How Many Measured Genes Is a Prior Worth? Unseen-Perturbation Prediction as Decoding

Anonymous authors
Paper under double-blind review

Abstract

Models of transcriptional responses to genetic and chemical perturbations rarely beat the mean training response on unseen perturbations. We ask what a small measurement of the new perturbation would add, and whether its value can be predicted. Treating prediction as decoding, we note that without perturbation-specific information beyond the training set no architecture beats the mean response, predict the expected held-out error of linear measurements (a few genes or cells, or another cell line's response) from training perturbations alone with the classical risk of least squares, and bound combinations of training responses by a coverage floor. On five screens from three studies, the best tested zero-shot predictor recovers 7% to 30% of the recoverable signal and larger tested decoders at most 1.0 percentage point more, whereas four greedily chosen genes of the new perturbation (a simulated targeted readout) recover 33% to 75%; the best prior is worth about half of the first greedy gene on four screens (0.4 to 0.6 gene equivalents) and 3.4 genes on the fifth. With eight or more training perturbations per measured dimension, predictions track held-out errors (slope 0.97, $R^2 = 0.98$, optimism 0.03) and choose panel sizes better than naive estimates, but are optimistic by about 0.11 below 100 training perturbations. Withholding a pathway raises the coverage floor from at most 13% to at least 33% of the held-out signal, and genes, priors and predictions fail; covering a quarter of it brings genes and predictions most of the way back.

1 Introduction

A virtual cell is expected to answer a counterfactual question: how would the transcriptome of a cell change under a perturbation that has never been applied to it (Bunne et al., 2024; Roohani et al., 2025b)? The hardest version, predicting the response to a genetic or chemical perturbation absent from the training data, is also where current models struggle. Deep perturbation models and single-cell foundation models rarely outperform the mean response of the training perturbations or a linear baseline on unseen perturbations (Ahlmann-Eltze et al., 2025; Csendes et al., 2025); simpler architectures are competitive (Wu et al., 2025), conclusions depend on the metric (Li et al., 2026), standard metrics reward systematic variation (Viñas Torné et al., 2026), and predictions compress the geometry that distinguishes perturbations (Huang et al., 2026). What is missing: capacity, or information about the new perturbation?

We treat a predictor as a *decoder* of a message about the perturbation, such as the identity of its target, the prior knowledge attached to it, or a small measurement of it. One statement is then universal: with no perturbation-specific information beyond the training set, the mean response is Bayes-optimal up to a factor $1 + 1/n$, whatever the architecture (Theorem 1). Everything else is operational and testable. For linear measurements of the new perturbation (a few of its genes, as a targeted assay would read them (Subramanian et al., 2017; Schraivogel et al., 2020), a few of its cells, or its response in another cell line or screen), we predict the expected held-out error from training perturbations alone, before the measurement is made. We price a prior against greedily chosen measured genes, and we show that predictors that combine training responses cannot cross a coverage floor: the part of a new response that lies outside the span of the training responses.

On CRISPR-activation, CRISPR-interference and chemical screens (Figure 1), the tested zero-shot predictors, from nearest neighbours to multilayer perceptrons, GEARS (Roohani et al., 2024) and frozen foundation-model embeddings, reach errors that more capacity does not lower, whereas four

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

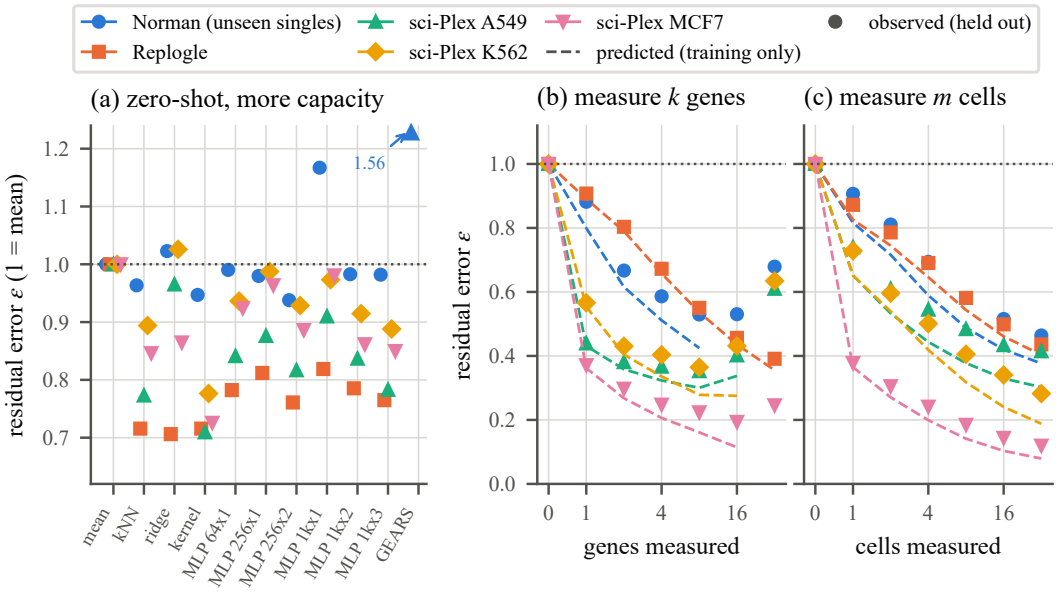


Figure 1: **At fixed side information, capacity barely moves the error; a few measurements of the new perturbation move it sharply.** Held-out normalized residual error ε (1: mean response of the training perturbations; 0: noise limit). (a) Zero-shot predictors with their best side information, by model family and parameter count: nearest neighbours, ridge and kernel ridge, MLPs of width w and depth d (1.5×10^5 to 7×10^6 parameters) and, on the Norman singles, GEARS at a reduced budget (off scale at 1.56). (b, c) Predictors that receive k genes or m cells of the held-out perturbation (regression decoder): markers are cross-validated observations, dashed lines predictions from training perturbations only (Theorem 3), drawn where $n/q \geq 8$.

well-chosen measured genes of the new perturbation beat every tested prior, by a wide margin on four of five screens. Unlike benchmarks that rank predictors (Ahlmann-Eltze et al., 2025; Wu et al., 2025) or diagnoses of what is missing (Viñas Torné et al., 2026; Huang et al., 2026), we ask how much a specified measurement will buy before it is made, and when the answer can be trusted.

Contributions.

- **Predicting what a measurement buys.** For linear channels we give the attainable error, a crossover cell count and a misattribution identity (Theorem 2), and the expected held-out error of the learned decoder from training perturbations alone (Theorem 3, Corollary 1); in its regime it is calibrated and chooses panel sizes better than naive estimates (Section 5.2).
- **A currency for priors.** The greedy-panel measurement equivalent (Definition 1) prices priors, including frozen foundation-model embeddings, at about one greedy gene or less on four of five screens (Table 2, Section 5.1); priors are redundant with measured genes on four screens and complementary on one, and a measurement of the same perturbation in another cell line is worth more than the best prior (Section 5.3).
- **A coverage boundary.** A coverage floor (Proposition 1) explains when priors, measured genes and error predictions fail together: it rises when a pathway is withheld and falls as training covers part of it (Section 5.4, Figure 3d, Table 1).

2 Related Work

Perturbation-response prediction. Models predict transcriptional responses from perturbation descriptors by latent arithmetic (Lotfollahi et al., 2019), compositional autoencoders (Lotfollahi et al., 2023; Hetzel et al., 2022), sparse mechanism shifts (Bereket & Karaletsos, 2023), optimal transport (Bunne et al., 2023), knowledge graphs (Roohani et al., 2024), single-cell foundation models (Cui et al., 2024; Theodoris et al., 2023), virtual-cell models (Adduri et al., 2026), flow matching (Yu et al.,

2026) and language-model gene embeddings (Chen & Zou, 2025). They improve the decoder; we ask how much information any decoder receives, and how to supply the rest.

Diagnosing the unseen-perturbation gap. Systema separates systematic from perturbation-specific variation (Viñas Torné et al., 2026), response decomposition assigns different information requirements to global and perturbation-specific components (Molina & Zhang, 2026), and Huang et al. (2026) show that predictions compress the geometry of unseen interventions, that a few directions learned from seen interventions capture much of a held-out response, and that same-target anchors transfer across contexts. These studies diagnose *what* is missing; we predict what a specified measurement supplies and bound when it cannot.

Measuring less, inferring more. The L1000 assay measures 978 landmark genes and infers the rest (Subramanian et al., 2017); geneBasis (Missarova et al., 2021) and PERSIST (Covert et al., 2023) select small gene panels that reconstruct transcriptomes, and composite (Cleary et al., 2017), targeted (Schraivogel et al., 2020) and compressed (Yao et al., 2024) Perturb-seq lower the cost of readouts. These methods choose panels that reconstruct cell states; we price a panel against a prior and predict its held-out error for a new perturbation before it is measured. Bayesian experimental design (Chaloner & Verdinelli, 1995) and its uses in functional genomics choose *which perturbations* to run (Mehrjou et al., 2022; Huang et al., 2024; Roohani et al., 2025a).

Information-theoretic limits and model selection. Rate-distortion theory (Shannon, 1959; Cover & Thomas, 2006) and lower bounds on the Bayes risk (Chen et al., 2016; Xu & Raginsky, 2022) bound error by the available information, usable information restricts it to a model family (Xu et al., 2020), greedy sensor selection has guarantees (Krause et al., 2008; Das & Kempe, 2011), and the optimism of least squares is classical (Breiman & Freedman, 1983). We use these tools as they are; the contribution is their operational composition.

3 Unseen-perturbation prediction as decoding

Setup. A perturbation p (a CRISPR target or a compound) is drawn from a population whose law Θ is unknown and random, with distribution Π. Its *response* $R_p \in \mathbb{R}^G$ is the expected change in log-normalized expression of G genes relative to control cells, and $Z_p = R_p - \mu_\Theta$, with $\mu_\Theta = \mathbb{E}[R \mid \Theta]$, is its *perturbation-specific residual*. Responses are observed through pseudobulk means of finitely many cells, so we split the cells of every perturbation into two disjoint halves, along batch labels whenever they exist. Half B gives the target $T_p = R_p + \xi_p$. A *message* M_p comes from half A (the responses of k genes or the profiles of m cells) or from another cell line or screen (the response of the same perturbation there). *Side information* S_p, available without any experiment on p, is a descriptor of p (for example Gene Ontology terms of its target gene or a chemical fingerprint) together with the n training perturbations $\mathcal{D}_n$. A predictor f has *distortion* $D(f) = \mathbb{E}\|T_p - f(S_p, M_p)\|^2$. Table 3 in Appendix A lists the notation.

Assumption 1 (Exchangeable perturbations and independent halves)**.** Given Θ, training and held-out perturbations (with their target noise) are i.i.d., and the target noise satisfies $\mathbb{E}[\xi_p \mid \Theta, R_p, S_p, M_p] = 0$.

Then $D(f) = \mathbb{E}\|R_p - f\|^2 + \operatorname{tr} N_B$ with $N_B = \operatorname{Cov}(\xi)$, the cell sampling and batch effects that the halves do not share (Lemma 1). We report the *normalized residual error*

$$\varepsilon(f) = \frac{D(f) - \operatorname{tr} N_B}{D_0 - \operatorname{tr} N_B}, \tag{1}$$

where D_0 is the distortion of the mean response of the training perturbations: $\varepsilon = 1$ is the mean-response (zero-bit) error, $\varepsilon = 0$ is perfect recovery up to measurement noise, and $1 - \varepsilon$ is the fraction of the *recoverable signal* (the response variance above the noise) that a predictor recovers. Being a ratio of squared errors, ε weighs perturbations by their response energy.

