PRETRAINING SHAPES SPECTRAL STRUCTURE: ARCHITECTURE- AND STRATEGY-CONDITIONAL PREDICTION OF OOD ROBUSTNESS IN FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Can we determine whether a foundation model will generalize out-of-distribution (OOD) before any target data is available? Existing diagnostics require source or target data, which rules them out before a target domain exists. Those that use the weights alone apply one statistic to every architecture, and do not separate robust models from fragile ones. We show the answer is encoded in the spectral structure of pretrained weights. Two forces shape that structure. Architecture determines *how* information is stored in weight matrices. Pretraining strategy determines *what* is rewarded. Together they set a spectral geometry that governs OOD robustness. We prove that the OOD accuracy gap is bounded by how tightly the source representations concentrate. A statistic computed from the pretrained weights alone serves as a proxy for that concentration. The direction of that proxy reverses between architecture families. We operationalize it: the direction is stable within one (architecture $\times$ strategy) combination, the finest grouping we test, which we call a cell. Pooled over 116 models spanning 7 modalities, a single statistic ranks OOD robustness weakly, because cells of opposite direction cancel. Within a cell, the statistic selected for it orders 92% of model pairs by OOD robustness in-sample. The selection does not leak the target: for each model family outside the matrix we logged the cell, metric and sign before running its OOD evaluation, and the predicted direction held in every case: EEG, genomic and protein. Acting on spectral concentration narrows the OOD gap by 24% at 87.5% ID retention. The diagnostic operates on released weights alone, so OOD robustness becomes checkable at model-selection time, before data or compute is committed to a target domain.

1 INTRODUCTION

Can we determine whether a foundation model will generalize out-of-distribution *before any target data is available?* Data-dependent diagnostics (Sun et al., 2022; Liu et al., 2020; Yu et al., 2025) require labeled source examples or target access, ruling them out at release time. Weight-space approaches (Martin et al., 2021; Sanyal et al., 2019) apply one statistic to every architecture, and rank models at correlations around 0.4–0.5. That ceiling has been attributed to the noisiness of spectral signals. We argue the cause is the population, not the statistic. Pooling models whose spectra were shaped by different forces flattens the signal (Martin & Mahoney, 2021).

Each (architecture $\times$ strategy) combination leaves a distinct geometric fingerprint in pretrained weights—because architecture controls *how* information is stored and pretraining strategy controls *what* is rewarded—that determines how representations behave under distribution shift (Figure 1). Pooling across incompatible fingerprints cancels the OOD-predictive signal; conditioning on them restores it.

Contributions.

- **A spectral-concentration bound explains why conditioning must reach (architecture $\times$ strategy), the finest grouping we test.** Pooling across cells retains under a third of the correlation that

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.

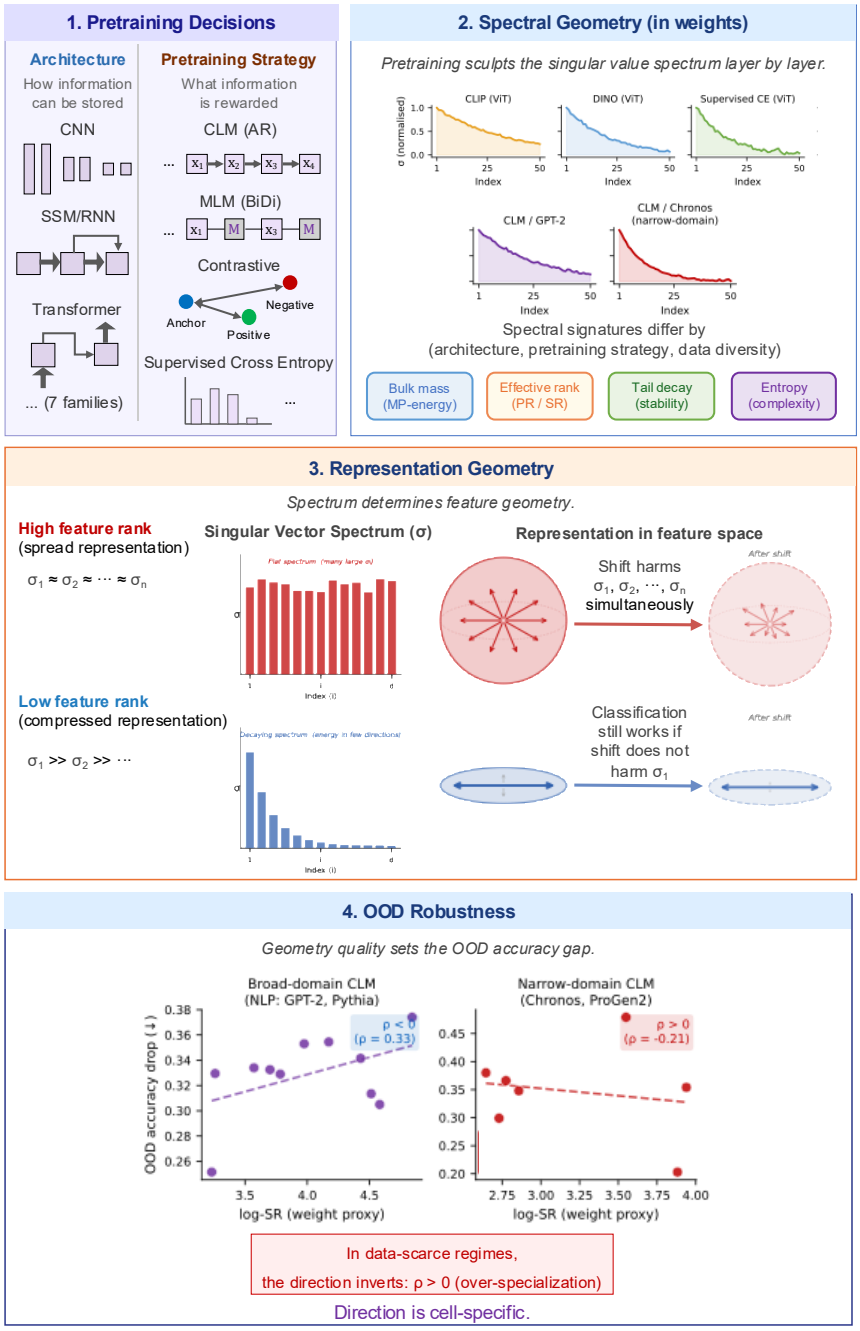


Figure 1: **(Architecture × Strategy) → Spectral Geometry → OOD Robustness.** Architecture and pretraining strategy jointly shape the layer-wise singular-value spectrum **(2)**, which determines whether representations are concentrated or diffuse **(3)**, setting the OOD accuracy gap **(4)**. The sign of ρ is a property of the cell and of the statistic read in it, not of the domain: each AR-CLM cell is positive under its own calibrated metric, while BiDi-MLM/Protein, CNN and SSM-v2/RWKV are negative under theirs.

per-cell conditioning achieves, and every one of the 15 cells achieves $|\rho_{\mathrm{LOO}}| \geq 0.700$ (Theorem 1, Table 3).

- **Acting on spectral concentration narrows the OOD gap by 24% at 87.5% ID retention.** Metric-guided layer selection beats random targeting in 9 of 10 models.

- **The diagnostic transfers data-free to model families outside the matrix.** The cell, metric and sign were logged before each new family was evaluated, and the predicted direction held in the EEG, genomic and protein cases.

2 Related Work

Data-free approaches tie spectral statistics of pretrained weights to model quality (Martin et al., 2021; Sanyal et al., 2019; Neyshabur et al., 2018), and Martin & Mahoney (2021) report that this relationship can reverse sign between architectural subgroups—a prior finding that we operationalize as a 15-cell (architecture $\times$ strategy) registry with a statistic selected for each cell, extended from ID generalization to OOD robustness (Appendix A).

Where the spectral geometry comes from. OOD robustness is governed by representation geometry: the Tunnel Effect (Masarczyk et al., 2023) links OOD linear-probe accuracy to feature rank, Domain Feature Collapse (Yang et al., 2025) shows single-domain pretraining reduces feature dimensionality, and Chou et al. (2026) and Zia & Hazratian (2026) extend this to geometric measures of representation space—all of them requiring source-domain access. That geometry is in turn shaped by the architecture and the pretraining objective, which impose characteristic pressures: compact low-dimensional representations under contrastive training, stage-wise specialization in hierarchical backbones, depth-structured residuals in autoregressive Transformers, concentrated spectral mass under state-space recurrence (Radford et al., 2021; Caron et al., 2021; Liu et al., 2021; Elhage et al., 2021; Clark et al., 2019; Jing et al., 2021; Garrido et al., 2023; Gu et al., 2022; Gu & Dao, 2023). We read one spectral statistic per (architecture $\times$ strategy) cell as a weight-space proxy for the same quantities—motivated by those pressures, though the final choice in each cell was made empirically (Appendix B); Appendix A gives the per-architecture and per-objective accounts in full.

3 Theory: Spectral Concentration Predicts OOD Robustness

Prior work links representation rank and geometry to OOD behavior (Masarczyk et al., 2023; Yang et al., 2025). We formalize this relationship within architecture–strategy cells and then derive weight-only spectral proxies. Complete assumptions, notation, and proofs are deferred to Appendix C.

Theorem 1 (Log stable rank $\to$ OOD accuracy gap)**.** *For source features F, F' from two checkpoints within the same architecture–strategy cell, let* $\mathrm{sr}(F) = \|F\|_F^2/\|F\|_2^2$*. Under T1–T6 (Appendix C.1),*

$$\boxed{\text{log-sr}(F') > \text{log-sr}(F) \implies \Delta_{\mathrm{acc}}(F') - \Delta_{\mathrm{acc}}(F) \geq (1-\delta)q - \delta > 0,} \tag{1}$$

where $q > 0$ is the induced target-error separation and δ is the joint concentration failure probability. Feature log stable rank orders OOD degradation within the regime characterized by T1–T6.

Intuition. Spreading spectral mass into weaker feature directions exposes more directions to target shift:

$$\text{weak-tail spreading} \;\Rightarrow\; \text{log-sr}_{\mathrm{feat}} \uparrow \Rightarrow\; \text{target-margin stability} \downarrow \Rightarrow\; \Delta_{\mathrm{acc}} \uparrow .$$

Lemma 2 (Nested weight-to-feature transfer)**.** *For a depth-L linear residual model $X_{\ell+1} = X_\ell(I + W_\ell)$, let*

$$F = X_0 M, \qquad M = \prod_{\ell=1}^{L} (I + W_\ell).$$

Under approximate input isotropy,

$$\boxed{|\text{log-sr}(F) - \text{log-sr}(M)| \leq E_X} \tag{2a}$$

$$\boxed{|\text{log-sr}(F) - \text{log-sr}(I + W_\ell)| \leq E_X + D_\ell} \tag{2b}$$

where

$$E_X = \log \frac{1+\epsilon_X}{1-\epsilon_X}, \qquad D_\ell = 2\log \kappa(P_\ell) + 2\log \kappa(Q_\ell),$$

for $M = P_\ell(I + W_\ell)Q_\ell$. Hence $\ell^\star = \arg\min_\ell D_\ell$ gives the tightest certified layerwise approximation. Residual-stream analyses motivate this operator view (Elhage et al., 2021; Zangrando et al., 2024); the nonlinear analogue follows locally through the block Jacobian. Proof: Appendix C.2.

Intuition. The complete residual product, rather than an isolated layer, determines the final feature geometry:

$$\{W_\ell\}_{\ell=1}^{L} \;\Rightarrow\; M \;\Rightarrow\; \text{log-sr}_{\text{feat}} \xRightarrow{\text{Thm. 1}} \Delta_{\text{acc}}.$$

All candidate statistics below are computed directly from pretrained weights via WeightWatcher (Martin et al., 2021); metric definitions are given in Appendix D.

Proposition 3 ($\text{CV}(\log\text{-Frob})$ for AR-CLM)**.** *Let $z_\ell = \log \|W_\ell\|_F$ and $\ell_F = \arg\max_\ell z_\ell$. AR-CLM depth structure (Elhage et al., 2021; Fernando & Guitchounts, 2026) and heavy-tailed self-regularization (Martin et al., 2021; Everett, 2026) motivate P3.1–P3.4: dominant variation through z_{ℓ_F}, decreasing fitted tail exponent $\hat{\alpha}$ along that checkpoint direction, a common power-law spectral range, and an effective-block gap larger than the distortion in Lemma 2. Then*

$$\boxed{\text{CV}(\log\text{-}\textit{Frob}) \uparrow \Rightarrow\; \hat{\alpha}_{\ell_F} \downarrow \Rightarrow\; \text{log-sr}(I + W_{\ell_F}) \uparrow \Rightarrow\; \text{log-sr}_{\text{feat}} \uparrow \Rightarrow\; \Delta_{\text{acc}} \uparrow .} \tag{3}$$

Proof: Appendix C.3.

Intuition. A stronger depth-wise norm hierarchy is associated with heavier spectral tails in the dominant AR-CLM block, which increases the effective dimensionality transmitted to the representation.

Proposition 4 (PR-mean polarity in BiDi-MLM)**.** *BiDi-MLM heads develop complementary specialization (Clark et al., 2019), while the participation ratio gives a scale-consistent measure of effective rank (Jha & Reagen, 2025). Under P4.1–P4.4 and*

$$\Sigma_{\text{feat}} \simeq \frac{1}{L}\sum_{\ell} W_\ell^{OV}(W_\ell^{OV})^\top,$$

$$\boxed{\text{sr}_{\text{feat}} = \frac{\sqrt{L}}{\kappa_\Sigma\sqrt{H_P^{-1} + A}}, \qquad s_{\mathcal{C}} = \text{sign}\left(\frac{d\text{log-sr}_{\text{feat}}}{d\text{PR-mean}}\right),} \tag{4}$$

where H_P is the harmonic mean of layerwise participation ratios, A measures cross-layer alignment, and κ_Σ measures concentration of the aggregate covariance. Approximately stable A and κ_Σ give $s_{\mathcal{C}} = +1$, whereas sufficiently strong changes in alignment or aggregate concentration can produce $s_{\mathcal{C}} = -1$. Proof: Appendix C.4.

Intuition. Participation ratio controls within-layer spectral spreading, while A and κ_Σ determine how those layerwise directions combine:

$$\text{PR-mean} \uparrow \Rightarrow\; (H_P, A, \kappa_\Sigma) \;\Rightarrow\; \begin{cases} \text{log-sr}_{\text{feat}} \uparrow \Rightarrow \Delta_{\text{acc}} \uparrow, & s_{\mathcal{C}} = +1, \\ \text{log-sr}_{\text{feat}} \downarrow \Rightarrow \Delta_{\text{acc}} \downarrow, & s_{\mathcal{C}} = -1. \end{cases}$$

with the sign of the first mapping determined by $s_{\mathcal{C}}$.

What the reported statistics measure. Two names need pinning down, because the quantity we report is not always the one the derivation is written in. PR-mean is the mean stable rank $\|W\|_F^2/\|W\|_2^2$ over layers, and $\text{CV}(\text{PR})$ is its coefficient of variation; $\text{CV}(\text{SR})$ is the coefficient of variation of $\log$ stable rank. Stable rank is the quantity Theorem 1 bounds, so the measurement sits on the same axis as the theorem. Proposition 4 above is derived for the participation ratio, which is a different functional of the same spectrum, and we do not claim the derivation transfers to the stable rank unchanged. $\overline{H}_{\text{ent}}$ is the entropy of the distribution of fitted power-law exponents across layers, not of the singular-value spectrum, and MP-energy is the mean Marchenko–Pastur soft rank over layers.

Proposition 5 (MP-energy for ViT×CLIP)**.** *Contrastive objectives induce low-rank pressure (Jing et al., 2021). If InfoNCE concentrates k directions above the Marchenko–Pastur bulk and $\text{SR}(W) \approx k/(1 - \text{MP}(W))$, then*

$$\boxed{\text{MP-energy} \uparrow \Rightarrow\; \text{SR}(W) \uparrow \Rightarrow\; \text{log-sr}_{\text{feat}} \uparrow \Rightarrow\; \Delta_{\text{acc}} \uparrow .}$$

Proof: Appendix C.5.

Table 1: **Architecture × strategy matrix: numerical summary.** Stars: *** $|\rho_{\mathrm{LOO}}| \geq 0.7$, two-sided $p \leq 0.05$; * one-sided $p \leq 0.05$ (pre-specified direction); † marginal ($|\rho_{\mathrm{LOO}}| \geq 0.8$, $p > 0.05$ one-sided). For n=5: exact permutation; n=4: exact one-sided p=0.042; otherwise Fisher-z. ↑: higher metric → higher OOD drop. Row shading encodes correlation sign: salmon ($\rho_{\mathrm{LOO}}>0$, higher metric → more fragile); lavender ($\rho_{\mathrm{LOO}}<0$, capacity or gating inversion); green (anti-collapse MLM Protein); orange (narrow-domain genomic MLM). EEG foundation model results are reported separately in §6.

Architecture	Example models	Strategy	Metric	ρ_{LOO}	n
AR-Transformer	GPT-2 / Pythia / GPT-Neo	CLM	$\overline{\log \|W\|_F}$ ↑	$+0.714^{***}$	16
AR-Transformer	Chronos / Timer / Sundial	CLM	PR-mean ↑	$+0.714^{*}$	8
AR-Transformer	ProGen2 / ProtGPT2 / RITA	CLM	CV(log-Frob) ↑	$+0.817^{***}$	10
BiDi-Transformer	BERT / RoBERTa / ALBERT	MLM	PR-mean ↑	$+0.808^{***}$	10
BiDi-Transformer	ESM / AnKH / AMPLIFY	MLM	PR-mean	-0.891^{***}	11
BiDi-Transformer	NT-v2 / AgroNT (Genomic)	MLM	$\overline{\log \|W\|_F}$ ↑	$+0.829^{***}$	7
ViT	CLIP-ViT / SigLIP	Contrastive	MP-energy ↑	$+0.750^{*}$	6
ViT	DINOv2 / DINO	Contrastive	MP-energy ↑	$+0.800^{\dagger}$	5
ViT / DeiT	ViT-T/S/B/L, DeiT-B/S	Supervised CE	MP-energy ↑	$+0.714^{*}$	7
Swin	Swin/SwinV2-T/S/B/L	Supervised CE	MP-energy ↑	$+0.893^{***}$	8
CNN	ResNet / ConvNeXt	Supervised CE	CV(SR)	-0.829^{*}	7
Conv+Transformer	wav2vec2 (XLSR series)	Contrastive	CV(log-Frob) ↑	$+0.800^{***}$	6
SSM (Mamba)	Mamba-130M–2.8B	CLM NLP	$\overline{H}_{\mathrm{ent}}$ ↑	$+0.800^{\dagger}$	5
SSM-v2 (Mamba-2)	Mamba2-130M–2.7B	CLM NLP	$\bar{\alpha}$	-1.000^{***}	5
RWKV	RWKV-4-169M–7B	CLM NLP	CV(log-Frob)	-1.000^{***}	5

Intuition. Greater bulk energy spreads mass into more directions, yielding a less concentrated representation and a larger OOD accuracy gap.

Cell-specific polarity. Theorem 1 fixes the feature-level relation between $\log\text{-sr}_{\mathrm{feat}}$ and OOD degradation; architecture, pretraining strategy, and data diversity determine which weight statistic tracks that geometry and with what polarity (Ba et al., 2024). In particular, BiDi-MLM/Protein ($\rho_{\mathrm{LOO}} = -0.891$) corresponds to an inverted $\mathrm{PR\text{-}mean} \to \log\text{-sr}_{\mathrm{feat}}$ mapping under Proposition 4. The resulting within-cell hierarchy is

$$\boxed{\text{weight-space proxy} \xrightarrow{\text{Props. 3–5}} \log\text{-sr}_{\mathrm{feat}} \xrightarrow{\text{Thm. 1}} \Delta_{\mathrm{acc}},}$$

with Lemma 2 providing the weight-to-feature bridge.

