star)^\top G\bar w_\kappa + \kappa\,w^{\star\top}(\Sigma_F + \kappa I)^{-1}G\bar w_\kappa.\end{aligned}$$

For $F = I_d$ the first term vanishes ($w^P = w^\star = \beta^\star$) and the second is $\kappa B_{1,2}$, recovering Eq. 76. In general the first term has no definite sign: when the prediction-optimal and reconstruction-optimal weights differ, the numerator can be negative for every κ, in which case no ridge level calibrates control at leading order. □

I Gauge invariance and gradient fragility: proofs

I.1 Distinct summary statistics for prediction and control

Proof of Corollary 16. By inspection of Proposition 36. The curve $E_{gen,DE}(\kappa)$ is built from $\kappa(\lambda)$, which depends on the spectrum of Σ_F and on p/n; from df_2^F, a function of the spectrum of Σ_F; from $B_{1,2}^F = w^{\star\top}\Sigma_F(\Sigma_F + \kappa I)^{-2}w^\star = h^\top\Sigma_F^\dagger\Sigma_F(\Sigma_F + \kappa I)^{-2}\Sigma_F^\dagger h$ with $h = F\Sigma\beta^\star$; and from $S_\perp = S - h^\top\Sigma_F^\dagger h$. Hence it is a function of $\mathcal{P} = (\Sigma_F, h, S, \sigma^2, n)$ only, and so are its minimizer κ_{CV} and $R_{gen}^2 = 1 - E_{gen,DE}/S$. By contrast $\mu_N = q^\top(\Sigma_F + \kappa I)^{-1}h$ with $q = F\beta^\star$, and $\|\bar\beta_{F,\kappa}\|^2 = h^\top(\Sigma_F + \kappa I)^{-1}G(\Sigma_F + \kappa I)^{-1}h$ and $T_F = \mathrm{Tr}[G(\Sigma_F + \kappa I)^{-2}\Sigma_F]$ involve G. Figure 27 displays these dependencies. □

I.2 The prediction-preserving gauge family

Since prediction depends on F only through (Σ_F, h), any transformation of F that fixes these two objects leaves prediction, and hence its cross-validated optimum, unchanged. We characterize such transformations and show that they generically change control.

Proposition 37 (Restatement of Proposition 18)**.** *Assume $\Sigma \succ 0$. For $Q \in O(d)$ with $Q\Sigma^{1/2}\beta^\star = \Sigma^{1/2}\beta^\star$, the map $F_Q := F\Sigma^{1/2}Q\Sigma^{-1/2}$ satisfies $F_Q\Sigma F_Q^\top = \Sigma_F$ and $F_Q\Sigma\beta^\star = h$. Consequently the joint Gaussian law of the training pairs $(F_Q\mathbf{x}, y)$, the full curve $E_{gen,DE}(\kappa)$, and the DE-based cross-validated penalty are the same for all F_Q in the family, whereas $G_Q = F_QF_Q^\top$ and $q_Q = F_Q\beta^\star$, and with them μ_N, μ_D and all accentuation metrics, vary in general. If in addition $Q\Sigma^{-1/2}\beta^\star = \Sigma^{-1/2}\beta^\star$, then $q_Q = q$ and only G_Q varies. For $\Sigma = I$ the family acts trivially on both prediction and control.*

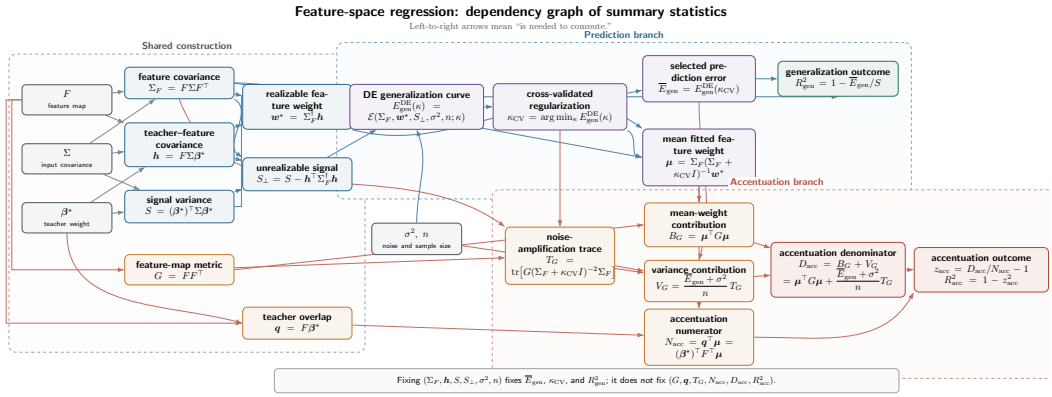

Figure 27: Dependency graph of the summary statistics of feature-space regression. Fixing $(\Sigma_F, h, S, S_\perp, \sigma^2, n)$ fixes the prediction branch; the accentuation branch additionally needs $G := FF^\top$ and $q = F\beta^\star$.

