LOCKED EVALUATION SURFACES TRANSFER FAILURE AND SAMPLING-DEPTH ENTANGLEMENT IN CRISPRI PERTURBATION-EFFECT PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting how held-out target genes respond to CRISPRi perturbation, and whether such predictions transfer across biological screens, is hard to evaluate: a representation can be informative within one screen yet fail across screens, while endpoint definitions and design factors such as sampling depth differ between datasets. We evaluate a frozen Geneformer representation under a locked, pre-registered protocol—heads and model selection frozen before test evaluation, external outcome labels withheld until final unblinding, and analysis-governing decisions fixed before the evaluations they govern. In-distribution on the Virtual Cell Challenge (VCC), the frozen representation carries measurable predictive information beyond a dimension-matched random-feature control ($\Delta R^2 = +0.1645$, 95% CI $[+0.1375, +0.1920]$), satisfying the pre-registered informativeness gate required before interpreting transfer. It then fails zero-shot transfer on both external screens (Spearman $\rho = -0.139$ and -0.267), lying below that control on each. Adding a predefined magnitude block improves the representation externally ($\Delta\rho = +0.032$ and $+0.143$) but, under the frozen primary head, does not rescue transfer: both remain negative. A pre-registered, count-adjusted max-response secondary is positively associated with the outcome on both screens; we report it as correlational and secondary, not as a recovered magnitude signal. Finally, the VCC endpoint is strongly sample-size associated—a count-only linear model reaches $R^2 = +0.4325$, versus $+0.2589$ for the four magnitude scalars; adding those scalars to cell count improves R^2 by only $+0.0017$—so much of the aggregate-magnitude signal overlaps with cell count. A locked evaluation thus surfaces a transfer failure and a sampling-depth entanglement that a less controlled evaluation could obscure.

1 INTRODUCTION

Predicting how a cell's transcriptome responds to a genetic perturbation is central to computational single-cell biology, with applications to target discovery and synthetic biology (Dixit et al., 2016; Replogle et al., 2022; Lotfollahi et al., 2019; Roohani et al., 2024). The hardest and most practically relevant variant, emphasized by benchmarks such as the Virtual Cell Challenge (VCC) (Roohani et al., 2025), is held-out target-gene generalization: train on one set of perturbations and evaluate on disjoint target genes, ideally in a new biological screen. The input is an observed post-perturbation expression profile and the target a scalar summary of its effect strength, so the task is to generalize that mapping to unseen target genes, not to forecast a response from perturbation identity alone. Strong performance within a single screen is not the same as a signal that transfers—a representation can be informative in-distribution yet fail across screens.

Foundation models and the evaluation problem. Single-cell foundation models such as scGPT (Cui et al., 2024) and Geneformer (Theodoris et al., 2023) are attractive here because their pretrained representations may encode transferable biological structure. But cross-screen evaluation is easy to get wrong. A representation may be informative only within the screen it is tuned on; the prediction endpoint distributed with different screens may measure different quantities; and simple summaries of the response—its overall magnitude—can correlate with nuisance or design factors such as the

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.

number of cells assayed per perturbation. Any of these can make an in-distribution success look like transferable signal.

Core question. We therefore ask not whether any single feature dominates, but: *what survives a locked cross-screen evaluation once in-distribution informativeness, zero-shot transfer, and endpoint/sample-size effects are separated?* We study a frozen Geneformer representation $\mathbf{z}$ on VCC and two external Replogle CRISPRi screens (Replogle et al., 2022), under the locked protocol of Sec. 4. We compare $\mathbf{z}$ against a dimension-matched random-feature control C_1 evaluated under the same selection protocol, and against shuffled controls, add a predefined four-scalar magnitude block $\widetilde{\boldsymbol{m}}$, and score external transfer against a harmonized target-gene endpoint rebuilt from single-cell data so that source and target measure the same quantity.

Findings. Headline findings: the frozen representation carries measurable in-distribution information, resolved even against a compressed-expression comparator, yet does not transport to either external screen; and the strong in-domain magnitude signal is entangled with sampling depth, so that reading requires refinement. Also reported: magnitude augmentation improves external correlation without rescuing transfer, and a count-adjusted max-response correlate is positive but correlational. We report the external pattern descriptively, without inferring a mechanism.

Contributions. The contribution is not a new architecture but an evaluation that changes the scientific conclusion a less controlled evaluation would reach.

1. **A locked, pre-registered cross-screen evaluation protocol** for perturbation-effect prediction, with selection, controls and amendments fixed in advance of the results they govern (Sec. 4).
2. **A foundation-model transfer result.** A frozen Geneformer representation carries measurable VCC in-distribution predictive information beyond a dimension-matched control ($\Delta R^2(\mathbf{z} - C_1) = +0.1645$, 95% CI $[+0.1375, +0.1920]$) yet fails zero-shot transfer on both external screens ($\rho = -0.139$ and -0.267), lying below that control—in-domain information is not evidence of portability.
3. **An endpoint / sampling-depth diagnosis.** The VCC endpoint is strongly sample-size associated, and a count-only linear model out-predicts the four magnitude scalars in validation ($R^2 = +0.4325$ vs. $+0.2589$, magnitude adding $+0.0017$ beyond count), so the aggregate-magnitude interpretation is substantially entangled with cell count.
4. **Magnitude augmentation, characterized precisely.** Adding the predefined magnitude block improves the frozen representation externally ($\Delta\rho = +0.032$ and $+0.143$) but, under the frozen primary head, does not rescue transfer (both remain negative).

2 Related work

Perturbation prediction. SPSM-style stable-prediction models descend from stable-prediction methods in causal inference (Kuang et al., 2018; Shen et al., 2018) adapted to single-cell data. Recent perturbation prediction methods include CPA (Lotfollahi et al., 2023), GEARS (Roohani et al., 2024), and pretrained foundation models such as scGPT (Cui et al., 2024), Universal Cell Embeddings (Rosen et al., 2026), and Geneformer (Theodoris et al., 2023). Many existing evaluations either pool across target-gene splits or use a held-out-pair setting; the strict disjoint target-gene split used here is more demanding, and our analysis is complementary: rather than proposing a new architecture we ask which features any model class needs to exploit on this split.

Linear baselines and mode collapse in perturbation prediction. A recent line of work documents that deep and foundation-model perturbation predictors frequently fail to outperform simple baselines. Ahlmann-Eltze et al. (2025) compared five foundation models and two deep architectures against linear baselines on Adamson, Norman, and Replogle, and found that none consistently outperformed the linear model on unseen-perturbation prediction; for held-out perturbations the deep models often failed to outperform predicting the unconditional training mean. They additionally show that a linear model with perturbation embeddings pretrained on Replogle K562 transfers across cell lines to RPE1 (and vice versa), establishing that cross-cell-line transfer with linear models is feasible. Wu et al. (2025) introduced PerturBench with covariate-transfer benchmarks across cell lines on Srivatsan20

and Jiang24, and identified mode collapse—predictions concentrating in a narrow region of output space—as a recurring failure mode, with simple baselines (Latent Additive, Decoder-Only) often matching or exceeding more complex autoencoder models. Csendes et al. (2025) found that a mean baseline exceeded scGPT and scFoundation (Hao et al., 2024) on post-perturbation Pearson(Δ), and Bendidi et al. (2024) reported that PCA-based baselines remained competitive with transcriptomics foundation models. Mejia et al. (2025) addressed mode collapse on the loss side with a DEG-weighted training objective. Our work sits in this lineage but differs in what it controls: we isolate response magnitude with pre-specified shuffle and Gaussian controls (Sec. 5.3), evaluate cross-screen transfer of a frozen foundation-model representation under a locked protocol (Sec. 6.2), and show that the endpoint distributed with those screens measures a different quantity from the training target, rebuilding a common one (Sec. 6.2). The evaluation uses the held-out target-gene split of the Virtual Cell Challenge (Roohani et al., 2025), distinct from the covariate-transfer splits of Wu et al. (2025) and the unseen-perturbation splits of Ahlmann-Eltze et al. (2025).

Imbalanced regression and tail prediction. Deep imbalanced regression frameworks include Label Distribution Smoothing (LDS) and Feature Distribution Smoothing (Yang et al., 2021), Balanced MSE (Ren et al., 2022), density-based reweighting (Steininger et al., 2021), and label-distribution-aware margin losses (Cao et al., 2019). We instantiate LDS-weighted and tail-weighted variants as baselines (Sec. 5); both fail in our setting, and App. U sets out how we read that failure.

Distribution shift and rank objectives. Cell-type transfer is a distribution shift between domains with distinct gene-expression baselines. Standard OOD techniques such as IRM (Arjovsky et al., 2019), GroupDRO (Sagawa et al., 2020), REx (Krueger et al., 2021), and the WILDS benchmark family (Koh et al., 2021) address related but distinct shifts and are complementary. Our use of magnitude as a deliberately low-dimensional, coordinate-permutation-invariant summary is conceptually allied with norm- and energy-based detection (Hendrycks & Gimpel, 2017; Liu et al., 2020) and norm-aware embeddings (Wang et al., 2017). The MAG-RANK extension (App. N) uses a margin-based pairwise rank loss in the spirit of LambdaRank-family approaches (Burges et al., 2005; Cao et al., 2007; Hadsell et al., 2006).

3 PROBLEM SETUP AND NOTATION

Data. The Virtual Cell Challenge (VCC) corpus (Roohani et al., 2025) consists of CRISPRi (dCas9–KRAB) (Gilbert et al., 2014) Perturb-seq (Dixit et al., 2016) screens in H1 human embryonic stem cells (H1 hESC), organized into 48 experimental batches indexed by $b \in \mathcal{B}$. Each row i corresponds to a (batch, target-gene) pair (b_i, g_i) with feature vector $\mathbf{x}_i \in \mathbb{R}^d$ and scalar label $y_i \in \mathbb{R}_{\geq 0}$. Features are log1p-normalized mean expression deltas relative to the NT-control mean of batch b_i, restricted to the $d = 2{,}000$ highly variable genes (HVGs) selected on the train split. The label is $y_i = \log(1 + \max\{0,\, A^2_{b_i,g_i}\})$, where $A^2_{b,g}$ is the two-sample Anderson–Darling statistic (Scholz & Stephens, 1987) between the log1p expression of gene g in cells perturbed at g within batch b and in non-targeting controls from the same batch; the AD statistic is computed on the target-gene column only, and $\mathbf{x}_i$ enters as predictor input only (App. W). The set of target genes is partitioned disjointly into $\mathcal{G}_{\text{tr}}$, $\mathcal{G}_{\text{val}}$, and $\mathcal{G}_{\text{te}}$ of sizes 140, 49, and 93, giving 5,740 training, 2,052 validation, and 3,972 test rows. These are counts after eligibility, which requires at least ten cells in some (target gene, batch) group and removes 10, 1, and 7 genes from the 150, 50, and 100 raw targets; every excluded gene is accounted for by that rule, with none left over.

Task and metrics. Given $(b_i, \mathbf{x}_i)$ for a test row, predict y_i. We report pooled test $R^2 = 1 - \sum_i (y_i - \hat{y}_i)^2 \big/ \sum_i (y_i - \bar{y}_{\text{te}})^2$, seed-level mean $\pm$ standard deviation across independent training runs (default: seeds $\{0, 1, 2\}$; MAG-Anchor uses five seeds), Spearman rank correlation, and per-quintile bias $\mathbb{E}[\hat{y} - y \mid q(y)]$. Standard deviations are sample standard deviations throughout.

Cross-dataset shift. For external transfer we use two Replogle CRISPRi (dCas9–KRAB) Perturb-seq screens (Replogle et al., 2022; Replogle & Weissman, 2022): RPE1 and K562-essential. Features are mapped into the VCC HVG namespace by gene symbol; coverage is screen-specific—K562-essential recovers 1,430 of 2,000 HVGs (71.5%) and RPE1 recovers 1,550 of 2,000 (77.5%); unmapped coordinates are imputed as zero, the value an unperturbed cell takes in this pipeline. For

each screen we rebuild the label from single-cell data using the definition above, so that source and target are scored on the same quantity—the Anderson–Darling column distributed with the Replogle release measures something else (Sec. 6.2). The rebuilt endpoint fixes subsample sizes at the VCC medians so the statistic is not driven by sample-size differences, and eligibility requires those cell counts and an unambiguous target mapping, a threshold whose residual selection pathway we return to in Sec. 7. We never train on either screen, and no external outcome label enters any fitting or preprocessing step: every preprocessing statistic is fit on the VCC training split and applied frozen, and no target-side scaling, calibration, or per-screen refit is performed (App. G). The pre-specified non-targeting reference, pooled within each external screen, enters only in feature and endpoint construction (App. A). A legacy training control on the shared-HVG subspace is reported in App. X.1.

4 LOCKED EVALUATION PROTOCOL

Whether a representation *transfers* is easy to overstate: an in-distribution success, a control too weak to be informative, or an outcome inspected after the fact can each inflate an apparent transfer signal. We therefore fix each analysis-governing choice before the result it governs and interpret transfer only through that recorded protocol (Fig. 3).

Pre-registration and freezing. Hypotheses, controls, metrics, the resampling procedures (App. F) and decision rules are recorded in advance. Here, "pre-registered" denotes an internal pre-analysis specification and lock rather than registration in an external registry; App. G.3 gives the chronology evidence and its limitations, and App. G the full governance record. Model and head selection use VCC training and validation only and are frozen before any test metric is computed; external outcome labels are withheld until final unblinding. Selection never uses a test or external metric.

An informativeness gate before transfer. Interpretation is staged. The in-distribution gate (H-FM1) asks whether $\mathbf{z}$ carries usable signal on VCC *beyond a dimension-matched random control* C_1, matched in input dimension and in protocol (App. B): it is informative iff the paired bootstrap interval on $\Delta R^2 = R^2(\text{FM-only}) - R^2(C_1)$ excludes zero on the positive side, with C_1 alone as the gate. Zero-shot transfer (H-FM2) is interpreted *only if* that gate passes; if it fails, a near-zero external correlation cannot be read as a transfer result (App. D). Passing the gate does not predict that transfer succeeds.

Controls, hypothesis family, and what stays open. Fixed controls—the dimension-matched Gaussian C_1 and shuffled-alignment controls—are generated from array shapes only, never from labels. Two taxonomies run in parallel and should not be conflated. The pre-registered *primary inferential family* is the four hypotheses H-FM1–H-FM4, and every one of them is reported in full. Separately, two *headline findings* organize the presentation: in-distribution informativeness together with the zero-shot transfer outcome, and the endpoint/sampling-depth diagnosis. H-FM3, the magnitude increment, therefore belongs to the primary inferential family and is reported in full, but is not promoted to a headline finding; the count-adjusted max-response analysis is a pre-registered secondary. Hypotheses left unresolved under the lock—including test-level magnitude recoverability (H-FM4)—are reported as unresolved. All four hypotheses, their gating controls and their outcomes are in App. D.

5 WHY THE PROBLEM IS HARD

Baseline failure mode. The SPSM V4 baseline, a stable-prediction MLP encoder with a within-context stability loss and a relation loss (App. L), achieves test pooled $R^2 = +0.0824 \pm 0.0100$ across three seeds. Its predictions exhibit mean collapse: the prediction range is a small fraction of the label range, and per-quintile bias is monotonic, over-predicting the lowest quintile and under-predicting the highest (Fig. 1c). We say "mean collapse" rather than "mode collapse" (Wu et al., 2025): predictions concentrate at the marginal training mean rather than at one of several modes of a multimodal output.

Five remedies that failed or stalled. Earlier exploratory work tried five plausible corrections under a matched protocol—an inductive-bias bilevel gate, cross-cell-type pretraining followed by VCC fine-tuning, a post-hoc tail correction, a tail-weighted MSE (Ren et al., 2022), and LDS-weighted

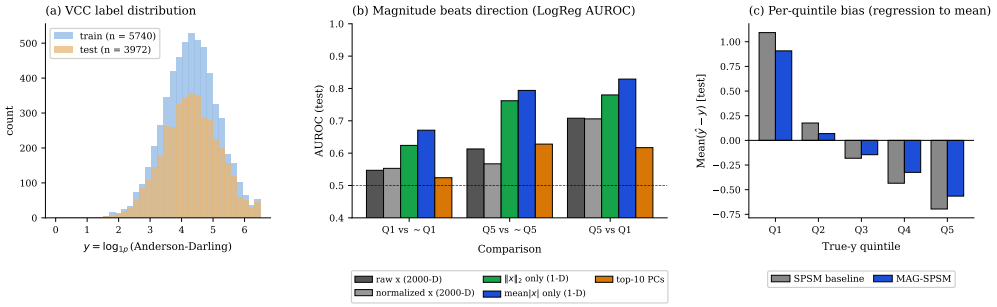


Figure 1: Diagnostic on VCC. **(a)** Label distribution on the training and test splits. **(b)** Separability of the highest from the lowest label quintile by logistic regression: one-dimensional magnitude summaries match or exceed the full 2,000-dimensional expression vector and its leading principal components. **(c)** Per-quintile bias: SPSM collapses predictions toward the marginal training mean; MAG-SPSM reduces the bias most at the tails.

MSE (Yang et al., 2021)—and all failed or stalled. Those runs predate the locked evaluation and their artifacts were not retained, so we use them as historical motivation and draw no quantitative or mechanistic conclusion from them. What they motivated is the step taken next: supplying response *magnitude* to the predictor as an explicit feature rather than through the loss.

The magnitude axis. A pre-specified diagnostic suite showed that one-dimensional magnitude summaries separate low- from high-response rows about as well as the full expression vector, while the baseline encoder collapses its predictions to a narrow window centered on $\bar{y}_{\text{tr}}$ (Fig. 1). This motivates exposing magnitude to the predictor explicitly.

5.1 Magnitude features and a proof-of-concept encoder

Magnitude features. For $\mathbf{x} \in \mathbb{R}^d$ with components $x_1, \ldots, x_d$, define the magnitude map $\boldsymbol{m} : \mathbb{R}^d \to \mathbb{R}^4$ by

$$\boldsymbol{m}(\mathbf{x}) = \Bigg(\underbrace{\tfrac{1}{d}\sum_j |x_j|}_{m_1:\ \text{mean}|x|},\ \underbrace{\Big(\sum_j x_j^2\Big)^{1/2}}_{m_2:\ \ell_2},\ \underbrace{\max_j |x_j|}_{m_3:\ \ell_\infty},\ \underbrace{\Big(\tfrac{1}{d}\sum_j (x_j - \bar{x})^2\Big)^{1/2}}_{m_4:\ \sigma(\mathbf{x})} \Bigg). \tag{1}$$

We z-score $\boldsymbol{m}$ component-wise using *training-set* statistics to obtain $\widetilde{\boldsymbol{m}}(\mathbf{x})$. Standardization removes any constant factor, so specifying m_2 as $\|\mathbf{x}\|_2$ or as $\|\mathbf{x}\|_2/\sqrt{d}$ gives an identical model input. All four statistics are coordinate-permutation invariant, m_1–m_3 are sign-flip invariant, and m_4 is invariant under translations along $\mathbf{1}$; these invariances emphasize *how much* a cell responded rather than *which genes* did, which motivates testing magnitude as a potentially more shift-robust summary across cell types.