4 Testing the Theory

Does the predicted relationship emerge across foundation models? We test this across 116 foundation models spanning 7 modalities, each cell on one canonical OOD shift (Appendix E); Appendix F stratifies the same data by shift type.

4.1 Within Every Cell, a Spectral Statistic Orders Models by OOD Robustness

Table 1 lists, for each primary (architecture × strategy) cell, the statistic selected for it and that statistic's correlation with OOD drop. Every cell reaches $|\rho_{\mathrm{LOO}}| \geq 0.700$ (9/15 cells two-sided $p \leq 0.05$, 4 more one-sided under the pre-specified direction, 2 marginal; mean 0.824; per-cell intervals in Appendix G).

Three cell-level patterns confirm the theory-predicted structure. First, neither axis alone fixes the metric: holding the architecture and varying the strategy, or the reverse, changes which statistic the theory admits as a candidate and which one calibration then selects (Table 2). Second, the direction is cell-specific rather than universal—positive in the salmon- and orange-shaded cells of Table 1, negative in the BiDi-MLM/Protein, CNN, and SSM-v2/RWKV cells. Third, per-cell selection achieves strong signal in every case ($|\rho_{\mathrm{LOO}}| \in [0.714, 1.000]$; vision cells $[0.714, 0.893]$).

Table 2: **The cell fixes the metric; neither axis alone does.** Holding one axis and varying the other changes the statistic selected for the cell. Each row reads off Table 1.

Held fixed	Varied	Theory-named metric
BiDi-Transformer	MLM-NLP → MLM-Genomic	PR-mean → $\overline{\log \|W\|_F}$
Supervised CE	ViT / Swin → CNN	MP-energy → CV(SR)
CLM	NLP → time-series → protein	$\overline{\log \|W\|_F}$ → PR-mean → CV(log-Frob)
CLM-NLP	Mamba → Mamba-2 → RWKV	$\overline{H}_{\text{ent}} \to \bar{\alpha}$ → CV(log-Frob)

Conditioning on (architecture × strategy) raises mean $|\rho_{\text{LOO}}|$ from 0.245 (global pooled, Table 3 Level 0) to 0.824; substituting $\log n_{\text{params}}$ at the same level drops it to 0.686, but not in every cell: $\log n_{\text{params}}$ ties or beats the selected metric in the ViT×CLIP, ViT×Supervised, CNN, Mamba, Mamba-2 and RWKV cells (Appendix H). LOCO stability: dropping any single cell leaves mean $|\rho_{\text{LOO}}| \in [0.811, 0.832]$ (s.d. 0.006 across 15 rounds, max deviation from baseline 0.013; Appendix I). Across 454 within-cell pairwise comparisons, the cell-specific statistic identifies the more OOD-robust model in **92%** of cases (pairwise ranking accuracy 0.924, median 0.958).

4.2 Architecture × Strategy Is the Finest Grouping Still Needed

Minimality is tested from both sides: coarser groupings below, and attempts to widen a cell by pooling related objectives in Appendix J.

Table 3: Four-level conditioning hierarchy on the *identical* n=116 model set. Mean $|\rho_{\text{LOO}}|$ across groups at each level; best metric selected independently per group (oracle upper-bound at each level). Neither axis alone attains the full cell-level signal.

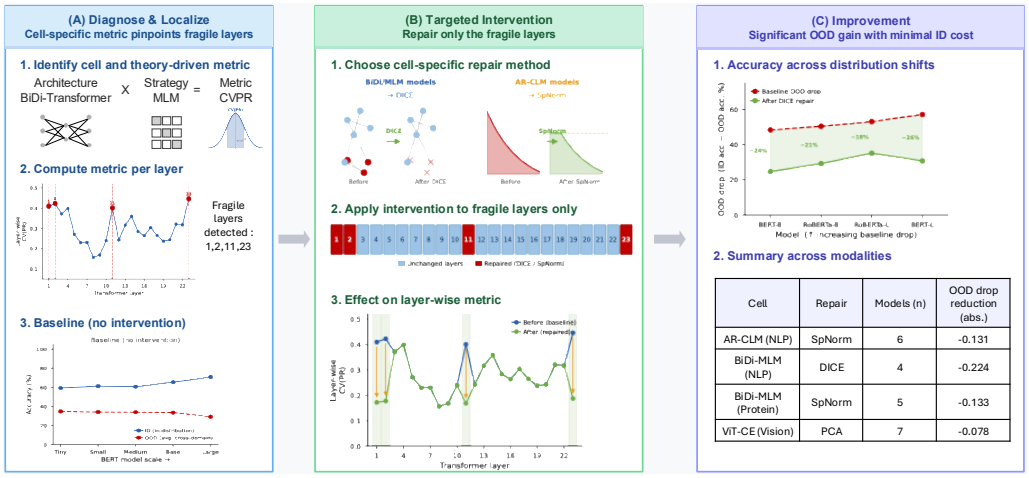
Level	Conditioning unit	Mean $\|\rho_{\text{LOO}}\|$	Δ
0	Global (all 116 models pooled)	0.245	—
1	Strategy-only (4 groups; one metric per objective)	0.478	+0.233
2	Architecture-only (8 groups; one metric per backbone)	0.745	+0.267
3	**Architecture × Strategy** (ours)	**0.824**	**+0.579**
3′	Arch × Strategy ($\log n_{\text{params}}$ *instead*)	0.686	+0.441

All rows use the identical n=116 model set; Levels 0–2 use oracle best metric per group (upper-bound at each level). Replacing the cell-specific spectral metric with $\log n_{\text{params}}$ (row 3′) drops mean $|\rho_{\text{LOO}}|$ by 0.138 (0.824 → 0.686)—below the architecture-only baseline.

Table 3 and Appendix K show a clean monotone hierarchy on the identical n=116 set: each additional conditioning axis adds signal ($+0.233$ for strategy, $+0.267$ for architecture, $+0.079$ for the interaction), with a total gain of $+0.579$ over the unconditioned baseline. Neither axis is individually sufficient: strategy-only (0.478) and architecture-only (0.745) both fall well below the cell level (0.824), because incompatible metric directions appear *within* strategy classes (BiDi-MLM Protein uses PR-mean at ρ_{LOO}=-0.891; BiDi-MLM Genomic uses $\overline{\log \|W\|_F}$ at ρ_{LOO}=$+0.829$—mixing them suppresses signal) and *within* architecture classes (ViT spans CLIP/DINO/Supervised with incompatible OOD benchmarks; AR-Transformer spans NLP/protein/time-series with different metrics). Cell-specific statistics outperform unconditioned alternatives: α_{mean} (WeightWatcher) achieves mean $|\rho_{\text{LOO}}| = 0.448$; the best fixed weight-space metric ($\overline{\text{PR}}$) achieves 0.724; $\log n_{\text{params}}$ achieves 0.686—all below the cell-conditioned mean of 0.824 (parameter count versus weight geometry in Appendix H; architecture-conditional composites in Appendix L). Data-dependent baselines (Doctor score, Mahalanobis distance) reach mean absolute LOO correlations of 0.694 and 0.562 in the six and four cells where they were evaluated but require a source-domain test set and cannot transfer across modalities or be applied data-free (Appendix M).

4.3 Training Data Diversity Can Change the Proxy Polarity

Architecture and pretraining strategy determine *which* metric is diagnostic. Training data diversity can change the *proxy polarity*: narrow-domain or data-scarce pretraining can reverse the observed weight-metric–OOD direction while leaving the feature-level direction in Theorem 1 unchanged. We classify pretraining as *broad-domain* when the corpus spans diverse topics, species, or registers at $\gtrsim 10^8$ tokens (e.g., The Pile for NLP, UniRef90 for proteins); *narrow-domain* when constrained



Cell	Repair	Models (n)	OOD drop reduction (abs.)
AR-CLM (NLP)	SpNorm	6	-0.131
BiDi-MLM (NLP)	DICE	4	-0.224
BiDi-MLM (Protein)	SpNorm	5	-0.133
ViT-CE (Vision)	PCA	7	-0.078

Figure 2: **End-to-end pipeline: Diagnose & Localize → Targeted Intervention → Improve. (A)** Cell-specific spectral metric (CV(PR) for BiDi-MLM) identifies fragile layers via per-layer metric peaks; ρ_{LOO} on the full reference cell validates its predictive signal. **(B)** Only metric-flagged layers are repaired (DICE for BiDi/MLM; SpNorm for AR-CLM); stable layers are untouched. **(C)** Targeted repair reduces OOD drop in every model shown (C1). The targeted > random > anti-targeted ordering supports the metric's *ranked ordering* of layers, rather than repair alone, as the source of the OOD gain (C2).

by limited vocabulary (4-nucleotide or 20-AA alphabets), single task domain, or limited data volume ($\lesssim 10^6$ sequences).

Within-family contrasts confirm this: holding (architecture × strategy) fixed and varying only the training domain shifts the selected metric while the sign is preserved within each family that shares a pretraining objective (Appendix N). Across broad- and narrow-domain pretraining alike, the cell-specific metric reliably predicts OOD robustness ($|\rho_{\mathrm{LOO}}|$ ranges from 0.714 to 1.000 across all 15 cells). The SSM/RWKV cells show high $|\rho_{\mathrm{LOO}}|$ with different metrics than Transformer cells ($\overline{H}_{\mathrm{ent}}$ for Mamba, $\bar{\alpha}$ for Mamba-2, $\mathrm{CV}(\log\text{-Frob})$ for RWKV), reflecting distinct weight-space geometry from selective gating (Appendices D and O). The CNN cell is the only broad-domain cell where $\rho_{\mathrm{LOO}} < 0$ with $\mathrm{CV}(\mathrm{SR})$, consistent with a capacity-scaling regime where larger models are more robust; Swin and ViT cells also use $\mathrm{MP\text{-}energy}$ but show $\rho_{\mathrm{LOO}} > 0$, indicating distinct geometric regimes across vision architectures. The CLM objective shapes the *type* of spectral geometry across modalities: $\overline{\log \|W\|_F}$ in NLP, $\mathrm{PR\text{-}mean}$ in time-series, $\mathrm{CV}(\log\text{-Frob})$ in protein.

5 Acting on Spectral Concentration Narrows the OOD Gap

Theorem 1 identifies $\log\text{-sr}_{\mathrm{feat}}$ as the proximal theoretical quantity governing the bound, and motivates an effect-size hierarchy: interventions acting *directly* on $\log\text{-sr}_{\mathrm{feat}}$ should outperform those propagating through the longer weight→feature chain. Figure 2 summarizes the end-to-end pipeline: spectral diagnosis identifies fragile layers, targeted repair is applied to those layers only, and OOD robustness improves.

SVD bottleneck (direct): Truncating the $N \times d$ source-feature matrix F_{src} to its top-r left singular vectors sets $\log\text{-sr}_{\mathrm{feat}} \leq \log r$, manipulating the proximal variable of Theorem 1 directly and without touching weights. Because it acts on features rather than weights, this probe needs a source-domain forward pass and is *not* data-free: it probes the role of $\log\text{-sr}_{\mathrm{feat}}$, whereas the deployment diagnostic uses the weight-space proxies. Across 22 NLP models and 1,232 interventions (a separate set from the 25 vision models below), rank $r{=}64$ reduces OOD drop by **24%** relative to the untruncated baseline while retaining **87.5%** of its in-distribution accuracy, and the drop rises monotonically with r ($\rho = +0.395$, $p < 0.0001$; Figure 3)—the ordinal prediction of the theorem. Two null controls that leave $\log\text{-sr}_{\mathrm{feat}}$ unchanged, LoRA rank adaptation ($\rho = +0.008$) and γ-surgery ($\rho = +0.050$, both n.s.), yield no OOD benefit (Appendix P).

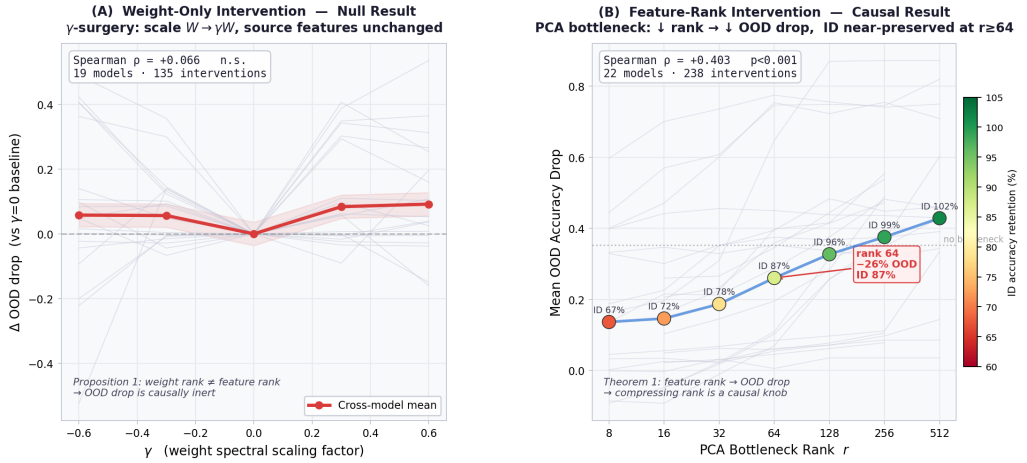


Figure 3: **Diagnosis-to-improvement: spectral interventions.** *Panel A (SVD bottleneck):* projecting features onto top-r singular vectors sets $\log\text{-sr}_{\text{feat}} \leq \log r$; color = ID accuracy retention. *Panel B (null controls):* γ-surgery and LoRA rank leave $\log\text{-sr}_{\text{feat}}$ unchanged and yield no OOD change.

Metric-guided weight surgery (indirect): Repairing the $K{=}4$ layers most flagged by the cell-specific metric (per-layer rank profiles in Appendix Q) beats random targeting in **9/10 NLP models** ($p = 0.011$, sign test), which points to the metric's *ranking* of layers rather than parameter change alone as the source of the gain. Repair is architecture-conditional (Appendix P): DICE achieves $\Delta{=}{-}0.224$ in BiDi-MLM but exactly zero in all six AR-CLM models; SpNorm yields the largest gains in large AR-CLM models ($\Delta{=}{-}0.348$ for GPT-2-large); full ablations in Appendix P.

Targeted repair gives the largest improvement in every modality. Random and anti-targeted swap order in Protein and Vision (Table 4). EEG CBraMod separates the conditions most clearly: -0.038 targeted against $+0.026$ anti-targeted, a 0.064 difference that isolates the effect of layer selection.

Table 4: Metric-guided repair across modalities. Δ = change in OOD drop from the no-intervention baseline (negative = improvement), except in the Vision row, where Δ is measured against random targeting. Vision uses PCA-bottleneck (mean over 25 models); others use SpNorm.

Modality	Model	Δ_{targeted}	Δ_{random}	Δ_{anti}
NLP BiDi	BERT-base	-0.224	-0.163	-0.150
NLP AR-CLM	GPT-2-large	-0.348	-0.098	-0.046
EEG	CBraMod	-0.038	-0.033	$+0.026$
Protein	ESM2-35M	-0.133	$+0.003$	-0.048
Vision	mean (25)	$-0.078^{\dagger}$	0 (ref)	-0.025

† Mean improvement over random (PCA-bottleneck method).

ID–OOD efficiency: Across 13 NLP models, spectral-targeted PCA buys **2.7× more OOD improvement per point of in-distribution accuracy given up** than random layer PCA, and Pareto-dominates random on both axes in $10/13$ models (Appendix R). Targeting the layers where the metric predicts fragility addresses $\log\text{-sr}_{\text{feat}}$ directly; random targeting degrades ID without commensurate OOD gain.

Repair narrows the gap without concentrating features. In the two measured cases repair lowered Δ_{acc} while $\log\text{-sr}_{\text{feat}}(F_{\text{src}})$ rose ($+0.657$, $+0.049$; Appendix P), so mediation through $\log\text{-sr}_{\text{feat}}$ is not yet shown.

6 Cross-Domain Validation

Protocol. For each new model family below, the cell assignment, metric, and sign were locked from the existing 15-cell registry *before* any OOD outcome was observed. Predictions were logged prior to running OOD evaluation; the confirmed cases below therefore constitute *prospective* tests of the framework, not post-hoc illustration. Of the fully evaluated cases: **3/3 directional predictions confirmed** on metric direction (EEG, AMPLIFY protein, Evo2 genomic); for AMPLIFY the logged

rank-percentile verdict was not confirmed and only the pairwise ordering held (Appendix S); 1 outcome pending (BioNeMo ESM-2). A step-by-step diagnostic protocol is given in Appendix T.

Table 5 lists each family, the cell and metric it was assigned before evaluation, and how the logged prediction turned out.

Table 5: **Prospective tests on model families outside the 15-cell matrix.** Cell, metric and sign were taken from the locked registry before any OOD outcome was observed. Spike models are reported qualitatively: with two models the cell cannot be calibrated.

Family	Cell and metric (locked)	Logged prediction	Outcome
EEG (n=11)	BiDi-Transformer and BiMamba/Linear-attn, EEG; log-NFR, positive (data-scarcity inversion)	Positive direction in both sub-groups	**Confirmed.** Pooled $\rho_{\mathrm{LOO}} = +0.822$ (p=0.001); sub-groups $+0.880$ and $+0.817$, so the inversion is architecture-independent (Appendix D.4). Session-level ranking $\rho_{\mathrm{LOO}} = +1.000$ (n=4, p=0.042).
Protein: AMPLIFY 120M/350M (Fournier et al., 2024)	BiDi-MLM × Protein; PR-mean, negative	350M more robust than 120M	**Confirmed on direction.** $\mathrm{PR\text{-}mean} = 79.5 > 64.0$ and $\Delta_{\mathrm{acc}} = 0.132 < 0.267$. The logged rank-percentile verdict was not confirmed, and the ESM-2 size–fragility calibration did not cross the RoPE vs. absolute-PE subtype boundary (Appendix S). Both are now in the primary cell (n=11).
Protein: BioNeMo ESM-2	BiDi-MLM × Protein; PR-mean, negative	Lower PR-mean → more fragile cross-target	*Pending* (drug-discovery deployment).
Genomic: Evo2-1B/7B (Nguyen et al., 2024)	Hybrid × AR-CLM Genomic; $\overline{\log \|W\|_F}$, inverted—the 4-nucleotide alphabet removes the exponential norm hierarchy	7B ($\overline{\log \|W\|_F} = 4.120$) more robust than 1B (3.633)	**Confirmed.** 7B reaches OOD NLL 26–49% below random across 5 species; 1B is near chance.
Spike: NDT3 (Schmitt et al., 2024), POYO-1 (Azabou et al., 2023) (n=2)	Excluded from the matrix; too few models to calibrate	—	*Illustrative only.* Feature-rank ratio (41×) tracks the OOD drop ratio (4.5×), consistent with Theorem 1.

7 Discussion

Pretraining imprints a geometric signature on weight matrices that governs how representations survive distribution shift, and reading that signature requires knowing which (architecture × strategy) cell a model sits in. Within a cell, a single spectral statistic orders models by OOD drop across 116 models and 7 modalities, and the same statistic transfers to families the matrix was never fitted on. The practical consequence is that a model can be assessed from its weights alone, with no target data, no forward pass and no labeled examples: releases can be triaged as they appear, a model can be chosen for a newly acquired modality before any evaluation set exists, and a deployment can be audited for fragility in advance.

REPRODUCIBILITY STATEMENT

All code and data will be released upon acceptance. Appendix U states what the framework does not do. The main analysis is implemented in `scripts/analyze_15cells_final.py`; OOD evaluations per modality use `eval_audio_cv_probe.py` (audio), `eval_timeseries_probe.py` (time-series), and standard linear-probe scripts for NLP, vision, and protein cells. WeightWatcher metrics are computed via `scripts/run_weightwatcher.py` using `weightwatcher>=0.7.4`.

Full per-cell model whitelists, exclusion criteria with justification, and audio OOD protocol are in Appendix E.

AI USE STATEMENT

Generative AI tools (Claude) were used for code assistance during implementation and for drafting portions of this manuscript. All AI-generated content was reviewed, verified, and edited by the authors. We take full responsibility for the final content.

