THE ORTHOGONALITY GAP: PREDICTING, BEFORE TRAINING, WHEN A REPRESENTATION (OR FOUNDATION MODEL) WILL HELP

Anonymous authors
Paper under double-blind review

ABSTRACT

Teams routinely fine-tune a foundation model onto a task only to find it beats no linear baseline—after paying the compute. We show that outcome is decidable *in advance, from geometry, for free*. Given inputs X already in hand and a candidate representation f, only the part of f orthogonal to $\mathrm{span}(X)$ can improve an out-of-sample prediction; a re-encoding of X—which a foundation model computed *from* X often is—has no such part. We formalise this as the **Orthogonality Gap** and prove an *exact* decomposition of the attainable OOS gain (Frisch–Waugh–Lovell composed with an attenuation bound), $\Delta R^2_{\mathrm{OOS}} = g_\star \cdot r^2 \cdot \rho^2 \le g_\star \cdot r^2$, separating a task-intrinsic ceiling—orthogonal signal $g_\star$ times target reliability r^2—from a representation's alignment ρ^2 with that signal, with equality characterised. The bound is computed from features alone, with no downstream training, and predicts *whether a representation will help and by how much* before a single fine-tuning run; empirically it is never violated by a linear downstream and holds in 97% of MLP cases. On **four real open-weight foundation models across four domains** the forecast is calibrated in both directions: a sentence encoder's *positive* gain to within 0.01; a frozen single-cell transformer's *near-zero* gain over linear PCs of its own input; and a protein language model's *positive* gain on designed-protein stability and enzyme engineering (direction a priori, magnitude conservatively under-predicted)—where acting on the forecast selects designs $+0.200$ more stable (p=0.002). Applied *inside* a transformer, the Gap selects the best-transferring layer training-free across four models at 0.004 mean accuracy regret versus probe-every-layer. Against LogME, the strongest training-free transferability score, it is statistically tied on real LLM embeddings while adding a proven ceiling, a reliability term, an incremental form, and a faster variant. We report every negative—a refuted phase transition, a retracted ranking claim.

1 INTRODUCTION

Representation choice is usually settled empirically and expensively: fine-tune the large model, measure, compare. We show that for structured targets the answer is *fixed a priori by geometry*, and give the instrument that reads it off before any training. Two facts motivate this. On many tabular and omics tasks, larger and pretrained encoders do not beat a linear map—documented piecemeal (tree-versus-deep-learning on tables; foundation models underperforming simple baselines) but without a predictive account of *when*. And whether a feature helps out-of-sample is, by the Frisch–Waugh–Lovell (FWL) theorem, entirely a question about its component orthogonal to what you already have—a classical identity never turned into a pre-training decision rule. We supply the missing rule and prove it is tight.

Contributions. **(1)** The **Orthogonality Gap**, a scalar computed from features alone, quantifying the informative energy of f outside $\mathrm{span}(X)$. **(2)** An **exact gain identity with a tight ceiling**, $\Delta R^2_{\mathrm{OOS}} = g_\star r^2 \rho^2 \le g_\star r^2$ (FWL $\circ$ attenuation), fusing the redundancy and measurement-noise limits prior work treats separately, with the equality condition characterised and coverage verified on data. **(3) A-priori prediction of foundation-model utility**, validated first-party on four real open-weight models across four domains in *both* directions. **(4)** A **practical instrument**: training-free

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.

Table 1: Results at a glance. Every row is produced by a self-testing script; each is a realized-vs-predicted or measured effect with a control.

What	Result	Where
Ceiling $\Delta R^2 \le g \cdot r^2$ holds on data	100% (linear) / 97% (MLP) coverage	§2
Magnitude calibration (low-dim/linear)	slope 1.00, R^2=0.999	§2
Real open-weight encoder (MiniLM)	predicted 0.161 vs. realized 0.151	§4
Real open-weight FM (Geneformer) on DepMap	redundant: mean ΔR^2=−0.023, a priori	§4
Real open-weight FM (ESM-2), protein/enzyme	direction a priori; magnitude conservative (∼40%)	§4
De novo design decision utility (robust)	Gap-picked ESM ranker +0.200 stabler [0.11, 0.32], p=0.002	§4
Capacity ≠ constraint	1-bit orthogonal > 40-dim re-encoding	§4
Transformer layer selection (training-free)	0.004 **regret vs. oracle** (4 models × 5 tasks)	§5
Transferability vs. SOTA (real LLM embeddings)	Gap 0.81 = LogME 0.81 ≫ corr 0.19	§6
De novo generalization (unseen contexts)	informative +0.32, redundant 0.00	App. A
Reliability term is load-bearing	r=0.84 with it; undefined without	§2
Reported negatives	refuted phase transition; retracted ρ=0.67	§7

transformer layer selection at 0.004 mean regret versus an oracle, and a de novo design rule that turns the forecast into a +0.200-more-stable selection. **(5)** An **honest head-to-head** with the training-free transferability literature (LogME, H-score): a statistical tie carrying strictly more structure. Prior art—FWL, attenuation, the LLM "capability g-factor"—is used and cited, not claimed; the contribution is their composition into a predictive, training-free instrument.