Three baselines built from $\widetilde{\boldsymbol{m}}$. We use $\widetilde{\boldsymbol{m}}$ in three ways: (i) MAG-LINEAR, a four-feature linear regression $\hat{y} = \mathbf{w}^\top \widetilde{\boldsymbol{m}}(\mathbf{x}) + b$; (ii) classical methods on $\mathbf{x}$, $\mathbf{x}/\|\mathbf{x}\|_2$, $\widetilde{\boldsymbol{m}}$, or $[\mathbf{x}; \widetilde{\boldsymbol{m}}]$ (App. H); and (iii) MAG-SPSM (below) as a proof of concept that the same signal can be exposed to a deep encoder.

MAG-SPSM (proof of concept). MAG-SPSM concatenates $\mathbf{x}$ with $\widetilde{\boldsymbol{m}}(\mathbf{x})$ to form $\mathbf{x}^+ = [\mathbf{x}; \widetilde{\boldsymbol{m}}(\mathbf{x})] \in \mathbb{R}^{2004}$ and feeds it to an MLP encoder $f_\theta : \mathbb{R}^{2004} \to \mathbb{R}^{256}$ (three hidden layers of width 256, ReLU, LayerNorm, no architectural dropout) followed by a linear head. The objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{MSE}} + \lambda_{\text{stab}}\,\mathcal{L}_{\text{stab}} + \lambda_{\text{rel}}\,\mathcal{L}_{\text{rel}}, \tag{2}$$

where $\mathcal{L}_{\text{stab}}$ and $\mathcal{L}_{\text{rel}}$ are the SPSM stability and relation losses (App. L; $\lambda_{\text{stab}} = \lambda_{\text{rel}} = 0.5$, inherited unchanged). Training schedule and optimizer are in App. L; model selection is by validation R^2. MAG-SPSM is *not* the strongest model here—classical tree ensembles on the same input outperform it (App. H)—but is retained as the simplest deep-architecture check that magnitude is recoverable when supplied explicitly, and as the substrate for the per-row alignment controls (Sec. 5.3).

5.2 MAGNITUDE-ANCHORED RESIDUAL PREDICTION

MAG-Anchor is a secondary deep architecture that splits the prediction into a magnitude anchor and a gated residual, $\hat{y} = g_\phi(\widetilde{\boldsymbol{m}}(\mathbf{x})) + \alpha\, h_\theta([\mathbf{x}; \widetilde{\boldsymbol{m}}(\mathbf{x})])$, with g_ϕ initialized from MAG-LINEAR and α learned. It was not reproduced under the locked revision harness and no value from it is carried into this revision; its architecture and per-row alignment controls are in App. J.

5.3 THE ROLE OF PER-ROW ALIGNMENT

The controls above rest on a simple population-level argument: if an augmenting block $M^\star$ is independent of (X, Y), then $\mathbb{E}[Y \mid X, M^\star] = \mathbb{E}[Y \mid X]$, so no measurable predictor using $(X, M^\star)$ can improve over the Bayes predictor on X alone under squared loss. This yields two consequences we test directly: (a) replacing $\widetilde{\boldsymbol{m}}(\mathbf{x}_i)$ by four i.i.d. $\mathcal{N}(0, I_4)$ features makes the augmenting block independent of $(\mathbf{x}_i, Y_i)$ by construction; (b) a row-shuffled augmentation preserves the marginal distribution of $\widetilde{\boldsymbol{m}}$ but breaks row-specific information about Y_i. A systematic R^2 gain over both controls therefore supports the reading that the four scalars contribute through per-row dependence on $\mathbf{x}_i$ rather than through added dimensionality alone (App. I).

6 EXPERIMENTS

The frozen representation. The representation $\mathbf{z}$ is Geneformer-V2-104M (Theodoris et al., 2023) used strictly as a frozen feature extractor: the pretrained weights are loaded in inference mode and never updated, so no gradient reaches the encoder anywhere in this study. Each cell is tokenized by the model's own rank-value encoding and embedded to 768 dimensions; a perturbation-level vector is the mean embedding over the cells of a construct minus a non-targeting reference mean, $\mathbf{z}_g = |g|^{-1} \sum_{i \in g} E_i - \mu_{\mathrm{NT}}$. The non-targeting reference is batch-local for VCC and globally pooled per external screen; that asymmetry, the pinned checkpoint digests, and the other recorded cross-screen differences are set out in App. A, together with the standardization, which is fit on the VCC training split alone.

Controls, heads, and other models. C_1 is the capacity floor that licenses every comparison below: a 768-dimensional matrix of independent standard Gaussian entries drawn per row from array shapes alone, matched to $\mathbf{z}$ in input dimension and passed through the identical preprocessing, candidate-head pool, selection rule and training budget. The remaining controls, the candidate heads, the selection rule and the freeze are specified in App. B and App. C, and all five controls were fixed in advance. The deep SPSM and MAG-SPSM models share the SPSM V4 backbone (App. L) and MAG-Anchor a two-branch architecture (App. J); their hyperparameters and compute are in App. L, and those of the classical baselines in App. K. We report mean $\pm$ std over seeds $\{0, 1, 2\}$ (MAG-Anchor uses five seeds).

6.1 IN-DISTRIBUTION INFORMATIVENESS

Before interpreting any transfer result we verify that the frozen representation carries usable in-distribution signal—the pre-registered informativeness gate (H-FM1). On the VCC held-out target-gene test set, the FM-only arm $\mathbf{z}$ improves pooled R^2 over its capacity-matched Gaussian control C_1 by a resolved paired margin, $\Delta R^2(\mathbf{z} - C_1) = +0.1645$ (95% CI $[+0.1375, +0.1920]$, sign-stable across all resamples; row-level paired bootstrap, App. F). The interval excludes zero on the positive side, so the gate passes with C_1 alone as the criterion: the frozen representation is informative in-distribution beyond a dimension-matched control. The frozen VCC test split was scored once (App. G.1). The pre-specified compressed-expression benchmark C_2 is reported alongside and non-gating, as the specification requires: $\Delta R^2(\mathbf{z} - C_2) = +0.1107$ (95% CI $[+0.0906, +0.1317]$). Absolute pooled R^2 on the frozen test split is $+0.1496$ for $\mathbf{z}$, $+0.0387$ for C_2 and -0.0150 for C_1.

Head selection ran independently per arm under one pre-registered rule, which resolved to a random forest for $\mathbf{z}$ and to histogram gradient boosting for C_1; the selected C_1 head attains $R^2 = -0.015$ on the frozen test split. We report this head-family asymmetry explicitly and read C_1 as the pre-registered random-feature floor rather than as a same-head comparator (App. C).

6.2 Zero-shot transfer across screens

A common endpoint. Transfer is scored only where source and target measure the same quantity. The Anderson–Darling column distributed with Replogle measures the breadth of a transcriptome-wide response, not target-gene effect strength, and we could not reproduce it from the single-cell data under any construction we tested (App. O). We therefore score transfer against the harmonized target-gene endpoint rebuilt for each external screen (Sec. 3), with subsample sizes fixed at the VCC medians so the statistic is not driven by sample-size differences. Eligibility retains 1,932 RPE1 and 1,903 K562-essential constructs for transfer evaluation, with exclusions dominated by target genes absent from the feature namespace rather than by insufficient cells (App. P). A small fraction of evaluated constructs target genes that also occur among the 150 raw VCC training targets—43 of 1,932 on RPE1 (2.2%) and 33 of 1,903 on K562-essential (1.7%). The primary retains all eligible constructs, as pre-registered; a pre-specified sensitivity that removes them leaves every verdict unchanged (App. D). Models trained only on VCC are applied zero-shot with no target-side calibration; uncertainty is a paired target-gene cluster bootstrap (App. F).

Transfer. Throughout, a comparison is *resolved* when its 95% interval excludes zero. Applied zero-shot to both external screens, the frozen representation does not transfer. Its rank correlation with the harmonized target-gene endpoint is resolvedly negative on both screens—Spearman $\rho = -0.139$ on RPE1 (95% CI $[-0.185, -0.092]$) and $\rho = -0.267$ on K562-essential ($[-0.310, -0.224]$), both intervals lying entirely below zero. The capacity-matched control C_1 is by contrast compatible with null on both screens ($\rho = -0.034$ $[-0.078, +0.012]$ and $+0.032$ $[-0.012, +0.077]$, each interval spanning zero), as is a shuffled-alignment control that destroys the representation's per-row identity while preserving its marginal ($\rho = +0.044$ $[-0.001, +0.088]$ and -0.004 $[-0.049, +0.041]$). The pre-specified compressed-expression control C_2 is also resolvedly negative on both screens: on RPE1, $\rho = -0.156$ (95% CI $[-0.202, -0.109]$) versus -0.139 for $\mathbf{z}$, and on K562-essential, $\rho = -0.264$ ($[-0.306, -0.221]$) versus -0.267. Because C_2 contains no foundation-model computation, the external failure is not specific to the frozen Geneformer representation; this comparison does not identify its mechanism. The representation is therefore not merely near zero: paired against its own noise floor it is resolvedly worse ($\rho_{\mathbf{z}} - \rho_{C_1} = -0.106$ $[-0.168, -0.044]$ on RPE1 and -0.299 $[-0.360, -0.238]$ on K562-essential). Because the in-distribution gate passed first, this is interpretable as a transfer failure—in-domain information that does not carry across screens—rather than an uninformative representation. In-distribution informativeness (Sec. 6.1) followed by this governed zero-shot outcome together constitute the paper's first headline finding. The complete locked per-arm external results and persisted paired contrasts are in App. D.1, Table 1.

6.3 Magnitude augmentation

Adding the predefined four-scalar magnitude block $\widetilde{\boldsymbol{m}}$ to the frozen representation improves it externally by a resolved paired margin. The paired increment $\Delta\rho([\mathbf{z}; \widetilde{\boldsymbol{m}}] - \mathbf{z})$ is $+0.032$ on RPE1 (95% CI $[+0.008, +0.056]$) and $+0.143$ on K562-essential ($[+0.111, +0.175]$)—positive with intervals excluding zero on both screens. The increment also remains resolved against both pre-registered controls, C_4 (row-shuffled magnitude) and C_5 (four Gaussian increment features), supporting dependence on aligned per-row magnitude rather than added dimensionality alone. The improvement does not, however, rescue transfer: the augmented representation remains resolvedly negative on both screens ($\rho = -0.107$ and -0.124) and still below the capacity-matched control. Adding magnitude makes the representation less anti-correlated with the outcome; it does not make it transfer. That holds under the frozen primary head and construction basis: the pre-specified secondary analyses show the augmented-arm result is head- and feature-basis-sensitive. Neither alters the $\mathbf{z}$-only transfer result, and neither was used to select the reported primary (App. D.2).

6.4 The endpoint is entangled with sampling depth

The second headline finding concerns the benchmark endpoint itself. The VCC Anderson–Darling label is strongly associated with the number of cells assayed per perturbation (Spearman $\rho(n_{\text{cells}}, y) = +0.705$, Fig. 2A): unlike the harmonized external endpoint, the VCC endpoint is computed over all cells in a group and is not size-controlled by construction. This association propagates into the magnitude features. In VCC validation, a linear model on cell count alone reaches $R^2 = +0.4325$, *above* the four magnitude scalars' $R^2 = +0.2589$, and adding the four scalars to cell count raises

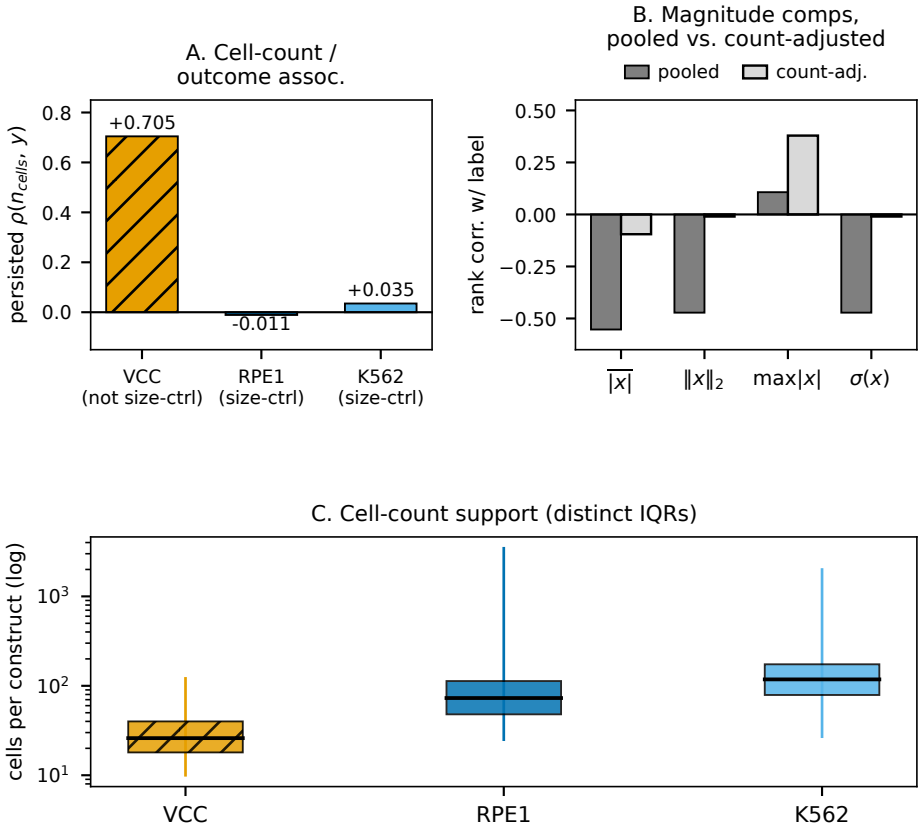


Figure 2: Sampling depth and the endpoint. **(A)** Rank correlation between cells assayed per construct and the outcome: $+0.705$ on VCC, whose endpoint is not size-controlled, against -0.011 on RPE1 and $+0.035$ on K562-essential, whose harmonized endpoint is. **(B)** The four magnitude summaries against the label, pooled and after adjusting for cell count; adjustment moves every summary toward zero except $\max_j |x_j|$, which strengthens. **(C)** Cell-count support per construct on the checkpoint $\mathbf{x}_{\text{cell}}$ accumulator population: the VCC interquartile range, $[18, 40]$ cells, is disjoint from both external ranges, while RPE1, $[48, 113]$, and K562-essential, $[79, 174]$, overlap each other. App. E.1 uses the persisted-embedding set, on which the RPE1 summary differs.

R^2 by only $+0.0017$. Much of the aggregate-magnitude signal that predicts this endpoint therefore overlaps with sampling depth on this benchmark.

This is an association on this benchmark, not a causal claim: partialling out an observed covariate does not establish a causal role, and it does not mean the label "is" cell count—the two are strongly associated on VCC. Nor does it extend automatically to the external screens, whose endpoint is size-controlled and where the association between cell count and outcome is near zero on the eligible sets ($\rho = -0.011$, $p = 0.63$ on RPE1; $\rho = +0.035$ on K562-essential)—though near zero is not proof of absence, and the eligibility threshold excludes sub-threshold constructs, a residual selection pathway whose pre-eligibility count is not recoverable from the record. The aggregate-magnitude reading is therefore entangled with a design factor—exactly what a locked evaluation should surface.

Unlike the endpoint, the magnitude and representation blocks are not size-matched across screens; under a pre-specified matched-count sensitivity the cross-screen ordering changes, so that comparison is depth-dependent—a property of the comparison, not an explanation of the transfer failure (App. E.1).

6.5 A COUNT-ADJUSTED MAX-RESPONSE CORRELATE

One pre-registered secondary survives count adjustment. After partialling out $\log(1 + n_{\text{cells}})$, the single largest-magnitude coordinate $\max_j |x_j|$ retains a positive association with the outcome on VCC ($+0.379$) and on both external screens ($+0.239$ on RPE1, $+0.355$ on K562-essential). We report point estimates only: no interval procedure was pre-registered for this secondary and none is persisted, so we do not attach uncertainty after the fact (App. F). The pre-registered robustness check that *was* specified—stratifying by cell count—finds the association positive within every n_{cells} quintile on both external screens. We report this as a narrow, correlational, count-adjusted secondary: positive on both screens, but not a recovered magnitude signal, not a new headline finding, and not evidence that magnitude successfully transfers.

6.6 WHAT THE LOCKED EVALUATION ESTABLISHES

Read together, the results form one argument: in-distribution information does not imply portability, and the benchmark endpoint that motivated the magnitude reading is itself entangled with a design factor. What the protocol does not settle it leaves open, and we say so rather than filling the gap post-hoc (Sec. 7). The contribution is the locked evaluation itself, which separates in-distribution informativeness, transferable signal, endpoint construction, and design factors, and thereby changes the conclusion a less controlled analysis would reach.

7 DISCUSSION AND LIMITATIONS

What the transfer failure means. The frozen Geneformer representation is informative on VCC yet fails on both external screens, below its capacity-matched control. In-domain predictive information is therefore not sufficient evidence of cross-screen portability, and strong held-out performance within one screen should not be read as transferable signal. This is a statement about this representation under this evaluation, not a claim that foundation models are generally unhelpful.

Why the magnitude interpretation changed, and what survives it. The count-only comparison of Sec. 6.4 refines the earlier aggregate-magnitude reading: the block is entangled with sampling depth rather than a clean measure of response strength. What survives is narrower—the pre-registered, count-adjusted max-response correlate of Sec. 6.5, which we do not promote to a headline finding.

Residual selection, and comparator sensitivity. Both caveats of Sec. 6.4 stand: the external null is not proof of absence, and the eligibility threshold's residual selection pathway is not quantifiable from the record. Separately, the primary and secondary heads can differ materially, as can the two external feature comparators; we fixed both in advance rather than selecting them after seeing outcomes.

The pre-registered interval on VCC is unclustered. The in-distribution interval resamples rows, whereas the external intervals resample target-gene clusters. Because VCC rows can share a target gene or a batch, the row-level procedure may understate uncertainty relative to a clustered alternative; we retained the pre-specified analysis rather than substituting one after the test evaluation.

What was not tested. One explanation is available but untested: $\mathbf{z}$ lands *below* its capacity-matched control externally rather than at zero, which could arise if $\mathbf{z}$ partly encodes sampling depth while the VCC label is depth-associated and the external one is not. We did not test it, deliberately: any analysis designed after unblinding would be post-hoc (a later post-unblinding probe of linear depth recovery, App. E.2, settles nothing here). The pre-specified C_2 comparator bears on this question; we report it descriptively and infer no mechanism from it. Magnitude recoverability (H-FM4) was evaluated on train/validation only; its test-level statement remains unresolved.

Lesson. The separation this evaluation enforces—in-domain informativeness, transferable signal, endpoint construction, design factors such as sampling depth, and headline versus supporting evidence—is what changes the conclusion (Sec. 6.6).

AI USE STATEMENT

We used generative AI assistants (ChatGPT and Claude, the latter including its command-line coding interface) at several stages of this work, and we describe their role in some detail because the contribution of this paper is a protocol whose credibility depends on how it was produced.

Assistants were used to help draft and refine the pre-registration document and its recorded amendments; to write and review analysis code; to execute the analysis pipeline under author direction; to audit the persisted record for consistency between reported values and stored artifacts; and to draft and revise the text of this manuscript. The research question, the decisions taken at each gate of the protocol, and the interpretation of every result are the authors'. Assistant-produced specifications, scripts, and manuscript text were reviewed by the authors before adoption, with quantitative claims checked against the persisted record where applicable.