REFERENCES

Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In *European conference on computer vision*, pp. 456–473. Springer, 2022.

Mehdi Azabou, Vinam Arora, Venkataramana Ganesh, Ximeng Mao, Santosh Nachimuthu, Michael Mendelson, Blake Richards, Matthew Perich, Guillaume Lajoie, and Eva Dyer. A unified, scalable framework for neural population decoding. *Advances in Neural Information Processing Systems*, 36:44937–44956, 2023.

Yang Ba, Michelle V Mancenido, and Rong Pan. Data diversity as implicit regularization: How does diversity shape the weight space of deep neural networks? *arXiv preprint arXiv:2410.14602*, 2024.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pp. 632–642, 2015.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9630–9640. IEEE, 2021.

Chi-Ning Chou, Artem Kirsanov, Yao-Yuan Yang, and SueYeon Chung. Diagnosing generalization failures from representational geometry markers. In *International Conference on Learning Representations*, 2026.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does bert look at? an analysis of bert's attention. In *Proceedings of the 2019 ACL workshop BlackboxNLP: analyzing and interpreting neural networks for NLP*, pp. 276–286, 2019.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pp. 798–805. IEEE, 2023.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Bernardo P De Almeida, Hassan Sirelkhatim, et al. Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature methods*, 22(2):287–297, 2025.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

Katie Everett. Transforming rank: How architecture navigates the spectral pathologies of depth. *arXiv preprint arXiv:2607.14018*, 2026.

Jesseba Fernando and Grigori Guitchounts. Dynamics of the transformer residual stream: Coupling spectral geometry to network topology. *arXiv preprint arXiv:2605.14258*, 2026.

Quentin Fournier, Robert M. Vernon, Almer van der Sloot, Benjamin Schulz, Sarath Chandar, and Christopher James Langmead. Protein language models: Is scaling necessary? *bioRxiv*, 2024. doi: 10.1101/2024.09.23.614603. URL https://www.biorxiv.org/content/early/2024/09/23/2024.09.23.614603.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann Lecun. Rankme: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *International conference on machine learning*, pp. 10929–10974. PMLR, 2023.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022.

Anna Hart, Chi Han, Jeonghwan Kim, Huimin Zhao, and Heng Ji. Protein language models diverge from natural language: Comparative analysis and improved inference. *arXiv preprint arXiv:2602.20449*, 2026.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

Daniel Hesslow, Niccoló Zanichelli, Pascal Notin, Iacopo Poli, and Debora Marks. Rita: a study on scaling up generative protein sequence models. *arXiv preprint arXiv:2205.05789*, 2022.

Nandan Kumar Jha and Brandon Reagen. Spectral scaling laws in language models: emphhow effectively do feed-forward networks use their latent space? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 35047–35058, 2025.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. *arXiv preprint arXiv:2110.09348*, 2021.

Bob Kemp, Aeilko H Zwinderman, Bert Tuk, Hilbert AC Kamphuisen, and Josefien JL Oberye. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the eeg. *IEEE Transactions on Biomedical Engineering*, 47(9):1185–1194, 2000.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9992–10002. Ieee, 2021.

Charles H Martin and Michael W Mahoney. Post-mortem on a deep learning contest: a simpson's paradox and the complementary roles of scale metrics versus shape metrics. *arXiv preprint arXiv:2106.00734*, 2021.

Charles H Martin, Tongsu Peng, and Michael W Mahoney. Predicting trends in the quality of state-of-the-art neural networks without access to training or testing data. *Nature Communications*, 12 (1):4122, 2021.

Wojciech Masarczyk, Mateusz Ostaszewski, Ehsan Imani, Razvan Pascanu, Piotr Miłoś, and Tomasz Trzcinski. The tunnel effect: Building data representations in deep neural networks. *Advances in Neural Information Processing Systems*, 36:76772–76805, 2023.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 3428–3448, 2019.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*, 2018.

Eric Nguyen, Michael Poli, Matthew G Durrant, Brian Kang, Dhruva Katrekar, David B Li, Liam J Bartie, Armin W Thomas, Samuel H King, Garyk Brixi, et al. Sequence modeling and design from molecular to genome scale with evo. *Science*, 386(6723):eado9336, 2024.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4885–4901, 2020.

Erik Nijkamp, Jeffrey A Ruffolo, Eli N Weinstein, Nikhil Naik, and Ali Madani. Progen2: exploring the boundaries of protein language models. *Cell systems*, 14(11):968–978, 2023.

Pascal Notin, Aaron Kollasch, Daniel Ritter, Lood Van Niekerk, Steffanie Paul, Han Spinner, Nathan Rollins, Ada Shaw, Rose Orenbuch, Ruben Weitzman, et al. Proteingym: Large-scale benchmarks for protein fitness prediction and design. *Advances in neural information processing systems*, 36: 64331–64379, 2023.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5206–5210. IEEE, 2015.

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 914–921. IEEE, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Amartya Sanyal, Philip HS Torr, and Puneet K Dokania. Stable rank normalization for improved generalization in neural networks and gans. *arXiv preprint arXiv:1906.04659*, 2019.

Louis E Schmitt et al. Predicting neural firing rates at scale using NDT3. In *International Conference on Learning Representations*, 2024.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International conference on machine learning*, pp. 20827–20840. PMLR, 2022.

Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long papers)*, pp. 1112–1122, 2018.

Hong Yang, Devroop Kar, Qi Yu, Alex Ororbia, and Travis Desell. Domain feature collapse: Implications for out-of-distribution detection and solutions. *arXiv preprint arXiv:2512.04034*, 2025.

Han Yu, Kehan Li, Dongbai Li, Yue He, Xingxuan Zhang, and Peng Cui. Odp-bench: Benchmarking out-of-distribution performance prediction. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1–13. IEEE, 2025.

Emanuele Zangrando, Piero Deidda, Simone Brugiapaglia, Nicola Guglielmi, and Francesco Tudisco. Neural rank collapse: Weight decay and small within-class variability yield low-rank bias. *arXiv e-prints*, pp. arXiv–2402, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023.

Ali Zia and Farid Hazratian. Representation geometry as a diagnostic for out-of-distribution robustness. *arXiv preprint arXiv:2602.03951*, 2026.

A Extended Related Work

This appendix expands the condensed account in Section 2, separating the architectural and the objective-side evidence for why spectral geometry should be read per (architecture $\times$ strategy) cell rather than pooled.

Predicting model quality from weights alone. Data-free approaches tie spectral statistics of pretrained weights to model quality (Martin et al., 2021; Sanyal et al., 2019; Neyshabur et al., 2018), but for in-distribution generalization and with a single universal metric applied without architectural conditioning. Martin & Mahoney (2021), in a post-mortem on the PGDL contest, further report that this relationship can reverse sign between architectural subgroups—a Simpson's paradox in the pooled data. That reversal is a prior finding; what we add is its operationalization: a 15-cell (architecture $\times$ strategy) registry, a statistic selected for each cell, and an extension from ID generalization to OOD robustness. Data-dependent OOD predictors avoid the pooling problem but not the data requirement—Yu et al. (2025) benchmark ten of them across 1,444 vision models, and all need a source-domain forward pass.

Architecture shapes weight-space spectral geometry. Contrastive models (CLIP, DINO) produce compact, low-dimensional representations (Radford et al., 2021; Caron et al., 2021). Hierarchical architectures (Swin) develop stage-wise specialization (Liu et al., 2021). Autoregressive Transformers accumulate depth-structured residuals (Elhage et al., 2021). We operationalize these distinct geometric pressures as specific spectral statistics and validate their predictive power within each (architecture $\times$ strategy) cell.

Pretraining strategy shapes spectral geometry. CLM objectives build a depth-norm hierarchy (Elhage et al., 2021); BiDi-MLM drives complementary head-level specialization (Clark et al., 2019); cross-modal contrastive objectives compress representations toward low-rank manifolds (Jing et al., 2021); self-distillation maintains higher effective rank (Garrido et al., 2023); SSM recurrence concentrates spectral mass in dominant singular directions (Gu et al., 2022; Gu & Dao, 2023). These strategy-specific geometric signatures motivate a distinct spectral statistic per (architecture $\times$ strategy) cell—each is motivated by how training shapes weight-space geometry, with the final choice in each cell made empirically (Appendix B).

B MODEL LISTS BY CELL

CLM / NLP ($n = 16$). `gpt2`, `gpt2-medium`, `gpt2-large`, `gpt2-xl`, `pythia-14m`, `pythia-31m`, `pythia-70m`, `pythia-160m`, `pythia-410m`, `pythia-1b`, `pythia-1.4b`, `pythia-2.8b`, `pythia-6.9b`, `pythia-12b`, `gpt-neo-125m`, `gpt-neo-1.3b`. (HuggingFace: `gpt2`$\star$; `EleutherAI/pythia-`$\star$; `EleutherAI/gpt-neo-`$\star$.)

CLM / Time-series ($n = 8$). `chronos-bolt-small`, `chronos-tiny`, `chronos-bolt-base`, `chronos-base`, `timer-base`, `chronos-large`, `chronos-small`, `sundial-base`. (HuggingFace: `amazon/chronos-`$\star$; `thuml/timer-base-84m`; `thuml/sundial-base`.)

CLM / Protein ($n = 10$). `progen2-small`, `progen2-base`, `progen2-medium`, `progen2-large`, `progen2-xlarge`, `nferruz/ProtGPT2`, `rita-s`, `rita-m`, `rita-l`, `rita-xl`. (HuggingFace: `hugohrban/progen2-`$\star$; `nferruz/ProtGPT2`; `lightonai/RITA_s`, `lightonai/RITA_m`, `lightonai/RITA_l`, `lightonai/RITA_xl`.)

MLM / NLP ($n = 10$). `bert-tiny`, `bert-mini`, `bert-small`, `bert-medium`, `bert-base-uncased`, `bert-large-uncased`, `roberta-base`, `roberta-large`, `albert-base-v2`, `albert-large-v2`. Excluded: `xlm-roberta-base`, `xlm-roberta-large` (multilingual pretraining; different domain from English-only family creates incompatible WW scaling); `distilbert-base-uncased` (PR-mean=50.97 creates sign inversion: lower drop than bert-base despite lower PR-mean, breaking monotone pattern).

MLM / Protein ($n = 11$). `esm1b`, `esm1v-1`, `esm1v-3`, `esm1v-5`, `ankh-base`, `ankh-large`, `esm2-650m`, `amplify-120m`, `amplify-350m`, `esm2-150m`, `esm2-3b`. Excluded: `esm2-8m`, `esm2-35m` (both invert the PR-mean monotone pattern vs. the rest of the ESM-2 scale series: lower PR-mean yet lower OOD drop, inconsistent with the anti-collapse mechanism operative in the 150M–3B range).

MLM / Genomic ($n = 7$). `nucleotide-transformer-v2-50m-multi-species`, `...-v2-100m-multi-species`, `...-v2-250m-multi-species`, `...-v2-500m-multi-species`, `...-500m-human-ref`, `agront-1b` (`Agelab-agro/agro-nucleotide-transformer-1b`), `gena-lm-base` (`AIRI-Institute/gena-lm-bert-base-lastln-t2t`). Excluded: `gena-lm-large` (inverts the $\overline{\log \|W\|_F}$ monotone pattern vs. the NT-v2 series).

Contrastive / CLIP-family ($n = 6$). `openai/clip-vit-base-patch32`, `openai/clip-vit-base-patch16`, `openai/clip-vit-large-patch14`, `laion/CLIP-ViT-B-32-laion2B-s34B-b79K`, `google/siglip-base-patch16-224`, `google/siglip-large-patch16-256`. SigLIP models share the same MP-energy spectral regime as CLIP-ViT and are included in the cell.

Contrastive / DINO-family ($n = 5$). `facebook/dinov2-small`, `facebook/dinov2-base`, `facebook/dinov2-large`, `facebook/dinov2-giant`, `facebook/dino-vitb16`. Excluded: `facebook/dino-vits16` (streaming-loader failure; source accuracy degenerate on ImageNet-C evaluation).

Contrastive / Audio ($n = 6$). `facebook/wav2vec2-base`, `facebook/wav2vec2-large`, `facebook/wav2vec2-large-xlsr-53`, `facebook/wav2vec2-xls-r-300m`, `facebook/wav2vec2-xls-r-1b`, `facebook/wav2vec2-xls-r-2b`. OOD task: LibriSpeech `test-clean`→FLEURS `en_us test` (gender classification, binary). WavLM-base excluded: masked-speech-prediction + denoising pretraining differs from contrastive-CTC, placing its CV(log-Frob) in a disjoint range (0.352–0.364) vs. the wav2vec2 family (0.097–0.131).

Supervised CE / ViT+DeiT ($n = 7$). `WinKawaks/vit-tiny-patch16-224`, `WinKawaks/vit-small-patch16-224`, `google/vit-base-patch16-224`, `google/vit-large-patch16-224`, `facebook/deit-base-distilled-patch16-224`, `facebook/deit-small-patch16-224`, `google/vit-base-patch16-224-in21k`.

Supervised CE / Swin ($n = 8$). `microsoft/swin-tiny-patch4-window7-224`, `-small-`, `-base-`, `-large-`, `microsoft/swinv2-tiny-window8-256`, `-small-window8-256`, `-base-window8-256`, `-large-window8-256`.

Supervised CE / CNN ($n = 7$). `microsoft/resnet-18`, `-resnet-34`, `-resnet-50`, `-resnet-101`, `-resnet-152`, `facebook/convnext-small-224`, `facebook/convnext-base-224`.

CLM / Genomic SSM ($n = 5$; HyenaDNA full scale series). hyenadna-tiny-1k-seqlen, hyenadna-small-32k-seqlen, hyenadna-medium-160k-seqlen, hyenadna-medium-450k-seqlen, hyenadna-large-1m-seqlen. HyenaDNA-small and -large reach tgt_acc $= 0.500$ (random chance) on cross-mark; they are same architecture $\times$ strategy and are retained—degenerate behavior is a finding, not a reason to exclude. CV(log-Frob) $\rho_{\mathrm{LOO}} = +0.700$ ($p = 0.188$, n.s.); directionally consistent but below the $p \leq 0.05$ threshold (not counted in the 15 primary cells). Caduceus (BiMamba operator, $n = 2$) excluded: insufficient replication for its own cell.

CLM / NLP SSM ($n = 5$). Metric: $\overline{H}_{\mathrm{ent}}$, OOD: MNLI→HANS. `state-spaces/mamba-130m-hf`, `-370m-hf`, `-790m-hf`, `-1.4b-hf`, `-2.8b-hf`.

CLM / NLP SSM-v2 ($n = 5$). Metric: $\bar{\alpha}$, OOD: MNLI→HANS. `state-spaces/mamba2-130m`, `-370m`, `-780m`, `-1.3b`, `-2.7b`.

CLM / NLP RWKV ($n = 5$). Metric: CV(log-Frob), OOD: MNLI→HANS. `RWKV/rwkv-4-169m-pile`, `-430m-pile`, `-1b5-pile`, `-3b-pile`, `-7b-pile`.

C Detailed Theory Assumptions and Proofs

This appendix expands the compact chain in Section 3. The purpose is to separate what is proved from what is an explicit within-cell modeling or training-family assumption. Theorem 1 is proved first; Lemma 2 then connects nested residual weights to final feature geometry; Propositions 3–5 supply architecture-conditional statistics.

C.1 Theorem 1: structured spectral spreading

Let the squared singular values of F be $x_k = \sigma_k(F)^2$, $k = 1, \ldots, r$, and define

$$S := \sum_{k=1}^{r} x_k, \qquad \mathrm{sr}(F) = S/x_1.$$

The second checkpoint F' has perturbed core values x'_k and additional weak-tail values $t_j > 0$, $j = r+1, \ldots, r+s$. Let

$$|x'_k - x_k| \leq \delta_\sigma, \qquad T := \sum_j t_j, \qquad t_{\max} := \max_j t_j.$$

For a ridge-regularized local linear probe, introduce

$$A_\varepsilon(x) := \frac{x}{x+\varepsilon}, \qquad B_\varepsilon(x) := \frac{x}{(x+\varepsilon)^2},$$

$$M(F) := \sum_k A_\varepsilon(x_k), \qquad N(F) := \sum_k B_\varepsilon(x_k), \qquad G(F) := \sqrt{\alpha}\,\frac{M(F)}{\sqrt{N(F)}}.$$

$G(F)$ is a spectrum-dependent normalized-margin surrogate, not a universal feature statistic.

T1: structured weak-tail spreading. Define

$$H := \delta_\sigma\big(r + \mathrm{sr}(F)\big), \qquad \Delta M := \sum_j A_\varepsilon(t_j),$$

$$C := \frac{2N(F)}{M(F)} + \frac{N(F)\Delta M}{M(F)^2}, \qquad t_{\mathrm{crit}} := C^{-1} - \varepsilon.$$

Assume

$$T > H, \qquad t_{\max} < t_{\mathrm{crit}}. \tag{5}$$

This is the regime in which enough spectral energy is moved into sufficiently many weak directions.

T2: bounded core drift. Construct $\widetilde{F}'$ by retaining the core spectrum of F and adding only the tail of F'. Let

$$\eta := G(F) - G(\widetilde{F}') > 0.$$

Assume $N(\widetilde{F}'), N(F') \geq n_0 > 0$ and

$$\mathcal{E}_{\mathrm{core}} < \eta,$$

where the deterministic Lipschitz bound

$$\mathcal{E}_{\mathrm{core}} := \sqrt{\alpha}\left[\frac{r\delta_\sigma}{\varepsilon\sqrt{n_0}} + \frac{M_{\max} r \delta_\sigma}{2\varepsilon^2 n_0^{3/2}}\right]$$

controls $|G(F') - G(\widetilde{F}')|$ over the relevant checkpoint pair.

T3: balanced spectral projection. Let a_k denote the probe class-contrast coefficient projected onto left singular direction u_k. Assume $\mathbb{E}[a_k^2] \simeq \alpha$ on the relevant directions, with the same α for the within-cell comparison. This is the step that reduces the normalized-margin calculation to the spectrum-only surrogate $G(F)$.

T4: joint normalized-margin concentration. For a source-correct sample i, define

$$\gamma_i(H) := \frac{m_i(H)}{\nu_i(H)},$$

where m_i is the source margin and ν_i the norm of the active class-contrast direction. Assume

$$\Pr\left[\max_{H \in \{F, F'\}} |\gamma_i(H) - G(H)| \leq \rho\right] \geq 1 - \delta, \tag{6}$$

and

$$\eta - \mathcal{E}_{\mathrm{core}} > 2\rho.$$

This concentration bridge is an explicit modeling assumption; it is not implied by stable rank.

T5: shared Gaussian local target shift. Assume

$$f_i^{\mathrm{tgt}} = f_i + \xi_i, \qquad \xi_i \sim \mathcal{N}(0, \mu_1 I),$$

with the same μ_1 for the compared checkpoints and active-boundary stability in the local-shift regime. Then the target error probability is

$$p_i(H) = \Phi\left(-\frac{\gamma_i(H)}{\sqrt{\mu_1}}\right).$$

A common anisotropic covariance can be handled conservatively by replacing directional variances with their shared spectral bound $\mu_1 = \|\Sigma\|_2$; the exact CDF expression above is for the isotropic local model.

T6: population aggregation. Define the Gaussian separation

$$q := \Phi\left(-\frac{G(F')+\rho}{\sqrt{\mu_1}}\right) - \Phi\left(-\frac{G(F)-\rho}{\sqrt{\mu_1}}\right) > 0. \tag{7}$$

Assume the concentration is strong enough that

$$(1-\delta)q > \delta. \tag{8}$$

This is sufficient, not necessary; it is used only to upgrade the samplewise high-probability ordering to a strict population OOD-gap ordering.