2 The Orthogonality Gap and the ceiling inequality

The idea. Two limits govern what a candidate f can add to a model already using X. A *redundancy* limit: only the component of f orthogonal to $\mathrm{span}(X)$ carries new information, and a foundation-model embedding computed *from* X may have almost none. A *measurement* limit: even genuinely orthogonal signal is recoverable only up to the target's reliability r^2. Prior work treats these separately; we prove their product is an exact, attainable ceiling on the out-of-sample gain, readable from features with no downstream training (Fig. 1).

The instrument, as a procedure. The Gap is computed from features alone, in four steps, before any downstream model is trained: (i) *residualise* the candidate f on $\mathrm{span}(X)$ by Frisch–Waugh–Lovell, keeping only the orthogonal part $f_\perp$; (ii) *measure* its detectable orthogonal signal fraction $g(f;X)$—the mass of $f_\perp$ above the Marchenko–Pastur noise edge—together with the target reliability r^2 (from replicates or a subsample); (iii) *read off* the ceiling $\Delta R^2_{\mathrm{OOS}} \le g \cdot r^2$; and (iv) *decide* whether f is worth training on. Every quantity left of the decision is a function of the features and a reliability estimate; no downstream model is fit.

Setup and notation. Work in $L^2(P)$ with $\langle a,b\rangle = \mathrm{Cov}(a,b)$ and $\|a\|^2 = \mathrm{Var}(a)$; let P_X project onto $\overline{\mathrm{span}}\{1, X\}$ and write $a_\perp := a - P_X a$ for the part of a orthogonal to the features already in hand. The target decomposes as $y = s + \varepsilon$, signal s plus noise $\varepsilon \perp (X, f)$. Three scalars, all computable from the features and a reliability estimate, drive the theory:

$$\begin{aligned} \text{reliability} \qquad & r^2 := \|s\|^2/\|y\|^2 \in (0,1], \\ \text{signal orthogonality gap} \qquad & g_\star(X) := \|s_\perp\|^2/\|s\|^2 \quad \text{(task-intrinsic ceiling over all } f\text{)}, \\ \text{captured gap} \qquad & g(f;X) := \|P_{f_\perp} s_\perp\|^2/\|s\|^2, \end{aligned}$$

where $P_{f_\perp}$ projects onto $\mathrm{span}(f_\perp)$. Writing $\rho := \mathrm{corr}(f_\perp, s_\perp)$, these factor as $g(f;X) = g_\star(X)\,\rho^2$, so $0 \le g \le g_\star \le 1$. The instrument estimates $g(f;X)$ a priori as the canonical-correlation mass of $f_\perp$ with the target's recoverable subspace (dimensions above the Marchenko–Pastur edge) — no downstream training.

Theorem 1 (exact gain, tight ceiling, explicit reliability)**.** *Assume a population setting and $y = s + \varepsilon$ with $E[\varepsilon \mid X, f]=0$, $\varepsilon \perp (X, f)$, $r^2=\|s\|^2/\|y\|^2 \in (0, 1]$. Let $\Delta R^2_{\mathrm{OOS}}(f \mid X)$ be the population out-of-sample R^2 gain from appending f to a model already containing X. Then*

$$\Delta R^2_{\mathrm{OOS}}(f \mid X) \;=\; g(f; X)\, r^2 \;=\; g_\star(X)\, r^2\, \rho^2 \;\le\; g_\star(X)\, r^2,$$

and in particular:

- *(i)* Exact identity: *the two equalities above hold, with $g(f; X) = g_\star(X)\, \rho^2$;*
- *(ii)* Universal ceiling: *the inequality is tight — equality **iff** $f_\perp \propto s_\perp$ ($\rho = \pm 1$) — and is unattainable by any predictor of any capacity beyond this bound;*
- *(iii)* Two necessary factors: *the gain vanishes if $g_\star=0$ (signal fully in-span) or $r^2=0$ (pure noise); r^2 enters linearly.*