Proof. Direct computation: $F_Q \Sigma F_Q^\top = F\Sigma^{1/2} QQ^\top \Sigma^{1/2} F^\top = \Sigma_F$ and $F_Q \Sigma \beta^\star = F\Sigma^{1/2} Q \Sigma^{1/2} \beta^\star = F\Sigma^{1/2}\Sigma^{1/2}\beta^\star = h$. The pairs $(F_Q \mathbf{x}, y)$ are jointly Gaussian with covariance blocks $\mathrm{Cov}(F_Q \mathbf{x}) = \Sigma_F$, $\mathrm{Cov}(F_Q \mathbf{x}, y) = h$, $\mathrm{Var}(y) = S + \sigma^2$, all invariant; hence every functional of the training data, including the fitted $\hat{w}$, any K-fold or leave-one-out cross-validation score, and the population risk $E_{gen} = (\hat{w} - w^\star)^\top \Sigma_F (\hat{w} - w^\star) + S_\perp$, has the same law across the family. (Two members need not give the same CV choice on the same realization of X, only the same distribution of choices.) On the other hand $G_Q = F\Sigma^{1/2} Q \Sigma^{-1} Q^\top \Sigma^{1/2} F^\top$ and $q_Q = F\Sigma^{1/2} Q \Sigma^{-1/2} \beta^\star$ are not invariant unless Q commutes with Σ on the relevant subspace, respectively fixes $\Sigma^{-1/2}\beta^\star$. For $\Sigma = I$, $F_Q = FQ$ with $Q\beta^\star = \beta^\star$, so $G_Q = G$ and $q_Q = q$. □

I.3 Constructing the family

Whitened coordinates. The constraints are transparent in whitened coordinates. Let $A := F\Sigma^{1/2}$, $b := \Sigma^{1/2}\beta^\star$ and $c := \Sigma^{-1/2}\beta^\star$, so that $\Sigma_F = AA^\top$, $h = Ab$, $q = Ac$ and $G = A\Sigma^{-1}A^\top$. Two maps F, F' are prediction-equivalent iff $A'b = Ab$ and $A'A'^\top = AA^\top$; the first says that the rows of $A - A'$ are orthogonal to the whitened teacher b. A left rotation $F \mapsto O_p F$, $O_p \in O(p)$, is a change of feature coordinates that is undone by $w \mapsto O_p w$ and produces no new control behavior; the nontrivial freedom is on the right.

Right rotations with a fixed point. Let $W \in \mathbb{R}^{d \times (d-1)}$ be an orthonormal basis of $b^\perp$ and $R \in O(d-1)$. Then

$$O := \frac{bb^\top}{\|b\|^2} + WRW^\top \in O(d), \qquad Ob = b, \tag{94}$$

and $A_O := AO$, $F_O := A_O \Sigma^{-1/2} = F\Sigma^{1/2} O \Sigma^{-1/2}$ is the family of Proposition 37: a rotation of the whitened feature map within the subspace orthogonal to the whitened teacher. To fix q as well, take W to span the complement of $\mathrm{span}(b, c)$.

Example 38 (Two-dimensional rotation)**.** Let $\mathbf{v}_1, \mathbf{v}_2 \perp b$ be orthonormal and let $\tilde{A}$ have rows orthogonal to both. Then $A(\theta) := \tilde{A} + u(\theta) e_1 \mathbf{v}_1^\top + v(\theta) e_1 \mathbf{v}_2^\top$ has $A(\theta)A(\theta)^\top = \tilde{A}\tilde{A}^\top + (u^2 + v^2) e_1 e_1^\top$ and $A(\theta)b = \tilde{A}b$; any $u^2(\theta) + v^2(\theta) = \text{const}$, e.g. $(\cos\theta, \sin\theta)$, gives a one-parameter prediction-equivalent family $F(\theta) = A(\theta)\Sigma^{-1/2}$.

Example 39 (Resampling the null part)**.** Write $A = \frac{hb^\top}{\|b\|^2} + K\, Q_0$ with $K \in \mathbb{R}^{p \times r}$ and $Q_0 \in \mathbb{R}^{r \times d}$ having orthonormal rows orthogonal to b (the SVD of $A - hb^\top/\|b\|^2$). Replacing Q_0 by any Q_ϕ with orthonormal rows orthogonal to b leaves $A_\phi b = h$ and $A_\phi A_\phi^\top = hh^\top/\|b\|^2 + KK^\top$ unchanged. This gives a constructive recipe: (i) from $(\Sigma, F, \beta^\star)$ form A, b, h; (ii) take the SVD of the residual

$A - hb^\top/\|b\|^2$; (iii) replace its right singular vectors by a random orthonormal frame in $b^\perp$. Random draws from this recipe generate the "population" of equally predictive feature maps used in the experiments (Fig. 14).

Proof of Corollary 19. Let P carry the components of A that fix Ab (and optionally Ac), and let R, V be row frames with $PR^\top = PV^\top = 0$, $RV^\top = 0$, $RR^\top = VV^\top$, and $Rb = Vb = 0$. For $A(t) = P + \sqrt{1-t}\,R + \sqrt{t}\,V$ all cross terms vanish and

$$A(t)A(t)^\top = PP^\top + (1-t)RR^\top + tVV^\top = PP^\top + RR^\top, \qquad A(t)b = Pb,$$

so $F(t) = A(t)\Sigma^{-1/2}$ lies in the family for every $t \in [0,1]$. Choosing R supported on high-variance and V on low-variance eigendirections of Σ makes $G(t) = A(t)\Sigma^{-1}A(t)^\top$ acquire the factor s_k^{-1} of the tail directions as $t \to 1$, so $\mathrm{Tr}\, G(t)$ and T_F grow without bound while every prediction summary is constant. □

Example 40 (Aligned features, discrete version)**.** In the eigenbasis of $\Sigma = \mathrm{diag}(\lambda_1, \lambda_H, \lambda_L)$ with $\beta^\star = (1,0,0)^\top$, the two maps

$$F_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \lambda_H^{-1/2} & 0 \end{pmatrix}, \qquad F_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & \lambda_L^{-1/2} \end{pmatrix}$$

have the same $\Sigma_F = \mathrm{diag}(\lambda_1, 1)$ and the same h, hence the same generalization error and cross-validated penalty, but $G_1 = \mathrm{diag}(1, \lambda_H^{-1})$ versus $G_2 = \mathrm{diag}(1, \lambda_L^{-1})$: the second feature of F_2 reads a low-variance direction and back-propagates estimation noise into it with gain $\lambda_L^{-1} \gg \lambda_H^{-1}$.