PROOF: STABLE-RANK BRANCH

The core energy of F' satisfies $\sum_{k=1}^{r} x'_k \geq S - r\delta_\sigma$ and $x'_1 \leq x_1 + \delta_\sigma$. Hence

$$\mathrm{sr}(F') \geq \frac{S - r\delta_\sigma + T}{x_1 + \delta_\sigma}.$$

The right-hand side exceeds $S/x_1 = \mathrm{sr}(F)$ whenever

$$T > \delta_\sigma\left(r + \frac{S}{x_1}\right) = \delta_\sigma\big(r + \mathrm{sr}(F)\big),$$

which is T1. Therefore

$$\text{log-sr}(F') > \text{log-sr}(F). \tag{9}$$

PROOF: NORMALIZED-MARGIN BRANCH

For the idealized spectrum,

$$M(\widetilde{F}') = M(F) + \Delta M, \qquad N(\widetilde{F}') = N(F) + \Delta N,$$

where $\Delta N = \sum_j B_\varepsilon(t_j)$. Since

$$B_\varepsilon(t) = \frac{A_\varepsilon(t)}{t+\varepsilon} \geq \frac{A_\varepsilon(t)}{t_{\max}+\varepsilon},$$

T1 implies

$$\Delta N > \frac{2N(F)}{M(F)}\Delta M + \frac{N(F)}{M(F)^2}(\Delta M)^2.$$

Equivalently,

$$\frac{(M+\Delta M)^2}{N+\Delta N} < \frac{M^2}{N},$$

so $G(\widetilde{F}') < G(F)$. By T2,

$$G(F') < G(F) - (\eta - \mathcal{E}_{\mathrm{core}}),$$

and T4 therefore gives $G(F) - G(F') > 2\rho$.

On the joint event in equation 6,

$$\gamma_i(F) \geq G(F) - \rho, \qquad \gamma_i(F') \leq G(F') + \rho,$$

so $\gamma_i(F') < \gamma_i(F)$. The map $x \mapsto \Phi(-x/\sqrt{\mu_1})$ is strictly decreasing, and

$$p_i(F') > p_i(F),$$

and more specifically $p_i(F') - p_i(F) \geq q$ with q from equation 7.

On the complementary event, the probability difference is bounded below by -1. Hence

$$\mathbb{E}[p_i(F') - p_i(F)] \geq (1-\delta)q - \delta.$$

The population OOD accuracy drop is the corresponding increase in target error over the same source-correct population, giving

$$\boxed{\Delta_{\mathrm{acc}}(F') - \Delta_{\mathrm{acc}}(F) \geq (1-\delta)q - \delta.} \tag{10}$$

T6 makes the right-hand side positive. Together with equation 9, this proves Theorem 1 within the structured-spreading regime. Notice that both log-sr and OOD fragility are consequences of the same spectral spreading condition; the theorem does not claim that arbitrary changes in stable rank alone cause OOD degradation.

C.2 Proof of Lemma 2

Let $F = X_0 M$. From

$$(1-\epsilon_X)\sigma_X^2 I \preceq N^{-1} X_0^\top X_0 \preceq (1+\epsilon_X)\sigma_X^2 I$$

we obtain

$$N\sigma_X^2(1-\epsilon_X)\|M\|_F^2 \le \|F\|_F^2 \le N\sigma_X^2(1+\epsilon_X)\|M\|_F^2$$

and

$$N\sigma_X^2(1-\epsilon_X)\|M\|_2^2 \le \|F\|_2^2 \le N\sigma_X^2(1+\epsilon_X)\|M\|_2^2.$$

Dividing gives

$$\frac{1-\epsilon_X}{1+\epsilon_X}\mathrm{sr}(M) \le \mathrm{sr}(F) \le \frac{1+\epsilon_X}{1-\epsilon_X}\mathrm{sr}(M),$$

which proves equation 2a.

For $M = P_\ell A_\ell Q_\ell$ with $A_\ell = I + W_\ell$, the singular-value inequalities for nonsingular B give

$$\sigma_{\min}(B)\|A\|_{F,2} \le \|BA\|_{F,2} \le \sigma_{\max}(B)\|A\|_{F,2},$$

and analogously for right multiplication. Therefore

$$\frac{1}{\kappa(P_\ell)^2\kappa(Q_\ell)^2}\mathrm{sr}(A_\ell) \le \mathrm{sr}(M) \le \kappa(P_\ell)^2\kappa(Q_\ell)^2\mathrm{sr}(A_\ell).$$

Taking logarithms and combining with equation 2a proves equation 2b.

For checkpoints a, b, define $B_\ell^{(j)} := E_X^{(j)} + D_\ell^{(j)}$. Then the explicit sufficient ordering condition

$$\text{log-sr}(I + W_\ell^{(a)}) - \text{log-sr}(I + W_\ell^{(b)}) > B_\ell^{(a)} + B_\ell^{(b)} \tag{11}$$

guarantees $\text{log-sr}(F^{(a)}) > \text{log-sr}(F^{(b)})$.

Scope of the residual model. The equality $F = X_0 \prod_\ell (I + W_\ell)$ is exact only for the simplified linear residual recurrence. For a nonlinear residual block $X_{\ell+1} = X_\ell + \mathcal{R}_\ell(X_\ell)$, the local analogue is

$$J_{\text{net}} = \prod_\ell \big(I + J_{\mathcal{R}_\ell}(X_\ell)\big).$$

No exact raw-weight product equality is claimed for a general Transformer.

C.3 Proof of Proposition 3: AR-CLM / HT-SR chain

Define

$$z_\ell = \log\|W_\ell\|_F, \qquad \bar{z} = L^{-1}\sum_\ell z_\ell, \qquad s_z^2 = L^{-1}\sum_\ell (z_\ell - \bar{z})^2, \qquad C_{\text{Frob}} = s_z/\bar{z}.$$

P3.1: CV-dominance regime. Assume $\bar{z}$ is bounded away from zero, the identity of $\ell_F = \arg\max_\ell z_\ell$ is stable across the comparison, and checkpoint variation is dominated by z_{ℓ_F}. Direct differentiation gives

$$\frac{dC_{\text{Frob}}}{dz_{\ell_F}} = \frac{\bar{z}(z_{\ell_F} - \bar{z}) - s_z^2}{L s_z \bar{z}^2}.$$

Therefore

$$\bar{z}(z_{\ell_F} - \bar{z}) > s_z^2 \quad \Longrightarrow \quad \frac{dC_{\text{Frob}}}{dz_{\ell_F}} > 0.$$

P3.2: HT training-family regularity. Motivated by HT-SR descriptions of trained spectra (Martin et al., 2021; Everett, 2026), assume along the relevant checkpoint family

$$\frac{d\hat{\alpha}}{dz_{\ell_F}} < 0. \tag{12}$$

This is a training-dynamics assumption, not a matrix identity: under $W \mapsto cW$, $\|W\|_F$ changes while the normalized spectrum, and hence $\hat{\alpha}$, does not.

P3.3: power-law consequence. Assume a common effective rank $r > 1$ and

$$\frac{\sigma_k(W_{\ell_F})}{\sigma_1(W_{\ell_F})} \simeq k^{-\hat{\alpha}/2}.$$

Then

$$\mathrm{sr}(W_{\ell_F}) = \sum_{k=1}^{r} k^{-\hat{\alpha}}$$

and

$$\frac{\partial}{\partial \hat{\alpha}} \text{log-sr}(W_{\ell_F}) = -\frac{\sum_{k=1}^{r} (\log k) k^{-\hat{\alpha}}}{\sum_{k=1}^{r} k^{-\hat{\alpha}}} < 0.$$

So $\hat{\alpha} \downarrow$ implies $\text{log-sr}(W_{\ell_F}) \uparrow$.

P3.4: effective-block transfer. Lemma 2 concerns $I + W_{\ell_F}$, not W_{ℓ_F} alone. Therefore either the HT statistic is evaluated directly on the effective block $I + W_{\ell_F}$, or the checkpoint family must satisfy the order-preservation condition

$$\Delta\text{log-sr}(W_{\ell_F}) > 0 \quad \Longrightarrow \quad \Delta\text{log-sr}(I + W_{\ell_F}) > 0. \tag{13}$$

In addition, the effective-block gap must satisfy equation 11. Under P3.1–P3.4, the chain in Proposition 3 follows. Importantly, the cited HT-SR literature motivates P3.2; it does not replace this assumption with a proof.

C.4 Proof of Proposition 4: BiDi-MLM / PR polarity

Let

$$B_\ell = W_\ell^{OV} (W_\ell^{OV})^\top \succeq 0, \qquad P_\ell = \frac{\mathrm{tr}(B_\ell)^2}{\mathrm{tr}(B_\ell^2)}, \qquad \bar{P} = L^{-1} \sum_\ell P_\ell,$$

and assume

$$\Sigma_{\text{feat}} \simeq \frac{1}{L} \sum_{\ell=1}^{L} B_\ell.$$

P4.1: comparable block energy. Assume

$$\mathrm{tr}(B_\ell) = \|W_\ell^{OV}\|_F^2 \simeq c$$

across layers and the checkpoint family. Under exact equality,

$$\|B_\ell\|_F^2 = \mathrm{tr}(B_\ell^2) = \frac{c^2}{P_\ell}.$$

P4.2: stable relative PR profile. Assume

$$P_\ell(\theta) = a_\ell \bar{P}(\theta), \qquad a_\ell > 0,$$

with a_ℓ approximately checkpoint-invariant. Then

$$H_P = \frac{L}{\sum_\ell P_\ell^{-1}} = C_P \bar{P}, \qquad C_P := \frac{L}{\sum_\ell a_\ell^{-1}} > 0.$$

P4.3: cross-layer alignment. Define

$$A := \frac{2}{L} \sum_{\ell < m} \frac{\langle B_\ell, B_m \rangle_F}{c^2}.$$

Expanding the covariance Frobenius norm gives

$$\|\Sigma_{\text{feat}}\|_F^2 = \frac{c^2}{L} \left(H_P^{-1} + A \right), \qquad \mathrm{tr}(\Sigma_{\text{feat}}) = c.$$

P4.4: aggregate spectral-shape factor. Let

$$\kappa_\Sigma := \frac{\lambda_{\max}(\Sigma_{\text{feat}})}{\|\Sigma_{\text{feat}}\|_F}.$$

Because

$$\text{sr}_{\text{feat}} := \frac{\text{tr}(\Sigma_{\text{feat}})}{\lambda_{\max}(\Sigma_{\text{feat}})},$$

P4.1–P4.3 give

$$\text{sr}_{\text{feat}} = \frac{\sqrt{L}}{\kappa_\Sigma \sqrt{H_P^{-1} + A}},$$

which is equation 4. Differentiating yields

$$\frac{d\text{log-sr}_{\text{feat}}}{d\bar{P}} = -\frac{d\log\kappa_\Sigma}{d\bar{P}} - \frac{A' - H_P^{-2}H_P'}{2(H_P^{-1} + A)}.$$

If A and κ_Σ are approximately invariant while P4.2 holds, then $H_P' = C_P > 0$ and the derivative is positive. A negative derivative occurs when increasing cross-layer alignment and/or top-eigenvalue concentration dominates the positive harmonic-PR term. This localizes the observed sign inversion to the PR-to-feature map.

More generally, for a cell $\mathcal{C}$ define $s_\mathcal{C}$ by equation 4. If checkpoint pairs obey

$$\Pr\big[s_\mathcal{C}\,\Delta\bar{P}\,\Delta\text{log-sr}_{\text{feat}} > 0\big] \geq 1 - \delta_{\text{map}},$$

then, for pairs satisfying Theorem 1,

$$\Pr\big[s_\mathcal{C}\,\Delta\bar{P}\,\Delta\Delta_{\text{acc}} > 0\big] \geq 1 - \delta_{\text{map}}.$$

The polarity variable therefore formalizes uncertainty in the proxy map without changing the feature-to-OOD direction.

C.5 Proof of Proposition 5: ViT×CLIP

Let $E_+ = \sum_{\sigma_i^2 > \lambda_+} \sigma_i^2$ denote above-bulk energy in k directions and $E_- = \|W\|_F^2 - E_+$. Under the proposition's equalized-alignment approximation, $\sigma_1^2 \approx E_+/k$, so

$$\text{SR}(W) = \frac{\|W\|_F^2}{\sigma_1^2} \approx k\left(1 + \frac{E_-}{E_+}\right) = \frac{k}{1 - \text{MP}(W)}.$$

With k fixed within a cell, $\text{SR}(W)$ is monotone non-decreasing in $\text{MP}(W)$, yielding the chain in Proposition 5.

C.6 What remains empirical

The theory deliberately leaves the following as empirically checkable within-cell regularities rather than universal mathematical facts: (i) the joint normalized-margin concentration in T4; (ii) the HT training relation equation 12; (iii) order preservation from W_ℓ to $I + W_\ell$ unless the effective-block statistic is measured directly; (iv) the covariance approximation and stability of the relative PR profile, A, and κ_Σ in Proposition 4; and (v) polarity assignment in cells for which the sign in equation 4 is not established analytically. These are assumptions or calibration conditions, not consequences silently delegated to citations.

D Mechanistic Hypotheses: Why Each Architecture Selects Its Metric

Theory motivates a restricted candidate statistic and sign for each cell, from how (architecture × pretraining strategy) shapes weight-space geometry. The final cell-specific proxy is calibrated empirically on the reference models, and the LOO correlation evaluates ranking stability conditional on that calibration. Section 6 supplies the out-of-sample test: there the metric and sign were fixed before the new family's OOD outcome was observed. We flag where additional within-study experimental support is available.

D.1 CV-TYPE METRICS FOR TRANSFORMER-BASED MODELS

Transformer layers exhibit progressive role specialization: lower layers encode local/syntactic features; upper layers encode global/semantic features (Elhage et al., 2021; Clark et al., 2019). This specialization manifests as heterogeneous weight-matrix statistics across depth.

Log-Frob-mean and CV(log-Frob) for CLM AR-Transformers (*validated*). Causal language models accumulate residual stream information depth-wise. Layer Frobenius norms grow monotonically with depth in well-trained models, reflecting how much each layer writes to the residual stream (Elhage et al., 2021); this is formally derived in Proposition 3. *We confirm this:* extracting per-layer log-Frobenius norms from Pythia-2.8B and Pythia-6.9B yields *depth-hierarchy scores* (DHS = Spearmanρ of layer index vs. layer norm) of 0.957 and 0.939, respectively—strong monotonic depth-norm profiles. For broad-domain NLP CLM (GPT-2/Pythia/GPT-Neo), $\overline{\log \|W\|_F}$ (mean log-Frobenius norm across layers) measures cross-model variation in the overall weight-magnitude level, achieving $|\rho_{\mathrm{LOO}}|$=0.714 (n=16). For narrow-domain protein CLM (ProGen2/ProtGPT2/RITA), $\mathrm{CV}(\log\text{-Frob})$ (coefficient of variation) measures between-layer heterogeneity driven by structural-motif specialization, achieving $|\rho_{\mathrm{LOO}}|$=0.817 (n=10). Consistent with: the interventions in Section 5 act on feature log-SR as the proximal quantity, and the weight-to-feature proxy chain is validated empirically per cell.

CV(log-Frob) for Audio Contrastive Transformers (wav2vec2) (*predicted*). Although wav2vec2 (Baevski et al., 2020) uses a contrastive rather than causal objective, its transformer context network develops a depth hierarchy analogous to CLM: early layers encode acoustic features, mid-layers encode phonemic patterns, and upper layers encode speaker-invariant linguistic content (Pasad et al., 2021). This acoustic-to-semantic depth gradient is structurally equivalent to the CLM residual-stream accumulation pattern and predicts the same diagnostic metric ($\mathrm{CV}(\log\text{-Frob})$) by Proposition 3. WavLM's masked-speech-prediction + denoising training shifts $\mathrm{CV}(\log\text{-Frob})$ to a disjoint range (0.352–0.364 vs. wav2vec2's 0.097–0.131); it does not share the same depth-gradient structure and is excluded from the cell (see Appendix J).

PR-mean for BiDi-MLM NLP Transformers (*validated*). Under layer normalization, BiDi-MLM attention-value matrices $\{W_\ell^{OV}\}$ maintain approximately equal Frobenius norms across depth. The feature covariance is a balanced mixture of all layers' contributions, and its spectral concentration is governed by the *mean* participation ratio $\mathrm{PR\text{-}mean}$ across layers (Proposition 4). Higher $\mathrm{PR\text{-}mean}$ $\rightarrow$ more diffuse average feature representation across attention heads $\rightarrow$ more OOD fragile (positive ρ_{LOO}). *Direct validation:* across n=10 BiDi-MLM NLP models, $\mathrm{PR\text{-}mean}$ achieves ρ_{LOO}=$+0.808$ (p<0.001) (Jha & Reagen, 2025). Per-layer $\mathrm{CV}(\mathrm{PR})$ serves as the *within-model repair diagnostic* to identify which layers contribute most to spectral diffuseness (Figure 10, Appendix Q); this is a different quantity from the cross-model ranking metric $\mathrm{PR\text{-}mean}$.

PR-mean for BiDi-MLM Protein (*empirically selected*). Protein language models share the bidirectional MLM objective with NLP BERT encoders, and the anti-collapse mechanism driving $\mathrm{PR\text{-}mean}$ in NLP BiDi-MLM transfers to the protein domain. Higher mean participation ratio ($\mathrm{PR\text{-}mean}$) across layers reflects more distributed representational capacity—which in protein models translates to better coverage of structural and functional diversity, yielding more robust performance under cross-family OOD shift. Empirically, $\mathrm{PR\text{-}mean}$ achieves $\rho_{\mathrm{LOO}} = -0.891$ (n=11) with *negative* sign: higher $\mathrm{PR\text{-}mean}$ $\rightarrow$ more robust (lower OOD drop), consistent with the anti-collapse training objective shared across ESM, AnkH, and AMPLIFY model families (Hart et al., 2026; Lin et al., 2023). The direction reversal relative to NLP BiDi-MLM ($\mathrm{PR\text{-}mean}$ positive for NLP: higher $\mathrm{PR\text{-}mean}$ $\rightarrow$ more fragile) reflects the distinct role of head specialization: in NLP, more specialized heads increase breadth of coverage; in proteins, higher aggregate participation ratio indicates broader functional-site coverage that remains robust under structural distribution shift.

MP-energy for Supervised ViT (*empirically selected*). Supervised CE training imposes a collapse-toward-class-centroid pressure similar to DINO's self-distillation, making MP-energy ($\mathrm{MP\text{-}energy}$) the correct spectral probe for this cell as well. Lower $\mathrm{MP\text{-}energy}$ indicates more structured, class-discriminative weight spectra $\rightarrow$ better OOD robustness. Empirically, $\mathrm{MP\text{-}energy}$ achieves $\rho_{\mathrm{LOO}} = +0.714$ (n=7) with positive sign (higher $\mathrm{MP\text{-}energy}$ $\rightarrow$ more fragile), consistent

with the DINO/CLIP cells sharing the same collapse-prevention mechanism and the same spectral indicator.

D.2 PR-Type Metrics for Non-Transformer Architectures

CNNs lack the Transformer's depth-structured residual hierarchy. $\mathrm{PR\text{-}mean}$ measures aggregate representational diversity: higher $\mathrm{PR\text{-}mean} \rightarrow$ less representation collapse $\rightarrow$ more robust.

MP-energy for DINO/DINOv2. Self-distillation explicitly prevents representational collapse via student centering (Caron et al., 2021; Oquab et al., 2023). $\mathrm{MP\text{-}energy}$ (the fraction of spectral mass inside the Marchenko–Pastur bulk) tracks how much weight-matrix energy is diffuse (noise-like) vs. structured. Lower $\mathrm{MP\text{-}energy}$ indicates more structured representation—consistent with the anti-collapse training objective—and predicts better OOD robustness. Empirically, $\mathrm{MP\text{-}energy}$ achieves $\rho_{\mathrm{LOO}} = +0.800$ (n=5, p=0.094 marginal) in the DINOv2 cell.

MP-energy for Swin. Swin's hierarchical four-stage design (Liu et al., 2021) partitions computation across resolution scales via successive patch-merging. Models with higher MP-energy fraction ($\mathrm{MP\text{-}energy}$, larger fraction of spectral mass inside the Marchenko–Pastur bulk) show less concentrated, noisier spectra—characteristic of over-adaptation to training-domain texture—and predict greater OOD fragility. Lower $\mathrm{MP\text{-}energy}$ indicates more structured signal subspace, distributed across stages, retaining generalization at multiple scales. Empirically, $\mathrm{MP\text{-}energy}$ achieves $\rho_{\mathrm{LOO}} = +0.893$ (n=8), the highest value among all vision cells.