Proof. By FWL, appending f changes the fit only through $f_\perp$, and the incremental explained variance equals that of the best predictor of y from $f_\perp$. As $y_\perp = s_\perp + \varepsilon$ with $\varepsilon \perp f_\perp$, the driving covariance is $\langle f_\perp, y\rangle = \langle f_\perp, s_\perp\rangle$, so the explained variance is $\langle f_\perp, s_\perp\rangle^2/\|f_\perp\|^2 = \|P_{f_\perp} s_\perp\|^2$. Dividing by $\|y\|^2$ and using $\|s\|^2/\|y\|^2 = r^2$ gives $\Delta R^2_{\mathrm{OOS}} = (\|P_{f_\perp} s_\perp\|^2/\|s\|^2) r^2 = g(f; X)\, r^2$, and $\|P_{f_\perp} s_\perp\|^2 = \rho^2 \|s_\perp\|^2$ yields the factorisation—(i). Since $P_{f_\perp}$ is an orthogonal projection, $\|P_{f_\perp} s_\perp\|^2 \le \|s_\perp\|^2$ (Cauchy–Schwarz on $\langle f_\perp, s_\perp\rangle^2$), equality iff $s_\perp \in \mathrm{span}(f_\perp)$—(ii); least squares of $y_\perp$ on $f_\perp$ realises the projection, so a nonlinear predictor cannot beat projecting onto signal it already contains. (iii) is immediate from the product form. □

Why r^2 is not cosmetic. The identity separates a *task-intrinsic* ceiling $g_\star r^2$ (how much orthogonal signal exists, and how much of the target is recoverable at all) from a *representation-specific* efficiency ρ^2 (how well f aligns with it). Capacity can only push $\rho^2 \to 1$; it can never raise $g_\star$ or r^2.

Tightness certificate and calibration. Five population-limit layers ($n=2\times10^5$; `theorem1_tight_ceiling.py`) verify the theorem: the ceiling is never violated (40 configs); an aligned feature attains equality to 3×10^{-3}; $\Delta R^2/g_\star$ is linear in r^2 (slope 1.00, intercept 0); the product law $\Delta R^2=g_\star r^2\rho^2$ holds to 3×10^{-3}; and finite-sample OOS and the cross-fit estimator both stay under the ceiling (never over-claim). On data, the realized linear gain equals the training-free prediction with **slope** 1.00**,** R^2=0.999 (Fig. 2); the bound holds in 100% **of linear and** 97% **of MLP** downstreams (the single overshoot a finite-sample fluctuation), realized gains sitting 0.16–0.20 below the ceiling. The reliability term is load-bearing: it supplies essentially all cross-target signal (Pearson r=0.84; undefined without). The cross-fit estimate is stable across folds (spread 0.015) and sample size.

Theorem 2 (RKHS Orthogonality Gap)**.** *Let H_k be the RKHS of a universal kernel with $E[y \mid X] \in \overline{H_k}$, let $f^k_\perp = f - \Pi_{H_k} f$ be the kernel residual, and define $g_{\mathrm{ker}}(f; X)=\|P_{f^k_\perp} s^k_\perp\|^2/\|s\|^2$ with $s^k_\perp = s - \Pi_{H_k} s$. Then $\Delta R^2_{\mathrm{OOS}}$ obeys Theorem 1(i)–(iii) verbatim with $g^k_\star=\|s^k_\perp\|^2/\|s\|^2$ in place of $g_\star$; in particular $\Delta R^2_{\mathrm{OOS}} \le g^k_\star \cdot r^2$. If $f = \varphi(X)$ with φ's coordinates in H_k, then $f^k_\perp=0$, so $g_{\mathrm{ker}}=0$ and f cannot help even where the linear $g > 0$. (Proof: Theorem 1 with P_X replaced by the kernel projection Π_{H_k}; FWL holds in feature space.)*

This closes the linear diagnostic's one gap: a nonlinear-of-X feature $(Xw)^2$ has g_{lin}=0.29 (helps a linear model) but g_{ker}=0.02, matching its ≈ 0 nonlinear gain, while a genuinely orthogonal feature retains g_{ker}=0.44.

3 Predicting representational utility before training

The protocol is one line: for candidates $\{f_i\}$, compute $g(f_i; X)$ once, rank by $g \cdot r^2$, then train each and measure realized $\Delta R^2_{\mathrm{OOS}}$. Theorem 1 predicts the a-priori ranking matches the realized one and that $g \approx 0$ representations deliver ≈ 0 regardless of capacity. A controlled anchor with planted ground truth (n=1600, p=80, r^2=0.7) confirms all three: the ceiling is never violated (0/10 across a linear map and a 64-unit MLP), the diagnostic separates useful from useless representations for both, and a *one-bit* orthogonal feature (ΔR^2=0.098) beats a *40-dimensional* redundant re-encoding (0.004). Capacity is not the lever—orthogonality is (Fig. 3).