The zero-information predictor. Let $\mathcal{R}_{Z|S}(D) = \inf\{I(Z; \hat{Z} \mid S) : \mathbb{E}\|Z - \hat{Z}\|^2 \le D\}$ be the conditional rate-distortion function of the residual (Shannon, 1959; Cover & Thomas, 2006); expectations include $\Theta \sim \Pi$.

Theorem 1 (No capacity without information). *Let $\hat{Z} = g(S, M, U)$ for any measurable g and independent randomness U, i.e. any architecture with any number of parameters. If $\mathbb{E}\|Z - \hat{Z}\|^2 \le D$, then $I(Z; M \mid S) \ge \mathcal{R}_{Z|S}(D)$. In particular, if $I(Z; M \mid S) = 0$, then $\mathbb{E}\|Z - \hat{Z}\|^2 \ge F(S) := \mathbb{E}\,\mathrm{tr}\,\mathrm{Cov}(Z \mid S)$, with equality iff $\hat{Z} = \mathbb{E}[Z \mid S]$. If, beyond $\mathcal{D}_n$, the descriptor and M carry no information about (Θ, Z), the Bayes predictor of the target is $\mathbb{E}[\mu_\Theta \mid \mathcal{D}_n]$, and the mean of the training targets has at most $1 + 1/n$ times its distortion, provided $\mathbb{E}\,\mathrm{tr}\,\Sigma_\Theta$ and $\mathrm{tr}\,N_B$ are finite, with $\Sigma_\Theta = \mathrm{Cov}(Z \mid \Theta)$.*

Theorem 1 is the only universal statement about capacity here. When the descriptor is informative, $F(S)$ depends on Π and cannot be computed; a tested family's error bounds it from above and measures that family's usable information (Xu et al., 2020), so whether larger decoders extract more is an empirical question (P1 below).

Linear channels. From here on the law is fixed. A message is a noisy linear channel $y = HZ + e$, with zero-mean e independent of Z, $\Sigma = \mathrm{Cov}(Z) \succeq 0$ and $\Sigma_e = \mathrm{Cov}(e) \succ 0$, so that $C = H\Sigma H^\top + \Sigma_e$ is invertible. A k-gene panel selects k rows of the identity; the mean of m cells has $H = I$ and $\Sigma_e = \Sigma_c/m + N_{\mathrm{msg}}$, with Σ_c the cell-level covariance and N_{msg} the batch noise of the message.

Theorem 2 (Linear channels). *(a) $K = \Sigma H^\top C^{-1}$ is the best linear decoder, with risk $D^\star(H) = \mathrm{tr}\,\Sigma - \mathrm{tr}(\Sigma H^\top C^{-1} H \Sigma)$; if (Z, e) is Gaussian it is the Bayes decoder. (b)* Crossover cell count: *if $\mathrm{tr}\,N_{\mathrm{msg}} < \mathrm{tr}\,\Sigma$, the raw mean of m cells beats the zero-bit predictor iff $m > m_\times := \mathrm{tr}\,\Sigma_c/(\mathrm{tr}\,\Sigma - \mathrm{tr}\,N_{\mathrm{msg}})$; otherwise it never does. (c)* Misattribution identity: *applying K to the message of another, independent perturbation gives risk $2\,\mathrm{tr}\,\Sigma - D^\star(H)$: a wrong message costs exactly what the right message gains. All parts use second moments only; Proposition 2 (Appendix A) adds the information carried by a channel and a reverse water-filling envelope.*

Learning the decoder from n perturbations. The *regression decoder* fits $T \approx a + \beta y$ by least squares across the n training perturbations, with training messages produced by the same channel (for example m cells drawn from each training perturbation's half A).

Theorem 3 (Data-limited and information-limited error). *Let the message y have dimension q, $\mathrm{Cov}(y) \succ 0$, and a design (the genes of a panel, or a projection) fixed independently of the training responses, and let (y, T) be jointly Gaussian. The regression decoder trained on $n > q + 2$ perturbations has expected held-out distortion*

$$\mathbb{E}D_q^{(n)} = L_q\Big(1 + \frac{1}{n}\Big)\frac{n-2}{n-q-2}, \qquad \mathbb{E}\Big[\frac{\mathrm{RSS}}{n-q-1}\Big] = L_q, \tag{2}$$

where L_q is the risk of the best linear decoder on the target (including $\mathrm{tr}\,N_B$) and RSS the training residual sum of squares (Breiman & Freedman, 1983).

Corollary 1 (Designs selected on independent perturbations). *If the design is chosen, by any rule, from perturbations $\mathcal{D}_{\mathrm{sel}}$ independent of the n perturbations that fit the decoder and of the held-out ones, Equation (2) holds conditionally on $\mathcal{D}_{\mathrm{sel}}$. Random designs are the case $\mathcal{D}_{\mathrm{sel}} = \emptyset$.*

Equation (2), the classical optimism correction of least squares, applies to any linear message that can be simulated on the training perturbations before it is measured for a new one, and predicts an expected error. It is exact for random and split-selected panels under Gaussianity; for panels selected on the fitting perturbations, as a practitioner would choose them, and for heavy-tailed responses we test it as an approximation.

A coverage floor. Most predictors we test output a weighted combination of training responses; for them, what training never saw is out of reach whatever the message.

Proposition 1 (Coverage floor). *Let f output an affine combination $\sum_i w_i T_i$, $\sum_i w_i = 1$, of the training targets (the regression decoder of any linear message, ridge and kernel ridge regression, nearest-neighbour and class averages, the mean response). Then $\|R_0 - f\|^2 \ge \mathrm{dist}^2(R_0, \mathcal{A}_n)$ for a held-out response R_0, where $\mathcal{A}_n$ is the affine hull of the training targets. If R_0^A and R_0^B are the pseudobulk responses of two independent halves of the held-out perturbation, $d_0^2 = \langle P_\perp(R_0^A - \bar{T}), P_\perp(R_0^B - \bar{T})\rangle$ estimates $\mathrm{dist}^2(R_0, \mathcal{A}_n)$ without bias, where $\bar{T}$ is the mean training target and $P_\perp$ projects out the span of the centred training targets.*

Table 1: Predictions, their basis and regime, and their outcomes on five screens (details in Section 5).

Prediction	Basis and regime	Test	Outcome
(P1) At fixed side information, larger decoders do not lower the zero-shot error; measured related responses do	Theorem 1 frames it; empirical, tested families	Fig. 1a, Table 30, App. C	MLPs at most 0.010 better; seen singles reach $\varepsilon = 0.26$
(P2) Equation (2) predicts the expected held-out error before measuring	Theorem 3, Corollary 1: exact for random or split designs, else approximate	Fig. 2a, Tables 6, 7	Slope 0.97, $R^2 = 0.98$ at $n/q \geq 8$; optimistic by 0.11 below 100 perturbations
(P3) A misattributed message costs what the right one gains	Theorem 2c; well-estimated decoder	Fig. 2b	Slopes 0.86 to 1.16; fails on the 81 Norman singles
(P4) Raw means beat the mean response only beyond $m_\times$	Theorem 2b; second moments	Table 17	Crossovers bracket $m_\times$ on three screens; "never" holds
(P5) Combinations of training responses stay above a floor that rises without the pathway	Proposition 1; exact bound	Fig. 3d	Floor 7% to 13% (random folds), 33% to 47% without the pathway

Predictions. These results are standard in isolation (data processing, linear estimation, the random-design risk of least squares, and linear algebra; proofs in Appendix A); what they add is a set of predictions that can each fail on the screens (Table 1).

4 Experimental protocol

Screens. We use the CRISPR-activation screen of Norman et al. (2019) and the CRISPR-interference screen of essential genes of Replogle et al. (2022), both in K562 cells and read out by Perturb-seq (Dixit et al., 2016; Adamson et al., 2016), and the sci-Plex chemical screen of Srivatsan et al. (2020) (compounds at 10 μM in A549, K562 and MCF7), in harmonized public releases (Li et al., 2026): 226 Norman conditions with at least 30 cells (101 single-gene and 125 two-gene perturbations), 1,617 Replogle perturbations and 165, 181 and 181 compounds, 2,370 conditions in all. We model the (up to) 2,000 most variable genes of *control* cells, so the gene universe never depends on a held-out perturbation, and remove all perturbed target genes from it, so that no message or target contains the self-effect of a CRISPR guide.

Batch-disjoint halves and folds. The two halves of each perturbation come from different 10x lanes (Norman, Replogle) or replicate wells (sci-Plex); the median half-A pseudobulk rests on 45 to 280 cells per perturbation. We use repeated 5-fold cross-validation over perturbations (four repeats; two for Replogle), and everything a predictor or a predicted error uses (means, covariances, panels, decoders, noise levels) is estimated on the training perturbations of the fold. The signal covariance is the cross-covariance $\hat{\Sigma} = \widehat{\mathrm{Cov}}(R^A, R^B)$ of the half-A and half-B responses, unbiased because the halves have independent noise; the decoders use a factor model that keeps its eigen-directions above the noise edge (cf. Gavish & Donoho 2014). The noise $\mathrm{tr}\, N_B$ is the cell-sampling variance of the half-B mean plus the target variance that the halves do not share (Appendix B).

Measurement channels and decoders. Gene panels of size $k \in \{1, 2, 4, \ldots, 32\}$ are chosen on training perturbations by greedy variance reduction, which adds the gene that most lowers $D^\star(H)$ (Das & Kempe, 2011); by greedy information, which is $(1 - 1/e)$-optimal for the information of the diagonal (conditionally independent) readout-noise model used for selection (Krause & Guestrin, 2005; Nemhauser et al., 1978); by highest signal variance or mean response; as the most stable genes; or at random (10 draws). Random panels meet the design condition of Theorem 3; to test Corollary 1 we also select greedy panels on a random half of the training perturbations and fit the decoder on the other half. A k-gene message is the half-A pseudobulk of those genes, a simulated targeted readout. Cell budgets are $m \leq 32$ cells from half A (up to 128 where the screens allow). Cross-context messages are the same compound's response in another cell line, and the same target's response in the RPE1 essential-gene screen (Replogle et al., 2022). Every channel is decoded by the regression decoder (for cells and contexts, on the message projected onto q leading training directions); panels

Table 2: Held-out normalized residual error ε (Equation (1); 1: mean baseline, 0: noise limit) of the best zero-shot predictor, its greedy-panel measurement equivalent (Definition 1; 95% cluster-bootstrap interval), and predictors that receive k greedily chosen genes (read from the half-A pseudobulk, whose median number of cells is given) or m cells (perturbations with at least 32 half-A cells). Brackets: Equation (2) from training perturbations only, an approximation for these adaptive designs (optimistic by 0.01 to 0.10). P: perturbations (Norman: CRISPRa; Replogle: CRISPRi); n/a: held-out two-gene perturbations share genes with training ones, so this row is not an unseen-perturbation test. GO: Gene Ontology.