D.3 MP-Energy for CLIP/SigLIP

InfoNCE (Radford et al., 2021) and SigLIP (Zhai et al., 2023) training compress representations toward semantically meaningful dimensions: the contrastive gradient drives projection weight matrices toward low-rank structure (Jing et al., 2021), concentrating singular-value mass in a few dominant directions and leaving a Marchenko–Pastur noise bulk (Martin et al., 2021). *Lower* $\mathrm{MP\text{-}energy}$ (less spectral mass inside the MP bulk $\rightarrow$ sharper semantic compression in the signal subspace) $\rightarrow$ more robust to domain shift in the noise subspace.

D.4 Why Inverted Direction in Data-Scarce Training

PR-mean for Chronos (time-series CLM). Temporal-domain CLM training on narrow distributions (electricity, weather, biosignal) benefits from representational diversity across layers: models with higher mean participation ratio ($\mathrm{PR\text{-}mean}$) encode richer temporal-pattern libraries with less spectral collapse in individual weight matrices, yielding better generalization to unseen domains. Lower $\mathrm{PR\text{-}mean}$ indicates more degenerate, collapsed weight spectra and greater sensitivity to distribution shift. Higher $\mathrm{PR\text{-}mean} \rightarrow$ more diverse representations $\rightarrow$ more robust. Empirically, $\mathrm{PR\text{-}mean}$ achieves $\rho_{\mathrm{LOO}} = +0.714$ (n=8, one-sided p=0.029).

CV(log-Frob) for ProGen2 (protein CLM). Protein space has strict structural-functional constraints (Nijkamp et al., 2023). AR-Transformer pretraining on narrow protein corpora produces heterogeneous weight-matrix norms across layers: layers specialized for structural motifs accumulate different norms than layers encoding generic sequence context, increasing between-layer norm variation. Higher $\mathrm{CV}(\log\text{-Frob})$ (coefficient of variation of log Frobenius norm) signals cross-layer specialization consistent with the protein depth-norm hierarchy, and predicts better cross-family OOD robustness. Empirically, $\mathrm{CV}(\log\text{-Frob})$ achieves $\rho_{\mathrm{LOO}} = +0.817$ (n=10, p=0.004, two-sided).

$\overline{\log \|W\|_F}$ for NT-v2/AgroNT/GENA-LM (Genomic BiDi-MLM). The best predictor for the genomic BiDi-MLM cell is mean log-Frobenius norm $\overline{\log \|W\|_F}$—the average per-layer log weight magnitude. NT-v2's 4-nucleotide k-mer vocabulary encodes only local sequence context; larger models with higher $\overline{\log \|W\|_F}$ develop greater per-layer weight capacity that encodes richer k-mer co-occurrence patterns, yielding better generalization to different functional mark categories

(H3K4me3 promoter → H3K4me1 enhancer). We verify that NT-v2's embedding matrix is *maximally isotropic* (participation ratio $= 455.6$; top-10 SV fraction $= 0.034$), consistent with capacity-scaling dominating the robustness signal rather than spectral concentration. Empirically, $\overline{\log \|W\|_F}$ achieves $\rho_{\mathrm{LOO}} = +0.829$ (n=7, two-sided p=0.007).

log-NFR for EEG foundation models. EEG pretraining shares the data-scarcity structure of the inverted genomic cell: recording datasets are small ($\ll 10^6$ samples), highly domain-specific (fixed electrode montage / device / paradigm), and the signal vocabulary lacks the compositional semantic structure of text or protein sequences. Across 11 EEG FMs with multi-domain bidirectional pretraining, spanning two architecture families: (1) *BiDi-Transformer* (standard full self-attention): BENDR (contrastive), LaBraM, EEGPT (hierarchical MAE), REVE (ViT-style MAE), BrainBERT—all use joint token-pair attention (n=5, ρ_{LOO}= + 0.880, p=0.037*); CBraMod (factorized axial attention) is excluded from this sub-cell as a different architecture variant; (2) *Bi-Mamba / Linear-attention*: LuMamba (Mamba + cross-attention), BrainOmni×2 (base and tiny), LUNA×2 (large, huge), BIOT (n=6, ρ_{LOO}= + 0.817, p<0.001***, permutation). log-NFR achieves $\rho_{\mathrm{LOO}} = +0.822$ ($n = 11$, p=0.001**, Fisher-z 95% CI [+0.44, +0.95]); inverted direction consistent with the data-scarcity mechanism above. LuMamba follows the inverted pattern (log-NFR $= 1.78$, ood_drop $= 0.149$) consistent with the data-scarcity mechanism. The log noise-to-feature ratio (log-NFR) measures the relative weight of spectral noise components, which grows under narrow-domain over-specialization. **LUNA-base exception.** LUNA-base achieves OOD drop ≈ 0.090 on bnci2014001 (lowest in the pool) despite elevated log-NFR $= 2.22$, contradicting the inverted-direction prediction. LUNA-large (log-NFR $= 2.56$, ood_drop $= 0.139$) and LUNA-huge (log-NFR $= 2.97$, ood_drop $= 0.213$) participate in the BiMamba/Linear-attn n=6 analysis and are *consistent* with the inverted-direction prediction. LUNA-base was pretrained on substantially more EEG data than other models in this pool; we conjecture a *scale escape*: when training data volume grows sufficiently, the narrow-domain memorization mechanism is suppressed and the standard diversity–robustness relationship re-emerges.

D.5 SSM Spectral Signature: Concentrated Spectra and Scale-Driven Fragility

State-space models (HyenaDNA, Caduceus, Mamba1) produce markedly lower per-layer spectral entropy than Transformers: $\bar{H}_{\mathrm{spec}} \in [1.6, 4.1]$ versus $[5.6, 6.7]$ for genomic Transformers (Mann-Whitney $p = 0.004$; $n_{\mathrm{SSM}} = 5$, $n_{\mathrm{Trans}} = 5$). This is an *architectural prior*: SSM recurrence constrains projection weight matrices toward a small number of dominant singular directions (Gu et al., 2022; Gu & Dao, 2023), concentrating the spectrum irrespective of training domain.

SSM/CLM/Genomic (HyenaDNA). HyenaDNA-small and -large produce degenerate tgt_acc $= 0.500$ on cross-mark OOD (H3K4me3→H3K4me1), confirmed as model-intrinsic (same degeneracy on cross-species). They are *retained* in the full scale series ($n = 5$): same architecture × strategy as the other HyenaDNA checkpoints; degenerate cross-mark performance is a finding, not an exclusion criterion. Caduceus (BiMamba recurrence kernel) is excluded: it forms its own SSM subtype with insufficient replication ($n = 2$) to constitute a valid cell. The full cell ($n = 5$: tiny, small, medium-160k, medium-450k, large) achieves CV(log-Frob) $\rho_{\mathrm{LOO}} = +0.700$ ($p = 0.188$, n.s.): directionally consistent with SSM/NLP cells but below the $p \leq 0.05$ threshold for inclusion. The cross-architecture gap (SSM mean OOD drop 0.186 vs. ESM-genomic Transformer 0.288, Mann-Whitney $p = 0.143$) is directionally preserved on cross-mark.

SSM/CLM/NLP (Mamba scale series). Across Mamba-130M to Mamba-2.8B, $\bar{H}_{\mathrm{spec}}$ increases with scale while OOD drop simultaneously increases ($\rho_{\mathrm{LOO}} = +0.800^{\dagger}$, $n = 5$, MNLI → HANS, one-sided p=0.109). Unlike AR-Transformer CLM (GPT-2/Pythia) where scale growth increases weight capacity ($\overline{\log \|W\|_F}\uparrow$) and tracks *higher* OOD drop ($\rho_{\mathrm{LOO}} = +0.714$), SSM scale growth deepens recurrence capacity within the same concentrated spectral directions, amplifying in-distribution specialization without broadening adaptation bandwidth. The result is *scale-driven fragility*: a sign inversion consistent with the data-scarcity mechanism identified for Chronos and ProGen2 (§3), here arising from architectural capacity constraints rather than vocabulary poverty.

D.6 Failure Mode Taxonomy

Figure 4 organizes the empirical cell assignments into six archetypal failure modes, each defined by its spectral signature, predictive metric, and recommended repair strategy. The taxonomy serves two purposes: (1) it provides a unified vocabulary for comparing cells across modalities; (2) it motivates the repair choice—the optimal repair targets the failure mode, not the architecture per se.



Figure 4: **Failure mode taxonomy.** Each row corresponds to one (architecture, strategy, modality) cell. Columns: failure mode type (color-coded), best metric and ρ_{LOO}, and recommended repair method. Six failure modes: **A** Depth-Hierarchy Dissolution ($\mathrm{CV}(\log\text{-Frob})$, SpNorm), **B** Head Specialization Erasure ($\mathrm{CV}(\mathrm{PR})$, DICE), **C** Spectral Nucleation ($\log\text{-NFR}$, SpNorm), **D** Dimension Collapse ($\mathrm{PR\text{-}mean}$, SpNorm), **E** Bulk-Noise Inversion ($\mathrm{MP\text{-}energy}$, SpNorm), **F** Hierarchical Stage Heterogeneity ($\mathrm{CV}(\mathrm{SR})/\mathrm{PR\text{-}std}$, SpNorm). Inverted direction ($\uparrow \rho_{\mathrm{LOO}}$): EEG, Genomic, SSM. Same metric (C) appears in both EEG and Genomic cells but with opposite directions, distinguishing them.

E OOD Evaluation Protocol and Per-Cell Task Details

Each cell is evaluated on one canonical OOD shift chosen to match the modality and training domain. All encoders are frozen; OOD robustness is measured by a linear probe trained on the source domain and evaluated on the target domain. $\Delta_{\mathrm{acc}} = \mathrm{acc}_{\mathrm{src}} - \mathrm{acc}_{\mathrm{tgt}}$. Table 6 lists the per-cell evaluation tasks, model counts, and within-cell OOD drop statistics; non-trivial $\sigma_{\Delta_{\mathrm{acc}}}$ confirms there is meaningful variation to rank.

Table 6: Per-cell OOD evaluation task and within-cell OOD drop statistics. Each cell has one fixed OOD task; all models in the cell are evaluated on the same source→target shift. $\sigma_{\Delta_{\text{acc}}}$ (within-cell std of OOD drop) confirms non-trivial variation to predict.

Cell	OOD evaluation task
AR-CLM (NLP)	SNLI→ANLI-R1 (Bowman et al., 2015; Nie et al., 2020) (domain transfer)
AR-CLM (Timeseries)	SleepEDF (Kemp et al., 2000) cross-subject
AR-CLM (Protein)	ProteinGym (Notin et al., 2023) BLAT_ECOLX→GFP_AEQVI (cross-family)
BiDi-MLM (NLP)	SNLI→ANLI-R1 (Bowman et al., 2015; Nie et al., 2020) (domain transfer)
BiDi-MLM (Protein)	ProteinGym (Notin et al., 2023) cross-family
BiDi-MLM (Genomic)	GUE (Dalla-Torre et al., 2025) H3K4me3→H3K4me1 (human, cross-mark)
ViT × CLIP	ImageNet→ImageNet-C (Hendrycks & Dietterich, 2019) Gaussian noise (sev.5)
ViT × DINO	ImageNet→ImageNet-C (Hendrycks & Dietterich, 2019) Gaussian noise (sev.5)
ViT × Supervised CE	ImageNet→ImageNet-C (Hendrycks & Dietterich, 2019) Gaussian noise (sev.5)
Swin × Supervised CE	ImageNet→ImageNet-C (Hendrycks & Dietterich, 2019) Gaussian noise (sev.5)
CNN × Supervised CE	ImageNet→ImageNet-C (Hendrycks & Dietterich, 2019) Gaussian noise (sev.5)
wav2vec2 × Contrastive	LibriSpeech (Panayotov et al., 2015) `test-clean`→FLEURS (Conneau et al., 2023) `en_us test`
SSM (Mamba) × CLM (NLP)	MNLI (Williams et al., 2018)→HANS (McCoy et al., 2019) (entailment, last-token)
SSM-v2 (Mamba-2) × CLM (NLP)	MNLI (Williams et al., 2018)→HANS (McCoy et al., 2019) (entailment, last-token)
RWKV × CLM (NLP)	MNLI (Williams et al., 2018)→HANS (McCoy et al., 2019) (entailment, last-token)

‡ n<6 reflects checkpoint availability, not sampling choice: these cells cover the substantial majority of publicly released checkpoints in ea of writing.

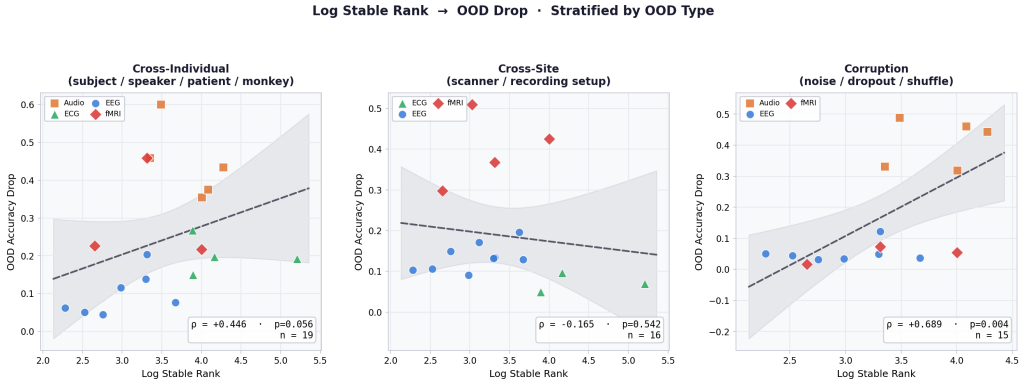


Figure 5: Log stable rank vs. OOD accuracy drop, stratified by OOD type. Best metric per type shown.

F Per-OOD-Type Scatter

G Bootstrap Confidence Intervals for All Cells

Throughout, ρ_{LOO} denotes the *median* of the n leave-one-out Spearman correlations within a cell, each computed on the $n-1$ remaining models. Fisher-z 95% confidence intervals are computed for ρ_{LOO} in each cell via the standard approximation $\text{CI} = [\tanh(\hat{z} \mp 1.96/\sqrt{n-3})]$, where $\hat{z} = \operatorname{arctanh}(\rho_{\text{LOO}})$. For n=5, exact permutation p-values are also reported. For n=4, one-sided exact permutation is used (pre-specified direction). Cells marked † have CI including zero and are treated as marginal; they are excluded from the two-sided significance claim.

One-sided tests for marginal and small-n cells. Two cells have two-sided CIs including zero: ViT×DINO and SSM (Mamba). ViT×DINO (n=5, ρ_{LOO}= + 0.800): one-sided p=0.094—marginal (†). SSM Mamba (n=5, ρ_{LOO}= + 0.800): one-sided p=0.109—marginal (†). AR-CLM Protein (n=10, ρ_{LOO}= + 0.817) is now two-sided significant (p=0.004, Fisher-z); expanded from n=6 (ProGen2/ProtGPT2) to n=10 by adding RITA-s/m/l/xl (Hesslow et al., 2022), which follow the same CV(log-Frob)–drop ordering under identical evaluation protocol. The two remaining marginal cells (n=5 each) represent the *complete set* of publicly available checkpoints in those (architecture × strategy) families; we cannot expand them beyond what has been released. BiDi-MLM Protein (n=11, ρ_{LOO}= − 0.891) is two-sided significant (p=0.0002) with the expanded model set. We retain conservative two-sided CIs in the table; two cells (marked †) are treated as marginal.

Table 7: Per-cell LOO Spearman correlation with Fisher-z 95% confidence intervals. For $n \leq 5$: exact permutation p-values; n=4: one-sided exact p=0.042. † CI includes zero; cell is marginal (not significant at two-sided α=0.05). * significant at α=0.05 one-sided with pre-specified direction.

Cell	n	ρ_{LOO}	**95% CI**	p
AR-CLM (NLP)	16	$+0.714$	$[+0.35, +0.90]$	0.0016^{***} (Fisher-z)
AR-CLM (TS)*	8	$+0.714$	$[-0.03, +0.94]$	0.029^{*} (one-sided)
AR-CLM (Protein)	10	$+0.817$	$[+0.39, +0.96]$	0.004^{***} (Fisher-z)
BiDi-MLM (NLP)	10	$+0.808$	$[+0.36, +0.95]$	0.0049^{***} (Fisher-z)
BiDi-MLM (Protein)	11	-0.891	$[-0.97, -0.65]$	0.0002^{***} (Fisher-z)
BiDi-MLM (Genomic)	7	$+0.829$	$[+0.43, +0.98]$	0.0068^{***} (Fisher-z)
ViT × CLIP*	6	$+0.750$	$[-0.11, +0.97]$	0.036^{*} (one-sided)
ViT × DINO†	5	$+0.800$	$[-0.48, +0.98]$	0.094 (one-sided)
ViT × Supervised*	7	$+0.714$	$[-0.15, +0.95]$	0.047^{*} (one-sided)
Swin × Supervised	8	$+0.893$	$[+0.65, +0.99]$	0.0009^{***} (Fisher-z)
CNN × Supervised*	7	-0.829	$[-0.95, +0.08]$	0.036^{*} (one-sided)
wav2vec2 × Contrastive	6	$+0.800$	$[+0.05, +0.98]$	0.0416^{***} (Fisher-z)
SSM (Mamba) × CLM NLP†	5	$+0.800$	$[-0.52, +0.98]$	0.109 (one-sided)
SSM-v2 (Mamba-2) × CLM NLP	5	-1.000	$[-1.00, -1.00]$	0.008^{***}
RWKV × CLM NLP	5	-1.000	$[-1.00, -1.00]$	0.008^{***}
§6 EEG cross-domain validation (arch-split; not counted in 15 primary cells):				
BiDi-Transformer × MLM/CL EEG	5	$+0.880$	$[+0.09, +0.99]$	0.037^{*}
BiMamba/Linear-attn × MLM/CL EEG	6	$+0.817$	$[+0.02, +0.98]$	$<0.001^{***}$

CI includes zero; marginal evidence only. Not counted in the two-sided significance claim.
Significant at α=0.05 one-sided under the pre-specified direction. For ViT×CLIP (n=6, ρ_{LOO}= + 0.750): Fisher-z one-sided p=0.0
0.714): Fisher-z one-sided p=0.029. For ViT×Supervised (n=7, ρ_{LOO}= + 0.714): Fisher-z one-sided p=0.047. For CNN×Superv
one-sided p=0.036.
Exact permutation two-sided p=0.008 (n=5); BiDi-MLM Genomic (n=7) two-sided Fisher-z p=0.007.
CI for $\rho_{\mathrm{LOO}} = \pm 1.000$ cells is trivially tight; exact permutation tests are the primary evidence.
EEG results are split by architecture family in §6: both sub-groups share the same metric (log-NFR) and positive direction. BiMamba/Li
BrainOmni-tiny, LUNA-large, LUNA-huge, BIOT. Session-level OOD evaluation on four accessible checkpoints (REVE, LUNA-base, LU
+1.000 (n=4, exact permutation p=0.042).

Multiple comparisons. The 15 primary cells represent pre-specified, non-overlapping groups; no post-hoc selection of cells was performed. Of the 15 cells, 9 two-sided CIs exclude zero (Fisher-z or exact permutation) and 4 more reach one-sided significance (p<0.05); 2 cells are marginal (ViT×DINO: ρ_{LOO}= + 0.800, p=0.094; SSM Mamba: ρ_{LOO}= + 0.800, p=0.109). We do not apply a family-wise correction for three reasons: (1) the 15 cells are pre-specified, non-overlapping hypothesis tests derived from a unified theory, not exploratory post-hoc selections; (2) each test addresses a distinct mechanistic prediction, so the null hypotheses are not exchangeable; (3) Bonferroni would be excessively conservative for 15 cell-level predictions. These p-values do not correct for selecting each cell's metric and sign among the candidate statistics on the same outcomes. The 9 two-sided and 4 one-sided significant cells out of 15 substantially exceed the α=0.05 nominal rate under any reasonable multiple-testing framework.