Figure 12 (App. A.3) realizes this continuous rotation on van Hateren images, and Fig. 14 shows random draws from the family; the numerical construction is given in App. B.5.

I.4 A spectral diagnostic for fragile feature maps

The variance contribution to the accentuation denominator carries the response-independent factor

$$T_F(\kappa) := \mathrm{Tr}\left[FF^\top(\Sigma_F + \kappa I)^{-2}\Sigma_F\right], \tag{95}$$

which makes precise when estimation error that is modest in prediction geometry becomes large after back-propagation to pixels.

Proposition 41 (Restatement of Proposition 20)**.** *Let $\Sigma_F = \sum_k s_k v_k v_k^\top$ and $g_k := F^\top v_k$ be the pixel gradient of feature PC k. Then*

$$T_F(\kappa) = \sum_{k:s_k>0} \frac{s_k}{(s_k+\kappa)^2}\|g_k\|^2 = \sum_{k:s_k>0} \left(\frac{s_k}{s_k+\kappa}\right)^2 \frac{\|g_k\|^2}{g_k^\top \Sigma g_k}. \tag{96}$$

Proof. Cyclic invariance of the trace and the eigendecomposition of Σ_F give the first equality; $s_k = v_k^\top F\Sigma F^\top v_k = g_k^\top \Sigma g_k$ gives the second. □

Remark 10 (Prediction-protected but control-fragile directions). The ratio $\|g_k\|^2/g_k^\top\Sigma g_k$ compares the Euclidean energy of the back-propagated PC gradient with its energy under the natural-image covariance; it is large when g_k concentrates on low-variance, typically high-spatial-frequency, input directions. The ridge factor $(s_k/(s_k+\kappa))^2$ weights feature PCs that survive regularization. Control is therefore fragile when even the well-resolved feature PCs back-propagate into low-variance input directions. T_F can vary while every prediction summary and $E_{gen,DE}$ remain fixed, and it can be computed from the map and the image statistics alone. In the nonlinear diagnostic of Section K the per-PC gradient energies $\|g_k\|^2$ are replaced by local Jacobian energies.

J Ratios of random quadratic forms, and peer review

Random-matrix theory controls N and D separately more cleanly than it controls their ratio, and every accentuation metric is a nonlinear function of N/D. The main text uses the leading approximation $\mathbb{E}[N/D] \approx \mu_N/\mu_D$. This section quantifies its error, gives the response-noise contribution to the variance of the ratio, describes a distributional surrogate, and then uses the same tools for peer review, where the ratio involves two independent fits.

J.1 Second-order expansion of a ratio

Proposition 42 (Restatement of Proposition 13, first part)**.** *Let $N = \mu_N + \delta_N$, $D = \mu_D + \delta_D$ with $\mu_N, \mu_D \neq 0$ and fluctuations small enough for a second-order expansion. Then*

$$\mathbb{E}\Big[\frac{N}{D}\Big] \approx \frac{\mu_N}{\mu_D}\Big(1 - \frac{\mathrm{Cov}(N,D)}{\mu_N\mu_D} + \frac{\mathrm{Var}(D)}{\mu_D^2}\Big), \tag{97}$$

$$\mathbb{E}\Big[\Big(\frac{N}{D}\Big)^2\Big] \approx \frac{\mu_N^2}{\mu_D^2}\Big(1 + \frac{\mathrm{Var}(N)}{\mu_N^2} - 4\frac{\mathrm{Cov}(N,D)}{\mu_N\mu_D} + 3\frac{\mathrm{Var}(D)}{\mu_D^2}\Big). \tag{98}$$

Proof. Expand $(\mu_D + \delta_D)^{-1} = \mu_D^{-1}(1 - \delta_D/\mu_D + \delta_D^2/\mu_D^2 - \dots)$. For the ratio,

$$\frac{N}{D} = \frac{\mu_N}{\mu_D} + \frac{\delta_N}{\mu_D} - \frac{\mu_N\delta_D}{\mu_D^2} - \frac{\delta_N\delta_D}{\mu_D^2} + \frac{\mu_N\delta_D^2}{\mu_D^3} + O(\delta^3),$$

and taking expectations, the first-order terms vanish and the second-order terms give Eq. 97. For the square,

$$\Big(\frac{N}{D}\Big)^2 = \frac{\mu_N^2 + 2\mu_N\delta_N + \delta_N^2}{\mu_D^2}\Big(1 - 2\frac{\delta_D}{\mu_D} + 3\frac{\delta_D^2}{\mu_D^2} + O(\delta^3)\Big),$$

whose expectation retains $\mu_N^2/\mu_D^2 + 3\mu_N^2\mathrm{Var}(D)/\mu_D^4 - 4\mu_N\mathrm{Cov}(N,D)/\mu_D^3 + \mathrm{Var}(N)/\mu_D^2$, which is Eq. 98. □

The leading term needs μ_N, μ_D only; the correction needs $\mathrm{Var}(N)$, $\mathrm{Var}(D)$ and $\mathrm{Cov}(N,D)$ relative to the means. The expansion fails when the distribution of N puts appreciable mass near zero, in which case moments of D/N may not exist and medians and quantiles should be compared instead.