During the locked revision, assistants were not used to train or refit predictive models, or to re-evaluate the held-out VCC test split. One pre-registered secondary, the non-overlap sensitivity of App. D, was executed after external unblinding under author direction, using stored predictions and the inference procedure inherited from the primary analysis; that execution and its ordering are disclosed there and in the amendment chronology of App. G.2. Numeric claims were checked against persisted artifacts, and values that could not be traced to one were removed rather than retained.

The authors take full responsibility for the content of this submission, including any errors it contains.

ETHICS STATEMENT

This work uses publicly released CRISPRi Perturb-seq screens of cultured cell lines—the Virtual Cell Challenge release and the Replogle RPE1 and K562-essential datasets. It involves no human subjects, no personally identifying information, and no new data collection. We introduce no deployed model or clinical decision system; the study is a methodological diagnostic on public benchmarks, and any societal effects would be indirect and mediated through future perturbation-prediction methods.

REPRODUCIBILITY STATEMENT

The evaluation is governed by a locked, pre-registered protocol (Section 4): heads and model selection are frozen on VCC before any test metric is computed, the locked protocol required the external outcome labels to be held out until final evaluation, and analysis-governing choices are fixed before the results they govern, with subsequent amendments recorded as dated documents. Controls, metrics, the paired target-gene cluster bootstrap, and decision rules are specified in advance; the problem setup, endpoint construction, magnitude features, seeds, and hyperparameters are given in Sections 3–6 and the appendices. The specification, its amendment chain, the environment record, and the artifact digests accompany this submission as supplementary material. Retained analysis code will be released upon acceptance; because some historical execution paths behind the persisted external-evaluation artifacts are not fully retained, we do not claim that every reported artifact can be regenerated from released code.

The record has further limits, and we state them rather than implying an archival guarantee we cannot support: which reported values do not re-derive under the locked revision harness, what the persisted artifacts bind and what they do not, which held-out accesses the durable record can and cannot locate, and how the specification and its amendments are evidenced. App. G sets all of these out in one place.

REFERENCES

Constantin Ahlmann-Eltze, Wolfgang Huber, and Simon Anders. Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. *Nature Methods*, 22(8): 1657–1661, 2025. doi: 10.1038/s41592-025-02772-6.

Martín Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Ihab Bendidi, Shawn Whitfield, Kian Kenyon-Dean, Hanene Ben Yedder, Yassir El Mesbahi, Emmanuel Noutahi, and Alisandra K. Denton. Benchmarking transcriptomics foundation models for

perturbation analysis: One PCA still rules them all. In *NeurIPS Workshop on AI for New Drug Modalities (AIDrugX)*, 2024.

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning (ICML)*, pp. 89–96, 2005. doi: 10.1145/1102351.1102363.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Aréchiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pp. 1567–1578, 2019.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th International Conference on Machine Learning (ICML)*, pp. 129–136, 2007. doi: 10.1145/1273496.1273513.

Gerold Csendes, Gema Sanz, Kristóf Z. Szalay, and Bence Szalai. Benchmarking foundation cell models for post-perturbation RNA-seq prediction. *BMC Genomics*, 26(1):393, 2025. doi: 10.1186/s12864-025-11600-2.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21(8):1470–1480, 2024. doi: 10.1038/s41592-024-02201-0.

Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P. Fulco, Livnat Jerby-Arnon, Nemanja D. Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, et al. Perturb-seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell*, 167(7): 1853–1866.e17, 2016. doi: 10.1016/j.cell.2016.11.038.

Luke A. Gilbert, Max A. Horlbeck, Britt Adamson, Jacqueline E. Villalta, Yuwen Chen, Evan H. Whitehead, Carla Guimaraes, Barbara Panning, Hidde L. Ploegh, Michael C. Bassik, Lei S. Qi, Martin Kampmann, and Jonathan S. Weissman. Genome-scale CRISPR-mediated control of gene repression and activation. *Cell*, 159(3):647–661, 2014. doi: 10.1016/j.cell.2014.09.029.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pp. 1735–1742, 2006. doi: 10.1109/CVPR.2006.100.

Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, and Le Song. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, 21(8):1481–1491, 2024. doi: 10.1038/s41592-024-02305-7.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations (ICLR)*, 2017.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5637–5664. PMLR, 2021.

David Krueger, Ethan Caballero, Jörn-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Rémi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (REx). In *International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5815–5826. PMLR, 2021.

Kun Kuang, Peng Cui, Susan Athey, Ruoxuan Xiong, and Bo Li. Stable prediction across unknown environments. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1617–1626, 2018. doi: 10.1145/3219819.3220082.

Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 21464–21475, 2020.

Mohammad Lotfollahi, F. Alexander Wolf, and Fabian J. Theis. scGen predicts single-cell perturbation responses. *Nature Methods*, 16(8):715–721, 2019. doi: 10.1038/s41592-019-0494-8.

Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L. Ibarra, Sanjay R. Srivatsan, Mohsen Naghipourfar, Riza M. Daza, Beth Martin, et al. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, 19(6):e11517, 2023. doi: 10.15252/msb.202211517.

Gabriel M. Mejia, Henry E. Miller, Francis J. A. Leblanc, Bo Wang, Brendan Swain, and Lucas Paulo de Lima Camillo. Diversity by design: Addressing mode collapse improves scRNA-seq perturbation modeling on well-calibrated metrics. In *2nd Generative AI for Biology Workshop at the 42nd International Conference on Machine Learning (ICML)*, 2025. Non-archival workshop paper. Preprint arXiv:2506.22641.

Jiawei Ren, Mingyuan Zhang, Cunjun Yu, and Ziwei Liu. Balanced MSE for imbalanced visual regression. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7926–7935, 2022.

Joseph M. Replogle and Jonathan S. Weissman. "Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq" Replogle et al. 2022 processed Perturb-seq datasets, 2022. Includes K562 genome-scale, K562 essential, and RPE1 essential Perturb-seq AnnData files, in pseudo-bulk and cell-level form. We use the RPE1 and K562-essential pseudo-bulk and raw single-cell releases.

Joseph M. Replogle, Reuben A. Saunders, Angela N. Pogson, Jeffrey A. Hussmann, Alexander Lenail, Alina Guna, Lauren Mascibroda, Eric J. Wagner, Karen Adelman, Gila Lithwick-Yanai, Nika Iremadze, Florian Oberstrass, Doron Lipson, Jessica L. Bonnar, Marco Jost, Thomas M. Norman, and Jonathan S. Weissman. Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell*, 185(14):2559–2575.e28, 2022. doi: 10.1016/j.cell.2022.05.013.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nature Biotechnology*, 42(6):927–935, 2024. doi: 10.1038/s41587-023-01905-6.

Yusuf H. Roohani, Tony J. Hua, Po-Yuan Tung, Lexi R. Bounds, Feiqiao B. Yu, Alexander Dobin, Noam Teyssier, Abhinav Adduri, Alden Woodrow, Brian S. Plosky, Reshma Mehta, Benjamin Hsu, Jeremy Sullivan, Chiara Ricci-Tam, Nianzhen Li, Julia Kazaks, Luke A. Gilbert, Silvana Konermann, Patrick D. Hsu, Hani Goodarzi, and Dave P. Burke. Virtual cell challenge: Toward a Turing test for the virtual cell. *Cell*, 188(13):3370–3374, 2025. doi: 10.1016/j.cell.2025.06.008.

Yanay Rosen, Yusuf Roohani, Ayush Agrawal, Leon Samotorčan, Tabula Sapiens Consortium, Stephen R. Quake, and Jure Leskovec. Universal cell embedding provides a foundation model for cell biology. *Nature*, 656(8126):183–191, 2026. doi: 10.1038/s41586-026-10689-z.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*, 2020.

F. W. Scholz and M. A. Stephens. K-sample Anderson-Darling tests. *Journal of the American Statistical Association*, 82(399):918–924, 1987. doi: 10.1080/01621459.1987.10478517.

Zheyan Shen, Peng Cui, Kun Kuang, Bo Li, and Peixuan Chen. Causally regularized learning with agnostic data selection bias. In *Proceedings of the 26th ACM International Conference on Multimedia*, pp. 411–419, 2018. doi: 10.1145/3240508.3240577.

Michael Steininger, Konstantin Kobs, Padraig Davidson, Anna Krause, and Andreas Hotho. Density-based weighting for imbalanced regression. *Machine Learning*, 110(8):2187–2211, 2021. doi: 10.1007/s10994-021-06023-5.

Christina V. Theodoris, Ling Xiao, Anant Chopra, Mark D. Chaffin, Zeina R. Al Sayed, Matthew C. Hill, Helene Mantineo, Elizabeth M. Brydon, Zexian Zeng, X. Shirley Liu, and Patrick T. Ellinor. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023. doi: 10.1038/s41586-023-06139-9.

Feng Wang, Xiang Xiang, Jian Cheng, and Alan L. Yuille. NormFace: L_2 hypersphere embedding for face verification. In *Proceedings of the 25th ACM International Conference on Multimedia*, pp. 1041–1049, 2017. doi: 10.1145/3123266.3123359.

Yan Wu, Esther Wershof, Sebastian M. Schmon, Marcel Nassar, Błażej Osiński, Ridvan Eksi, Zichao Yan, Rory Stark, Kun Zhang, and Thore Graepel. PerturBench: Benchmarking machine learning models for cellular perturbation analysis. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

Yuzhe Yang, Kaiwen Zha, Ying-Cong Chen, Hao Wang, and Dina Katabi. Delving into deep imbalanced regression. In *International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11842–11851. PMLR, 2021.

A THE FROZEN FOUNDATION-MODEL REPRESENTATION

Model and pinned artifacts. The representation is produced by Geneformer-V2-104M (Theodoris et al., 2023), pretrained and used frozen with no fine-tuning: twelve transformer layers, hidden width 768, twelve attention heads, a vocabulary of 20,275 tokens, and a maximum sequence length of 4,096. The artifacts below were pinned before any result was produced; digests are the first sixteen hexadecimal characters of the SHA-256 of each file, and the full digests accompany the code release.

Artifact	Digest
Geneformer repository commit	`ad8f66dfcda3ebbd`
V2-104M checkpoint, 417,571,156 bytes	`fff5cba29ddd8792`
gc104M token dictionary	`67c445f4385127ad`
gc104M gene-median dictionary	`a51c53f6a771d645`
gc104M Ensembl mapping dictionary	`0819bcbd869cfa14`

What "frozen" means here. The encoder is loaded in evaluation mode. No optimizer, backward pass, or gradient-enabled encoder parameter appears anywhere in the analysis pipeline, and re-embedding the same cells reproduces the stored features exactly, to a maximum absolute difference of zero. The only trained components are the downstream heads of App. C.

Vocabulary coverage. The Geneformer token vocabulary covers 17,975 of 18,080 measured genes in VCC (99.42%), 8,520 of 8,749 in RPE1 (97.38%), and 8,210 of 8,563 in K562-essential (95.88%). Coverage is therefore high and comparable across the three screens, so gross vocabulary loss is unlikely on its own to explain the negative external transfer of Sec. 6.2. Other recorded cross-screen differences, set out below, remain.

Input and tokenization. The input is the per-cell raw count vector. Genes are identified by Ensembl accession and passed to the model's own tokenizer, which applies per-cell count normalization and gene-wise median scaling and emits a rank-ordered token sequence truncated to 4,096 tokens. Duplicate, unmapped, and absent genes are handled inside the tokenizer and its pinned mapping dictionary; we impose no additional gene rule. Batching used a chunk size of 512, which affects throughput only.

Cell selection. Cells enter the embedding if they belong to an eligible construct or are non-targeting. For VCC, constructs are additionally required to carry at least ten cells in their batch and target-gene group. Beyond this we apply no additional explicit per-cell count, gene-count, or doublet filter outside the tokenizer; any tokenizer-internal filtering is not separately recorded. We state this as a limit of the record rather than as evidence that no such filtering occurred.

Pooling and aggregation. Per cell we take the class-token activation at the second-to-last transformer layer, giving $E_i \in \mathbb{R}^{768}$ in single precision. Padding is masked; no other token masking is applied. Perturbation-level vectors are

$$\mathbf{z}_g \;=\; \frac{1}{|g|}\sum_{i\in g} E_i \;-\; \mu_{\mathrm{NT}}, \qquad \mathbf{z}_g \in \mathbb{R}^{768}, \tag{3}$$

where E_i is the embedding of cell i and g ranges over the cells of one construct. For the external screens the grouping key is the construct identifier; for VCC, rows are joined on split, batch, and target gene.

The non-targeting reference. μ_{NT} is the mean embedding of cells annotated non-targeting. Its scope differs by screen, and the difference is recorded rather than incidental. For VCC it is batch-local: one non-targeting mean per Flex batch, subtracted from the constructs in that batch. For RPE1 it is a single globally pooled mean over $11{,}485$ non-targeting cells, and for K562-essential a single globally pooled mean over $10{,}691$, in both cases with no batch matching and no subsampling.

What is identical across screens, and what the record shows is not. The extraction configuration was verified identical across VCC, RPE1, and K562-essential in checkpoint, embedding mode, layer, numerical precision, sequence length, chunking, subtraction form, and output type. Four things are *not* identical, and we state them because each is a possible alternative explanation for the external results of Sec. 6.2.

1. *Scope of the non-targeting reference*—batch-local for VCC, globally pooled for both external screens.
2. *Size of the non-targeting pool*—$11{,}485$ cells for RPE1 against $10{,}691$ for K562-essential.
3. *Provenance of the embeddings*—those for VCC and for $1{,}889$ of the $1{,}932$ RPE1 constructs were read from a precomputed store generated with the same pinned checkpoint and the same extraction configuration; the remaining 43 RPE1 constructs and all K562-essential constructs were embedded afresh. The value 43 appears in three distinct contexts in this manuscript: raw VCC-training-target overlap, RPE1 constructs newly embedded during top-up, and RPE1 constructs for which the batch-matched control is computable. The first two are the same set of constructs: the upstream pipeline excluded the VCC-training-overlap constructs, so exactly those were absent from the precomputed store and were the ones embedded during top-up. The third is a different set over a different universe and should not be conflated with them. Under the recorded extraction configuration, deterministic re-embedding showed numerical agreement, so cached and freshly computed embeddings were treated as equivalent for aggregation.
4. *RPE1 cell sets*—the RPE1 embedding set and the RPE1 cell-count accumulator used elsewhere in the paper are distinct objects produced by different quality filters, with minima of 27 and 25 cells per construct and non-targeting pools of $11{,}485$ and $11{,}007$ cells respectively.

We did not test whether the asymmetry in the non-targeting reference contributes to the negative external transfer reported in Sec. 6.2. The frozen protocol admits no post-unblinding re-analysis, so this remains a live alternative explanation and we flag it as such rather than dismissing it.

Scaling after construction. Each feature block carries its own standardizer—centering and scaling to unit variance—fit on that block's VCC training rows alone and applied with frozen statistics to validation, test, RPE1, and K562-essential. The representation and each control carry separate standardizers. No ℓ_2 normalization, whitening, principal-component projection, or clipping is applied to $\mathbf{z}$; the clipping referenced elsewhere in this paper applies only to the expression and magnitude blocks.

Determinism and seeds. No random seed governs the construction of $\mathbf{z}$. Seeds enter only downstream: head training, the secondary matched-subsampling sensitivity analysis, and the bootstrap intervals of App. F.

A limit on provenance. The persisted result records do not carry a content digest binding a specific stored copy of $\mathbf{z}$ to the reported external correlations. App. G.4 states what is pinned instead, and what that does and does not support.

B CONTROL DEFINITIONS

All five controls were specified in the pre-result lock. Feature-generation seeds are disjoint from head-training seeds, and every control is generated from array shapes alone: no label, no validation

statistic, and no test statistic enters any control. Below, the split index runs over training, validation, test, RPE1, and K562-essential in that order, and each control's generator is seeded at its base value plus that index.

C_1: the capacity floor (gating). A matrix in $\mathbb{R}^{n\times 768}$ with independent standard Gaussian entries drawn per row in single precision, base seed 20270101, one deterministic draw per split, generated independently within each split from the row count alone. External splits receive their own draws; nothing is transferred from VCC. The expected row norm is $\sqrt{768} \approx 27.71$. Exactly one realization exists per split, so reported values are not averaged over several Gaussian draws and the floor is conditioned on that one pre-specified draw. Drawing more would have required choosing, after the fact, how many to draw and how to combine them, which the specification does not fix; we note the dependence rather than resolve it after unblinding. "Capacity-matched" here means matched to $\mathbf{z}$ in input dimension, and sharing its preprocessing, candidate-head pool, selection rule, seeds, and training budget. It is a zero-information floor, and dimension matching is the criterion the specification enforces; we claim no stronger notion of equal effective capacity.

C_2: compressed expression benchmark (secondary, non-gating). A random projection of the 2,000-dimensional expression vector to 768 dimensions with independent Gaussian entries of variance $1/768$, drawn once with seed 20270102 and applied unchanged to every split.

C_3: alignment control. The representation of Eq. 3 with its rows permuted. The permutation is drawn per split from base seed 20270103 and applied to that split's rows; the *same* permutation is used at training and at evaluation within a split, and different permutations are used across splits. The permutation is unconstrained, and the standardizer fit on the unpermuted representation is reused, so the marginal distribution of $\mathbf{z}$ is preserved exactly and only the row-to-label correspondence is destroyed.

C_4: alignment control on the magnitude block. The concatenation $[\,\mathbf{z}\,;$ row-permuted $\widetilde{\boldsymbol{m}}\,] \in \mathbb{R}^{n\times 772}$, with the magnitude rows permuted per split from base seed 20270104.

C_5: dimension-increment control. The representation concatenated with four independent standard Gaussian scalars, giving a matrix in $\mathbb{R}^{n\times 772}$, drawn per row per split from base seed 20270105. Together, C_4 and C_5 separate a real contribution of the magnitude block from the mere addition of four columns.

Pre-registration. All five controls appear in the original pre-result specification. The recorded amendments govern external baselines, the external head, cell-count-matching seeds, and secondary analyses; the remainder are editorial or record-keeping, and none adds, removes, or reseeds a control. The specification forbids introducing a control after primary results and forbids changing a control seed because a result looks unusual.

C HEAD FAMILIES, SELECTION, AND FREEZING

Candidate pool. Four head families with fixed configurations—not a grid, and no hyperparameter search of any kind: ridge regression with unit regularisation; a random forest of 200 trees with maximum depth 20 and at least two samples per leaf; histogram gradient boosting with 500 boosting iterations, maximum depth 6, learning rate 0.05, and early stopping enabled; and a 256-unit multilayer perceptron. Each is a single-output regressor on y and differs across arms only in input width, 768 or 772.

Budget. Seven arms, four families, and three training seeds give 84 fits, identical for every arm.

Selection rule. Per arm, the argmax of the seed-mean VCC *validation* R^2. No test metric influences selection. The rule is identical for $\mathbf{z}$, for $[\mathbf{z};\widetilde{\boldsymbol{m}}]$, and for every control.