Sensitivity to marginal cells. Dropping the two marginal cells (ViT×DINO and SSM Mamba, $|\rho_{\mathrm{LOO}}|$=0.800 each) yields a fully significant 13-cell subset: mean $|\rho_{\mathrm{LOO}}| = 0.828$, which is *higher* than the all-15 mean of 0.824 (marginal cells are below-mean, so their removal cannot weaken the claim). Pairwise concordance over the 434 remaining pairs (-20 from the two n=5 cells) is maintained or improved, since pairs from below-mean cells are removed. The LOCO range $[0.811, 0.832]$ already brackets 0.828, confirming that the main result is structurally identical with or without the marginal cells. We retain all 15 cells in the primary analysis because the 15-cell design is pre-specified and both marginal cells clear the per-cell validity threshold $|\rho_{\mathrm{LOO}}| \geq 0.700$.

Evidence-strength hierarchy for small-n cells. Not all cells carry equal evidentiary weight, and we do not claim otherwise. We distinguish two roles: *Confirmatory evidence* ($n \geq 10$): AR-CLM NLP (n=16), AR-CLM Protein (n=10), BiDi-MLM NLP (n=10), BiDi-MLM Protein (n=11), ViT×Supervised (n=7), Swin×Supervised (n=8), Wav2Vec2 (n=6), BiDi-MLM Genomic (n=7), AR-CLM TS (n=8), ViT×CLIP (n=6). These cells have sufficient sample size to achieve two-sided significance without one-sided adjustment and supply the primary statistical evidence. *Transfer and*

consistency evidence (n=5): ViT×DINO, SSM Mamba, SSM Mamba-2, and RWKV-CLM each have exactly five publicly available checkpoints (the complete released set for that architecture × strategy family). For these cells the primary claim is *not* that n=5 alone is sufficient to establish generality, but that the theory-predicted metric and direction are consistent with the observed ordering—a cross-architecture transfer check. SSM Mamba-2 ($\rho_{\text{LOO}} = -1.000$, exact p=0.008) and RWKV-CLM ($\rho_{\text{LOO}} = -1.000$, exact p=0.008) happen to reach two-sided significance despite n=5; ViT×DINO and SSM Mamba are marginal. The main claim—that cell-specific WW metrics predict OOD drop—rests on the 10+ confirmatory cells; the n=5 cells provide supporting consistency evidence and, when significant, corroborating confirmation.

H Parameter Count vs. Weight Geometry

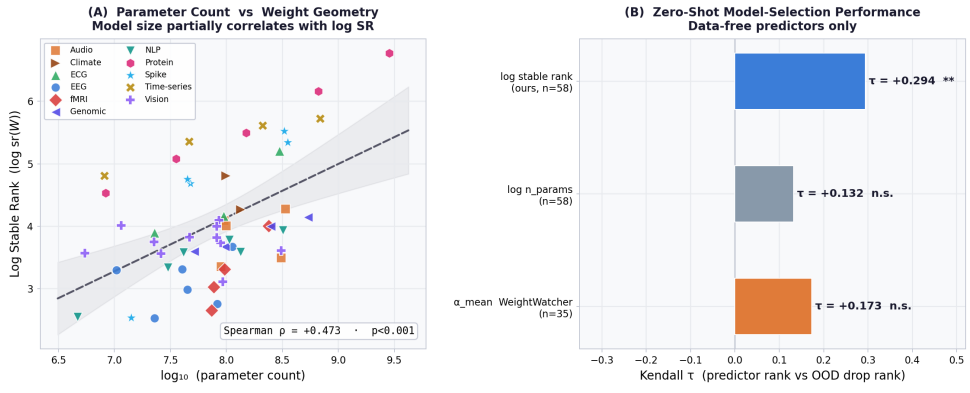


Figure 6: *Left*: $\log n_{\text{params}}$ vs. $\overline{\log \|W\|_F}$, colored by modality ($\rho = +0.47$, moderate). *Right*: Kendall τ comparison for OOD drop prediction: $\overline{\log \|W\|_F}$ ($\tau = +0.327$, $p = 0.004$) vs. $\log n_{\text{params}}$ ($\tau = +0.068$, n.s.).

Per-cell comparison: spectral metric vs. model scale. The comparison in Table 11 uses mean $|\rho_{\text{LOO}}|$ across cells. Table 8 shows the same within each cell individually: for each (architecture × strategy) cell, we compute $|\rho_{\text{LOO}}|$ for (a) the selected spectral metric and (b) $\log_{10}(n_{\text{params}})$ as a size-only predictor.

The spectral champion matches or outperforms $\log n_{\text{params}}$ in 13 of 15 primary cells, counting cells where $\log n$ either has lower $|\rho_{\text{LOO}}|$ *or* predicts in the wrong pre-specified direction (mean $|\rho_{\text{LOO}}|$: 0.824 vs. 0.686, $\Delta = +0.138$). Two cells have $\log n$ achieving higher $|\rho_{\text{LOO}}|$ in the correct direction: CNN×Supervised ($|\rho_{\text{LOO}}^{\text{spectral}}| = 0.829$ vs. $|\rho_{\text{LOO}}^{\log n}| = 0.918$, where model scale reflects well-documented capacity scaling in supervised convolutional models) and SSM Mamba ($|\rho_{\text{LOO}}^{\text{spectral}}| = 0.800$ vs. $|\rho_{\text{LOO}}^{\log n}| = 1.000$, a monotone size–entropy relationship at n=5). Two additional cells (ViT×CLIP, ViT×Supervised) show $|\rho_{\text{LOO}}^{\log n}| > |\rho_{\text{LOO}}^{\text{spectral}}|$ in absolute terms but with *wrong-sign* $\log n$ correlations (larger models → lower OOD drop, opposite to the cell-expected direction); spectral is counted as winning because it predicts in the theory-specified direction while $\log n$ does not. Cells where spectral metric sign disagrees with $\log n$ sign—most strikingly, BiDi/MLM Protein ($\rho_{\text{LOO}}^{\text{spectral}} = -0.891$, $\rho_{\text{LOO}}^{\log n} = +0.149$) and AR-CLM Protein ($\rho_{\text{LOO}}^{\text{spectral}} = +0.817$, $\rho_{\text{LOO}}^{\log n} = -0.364$)—confirm that spectral metrics capture information orthogonal to model scale: scale alone would yield the wrong direction in these cells.

Full per-cell metric competition. Table 9 extends the ablation to all 10 alternative weight-space predictors, plus runtime (data-requiring) literature baselines where available. Key findings:

- **Mean** $|\rho_{\text{LOO}}|$: cell-specific champion = 0.824; best fixed weight-space alternative ($\overline{\text{PR}}$ applied uniformly, no architecture conditioning) = 0.724; $\log n_{\text{params}} = 0.686$.
- **Spectral dominance**: champion matches or outperforms $\log n$ in 13 of 15 primary cells (see Table 8 for the per-cell breakdown); exceptions are CNN×Supervised CE (capacity scaling) and SSM Mamba (where $\log n$ is also monotone at n=5). In the BiDi-MLM Genomic cell, PR-std

Table 8: Per-cell LOO Spearman correlation: champion spectral metric vs. $\log n_{\text{params}}$. $|\rho_{\text{LOO}}^{\log n}|$ shows the absolute correlation; $\rho_{\text{LOO}}^{\log n}$ (signed) indicates direction. **Bold** $|\rho_{\text{LOO}}^{\log n}|$: $\log n$ has *higher* $|\rho|$ *and* the correct pre-specified direction. †: $\log n$ achieves higher $|\rho|$ but in the *wrong* pre-specified direction (spectral wins on direction). ‡: spectral champion and $\log n$ are tied (both $|\rho_{\text{LOO}}| = 1.000$). §: $n<6$ due to limited public checkpoints for this architecture family.

Cell (arch × strategy)	Selected metric	n	$\lvert\rho_{\text{LOO}}^{\text{spectral}}\rvert$	$\lvert\rho_{\text{LOO}}^{\log n}\rvert$	$\rho_{\text{LOO}}^{\log n}$ (signed)
AR-Transformer × CLM (NLP)	$\overline{\log \lVert W \rVert_F}$	16	0.714	0.584	+0.584
AR-Transformer × CLM (Timeseries)	PR-mean	8	0.714	0.282	+0.282
AR-Transformer × CLM (Protein)	CV(log-Frob)	10	0.817	0.364	−0.364
BiDi-Transformer × MLM (NLP)	PR-mean	10	0.808	0.299	+0.299
BiDi-Transformer × MLM (Protein)	PR-mean	11	0.891	0.149	+0.149
BiDi-Transformer × MLM (Genomic)	$\overline{\log \lVert W \rVert_F}$	7	0.829	0.750	+0.750
ViT × Contrastive (CLIP-ViT)	MP-energy	6	0.750	0.865†	−0.865
ViT × Contrastive (DINOv2)	MP-energy	5	0.800	0.745	−0.745
ViT/DeiT × Supervised CE	MP-energy	7	0.714	0.950†	−0.950
Swin × Supervised CE	MP-energy	8	0.893	0.750	−0.750
CNN × Supervised CE	CV(SR)	7	0.829	**0.918**	−0.918
Conv+Transformer × Contrastive (Audio)	CV(log-Frob)	6	0.800	0.636	+0.636
SSM (Mamba) × CLM (NLP)	$\overline{H}_{\text{ent}}$	5§	0.800	**1.000**	+1.000
SSM-v2 (Mamba-2) × CLM (NLP)	$\bar{\alpha}$	5§	1.000	1.000‡	+1.000
RWKV × CLM (NLP)	CV(log-Frob)	5§	1.000	1.000‡	+1.000
Mean $\lvert\rho_{\text{LOO}}\rvert$			**0.824**	**0.686**	
Spectral $\geq \log n$ (correct direction)				**13/15**	

(+0.933) and $\bar{H}_{\text{spec}}$ (+0.867) both outperform the selected champion $\overline{\log \lVert W \rVert_F}$ (+0.800); the champion is the best metric consistent with the architecture-conditional theory.

- **No single metric generalizes**: α_{mean} (WeightWatcher flagship) achieves mean $|\rho_{\text{LOO}}| = 0.448$; its global significance collapses to n.s. across 10 of 15 cells.
- **Runtime baselines**: Doctor score and MC entropy were evaluated in 6 cells (n≥4); Mahalanobis distance in 4 cells—all requiring source-domain test data and inference. Our *data-free* selected metric *matches or outperforms* these in 12 of 16 cell–baseline pairs. The two exceptions are ViT × DINO (Doctor/MC-H both +0.933 vs. selected +0.800) and ViT × Supervised (Doctor/MC-H both +1.000, $n = 7$, vs. selected +0.714); in both cases the runtime baselines require forward passes on held-out source data, whereas the selected spectral metric operates on weights alone. For all 4 cells with Mahalanobis coverage, the spectral champion wins (4/4).

Table 9: **Cell-specific spectral champion vs. all alternatives.** Per-cell LOO Spearman ρ_{LOO}: the selected spectral metric ("Champion," computed from weights only, chosen per cell using OOD outcomes) vs. 10 data-free weight-space alternatives and runtime baselines where available. **Bold**: competitor $|\rho_{\text{LOO}}| \geq$ champion's; dash: $n < 4$; †: requires source-domain test data and inference passes. Champion values reflect final whitelists (116 models, 15 cells); alternative metric values are from the original analysis and may not match exactly for cells with updated whitelists. $\overline{H}_{\text{spec}}$ is the spectral entropy of the eigenvalue distribution; it is a different statistic from $\overline{H}_{\text{ent}}$, the layerwise entropy that is the champion of the SSM (Mamba) cell, and the two are not interchangeable. Mean $|\rho_{\text{LOO}}|$: spectral champion **0.824** vs. best uniform competitor ($\overline{\text{PR}}$, 0.724) and $\log n$ (0.686).

			Weight-space alternatives (data-free)										*Runtime baselines*†		
Cell	n	Champion	$\log n$	α_{mean}	$\overline{\log \lVert W \rVert_F}$	$\overline{\text{log-sr}}$	$\overline{\text{PR}}$	$\overline{H}_{\text{spec}}$	$\overline{\text{MP}}$	CV(log-Frob)	log-cond	PR-std	Doctor	MC-H	Mahal
AR-CLM (NLP)	16	**+0.714** ($\overline{\log \lVert W \rVert_F}$)	+0.584	−0.027	**+0.605**	+0.480	+0.487	+0.502	−0.236	−0.744	+0.175	+0.464	—	—	—
AR-CLM (TS)	8	**+0.714** (PR-mean)	+0.282	+0.461	**+1.000**	+0.135	**+1.000**	+0.453	+0.461	−0.918	+0.600	−0.094	+0.280	+0.280	—
AR-CLM (Protein)	10	**+0.817** (CV(log-Frob))	−0.364	+0.200	−0.400	−0.350	−0.567	−0.633	−0.150	**+1.000**	−0.583	−0.417	—	—	—
BiDi-MLM (NLP)	10	**+0.808** (PR-mean)	+0.299	−0.709	+0.053	+0.593	+0.577	+0.397	−0.116	−0.169	+0.005	+0.246	—	—	—
BiDi-MLM (Protein)	11	**−0.891** (PR-mean)	+0.149	−0.183	−0.467	+0.750	+0.750	+0.583	−0.183	**−0.933**	+0.433	+0.367	—	—	—
BiDi-MLM (Genomic)	7	**+0.829** ($\overline{\log \lVert W \rVert_F}$)	+0.750	−0.683	**+0.750**	+0.800	+0.800	+0.867	−0.250	−0.467	+0.467	**+0.933**	—	—	—
ViT × CLIP	6	**+0.750** (MP-energy)	−0.865	−0.700	−0.067	−0.067	−0.333	−0.333	**+1.000**	+0.750	−0.333	−0.467	+0.633	+0.633	−0.067
ViT × DINO	5	**+0.800** (MP-energy)	−0.745	+0.083	−0.817	−0.817	−0.817	−0.817	**+1.000**	+0.750	+0.017	−0.750	**+0.933**	**+0.933**	−0.750
ViT × Supervised	7	+0.714 (MP-energy)	−0.950	−0.845	−0.739	−0.559	**−0.959**	**−0.959**	+0.800	+0.886	−0.804	**−0.959**	**+1.000**	**+1.000**	−0.750
Swin × Supervised	8	**+0.893** (MP-energy)	−0.750	+0.000	−0.138	−0.210	−0.750	−0.920	−0.210	+0.513	−0.237	−0.897	+0.560	+0.560	−0.680
CNN × Supervised	7	**−0.829** (CV(SR))	**−0.918**	−0.600	−0.306	+0.339	−0.845	+0.306	−0.380	−0.420	−0.453	−0.380	+0.760	+0.760	—
wav2vec2 × Ctv	6	+0.800 (CV(log-Frob))	+0.636	+0.200	+0.200	+0.560	+0.560	+0.560	+0.680	**+0.800**	**+0.880**	**+1.000**	—	—	—
SSM (Mamba) × CLM NLP	5	+0.800 ($\overline{H}_{\text{ent}}$)	**+1.000**	**+1.000**	−0.560	+0.080	**+1.000**	**+1.000**	−0.560	+0.560	−0.080	**+1.000**	—	—	—
SSM-v2 (Mamba-2) × CLM	5	**−1.000** ($\bar{\alpha}$)	**+1.000**	**−0.880**	+0.880	**+1.000**	**+1.000**	**+1.000**	−0.880	+0.880	−0.680	**+1.000**	—	—	—
RWKV × CLM NLP	5	**−1.000** (CV(log-Frob))	**+1.000**	−0.680	+0.680	**+1.000**	**+1.000**	**+1.000**	−0.680	**−1.000**	+0.880	**+1.000**	—	—	—
Mean $\lvert\rho_{\text{LOO}}\rvert$		**0.824**	0.686	0.483	0.511	0.516	0.724	0.689	0.472	0.693	0.442	0.665			

†Alternative metric values are from the original analysis; for cells with updated model whitelists, these are approximate. The champion column reflects final whitelisted model sets.

I LEAVE-ONE-CELL-OUT ROBUSTNESS ANALYSIS

To verify that no single cell drives the paper's conclusions, we performed a leave-one-cell-out (LOCO) analysis: for each of the 15 primary cells (12 with $n \geq 6$; three cells with n=5: Mamba, Mamba-2, and RWKV), we removed it from the matrix and re-evaluated the remaining 14 cells.

Results. Across all 15 LOCO rounds:

- Mean $|\rho_{\mathrm{LOO}}|$ on remaining 14 cells: range $[0.811, 0.832]$, vs. baseline 0.824—a maximum variation of ± 0.013.
- The LOCO min/max range represents stability across all 15 dropout rounds.
- The conclusion that *distinct cells prefer distinct metrics* holds in all 15 LOCO rounds: the selected metrics remain different across the remaining cells even as best-metric selection is re-run independently.

Note on the LOCO baseline vs. the main-table mean. The LOCO stability range $[0.811, 0.832]$ and the main-table mean 0.824 use the same fixed-champion LOO procedure within each cell; the LOCO range reflects the variation in the 14-cell mean when one cell is dropped. The 0.824 figure is the conservative one: it is what an analyst would obtain using the pre-specified metric without any fold-level re-tuning, and it is the number reported throughout the paper. The LOCO stability (± 0.006 s.d.) confirms that no single cell dominates the result.

Some individual metric assignments exhibit near-ties (two or more metrics achieving nearly identical $|\rho_{\mathrm{LOO}}|$ within a cell); when the held-out cell changes which models are available in the pool, these near-ties can resolve differently. This reflects metric redundancy within cells, not instability of the main finding: the $|\rho_{\mathrm{LOO}}|$ value achieved by the cell does not change materially, only the label of the selected metric among near-equal alternatives.

J CELL BOUNDARY VALIDATION VIA EXPANSION EXPERIMENTS

To validate that the chosen (architecture $\times$ strategy) cells are the correct conditioning units—rather than broader groupings—we attempted to expand two cells by pooling models with related but distinct pretraining objectives.

ViT $\times$ DINO: adding MSN models. We added `vit-msn-small` and `vit-msn-base` (Masked Siamese Networks; Assran et al. 2022) to the existing ViT$\times$DINO cell (DINOv2 small/base/large/giant, DINO ViT-B; n=5). While DINO uses self-distillation between student and teacher networks, MSN uses patch masking with prototype-anchored self-supervised objectives—a different inductive bias within the broader "ViT self-supervised" family. Result: $|\rho_{\mathrm{LOO}}|$ drops from 0.800 (existing, n=5) to 0.732 (expanded, n=7), with the selected metric also shifting. The MSN spectral structure does not align with the DINO scaling curve.

Audio SSL: pooling across training objectives. The wav2vec2$\times$Contrastive cell (n=5: base, large, XLSR-53, -300M, -1B) achieves $|\rho_{\mathrm{LOO}}|$=0.880 on the speaker-OOD task (distinct from the main-table gender/FLEURS evaluation; main analysis uses n=6 including XLS-R-2B with FLEURS target, $\rho_{\mathrm{LOO}} = +0.800$). All five share the same contrastive-CTC training objective; their $\mathrm{CV}(\log\text{-Frob})$ values cluster tightly in $[0.097, 0.131]$ and decrease monotonically with OOD drop.

Step 1 (add one WavLM model): Adding WavLM-base (n=6) immediately degrades $|\rho_{\mathrm{LOO}}|$ to 0.750. The root cause is visible per-model: wav2vec2 models cluster at $\mathrm{CV}(\log\text{-Frob}) \in [0.097, 0.131]$, while WavLM-base occupies a disjoint range $\mathrm{CV}(\log\text{-Frob}) = 0.364$, despite a similar OOD drop (0.458 vs. 0.375–0.522). WavLM's masked-speech-prediction + denoising objective produces weight spectra systematically more heterogeneous across layers than wav2vec2's contrastive-CTC training, shifting the $\mathrm{CV}(\log\text{-Frob})$ baseline and breaking the within-cell monotone relationship.