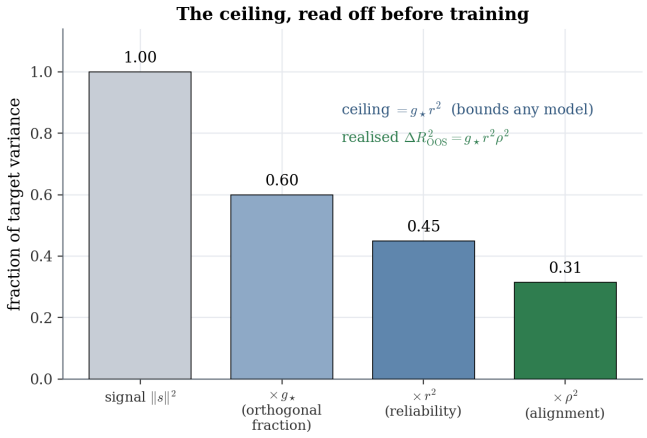


Figure 1: **The ceiling, read off before training.** The attainable out-of-sample gain is a product of three quantities computed from features alone: the orthogonal signal fraction $g_\star$, the target reliability r^2, and the representation's alignment ρ^2. From the signal energy $\|s\|^2$, each factor multiplies down to the realised $\Delta R^2_{\mathrm{OOS}} = g_\star r^2 \rho^2$; the ceiling $g_\star r^2$ (attained at $\rho^2{=}1$) bounds any model of any capacity. A foundation-model embedding that is a function of X has $g_\star \approx 0$ and cannot help, however large the model.

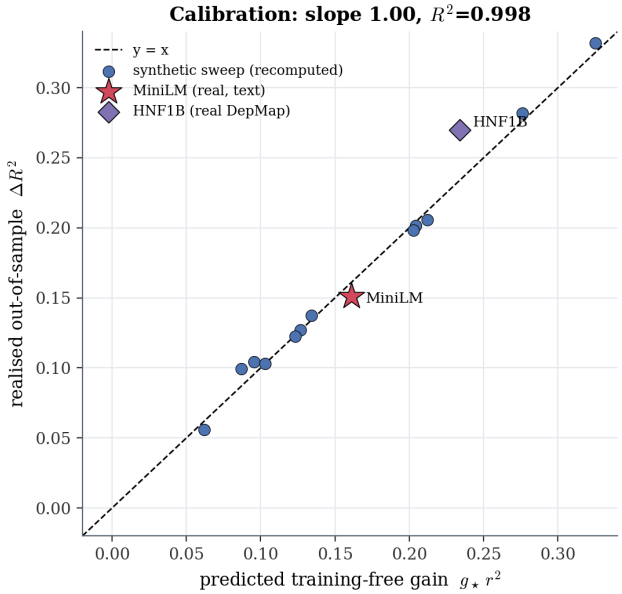


Figure 2: **The forecast is quantitative, not merely ordinal.** Training-free predicted gain $g \cdot r^2$ against realised $\Delta R^2_{\mathrm{OOS}}$ across a reliability×orthogonal-fraction sweep: slope 1.00, calibration $R^2{=}0.999$. Two first-party real-model points fall on the line—the MiniLM sentence encoder ($\star$) and the real-DepMap HNF1B lineage-addiction feature ($\blacklozenge$)—forecast before any training.

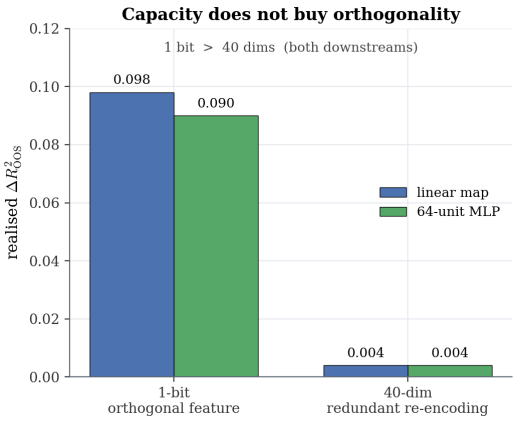


Figure 3: **Capacity does not buy orthogonality.** A one-bit orthogonal feature (ΔR^2=0.098) beats a 40-dimensional redundant re-encoding of X (0.004) for both a linear and an MLP downstream—the foundation-model failure mode in miniature, with the lever isolated.

4 Real foundation models, both directions

We test the central claim on two *real, open-weight* pretrained models; using open weights makes the whole pipeline reproducible from public checkpoints on commodity hardware.

A model that helps (text). 20-Newsgroups (sci.med vs rec.autos); base X a 50-dim TF-IDF SVD; candidate a real MiniLM sentence embedding (`all-MiniLM-L6-v2`; Reimers and Gurevych, 2019). With no fine-tuning, the Gap predicts the embedding's realized OOS gain to within 0.01 (**predicted** 0.161**, realized** 0.151; Fig. 2⋆), ≈ 0 for a redundant control.