Realised selections and their asymmetries. Under this single rule the arms resolved to different families: a random forest for $\mathbf{z}$, $[\mathbf{z};\widetilde{\boldsymbol{m}}]$, C_2, C_4, and C_5, and histogram gradient boosting for C_1 and

C_3. The multilayer perceptron diverged on every arm except C_2, so the effective candidate pool was three families for six arms and four for C_2. We report these asymmetries because they are the natural first objection to the primary contrast: the procedure was symmetric, the outcomes were not.

Freezing. The selection was written to disk before any VCC test label was scored. The ordering is established by the within-process write order and by the execution log, in which the freeze record precedes every test entry, rather than by filesystem timestamps. Test *predictions* were materialized during training, before the freeze; only the scoring of predictions against test labels occurs afterwards, and selection consumed validation predictions alone.

Robustness of the selection. The executed selection used an unguarded sort, so the ranking of divergent entries was not explicitly specified. In this run the divergent multilayer perceptron entries ranked below all finite maxima, and the selected family is the argmax over finite validation scores for every one of the seven arms. We record the fragility because it was real, and the outcome because it is verifiable from the persisted validation table.

D PRE-REGISTERED HYPOTHESES

Four hypotheses were fixed in the pre-result specification, together with their gating controls, their inference procedures, and their pass criteria. We reproduce each criterion and report its outcome.

H-FM1: in-distribution information. Criterion: the frozen representation is informative if and only if the paired bootstrap interval on $\Delta R^2 = R^2(\text{FM-only}) - R^2(C_1)$ excludes zero on the positive side, with C_1 alone as the gate. Inference: row-level paired bootstrap, $B = 10{,}000$, seed 42, 95% percentile interval (App. F). **Outcome: passed**, with $\Delta R^2 = +0.1645$ and interval $[+0.1375, +0.1920]$. Reported alongside and non-gating, as the specification requires: $\Delta R^2(\mathbf{z} - C_2) = +0.1107$, interval $[+0.0906, +0.1317]$.

H-FM2: zero-shot transfer. Criterion: interpretable *only if* the C_1 gate of H-FM1 passed. The pre-registered expectation was that the frozen representation would behave like the expression-only group, with an external Spearman correlation negative or unresolved on *both* screens. The specification further states that if the gate fails, a near-zero external correlation must not be reported as a transfer finding; the admissible statement in that case would have been that usable in-distribution information had not been established. **Outcome: the gate passed, so the external result is interpretable**, and the pre-registered expectation was met, with $\rho = -0.139$ on RPE1 and -0.267 on K562-essential, negative and resolved on both screens. The expected direction was recorded before the first documented external-outcome access.

H-FM3: magnitude increment. Criterion: compare the magnitude-augmented representation against the frozen representation on both screens, with the paired difference in Spearman correlation as the primary quantity; a resolved positive increment is required on *both* screens, mirroring the increment previously observed from SPSM to MAG-SPSM, and the increment must survive C_4 and C_5. If a row-permuted magnitude block or four Gaussian scalars reproduce the gain, no claim of per-row magnitude alignment may be made, and no single-screen success may be claimed. Inference: paired target-gene cluster bootstrap, $B = 10{,}000$, 95% percentile interval. **Outcome: the criterion was met on both screens**, with $\Delta\rho = +0.032$ $[+0.008, +0.056]$ on RPE1 and $+0.143$ $[+0.111, +0.175]$ on K562-essential, both resolved positive and both surviving C_4 and C_5. The RPE1 increment falls below, and the K562-essential increment above, the mirroring values quoted in the pre-registered expectation; the criterion was a resolved positive increment on both screens rather than a magnitude threshold, and we do not reinterpret it as one.

Pre-specified non-overlap sensitivity. A small fraction of externally evaluated constructs target genes that also occur among the 150 raw VCC training targets: 43 of 1,932 on RPE1, spanning 42 genes, and 33 of 1,903 on K562-essential, spanning 32 genes. The exclusion universe is the raw training targets, not the 140 that survive the eligibility filter of Sec. 3, and the two are not interchangeable here: the eligible set would have removed 37 constructs on RPE1 and 29 on K562-essential rather than 43 and 33; these counterfactual counts were reconstructed under a frozen

metadata-only procedure from the pre-specified set definitions, after first reproducing the historical 43 and 33 counts (App. G.3). Excluding on the raw set is the more conservative of the two rules, simply because the raw set contains the eligible one and therefore removes strictly more. The additional six constructs on RPE1 and four on K562-essential carry no exposure risk in any case: a target gene that the cell-count floor removed from training contributes no training row, so the model never saw it. Occurring among the raw 150 is therefore not the same as having been seen during training. The specification fixes the primary on all eligible constructs and pre-registers a secondary that repeats the external comparison with these constructs removed, so that transfer can be read on target genes the model has never seen. This secondary was not executed at the time the primary was persisted. It was run afterwards, after external unblinding, and that ordering is recorded in the amendment chain. The specification fixes no inference procedure for it, so we inherited the primary's: a paired target-gene cluster bootstrap with $B = 10{,}000$, percentile intervals, and the primary's seeds. Before subsetting, re-scoring the full sets from the stored predictions reproduced every reported correlation, interval, and contrast exactly, so any difference below is attributable to the exclusion rather than to the re-scoring.

On the retained sets—1,889 constructs for RPE1 and 1,870 for K562-essential—no verdict changes. The frozen representation remains resolvedly negative on both screens ($\rho = -0.147$ and -0.273) and remains resolvedly below its capacity floor; C_1 and C_3 remain compatible with null; and the magnitude increment remains resolved positive ($\Delta\rho = +0.033$ and $+0.135$). No quantity moves by more than 0.014, and the external correlations shift slightly away from zero rather than toward it. This outcome is inconsistent with target overlap being the explanation for the negative transfer. The shifts are small relative to the intervals, and we do not read a direction into their magnitude.

H-FM4: magnitude recoverability (mechanism diagnostic). Criterion: four separate single-output ridge probes, one per magnitude scalar rather than a shared multi-output fit, trained on the VCC training split with regularisation selected on validation. Each probe is read against matched references from the same probe family—a ceiling given by predicting the scalar from the expression vector, and two floors given by a random projection of the expression vector and by the row-permuted representation. No absolute threshold was set: all four components were to be reported by their position between the matched floors and the ceiling, without summarizing to a single number and without cherry-picking. **Outcome: evaluated on train/validation only.** The VCC test split was closed before the budgeted test-level evaluation could be run; the diagnostic was later executed under the frozen design on the training and validation splits alone, as a supporting diagnostic. The test-level statement therefore remains unresolved, and the mechanism it was designed to probe—whether $\mathbf{z}$ carries the magnitude information that $\widetilde{\boldsymbol{m}}$ supplies—is not settled by train/validation evidence. We report this rather than omitting it, because a pre-registered analysis and the restricted form in which it was eventually run are both part of the record.

D.1 Locked external evaluation, in full

Table 1 gives the current locked external evaluation for every pre-registered arm and contrast on both screens. It is the evaluation Sec. 6.2 reports; the legacy classical-transfer table elsewhere in this appendix is a separate, historical object and is not this evaluation.

D.2 Pre-specified secondary external analyses

The locked protocol contained several pre-specified external secondary analyses. The two reporting obligations omitted from the previous manuscript version were A1's $\mathbf{x}_{\mathrm{release}}$ feature-basis comparisons and A2's ordinary linear head on $[\mathbf{z}; \widetilde{\boldsymbol{m}}]$; Table 2 also reports the two pre-specified A4 secondaries, MAX-LINEAR and MAG+COUNT. All were fixed before external evaluation and executed only after the frozen primary had been persisted, following the recorded execution ordering. All remain secondary: none replaces the frozen primary result, and disagreements are reported rather than resolved by post-outcome selection.

Their scope is bounded by the amendments that fixed them. The A2 linear head applies only to arms containing $\widetilde{\boldsymbol{m}}$, not to $\mathbf{z}$-only arms; it therefore re-evaluates no hypothesis, in particular not H-FM3, and leaves the $\mathbf{z}$-only external results of Sec. 6.2 untouched. The A1 rows repeat the corresponding external arm and control definitions on the submitted $\mathbf{x}_{\mathrm{release}}$ feature basis rather than the frozen-primary $\mathbf{x}_{\mathrm{cell}}$ basis; the $\mathbf{x}_{\mathrm{cell}}$ counterparts are in Table 1, so the two bases can be compared arm by

Table 1: Primary zero-shot external evaluation: frozen VCC heads applied unchanged (no target-side calibration) to RPE1 and K562-essential on the construction-matched $\mathbf{x}_{\text{cell}}$ basis. Entries are construct-level Spearman ρ against the harmonized target-gene endpoint with 95% paired target-gene cluster-bootstrap intervals ($B = 10{,}000$; per-screen seeds RPE1 20260901, K562 20260902). $\mathbf{z}$ lands *below* its capacity-matched control C1 on both screens.

Arm	RPE1 ρ [95% CI]	K562-essential ρ [95% CI]
$\mathbf{z}$ (FM-only)	-0.139 $[-0.185, -0.092]$	-0.267 $[-0.310, -0.224]$
$[\mathbf{z}; \widetilde{\boldsymbol{m}}]$	-0.107 $[-0.152, -0.062]$	-0.124 $[-0.170, -0.080]$
C1 (Gaussian)	-0.034 $[-0.078, +0.012]$	$+0.032$ $[-0.012, +0.077]$
C2 (proj. $\mathbf{x}$)	-0.156 $[-0.202, -0.109]$	-0.264 $[-0.306, -0.221]$
C3 (shuffled $\mathbf{z}$)	$+0.044$ $[-0.001, +0.088]$	-0.004 $[-0.049, +0.041]$
C4	-0.138 $[-0.184, -0.091]$	-0.266 $[-0.309, -0.223]$
C5	-0.139 $[-0.184, -0.091]$	-0.269 $[-0.312, -0.226]$
Pre-registered paired contrasts ($\Delta\rho$, 95% CI)		
$\mathbf{z}-$C1	-0.106 $[-0.168, -0.044]$	-0.299 $[-0.360, -0.238]$
$[\mathbf{z}; \widetilde{\boldsymbol{m}}] - \mathbf{z}$ (H-FM3)	$+0.032$ $[+0.008, +0.056]$	$+0.143$ $[+0.111, +0.175]$
$[\mathbf{z}; \widetilde{\boldsymbol{m}}]-$C4	$+0.031$ $[+0.007, +0.056]$	$+0.142$ $[+0.111, +0.174]$
$[\mathbf{z}; \widetilde{\boldsymbol{m}}]-$C5	$+0.031$ $[+0.008, +0.056]$	$+0.144$ $[+0.113, +0.177]$

arm. What these establish is that the magnitude-augmented arm is sensitive to the pre-specified choice of head and of feature basis: the linear head is resolvedly positive on RPE1 and inconclusive on K562-essential, while the $\mathbf{x}_{\text{release}}$ basis weakens the negative RPE1 association and is inconclusive on K562-essential. The persisted paired comparison between the two bases is in the table caption.

MAX-LINEAR. MAX-LINEAR is the pre-specified one-feature zero-shot linear predictor fit on VCC train using $\max_j |x_j|$ alone. It is distinct from the count-adjusted partial-Spearman diagnostic of $\max_j |x_j|$ in Sec. 6.5, which partials out $\log(1 + n_{\text{cells}})$; the two are different estimands and should not be read as one quantity. MAX-LINEAR is positive on both external screens, whereas the frozen $\mathbf{z}$-only predictor is negative. Because MAX-LINEAR uses a different one-dimensional magnitude signal and readout, this contrast shows that external transfer behaviour depends on the predictive signal and estimator being evaluated; it does not identify the mechanism of the $\mathbf{z}$-only failure, and it is not promoted.

MAG+COUNT, and a limitation recorded in advance. MAG+COUNT appends $\log(1 + n_{\text{cells}})$ to the magnitude block. The external count ranges extend far beyond the VCC-training range (VCC 10–121; RPE1 25–3,451; K562-essential 27–1,996), so the frozen external linear head extrapolates on this feature. This limitation was specified before execution and is reported here alongside the result; it is not an explanation inferred from the observed outcome.

A metadata caution. The persisted secondary artifact records a cell-set field, `RPE1_cellset_for_counts`, identifying the checkpoint population behind its count-derived quantities. That field describes count provenance only. It is not a qualifier on the correlations reported here, which are computed over the persisted evaluated construct sets, and it is a different population from the persisted-embedding cell set used by the matched sensitivity of App. E.1. No separately persisted paired contrast between any of these secondaries and the $\mathbf{z}$-only arm is reported here.

D.3 Position in the pre-specified interpretation table

The specification fixed a two-way interpretation table before any number existed, crossing whether $\mathbf{z}$ transfers well or poorly with whether $\widetilde{\boldsymbol{m}}$ is recoverable from $\mathbf{z}$. Under that table the observed result occupies the "$\mathbf{z}$ transfers poorly" row. The recoverability column remains undetermined, because H-FM4 was evaluated on train/validation only and was never run on the held-out VCC test split; accordingly, neither test-level column is assigned. H-FM4 is a VCC in-distribution diagnostic, and the specification defines no external stage for it.

Table 2: Pre-specified secondary external analyses (construct-level Spearman ρ with 95% paired target-gene cluster-bootstrap intervals, $B = 10{,}000$). Every row is a SECONDARY analysis; none is promoted, and none replaces a frozen primary result. The three A1 rows repeat the corresponding external arm and control definitions on the submitted $\mathbf{x}_{\text{release}}$ feature basis; their frozen-primary $\mathbf{x}_{\text{cell}}$ counterparts are in Table 1. For $[\mathbf{z}; \widetilde{\boldsymbol{m}}]$ the persisted $\mathbf{x}_{\text{cell}}$ minus $\mathbf{x}_{\text{release}}$ paired contrasts are -0.056 $[-0.084, -0.028]$ on RPE1 and -0.131 $[-0.179, -0.084]$ on K562-essential; the disagreement is reported without selecting either basis post hoc. MAX-LINEAR is the pre-specified one-feature zero-shot linear predictor fit on VCC train using $\max_j |x_j|$ alone; it is distinct from the count-adjusted partial-Spearman diagnostic of $\max_j |x_j|$ in Sec. 6.5, which partials out $\log(1 + n_{\text{cells}})$. The intervals shown for MAX-LINEAR are the persisted intervals for this zero-shot predictive secondary; they are not intervals for that separate diagnostic, for which no interval procedure was pre-specified or attached.

Secondary analysis	RPE1 ρ [95% CI]	K562-essential ρ [95% CI]
A4: MAX-LINEAR	$+0.232$ $[+0.185, +0.278]$	$+0.354$ $[+0.310, +0.395]$
A4: MAG+COUNT	-0.027 $[-0.072, +0.017]$	$+0.033$ $[-0.013, +0.077]$
A2: linear head on $[\mathbf{z}; \widetilde{\boldsymbol{m}}]$	$+0.097$ $[+0.054, +0.140]$	$+0.020$ $[-0.025, +0.064]$
A1: $[\mathbf{z}; \widetilde{\boldsymbol{m}}]$ on $\mathbf{x}_{\text{release}}$	-0.052 $[-0.097, -0.006]$	$+0.007$ $[-0.041, +0.054]$
A1: C_2 on $\mathbf{x}_{\text{release}}$	-0.175 $[-0.221, -0.128]$	-0.274 $[-0.317, -0.231]$
A1: C_4 on $\mathbf{x}_{\text{release}}$	-0.139 $[-0.184, -0.091]$	-0.267 $[-0.310, -0.224]$

E SAMPLING-DEPTH DIAGNOSTICS

E.1 A PRE-SPECIFIED FEATURE-SIDE SENSITIVITY BEFORE DOCUMENTED EXTERNAL-OUTCOME ACCESS

Why this exists, and when it was fixed. The harmonized external endpoint is built from fixed-size subsamples, but the magnitude and representation blocks are means over all of a construct's cells, so cells per construct differ across screens by construction. A recorded amendment fixed a sensitivity for exactly this asymmetry on 2026-08-13, before any of its own measurements existed. The surviving local record places this sensitivity before the first documented external-outcome access.

Design. Two matched cell counts were fixed mechanically from the observed per-construct minima, with no discretionary choice: $n_{\text{match}} = 10$, the feature path's existing floor, at which every construct in every screen qualifies; and $n_{\text{match}} = 27$, the maximum of the three screens' minima (10, 25, 27) in the table available at specification time. For RPE1 that minimum of 25 was measured on the checkpoint $\mathbf{x}_{\text{cell}}$ accumulator; the executed construction-matched sensitivity instead operated on the persisted-embedding cell set, whose RPE1 minimum is 27. For each construct and each draw, n_{match} perturbed cells are sampled once and *both* blocks are rebuilt from that same cell set, so construction parallelism is preserved. The non-targeting reference is never subsampled: resampling it would perturb a second noise source that enters every construct in a screen. Three sampling seeds were pre-registered (20270201, 20270202, 20270203). No embedding was recomputed—the representation is re-aggregated from the persisted per-cell embeddings.

This is a different intervention from the endpoint's fixed-size subsampling of App. P. The two share the number 27 by coincidence of derivation—the endpoint's is a VCC median, this one a maximum over screen minima—and they act on opposite sides of the pipeline. They should not be conflated.

Retention is part of the intervention, and was declared in advance. At $n_{\text{match}} = 10$ every construct is retained in all three screens (5,740 VCC, 1,932 RPE1, 1,903 K562-essential). At $n_{\text{match}} = 27$, VCC retains 2,843 of 5,740 and drops 2,897, while RPE1 retains 1,932 of 1,932 (none dropped) and K562-essential 1,903 of 1,903. The drop the amendment anticipated for RPE1's 25–26-cell constructs therefore did not occur on the executed population. The amendment stated before the measurement that VCC's median is 26, that roughly half of it would therefore drop, and that the survivors would be its high-count constructs: the retained VCC median rises from 26 (IQR $[18, 40]$) to 40 (IQR $[32, 50]$), while RPE1 and K562-essential are unselected at 76 (IQR $[50, 118]$) and 118 (IQR $[79, 174]$). The RPE1 summary here—median 76, IQR $[50, 118]$—is that of the persisted-embedding

cell set this sensitivity used; the 73 with IQR $[48, 113]$ reported for the sampling-depth accumulator (Fig. 2C) describes the distinct checkpoint $\mathbf{x}_{\text{cell}}$ population over the same $1{,}932$ constructs. Only VCC's retained set is a biased high-count sample. This is a declared property of the intervention rather than a discovery about it, and both counts are reported precisely because they trade off in opposite directions: $n = 10$ carries no retention-induced selection but leaves more estimation noise, while $n = 27$ leaves less noise but selects on cell count.

Result: the ordering changes with the matched count. Over all cells the cross-screen ordering of median magnitude is K562-essential $<$ VCC $<$ RPE1. At *both* matched counts it becomes VCC $<$ K562-essential $<$ RPE1: RPE1 stays largest while VCC and K562-essential exchange places. The change is stable across all three seeds with non-overlapping intervals, and the representation block shows the same reordering as the magnitude block. Under the interpretation fixed before these numbers existed, an ordering that changes at either count makes the cross-screen magnitude comparison sampling-depth dependent and obliges us to state that the magnitude features, unlike the endpoint, are not size-matched across screens. We state it.