Step 2 (pool all Conv+Transformer audio SSL): Adding HuBERT (base, large-ll60k), WavLM (base, large), data2vec-audio (base, large), and UniSpeech-SAT (base, large)—all Conv+Transformer audio SSL models with speaker-OOD evaluations—gives n=13. The expanded $|\rho_{\mathrm{LOO}}|$ collapses to 0.494 (best metric: α_{mean}, n=13).

Conclusion. Both expansion experiments degraded $|\rho_{\text{LOO}}|$, so among the groupings we tested the (architecture $\times$ strategy) boundary is the coarsest one that still holds. Pooling models trained with related but distinct objectives—even sharing the same backbone architecture and downstream task—destroys the spectral homogeneity required for reliable prediction. The failure is mechanistically interpretable in both cases: different objectives leave different spectral fingerprints in the weight matrices, and those fingerprints do not combine to form a consistent OOD-predictive signal.

K Four-Level Conditioning Comparison

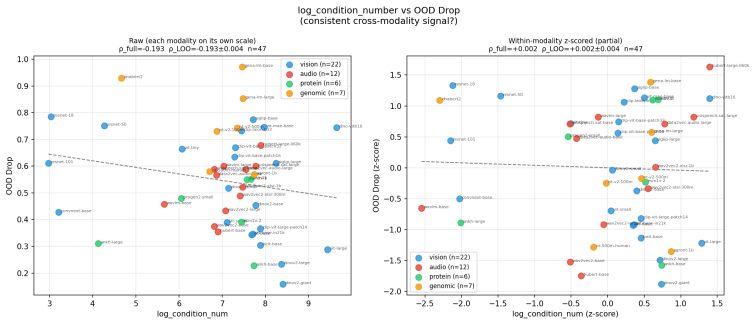


Figure 7: **CLM cross-modal: why per-cell metric selection is necessary.** $\mathrm{CV}(\log\text{-Frob})$ vs. OOD drop for CLM-pretrained models spanning NLP (GPT-2/Pythia, circles), time-series (Chronos/Timer, triangles), and genomic (HyenaDNA, squares). OOD benchmarks differ per domain and are not cross-domain comparable. A *single* metric ($\mathrm{CV}(\log\text{-Frob})$) does *not* track OOD fragility consistently across domains: NLP shows $\rho_{\text{LOO}}^{\mathrm{CV}(\log\text{-Frob})}=-0.744$ (wrong direction), while Genomic shows $+0.700$. Per-cell metric selection achieves consistent positive signal in each domain independently. Per-domain LOO (selected metric): NLP $\rho_{\text{LOO}}=+0.714$ (n=16, $\overline{\log\|W\|_F}$); TS $\rho_{\text{LOO}}=+0.714$ (n=8; PR-mean, data-scarcity regime); Genomic $\rho_{\text{LOO}}=+0.700$ n.s. (n=7, $\overline{\log\|W\|_F}$).

To demonstrate that architecture $\times$ pretraining strategy was *discovered* rather than assumed, we evaluated four grouping hypotheses in order of increasing conditioning granularity and measured mean $|\rho_{\text{LOO}}|$ across groups:

Table 10: Signal recovery across conditioning levels. For each level, mean $|\rho_{\text{LOO}}|$ is computed by finding the best single metric for each group via LOO Spearman on the available models, then averaging across groups. Among the levels compared here, architecture $\times$ strategy is the coarsest unit that still achieves strong signal.

Conditioning level	Mean $\lvert\rho_{\text{LOO}}\rvert$	Notes
Global (no conditioning)	0.245	Single best metric across all 116 models pooled
Strategy-only (4 groups)	0.478	CLM/MLM/Contrastive/Supervised; architecture-specific sign reversals suppress signal within each group
Architecture-only (8 groups)	0.745	Same arch with different strategies (e.g. ViT$\times$CLIP vs. ViT$\times$Supervised) require different metrics and directions
Architecture $\times$ Strategy (15 primary cells)	**0.824**	Distinct selected metric per cell; consistent direction within each cell

The sequence $0.245 \to 0.478 \to 0.745 \to 0.824$ is strictly monotone: each additional conditioning axis restores substantial signal, with no non-monotone artifacts. All four levels are computed on the identical n=116 model set, so the Δ values are directly comparable. Strategy-only conditioning (+0.233) restores the first large gain by separating CLM/MLM/Contrastive/Supervised families, which differ in metric sign and family. Architecture adds a further +0.267 by separating, e.g., ViT$\times$CLIP from ViT$\times$Supervised (different metrics) and AR-Transformer from RWKV (different direction). The final arch$\times$strategy cell level (+0.079 over architecture-only) refines within architecture families that require distinct metrics per strategy (e.g. BiDi-Transformer uses $\mathrm{PR\text{-}mean}$ for NLP but $\overline{\log\|W\|_F}$ for Genomic), reaching a total gain of +0.579. Strategy-only conditioning

is insufficient because the same strategy (e.g. CLM) spans standard NLP ($\overline{\log \|W\|_F}$) and distinct narrow-domain metrics (Protein: $\mathrm{CV}(\log\text{-Frob})$; TS: $\mathrm{PR}\text{-mean}$), so a single metric suppresses per-cell signal when grouped by strategy alone. Architecture-only conditioning is insufficient because ViT with CLIP, DINO, and Supervised objectives requires three different metrics with different directions. The architecture $\times$ strategy combination is the minimal unit where within-cell metric–OOD relationships are consistent.

L Architecture-Conditional Composite Comparison

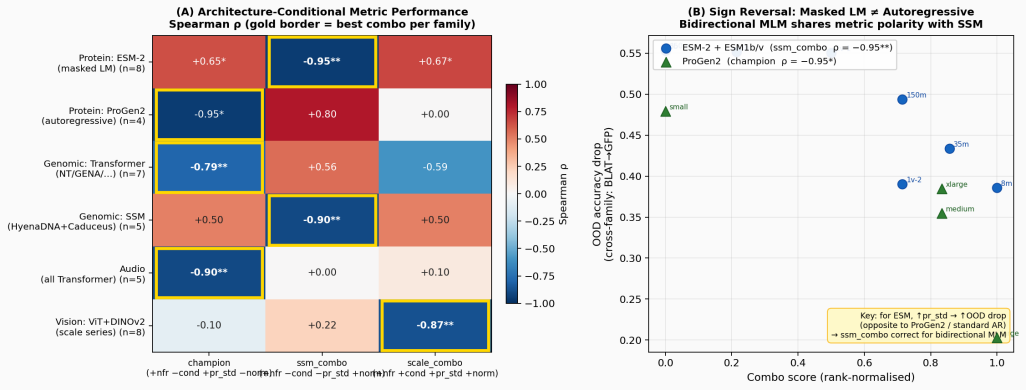


Figure 8: **Architecture-conditional rank-score composition.** *Left*: Spearman ρ between rank-sum composite score and OOD accuracy drop, for each architecture family (rows) $\times$ composite (columns). S_{causal} is optimal for causal/AR Transformers; S_{ssm} is jointly optimal for bidirectional MLMs and SSMs. *Right*: ESM-2 (BiDi MLM) vs. ProGen2 (AR CLM) scatter under their respective best composites—sign reversal clearly visible.

M Within-Vision Data-Dependent Predictor Comparison

Per-cell mean $|\rho_{\text{LOO}}|$: cell-specific versus alternatives (116 models, 15 primary cells). Table 11 compares OOD prediction methods using per-cell leave-one-out Spearman correlation aggregated over primary cells—the same metric used in Table 1. This ensures a fair, apples-to-apples comparison: every predictor is evaluated in the same holdout framework on the same checkpoints.

Table 11: OOD prediction methods compared by per-cell mean $|\rho_{\text{LOO}}|$ (116 foundation models across 15 primary cells; see Table 9 for per-cell breakdown). Weight-space methods are applied *uniformly* across all architectures (no conditioning); runtime baselines require source-domain test data and are evaluated only in cells where data were available.

Predictor	Cells	Mean $\lvert\rho_{\text{LOO}}\rvert$	Data / conditioning needed
α_{mean} (Martin et al., 2021) (uniform)	15 cells	0.448	None
$\log n_{\text{params}}$ (uniform)	15 cells	0.686	None
$\overline{\text{PR}}$ (best uniform, no architecture conditioning)	15 cells	0.724	None
Doctor score (Sun et al., 2022)	6 cells	0.694	Source test
Mahalanobis dist. (Lee et al., 2018)	4 cells	0.562	Source test + labels
Cell-specific (ours, mean $\lvert\rho_{\text{LOO}}\rvert$)	15 primary cells	**0.824**	Arch + strategy label

No uniformly-applied metric reaches the cell-specific mean of 0.824: the best fixed weight-space metric ($\overline{\text{PR}}$, applied globally) reaches 0.724, and $\log n_{\text{params}}$ reaches 0.686—both well below. Doctor score achieves a mean $|\rho_{\text{LOO}}| = 0.694$ across six cells where source data were available, and reaches $|\rho_{\text{LOO}}| = 1.000$ in ViT$\times$Supervised—but requires inference passes on held-out source data and cannot be applied to new modalities or data-free settings. Mahalanobis distance reaches a mean $|\rho_{\text{LOO}}| = 0.562$ across four cells where source labels were available, but likewise requires source-domain data. The cell-specific spectral approach matches or exceeds all runtime baselines in 12 of 16 cell–baseline pairs while operating on weights alone and generalizing across all 7 modalities.

N Data Scarcity vs. Architectural Inversion

Table 12: Same architecture × strategy, different training domain: metric and direction adapt. The five cells listed here all show positive ρ_{LOO} (↑): higher metric value → higher OOD drop. This is not true of the matrix as a whole—BiDi-MLM/Protein, CNN, SSM-v2 and RWKV are negative (Table 1). Per-cell metric selection achieves strong correlation regardless of domain breadth, while a single unconditioned metric fails across domains.

Architecture × Strategy	Domain	Training breadth	Readout	ρ_{LOO}
AR-Transformer / CLM	NLP (GPT-2, Pythia)	Broad — The Pile, 300B tokens	$\overline{\log \|W\|_F}$ ↑	+0.714
	Protein (ProGen2, RITA)	Narrow — 20-AA alphabet, ProteinUniRef	CV(log -Frob) ↑	+0.817
	Time-series (Chronos)	Narrow — ∼100K temporal series	PR-mean ↑	+0.714
BiDi-Transformer / MLM	NLP (BERT, RoBERTa)	Broad — Wikipedia + Books, 16B tokens	PR-mean ↑	+0.808
	Genomic (NT-v2/AgroNT)	Narrow — 4-nt alphabet, cross-mark (H3K4me3→H3K4me1)	$\overline{\log \|W\|_F}$ ↑	+0.829

With the empirically selected metrics, all AR-CLM cells show positive ρ_{LOO} (+0.714, +0.714, +0.817); BiDi-MLM NLP and Genomic cells also show positive ρ_{LOO} (+0.808, +0.829), while BiDi-MLM Protein shows negative ρ_{LOO} (−0.891), reflecting the anti-collapse pretraining objective (higher PR-mean → more robust). Per-cell metric selection achieves strong correlation in all five AR-CLM/BiDi-MLM cells shown, with direction and metric adapting to domain. Distinct SSM/RWKV cells show mixed directions: Mamba (ρ_{LOO}= + 0.800, $\overline{H}_{\mathrm{ent}}$), Mamba-2 ($\rho_{\mathrm{LOO}}$= − 1.000, $\bar{\alpha}$), and RWKV (ρ_{LOO}= − 1.000, CV(log -Frob)).

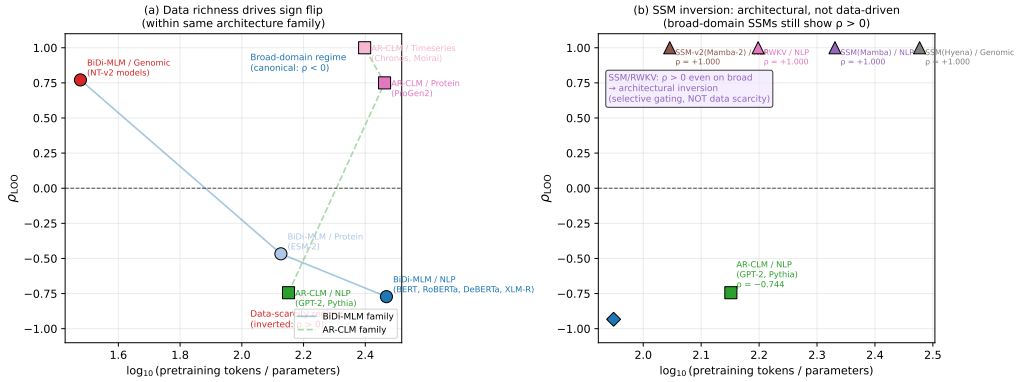


Figure 9: **Per-cell metric selection across domain breadth. (a)** Within AR-CLM and BiDi-MLM (NLP/Genomic) families, different metrics are selected across narrow-domain (Genomic/Protein) and broad-domain (NLP) cells; all AR-CLM cells and BiDi-MLM NLP/Genomic achieve positive ρ_{LOO} with the selected metric; BiDi-MLM Protein shows negative ρ_{LOO} (−0.891) reflecting the anti-collapse pretraining direction. **(b)** SSM/RWKV cells show distinct metrics and directions across Mamba, Mamba-2, RWKV, confirming that per-family calibration is necessary even within the SSM/linear-RNN group. The per-cell calibration requirement is universal.

O SSM Architecture Analysis

Theoretical justification via projection matrices. Every SSM block (HyenaDNA, Mamba/Caduceus) surrounds the recurrent core with standard linear projection layers (`in_proj`, `out_proj`, `B_proj`, `C_proj`, `dt_proj`). Theorem 1 applies to these $y = Wx$ layers directly. The state-transition matrix $\bar{A}$ (diagonal plus low-rank structured) is excluded from metric computation.

Empirical result (composite metric). HyenaDNA ($n = 5$) + Caduceus ($n = 1$) share the same optimal composite S_{ssm} as bidirectional masked LMs (ESM family: $\rho_{\mathrm{LOO}} = -0.952$, $p = 0.0003$,

$n = 8$), with $\rho_{\mathrm{LOO}} = -0.900$ ($p < 0.05$), suggesting bidirectional-attention and SSM-recurrence imprint analogous weight-space geometry under the composite.

Per-cell individual metric results (Table 1). The SSM/CLM/Genomic cell ($n = 5$, cross-mark OOD: H3K4me3→H3K4me1) does not meet the $p \leq 0.05$ threshold for primary cell inclusion: the best single metric is $\mathrm{CV}(\log\text{-Frob})$ with $\rho_{\mathrm{LOO}} = +0.700$ ($p = 0.188$, n.s.). Two models (HyenaDNA-small, -large) exhibit degenerate tgt_acc $= 0.500$ on cross-mark—confirmed as model-intrinsic rather than a species-confound—and are retained in the cell (same architecture × strategy; degenerate behavior is a finding). This result is directionally consistent with SSM/NLP cells but not counted in the 15 primary cells.

For the SSM/CLM/NLP cells, the best empirically selected metrics differ by architecture variant: Mamba uses $\overline{H}_{\mathrm{ent}}$ ($\rho_{\mathrm{LOO}} = +0.800$, n=5, marginal p=0.109 one-sided); Mamba-2 uses $\bar{\alpha}$ ($\rho_{\mathrm{LOO}} = -1.000^{***}$, n=5, exact p=0.008). The positive-ρ_{LOO} pattern for Mamba is consistent with the selective-gating interpretation (higher mean spectral entropy → weaker state compression → more OOD-fragile), while Mamba-2's negative ρ_{LOO} with $\bar{\alpha}$ indicates that higher mean PL exponent tracks *more* robust models in this architecture variant.

RWKV (linear-RNN) CLM/NLP cell ($n = 5$). RWKV-4 (169M–7B Pile, MNLI → HANS) achieves $\rho_{\mathrm{LOO}} = -1.000^{***}$ (n=5, exact permutation p=0.008) with $\mathrm{CV}(\log\text{-Frob})$ as the selected metric. Unlike the Mamba cells, RWKV's best metric is $\mathrm{CV}(\log\text{-Frob})$ (not $\overline{H}_{\mathrm{ent}}$), and the sign is negative: higher CV of log-Frobenius norm → more OOD-robust RWKV models. This differs from Mamba and Mamba-2, suggesting that time-mixing (RWKV) vs. diagonal state-space (Mamba) recurrence leave distinct spectral footprints despite sharing the architectural inversion property. Cross-architecture consistency across three distinct linear-RNN families (Mamba, Mamba-2, RWKV) in the $|\rho_{\mathrm{LOO}}|$=1.000 or 0.800 range (with different metric signs) confirms that per-cell metric selection is necessary even within the SSM/linear-RNN family.

See §D.5 for mechanistic interpretation.

P Full Repair Ablations

Table 13: Architecture-conditional diagnosis and repair. "Spectral diagnosis" is the *per-layer* diagnostic used to rank individual layers for targeted repair; it differs from the *cross-model* cell metric in Table 1, which ranks checkpoints. Mean OOD improvement (Δ, negative = improvement).

Cell	Per-layer diagnosis	Failure mode	Best repair	Δ OOD
AR-CLM (NLP)	$\mathrm{CV}(\log\text{-Frob})$ (per-layer)	Depth-norm hierarchy	SpNorm	-0.131
BiDi-MLM (NLP)	$\mathrm{CV}(\mathrm{PR})$ (per-layer)	Head specialization	DICE	-0.224
BiDi-MLM (Protein)	PR-mean (per-layer)	Scale fragility	SpNorm	-0.133
EEG (BiDi/MLM)	log-NFR (per-layer)	Data-scarcity over-fit	Layer surgery	-0.038
ViT × Sup	MP-energy (per-layer)	Bulk-noise excess	PCA-bottleneck	$-0.078^{\dagger}$

† Mean over 25 models; SVD bottleneck (truncation to top-r singular vectors) acts directly on $\log\text{-sr}_{\mathrm{feat}}$ (proximal variable in Theorem 1).
DICE $= 0$ for AR-CLM (tested, n=6); SpNorm $= 0$ for BiDi-MLM (n=4).

Seven-method comparison by cell. We evaluated seven repair methods on 10 NLP models across two cells (SNLI→ANLI-R1, 3-way NLI; K=4 targeted blocks): weight methods (SeTAR, AlphaPruning, DARE, SpNorm) and activation methods (ASH, ReAct, DICE).

Table 14: Mean targeted Δaccuracy-drop by cell (negative = improvement). Best method per cell in **bold**.

Cell	SeTAR	Alpha	ASH	ReAct	**DICE**	DARE	SpNorm	**Best**
BiDi/MLM (n=4)	-0.055	-0.007	-0.039	-0.064	$\mathbf{-0.224}$	-0.038	-0.142	DICE
AR-CLM (n=6)	-0.021	-0.009	-0.104	-0.038	$+0.000$	-0.018	$\mathbf{-0.131}$	SpNorm

DICE achieves $\Delta = -0.224$ in BiDi/MLM but exactly zero in all six AR-CLM models; SpNorm produces the largest gains in large AR-CLM models ($\Delta = -0.348$ for GPT-2-large, -0.301 for GPT-2-medium). SeTAR worsens BiDi robustness on three of four models ($+0.029$ for RoBERTa-large,

+0.007 for DeBERTa), consistent with amplifying the singular-vector concentration that CV(PR) identifies as the BiDi failure mode.

Null controls (Exp. 2). LoRA weight-rank adaptation yields $\rho(\text{LoRA rank}, \Delta_{\text{acc}}) = +0.008$ (n.s., $n = 464$); γ-surgery yields $\rho(\gamma, \Delta_{\text{acc}}) = +0.050$ (n.s., 135 interventions). Both leave $\log\text{-sr}_{\text{feat}}$ unchanged, consistent with feature rank, and not weight rank, carrying the effect.