Screen	P	cells (half A)	best zero-shot ε (prior)	prior worth (greedy genes)	1 gene ε [pred.]	4 genes ε [pred.]	1 cell ε [pred.]	16 cells ε [pred.]
Norman (all)	226	191	n/a	n/a	0.77 [0.73]	0.42 [0.39]	0.72 [0.68]	0.34 [0.30]
Norman (singles)	101	280	0.93 (GO)	0.6 [0.0, 1.2]	0.88 [0.80]	0.59 [0.51]	0.91 [0.82]	0.52 [0.42]
Replogle	1,617	68	0.71 (GO)	3.4 [2.9, 3.9]	0.91 [0.89]	0.67 [0.66]	0.87 [0.83]	0.50 [0.46]
sci-Plex A549	165	45	0.71 (Morgan)	0.4 [0.3, 0.6]	0.44 [0.43]	0.37 [0.32]	0.74 [0.65]	0.43 [0.33]
sci-Plex K562	181	106	0.76 (Morgan)	0.5 [0.3, 0.7]	0.57 [0.55]	0.40 [0.34]	0.73 [0.65]	0.34 [0.24]
sci-Plex MCF7	181	105	0.70 (Morgan)	0.4 [0.2, 0.5]	0.37 [0.36]	0.25 [0.21]	0.38 [0.36]	0.14 [0.10]

and cells also by the plug-in Bayes decoder of Theorem 2, and cells also by their raw mean and by per-gene shrinkage (the Bayes decoder with a diagonal covariance).

Definition 1 (Greedy-panel measurement equivalent)**.** The measurement equivalent of a predictor is the number k of greedily chosen genes (greedy variance reduction) at which the held-out error of the panel, interpolated linearly in $\log_2(1+k)$, equals the predictor's error on the same perturbations. It is an exchange rate between two ways of lowering the error, not an additive quantity; below one gene it is $2^\eta - 1$ with η the ratio of the signal recovered by the predictor to that recovered by the first gene; and it does not depend on the noise floor. It depends on the panel rule: other informed rules and interpolations price the best priors at 0.3 to 1.6 genes on four screens and 3.4 to 6.4 on Replogle (Appendix Table 15).

Zero-shot predictors. For genetic screens the side information is the target gene's control-cell co-expression profile, its Gene Ontology annotations (Ashburner et al., 2000) and its GenePT embedding (Chen & Zou, 2025); for chemical screens, a 2,048-bit Morgan fingerprint and class annotations. These feed nearest-neighbour averaging, ridge and kernel ridge regression, and multilayer perceptrons of width 64 to 1,024 and depth 1 to 3 (1.5×10^5 to 7×10^6 parameters) that start at the mean response and stop early; frozen ESM-2 (Lin et al., 2023), Geneformer (Theodoris et al., 2023) and ChemBERTa (Ahmad et al., 2022) embeddings feed the first three. GEARS (Roohani et al., 2024) is trained on the unseen single-gene folds of Norman at a reduced budget (at most 100 cells per perturbation, six epochs; its validation error was lowest by the fourth epoch in every fold), so it illustrates rather than bounds what graph models achieve. The best prior of each screen is chosen by held-out error, which favours the priors. To test Assumption 1 and (P5), we hold out one sci-Plex pathway at a time (16 pathways) while a fraction $c \in \{0, 0.25, 0.5, 0.75\}$ of its compounds is moved into training.

5 Results

5.1 At fixed side information, larger tested decoders buy little

Figure 1a and Table 2 test (P1). For unseen single-gene activations on Norman, the best genetic prior leaves $\varepsilon = 0.93$; on Replogle, whose essential genes fall into complexes that Gene Ontology annotates, the same prior reaches 0.71; on sci-Plex, fingerprints reach 0.70 to 0.76. Each is the best of all tested predictors on held-out error, a choice that favours the priors. On identical folds, multilayer perceptrons with 1.5×10^5 to 7×10^6 parameters improve ε by at most 0.010 over the best kernel or linear predictor, and none of the controls (default initialization; 1,000 epochs) does better by more than 0.010 (Appendix Table 30). Frozen embeddings of large pretrained models do no better: ESM-2 protein and Geneformer gene embeddings reach ε of 0.93 to 1.00 on the genetic screens and ChemBERTa molecular embeddings 0.88 to 0.90 on sci-Plex, worth at most 0.7 measured genes (Appendix Table 28). GEARS at its reduced budget reaches $\varepsilon = 1.56$, and its correlation with the observed change from control (0.37 on all genes) is below that of the mean response (0.48; Appendix Table 32); like the deep models of Ahlmann-Eltze et al. (2025), it does not outperform the mean.

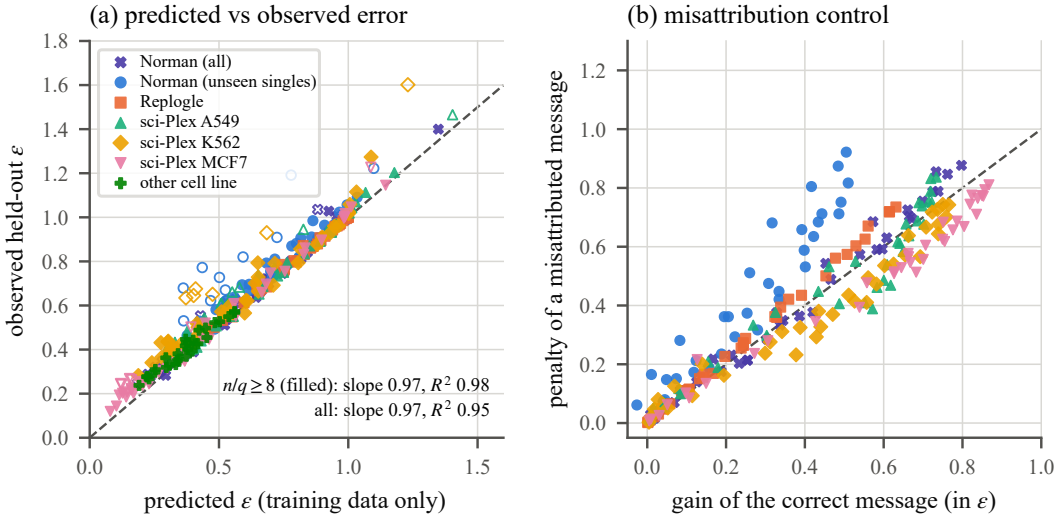


Figure 2: **Expected held-out error is predicted before measuring.** (a) Observed against predicted normalized residual error for every screen, channel, design and budget (regression decoder, raw mean and per-gene shrinkage; predictions from training perturbations only). Hollow markers: $n/q < 8$; 33 combinations with $\varepsilon > 1.8$ lie outside the range. (b) Misattribution control for the plug-in Bayes decoder of the gene panels: the message of another perturbation raises the error by what the correct message lowers it (dashed: identity).

The same model classes succeed once training contains measured responses of the same genes: for held-out two-gene perturbations whose single-gene responses are in training, ridge regression reaches $\varepsilon = 0.26$ (Appendix C). These statements concern the side information and decoder families tested, not fine-tuned foundation models.

5.2 Expected held-out error is predictable before measuring, when training is large

Figure 2a tests (P2) on 321 combinations. The design condition of the theorem matters less than the response distribution: random panels, whose design ignores the responses, are optimistic by 0.036, adaptive designs by 0.039, and selecting greedy panels on a separate half (Corollary 1) changes the optimism from 0.057 to 0.038, whereas on Gaussian responses the whole pipeline, adaptive selection included, is unbiased (0.002 against 0.043; Appendix Tables 4, 9 and 10). Over all designs at $n/q \geq 8$, the predictions match the cross-validated errors with slope 0.97 and $R^2 = 0.98$ and are optimistic by 0.03 on average (0.014 to 0.055 per screen). The leverage factor of Equation (2) does work: without it, the residual variance $\mathrm{RSS}/(n-q-1)$ and the in-sample error RSS/n fit worse (R^2 0.92 and 0.83, optimism 0.09 and 0.13), also within a fixed budget, and Equation (2) picks the best panel size on four of six screens against two (mean regret 0.04 against 0.12; Appendix Table 6). The regime needs n as well as n/q: in the training-size runs at $n/q \geq 8$, the optimism is 0.11 on average below 100 training perturbations (up to 0.29) and 0.01 above 200 (Appendix Table 7), and extrapolating from a subsample to the full training set beats an n-independent guess on Replogle (mean absolute error 0.03 against 0.10) but not on the smaller screens (Appendix Table 8). The predictions are expectations: per fold they miss by 0.12 on average, as nested cross-validation and leave-one-out do, and leave-one-out is less biased (Appendix Table 9). They are also less accurate where little signal remains: without the top decile of response energy, they are optimistic by 0.06 to 0.12 in ε (Appendix Table 12).

Prediction (P3) is the mechanism control (Figure 2b). The message of another perturbation raises the error of the plug-in Bayes decoder 0.86 to 1.16 times as much as the correct message lowers it, except on the Norman singles, whose 81 training perturbations estimate the decoder poorly (slope 1.52); had the gain come from variation shared across perturbations, such as systematic or batch effects, another

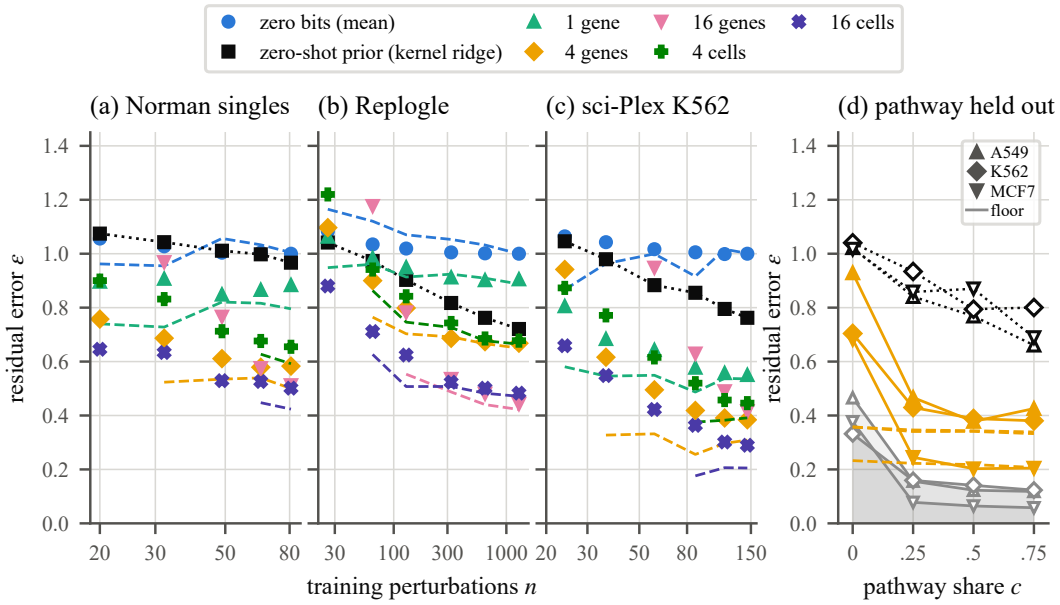


Figure 3: **Data-limited, information-limited and coverage-limited error.** (a to c) Residual error against the number of training perturbations n (markers: observed; dashed: Theorem 3 where $n \geq 8q$; dotted: zero-shot prior). (d) One sci-Plex pathway held out at a time with a fraction c of its compounds moved into training: coverage floor of Proposition 1 (grey), observed four-gene error (filled), Equation (2) (dashed) and best prior (hollow).

perturbation's message would have kept much of it. The least-squares decoder, which overfits at large k, follows a first-order finite-sample form instead (Appendix A; slope 0.91, $R^2 = 0.98$).