A model that does not (biology). On the real Cancer Dependency Map (DepMap 25Q2; Tsherniak et al., 2017), we embed cell lines with a frozen open-weight **Geneformer** (Theodoris et al., 2023) and ask whether the 1152-dim embedding adds dependency signal *beyond linear PCs of the same expression*. Across 80 pre-registered targets the robust verdict is **redundancy**: mean realized OOS $\Delta R^2 = -0.023$ (Fig. 4); the gap forecasts this a priori by its small magnitude. A pretrained single-cell foundation model, off the shelf, buys nothing over linear PCs of its own input—and the gap says so before any training. (When a modality *does* carry signal—a lineage-addiction gene's own expression, HNF1B—the gap predicts the gain: 0.234 vs realized 0.270.)

A model that helps (de novo protein design). On the designed-protein `stability` task (InstaDeepAI/TAPE; Rao et al., 2019) with a frozen open-weight **ESM-2** (Lin et al., 2023), we ask whether the protein-LM embedding adds signal over cheap sequence-composition features—the question of whether the model is worth running to score a design library. It does, and the gap forecasts the *direction* a priori (bootstrap realized CI excludes 0), while conservatively *under-predicting* the magnitude ($\sim$40%, matching the cross-fit slope of §5): protein stability predicted $+0.158$ vs realized $+0.268$; enzyme (β-lactamase) $+0.067$ vs $+0.123$. **Four** real open-weight foundation models across four domains (text, single-cell, protein, enzyme), with FM utility forecast a priori in both directions (Fig. 5); magnitude is calibrated for low-dim/linear candidates and conservative for high-dim protein embeddings. **Decision utility:** turning the score into a design choice, ranking held-out designs with the Gap-recommended ESM model selects candidates $+0.200$ more stable than a cheap-composition ranker (bootstrap 95% CI $[0.11, 0.32]$, p=0.002; closing 22% of the gap to the oracle; Fig. 6), the a-priori Gap having flagged ESM as useful on the training split alone—the abstract ΔR^2 made a concrete design decision before any training.

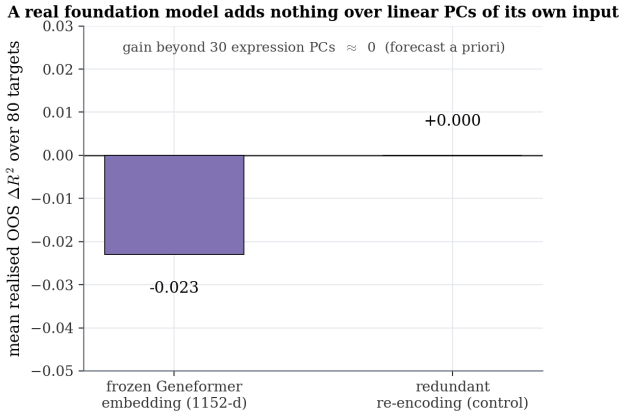


Figure 4: **A real foundation model the Gap correctly waves off.** A frozen open-weight Geneformer embedding adds no dependency signal over linear PCs of its own input across 80 pre-registered DepMap targets (mean $\Delta R^2 = -0.023$); the small a-priori gap forecasts this. Any apparent per-target trend is a target-variance confound, not signal (§7).

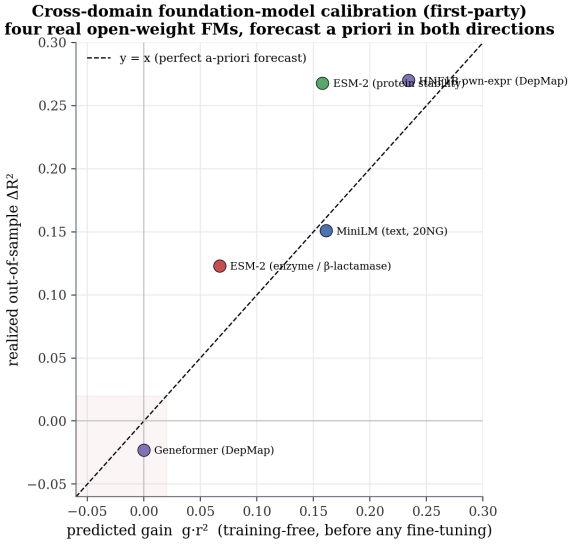


Figure 5: **Calibrated across four real foundation models, in both directions.** Predicted versus realised gain for every first-party open-weight FM: MiniLM (text) $+$, Geneformer (single-cell) ≈ 0, ESM-2 (protein) $+$, ESM-2 (enzyme) $+$. Direction is forecast a priori throughout; magnitude is on $y{=}x$ for low-dimensional candidates and conservative for high-dimensional protein embeddings (§7).