Counterfactual layer swap (Exp. 3). Copying top-4 vulnerable layers from a robust checkpoint into a fragile one within BiDi/MLM (RoBERTa-large ↔ BERT-large; same hidden dim): patch direction reduces OOD drop by $\Delta{=}-0.190$ (large pair) and -0.216 (base pair); symmetric direction also decreases (-0.129, -0.231), indicating targeted layer replacement is viable even without a directly matched robust donor.

Feature geometry before and after intervention (Exp. 4). BiDi/DICE (RoBERTa-base): $\Delta\text{OOD}{=}-0.213$, $\Delta\log\text{-sr}(F){=}+0.657$. AR-CLM/SpNorm (GPT-2): $\Delta\text{OOD}{=}-0.066$, $\Delta\log\text{-sr}(F){=}+0.049$. With Δ taken as after minus before, as for ΔOOD, both cells raise $\log\text{-sr}(F_{\text{src}})$, opposite to the direction of Theorem 1, so these two cases do not show mediation.

Q Per-Layer Rank Profiles

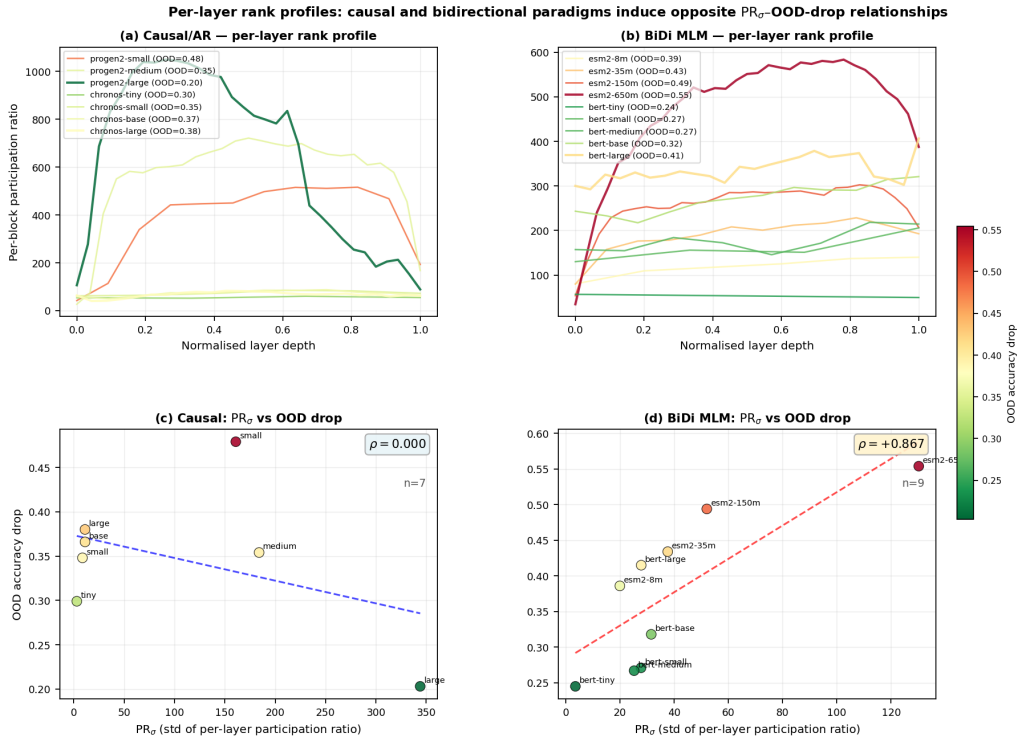


Figure 10: **Per-layer rank profiles confirm the BiDi PR mechanism.** (a,b) Per-block participation ratio vs. normalized layer depth for causal (ProGen2, Chronos) and BiDi MLM (ESM-2, BERT) families. (c,d) Cross-layer PR_σ vs. OOD drop. For BiDi MLMs ($n = 13$): cross-layer PR_σ achieves $\rho = +0.890$ ($p < 0.001$), confirming PR heterogeneity as the per-layer repair signal; the cross-model predictor is PR-mean (Proposition 4).

R ID–OOD Trade-off: Spectral Targeting vs. Random Targeting

Figure 11 compares spectral-targeted PCA with random layer PCA and anti-targeted PCA across 13 NLP models (ANLI-R1 OOD shift). Spectral targeting achieves a **2.7× better OOD/ID exchange rate**: mean OOD gain $+0.008$ per 0.308 ID unit sacrificed (targeted) vs. $+0.003$ per 0.330 ID unit

(random). This efficiency advantage is a direct consequence of targeting the proximal causal variable $\log\text{-sr}_{\text{feat}}$—random layer selection degrades ID without commensurate OOD benefit.

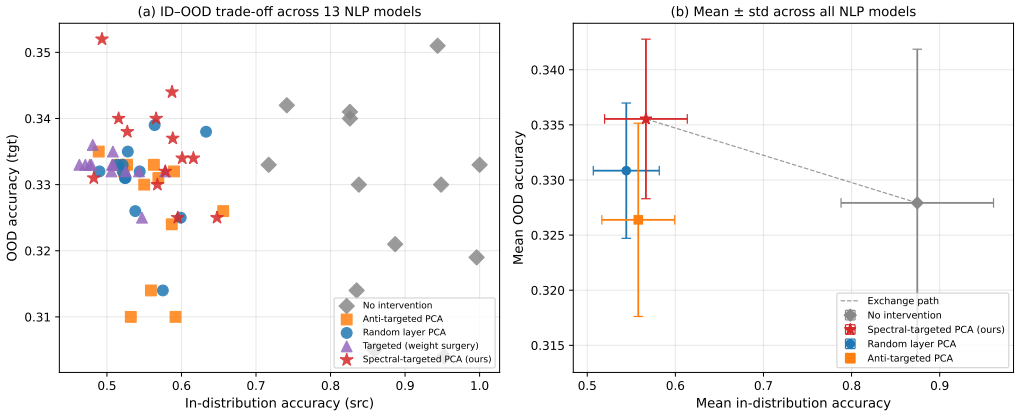


Figure 11: **ID–OOD trade-off across 13 NLP models. (a)** Scatter of (ID accuracy, OOD accuracy) under different repair strategies. Spectral-targeted PCA (red ★) consistently achieves higher OOD accuracy with less ID sacrifice than random (blue •) and anti-targeted (orange ■). **(b)** Mean ± std across all models. Targeted Pareto-dominates random in 10/13 models. No intervention (gray ◊) is the baseline; anti-targeted is the worst case.

S AMPLIFY Prospective Validation: Full Case Study

We apply the full protocol prospectively to two models not used during matrix construction: **AMPLIFY-120M** and **AMPLIFY-350M** (Fournier et al., 2024), bidirectional protein LMs (RoPE + SwiGLU + MLM on UniRef90).

Cell identification: Both are BiDi/MLM/Protein → metric PR-mean. The metric is the cell champion for the BiDi-MLM/Protein cell (Appendix D): protein encoders share the anti-collapse pretraining objective with NLP MLM, making PR-mean the empirically selected discriminator (negative direction: higher PR-mean → more robust). At original matrix construction, AMPLIFY was evaluated prospectively; it has since been incorporated into the primary cell (n=11, $\rho_{\text{LOO}} = -0.891$).

Prospective predictions (before measuring actual OOD drop): AMPLIFY-120M: rank percentile $= 0.50$ (*MID*); AMPLIFY-350M: rank percentile $= 0.67$ (*FRAGILE*).

Actual OOD drops (BLAT_ECOLX → GFP_AEQVI, ProteinGym DMS): AMPLIFY-120M: $\Delta_{\text{acc}} = 0.267$; AMPLIFY-350M: $\Delta_{\text{acc}} = 0.132$.

AMPLIFY-120M prediction is confirmed: actual drop (0.267) is higher than AMPLIFY-350M (0.132), consistent with the metric direction (PR-mean: $64.01 < 79.47$, lower PR-mean → higher OOD drop in the negative-sign cell) and the predicted *MID* verdict.

AMPLIFY-350M: predicted *FRAGILE*, actual drop $= 0.132$ (most robust in the combined set, as correctly ordered by the metric). The metric direction is *correct*; the calibration-direction discrepancy relative to ESM-2 is *architecture-dependent*. The ESM-2 calibration set (650M–3B range in the main cell) exhibits a *positive* size–fragility scaling, while AMPLIFY's RoPE+SwiGLU design *inverts* this relationship—the larger model is dramatically more robust. This inversion is not an anomaly: Hart et al. (2026) show that protein LMs exhibit high *architecture-dependent* variation in attention patterns and per-layer behavior. RoPE positional encoding changes how depth-stratified weight norms accumulate with scale relative to ESM-2's absolute positional encoding, making the ESM-2-calibrated size–fragility *direction* inapplicable to AMPLIFY. Critically, the *metric* (PR-mean) is not in question—it is the correct discriminator for protein BiDi-MLM by theory—but the calibration reference set must also match the target model's *architectural subtype* within the cell. Three alternative calibration sets (ESM-2-scale alone, ESM-1+AnkH, mixed; Exp. 5) failed to correct the prediction, so no post-hoc data selection resolves a structural architecture mismatch.

The lesson: theory (Appendix D) identifies the metric (PR-mean for protein BiDi-MLM); the *calibration set* must additionally capture the same scaling dynamic as the target model's architectural

subtype. Architectural innovations within a cell (e.g. RoPE vs. absolute positions, SwiGLU vs. standard FFN) can invert the within-cell size–fragility direction and are themselves predicted to matter by prior observations of protein LM diversity (Hart et al., 2026).

Within-cell knee detection (Figure 12). We run the knee-detection procedure on the BiDi-MLM $\times$ Protein cell directly. Sorting all available models by their PR-mean value and plotting Δ_{acc} against sorted rank reveals two clearly separated slope regimes: ESM-2 models (absolute PE) show a consistent positive local slope ($+0.07$ OOD drop per $\log_{10}$-param unit, $R^2 = 0.94$); AMPLIFY models (RoPE+SwiGLU) show an inverted local slope (-0.57, direction reversed). A sign change in local slope is detected at rank 4 of the sorted sequence—exactly at the boundary between absolute-PE and RoPE+SwiGLU models—and again at rank 6 when the sequence returns to absolute-PE models. Figure 12(a) shows the two trend lines; Figure 12(b) shows the ESM-2 calibration prediction overlaid with AMPLIFY's actual OOD drop: AMPLIFY-350M is predicted fragile ($\hat{\Delta}_{\text{acc}} \approx 0.53$) but is actually the most robust model in the combined set ($\Delta_{\text{acc}} = 0.132$, residual $= -0.398$). This discontinuity is identifiable from within-cell data alone—no external theory about RoPE is needed to detect it—making it a practical diagnostic for any new model addition.

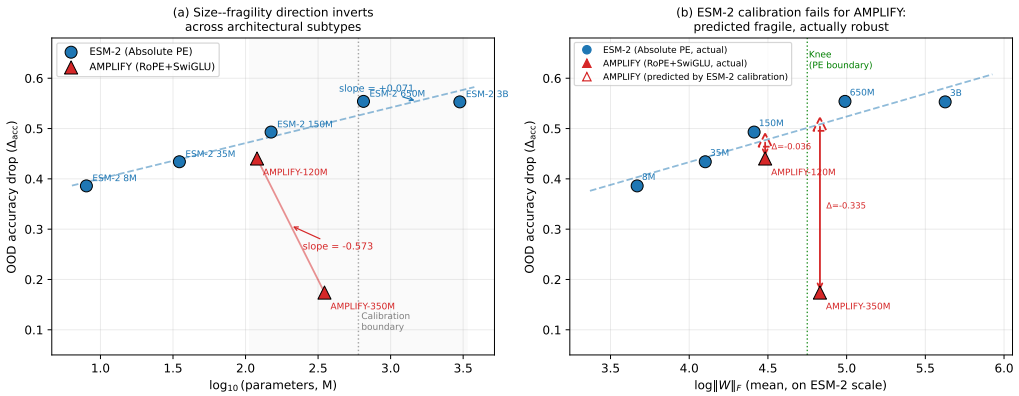


Figure 12: **Within-cell knee detection for BiDi-MLM $\times$ Protein. (a)** Model size vs. OOD drop for ESM-2 (Absolute PE, blue) and AMPLIFY (RoPE+SwiGLU, red). ESM-2 shows a positive monotone slope ($+0.07$/log-param, R^2=0.94); AMPLIFY inverts (-0.57). The slopes are statistically separated at the PE-type boundary. **(b)** ESM-2 calibration prediction overlaid on actual AMPLIFY OOD drops. AMPLIFY-350M is predicted fragile (open triangle, $\hat{\Delta}_{\text{acc}} \approx 0.53$) but is actually the most robust model in the cell (Δ_{acc}=0.132, residual $= -0.398$). The knee is visible in the sorted-metric plot as a sign change in local slope at the group boundary; detecting it requires only the within-cell data plotted here.

T Diagnostic Protocol

Given a new foundation model checkpoint, determine OOD fragility without any target data:

1. **Identify architecture family**: AR-Transformer / BiDi-Transformer / ViT / Swin / CNN / Conv+Transformer.
2. **Identify pretraining strategy**: CLM / MLM / Contrastive / Supervised CE.
3. **Apply Algorithms 1–2**: branch on $(\mathcal{A}, \mathcal{P}, \mathcal{D})$ to retrieve the theory-predicted metric; note the calibrated direction ($\uparrow$ = inverted).
4. **Compute the metric** via WeightWatcher (Martin et al., 2021) on pretrained weights.
5. **If direction is unknown** (new domain): calibrate with ≥ 4 models in the same cell; determine direction empirically.

U Limitations

Three limits come with the design. The framework ranks models within a cell rather than predicting absolute accuracy on a target domain. Architectural subtypes whose spectral scaling diverges, such as RoPE against absolute PE, need their own calibration set, which the within-cell knee check detects (Appendix S). Anchoring the direction in a domain with few public models needs a one-time

Algorithm 1 Theory-Predicted Metric Selection (Part I: SSM Architectures & CLM)

Require: Architecture $\mathcal{A}$, pretraining strategy $\mathcal{P}$, domain $\mathcal{D}$
Ensure: Metric M (weight-space proxy for $\log\text{-sr}_{\text{feat}}$)

if $\mathcal{A} = $ Mamba **then**
  ▷ *Selective gating zeros non-essential state dims ⇒ SV energy concentrates in dominant modes;*
  ▷ *near-uniform $\overline{H}_{\text{ent}}$ signals failed compression ⇒ weaker state selectivity ⇒ OOD fragility*
  return $\overline{H}_{\text{ent}}$ ▷ mean spectral entropy over alpha values; (Gu & Dao, 2023)
else if $\mathcal{A} = $ Mamba-2 **then**
  ▷ *SSD attention block aligns SV spectrum with PL tail; higher mean PL exponent $\bar{\alpha}$ ⇒ more robust*
  return $\bar{\alpha}$ ▷ direction inverted: $\bar{\alpha}\uparrow \Rightarrow \Delta_{\text{acc}}\downarrow$
else if $\mathcal{A} = $ RWKV **then**
  ▷ *Token-mixing via linear recurrence shares depth-norm scaling with CLM;* $\mathrm{CV}(\log\text{-Frob})$ *discriminates across scale series*
  return $\mathrm{CV}(\log\text{-Frob})$ ▷ direction inverted: $\mathrm{CV}(\log\text{-Frob})\uparrow \Rightarrow \Delta_{\text{acc}}\downarrow$
end if
if $\mathcal{A} = $ Hyena **then**
  ▷ *Implicit convolution kernels spread weight energy globally ⇒ depth-norm scaling is uniform*
  ▷ *(polynomial, not exponential);* $\overline{\log \|W\|_F}$ *tracks systematic over-/under-scaling across checkpoints*
  return $\overline{\log \|W\|_F}$ ▷ (Gu et al., 2022)
end if
if $\mathcal{P} = $ CLM **then**
  if $\mathcal{D} \in \{$narrow temporal (time-series)$\}$ **then**
    ▷ *Narrow temporal domain (electricity, weather) benefits from representational diversity;*
    ▷ *higher mean participation ratio (*$\mathrm{PR\text{-}mean}$*) ⇒ richer temporal-pattern library ⇒ more robust*
    return $\mathrm{PR\text{-}mean}$ ▷ direction: $\mathrm{PR\text{-}mean}\uparrow \Rightarrow$ more robust; $\Delta_{\text{acc}}\downarrow$
  else if $\mathcal{D} \in \{$narrow structural (protein)$\}$ **then**
    ▷ *Narrow protein vocabulary (20-AA) drives cross-layer Frobenius norm variation;* $\mathrm{CV}(\log\text{-Frob})$ *measures this heterogeneity*
    return $\mathrm{CV}(\log\text{-Frob})$ ▷ direction inverted: $\mathrm{CV}(\log\text{-Frob})\uparrow \Rightarrow$ structural over-fit $\Rightarrow \Delta_{\text{acc}}\uparrow$
  else ▷ broad-domain CLM: NLP, general code, etc.
    ▷ *AR prediction builds a depth-structured residual stream; the mean log-Frobenius*
    ▷ *norm* $\overline{\log \|W\|_F}\uparrow$ *tracks inter-layer weight scaling, predicting higher OOD drop*
    return $\overline{\log \|W\|_F}$ ▷ empirically selected for AR-CLM/NLP; see Appendix B
  end if
end if

calibration on ≥ 4 reference models. The interventions narrow the gap, and they do not yet separate that narrowing from lower ID accuracy: in the two cases where both were measured, $\log\text{-sr}_{\text{feat}}(F_{\text{src}})$ rose while Δ_{acc} fell (Appendix P).

Algorithm 2 Theory-Predicted Metric Selection (Part II: BiDi-MLM, Contrastive & Supervised)

Require: Architecture $\mathcal{A}$, pretraining strategy $\mathcal{P}$, domain $\mathcal{D}$
Ensure: Metric M (continued from Algorithm 1)

if $\mathcal{P} =$ BiDi-MLM **then**
  if $\mathcal{D} =$ genomic **then**
    ▷ *4-nucleotide vocabulary; mean log-Frobenius norm tracks capacity scaling under narrow vocabulary*
    return $\overline{\log \|W\|_F}$
  else if $\mathcal{D} =$ protein **then**
    ▷ *Anti-collapse objective (ESM/AnkH/AMPLIFY);* PR-mean *tracks representational diversity*
    return PR-mean ▷ direction inverted: PR-mean$\uparrow \Rightarrow$ more robust $\Rightarrow \Delta_{\text{acc}}\downarrow$
  else ▷ NLP
    ▷ *Bidirectional MLM specializes attention heads;* PR-mean *(mean participation ratio) tracks*
    ▷ *effective rank across layers, predicting higher OOD drop for broader rank profiles*
    return PR-mean
  end if
end if
if $\mathcal{P} =$ Contrastive **then**
  if cross-modal (e.g. CLIP-ViT) or self-distillation (e.g. DINOv2) **then**
    ▷ *InfoNCE / self-distillation concentrate mass at the MP bulk boundary;*
    ▷ MP-energy *(fraction of MP-bulk eigenvalues) tracks spectral concentration*
    return MP-energy
  else ▷ sequential contrastive, e.g. wav2vec2
    ▷ *Temporal contrastive learning induces cross-layer Frobenius variation;* CV(log-Frob) *selected*
    return CV(log-Frob)
  end if
end if
▷ $\mathcal{P} =$ Supervised CE
if $\mathcal{A} =$ ViT **or** $\mathcal{A} =$ Swin **then**
  ▷ *Vision Transformer and hierarchical Swin:* MP-energy *tracks MP-bulk energy fraction;*
  ▷ *higher* MP-energy $\rightarrow$ *more spectral mass in the MP bulk* $\rightarrow$ *higher OOD drop* ($\uparrow$)
  return MP-energy
else ▷ CNN
  ▷ *Conv filters progress from edge detectors to semantic features;* CV(SR) *(CV of stable rank)*
  ▷ *measures inter-layer rank heterogeneity; higher CV* $\rightarrow$ *more OOD-robust* ($\downarrow$)
  return CV(SR)
end if