5.3 A FEW MEASURED GENES BEAT EVERY TESTED PRIOR, BY A MARGIN THAT DEPENDS ON THE CELLS BEHIND THEM

One greedily chosen gene of the held-out perturbation lowers ε to 0.37 to 0.57 on the chemical screens and 0.91 on Replogle, and four genes to 0.25 to 0.67 (0.59 on the Norman singles; Table 2). Four measured genes beat the best prior on every screen, with paired 95% bootstrap intervals that exclude zero, also without the histone deacetylase inhibitors and on 58% to 87% of individual perturbations (Appendix Tables 22 and 24); the genes do not merely predict themselves (excluding them from the target raises ε by 0.03 on average, about 0.1 at most; Appendix Table 23). The margin depends on the cells behind each gene: read from 5 to 20 cells instead of the pseudobulk, four genes still beat the prior on four screens but not on Replogle (Appendix Table 18). Priors and genes are largely redundant on sci-Plex and on the Norman singles, where adding the prior's prediction to four genes changes ε by at most 0.02, but complementary on Replogle, where Gene Ontology and four genes together reach 0.53 against 0.68 and 0.71 alone (Appendix Table 31). The RPE1 response of the same target, the cross-screen baseline of Ahlmann-Eltze et al. (2025), measures the perturbation in another context: on the 1,036 Replogle perturbations it covers, it reaches $\varepsilon = 0.66$ at the $q = 64$ that Equation (2) predicts best (predicted 0.62), worth 4.7 greedy genes against 2.4 for Gene Ontology (Appendix Table 20). A compound's response in another cell line, reduced to its leading coordinate, reaches 0.36 to 0.57, and the wrong compound's response 1.46 to 1.61, close to the predicted $2 - \varepsilon$ (Appendix Table 21).

We test (P4) beyond the original budget of 32 cells (Appendix Table 17). The raw mean of m cells overtakes the mean response between 16 and 24 cells on Norman (predicted $m_\times = 23$), 32 and 48 on Replogle (41) and 32 and 48 on sci-Plex MCF7 (41); on the 39 K562 compounds with at least 128 cells, where the estimated batch noise of the message exceeds the signal, it never does, as predicted, though its error is underpredicted by 24%.

5.4 Training size, coverage, and where the account breaks

In Figure 3a to c, the zero-bit error is flat in the number of training perturbations n, as Theorem 3 predicts; on the Norman singles the zero-shot prior also stays at the zero-bit error, whereas on Replogle and sci-Plex K562 it falls to 0.72 and 0.76 and is still falling at the largest n: these priors are data-limited as well as information-limited, yet they trail four measured genes. Errors of predictors that receive a message fall towards their information limit, and the best message size grows with n: sixteen genes overtake four by 81 (Norman singles) and 129 (Replogle) training perturbations, but not within the 145 compounds of sci-Plex K562.

Figure 3d tests (P5). Under random folds, 7% to 13% of the held-out signal energy lies outside the span of the training responses (the coverage floor of Proposition 1); with the pathway held out, it is 33% to 47%. No predictor that combines training responses can cross this floor: the fingerprint and annotation priors reach 1.02 to 1.04 and four measured genes 0.68 to 0.93, whereas Equation (2), which cannot see the floor, predicts 0.23 to 0.36. Moving a quarter of the pathway into training lowers the floor to 8% to 16% and restores four genes to 0.24 to 0.47, with optimism 0.02 to 0.13; the priors recover more slowly (best prior 0.84 to 0.93; Appendix Tables 26 and 27). The leverage of the four-gene message does not flag the shift (8% to 12% of held-out perturbations above the training 95th percentile, against 4% to 8% under random folds).

6 Discussion and limitations

What is learned. The universal statement is Theorem 1: without perturbation-specific information beyond the training set, no decoder beats the mean response by more than a factor $1 + 1/n$. The empirical one is narrower: with the side information and decoder families we tested, extra capacity recovered little, informative priors mostly recognized programs already in training (essential-gene complexes, one class of inhibitors), and a small measurement of the new perturbation supplied what was missing. For a covered class, four well-chosen genes read from the pseudobulk recovered a third to three quarters of the recoverable signal, against 7% to 30% for the best prior, at an expected value predictable from a few hundred training perturbations. The first members of a new class should be profiled in full; their coverage floor (Proposition 1) then shows whether the class is covered, and in our hold-out a quarter of a pathway sufficed.

Operating envelope. The account is reliable with a few hundred training perturbations and $n/q \geq 8$, for strong responders, and inside the programs that training covers. It fails, in the direction the theory predicts, when a pathway is absent (the floor rises and Equation (2) turns optimistic; the message's leverage does not flag it), below 100 training perturbations (optimism near 0.1), where little signal remains, and for plug-in decoders with many estimated parameters.

Limitations. Our theory is linear and its finite-sample part Gaussian; greedy panels are tested as an approximation. The capacity statement concerns the decoder families tested: we did not fine-tune single-cell foundation models, run chemCPA, or train GEARS at its full budget. Gene panels are simulated from whole-transcriptome pseudobulks, so their value depends on the cells behind each gene, and we do not model cost: a targeted readout lowers sequencing cost but still requires running the perturbation. Below one gene, the measurement equivalent is an interpolated ratio. The error ε is energy-weighted, so the chemical results rest largely on the strongest responders (per-perturbation comparisons agree), all genetic screens are in K562, and the compounds come from one library. Class-mean priors fail under pathway hold-out by construction.

7 Conclusion

Predicting an unseen perturbation is a decoding problem: without perturbation-specific information beyond the training set no decoder beats the mean response, and on the screens and decoder families we tested, capacity did not substitute for that information. For perturbations within response programs represented in training, what a small gene, cell or cross-context measurement buys can be predicted from training data when training is large enough, and a prior can be priced against greedily chosen genes. Outside those programs a coverage floor, which no combination of training responses can cross, marks where measurements, priors and error predictions fail together.

AI USE STATEMENT

In this paper, AI helped to aid or polish the writing, review the draft, act as coding agents, take part in discussions and generate plots; it also helped to create slides for internal meetings. Specifically, large-language-model chat and coding assistants took part in the discussions that shaped the experimental design; helped formulate the formal statements; implemented the data processing, decoders and baselines, ran the experiments and produced the plots; and drafted, polished and reviewed parts of the text and searched for and formatted references. They were not used to generate or modify data. Theorem 3 was verified by Monte Carlo simulation (Appendix A); every number in the paper was traced to the code output that produced it, every reference was verified against its publisher, proceedings or preprint record, and the code was tested (for example, that every predicted quantity uses training perturbations only). The authors take full responsibility for the content, including text, claims, code and figures produced with the aid of generative AI.

ETHICS STATEMENT

This work analyses publicly released single-cell perturbation screens of human cancer cell lines; it involves no human participants, no patient data and no new experiments. The measurement designs we study are intended to reduce the experimental cost of characterising perturbations; they are research tools and not clinical recommendations. We see no specific risk of harm beyond the general dual-use considerations of perturbation biology.

REPRODUCIBILITY STATEMENT

Section 4 specifies the datasets, gene selection, batch-disjoint halves, cross-validation folds, channels, decoders and baselines, and Appendix B lists every remaining setting (preprocessing, covariance and noise estimation, panel selection, model grids and their controls, bootstrap, stress-test, coverage and cross-screen protocols, random seeds and compute). Appendix A contains complete proofs of all results and Appendix C the sensitivity analyses. All datasets are public: the harmonized screens of Li et al. (2026), the Gene Ontology annotations (Ashburner et al., 2000) and the GenePT embeddings (Chen & Zou, 2025).

REFERENCES

Charlotte Bunne, Yusuf Roohani, Yanay Rosen, Ankit Gupta, Xikun Zhang, Marcel Roed, Theo Alexandrov, Mohammed AlQuraishi, Patricia Brennan, Daniel B. Burkhardt, et al. How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell*, 187(25):7045–7063, 2024. doi: 10.1016/j.cell.2024.11.015.

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, et al. Virtual Cell Challenge: Toward a Turing test for the virtual cell. *Cell*, 188(13):3370–3374, 2025b. doi: 10.1016/j.cell.2025.06.008.

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.

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.

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 38 (Datasets and Benchmarks Track)*, 2025.

Lanxiang Li, Yue You, Yunlin Fu, Wenyu Liao, Xueying Fan, Shihong Lu, Ye Cao, Bo Li, Wenle Ren, Jiaming Kong, et al. A Systematic Comparison of Single-Cell Perturbation Response Prediction Models. *Science Advances*, 12(37):eaed3414, 2026. doi: 10.1126/sciadv.aed3414.

Ramon Viñas Torné, Maciej Wiatrak, Zoe Piran, Shuyang Fan, Liangze Jiang, Sarah A. Teichmann, Mor Nitzan, and Maria Brbić. Systema: a framework for evaluating genetic perturbation response prediction beyond systematic variation. *Nature Biotechnology*, 44(6):1050–1059, 2026. doi: 10.1038/s41587-025-02777-8.

Yongqi Huang, Hanzhi Wang, and Parker Wilson. Virtual-cell models compress unseen intervention geometry through a target-specific generalization bottleneck. *bioRxiv*, 2026. doi: 10.64898/2026.08.21.746243.

Aravind Subramanian, Rajiv Narayan, Steven M. Corsello, David D. Peck, Ted E. Natoli, Xiaodong Lu, Joshua Gould, John F. Davis, Andrew A. Tubelli, Jacob K. Asiedu, et al. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell*, 171(6):1437–1452.e17, 2017. doi: 10.1016/j.cell.2017.10.049.

Daniel Schraivogel, Andreas R. Gschwind, Jennifer H. Milbank, Daniel R. Leonce, Petra Jakob, Lukas Mathur, Jan O. Korbel, Christoph A. Merten, Lars Velten, and Lars M. Steinmetz. Targeted Perturb-seq enables genome-scale genetic screens in single cells. *Nature Methods*, 17(6):629–635, 2020. doi: 10.1038/s41592-020-0837-5.

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.

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.

Leon Hetzel, Simon Boehm, Niki Kilbertus, Stephan Günnemann, Mohammad Lotfollahi, and Fabian J. Theis. Predicting cellular responses to novel drug perturbations at a single-cell resolution. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Michael Bereket and Theofanis Karaletsos. Modelling cellular perturbations with the sparse additive mechanism shift variational autoencoder. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Charlotte Bunne, Stefan G. Stark, Gabriele Gut, Jacobo Sarabia del Castillo, Mitch Levesque, Kjong-Van Lehmann, Lucas Pelkmans, Andreas Krause, and Gunnar Rätsch. Learning single-cell perturbation responses using neural optimal transport. *Nature Methods*, 20(11):1759–1768, 2023. doi: 10.1038/s41592-023-01969-x.

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.

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.

Abhinav K. Adduri, Dhruv Gautam, Beatrice Bevilacqua, Mohsen Naghipourfar, Alishba Imran, Rohan Shah, Noam Teyssier, Rishi Verma, Christopher Carpenter, Basak Eraslan, et al. Predicting cellular responses to perturbation across diverse contexts with State. *Cell*, 189(19):5914–5931.e20, 2026. doi: 10.1016/j.cell.2026.07.052.

Chenglei Yu, Chuanrui Wang, Bangyan Liao, and Tailin Wu. scDFM: Distributional flow matching model for robust single-cell perturbation prediction. In *International Conference on Learning Representations*, 2026.