What this does not show. It does not establish a causal role for sampling depth in the external transfer failure, it does not show that matching features repairs transfer, and it licenses no claim of robustness to sampling depth. A matched-feature external *evaluation* was deliberately placed outside the amendment's scope; it would have required its own amendment committed before the external labels were opened, and it was never run. What is established is narrower and wholly descriptive: a comparison this paper makes across screens moves when feature-side sampling depth is equalised, so that comparison should be read as depth-dependent.

E.2 Two later diagnostics, recorded and executed after unblinding

Both diagnostics below were specified and executed on 2026-08-29, after external unblinding. Each was locked before its own result existed, but neither belongs to the original pre-result specification, neither gates or revises any hypothesis, and neither touches the VCC test split or any external outcome. We report them because they bear on limitations stated elsewhere in this paper.

A linear depth-recovery probe on the representation. Target: $\log(1+n_{\text{cells}})$ over a construct's perturbed cells. Data: VCC training and validation rows only. Predictor: the frozen representation $\mathbf{z}$ alone. Estimator: ridge regression over the pre-specified penalty grid $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$, selected on validation R^2. Validation R^2 rose monotonically across the grid and stayed negative throughout, from -0.140 at 10^{-3} to -0.015 at 10^3; the selected penalty was the grid's upper endpoint, 10^3.

Both clauses of the result travel together: this probe did not linearly recover sampling-depth-associated information from the frozen representation on VCC train/validation under the pre-specified estimator and grid, *and* the regularisation optimum was not bracketed above within that grid, so non-recovery is not established even for this estimator family. It is in particular not evidence that $\mathbf{z}$ fails to encode sampling depth, that depth information is absent, or that the depth explanation considered in Sec. 7 has been ruled out. The grid was frozen in advance and was not widened after the boundary selection was observed.

Target-gene structure in the cell-count target. This sharpens the clustering limitation noted in Sec. 7 and App. F. Treating $\log(1 + n_{\text{cells}})$ as the response and applying a one-way method-of-moments variance decomposition to VCC training and validation rows separately: grouping by target gene gives an intraclass correlation of 0.86 on training (140 groups) and 0.82 on validation (49 groups), while grouping by the recorded batch gives small negative estimates on both (-0.001 and -0.011, over 48 groups each). The batch estimates are computed from $N_{\text{batches}} = 48$ groups; when the number of groups is limited, the between-batch mean square and the resulting method-of-moments between-group component may be unstable. Cell count is therefore strongly clustered by target gene, with no positive one-way clustering along batch.

Four qualifications are mandatory. The two intraclass correlations are separate one-way estimates, mutually confounded to an unquantified degree; they are not components of a joint decomposition and do not sum to a total. The associated design-effect figures (roughly 34.4–39.3 for target gene) are

conventional approximations, not literal effective sample sizes, and a negative estimate mechanically drives the corresponding figure below one. Nothing here concerns the VCC test split. And nothing here shows that the outcome y is itself target-gene clustered, that target gene is the true statistical unit, or that the bootstrap coverage of App. F is thereby validated: the pre-specified row-level interval was retained rather than replaced, as stated in Sec. 7.

F UNCERTAINTY QUANTIFICATION

Two distinct resampling procedures are used in this paper. They are not interchangeable, and we set them out separately.

In-distribution comparison on VCC (Sec. 6.1). The statistic is the pooled ΔR^2 between two arms on the frozen VCC test split. The resampling unit is the test row. The procedure is an independent row bootstrap with $B = 10{,}000$ resamples and seed 42; one resample index is applied to *both* arms in each replicate, so the comparison is paired, and the three head-training seeds are averaged *inside* each replicate. Intervals are percentile intervals at the 2.5th and 97.5th points. No clustering is used. For $\Delta R^2(\mathbf{z} - C_1)$ this gives a point estimate of $+0.1645$, a bootstrap median of $+0.1649$, a 95% interval of $[+0.1375, +0.1920]$, and sign-stability across every resample.

External transfer (Sec. 6.2). The statistic is the Spearman correlation between prediction and outcome, and paired differences in that correlation between arms. The resampling unit is the target-gene cluster. Each screen contains $1{,}811$ target-gene clusters, but these are distinct screen-specific gene sets of equal cardinality, not a shared set. Constructs sharing a target gene are resampled together. The procedure is a paired cluster bootstrap with $B = 10{,}000$ resamples and seeds 20260901 for RPE1 and 20260902 for K562-essential, with percentile intervals at the 2.5th and 97.5th points; the same cluster resample is applied to both arms of a contrast within each replicate, and head seeds are averaged *outside* the bootstrap.

These are not the same procedure. The two differ in statistic—pooled ΔR^2 against Spearman $\Delta\rho$—in resampling unit, row against target-gene cluster; in whether clustering is applied at all; in seed; and in whether head seeds are averaged inside or outside the resample. They share only the number of resamples and the percentile method. Intervals from the two settings should not be read as though produced by a common procedure.

Secondary magnitude correlations (Sec. 6.5). The count-adjusted partial Spearman correlations of $\max_j |x_j|$ with y are reported without intervals. No interval procedure was pre-specified for them and none is persisted, so we report the point values and do not attach uncertainty after the fact. The pre-specified robustness check for these correlations was a within-quintile stratification of cell count; it was executed, and the association is positive within every quintile on all three screens.

G EVALUATION INTEGRITY RECORD

This appendix is the single record of how the evaluation was governed and of what the durable evidence does and does not support: the protocol timeline (Fig. 3), the held-out data access ledger, the amendment chain, the chronology evidence with its limits, and the provenance limits that attach to the frozen representation and to results carried over from the earlier submission. It reports no model result; the one derivation it contains is the provenance reconstruction of App. G.3. It is the record against which the protocol of Sec. 4 and the results of Sec. 6 should be read.

G.1 HELD-OUT DATA ACCESS LEDGER

The VCC test split. Exactly one scientific evaluation of the frozen VCC test split was performed: the frozen test table, produced after head selection had been written to disk. Two further accesses are on the record and we list them rather than fold them into a single count—a reproduction-gate run before the freeze, which did not complete, and one post-hoc diagnostic access described in the session record. Neither contributed a reported number, and every phase after the specification lock

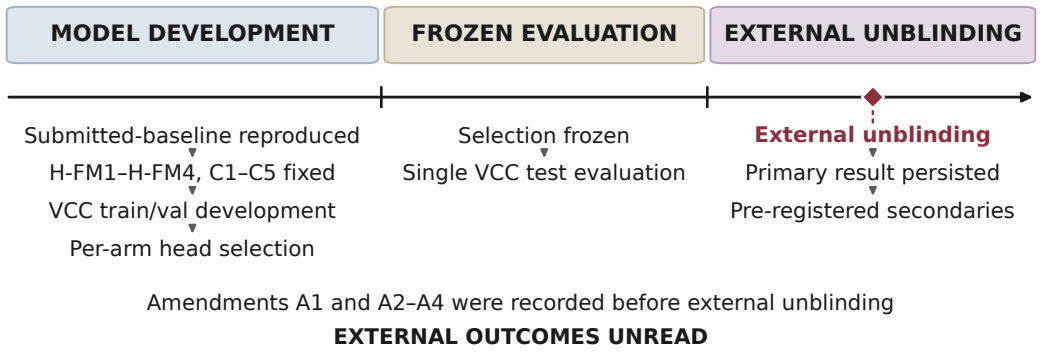


Figure 3: Locked, pre-registered evaluation protocol. Model and head selection are frozen on VCC before a single test evaluation; under the locked protocol the external outcome labels were to be held out until final evaluation; amendments that govern an analysis precede its results, and later amendments are recorded as dated documents. The in-distribution gate (H-FM1) precedes and conditions the zero-shot transfer reading (H-FM2). The figure's label EXTERNAL OUTCOMES UNREAD states that protocol requirement. The surviving local record positively locates the first documented external-outcome access at final evaluation and does not exclude undocumented earlier access (App. G.3 below). The pre-registered secondaries stage shown here is reported in App. D.2.

records the test split as untouched. A single unqualified access count is not derivable from the durable record, and we do not assert one.

The external screens. Under the locked protocol, external outcomes were to remain held out until final evaluation. The surviving local record positively locates the first documented external-outcome access at that point, while not excluding undocumented earlier access. Head selection was frozen on one day; external predictions were materialized blind, and the primary result persisted, on the next; and the first documented external-outcome access follows both. Every prior session record states that the external outcome had not been read.

No external fitting. Training and head selection used VCC labels alone. All preprocessing statistics were fit on the VCC training split and applied frozen. No external calibration, label-derived scaling, screen-specific refit, per-screen head choice, or target-side tuning of any kind was performed.

G.2 Amendment chronology

Seventeen numbered amendments to the specification exist—sixteen as files, since A5 is a fileless editorial event—and we separate them by timing rather than describing the record with a single claim. A1 through A4 were authored and recorded between 2026-08-11 and 2026-08-13, each before the measurements it governs existed and before the first positively located external-outcome access; A1's substance is supported from 2026-08-11, while its current text reflects a recorded edit of 2026-08-13 that added a classification note without altering substance. They add pre-registered secondaries and sensitivities, and none adds, removes, or reseeds a control; they first entered version control later, on 2026-08-25. A5 and A6 were recorded after external unblinding and are editorial: A5 rewords a paragraph of the introduction and A6 replaces four occurrences of one phrase, and neither changes a hypothesis, a control, a seed, an endpoint, a statistic, or a reported value. A5 is a fileless governance event with no artifact; A6 first entered version control on 2026-08-26. A7 was also recorded after unblinding, entering version control on 2026-08-26, and is not editorial: it records the execution of the pre-registered non-overlap sensitivity of App. D, states that the analysis was specified before unblinding but run after the primary was persisted, and reports its outcome in full. A8 was recorded on 2026-08-27 and remained outside version control until the filing accompanying this revision; it is record-keeping, closing two documentation items the audit record had left open—the eligibility reconciliation behind the split counts of Sec. 3, together with the universe against which the non-overlap exclusion was defined, and an encoder-width inconsistency carried by an earlier draft and since corrected. A9 is a retrospective governance and reporting reconciliation recorded with this

revision. Neither changes a hypothesis, control, seed, endpoint, statistic, reported value, or conclusion. A10, A11 and A12 are later still, all recorded on 2026-08-29, well after external unblinding: they fix the frozen preprocessing state for post-hoc work and govern the two supporting diagnostics of App. E.2. Each was locked before its own measurement existed, but none belongs to the original pre-result specification, and none changes a hypothesis, control, seed, endpoint, or reported value. A13, recorded after external evaluation, reconciled the A3 RPE1 population description and restored previously omitted pre-specified secondary reporting without changing any frozen primary result. A14, also post-unblinding, records the final reporting-completion corrections identified by release audit, including full A1 secondary reporting and reconciliation of the A4 secondary descriptions. A15, also post-unblinding, is a wording reconciliation: it aligns chronology phrasing to the documented-access evidence bound and changes no value. A16, recorded on 2026-08-30, is also post-unblinding and repairs two release-level defects: it redirects the supplement's references to a non-shipped internal evidence record onto shipped artifacts that support the same statements, removing the one reference no shipped artifact supports, and it rescopes the H-FM2 chronology sentence from "fixed in advance of seeing it" to the evidence-bounded statement that the expected direction was recorded before the first documented external-outcome access. It performs no computation and changes no result, numerical value, experiment, model, head, metric, control, seed, split, endpoint, estimator, or figure, and strengthens no claim. A17, recorded on 2026-09-08, is post-unblinding and documentation-only: it records the author-and-advisor decision, which A4 deliberately left open, to present the endpoint/sampling-depth diagnosis as the second headline finding, and for presentation only supersedes the specification's original restriction against a second headline. The pre-registered primary inferential family remains exactly H-FM1–H-FM4, no secondary analysis is promoted, and no computation, reported value, hypothesis, control, seed, endpoint or inferential classification changes; the decision is not represented as pre-registered.

A gap in the amendment token taxonomy. The change-control clause enumerates three classification tokens but never defines them, and none of the three fits an additive pre-registered sensitivity. A2 and A3 therefore carry the token SENSITIVITY_ADDITION, which is not among the three, and A4 records the mismatch in place of a token. Each of the three was recorded before the measurements it governs. We report this as a formal gap in the taxonomy; it does not bear on what was run, when it was run, or any reported value.

G.3 Chronology evidence and its limits

The ordering of the specification, its amendments, model and head selection, and the external unblinding is reconstructed from records local to the authoring machine. In the surviving local evidence, amendments A1 through A4 precede the first positively located external-outcome access: the latest of the four is supported from 2026-08-13T05:37:46Z, and the first positively located external-outcome access lies between 2026-08-13T06:15:09.591Z and 2026-08-13T06:15:35.490Z, a margin of approximately 37 minutes 24 seconds. A1's substance is supported from 2026-08-11T19:09:21.979Z, while its current text reflects a recorded edit of 2026-08-13T00:59:34.920Z that added a classification note without altering substance.

The evidence classes are contemporaneous machine-written session records, filesystem creation metadata, and process-written run artifacts. These are distinct local evidence mechanisms that corroborate one another, but they are local to the authoring machine: they are not externally timestamped, not cryptographically signed, and not maintained in an append-only ledger, and they do not exclude an undocumented earlier access. Pinned SHA-256 digests cover the frozen foundation-model artifacts and the two harmonized external endpoint files; the specification, lock and amendment chain do not carry an equivalent externally anchored digest history.

Two gaps are disclosed rather than resolved. The session records provide substantial but incomplete coverage of the experimental window: a 28.04-hour interval spanning 2026-08-12 has no session-record entries, and the head-selection freeze falls inside it, resting instead on filesystem metadata and the process-written phase-1 execution log, whose internal write order places the freeze record before every test entry. Separately, no positive `EXTERNAL_Y_READ = YES` runtime marker exists anywhere in the record, so the unblinding moment is located by the surrounding artifacts rather than by the project's own declared marker.

Table 3: Strong classical baselines on the VCC held-out target-gene test set (mean ± std, three seeds; Ridge and ElasticNet are deterministic at the displayed precision). Removing row-level magnitude (DIR, $\mathbf{x}/\|\mathbf{x}\|_2$) sharply reduces performance, while concatenating $\mathbf{x}$ with the four magnitude statistics (X+M) enables tree ensembles to substantially outperform MAG-SPSM. Bold marks the best value within each feature column.

Method	$\mathbf{x}$ only	Direction only	Magnitude only	$\mathbf{x}$ + magnitude
Ridge	-0.306	-0.319	$\mathbf{+0.187}$	-0.107
ElasticNet	-0.247	-0.252	$+0.175$	-0.037
Random Forest	$+0.090 \pm 0.004$	-0.062 ± 0.003	-0.042 ± 0.005	$\mathbf{+0.370 \pm 0.004}$
HistGradientBoosting	$\mathbf{+0.190 \pm 0.006}$	$\mathbf{+0.062 \pm 0.008}$	$+0.054 \pm 0.047$	$+0.365 \pm 0.008$
k-NN (k=10)	-1.124	-2.256	-0.041	-1.000
SPSM (deep)	$+0.082 \pm 0.010$ (no magnitude features)			
MAG-SPSM (deep, ours)	$+0.225 \pm 0.004$ (deep proof-of-concept; uses $[\mathbf{x}; \widetilde{\boldsymbol{m}}]$)			
MAG-LINEAR	$+0.208$ (four-feature linear regression on $\widetilde{\boldsymbol{m}}$)			

The counterfactual exclusion counts of App. D were reconstructed under a frozen metadata-only procedure from the pre-specified metadata and set definitions, after first reproducing the historical 43 and 33 construct overlap counts. The matching rule was recovered from the executed analysis code rather than chosen, the procedure was fixed before execution and run once, and no model was run and no external outcome column was accessed. The frozen script has SHA-256 `9ed9deb1bd9123af4b136fd658c737f110f50102b9cca165e1e1ead548ee1ca9`. This is a provenance reconstruction, not a proof and not a scientific validation.

G.4 PROVENANCE LIMITS

The frozen representation. The persisted result records do not carry a content digest binding a specific stored copy of $\mathbf{z}$ to the reported external correlations; the link between the feature arrays and the persisted predictions is recorded in prose rather than in a manifest. What *is* pinned is the complete generating specification— checkpoint, repository commit, tokenizer and vocabulary digests, layer, pooling, aggregation, and scaling—together with the observed determinism of the embedding step. That specification supports deterministic regeneration under the recorded pipeline, but we cannot demonstrate that a regenerated array is bitwise identical to the exact feature array behind the reported correlations. We state this rather than implying an archival guarantee we cannot support.

Results carried over from the earlier submission. As disclosed in App. X, the classical baseline tables from the earlier submission were not re-derived under the locked revision harness and the artifacts required to reproduce them were not retained. No result in this paper depends on them; where a legacy value is cited, it is as historical context for a limitation rather than as evidence. The deep SPSM and MAG-SPSM values reported here are the persisted lock-era values, which re-derive from retained per-row prediction artifacts. A later revision reproduction gave a larger seed dispersion; that reproduction survives only through summary-level artifacts and cannot be independently re-derived from equivalently retained per-row prediction artifacts currently preserved, no amendment authorised substituting it, and the cause of the difference remains unresolved—nondeterministic GPU or environment behaviour is a hypothesis only.

Code and artifact release. The specification, its amendment chain, the environment record, and the artifact digests accompany this submission as supplementary material. Retained analysis code will be released upon acceptance; some historical execution paths behind the persisted external-evaluation artifacts are not fully retained, so we do not claim that every reported artifact can be regenerated from released code. The persisted artifacts and their digests are unaffected by this limitation.

H PRE-SPECIFIED HYPOTHESES FOR THE STRONG-BASELINE AUDIT

We test whether the magnitude effect is specific to the deep SPSM architecture or a more general property of held-out target-gene prediction. We evaluated five classical methods (Ridge, ElasticNet,

Random Forest, HistGradientBoosting, k-NN) on four feature sets: the full $\mathbf{x}$, the row-normalized direction $\mathbf{x}/\|\mathbf{x}\|_2$ (DIRECTION-ONLY), $\widetilde{\boldsymbol{m}}$ alone (MAG-ONLY), and the concatenation $[\mathbf{x};\, \widetilde{\boldsymbol{m}}]$ (X+MAG). All preprocessing used training-split statistics; the four magnitude features are pre-specified before any model training. Five hypotheses were pre-specified before the experiment was run; we report all five verdicts below.

The result (Table 3) reframes the interpretation of the deep model. *(i) Four magnitude scalars match the full input.* Ridge on MAG-ONLY reaches $+0.187$, only 0.003 below the strongest classical model on $\mathbf{x}$ (HistGradientBoosting, $+0.190 \pm 0.006$); MAG-LINEAR reaches $+0.208$. Four scalars carry approximately as much held-out predictive signal as the full $2{,}000$-dimensional vector under this split. *(ii) Removing magnitude collapses performance.* Row-normalizing $\mathbf{x}$ reduces the strongest classical model from $+0.190$ to $+0.062 \pm 0.008$. All linear methods become strongly negative on DIRECTION-ONLY. *(iii) Combining the two outperforms the deep encoder.* Random Forest on $[\mathbf{x};\, \widetilde{\boldsymbol{m}}]$ reaches $+0.370 \pm 0.004$ and HistGradientBoosting reaches $+0.365 \pm 0.008$, both substantially exceeding our deep MAG-SPSM encoder ($+0.225 \pm 0.004$). Direction combined with explicit magnitude adds $\sim 0.18\, R^2$ that tree ensembles, but not our deep encoder, exploit. That a low-dimensional response-magnitude summary is recoverable by classical models here is therefore not specific to a single MLP-encoder configuration; whether attention-based, set-equivariant, or foundation-model encoders behave similarly is the question the locked foundation-model evaluation takes up. The Random Forest result on $[\mathbf{x};\, \widetilde{\boldsymbol{m}}]$ is robust across reasonable RF hyperparameter choices (App. T).