J.2 Gaussian distributional surrogate

A third approximation treats the fitted weight itself as Gaussian, $\hat\beta \dot\sim \mathcal{N}(m_{DE}, V_{DE})$, with the DE mean $m_{DE} = \Sigma(\Sigma + \kappa I)^{-1}\beta^\star$ and a covariance that is diagonal in the eigenbasis of Σ with the per-mode variances of Eq. 69 (signal and noise fluctuation terms), and evaluates any metric by Monte Carlo over this surrogate: $\mathbb{E}_{\hat\beta\sim\mathcal{N}(m_{DE},V_{DE})}[\,\hat\beta^\top\beta^\star/\hat\beta^\top\hat\beta\,]$ and likewise for R^2_{acc} and the slope. It yields medians, quantiles and the probability of sign changes of N, which the moment expansions cannot. It is a surrogate rather than a proved distributional deterministic equivalent: matching per-coordinate second moments does not match the joint fluctuations of quadratic forms; it omits off-diagonal, design-induced covariance; and it does not enforce the exact ridgeless identity $N = D$ that holds for the minimum-norm interpolator ($\hat\beta = P_X\beta^\star$ with P_X a projector, so $N = D$ for every design). In practice it agrees with the leading DE where the latter is accurate and inherits its failures near $\lambda \to 0$.

J.3 Peer review between independent fits

Peer review evaluates $\hat\beta$ on the path generated by another model β'. Its metrics (Eq. 61) involve the two cross products $\hat\beta^\top\beta'$ and $\beta^{\star\top}\beta'$. When the two models are independent ridge fits, both have simple leading equivalents.

Proposition 43 (Restatement of Proposition 12)**.** *Let $\hat\beta$ and β' be ridge fits from independent samples with effective penalties κ and κ'. Then*

$$\hat\beta^\top\beta' \asymp \sum_k \frac{s_k^2\, b_k^2}{(s_k+\kappa)(s_k+\kappa')}, \qquad \beta^{\star\top}\beta' \asymp B_{1,1}(\kappa'), \tag{99}$$

so that for equal penalties

$$\begin{aligned} G_{peer} &:= \frac{\hat\beta^\top\beta'}{\beta^{\star\top}\beta'} \asymp \frac{B_{2,2}}{B_{1,1}} \in (0,1], \\ R^{2,(0)}_{peer} &= 1 - \Big(1 - \frac{B_{2,2}}{B_{1,1}}\Big)^2 \in (0,1], \qquad \mathrm{slope}^{(0)}_{peer} = \frac{B_{1,1}}{B_{2,2}} \geq 1. \end{aligned} \tag{100}$$

Equivalently, with $z_{peer} := \hat{\beta}^\top \beta' / \beta^{\star\top} \beta' - 1 \asymp B_{2,2}/B_{1,1} - 1 \in (-1, 0]$*, the self-review error of Proposition 10 decomposes as* $z_{acc}^{(0)} = z_{peer} + V_{\mathrm{pix}}/B_{1,1}$*: self-review is peer review plus the norm inflation.*

Proof. Write $\hat{\beta} = g + h$ and $\beta' = g' + h'$ as in the previous subsection. By independence of the two samples and $\mathbb{E}[h \mid X] = \mathbb{E}[h' \mid X'] = 0$, the cross products $g^\top h'$, $h^\top g'$, $h^\top h'$ have zero mean, and their fluctuations are of lower order, so $\hat{\beta}^\top \beta' \asymp g^\top g'$. By the one-point DE, $g \asymp \Sigma(\Sigma + \kappa I)^{-1} \beta^\star$ in the sense of quadratic forms, giving $g^\top g' \asymp \beta^{\star\top} \Sigma(\Sigma + \kappa I)^{-1} \Sigma(\Sigma + \kappa' I)^{-1} \beta^\star$, which is the stated sum; similarly $\beta^{\star\top} \beta' \asymp \beta^{\star\top} \Sigma(\Sigma + \kappa' I)^{-1} \beta^\star = B_{1,1}(\kappa')$. For $\kappa = \kappa'$ the first is $B_{2,2}$, and $0 < B_{2,2} \le B_{1,1}$ since $s_k/(s_k + \kappa) \le 1$. □

Remark 11 (Why peer review is benign and self-review is not). The Euclidean noise energy $h^\top h$ that inflates the self-denominator D is absent from $\hat{\beta}^\top \beta'$ because independent noise realizations are uncorrelated. Peer review therefore measures only the shared shrinkage bias: at leading order R^2_{peer} is never negative, degrades gracefully with noise (through κ), and is symmetric between the two fits. Self-review contains the estimator's own noise energy and $R^2_{acc} = 1 - z^2_{acc}$ is unbounded below. This is the linear form of "recognition without generation". With observation noise on the measured responses, a path that barely modulates the true response (small $\beta^{\star\top} \beta'$) yields poor empirical scores for every evaluator, which is why the peer scores of a weak generator's images are also low in the data; and two models with different feature maps have different noise energies, which breaks the symmetry.

Remark 12 (Perfect peer, and second-order corrections). If the generator is the teacher itself, $\beta' = \beta^\star$, then $E_{peer} = \mathrm{Var}[\alpha](\beta^{\star\top}(\hat{\beta} - \beta^\star))^2$, $R^2_{peer} = 1 - (\beta^{\star\top} \hat{\beta}/\|\beta^\star\|^2 - 1)^2$ and $\mathrm{slope}_{peer} = \|\beta^\star\|^2 / \beta^{\star\top} \hat{\beta}$, to be compared with the self-review slope $\hat{\beta}^\top \beta^\star / \hat{\beta}^\top \hat{\beta}$: the generation side divides by the student's own norm, the evaluation side by the teacher's. The peer slope is again an expectation of a random ratio and is handled by Proposition 42. For two independent fits, writing $m_k = s_k b_k/(s_k + \kappa)$ and v_k for the per-mode fluctuation variance of Eq. 69, the delta method with $U = \hat{\beta}^\top \beta'$, $V = \beta^{\star\top} \beta'$ uses $\mathrm{Var}(U) = 2\sum_k m_k^2 v_k + \sum_k v_k^2$, $\mathrm{Var}(V) = \sum_k b_k^2 v_k$, $\mathrm{Cov}(U, V) = \sum_k m_k b_k v_k$; the resulting corrections to R^2_{peer} are small (maximum gap below 0.005 in the validation experiments), consistent with the benign character of the peer ratio.