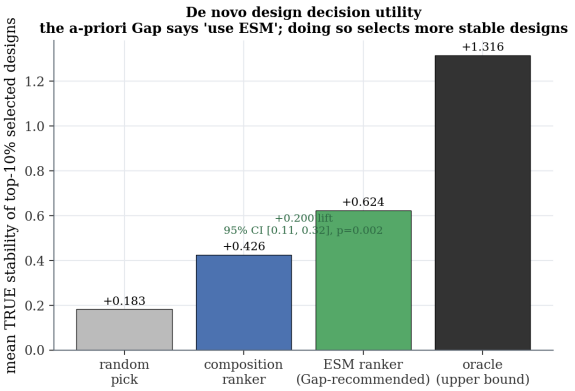


Figure 6: **The abstract gain makes a concrete design decision.** Ranking held-out designs with the Gap-recommended ESM model selects candidates $+0.200$ more stable than a cheap-composition ranker (bootstrap 95% CI $[0.11, 0.32]$, p=0.002), closing 22% of the gap to the oracle—the a-priori Gap having flagged ESM as useful on the training split alone.

5 Applying the Gap inside a transformer: training-free layer selection

Which layer of a pretrained transformer should you transfer from—normally answered by probing every layer? We extract per-layer mean-pooled embeddings and ask whether the training-free Gap picks the best layer a priori, across **four transformers (distilbert, BERT-base, RoBERTa, MiniLM-L12)** $\times$ **five tasks**. Pooled, the Gap ranks layers by realized probe accuracy at weighted-Kendall τ_w=0.77, and its a-priori chosen layer has mean **accuracy regret** 0.004 versus the probe-every-layer oracle (exact-best hit-rate 50%; when it misses it picks a near-equivalent layer). It is exact on distilbert (τ_w 0.93, regret 0.000) and consistent on the deeper models, with chosen layers in the transfer-favoured late-middle band (Fig. 7). LogME performs equivalently (pooled τ_w 0.77). The Gap replaces per-layer probing: L training runs collapse to one training-free score at $\approx 0.4\%$ accuracy cost across four real transformers.

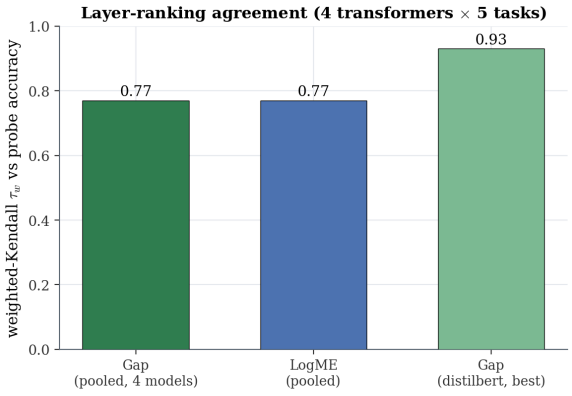


Figure 7: **Training-free layer selection.** Layer-ranking agreement (weighted-Kendall τ_w) between each training-free score and realised probe accuracy across four transformers $\times$ five tasks: the Gap ties LogME pooled (0.77) and is exact on distilbert (0.93). The a-priori chosen layer's mean accuracy regret versus probing every layer is 0.004 (§5).

6 Head-to-head with training-free transferability estimation

The closest prior art predicts representation utility without training the downstream: LogME (You et al., 2021), LEEP (Nguyen et al., 2020), NCE (Tran et al., 2019), H-score (Bao et al., 2019), TransRate (Huang et al., 2022). Our instrument differs in being *incremental* (it scores the gain of a candidate *over what you already have*) and in carrying a *proven tight ceiling* with a reliability term, not a correlate. We compare under the field-standard weighted-Kendall τ_w against a *nonlinear* downstream (gradient boosting), so no scorer predicts its own estimator. On **real open-weight LLM embeddings** (six encoders, five 20NG tasks) the Gap **ties LogME** (τ_w 0.81=0.81) and both decisively beat H-score (0.75) and a trivial correlation baseline (0.19)—in the high-dimensional regime these methods are built for, naive correlation collapses (Fig. 8). A cross-field study (vision, medical, chemistry, finance) shows the principled scores earn their keep specifically in high-dimensional representation spaces, while a fast closed-form Gap variant is *faster than LogME* (App. B).