Yiqun Chen and James Zou. Simple and effective embedding model for single-cell biology built from ChatGPT. *Nature Biomedical Engineering*, 9(4):483–493, 2025. doi: 10.1038/s41551-024-01284-6.

Alexis Molina and Xinyi Zhang. Perturbation response decomposition enables biologically aligned generalization to unseen perturbations and cellular contexts. *bioRxiv*, 2026. doi: 10.64898/2026.07.24.740459.

Alsu Missarova, Jaison Jain, Andrew Butler, Shila Ghazanfar, Tim Stuart, Maigan Brusko, Clive Wasserfall, Harry Nick, Todd Brusko, Mark Atkinson, Rahul Satija, and John C. Marioni. geneBasis: an iterative approach for unsupervised selection of targeted gene panels from scRNA-seq. *Genome Biology*, 22(1):333, 2021. doi: 10.1186/s13059-021-02548-z.

Ian Covert, Rohan Gala, Tim Wang, Karel Svoboda, Uygar Sümbül, and Su-In Lee. Predictive and robust gene selection for spatial transcriptomics. *Nature Communications*, 14(1):2091, 2023. doi: 10.1038/s41467-023-37392-1.

Brian Cleary, Le Cong, Anthea Cheung, Eric S. Lander, and Aviv Regev. Efficient Generation of Transcriptomic Profiles by Random Composite Measurements. *Cell*, 171(6):1424–1436.e18, 2017. doi: 10.1016/j.cell.2017.10.023.

Douglas Yao, Loic Binan, Jon Bezney, Brooke Simonton, Jahanara Freedman, Chris J. Frangieh, Kushal Dey, Kathryn Geiger-Schuller, Basak Eraslan, Alexander Gusev, Aviv Regev, and Brian Cleary. Scalable genetic screening for regulatory circuits using compressed Perturb-seq. *Nature Biotechnology*, 42(8):1282–1295, 2024. doi: 10.1038/s41587-023-01964-9.

Kathryn Chaloner and Isabella Verdinelli. Bayesian Experimental Design: A Review. *Statistical Science*, 10(3):273–304, 1995. doi: 10.1214/ss/1177009939.

Arash Mehrjou, Ashkan Soleymani, Andrew Jesson, Pascal Notin, Yarin Gal, Stefan Bauer, and Patrick Schwab. GeneDisco: A benchmark for experimental design in drug discovery. In *International Conference on Learning Representations*, 2022.

Kexin Huang, Romain Lopez, Jan-Christian Hütter, Takamasa Kudo, Antonio Rios, and Aviv Regev. Sequential optimal experimental design of perturbation screens guided by multi-modal priors. In *Research in Computational Molecular Biology (RECOMB 2024)*, volume 14758 of *Lecture Notes in Computer Science*, pp. 17–37. Springer, 2024. doi: 10.1007/978-1-0716-3989-4_2.

Yusuf Roohani, Andrew Lee, Qian Huang, Jian Vora, Zachary Steinhart, Kexin Huang, Alexander Marson, Percy Liang, and Jure Leskovec. BioDiscoveryAgent: An AI agent for designing genetic perturbation experiments. In *International Conference on Learning Representations*, 2025a.

Claude E. Shannon. Coding theorems for a discrete source with a fidelity criterion. In *IRE National Convention Record, Part 4*, volume 7, pp. 142–163, 1959.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, second edition, 2006. doi: 10.1002/047174882X.

Xi Chen, Adityanand Guntuboyina, and Yuchen Zhang. On Bayes risk lower bounds. *Journal of Machine Learning Research*, 17(218):1–58, 2016.

Aolin Xu and Maxim Raginsky. Minimum Excess Risk in Bayesian Learning. *IEEE Transactions on Information Theory*, 68(12):7935–7955, 2022. doi: 10.1109/TIT.2022.3176056.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. In *International Conference on Learning Representations*, 2020.

Andreas Krause, Ajit Singh, and Carlos Guestrin. Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9:235–284, 2008.

Abhimanyu Das and David Kempe. Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 1057–1064, 2011.

L. Breiman and D. Freedman. How Many Variables Should Be Entered in a Regression Equation? *Journal of the American Statistical Association*, 78(381):131–136, 1983. doi: 10.1080/01621459.1983.10477941.

Thomas M. Norman, Max A. Horlbeck, Joseph M. Replogle, Alex Y. Ge, Albert Xu, Marco Jost, Luke A. Gilbert, and Jonathan S. Weissman. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, 365(6455):786–793, 2019. doi: 10.1126/science.aax4438.

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, et al. 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.

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.

Britt Adamson, Thomas M. Norman, Marco Jost, Min Y. Cho, James K. Nuñez, Yuwen Chen, Jacqueline E. Villalta, Luke A. Gilbert, Max A. Horlbeck, Marco Y. Hein, et al. A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell*, 167(7):1867–1882.e21, 2016. doi: 10.1016/j.cell.2016.11.048.

Sanjay R. Srivatsan, José L. McFaline-Figueroa, Vijay Ramani, Lauren Saunders, Junyue Cao, Jonathan Packer, Hannah A. Pliner, Dana L. Jackson, Riza M. Daza, Lena Christiansen, et al. Massively multiplex chemical transcriptomics at single-cell resolution. *Science*, 367(6473):45–51, 2020. doi: 10.1126/science.aax6234.

Matan Gavish and David L. Donoho. The optimal hard threshold for singular values is $4/\sqrt{3}$. *IEEE Transactions on Information Theory*, 60(8):5040–5053, 2014. doi: 10.1109/TIT.2014.2323359.

Andreas Krause and Carlos Guestrin. Near-optimal nonmyopic value of information in graphical models. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, pp. 324–331, 2005.

G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions-I. *Mathematical Programming*, 14(1):265–294, 1978. doi: 10.1007/BF01588971.

Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, et al. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000. doi: 10.1038/75556.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574.

Walid Ahmad, Elana Simon, Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. ChemBERTa-2: Towards chemical foundation models. *arXiv preprint arXiv:2209.01712*, 2022.

A PROOFS

Table 3: Notation.

Symbol	Meaning
p, n, G	perturbation; number of training perturbations; number of modelled genes
Θ, Π, μ_Θ	population law; its distribution; mean response under Θ
R_p, $Z_p = R_p - \mu_\Theta$	response (expected expression change); perturbation-specific residual
$T_p = R_p + \xi_p$, N_B	target from half B; covariance of its noise ξ_p (cells and batch)
R^A, R^B, $\hat{\Sigma}$	half-A and half-B responses; their cross-covariance across training perturbations
S_p, $\mathcal{D}_n$	side information (a descriptor of p and the training perturbations)
M_p, y	message from half A, another cell line or another screen; its vector form
Σ, λ_i, Σ_Θ	covariance of Z (given the law); its eigenvalues; $\mathrm{Cov}(Z \mid \Theta)$
H, Σ_e, C, K	channel matrix; message noise covariance; $H\Sigma H^\top + \Sigma_e$; $\Sigma H^\top C^{-1}$
Σ_c, N_{msg}	cell-level covariance; batch noise of a message
k, m, q	measured genes; measured cells; message dimension of the regression decoder
$D(f)$, D_0	distortion of f; distortion of the mean response of training perturbations
$F(S)$, $D^\star(H)$	floor $\mathbb{E}\,\mathrm{tr}\,\mathrm{Cov}(Z \mid S)$; risk of the best linear decoder
L_q, RSS, $D_q^{(n)}$	best linear risk on the target; training residual sum of squares; held-out distortion
$\mathcal{D}_{\mathrm{sel}}$, $\mathcal{A}_n$, d_0^2	perturbations that select a design; affine hull of training targets; coverage estimate
ε, r	normalized residual error, Equation (1); residual correlation (cosine similarity)
$\mathcal{R}_{Z\mid S}(D)$, $m_\times$	rate-distortion function of Z given S; crossover cell count
$I_{\mathrm{G}}(H)$, $D_{\mathrm{WF}}(\rho)$	(Gaussian-equivalent) information of a channel; water-filling distortion
c	fraction of a held-out pathway moved into training
η	signal recovered by a predictor relative to the first greedy gene (Definition 1)
$\bar{T}$, $P_\perp$	mean training target; projection off the span of the centred training targets

Lemma 1 (Observable target)**.** *Under Assumption 1, for any predictor $f = f(S, M)$, $\mathbb{E}\|T - f\|^2 = \mathbb{E}\|R - f\|^2 + \mathrm{tr}\, N_B$.*

Proof. $T - f = (R - f) + \xi$, and $\mathbb{E}[(R-f)^\top \xi] = \mathbb{E}[(R-f)^\top \mathbb{E}(\xi \mid R, S, M)] = 0$ because $R - f$ is a function of (R, S, M) and $\mathbb{E}[\xi \mid R, S, M] = \mathbb{E}\big[\mathbb{E}[\xi \mid \Theta, R, S, M] \mid R, S, M\big] = 0$. Finally $\mathbb{E}\|\xi\|^2 = \mathrm{tr}\, N_B$. □

Proof of Theorem 1. All expectations are over $\Theta \sim \Pi$, the training perturbations and the held-out perturbation. Since $\hat{Z} = g(S, M, U)$ with U independent of (Z, S, M), the data-processing inequality conditional on S gives $I(Z; \hat{Z} \mid S) \le I(Z; M, U \mid S) = I(Z; M \mid S)$. The conditional law $P_{\hat{Z}\mid Z,S}$ is feasible in the definition of $\mathcal{R}_{Z\mid S}(D)$ because its distortion is at most D, hence $I(Z; \hat{Z} \mid S) \ge \mathcal{R}_{Z\mid S}(D)$. If $I(Z; M \mid S) = 0$, then $I(Z; \hat{Z} \mid S) = 0$, i.e. $\hat{Z} \perp Z \mid S$. Writing $Z - \hat{Z} = (Z - \mathbb{E}[Z \mid S]) + (\mathbb{E}[Z \mid S] - \hat{Z})$, the cross term vanishes because, given S, the first factor has mean zero and is independent of the second; hence $\mathbb{E}\|Z - \hat{Z}\|^2 = F(S) + \mathbb{E}\|\mathbb{E}[Z \mid S] - \hat{Z}\|^2$. Now let the descriptor and M be conditionally independent of (Θ, Z) given $\mathcal{D}_n$ (they carry no information about the law or the residual beyond the training set). Then $\mathbb{E}[\mu_\Theta \mid S, M] = \mathbb{E}[\mu_\Theta \mid \mathcal{D}_n]$ and $\mathbb{E}[Z \mid S, M] = \mathbb{E}[Z \mid \mathcal{D}_n] = \mathbb{E}\big[\mathbb{E}[Z \mid \Theta, \mathcal{D}_n] \mid \mathcal{D}_n\big] = 0$, because given Θ the held-out perturbation is independent of $\mathcal{D}_n$ and $\mathbb{E}[Z \mid \Theta] = 0$. Since $\mathbb{E}[\xi \mid R, S, M] = 0$, the Bayes predictor of the target $T = \mu_\Theta + Z + \xi$ is $\mathbb{E}[\mu_\Theta \mid \mathcal{D}_n]$, and its distortion is $\mathbb{E}\,\mathrm{tr}\,\Sigma_\Theta + \mathbb{E}\,\mathrm{tr}\,\mathrm{Cov}(\mu_\Theta \mid \mathcal{D}_n) + \mathrm{tr}\, N_B \ge \mathbb{E}\,\mathrm{tr}\,\Sigma_\Theta + \mathrm{tr}\, N_B$ (the cross terms vanish by Lemma 1 and because, given Θ, Z has mean zero and is independent of $\mathcal{D}_n$). Given Θ, the training targets $T_i = \mu_\Theta + Z_i + \xi_i$ and the held-out target are i.i.d. (Assumption 1) with $\mathbb{E}[\xi_i \mid \Theta, R_i] = 0$ (Assumption 1), so the mean $\bar{T}_n$ of the n training targets has $\mathbb{E}[\|T - \bar{T}_n\|^2 \mid \Theta] = (1 + 1/n)\big(\mathrm{tr}\,\Sigma_\Theta + \mathrm{tr}\,\mathrm{Cov}(\xi \mid \Theta)\big)$. Its distortion $(1 + 1/n)(\mathbb{E}\,\mathrm{tr}\,\Sigma_\Theta + \mathrm{tr}\, N_B)$ is therefore at most $1 + 1/n$ times the Bayes distortion whenever $\mathbb{E}\,\mathrm{tr}\,\Sigma_\Theta$ and $\mathrm{tr}\, N_B$ are finite. □