J.4 Peer review on shared data across feature maps

In the neuronal experiment all encoding models are fitted to the *same* recordings, and the models differ in their feature maps. Peer review then involves two fits that share the design X and the noise realization η, and the independence used in Proposition 43 is lost: the evaluator can recognize its own estimation noise in the generator's path. The following statement is exact conditionally on the design and isolates the term that carries this effect.

Let two students $i \in \{1, 2\}$ read out from fixed maps $F_i \in \mathbb{R}^{p_i \times d}$ with penalties λ_i, both trained on (X, η). Define the pixel-space smoother of student i,

$$M_i := F_i^\top \left(F_i \hat{\Sigma} F_i^\top + \lambda_i I_{p_i}\right)^{-1} F_i \in \mathbb{R}^{d \times d}, \qquad \hat{\beta}_i = \underbrace{M_i \hat{\Sigma} \beta^\star}_{g_i} + \underbrace{\tfrac{\sigma}{n} M_i X^\top \eta}_{\varepsilon_i}, \tag{101}$$

which is Eq. 81 written in pixel space, and let student 1 evaluate the path generated by student 2, so that by Eq. 61 the peer gain is $G_{1\leftarrow 2} := \hat{\beta}_2^\top \hat{\beta}_1 / \hat{\beta}_2^\top \beta^\star$, $R^2_{1\leftarrow 2} = 1 - (G_{1\leftarrow 2} - 1)^2$ and $\mathrm{slope}_{1\leftarrow 2} = 1/G_{1\leftarrow 2}$.

Proposition 44 (Restatement of Proposition 21)**.** *Conditionally on the design* X*,*

$$\mathbb{E}_\eta\left[\hat{\beta}_1^\top \hat{\beta}_2\right] = g_1^\top g_2 + \Omega_{12}, \qquad \Omega_{12} := \mathbb{E}_\eta\left[\varepsilon_1^\top \varepsilon_2\right] = \frac{\sigma^2}{n} \mathrm{Tr}\left[M_1 \hat{\Sigma} M_2\right], \qquad \mathbb{E}_\eta\left[\beta^{\star\top} \hat{\beta}_2\right] = \beta^{\star\top} g_2. \tag{102}$$

The shared-noise overlap satisfies $|\Omega_{12}| \le \sqrt{\Omega_{11}\Omega_{22}}$*, where* $\Omega_{ii} = \frac{\sigma^2}{n} \mathrm{Tr}[M_i \hat{\Sigma} M_i]$ *is the Euclidean noise energy that inflates the self-denominator* D_i*; we write* $\Omega_{12} = \rho_{12}\sqrt{\Omega_{11}\Omega_{22}}$ *with* $\rho_{12} \in [-1, 1]$*. Consequently, at the level of ratios of conditional means,*

$$G_{1\leftarrow 2} \approx \frac{g_1^\top g_2 + \rho_{12}\sqrt{\Omega_{11}\Omega_{22}}}{\beta^{\star\top} g_2}, \qquad \textit{versus} \qquad G_{2\leftarrow 2} = \frac{D_2}{N_2} \approx \frac{\|g_2\|^2 + \Omega_{22}}{\beta^{\star\top} g_2}. \tag{103}$$

Two limits are exact: for identical students ($F_1 = F_2$, $\lambda_1 = \lambda_2$) Eq. 102 reduces to the self-accentuation denominator of Proposition 36; for fits on independent data the overlap term is absent and, at leading order, $G_{1\leftarrow 2} \asymp \bar{\beta}_1^\top \bar{\beta}_2 / \beta^{\star\top} \bar{\beta}_2$ with $\bar{\beta}_i = F_i^\top (\Sigma_{F_i} + \kappa_i I)^{-1} F_i \Sigma \beta^\star$, which generalizes Proposition 43 to different feature maps.

Proof. Expanding $\hat{\beta}_1^\top \hat{\beta}_2 = (g_1 + \varepsilon_1)^\top (g_2 + \varepsilon_2)$, the cross terms are linear in η and vanish under $\mathbb{E}_\eta$, while $\mathbb{E}_\eta[\varepsilon_1^\top \varepsilon_2] = \frac{\sigma^2}{n^2} \mathbb{E}_\eta[\eta^\top X M_1 M_2 X^\top \eta] = \frac{\sigma^2}{n^2} \operatorname{Tr}[X M_1 M_2 X^\top] = \frac{\sigma^2}{n} \operatorname{Tr}[M_1 \hat{\Sigma} M_2]$, using $\mathbb{E}[\eta\eta^\top] = I_n$ and $M_i = M_i^\top$. The bound on Ω_{12} is the Cauchy–Schwarz inequality for $\mathbb{E}_\eta[\varepsilon_1^\top \varepsilon_2]$. For identical students $g_1 = g_2$ and $\Omega_{12} = \Omega_{22}$, which is $\mathbb{E}_\eta[D_2]$. For independent data ε_1 and ε_2 are independent with zero mean, g_1 and g_2 are independent, and the one-point DE gives $g_i \asymp \bar{\beta}_i$ in quadratic forms (Eq. 90). □

Remark 13 (Recognition, agreement and the overlap of noise directions).