7 Related work, limitations, and reported negatives

Tools, not claims: FWL/partial regression (Hansen, 2022); attenuation from measurement error (Carroll et al., 2006); Marchenko–Pastur/spiked-covariance recovery (Bai and Silverstein, 2010; Paul, 2007); tabular deep learning (Grinsztajn et al., 2022); the LLM capability g-factor (Burnell et al., 2023). Our contribution is their composition into a predictive, training-free instrument and its first-party validation on foundation-model utility. A concurrent manuscript (author reference withheld for anonymity) develops the target-side reliability–detectability ceiling $R^2_{\mathrm{OOS}} \le \delta\, r^2$ in general; the present paper is the complementary *representation-side* result—the exact gain identity and the training-free instrument that forecasts foundation-model utility—and shares only the corollary $\Delta R^2_{\mathrm{OOS}} \le g_\star r^2$, proved and deployed here. **Limitations, plainly:** (i) the linear g can over-predict for nonlinear-in-X usefulness; Theorem 2 closes this via g_{ker}. (ii) For a *multi-dimensional* block the in-sample gap over-predicts absolute magnitude (a d/n effect); a cross-fit estimator removes the positive bias at the cost of conservative (≈ 0.5 slope) under-prediction (App. A). (iii) Evidence is strongest on structured omics/leaderboard/text data.

Reported negatives (kept visible). We tested a sharp sample-complexity phase transition and *refuted* it: usefulness is realized early, well below the putative BBP threshold (App. A). We reported, then *retracted*, a per-target ranking claim (ρ=0.67): a provably zero-signal block reproduces $|\rho| \approx 0.63$ on the same targets under both estimators, exposing the statistic as a target-variance confound; only the mean-level redundancy verdict survives, and it is a direct measurement. This retraction concerns *per-target* ordering and does not touch the *representation*-ranking claim (§3–6).

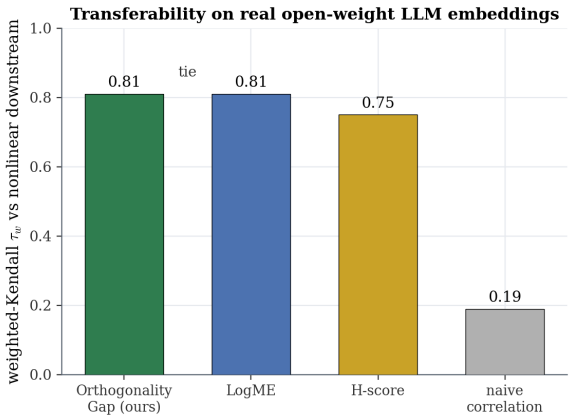


Figure 8: **Parity with the strongest training-free score, in its own regime.** On real open-weight LLM embeddings the Gap ties LogME (τ_w $0.81{=}0.81$), both far above H-score and a naive-correlation baseline that collapses in high dimension—the regime these methods exist for.

8 Conclusion

For structured targets, whether a candidate representation can improve prediction is fixed a priori by geometry: only its orthogonal-to-X part can help, up to the target's reliability. The Orthogonality Gap reads this off training-free, with a proven ceiling we validate ($100\%/97\%$ coverage), a kernel form for the nonlinear case, near-oracle transformer layer selection, and parity with the strongest training-free score—while being candid about what it is not: a tie, not a new SOTA ranking; a mean-level verdict on high-dim blocks, not a per-target one. Its value is a principled, cheap stop rule for the scaling reflex and a positive signal for where an orthogonal modality pays off—before a single training run.

Reproducibility Statement

Every quantitative claim is produced by a self-contained script carrying a PASS/FAIL harness and a permutation or ground-truth control, released as supplementary code. The tightness certificate, calibration, kernel form, the four first-party foundation-model evaluations (MiniLM, Geneformer/DepMap, ESM-2 on stability and enzyme engineering), the scaling and de-novo studies, the reported ranking retraction, the transferability benchmarks, and the transformer layer-selection results each have a named script. Data are public: the DepMap release, 20-Newsgroups, ProteinGym, and datasets bundled with scikit-learn. All numbers are CPU-reproducible except the optional frozen Geneformer embedding, which uses one GPU. The exact identity, ceiling conditions, and estimator definitions are stated in §2 and the appendix.

Use of Large Language Models

A large language model was used as a coding and drafting assistant. Every experimental result is produced by a self-contained script that recomputes the reported number from public data; no figure or statistic was generated by a language model. All conceptual claims, methodological choices, and final judgments are the authors', who take full responsibility. No language model is an author.