The five hypotheses below were written in the experiment report before the corresponding scripts were run; we have no formal external preregistration, but each hypothesis is dated in the project log preceding the run. We report each verdict honestly.

- H1 Classical models on $\mathbf{x}$ alone will not clearly beat MAG-LINEAR. *Verdict: supported.* Best $\mathbf{x}$-only model (HistGradientBoosting, $R^2 = +0.190$) is below MAG-LINEAR ($+0.208$) and within 0.003 of Ridge on MAG-ONLY ($+0.187$); no $\mathbf{x}$-only classical model exceeds MAG-LINEAR by a margin larger than seed noise.
- H2 DIRECTION-ONLY will perform worse than $\mathbf{x}$-only and MAG-ONLY. *Verdict: supported.* Best DIRECTION-ONLY ($+0.062$) is well below both anchors.
- H3 Linear methods on MAG-ONLY will approximately reproduce MAG-LINEAR. *Verdict: supported.* Ridge on MAG-ONLY reaches $+0.187$ ($\geq +0.180$ pre-specified threshold).
- H4 Tree ensembles on X+MAG may match or beat MAG-SPSM. *Verdict: clearly beats.* A row-level paired bootstrap gives Random Forest on $[\mathbf{x};\, \widetilde{\boldsymbol{m}}]$ over MAG-SPSM a margin of $\Delta R^2 = +0.145$ (95% CI $[+0.127, +0.163]$, sign-stable in 100% of resamples).
- H5 If a non-deep model on X+MAG clearly beats MAG-SPSM, the paper should be reframed as diagnostic discovery rather than method paper. *Verdict: triggered.* The reframing applied to this paper is the consequence of H5.

I PER-ROW ALIGNMENT: FULL POPULATION-LEVEL ARGUMENT

In the population view, let $M^\star$ be an auxiliary random vector with the same marginal distribution as $\widetilde{\boldsymbol{m}}(X)$ but independent of (X, Y). Then the Bayes predictor with access to $(X, M^\star)$ coincides with the Bayes predictor with access to X alone:

$$\mathbb{E}[Y \mid X, M^\star] \;=\; \mathbb{E}[Y \mid X].$$

By the optimality of conditional expectation under squared loss, no measurable predictor using $(X, M^\star)$ can improve over the Bayes risk achieved by $\mathbb{E}[Y \mid X]$ in squared loss. This yields two consequences we test directly: (a) replacing $\widetilde{\boldsymbol{m}}(\mathbf{x}_i)$ by four i.i.d. $\mathcal{N}(0, I_4)$ features makes the augmenting block independent of $(\mathbf{x}_i, Y_i)$ by construction; (b) under a row-shuffled augmentation $\mathbf{x}_i^\pi = [\,\mathbf{x}_i;\, \widetilde{\boldsymbol{m}}(\mathbf{x}_{\pi(i)})\,]$ with π a uniformly random permutation independent of the data, the augmenting block preserves the marginal distribution of $\widetilde{\boldsymbol{m}}$ but no longer carries row-specific information about Y_i. The shuffled-magnitude and random-feature controls of Sec. 5.3 are finite-sample diagnostics of this principle: they preserve dimensionality and marginal scale while breaking row-specific alignment with the label. A systematic R^2 gain over both controls therefore supports attributing the contribution of the four scalars to their per-row dependence on $\mathbf{x}_i$ rather than to added input dimensionality, scale, or noise diversity. We do not promote this observation to a formal theorem;

Table 4: Per-row alignment controls (seed 0). Both shuffled-magnitude (per-row alignment broken) and four random Gaussian features (independent of input by construction) return performance to baseline within seed noise, consistent with the population-level argument of this appendix.

Arm	Test R^2	Spearman	Q1 bias	Q5 bias
BASELINE-MSE (no $\widetilde{\boldsymbol{m}}$)	$+0.0768$	$+0.399$	$+1.061$	-0.735
MAG-SPSM (4 magnitude features)	$\mathbf{+0.2224}$	$\mathbf{+0.576}$	$+0.918$	-0.580
SHUFFLED-MAGNITUDE (per-row π random)	$+0.0712$	$+0.402$	$+1.191$	-0.592
RANDOM-FEATURE (four $\mathcal{N}(0,1)$)	$+0.0869$	$+0.377$	$+1.144$	-0.660

Table 5: VCC in-distribution held-out target-gene results, reproduced under the locked revision harness (mean $\pm$ std across seeds, $\mathrm{ddof}{=}1$, where the model is stochastic). MAG-Anchor was not reproduced under the locked revision harness and is described qualitatively (App. J); its as-submitted performance figures are not carried into this revision.

Model	VCC R^2
SPSM baseline	$+0.082 \pm 0.010$
MAG-SPSM	$+0.225 \pm 0.004$
MAG-LINEAR	$+0.208$
RF (x + mag)	$+0.370 \pm 0.004$

its role is to motivate the two pre-specified experimental controls (row-shuffling and i.i.d. Gaussian features) reported in Sec. 5.3.

Per-row alignment: empirical controls (Table 4). We instantiated three additional arms at seed 0 (otherwise identical training): (i) BASELINE-MSE: SPSM V4 with no magnitude features ($\mathbf{x}^+ = \mathbf{x}$); (ii) SHUFFLED-MAGNITUDE: MAG-SPSM but with per-row correspondence between $\mathbf{x}_i$ and $\widetilde{\boldsymbol{m}}$ broken by a uniformly random permutation π; (iii) RANDOM-FEATURE: MAG-SPSM but with the magnitude block replaced by four i.i.d. $\mathcal{N}(0,1)$ features. The result (Table 4) is unambiguous: both controls return performance to baseline within seed noise. Shuffling removes the gain entirely ($\Delta = -0.151$ vs. MAG-SPSM); four random Gaussian features fail to recover any gain. Under shuffling, $\widetilde{\boldsymbol{m}}(\mathbf{x}_{\pi(i)}) \perp\!\!\!\perp (\mathbf{x}_i, y_i)$ in the sample limit, so the augmenting block carries no information about y_i.

J MAG-ANCHOR: FULL ARCHITECTURE, HYPERPARAMETERS, AND CONTROLS

MAG-Anchor is a secondary proof-of-concept deep method, not a state-of-the-art in-distribution predictor, and was not reproduced under the locked revision harness; its as-submitted performance figures are not carried into this revision. Its architecture and the pre-specified per-row alignment controls (eval-time row shuffling, train-time row shuffling, and random-Gaussian replacement of $\widetilde{\boldsymbol{m}}$, each of which collapses the gain, so the improvement depends on correctly aligned per-row magnitude features rather than on added capacity) are given below. Table 5 lists the in-distribution results that were reproduced under the locked harness, with MAG-Anchor named but its as-submitted figures withheld.

Architecture. MAG-Anchor decomposes the predicted perturbation effect into a direct magnitude anchor and a gated residual correction:

$$\hat{y} = g_\phi(\widetilde{\boldsymbol{m}}(\mathbf{x})) + \alpha\, h_\theta([\mathbf{x}; \widetilde{\boldsymbol{m}}(\mathbf{x})])\,, \tag{4}$$

where $\widetilde{\boldsymbol{m}}(\mathbf{x}) \in \mathbb{R}^4$ is the standardized magnitude vector defined in Eq. 1. The anchor branch $g_\phi(\widetilde{\boldsymbol{m}}) = \mathbf{w}^\top \widetilde{\boldsymbol{m}} + b$ is a single linear layer with $\mathbf{w} \in \mathbb{R}^4$, $b \in \mathbb{R}$; its weights are warm-started from a sklearn `LinearRegression` fitted on $(\widetilde{\boldsymbol{m}}, y)$ and remain trainable thereafter (no freezing). The residual branch h_θ is a small MLP with layers $d_{\text{in}} \to 256 \to 128 \to 1$, ReLU activations, and architectural dropout $p{=}0.1$ after each ReLU; no LayerNorm or BatchNorm is used. This

dropout setting applies only to the MAG-Anchor residual branch; the SPSM/MAG-SPSM backbone (App. L) uses dropout p=0. For the headline MAGAnchor-xmag the residual input is $[\mathbf{x};\, \widetilde{\boldsymbol{m}}] \in \mathbb{R}^{2004}$ ($d_{\text{in}} = 2004$); for the MAGAnchor-x ablation the residual input is $\mathbf{x}$ alone ($d_{\text{in}} = 2000$). The scalar gate α is parameterized as $\alpha = \text{softplus}(\rho)$ with $\rho_{\text{init}} = \ln(\mathrm{e}^{0.1} - 1)$, so $\alpha_{\text{init}} = 0.1$ exactly and $\alpha \in (0, \infty)$ throughout training. This design preserves a direct pathway from magnitude features to the output while allowing the residual branch to learn corrections from both the expression vector and the magnitude summary.

Training hyperparameters. Unless otherwise stated, MAG-Anchor uses a trainable anchor, no residual penalty, Adam with learning rate 10^{-3} and weight decay 10^{-5}, batch size 64, 25 training epochs, and checkpoint selection by validation R^2. Frozen-anchor and residual-penalty variants were evaluated only as sensitivity checks; they were not reproduced under the locked revision harness and no value from them is reported.

Per-row alignment is load-bearing (controls). We ran three pre-specified controls on the MAG-Anchor recipe to test whether its gain depends on per-row magnitude alignment or merely on added input dimensionality: permuting $\widetilde{\boldsymbol{m}}(\mathbf{x})$ across rows at evaluation time, permuting it across training rows, and replacing it with random Gaussian features. Each control collapses VCC performance and removes the positive Replogle association, so the gain depends on correctly aligned per-row magnitude features and is not a generic capacity effect.

K CLASSICAL-BASELINE HYPERPARAMETERS

- Ridge: `Ridge(alpha=1.0)`.
- ElasticNet: `ElasticNet(alpha=0.001, l1_ratio=0.5, max_iter=10000)`.
- Random Forest: `RandomForestRegressor(n_estimators=200, max_depth=20, min_samples_leaf=2, n_jobs=8, random_state=seed)`.
- HistGradientBoosting (XGBoost fallback): default sklearn `HistGradientBoostingRegressor` with `max_iter=500, learning_rate=0.05, max_depth=6, random_state=seed`, early stopping on validation with patience 20.
- k-NN: `KNeighborsRegressor(n_neighbors=10, weights="distance")`.

For linear models and k-NN, features are standardized using training-split mean/std exclusively; for tree models, raw features are used (with the four magnitude features z-scored regardless of model).

L SPSM V4 BACKBONE AND LOSS FORMULAS

The SPSM V4 backbone is a three-layer MLP encoder with hidden width 256, ReLU activations, and LayerNorm; architectural dropout is $p = 0$ (no active dropout regularization). The encoder produces a 256-dimensional representation $\mathbf{z}_i = f_\theta(\mathbf{x}_i^+)$. The total objective combines a prediction loss with two SPSM regularizers, $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{MSE}} + \lambda_{\text{stab}}\,\mathcal{L}_{\text{stab}} + \lambda_{\text{rel}}\,\mathcal{L}_{\text{rel}}$, with $\lambda_{\text{stab}} = \lambda_{\text{rel}} = 0.5$, both inherited from the SPSM V4 backbone unchanged.

Stability loss. The stability term is a Gaussian-kernel maximum-mean-discrepancy (MMD) penalty on encoder features across the experimental Flex batches that co-occur within the same SGD mini-batch; no stochastic input perturbation, dropout-perturbed forward pass, or input noising is used. Concretely, for an SGD minibatch $\mathcal{I}$ of size $B = 64$, let $\mathcal{B}(\mathcal{I}) = \{b_i : i \in \mathcal{I}\}$ denote the set of experimental Flex batches represented in $\mathcal{I}$, and let $Z_b(\mathcal{I}) = \{\mathbf{z}_i : i \in \mathcal{I},\, b_i = b\}$ collect the encoder outputs from rows of batch b. The stability loss averages the squared MMD across all unordered pairs of co-represented batches:

$$\mathcal{L}_{\text{stab}} \;=\; \frac{2}{|\mathcal{B}(\mathcal{I})|\,(|\mathcal{B}(\mathcal{I})|-1)} \sum_{\substack{b,b' \in \mathcal{B}(\mathcal{I}) \\ b<b'}} \text{MMD}^2_\kappa\big(Z_b(\mathcal{I}),\, Z_{b'}(\mathcal{I})\big)\,,$$

where κ is a Gaussian kernel whose bandwidth is set by the median heuristic on the in-minibatch pairwise distances, computed with gradients disabled. For each SGD minibatch $\mathcal{I}$, $\mathcal{B}(\mathcal{I})$ denotes the experimental batches represented in that minibatch. The sampler constructs context-stratified minibatches so that each represented batch contributes enough rows for the MMD pairs in the stability loss below to be well-defined.

Relation loss. The relation loss is a soft-Spearman ranking penalty applied to the scalar predictions $\hat{y}$ across pairs of experimental batches that co-occur in the same minibatch; it does not use a learnable projection on $\mathbf{z}$ and does not enter the labels y_i (labels enter only through $\mathcal{L}_{\mathrm{MSE}}$). For an SGD minibatch $\mathcal{I}$, we construct a small set $\mathcal{P}$ of matched context-pairs by iterating over unordered pairs of distinct Flex batches $(b, b') \in \mathcal{B}(\mathcal{I}) \times \mathcal{B}(\mathcal{I})$ and, for each pair, taking the first applicable of three tiers: tier 1 (weight $w = 1.00$) matches rows with the same target-gene identifier shared between b and b'; tier 2 (weight $w = 0.50$) matches rows with the same auxiliary cluster index, paired greedily; tier 3 (weight $w = 0.25$) takes k=1 nearest neighbors in an auxiliary u-space. On VCC, only tier 1 contributes informative matches, so in practice $\mathcal{P}$ contains pairs of Flex batches with at least one shared target gene, aligned by target gene. Each matched pair $p \in \mathcal{P}$ yields two aligned prediction vectors $\hat{\boldsymbol{y}}_a^{(p)}, \hat{\boldsymbol{y}}_b^{(p)} \in \mathbb{R}^{m_p}$ and a weight w_p. With soft ranks $\widetilde{r}_k(\boldsymbol{v}; \tau) = \sum_{\ell \neq k} \mathrm{sigmoid}\big((v_k - v_\ell)/\tau\big)$, the soft-Spearman correlation

$$\widetilde{\rho}(\boldsymbol{a}, \boldsymbol{b}; \tau) \;=\; \frac{\sum_k (\widetilde{r}_k(\boldsymbol{a}) - \bar{\widetilde{r}}(\boldsymbol{a}))(\widetilde{r}_k(\boldsymbol{b}) - \bar{\widetilde{r}}(\boldsymbol{b}))}{\sqrt{\sum_k (\widetilde{r}_k(\boldsymbol{a}) - \bar{\widetilde{r}}(\boldsymbol{a}))^2 \, \sum_k (\widetilde{r}_k(\boldsymbol{b}) - \bar{\widetilde{r}}(\boldsymbol{b}))^2}},$$

the relation loss is

$$\mathcal{L}_{\mathrm{rel}} \;=\; \frac{\sum_{p \in \mathcal{P}} w_p \left(1 - \widetilde{\rho}(\hat{\boldsymbol{y}}_a^{(p)}, \hat{\boldsymbol{y}}_b^{(p)}; \tau)\right)}{\sum_{p \in \mathcal{P}} w_p}, \qquad \tau = 0.1.$$

$\mathcal{L}_{\mathrm{rel}}$ thus drives the model to produce a consistent rank order of predictions across pairs of experimental batches in the same minibatch.

Optimization. Adam at lr 10^{-3}, weight decay 10^{-5}, batch size 64, 25 epochs; HVG selection on the training split only; magnitude z-score statistics from the training split only; model selection by validation R^2.

Architecture overview (visual). Figure 4 summarizes the MAG-SPSM pipeline. The four magnitude statistics $\boldsymbol{m}(\mathbf{x})$ defined in Eq. (1) are z-scored using VCC training statistics to obtain $\widetilde{\boldsymbol{m}}(\mathbf{x}) \in \mathbb{R}^4$, then concatenated with the d=2,000-dimensional log1p delta input $\mathbf{x}$ to form $\mathbf{x}^+ \in \mathbb{R}^{2004}$, which is fed to the SPSM V4 backbone and a linear head. The same backbone, the same losses, and the same optimizer are used for the SPSM baseline (which receives only $\mathbf{x}$) and for MAG-SPSM (which receives $\mathbf{x}^+$); the only difference is whether the four magnitude features are exposed to the encoder.

M MAGNITUDE STATISTICS ON REPLOGLE

For the primary 2,000-HVG transfer evaluation, each Replogle row $\mathbf{x}^{(R)}$ is first embedded into the VCC HVG namespace by zero-imputing the VCC-only coordinates missing from that screen: 570 for K562-essential and 450 for RPE1 (Sec. 3). The magnitude statistics $\boldsymbol{m}(\mathbf{x}^{(R)})$ are then computed using Eq. (1) over this 2,000-dimensional vector, including the zero-imputed entries. In the HVG-coverage controls (App. X.1), by contrast, models are trained and evaluated directly in the 1,430-dimensional subspace shared between VCC and Replogle K562. We then z-score using *VCC training* statistics $\boldsymbol{\mu}_{\mathrm{tr}}, \boldsymbol{\sigma}_{\mathrm{tr}}$ (Sec. 5.1), preserving the strict zero-shot constraint. The Replogle z-scored magnitudes have empirical means roughly $+0.8\,\sigma_{\mathrm{tr}}$ to $+1.9\,\sigma_{\mathrm{tr}}$ relative to the VCC training baseline, and component-wise standard deviations $1.6\times$ to $2.4\times$ wider, reflecting the larger feature scale of the Replogle release. The deep MAG-SPSM sees these as out-of-distribution; the linear baseline, having no encoder, does not suffer the same OOD penalty.

Optimization, training schedule, and compute. All deep models (SPSM, MAG-SPSM, MAG-Anchor) use Adam at learning rate 10^{-3} with weight decay 10^{-5}, batch size 64, and 25 training

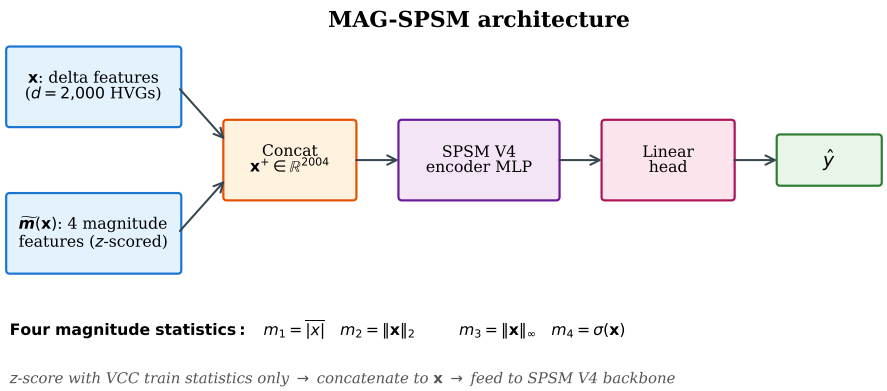


Figure 4: MAG-SPSM architecture. The expression delta vector $\mathbf{x}$ (d=2,000 HVGs) is concatenated with the four z-scored magnitude features $\widetilde{\boldsymbol{m}}(\mathbf{x}) = (\overline{|x|},\ \|\mathbf{x}\|_2,\ \|\mathbf{x}\|_\infty,\ \sigma(\mathbf{x}))$ to form $\mathbf{x}^+ \in \mathbb{R}^{2004}$, which is passed through the SPSM V4 encoder MLP (three hidden layers of width 256, ReLU, LayerNorm, no architectural dropout) and a linear head. All z-score statistics use VCC training data exclusively.