Proof of Theorem 2. $\Sigma_e \succ 0$ makes $C \succ 0$, so C^{-1} exists. (If the message noise were singular, the same statements hold with the Moore-Penrose inverse of C on its range.) (a) For any distribution with covariances Σ and Σ_e, the linear decoder Ky has risk $\mathrm{tr}\,\Sigma - 2\,\mathrm{tr}(KH\Sigma) + \mathrm{tr}(KCK^\top)$, which is minimized at $K = \Sigma H^\top C^{-1}$ (the normal equations) with risk $D^\star(H)$. For jointly Gaussian (Z, y), $\mathbb{E}[Z \mid y] = Ky$, so K is the Bayes decoder. (b) The raw mean of m cells is $\bar{y} = Z + \bar{e} + b$,

where $\bar{e}$ is the average cell noise, with covariance Σ_c/m, and b the batch noise of the message, with covariance N_{msg}, both independent of Z. Its error $\bar{e}+b$ has trace covariance $\operatorname{tr}\Sigma_c/m+\operatorname{tr}N_{\mathrm{msg}}$, and the zero-bit predictor has risk $\operatorname{tr}\Sigma$. The raw mean is better iff $\operatorname{tr}\Sigma_c/m < \operatorname{tr}\Sigma - \operatorname{tr}N_{\mathrm{msg}}$, which requires $\operatorname{tr}N_{\mathrm{msg}} < \operatorname{tr}\Sigma$ and is then equivalent to $m > m_\times$. (c) Let y' be the message of an independent perturbation, with the same second moments. Then $\mathbb{E}\|Z-Ky'\|^2 = \operatorname{tr}\Sigma + \operatorname{tr}(KCK^\top) - 2\operatorname{tr}(K\operatorname{Cov}(y',Z))$ and the last term is zero. Since $KCK^\top = \Sigma H^\top C^{-1}H\Sigma$, $\operatorname{tr}(KCK^\top) = \operatorname{tr}\Sigma - D^\star(H)$, so the risk equals $2\operatorname{tr}\Sigma - D^\star(H)$. □

Proposition 2 (Information of a linear channel and an envelope)**.** *Under the conditions of Theorem 2, let* $I_{\mathrm{G}}(H) = \frac{1}{2}\log\det(I+\Sigma_e^{-1/2}H\Sigma H^\top\Sigma_e^{-1/2})$ *nats, the mutual information* $I(Z;y)$ *if* (Z,e) *is Gaussian and the Gaussian-equivalent information of the channel otherwise. Then* $D^\star(H) \geq D_{\mathrm{WF}}(I_{\mathrm{G}}(H))$*, where* $D_{\mathrm{WF}}(\rho) = \sum_i \min(\lambda_i,\theta)$ *and* θ *solves* $\frac{1}{2}\sum_i \log(\lambda_i/\theta)_+ = \rho$ *(reverse water-filling). If* (Z,e) *is Gaussian the bound holds for every decoder; otherwise it bounds linear decoders. It is attained by a channel that reveals the leading eigen-coordinates of* Σ *with noise variances* $\lambda_i\theta/(\lambda_i-\theta)$*.*

Proof. For Gaussian (Z,e), $I(Z;y) = h(y) - h(y \mid Z) = \frac{1}{2}\log\det C - \frac{1}{2}\log\det\Sigma_e = I_{\mathrm{G}}(H)$. Theorem 1 with trivial S and $M=y$ shows that any decoder achieving D needs $I(Z;y) \geq \mathcal{R}_Z(D)$; the rate-distortion function of a Gaussian vector is given by reverse water-filling on the eigenvalues of Σ (Cover & Thomas, 2006, Thm. 10.3.3), and D_{WF} is its inverse, so every decoder satisfies $D \geq D_{\mathrm{WF}}(I(Z;y))$. $D^\star(H)$ and $I_{\mathrm{G}}(H)$ depend only on Σ, H and Σ_e, so for other laws the inequality holds between the risk of the best linear decoder and the Gaussian-equivalent information; it bounds nonlinear decoders only under Gaussianity, since non-Gaussian sources and channels can have a smaller rate-distortion function. For the attaining channel $y_i = u_i^\top Z + e_i$ with $\operatorname{Var}(e_i) = \lambda_i\theta/(\lambda_i-\theta)$ for $\lambda_i > \theta$, the coordinates are independent, the posterior variance of coordinate i is θ and its information is $\frac{1}{2}\log(\lambda_i/\theta)$. Figure 4 evaluates both sides with the same covariance estimate, in bits. □

Proof of Theorem 3. Treat one output coordinate t (the argument applies to each of the G genes and the risks add). With $\mu_y = \mathbb{E}y$ and $\Sigma_y = \operatorname{Cov}(y) \succ 0$, joint Gaussianity gives $t = \alpha + \beta^\top(y-\mu_y) + \zeta$ with $\zeta \sim \mathcal{N}(0,\sigma^2)$ independent of y, and σ^2 is the risk of the best linear decoder for this coordinate; summed over coordinates it is L_q. Least squares with intercept on $(y_i,t_i)_{i\leq n}$ gives $\hat{t}_0 = \bar{t} + \hat{\beta}^\top(y_0-\bar{y})$ and

$$t_0 - \hat{t}_0 = \zeta_0 - \bar{\zeta} - (\hat{\beta}-\beta)^\top(y_0-\bar{y}).$$

Conditionally on the training messages (y_i), $\hat{\beta}-\beta \sim \mathcal{N}(0,\sigma^2W^{-1})$ with $W = \sum_i (y_i-\bar{y})(y_i-\bar{y})^\top$, $\bar{\zeta} \sim \mathcal{N}(0,\sigma^2/n)$ is independent of $\hat{\beta}$, and ζ_0 is independent of everything. Hence $\mathbb{E}(t_0-\hat{t}_0)^2 = \sigma^2(1+1/n) + \sigma^2\,\mathbb{E}[(y_0-\bar{y})^\top W^{-1}(y_0-\bar{y})]$. For Gaussian samples $\bar{y}$ and W are independent, $y_0-\bar{y} \sim \mathcal{N}(0,(1+1/n)\Sigma_y)$ is independent of W, and $W \sim \mathcal{W}_q(\Sigma_y, n-1)$ has $\mathbb{E}W^{-1} = \Sigma_y^{-1}/(n-q-2)$ for $n > q+2$. The last expectation is therefore $(1+1/n)\,q/(n-q-2)$ and $\mathbb{E}(t_0-\hat{t}_0)^2 = \sigma^2(1+1/n)(n-2)/(n-q-2)$. The residual sum of squares satisfies $\mathrm{RSS} \sim \sigma^2\chi^2_{n-q-1}$, so $\mathrm{RSS}/(n-q-1)$ is unbiased for σ^2. Summing over coordinates gives Equation (2); with $q=0$ it reduces to $L_0(1+1/n)$. The argument conditions on a design fixed before the responses are seen. When the design is chosen on the same training responses, RSS tends to be biased downwards, and heavy tails can inflate $\mathbb{E}[(y_0-\bar{y})^\top W^{-1}(y_0-\bar{y})]$ beyond its Gaussian value; both make the prediction optimistic. A Monte Carlo check with Gaussian (y,T), n from 20 to 400 and q from 0 to 32 (20,000 draws each) reproduces Equation (2) within 1.5% (within simulation error) and the mean RSS-based prediction within 0.2%. □

Proof of Corollary 1. Condition on $\mathcal{D}_{\mathrm{sel}}$. The design is then a fixed function of $\mathcal{D}_{\mathrm{sel}}$, and the fitting and held-out perturbations are independent of $\mathcal{D}_{\mathrm{sel}}$, so their conditional law is the population law and the design does not depend on their responses. Theorem 3 applies to this conditional problem, which gives Equation (2) conditionally on $\mathcal{D}_{\mathrm{sel}}$, and therefore also on average over it. A random design is independent of all responses. □

Proof of Proposition 1. If $\sum_i w_i = 1$, then $f = \bar{T} + \sum_i w_i(T_i-\bar{T})$ lies in $\mathcal{A}_n = \bar{T} + \operatorname{span}\{T_i-\bar{T}\}$, so $\|R_0-f\|^2 \geq \min_{a\in\mathcal{A}_n}\|R_0-a\|^2 = \|P_\perp(R_0-\bar{T})\|^2$. The regression decoder qualifies: with X_c the centred training messages, its prediction is $\bar{T} + (y_0-\bar{y})^\top(X_c^\top X_c)^{-1}X_c^\top T_c = \bar{T} + \sum_i w_i(T_i-\bar{T})$

with $w = X_c(X_c^\top X_c)^{-1}(y_0 - \bar{y})$, whose entries sum to zero; ridge and kernel ridge predictions are $\bar{T} + \sum_i \beta_i(T_i - \bar{T})$ in the same way, and nearest-neighbour and class averages are convex combinations. (Predictors trained on full-pseudobulk targets combine those instead, and the bound holds for their hull.) For the estimate, $R_0^A = R_0 + \xi_0^A$ and $R_0^B = R_0 + \xi_0^B$ with independent zero-mean noises that are independent of the training data, so $\mathbb{E}\langle P_\perp(R_0^A - \bar{T}), P_\perp(R_0^B - \bar{T})\rangle = \|P_\perp(R_0 - \bar{T})\|^2$ given the training data. □

Misattributed messages for the regression decoder (first-order approximation). If the decoder of Theorem 3 receives the message y' of another, independent perturbation, its error is $(T_0 - \bar{T}) - \hat{\beta}^\top(y' - \bar{y})$. The two terms are uncorrelated up to $O(1/n)$, so the expected distortion is $(1 + 1/n)\big(\operatorname{tr}\widehat{\operatorname{Cov}}(T) + \operatorname{tr}(\hat{\beta}^\top\widehat{\operatorname{Cov}}(y)\hat{\beta})\big)$ to first order: the total variance of the target plus the variance the decoder explains on the training perturbations. This is the finite-sample counterpart of Theorem 2c, a heuristic rather than an identity. It equals twice the zero-bit risk minus the aligned risk only when the decoder is well estimated; an overfitted decoder explains more variance on the training perturbations than it recovers on held-out ones, so its misattribution penalty exceeds its gain. Against the observed errors of misattributed least-squares decoders (all screens, designs and panel sizes), this form fits with slope 0.91 and $R^2 = 0.98$, with a median relative error of 0.2% to 3.4% per screen.