1. *What replaces independence.* On shared data the numerator of the peer gain contains the overlap Ω_{12} of the two models' noise-driven pixel weights. ε_i is the response noise back-propagated through map i, so Ω_{12} measures whether the two maps route the same noise into the same pixel directions. Two models with similar gradient geometry ($\rho_{12} \approx 1$) share their noise and agree on inflated paths: each over-predicts on the other's stimuli almost as much as on its own. Two models with dissimilar gradient geometry, e.g. one whose gradients live in high-spatial-frequency directions and one whose gradients are natural-image-like, have small ρ_{12} and evaluate each other with little noise-induced inflation, so each can recognize what the other's path does to the teacher. Peer review across dissimilar models therefore behaves like the independent-fit case even though the data are shared.

2. *Asymmetry.* The numerator of Eq. 103 is symmetric in the two students; the asymmetry between $G_{1\leftarrow 2}$ and $G_{2\leftarrow 1}$ comes from the denominators, the true gains $\beta^{\star\top} g_i$ of the two generators, and, once measured responses carry trial noise, from the different signal levels along the two paths. In the noiseless linear model the only source of asymmetry is the generator's true gain.

3. *What is computable without random-matrix theory.* All terms of Eq. 102 are explicit functions of the empirical $\hat{\Sigma}$ and of the two maps, and can be evaluated on a given dataset for any number of models without simulation: the conditional-on-design expectation is the relevant average when, as in the experiment, the same recordings are used for every model. Replacing g_i by $\bar{\beta}_i$ and the noise traces by population quantities requires deterministic equivalents for products of resolvents of two *different* sample covariances $F_1 \hat{\Sigma} F_1^\top$ and $F_2 \hat{\Sigma} F_2^\top$ built from the same design; these are not derived here. The naive replacement of each resolvent by its own one-point equivalent misses the coupling between the two: in the identical-map limit it gives $\Omega_{ii} \approx \sigma^2 (\kappa/\lambda) T_F$ instead of the correct $\sigma^2 \, n T_F / (n - \mathrm{df}_2)$, so it should be used only as a qualitative guide to the overlap.

K Nonlinear gradient diagnostic

The linear-feature theory does not become a theorem for a nonlinear map $\phi : \mathbb{R}^d \to \mathbb{R}^p$, but its variance term suggests a local diagnostic. Let

$$J_\phi(x) := \frac{\partial \phi(x)}{\partial x} \in \mathbb{R}^{p \times d}, \qquad \Sigma_\phi := \operatorname{Cov}_{x \sim P_x}[\phi(x)] = \sum_k s_k v_k v_k^\top. \tag{104}$$

Replacing the constant Jacobian F by $J_\phi(x)$ in Eq. 96 gives

$$T_\phi(x, \kappa) := \sum_k \frac{s_k}{(s_k + \kappa)^2} \| J_\phi(x)^\top v_k \|^2, \qquad V_\phi := \frac{E_{val}}{n} \, \mathbb{E}_x[T_\phi(x, \kappa)], \tag{105}$$

where $J_\phi(x)^\top v_k = \nabla_x [v_k^\top \phi(x)]$ is the input gradient of feature PC k and E_{val} is the held-out prediction error. Two remarks fix its status. First, for any differentiable ϕ and any evaluation distribution Q over inputs, the trace identity $\operatorname{Tr}[G_Q (\Sigma_\phi + \kappa I)^{-2} \Sigma_\phi] = \sum_k q_k \, s_k / (s_k + \kappa)^2$ with $G_Q := \mathbb{E}_Q[J_\phi J_\phi^\top]$ and $q_k := \mathbb{E}_Q \| \nabla_x (v_k^\top \phi) \|^2$ is exact, so $\mathbb{E}_Q[T_\phi]$ can be computed from one backward pass per PC without materializing the Jacobian, averaging *squared* gradients. Second, unlike

the linear case, $s_k \neq \nabla_x[v_k^\top \phi]^\top \Sigma \nabla_x[v_k^\top \phi]$ in general, so Eq. 105 is a diagnostic motivated by the theory, not a deterministic equivalent for nonlinear regression: it compares a local gradient energy with a global response variance.

Finite-neighborhood versions. An exact Jacobian at the seed may be too local for an accentuation path that travels far from it. For $z \sim \mathcal{N}(0, I_d)$ and a neighborhood scale τ, natural PC-wise energies are

$$M_{smooth,k} := \left\| \mathbb{E}_z[J_\phi(x_0 + \tau z)^\top v_k] \right\|^2, \quad M_{neighbor,k} := \mathbb{E}_z\left[\|J_\phi(x_0 + \tau z)^\top v_k\|^2 \right],$$
$$M_{var,k} := \tau^{-2} \mathrm{Var}_z[v_k^\top \phi(x_0 + \tau z)], \qquad M_{step,k} := \tau^{-2} \mathbb{E}_z\left[(v_k^\top \phi(x_0 + \tau z) - v_k^\top \phi(x_0))^2 \right]. \tag{106}$$

They coincide in the locally linear, small-τ limit and otherwise answer different questions: squared mean gradient (the gradient of the Gaussian-smoothed function), mean squared gradient, response variance, and finite-step sensitivity, the last of which also contains the mean drift of the response.

Stein estimator. Stein's identity gives, for differentiable ϕ with suitable integrability,

$$\mathbb{E}_z[J_\phi(x_0 + \tau z)^\top v_k] = \frac{1}{\tau} \mathbb{E}_z\left[(v_k^\top \phi(x_0 + \tau z))\, z \right], \tag{107}$$

a forward-only estimator of the smoothed gradient. Centering by $\phi(x_0)$ is a control variate and antithetic pairs $x_0 \pm \tau z$ replace the response by $v_k^\top [\phi(x_0 + \tau z) - \phi(x_0 - \tau z)]/2$; both preserve the population identity and reduce Monte Carlo variance. The one-sided form is unstable in 10^5 input dimensions (its finite-sample off-diagonal U-statistic is often negative), whereas the antithetic form with several hundred directions is positive on every site–model pair and agrees with the smoothed-Jacobian energy to within a few percent. Empirical use should report the neighborhood scale, the number of probes and the estimator variance rather than treat a noisy Jacobian estimate as exact.