Table 6: Magnitude feature ablation on VCC test (seed 0). Δ vs full = R^2 minus full 4-feature MAG-SPSM. Bold marks the full four-feature model and the best leave-one-out variant. Differences are computed from unrounded values and may therefore differ in the last digit from the subtraction of the rounded entries shown.

Variant	Test R^2	Δ vs base	Δ vs full	Spearman
SPSM baseline (no MAG)	+0.0768	+0.0000	−0.1456	+0.3986
MAG-SPSM (full 4)	+0.2224	+0.1456	+0.0000	+0.5763
only m_1 (mean $\lvert x\rvert$)	+0.1278	+0.0510	−0.0946	+0.4987
only m_2 (ℓ_2)	+0.0950	+0.0181	−0.1274	+0.4557
only m_3 (max $\lvert x\rvert$)	+0.1269	+0.0501	−0.0955	+0.4406
only m_4 (std)	+0.1001	+0.0233	−0.1223	+0.4531
drop m_1	+0.1986	+0.1218	−0.0238	+0.5730
drop m_2	+0.2303	+0.1535	+0.0079	+0.5884
drop m_3	+0.1365	+0.0597	−0.0859	+0.5084
drop m_4	+0.2084	+0.1316	−0.0140	+0.5784

epochs. The loss (2) applies to SPSM/MAG-SPSM; MAG-Anchor uses MSE on $\hat{y}$. HVG selection and all z-score statistics use the VCC training split exclusively. All deep experiments fit on a single TITAN RTX 24 GB GPU; classical baselines were run on CPU. Classical baselines use `scikit-learn` defaults (detailed in App. K).

N Additional results: ablations, rank extension, and negative results

Single-feature and leave-one-out ablation of MAG-SPSM. We performed a complete single-feature and leave-one-out (LOO) ablation of the four magnitude features at seed 0 (Table 6). Across the four single-feature variants, the VCC test R^2 ranges from $+0.095$ (single $\|\mathbf{x}\|_2$) to $+0.128$ (single $\overline{|\mathbf{x}|}$), all well below the full 4-feature MAG-SPSM at $+0.222$. Across the four leave-one-out variants, the test R^2 ranges from $+0.137$ (drop $\max|\mathbf{x}|$) to $+0.230$ (drop $\|\mathbf{x}\|_2$); the latter is the best LOO arm and is within 0.008 of the full 4-feature model, while removing $\max|\mathbf{x}|$ costs 0.086, indicating that $\max|\mathbf{x}|$ contributes the most among the four magnitude features.

MAG-RANK: a small rank-margin extension of MAG-SPSM. MAG-RANK adds a margin-based pairwise rank loss $\mathcal{L}_{\text{rank}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \max(0,\ \alpha - \text{sign}(y_i - y_j)\cdot(\hat{y}_i - \hat{y}_j))$ over within-batch close-call pairs, with $\alpha = 0.1$, in-batch close-call pairs dropped at $\varepsilon_y = 0.05$, and weight $\lambda_{\text{rank}} = 0.1$. The 3-seed test R^2 rises to $+0.2396 \pm 0.0078$, a paired improvement over MAG-SPSM of $+0.0142 \pm 0.0040$. Spearman is statistically unchanged; we report MAG-RANK as a small reproducible extension and do not promote it as a contribution. The strong-baseline result of App. H dominates this gain by an order of magnitude.

Per-quintile bias. Relative to SPSM, MAG-SPSM reduces the signed tail biases at seed 0 from $+1.061$ to $+0.918$ in Q1 and from -0.735 to -0.580 in Q5 (Table 4), consistent with the diagnostic that magnitude features expand the prediction range, but substantial residual calibration bias remains. The residual per-quintile bias is not closed by MAG-SPSM, motivating the strong-baseline audit (App. H) and the cross-dataset experiment (Sec. 6.2).

The single-seed feature ablation is summarized in Fig. 5; it is diagnostic rather than a primary multi-seed result.

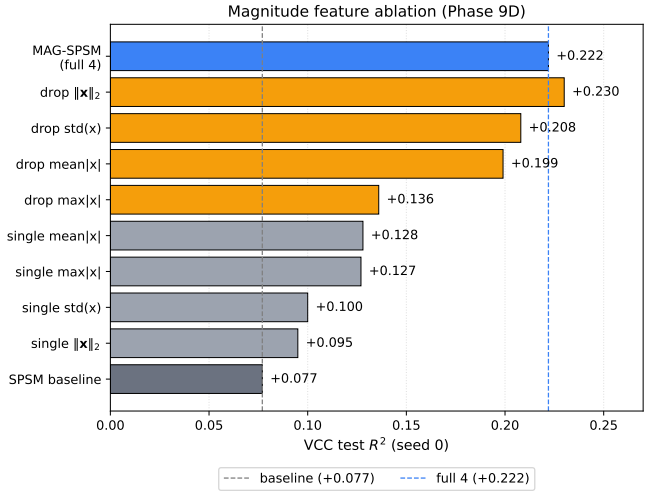


Figure 5: Single-feature and leave-one-out ablations of the four magnitude features. The four form a partially redundant low-dimensional summary: dropping $\|\mathbf{x}\|_2$ leaves performance essentially unchanged, whereas dropping $\max_j |x_j|$ costs the most.

Dataset-level split and scale summaries are provided in Table 7.

O ENDPOINT PROVENANCE FOR THE DISTRIBUTED REPLOGLE COLUMN

Reconstruction attempt. We tested whether the Anderson–Darling column distributed with the Replogle releases can be reproduced from the corresponding single-cell data under the definition of Sec. 3. Using RPE1, where both the distributed column and single-cell counts are available, we selected 30 constructs by a deterministic rule—sorted by the stored value, binned into ten equal-frequency strata, three per stratum—and computed the two-sample Anderson–Darling statistic at each target gene's own column under eight pre-declared variants: aggregation (pooled across batches; per-batch then unweighted mean, count-weighted mean, or median), transform (raw statistic; $\log 1p(\max(0, \cdot))$), input matrix (raw counts; normalized expression), and control sample (non-targeting cells pooled globally; restricted to the same batch).

No variant reproduced the distributed values. The closest was per-batch median on normalized expression with globally pooled controls, reaching Pearson 0.335 and Spearman 0.268 with normalized RMSE 0.30. We classify the outcome as divergent: the distributed column is not a target-gene Anderson–Darling effect size under any tested construction.

Table 7: Dataset statistics for the raw releases and the earlier distributed-endpoint analysis. Both screens use CRISPRi perturbation and differ in cell line, biological-row unit, and feature normalization. **These are not the eligible construct sets used by the locked harmonized transfer evaluation**, which retains 1,932 RPE1 and 1,903 K562-essential constructs (Sec. 6.2), and the gene-overlap row below counts a different quantity from the training overlap reported there.

	VCC	Replogle K562 (essential)
Source	Virtual Cell Challenge	Replogle et al. 2022 (Cell)
Cell type	H1 hESC (stem)	K562 (CML)
Perturbation	CRISPRi (dCas9-KRAB)	CRISPRi (dCas9-KRAB)
Train / Val / Test rows	5,740 / 2,052 / 3,972	— (held-out test only)
Non-NT rows used (test)	3,972	2,176
Unique target genes	282 (all splits)	2,057
Train/Val/Test target-gene split	140 / 49 / 93 (*disjoint*)	—
Feature dim (d_x used)	2,000 (HVG)	8,563 full; 1,430 in HVG
Feature normalization	log1p Δ vs per-batch NT	pre-z-scored log1p change
Feature std (post-HVG)	0.085	0.161 ($\sim 1.9\times$)
Label	log1p(AD) per (batch, gene)	log1p(AD) per (pert, prom)
Label range	[0, 5.72]	[0, 8.71]
Label mean (std)	4.23 (0.81)	3.99 (2.75)
Per-row biological unit	(batch $\times$ gene) pseudobulk	(pert, prom) batch-aggregated
VCC target genes (all splits) also assayed	—	68 / 282 (24.1%)

Evidence for a transcriptome-wide interpretation. The distributed values are integer-valued and bounded well below the gene count in every screen ($[0, 6036]$, $[0, 5547]$, and $[0, 5528]$ for K562-essential, RPE1, and K562 genome-wide), which is consistent with a per-perturbation count over genes. We therefore computed transcriptome-wide per-gene Anderson–Darling statistics for ten deterministically selected constructs and compared eight pre-declared count definitions against the stored values: counts of genes with statistic above zero, above 1.0, or above 2.0; counts with $p < 0.05$ or $p < 0.01$, uncorrected or after Benjamini–Hochberg correction; and a detection-count null.

All count candidates were rank-monotone with the stored column ($\rho = 0.96$–0.98). The Benjamini–Hochberg-corrected count at $q < 0.05$ on normalized expression reached Pearson 0.986; the detection-count null was rejected, being constant at 8,749. The fitted slope was approximately 0.60: our reconstruction runs roughly $1.6\times$ the stored value. The exact upstream preprocessing, significance threshold, control set, and tested gene universe therefore remain unresolved, and we do not claim to have reconstructed the column. We note that `scipy.stats.anderson_ksamp` floors its reported significance level at 0.001 and caps it at 0.25, which makes any threshold-based count sensitive to implementation detail.

Association with the target-gene endpoint. Within each screen, the distributed column and the rebuilt target-gene endpoint are only weakly associated at construct level: Spearman 0.165 $[0.118, 0.211]$ in RPE1, 0.301 $[0.257, 0.343]$ in K562-essential, and 0.342 $[0.322, 0.362]$ at target level in the K562 genome-wide screen. The two endpoints capture largely distinct aspects of perturbation effect.

P HARMONIZED ENDPOINT CONSTRUCTION

Protocol. The subsample sizes are the medians of the VCC distribution: 27 perturbed cells per row and 788 non-targeting cells per batch. For each eligible construct we draw 27 perturbed cells without replacement from its pooled cells and 788 non-targeting cells without replacement from the pooled control set, compute the log1p Anderson–Darling statistic at the target gene's column, and repeat 25 times with pre-specified seeds. The endpoint is the median across draws; we also record the mean, standard deviation, interquartile range, and 5th and 95th percentiles. Quantiles use `numpy.quantile` with linear interpolation throughout.

Eligibility requires at least $\max(20, 27)$ perturbed cells, at least $\max(200, 788)$ pooled control cells, an unambiguous mapping from target gene to exactly one feature column, and a verified one-to-one join between the bulk and single-cell artifacts.

Coverage. Eligibility retains 1,932 constructs for RPE1 and 1,903 for K562-essential. Exclusions are dominated by target genes absent from the feature namespace rather than by insufficient cells. We report retained counts rather than a retention fraction: the pre-eligibility denominator is the output of an internal construct-selection step and is not recoverable from a governed artifact, so a percentage would carry a denominator we cannot source.

Resampling spread. The ratio of the median within-construct interquartile range to the between-construct standard deviation is 0.174 at construct level and 0.184 at target level for RPE1, and 0.236 and 0.258 for K562-essential. Within-construct resampling variation is therefore small relative to between-construct variation in both screens.

Pooled-scale comparison. A simple pooled construction using all eligible cells, without subsampling, correlates with the matched-size endpoint at Spearman 0.884 $[0.871, 0.895]$ in RPE1 and 0.874 $[0.860, 0.886]$ in K562-essential. Rank agreement is strong but not exact: matching the subsample size changes some rankings, which is why we fix it rather than pooling.

Control-matching sensitivity. Restricting control cells to the same batch as the perturbed cells, rather than pooling globally, is possible for 43 constructs in RPE1 and 194 in K562-essential under a minimum of ten perturbed and fifty control cells per batch. Where computable, the batch-matched endpoint correlates with the pooled-control endpoint at Spearman 0.910 in RPE1 and 0.973 in K562-essential. The two constructions therefore show high rank agreement in the subsets where both are computable; we do not extrapolate that to an equivalence across the full construct sets. We note that the batch variable available in these releases is a technical grouping whose status as an independent experimental replicate is not established by the accompanying metadata.

Q MAG-ANCHOR AND EXTERNAL TRANSFER

MAG-Anchor is described qualitatively in App. J, with no value carried into this revision, and it does not appear in Table 10. The per-construct prediction vectors required to evaluate it against a rebuilt endpoint were not retained for the external screens; only aggregate metrics were. Since our transfer evaluation re-scores stored predictions rather than retraining, and no stored MAG-Anchor prediction exists to validate a reconstruction against, we exclude it rather than introduce an unvalidated refit into the comparison. This is an availability constraint, not a selection based on outcome.

R BOOTSTRAP INTERVALS FOR THE EARLIER CLASSICAL COMPARISONS

The intervals in this section belong to the classical-baseline analysis of the earlier version of this work and are distinct from the resampling procedures that govern the results of Sec. 6.1 and Sec. 6.2, which are specified in App. F. We computed row-level paired bootstrap 95% confidence intervals ($B = 10{,}000$ resamples, RNG seed 42, percentile method) for three pre-specified comparisons. The summary is in Table 8. (C5) Random Forest on $[\mathbf{x}; \widetilde{\boldsymbol{m}}]$ versus MAG-SPSM on VCC test pooled R^2, paired across the same three seeds: median $\Delta R^2 = +0.145$, 95% CI $[+0.127, +0.163]$, sign-stable in 100% of resamples (STRONG). (C7) MAG-LINEAR versus HistGradientBoosting on $\mathbf{x}$-only on VCC test R^2 against the mean of three HistGB seeds: median $\Delta R^2 = +0.012$, 95% CI $[-0.009, +0.032]$ (WEAK). (C8) MAG-LINEAR versus the SPSM baseline on K562 Spearman ρ, row-resampled with no seed component: median $\Delta\rho = +0.750$, 95% CI $[+0.700, +0.800]$ (STRONG).

Table 8: Row-level paired bootstrap 95% confidence intervals for three pre-specified comparisons. $B = 10{,}000$ resamples, RNG seed 42, percentile method. Sign-stability is the fraction of bootstrap resamples in which the sign of the paired difference matches the median.

ID	Comparison	Metric	Median Δ	95% CI
C5	RF $[\mathbf{x}; \widetilde{\boldsymbol{m}}]$ vs. MAG-SPSM (VCC)	R^2	$+0.145$	$[+0.127, +0.163]$
C7	MAG-LINEAR vs. HistGB $\mathbf{x}$-only (VCC)	R^2	$+0.012$	$[-0.009, +0.032]$
C8	MAG-LINEAR vs. SPSM baseline (K562)	ρ	$+0.750$	$[+0.700, +0.800]$

We do not report per-row bootstrap intervals for comparisons involving MAG-Anchor or for the asymmetric three-seed-vs-one-seed RF–MAG-SPSM K562 comparison. A seed-0 calibration check on regenerated deep-model predictions did not reproduce the original predictions to the pre-specified tolerance; we therefore avoid uncertainty estimates that would mix the original paper predictions with regenerated deep-model outputs. This analysis is reported as supporting uncertainty evidence for the main comparisons, not as a new contribution.

S ARCHITECTURE CONTROLS FOR $\mathbf{x}$-ONLY MLP

To test whether the $\mathbf{x}$-only MLP failure to recover the magnitude-like signal is explained by normalization, dropout, capacity, or the absence of a simple scale-preserving path, we ran an architecture-control audit. We held the data, target, optimizer, and training schedule fixed and varied only the encoder: (B) remove LayerNorm; (C) remove LayerNorm and Dropout—the backbone already sets dropout to $p = 0$ (App. L), so (C) is expected to coincide with (B) and we retain it only as the audit enumerated it; (D) widen the MLP to 1024 hidden units; (E) add a learnable scalar shortcut $f(\mathbf{x}) = \mathrm{MLP}(\mathbf{x}) + \beta \cdot \|\mathbf{x}\|_2$ with β initialized at 0. Variants (B)–(D) are strictly $\mathbf{x}$-only; variant (E) augments $\mathbf{x}$ with the single hand-designed magnitude scalar $\|\mathbf{x}\|_2$ (i.e., m_2 from Eq. 1) routed through an additive shortcut, so it is the strongest x-with-magnitude-shortcut control we tested. Across three seeds each, no variant crosses VCC test $R^2 = +0.150$ (best individual seed: wide-no-norm seed 1 at $R^2 = +0.034$); means range from -0.040 ± 0.021 to $+0.016 \pm 0.018$ (sample standard deviation, $\mathrm{ddof}{=}1$). Per a pre-specified rule (PARTIAL-RESCUE if mean $R^2 \geq +0.150$), the outcome is NO-RESCUE. We conclude that the deep $\mathbf{x}$-only MLP failure is not explained by LayerNorm, Dropout, limited capacity, or even a hand-designed magnitude shortcut: even when the network is given direct access to m_2 via an additive path, it fails to recover the signal. This audit is reported as a diagnostic robustness analysis covering four common architecture variants and is not exhaustive; activation, optimizer, and skip-connection variants remain untested.

T RF HYPERPARAMETER ROBUSTNESS

To test whether the Random Forest result on $[\mathbf{x}; \widetilde{\boldsymbol{m}}]$ depends on a narrow hyperparameter choice, we ran a targeted hyperparameter robustness sweep. We held the data, features, and seeds fixed and varied only the RF: paper reference S0 ($n_{\text{est}}{=}200$, $\max_{\text{depth}}{=}20$, $\min_{\text{leaf}}{=}2$); S1 shallower trees ($\max_{\text{depth}}{=}10$); S2 unbounded depth; S3 stronger leaf regularization ($\min_{\text{leaf}}{=}5$); S4 more trees ($n_{\text{est}}{=}500$); S5 intermediate depth ($\max_{\text{depth}}{=}15$). Across three seeds $\times$ six settings (18 fits), VCC test R^2 remained in $[+0.361, +0.372]$ (Table 9), every setting meeting the pre-specified robustness criterion (VCC $R^2 \geq +0.330$). We report this sweep as a robustness check and keep the original S0 setting as the paper's primary RF configuration. (The original sweep also recorded a legacy distributed-endpoint K562 transfer Spearman per setting; those values are not reproduced under the locked revision harness and are omitted.)

Table 9: RF hyperparameter robustness sweep on $[\mathbf{x}; \widetilde{\boldsymbol{m}}]$. Mean $\pm$ sample standard deviation of VCC test R^2 across three seeds (ddof=1) for each setting. All settings meet the robustness criterion used for this sweep (VCC $R^2 \geq +0.330$).