B Experimental details

Data and preprocessing. All screens were taken from the harmonized AnnData files of Li et al. (2026), which store log-normalized expression ($\log(1 + 10^4\, c/\sum c)$). For Norman we used all 226 conditions with at least 30 cells (97k cells, 2,000 control cells, 5,025 genes); the 10x lane of every cell is encoded in its barcode suffix (lanes 1 to 8, every condition present in all lanes). Half A collects lanes in a fixed order until it holds about half of a condition's cells (lanes 1 to 4 for 138 conditions, 1 to 5 for 87, 1 to 3 for one), and half B the remaining lanes. For Replogle (K562, essential genes) we used the 1,617 perturbations with at least 30 cells (261k cells; 2,000 control cells) and split each perturbation's 10x GEM groups (27 to 48 per perturbation) between the halves by the same rule, requiring at least 15 cells per half. For sci-Plex we used the 10 μM dose of each compound, the vehicle wells as controls, and the two replicate wells of each compound as halves; compounds with fewer than 15 cells in either replicate were removed. The gene universe consists of the (at most) 2,000 genes with the highest normalized dispersion among control cells within the expression window (mean log expression between 0.0125 and 3), after removing every gene targeted in the screen: 2,000 genes for Norman, Replogle and sci-Plex K562, 1,931 for A549 and 1,719 for MCF7. Pseudobulk responses are mean log-normalized expression minus the control mean. Compound classes (pathways, targets) come from the sci-Plex annotation; 24 to 27 compounds per line target histone deacetylases.

Covariance and noise estimation. Within each training fold, $\hat{\Sigma} = \frac{1}{2}(\Gamma_{AB} + \Gamma_{AB}^\top)$ with Γ_{AB} the cross-covariance of the half-A and half-B responses across training perturbations. Its trace and diagonal are used as is; the factor-model estimate used by the Bayes decoders and by greedy selection keeps the eigen-directions whose eigenvalue exceeds $|\lambda_{\min}(\hat{\Sigma})|$ and adds a non-negative diagonal remainder. The cell-sampling noise $\hat{\sigma}^2_{B,g}$ of gene g in half B is the average within-perturbation cell variance divided by the number of cells in the half. Batch noise enters in two places. The normalization of ε uses the aggregate estimate $\max\{\mathrm{TSS} - \operatorname{tr}\hat{\Sigma} - \sum_g \hat{\sigma}^2_{B,g}, 0\}$, where TSS is the total variance of the half-B responses across training perturbations (Table 34); summing per-gene clipped estimates would keep only positive fluctuations and bias the noise floor upwards. The Bayes, raw-mean and shrinkage decoders and $m_\times$ use the per-gene estimate $\max\{\widehat{\operatorname{Var}}(T_g) - \hat{\Sigma}_{gg} - \hat{\sigma}^2_{B,g}, 0\}$ in their noise model. For the Bayes few-cell decoder the cell-level covariance is the pooled within-perturbation covariance of the half-A cells of at most 300 training perturbations, shrunk by 5% towards its diagonal; the raw-mean, shrinkage and Bayes cell decoders average three draws of m cells per held-out perturbation. Table 13 replaces the aggregate batch-noise estimate by zero (cell sampling only), by the per-gene clipped sum, and by half and one and a half times its value. These noise models are simplifications that the theory takes as given: the channel noise is treated as independent of the response, although cell-sampling noise grows with expression; one pooled cell-level covariance serves every perturbation; and messages from perturbations with different numbers of cells (median 45 to 280 per half) are pooled by the regression decoder, which makes them heteroscedastic and so

part of the non-Gaussianity. We do not model these effects; they enter the calibration errors reported in Section 5.2 and the crossover brackets of Table 17.

Decoders and predictions. The regression decoder is ordinary least squares with intercept on the message; its predicted risk is Equation (2) with L_q replaced by $\mathrm{RSS}/(n-q-1)$. For cells the message is projected on the q leading eigenvectors of the training $\hat{\Sigma}$ ($q \in \{4, 8, 16\}$, main text $q = 8$); for cross-context messages, the compound's full pseudobulk response in the source line is projected on the q leading training eigenvectors of that line ($q \in \{1, 2, 4, 8, 16\}$, main text $q = 1$). For each of two independent draws of m half-A cells per perturbation the decoder is refit on the n training perturbations and its prediction recomputed, and both are averaged over draws. The plug-in Bayes decoder uses $K = \hat{\Sigma}H^\top(H\hat{\Sigma}H^\top + \hat{\Sigma}_e)^{-1}$ and predicts $\mathrm{tr}\,\hat{\Sigma} - \mathrm{tr}(KH\hat{\Sigma}) + \mathrm{tr}\,N_B$ plus the $O(1/n)$ mean-estimation term. Misattributed messages are obtained with a random derangement of the held-out perturbations within a fold. The RBF decoders of Table 33 use the kernel $\exp(-s\|z - z'\|^2)$ on standardized messages. Designs are classified for Figure 2a and Table 4 as random (ten random panels per size, independent of all responses: Theorem 3 applies), adaptive (greedy variance and information, top variance, top mean, stable genes, and the leading training directions used by the cell and cross-context channels, all chosen on the fitting perturbations), and Theorem 2 predictions (raw mean and per-gene shrinkage, which need no regression).

Zero-shot models. Nearest-neighbour predictors take the similarity-weighted average of the training responses of the 1, 5 or 20 most similar perturbations (cosine similarity; Tanimoto for fingerprints). Gene Ontology annotations come from the GOA human release of 20 May 2026; none of them cites the publication of the Replogle screen. Ridge and kernel ridge regression choose the regularization among 15 and 13 log-spaced values by exact leave-one-out error on the training perturbations. MLPs (GELU, dropout 0.1, AdamW with learning rate 10^{-3} and weight decay 10^{-4}, full batch, early stopping with patience 40 on 15% of the training perturbations, at most 400 epochs, output layer initialized at zero so that every network starts at the mean response) are run on the first repeat of the folds. The two controls of Table 30 use the default initialization of the output layer, on the best side information of each screen (Gene Ontology for the genetic screens, fingerprints for sci-Plex), once with the same early stopping and once trained on all training perturbations for 1,000 full-batch epochs without early stopping, after which the training loss had converged. GEARS (Roohani et al., 2024) (package version 0.1.2, hidden size 64, six epochs, batch size 64, learning rate 10^{-3}) was trained on the lane-disjoint Norman singles with at most 100 cells per perturbation, on the first five folds, a reduced budget for compute reasons (about 45 minutes per fold on our hardware); 10% of the training perturbations of each fold served as its validation set, whose error reached its minimum at epoch 4, 2, 3, 1 and 2 of six in the five folds, and GEARS keeps the model with the lowest validation error. Its co-expression graph was rebuilt from each fold's training perturbations. It is evaluated on the 100 of 101 perturbations whose gene is in its Gene Ontology graph, and its ε is normalized by the mean response on the same perturbations; the other zero-shot models use all 101.

Uncertainty and error prediction. Intervals are 95% cluster-bootstrap intervals (2,000 resamples of perturbations, all records of a perturbation moving together) of ratio-of-sums errors; the screen-clustered interval of the calibration slope resamples the five screens (the Norman singles are part of the Norman screen). The measurement equivalent interpolates the greedy panel's error linearly in $\log_2(1+k)$ between $k \in \{0, 1, 2, 4, \ldots, 32\}$ ($\varepsilon = 1$ at $k = 0$), on the folds on which the prior was evaluated, and stops where the panel's error curve turns up; Table 15 replaces the greedy panel by the other designs, and the interpolation by one linear in k or by a monotone cubic (PCHIP) in $\log_2(1+k)$. The error-prediction comparison (Table 9) uses the first repeat of the folds: nested cross-validation splits the training perturbations into five inner folds and re-selects the greedy panel in each; leave-one-out uses the hat matrix of the fitted decoder; split selection chooses the panel on a random half of the training perturbations and fits it on the other half (Corollary 1). The top-decile experiment removes, from each sci-Plex line, the 10% of compounds with the largest response energy $\|R_p - \bar{R}\|^2$ (17 to 19 compounds) and reruns the whole pipeline on the rest.

Crossover cell counts. For Table 17 the raw mean of m half-A cells is evaluated, with three draws per held-out perturbation, on the perturbations with at least 64 (Norman, sci-Plex MCF7) or 128 (Replogle, sci-Plex K562) stored half-A cells, on the first two repeats of the folds; $m_\times$, the predicted errors and the noise floor are computed from the training perturbations of the same subset.

Pathway hold-out and coverage. The pathway stress test holds out one of the 16 first-level pathways of the sci-Plex annotation at a time. For Table 26 and Figure 3d, a random fraction $c \in \{0.25, 0.5, 0.75\}$ of the held-out pathway's compounds (two random draws each; at least one compound stays held out) is moved into training, together with all other pathways. We evaluate the greedy one- and four-gene panels with the regression decoder and their training-only predictions, fingerprint kernel ridge and the second-level-pathway class mean; errors are ratios of sums over all held-out compounds and draws.

Coverage floor, leverage, and priors with genes. For Table 27 each fold's training targets (half-B responses of the training perturbations) are centred and their span is computed by a singular value decomposition; $P_\perp$ removes it, and the floor of a fold is $\sum_p d_p^2 / \sum_p (D_{0,p} - \operatorname{tr} N_B)$ over its held-out perturbations, with d_p^2 the split-half estimate of Proposition 1. The floor depends on the training size, since the hull has dimension at most $n - 1$ and would span the gene space as n approaches G; here n is below 180 against G of 1,719 to 2,000 genes, and the pathway hold-out trains on about as many compounds as the random folds (123 to 179 against 132 to 144), so its higher floor is not an effect of n. The leverage of a held-out four-gene message is $(y_p - \bar{y})^\top W^{-1} (y_p - \bar{y})$ with W the scatter matrix of the training messages. For Table 31, the prior (kernel ridge on the cosine or Tanimoto kernel, or ridge on Gene Ontology for Replogle, regularization chosen by leave-one-out) predicts the held-out perturbations from all training perturbations and each training perturbation by exact leave-one-out; the decoder of gene g regresses its half-B target on the k measured genes and on the prior's prediction of gene g.

Naive predictions, extrapolation and a standard metric. The naive predictions of Table 6 reuse the residual sum of squares of each fitted decoder. Decisions use all budgets of each screen. For Table 8, $\mathrm{RSS}/(n - q - 1)$ of a subsample is multiplied by the factor of Equation (2) at the full size. The Pearson correlations of Table 32 use the half-B response as the observed change from control and the full pseudobulk to rank the 20 genes with the largest change.

Cross-screen messages. The Replogle RPE1 essential-gene screen was taken from the same harmonized release as the K562 screen; its pseudobulk responses (log-normalized mean minus the control mean, perturbations with at least 30 cells) were computed on the K562 gene universe, and each K562 perturbation whose target was also perturbed in RPE1 receives the RPE1 response of that target as its message, projected on q leading directions of the training perturbations' RPE1 responses. The matched perturbations are split into two repeats of five folds, on which the Gene Ontology ridge prior and the greedy panels are evaluated on the same perturbations; the misattributed message is the RPE1 response of another held-out target.