K.1 Empirical validation against neuronal control

The empirical test of this diagnostic on the 250 site–model pairs of the neuronal experiment is reported in Sec. 6 and Fig. 5, with data, models and computation in Apps. B.6–B.8 and the full comparison of estimators in App. A.5.

L Random-matrix toolkit

This section records the deterministic equivalents and the spectral-sum identities used in the proofs. The design is $X \in \mathbb{R}^{n \times d}$ with i.i.d. rows $\mathbf{x}_i \sim \mathcal{N}(0, \Sigma)$, $\hat{\Sigma} = X^\top X/n$, and $\gamma = d/n$; all statements hold in the proportional limit and $\asymp$ denotes almost-sure asymptotic equivalence of the two sides. For feature-space regression the same statements apply to the design $Z = XF^\top$ with (Σ, d) replaced by (Σ_F, p).

L.1 Fixed point and deterministic equivalents

Effective penalty. For $\lambda > 0$ let $\kappa = \kappa(\lambda)$ be the unique positive solution of

$$\kappa - \lambda = \gamma\, \kappa\, \mathrm{tr}\left[\Sigma(\Sigma + \kappa I)^{-1}\right] = \frac{\kappa}{n} \mathrm{df}_1(\kappa) \quad \Longleftrightarrow \quad n\lambda = \kappa\big(n - \mathrm{df}_1(\kappa)\big), \tag{108}$$

where $\mathrm{tr} = \mathrm{Tr}/d$ is the normalized trace. This is the Silverstein equation in the parametrization $\kappa(z) = 1/\varphi(-z)$ with $\varphi(z) = \mathrm{tr}[(XX^\top - nzI)^{-1}]$; in terms of the spectral measure μ of Σ,

$$\kappa(\lambda) - \lambda = \gamma\, \kappa(\lambda) \int_0^\infty \frac{s\, d\mu(s)}{\kappa(\lambda) + s}. \tag{109}$$

κ is increasing in λ, $\kappa \geq \lambda$, and $\kappa \to \kappa_0 > 0$ as $\lambda \to 0$ when $\gamma > 1$ (or when Σ has rank exceeding n), which is the ridgeless limit.

One-point equivalent. For deterministic A of bounded norm,

$$\begin{aligned} \operatorname{tr}\left[A(\hat{\Sigma}+\lambda I)^{-1}\right] &\asymp \frac{\kappa}{\lambda}\operatorname{tr}\left[A(\Sigma+\kappa I)^{-1}\right], \\ \text{equivalently} \quad \operatorname{tr}\left[A\hat{\Sigma}(\hat{\Sigma}+\lambda I)^{-1}\right] &\asymp \operatorname{tr}\left[A\Sigma(\Sigma+\kappa I)^{-1}\right]. \end{aligned} \tag{110}$$

The two forms are equivalent by $\hat{\Sigma}(\hat{\Sigma}+\lambda I)^{-1} = I - \lambda(\hat{\Sigma}+\lambda I)^{-1}$ and $I - \kappa(\Sigma+\kappa I)^{-1} = \Sigma(\Sigma+\kappa I)^{-1}$. We use them also for rank-one $A = \beta^\star\beta^{\star\top}$ (quadratic forms), for which the same equivalence holds without normalization.

Two-point equivalent. For deterministic A, B,

$$\operatorname{Tr}\left[A(\hat{\Sigma}+\lambda I)^{-1}B(\hat{\Sigma}+\lambda I)^{-1}\right] \asymp \frac{\kappa^2}{\lambda^2}\Big\{ \operatorname{Tr}\left[A(\Sigma+\kappa I)^{-1}B(\Sigma+\kappa I)^{-1}\right] \tag{111}$$

$$+ \frac{\operatorname{Tr}\left[A(\Sigma+\kappa I)^{-2}\Sigma\right]\operatorname{Tr}\left[B(\Sigma+\kappa I)^{-2}\Sigma\right]}{n-\mathrm{df}_2(\kappa)}\Big\}, \tag{112}$$

$$\operatorname{Tr}\left[A\hat{\Sigma}(\hat{\Sigma}+\lambda I)^{-1}B\hat{\Sigma}(\hat{\Sigma}+\lambda I)^{-1}\right] \asymp \operatorname{Tr}\left[A\Sigma(\Sigma+\kappa I)^{-1}B\Sigma(\Sigma+\kappa I)^{-1}\right] \tag{113}$$

$$+ \frac{\kappa^2\operatorname{Tr}\left[A(\Sigma+\kappa I)^{-2}\Sigma\right]\operatorname{Tr}\left[B(\Sigma+\kappa I)^{-2}\Sigma\right]}{n-\mathrm{df}_2(\kappa)}. \tag{114}$$

The second term is the finite-sample correction that couples the two resolvents; it is what produces the $n/(n-\mathrm{df}_2)$ factors and the finite-sample signal-variance terms in the proofs. For two different penalties λ, λ' the same form holds with $T_\Sigma = \Sigma(\Sigma+\kappa I)^{-1}$, $T'_\Sigma = \Sigma(\Sigma+\kappa' I)^{-1}$, $G_\Sigma = (\Sigma+\kappa I)^{-1}$, $G'_\Sigma = (\Sigma+\kappa' I)^{-1}$:

$$\operatorname{Tr}\left[A\hat{\Sigma}(\hat{\Sigma}+\lambda I)^{-1}B\hat{\Sigma}(\hat{\Sigma}+\lambda' I)^{-1}\right] \asymp \operatorname{Tr}[AT_\Sigma BT'_\Sigma]$$

$$+ \frac{\kappa\kappa'}{n-\mathrm{df}_2(\kappa,\kappa')}\operatorname{Tr}[AG_\Sigma\Sigma G'_\Sigma]\operatorname{Tr}[G'_\Sigma\Sigma G_\Sigma B], \tag{115}$$

$$\mathrm{df}_2(\kappa,\kappa') := \operatorname{Tr}[\Sigma^2 G_\Sigma G'_\Sigma]. \tag{116}$$

These are the standard anisotropic ridge equivalents (Dobriban & Wager, 2018; Hastie et al., 2022); the two-point form in the normalization used here follows the deterministic-equivalent calculus for products of resolvents.