ID	Configuration	VCC test R^2
S0	Paper reference ($n_{\text{est}}{=}200$, depth$=20$, leaf$=2$)	$+0.370 \pm 0.004$
S1	Shallower trees (depth$=10$)	$+0.361 \pm 0.003$
S2	Unbounded depth	$+0.369 \pm 0.004$
S3	Leaf regularization (leaf$=5$)	$+0.367 \pm 0.003$
S4	More trees ($n_{\text{est}}{=}500$)	$+0.372 \pm 0.003$
S5	Intermediate depth (depth$=15$)	$+0.370 \pm 0.004$

U ADDITIONAL DISCUSSION: MAG-SPSM AND MAG-ANCHOR AS A SECONDARY DEEP METHOD

Why we exposed magnitude as a feature rather than as a loss. Reweighting losses (Balanced MSE (Ren et al., 2022), LDS (Yang et al., 2021)) act on the objective while leaving the encoder's input unchanged. In the exploratory runs of Sec. 5 they did not resolve prediction-range collapse. Those runs predate the locked evaluation and their artifacts were not retained, so we treat this as the motivation for supplying magnitude as an explicit feature rather than as an established mechanism. Reweighting can in principle reshape a learned representation through its gradients; our claim is only that the variants we tried did not.

Foundation models scope. The diagnostic claims of this section concern the MLP-encoder family and may not generalize to other foundation-model-based perturbation predictors. Geneformer is evaluated here as a *frozen* representation (Sec. 6.1, Sec. 6.2); its behavior under fine-tuning is a separate question that we did not address. Whether other large pretrained single-cell encoders such as scGPT (Cui et al., 2024) and Universal Cell Embeddings (Rosen et al., 2026) recover magnitude on the disjoint target-gene split, and whether any of these encoders recovers it once its weights are allowed to adapt, was not evaluated. Compute and preprocessing costs (gene-vocabulary alignment, and the memory budget for fine-tuning on a single 24 GB device) placed a fairness-controlled fine-tuning comparison outside the scope of this submission; it is an important direction for follow-up work.

MAG-SPSM as a proof of concept, not the final model. The strong-baseline audit (App. H) decisively shows that MAG-SPSM is not the strongest exploiter of the magnitude signal. Random Forest on the same features outperforms it by $\Delta R^2 = +0.145$ in-distribution. The earlier submission also reported a gap on the legacy distributed endpoint; that value was not re-derived under the locked harness and we do not use it as evidence here. We retain MAG-SPSM in the main text as a diagnostic deep augmentation: the simplest verification that the same magnitude signal can be exposed to a deep encoder, and the substrate for the per-row alignment controls in Sec. 5.3. We do not recommend MAG-SPSM as a final architecture.

MAG-Anchor as a secondary deep method. MAG-Anchor is a secondary proof-of-concept that refines, rather than replaces, the diagnostic story; it was not reproduced under the locked revision harness, so we rest no quantitative claim on it. Its per-row-alignment controls (App. J) indicate qualitatively that exposing magnitude at the output alone is not enough—the residual branch must also condition on the magnitude summary—while Random Forest on $[\mathbf{x}; \widetilde{m}]$ remains the stronger in-distribution model. We therefore treat MAG-Anchor as a secondary deep-architecture contribution rather than a replacement for the strongest classical baseline.

Two readings of the diagnostic. A reasonable alternative reading of our finding is that mean-collapse arises from optimization or normalization behavior of the deep MLP when the label scale couples to input scale through normalization layers, rather than from a representational property of perturbation prediction *per se*. The architecture-control analysis in App. S tests this reading directly: holding the data, target, optimizer, and training schedule fixed and varying only the encoder (no LayerNorm; no LayerNorm and no Dropout; widened to 1024 hidden units; an additive scale-preserving shortcut), no $\mathbf{x}$-only variant crosses VCC test $R^2 = +0.150$, well below the $+0.208$ obtained by a four-feature linear regression on $\widetilde{m}(\mathbf{x})$. This empirically disfavors the optimization-only reading within the MLP family we tested, though it does not exclude it for architectures we did not test (set transformers, attention-based encoders, foundation-model fine-tuning). Under both readings, explicit magnitude exposure is the empirical separator between models that recover the magnitude signal and models that collapse, and the practical recommendation is the same—supply magnitude as features rather than rely on the encoder to recover it from the expression vector alone. The diagnostic is therefore most precisely stated as a finding about *which models, on the disjoint target-gene split with this label parameterization, exploit a known low-dimensional predictive signal*.

V ADDITIONAL LIMITATIONS

1. *The label is magnitude-like by construction.* Because the target is the log Anderson–Darling distance from NT, magnitude features are in principle informative. Our claim is not that this relationship is biologically surprising in isolation; it is that direction is fragile under the disjoint target-gene split, that standard encoders fail to exploit magnitude reliably, and that explicit magnitude exposure is the empirical separator under cross-cell transfer.

2. *Replogle preprocessing residuals.* Replogle K562 uses a different per-row biological unit than VCC ($1{,}430$ of $2{,}000$ HVGs mapped; feature scales differ by $\sim 1.9\times$). We align via VCC-side normalization (App. M); residual confounding cannot be fully ruled out.

3. *Cross-dataset R^2 scale disparity.* VCC test, Replogle K562 transfer, and Replogle RPE1 transfer have AD-based labels with substantially different marginal standard deviations (sample std, $\mathrm{ddof}{=}1$): VCC $\sigma_y = 0.758$, K562 $\sigma_y = 2.750$ ($3.63\times$ the VCC value), and RPE1 $\sigma_y = 2.372$ ($3.13\times$ the VCC value). Since R^2 denominators scale with the squared label std, cross-dataset R^2 magnitudes are not directly comparable across the three benchmarks; the K562 R^2 denominator is approximately $13.2\times$ the VCC denominator and the RPE1 denominator is approximately $9.8\times$. Cross-dataset R^2 values should therefore be read as within-dataset goodness-of-fit measures, not as a single calibrated quantity transferable from one dataset to the other; rank-based metrics (Spearman ρ) are unaffected by this scale disparity and are the metric we emphasize in cross-cell claims.

4. *HVG-coverage confound on cross-cell transfer.* Models are trained on the full $2{,}000$-dimensional VCC HVG namespace. In the primary Replogle K562 transfer evaluation, Replogle rows are embedded into this $2{,}000$-dimensional namespace by zero-imputing the 570 VCC-only coordinates missing from Replogle. This is a structural input shift that is logically distinct from the biological cross-cell shift. The HVG-coverage control in App. X.1 removes the zero-padding by training and evaluating directly in the $1{,}430$ shared-HVG subspace, and K562 rank association is preserved and slightly higher ($\rho = +0.377 \pm 0.010$ vs. $+0.314 \pm 0.009$). This addresses the coverage question only for the legacy distributed-endpoint analysis; the corresponding prediction vectors were not retained, so the control cannot be re-scored against the rebuilt target-gene endpoint without an unvalidated refit.

5. *Modest transfer magnitudes.* On the distributed Replogle endpoint the fits we observe are modest (legacy distributed-endpoint values are not reproduced under the locked revision harness and are omitted). On the rebuilt target-gene endpoint we report rank association rather than pooled R^2, since predictions produced on one outcome scale are not calibrated to another. The magnitude increment is modest and does not solve cross-context generalization.

6. *Tree behavior on low-dimensional input.* Random Forest on MAG-ONLY has VCC test $R^2 = -0.042$ but positive Replogle rank association; the negative VCC R^2 reflects poor calibration of leaf means rather than absence of signal (the in-distribution Spearman is $+0.555$, and tree predictions are bounded by the training label range). We therefore interpret tree results on MAG-ONLY as evidence of preserved rank association, not as calibrated in-distribution predictors.

7. *No exhaustive deep-architecture search.* The trained encoders we tested are MLP-family (SPSM/MAG-SPSM, MAG-Anchor, and the MAGAnchor-x ablation); Geneformer enters only as a frozen representation, with no encoder parameter updated. Set Transformers, attention-based encoders, foundation-model *fine-tuning* (scGPT (Cui et al., 2024), Geneformer (Theodoris et al., 2023), UCE (Rosen et al., 2026)), and width and depth sweeps were not evaluated. The conclusion that no deep encoder we trained matches the tree-ensemble baseline is therefore empirical and limited to the architectures listed. Within the MLP family, additional architecture controls (no-LayerNorm, no-Dropout, wider, and scale-aware variants) did not recover signal from $\mathbf{x}$-only input (App. S). Whether other architecture families recover magnitude from $\mathbf{x}$ alone, or whether the gap to RF on $[\mathbf{x}; \widetilde{\boldsymbol{m}}]$ narrows under such architectures, remains an important open question.

8. *MAG-Anchor is not state of the art.* As submitted, MAG-Anchor obtained positive H1 $\rightarrow$ K562 rank association when scored against the legacy distributed endpoint, and improved over MAG-SPSM in-distribution; those figures were not reproduced under the locked revision harness and no value from them is carried into this revision (App. J). It is not the strongest in-distribution predictor: Random Forest on $[\mathbf{x}; \widetilde{\boldsymbol{m}}]$ remains the stronger in-distribution model. We therefore

view MAG-Anchor as a secondary method contribution that operationalizes the diagnostic finding, not as a new state-of-the-art perturbation predictor.

9. *Statistical reporting choices and their scope.* We report all stochastic results as mean ± sample standard deviation (ddof=1) across seeds, and report paired three-seed deltas for matched-architecture comparisons (e.g., MAG-RANK vs. MAG-SPSM, $\Delta R^2 = +0.0142 \pm 0.0040$; App. N). For three central comparisons where reliable per-row predictions are available, we report row-level paired bootstrap 95% confidence intervals on the pooled test set (n_{VCC}=3,972, n_{K562}=2,176) in App. R. We do not report formal hypothesis tests, paired Wilcoxon signed-rank statistics, or multiple-comparison corrections across the full method grid; with $n_{\text{seeds}} \in \{3, 5\}$, such tests would be underpowered for small effects, and we judged that qualitative pattern descriptions are more honest than artificially sharp p-values. The central qualitative findings of the paper do not depend on small-effect comparisons: a row-level paired bootstrap places the Random Forest advantage on $[\mathbf{x}; \widetilde{\boldsymbol{m}}]$ over MAG-SPSM at $\Delta R^2 = +0.145$ (95% CI $[+0.127, +0.163]$, sign-stable across resamples); the contrast between expression-only models and magnitude-aware models on cross-screen transfer is qualitative rather than marginal; and the harmonized transfer verdicts of Sec. 6.2 are determined by whether a bootstrap interval excludes zero, not by a p-value cutoff. Pairwise deltas with overlapping intervals are flagged as unresolved in the Discussion and in App. U; we do not claim these are statistically distinguishable. A targeted hyperparameter robustness sweep for the RF result is reported in App. T.

W ANDERSON–DARLING LABEL CONSTRUCTION

For each row $i = (b_i, g_i)$ with experimental Flex batch b_i and target gene g_i, the regression label is

$$y_i \;=\; \log\!\big(1 + \max\{0,\, A^2_{b_i,g_i}\}\big),$$

where $A^2_{b,g}$ is the two-sample Anderson–Darling statistic returned by `scipy.stats.anderson_ksamp` (Scholz & Stephens, 1987) applied to two one-dimensional samples: (i) the log1p-normalized expression values of the target gene g in cells perturbed at g within batch b, and (ii) the log1p-normalized expression values of the same gene g in non-targeting (NT) control cells from the same batch b. The Anderson–Darling test is therefore applied to the *target-gene column only*; the 2,000-HVG vector $\mathbf{x}_i$ is the predictor input and does not enter label construction. Negative AD statistics (rare, possible with the AD null-bias subtraction) are clamped to zero before the $\log(1 + \cdot)$ transform. This per-row (b, g)-indexed convention, with b playing the role of the Replogle gemgroup-style batch identifier, is the one used in the earlier distributed-endpoint analysis. The harmonized endpoint that the locked transfer evaluation of Sec. 6.2 scores against is built differently: it is construct-level, uses a pooled non-targeting control, and is computed by repeated fixed-size subsampling (App. P).

X RESULTS FROM THE EARLIER SUBMISSION, NOT RE-DERIVED

The tables in this section reproduce results from the earlier version of this work, computed with classical baselines. Their numeric values were not re-derived under the locked evaluation described in Sec. 4, and the artifacts required to reproduce them were not retained. They are included for continuity with the prior version. No claim in this paper depends on them. Each table's caption states the endpoint against which it was computed.

Table 10 reports point estimates and verdicts. The corresponding 95% intervals come from a paired cluster bootstrap over target genes with $B = 10{,}000$ replicates and a pre-specified seed. Each replicate samples target-gene clusters with replacement; a cluster drawn more than once contributes all of its constructs once per occurrence; and the same draw is applied jointly to every prediction, every seed, and every contrast within that replicate, so that paired comparisons share resampling noise.

For the nine classical families the metric is computed per seed and averaged, with the averaging repeated inside each replicate; for the three single-seed predictors the metric comes from the one archived prediction vector, so the interval reflects target-gene sampling conditional on that seed. No replicate was discarded for an undefined correlation in either screen.

Table 10: Legacy classical-model transfer results, retained for historical comparison. These values are not the current locked foundation-model evaluation (Table 1) and were not independently re-derived in the revision harness. Zero-shot transfer to two external screens on a common target-gene endpoint. Entries are construct-level Spearman correlations between each VCC-trained prediction and the harmonized endpoint. Verdicts follow a pre-specified rule: Pos if the 95% target-gene cluster-bootstrap interval lies wholly above zero, Neg if wholly below, Inc if it includes zero; an inconclusive verdict is not evidence of absent transfer. Panel A families are evaluated per seed and averaged (3 seeds for HistGB and RF, 1 for Ridge). Panel B predictors have a single training seed, so their intervals are conditional on it. Feature sets: $\mathbf{x}$ expression only, $\widetilde{\boldsymbol{m}}$ magnitude only, $[\mathbf{x}; \widetilde{\boldsymbol{m}}]$ both. Interval widths are not comparable across screens: the two evaluations differ in cluster structure (App. X).

		RPE1		K562-essential	
Model	Features	ρ	Verdict	ρ	Verdict
Panel A: classical families, seed-averaged					
Ridge	$\mathbf{x}$	-0.091	Neg	$+0.009$	Inc
Ridge	$\widetilde{\boldsymbol{m}}$	$+0.097$	Pos	$+0.133$	Pos
Ridge	$[\mathbf{x}; \widetilde{\boldsymbol{m}}]$	-0.068	Neg	-0.043	Inc
HistGB	$\mathbf{x}$	-0.116	Neg	-0.211	Neg
HistGB	$\widetilde{\boldsymbol{m}}$	$+0.009$	Inc	$+0.160$	Pos
HistGB	$[\mathbf{x}; \widetilde{\boldsymbol{m}}]$	-0.038	Inc	$+0.018$	Inc
RF	$\mathbf{x}$	-0.105	Neg	-0.218	Neg
RF	$\widetilde{\boldsymbol{m}}$	$+0.053$	Pos	$+0.178$	Pos
RF	$[\mathbf{x}; \widetilde{\boldsymbol{m}}]$	-0.018	Inc	$+0.166$	Pos
Panel B: deep and magnitude-only predictors					
SPSM baseline	—	-0.050	Neg	-0.070	Neg
MAG-SPSM	—	$+0.067$	Pos	$+0.016$	Inc
MAG-LINEAR	$\widetilde{\boldsymbol{m}}$	$+0.057$	Pos	$+0.147$	Pos

Table 11: HVG-coverage control for K562 transfer against the legacy distributed-breadth endpoint. Mean $\pm$ sample standard deviation over three seeds.

Training namespace	VCC test R^2	K562 Spearman ρ
2,000 VCC HVGs	$+0.370 \pm 0.004$	$+0.314 \pm 0.009$
1,430 shared HVGs	$+0.340 \pm 0.002$	$+0.377 \pm 0.010$

The two evaluations differ in cluster structure: RPE1 has 1,932 constructs over 1,811 target genes with at most three constructs per gene, and K562-essential has 1,903 over 1,811 with at most two. These are distinct screen-specific gene sets of equal cardinality. Interval widths are therefore not mechanically comparable across screens.

X.1 HVG-COVERAGE CONTROL

To test whether the legacy K562 distributed-endpoint transfer result is explained by the 2,000-to-1,430 HVG dimension drop, we reran the RF $[\mathbf{x}; \widetilde{\boldsymbol{m}}]$ baseline using only the 1,430 HVGs shared between VCC and Replogle K562. The shared genes were selected deterministically by gene name in VCC HVG order, without using Replogle target labels or outcomes. Magnitude features were recomputed from the 1,430-dimensional inputs and z-scored using VCC training statistics only. Across three seeds with the same RF hyperparameters as the main experiment (n_{est}=200, $\max_{\text{depth}}$ =20, $\min_{\text{leaf}}$ =2), the 1,430-HVG control achieved VCC test $R^2 = +0.340 \pm 0.002$ and K562 transfer Spearman $\rho = +0.377 \pm 0.010$ (Table 11). This passes our pre-specified robustness rule ($\rho \geq +0.20$) and shows that, *when transfer is scored against the distributed endpoint*, the positive K562 rank association is not explained by the missing 570 HVGs alone. The in-domain VCC R^2 decreases relative to the 2,000-HVG result ($+0.370 \pm 0.004$), as expected from the reduced feature set, but cross-cell rank transfer is preserved.

Pre-specified HVG-coverage control: x-only models. We additionally pre-specified a symmetric x-only counterpart of the control above to test whether the negative cross-cell transfer of expression-only models is a feature-truncation artifact rather than a direction-vs-magnitude phenomenon. The pre-registration committed four hypotheses with PASS/WARN/FAIL thresholds and the consequent paper-text edits before any model was trained in the pre-registration file committed to the project repository. We trained Ridge, Random Forest, HistGradientBoosting, and the deep SPSM V4 baseline on $\mathbf{x}$ alone restricted to the same $1{,}430$ shared HVGs, three seeds each, with hyperparameters frozen at their main-paper values. Three-seed mean K562 Spearman values were: Ridge $\rho = -0.229$, Random Forest $\rho = -0.407 \pm 0.003$, HistGradientBoosting $\rho = -0.305 \pm 0.019$, and SPSM V4 baseline $\rho = -0.389 \pm 0.007$. All four pre-specified hypotheses passed: H1 (every $\mathbf{x}$-only classical model retains negative mean Spearman) passed with the largest value at Ridge -0.229; H2 ($\mathbf{x}$-only deep baseline mean $\rho < +0.05$) passed at -0.389; H3 (best magnitude-aware minus best $\mathbf{x}$-only ρ gap on the $1{,}430$-HVG sub-space $\geq +0.20$) passed with gap $+0.607$ ($+0.377$ vs -0.229); and H4 (within-VCC $|\Delta R^2| \leq 0.030$ relative to the $2{,}000$-HVG baselines) passed with worst-case $|\Delta R^2| = 0.011$ for Ridge. Per the pre-specified decision rule, the legacy distributed-endpoint $\mathbf{x}$-only negative-transfer pattern is robust to feature truncation and the magnitude–direction asymmetry holds on the shared-HVG sub-space.

Scope. Per-construct prediction vectors from this control were not retained, so it cannot be re-scored against the rebuilt target-gene endpoint without an unvalidated refit. We therefore treat it as a legacy distributed-endpoint sensitivity analysis and do not use it to support Table 10.