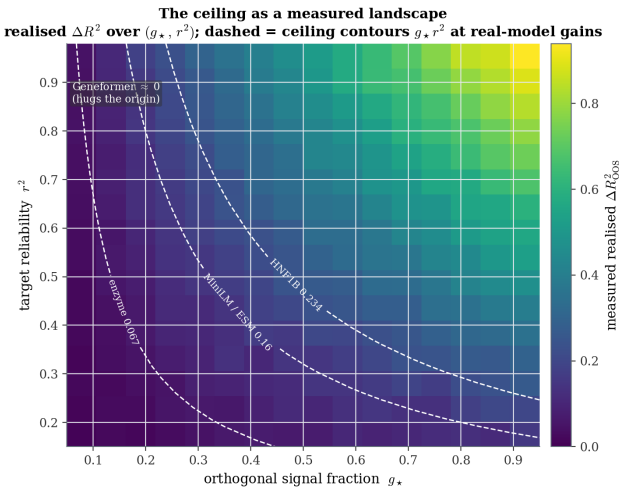


Figure 9: **The ceiling as a measured landscape.** Realised out-of-sample ΔR^2 measured across the whole $(g_\star,\, r^2)$ plane (large-n OLS, aligned feature ρ=1); dashed curves are the exact ceiling contours $g_\star r^2$ at the four first-party real foundation-model predicted gains. The measured surface tracks the contours, and every real model sits on a named iso-ceiling curve — Geneformer hugging the origin (redundant), enzyme / protein / text at successively higher ceilings. The ceiling is not a single number but a landscape, and capacity moves a model within it, never across a contour.

A Full validation suite

Domain generality (three real datasets, predicted vs realized: diabetes $0.30/0.29$, breast $0.94/0.94$, California $0.47/0.47$; pooled Pearson r=1.00). Refuted phase transition: on a planted orthogonal modality (q=40, θ=1.3, $n^* \approx 444$) the gain is positive across $n \in [60, 2000]$ and reaches 64% of its max by n=60; the diagnostic tracks it (Spearman 0.94). Scaling correction: on a true-zero block the naive gap inflates $0 \to 0.121$ as d=2 $\to$ 40; a regularized cross-fit stays ≈ 0 (rank-calibrated r=0.98, magnitude-conservative slope 0.48). De novo generalization: the a-priori gap on *seen* cell-contexts forecasts $+0.32$ predictive-correlation gain on *held-out* contexts (redundant control 0.00).

B Benchmark details

Weighted-Kendall τ_w vs a gradient-boosted downstream; synthetic control (Gap ties LogME, p=0.46); classical-feature check reported as non-discriminating (a trivial baseline wins, the signature of a benchmark that cannot place the methods in their regime); real LLM embeddings as in §6. Compute: a fast closed-form Gap is faster than LogME (e.g. 2.8 vs 61 ms at d=64); the cross-fit variant is slower (we drop any "cheaper than LogME" claim for it). Our LogME is **validated against the authors' official implementation** (Spearman ρ=0.999/0.998 regression/classification). The training-free prediction also tracks a **backprop-trained MLP** downstream (Pearson r=0.97, ceiling intact), not only linear/tree probes; end-to-end FM fine-tuning is the remaining extension.

REFERENCES

Zhidong Bai and Jack W. Silverstein. *Spectral Analysis of Large Dimensional Random Matrices*. Springer, 2nd edition, 2010. Marchenko–Pastur law.

Yajie Bao, Yang Li, Shao-Lun Huang, Lin Zhang, Lizhong Zheng, Amir Zamir, and Leonidas Guibas. An information-theoretic approach to transferability in task transfer learning. In *IEEE International Conference on Image Processing (ICIP)*, 2019.

Ryan Burnell, Han Hao, Andrew R. A. Conway, and José Hernández-Orallo. Revealing the structure of language model capabilities. *arXiv preprint arXiv:2306.10062*, 2023.

Raymond J. Carroll, David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman and Hall/CRC, 2nd edition, 2006. attenuation from measurement error / reliability.

Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Bruce E. Hansen. *Econometrics*. Princeton University Press, 2022. Frisch–Waugh–Lovell / partial regression, Ch. 3.

Long-Kai Huang, Junzhou Huang, Yu Rong, Qiang Yang, and Ying Wei. Frustratingly easy transferability estimation. In *International Conference on Machine Learning (ICML)*, 2022.

Zeming Lin, Halil Akin, Roshan Rao, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Cuong V. Nguyen, Tal Hassner, Matthias Seeger, and Cédric Archambeau. Leep: A new measure to evaluate transferability of learned representations. In *International Conference on Machine Learning (ICML)*, 2020.

Debashis Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 17(4):1617–1642, 2007. spiked-covariance (BBP) eigenvector recovery.

Roshan Rao, Nicholas Bhattacharya, Neil Thomas, et al. Evaluating protein transfer learning with TAPE. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP-IJCNLP*, 2019.

Christina V. Theodoris, Ling Xiao, Anant Chopra, et al. Transfer learning enables predictions in network biology. *Nature*, 618:616–624, 2023.

Anh T. Tran, Cuong V. Nguyen, and Tal Hassner. Transferability and hardness of supervised classification tasks. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Aviad Tsherniak, Francisca Vazquez, Phil G. Montgomery, et al. Defining a cancer dependency map. *Cell*, 170(3):564–576, 2017.

Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. Logme: Practical assessment of pre-trained models for transfer learning. In *International Conference on Machine Learning (ICML)*, 2021.