Derivative of the effective penalty. Differentiating Eq. 108 in the form $\lambda = \kappa - \gamma\kappa\operatorname{tr}[\Sigma(\Sigma+\kappa I)^{-1}]$,

$$\begin{aligned} \frac{d\lambda}{d\kappa} &= 1 - \gamma\operatorname{tr}\left[\Sigma(\Sigma+\kappa I)^{-1}\right] + \gamma\kappa\operatorname{tr}\left[\Sigma(\Sigma+\kappa I)^{-2}\right] \\ &= 1 - \gamma\operatorname{tr}\left[\Sigma^2(\Sigma+\kappa I)^{-2}\right] \\ &= 1 - \frac{\mathrm{df}_2(\kappa)}{n}, \end{aligned}$$

so that

$$\frac{d\kappa}{d\lambda} = \frac{n}{n-\mathrm{df}_2(\kappa)} > 0. \tag{117}$$

Combined with the one-point equivalent this gives the trace identities used for the noise terms: $\operatorname{Tr}[\Sigma(\hat{\Sigma}+\lambda I)^{-2}\hat{\Sigma}] \asymp \frac{n}{n-\mathrm{df}_2}\mathrm{df}_2$ and $\operatorname{Tr}[(\hat{\Sigma}+\lambda I)^{-2}\hat{\Sigma}] \asymp \frac{n}{n-\mathrm{df}_2}\mathrm{df}_{1,2}$, both obtained by writing the left side as $-\frac{d}{d\lambda}$ of a one-point trace.

L.2 Spectral sums

Definitions and interpretation. With $\Sigma = \sum_k s_k \mathbf{u}_k\mathbf{u}_k^\top$ and $b_k = \mathbf{u}_k^\top\beta^\star$,

$$\begin{aligned} B_{p,q}(\kappa) &:= \sum_{k=1}^{d}\frac{s_k^p}{(s_k+\kappa)^q}b_k^2 = \beta^{\star\top}\Sigma^p(\Sigma+\kappa I)^{-q}\beta^\star, \\ \mathrm{df}_{p,q}(\kappa) &:= \sum_{k=1}^{d}\frac{s_k^p}{(s_k+\kappa)^q} = \operatorname{Tr}\left[\Sigma^p(\Sigma+\kappa I)^{-q}\right], \end{aligned} \tag{118}$$

with the classical shorthands $\mathrm{df}_1 = \mathrm{df}_{1,1}$, $\mathrm{df}_2 = \mathrm{df}_{2,2}$, and the endpoints $\mathrm{df}_{0,0} = d$, $B_{0,0} = \|\beta^\star\|^2$. For $p = q$ the weight $(s_k/(s_k+\kappa))^p \in (0,1)$ and $\mathrm{df}_{p,p}$ is a soft count of the modes above the resolution κ; $B_{p,p}$ is the teacher power in those modes. The off-diagonal sums are band-limited counts: $s_k/(s_k+\kappa)^2$ peaks at $s_k = \kappa$ and $s_k^2/(s_k+\kappa)^3$ at $s_k = 2\kappa$, each with constant width on a logarithmic axis.

Inequalities. Since $s_k/(s_k+\kappa) < 1$,

$$B_{p,q} \geq B_{p+1,q+1}, \qquad \mathrm{df}_{p,q} \geq \mathrm{df}_{p+1,q+1}, \qquad \text{in particular} \quad \|\beta^\star\|^2 = B_{0,0} \geq B_{1,1} \geq B_{2,2}, \tag{119}$$

which contains $d \geq \mathrm{df}_1 \geq \mathrm{df}_2$ and $\mathrm{df}_{0,1} \geq \mathrm{df}_{1,2} \geq \mathrm{df}_{2,3}$ as special cases; also $\min(n,d) > \mathrm{df}_2$.

Recursion. From $s_k = (s_k+\kappa) - \kappa$,

$$\mathrm{df}_{p,q} = \mathrm{df}_{p+1,q+1} + \kappa\,\mathrm{df}_{p,q+1}, \qquad B_{p,q} = B_{p+1,q+1} + \kappa\,B_{p,q+1}, \tag{120}$$

in particular $\mathrm{df}_1 - \mathrm{df}_2 = \kappa\,\mathrm{df}_{1,2}$ and $B_{1,1} - B_{2,2} = \kappa\,B_{1,2}$.

Derivatives.

$$\partial_\kappa \mathrm{df}_{p,q} = -q\,\mathrm{df}_{p,q+1}, \qquad \partial_\kappa\big(\kappa^r \mathrm{df}_{p,q}\big) = r\kappa^{r-1}\mathrm{df}_{p+1,q+1} + (r-q)\kappa^r \mathrm{df}_{p,q+1}, \tag{121}$$

and identically for $B_{p,q}$. The second identity follows from the first and the recursion. The cases used in the tuning proofs are $\partial_\kappa(\kappa^2 B_{1,2}) = 2\kappa B_{2,3}$ and $\partial_\kappa \mathrm{df}_2 = -2\,\mathrm{df}_{2,3}$.