Training-size experiment. For Figure 3a to c each training fold is subsampled at random to a fraction of its perturbations (Norman singles 0.25 to 1, two repeats of the five folds; Replogle 0.02 to 1, one repeat; sci-Plex K562 0.15 to 1, two repeats; 20 to 26 perturbations at the smallest fraction), the greedy panel is re-selected on the subsample, cells use one draw, and the zero-shot prior is kernel ridge on Gene Ontology (cosine kernel) or fingerprints (Tanimoto kernel). Errors are normalized by the zero-bit error and the per-gene noise floor of the full training sets.

Compute. All experiments ran on a laptop (Intel i7-10610U, 32 GB RAM, NVIDIA Quadro P520 with 2 GB); the complete pipeline takes about 20 hours, dominated by disk input/output and GEARS training. Random seeds are fixed for folds, panels, cell draws, bootstrap resamples and model initialization.

C Additional results

Calibration by screen and by design regime. Table 4 treats the five screens as the units of evidence, since the combinations of one screen share folds, covariance estimates and panels. Every screen is calibrated on its own, a bootstrap over the five screens (too few clusters for a precise interval) keeps the pooled slope between 0.94 and 0.99, and a linear recalibration fitted on four screens and applied to the fifth leaves a mean absolute difference of 0.02 to 0.04 on the fifth: the residual optimism is shared across screens rather than specific to one. Random panels, for which Equation (2) is exact

under Gaussianity (Corollary 1), are about as optimistic as adaptive designs; their slope is less precise because random panels span a narrow range of errors. The Norman singles alone, with 81 training perturbations, are the most optimistic subset (slope 0.94, bias +0.055), as the dependence on n in Section 5.2 predicts; leaving out the Norman (all) combinations, whose held-out doubles are not unseen perturbations, changes the pooled slope and R^2 to 0.96 and 0.98.

Table 4: Calibration of the training-only predictions ($n/q \geq 8$) per screen (target screen for cross-context messages), averaged over screens, with a screen-clustered bootstrap interval, and by design regime.

Combinations ($n/q \geq 8$)	number	slope	R^2	mean abs. diff.	bias
Norman	81	1.00	0.98	0.039	+0.038
singles	37	0.94	0.97	0.055	+0.055
all conditions	44	1.00	0.99	0.026	+0.023
Replogle	48	0.95	1.00	0.014	+0.014
sci-Plex A549	52	0.97	0.99	0.034	+0.033
sci-Plex K562	52	0.93	0.96	0.055	+0.053
sci-Plex MCF7	53	0.96	0.99	0.030	+0.028
mean over screens		0.96	0.98	0.035	+0.033
pooled (95% interval)	286	[0.94, 0.99]	[0.97, 0.99]	[0.025, 0.046]	[+0.024, +0.044]
random panels (Thm. 3)	30	0.84	0.96	0.036	+0.036
adaptive designs	216	0.97	0.98	0.040	+0.039
raw mean, shrinkage (Thm. 2)	40	1.04	1.00	0.011	+0.002

Calibration regimes. Table 5 repeats the fit of Figure 2a for several thresholds on the number of training perturbations per message dimension. The slope is stable, and the fit tightens as n/q grows because the estimation factor of Equation (2) approaches 1; below $n/q = 8$ the predictions are optimistic by 0.17 in ε on average. The points with $n/q = \infty$ are raw means and per-gene shrinkage, whose risks follow from Theorem 2 rather than Equation (2). The figure and the fits omit 33 combinations with $\varepsilon > 1.8$: 32 raw means of 1 to 32 cells, whose risk is predicted to within 7% (Table 19), and one stable-gene panel at $n/q < 8$. In the training-size runs of Figure 3a to c, with $n/q \geq 8$, the predictions of greedy panels are optimistic by 0.11 on average with fewer than 100 training perturbations (at most 0.29) and by 0.01 with more than 200 (at most 0.05).

Table 5: Calibration of the training-only predictions by regime: least-squares fit of observed on predicted ε over all screens, channels, designs and budgets with at least the stated number of training perturbations per message dimension. Bias is the mean of observed minus predicted ε.

Combinations	number	slope	R^2	mean abs. difference	bias
all	321	0.97	0.95	0.049	+0.048
$n/q \geq 4$	316	0.97	0.97	0.045	+0.044
$n/q \geq 8$	286	0.97	0.98	0.035	+0.034
$n/q \geq 16$	244	0.97	0.99	0.028	+0.026
$n/q \geq 32$	184	0.99	0.99	0.019	+0.017
$n/q < 8$	35			0.166	+0.166

Does the leverage factor matter? Table 6 compares Equation (2) with two naive training-only predictions on the same regression-decoder combinations ($n/q \geq 8$): the residual variance $\mathrm{RSS}/(n-q-1)$, which drops the factor $(1+1/n)(n-2)/(n-q-2)$, and the in-sample error RSS/n. Each is normalized by its own zero-bit prediction. The pooled fit, the fit within a budget (after removing the mean of each channel and size, so that the range of budgets does not drive R^2) and the decisions all favour Equation (2): it picks the panel size with the lowest held-out error on four of six screens, against two for the naive predictions, which cannot see overfitting. At a fixed size all three rank the designs identically, since the factor does not depend on the design.

Table 6: Equation (2) against naive training-only predictions (regression decoder, $n/q \geq 8$): pooled calibration, R^2 within a budget, and decisions (the panel size of the greedy design, and the design at each fixed size, with the lowest held-out error; regret in ε).

	pooled ($n/q \geq 8$)			within budget	decisions: correct / mean regret	
Prediction	slope	R^2	bias	R^2	panel size	design at fixed k
Equation (2)	0.96	0.98	+0.040	0.99	4/6, 0.036	30/36, 0.004
$\mathrm{RSS}/(n-q-1)$	0.90	0.92	+0.088	0.94	2/6, 0.121	30/36, 0.004
RSS/n	0.81	0.83	+0.133	0.85	2/6, 0.121	30/36, 0.004

Optimism by training size. Table 7 bins the optimism of Equation (2) in the training-size runs (greedy panels on the Norman screen and its singles, Replogle and sci-Plex K562; every combination of panel size and subsample with $n/q \geq 8$), the source of the rule on n in Section 5.2.

Table 7: Optimism of Equation (2) (observed minus predicted error, in units of the full-size zero-bit error) in the training-size runs with $n/q \geq 8$, by the number of training perturbations.

Training perturbations	combinations	mean optimism	largest optimism
fewer than 100	26	0.11	0.29
100 to 199	14	0.07	0.23
200 or more	9	0.01	0.05

Extrapolation in n. Table 8 predicts the error at the full training size from each subsample of the training-size runs with $n/q \geq 8$, by Equation (2) and by the n-independent guess that the error stays as observed at the smaller n. On Replogle, where n reaches the hundreds, Equation (2) is the better extrapolation (mean absolute error 0.03 against 0.10); on the Norman singles and on sci-Plex K562, whose subsamples have at most 65 and 116 perturbations, its optimism at small n makes it no better (0.12 against 0.03 and 0.11 against 0.10).

Designs selected on independent perturbations. Table 9 compares training-only predictions of the held-out error of greedy panels ($n/q \geq 8$ with $q = k$) on the first repeat of the folds. On the full training set, Equation (2) is as accurate per fold as nested cross-validation, which re-selects the panel in five inner folds, and as leave-one-out (mean absolute differences 0.12, 0.12 and 0.12). It is optimistic on every screen (by 0.01 to 0.06); nested cross-validation is pessimistic because each inner decoder sees 80% of the perturbations; and leave-one-out, which for least squares needs no refit either, is the least biased (-0.05 to $+0.03$). What Equation (2) adds is an explicit dependence on n and q, which chooses panel sizes (Table 6) and extrapolates in n once n is large (Table 8). Selecting the panel on one half of the training perturbations and fitting it on the other, the setting of Corollary 1, removes the optimism on the Norman singles and on MCF7 but not elsewhere: the design selection explains part of the optimism on these two screens, and the rest persists when the corollary holds.

Gaussian responses. Table 10 runs the full pipeline (split-half covariance, noise estimates, greedy panel, least squares, Equation (2)) on Gaussian responses simulated with each training fold's covariance and noise and the real numbers of perturbations (three simulations per fold, first repeat). On Gaussian data the predictions are unbiased up to 0.03 even though the panel is chosen on the training responses; on the real data they are optimistic. Selection alone therefore does not explain the optimism, which is consistent with heavy-tailed, non-Gaussian responses being its main source; covariance estimation and the noise model may contribute.

Without the strongest compounds. Table 12 removes the top decile of response energy from each sci-Plex line and reruns the pipeline. Most of the recoverable signal leaves with these compounds: the signal trace falls by 47% to 72%, and four genes recover little. Equation (2) still predicts the held-out distortion within 1% to 5% on average, as it does with all compounds, but in units of the much smaller recoverable signal this is an optimism of 0.06 to 0.12, with slopes of 0.74 to 0.88. The calibration of Figure 2a thus owes part of its range to the strongest compounds, and it is weaker

Table 8: Extrapolation of the held-out error of greedy k-gene panels to the full training size from a subsample of n training perturbations ($n/q \geq 8$): observed error at full size, Equation (2) from the subsample, the n-independent guess (the error observed at the subsample size) and Equation (2) at the subsample size, all expressed on one scale (distortions normalized by the observed full-size zero-bit error).

Screen	n	k	observed, full n	Eq. (2)	unchanged	Eq. (2) at n
Norman (singles)	20	1	0.88	0.67	0.90	0.74
Norman (singles)	32	1	0.88	0.69	0.91	0.73
Norman (singles)	32	4	0.58	0.45	0.69	0.52
Norman (singles)	49	1	0.88	0.80	0.85	0.82
Norman (singles)	49	4	0.58	0.50	0.61	0.53
Norman (singles)	65	1	0.88	0.81	0.87	0.82
Norman (singles)	65	4	0.58	0.53	0.58	0.54
Replogle	26	1	0.91	0.79	1.06	0.95
Replogle	65	1	0.91	0.90	0.98	0.96
Replogle	65	4	0.67	0.62	0.90	0.76
Replogle	129	1	0.91	0.88	0.95	0.91
Replogle	129	4	0.67	0.64	0.80	0.70
Replogle	129	16	0.44	0.35	0.78	0.55
Replogle	324	1	0.91	0.91	0.91	0.92
Replogle	324	4	0.67	0.67	0.69	0.69
Replogle	324	16	0.44	0.42	0.54	0.49
Replogle	647	1	0.91	0.91	0.90	0.91
Replogle	647	4	0.67	0.66	0.67	0.67
Replogle	647	16	0.44	0.42	0.48	0.44
sci-Plex K562	24	1	0.55	0.45	0.81	0.58
sci-Plex K562	36	1	0.55	0.47	0.68	0.55
sci-Plex K562	36	4	0.38	0.14	0.62	0.33
sci-Plex K562	58	1	0.55	0.51	0.64	0.55
sci-Plex K562	58	4	0.38	0.24	0.50	0.33
sci-Plex K562	87	1	0.55	0.47	0.58	0.49
sci-Plex K562	87	4	0.38	0.22	0.42	0.26
sci-Plex K562	116	1	0.55	0.53	0.56	0.54
sci-Plex K562	116	4	0.38	0.28	0.39	0.30

Table 9: Bias (mean of observed minus predicted ε) and, in parentheses, mean absolute difference of training-only predictions of the held-out error of greedy-variance panels, per fold and size. Columns 2 and 3 select and fit the panel on the same perturbations (all training perturbations, or a random half); column 4 selects on one half and fits on the other (Corollary 1).

	selected on the fitting set		split selection	competitors (full set)	
Screen	full n	half n	half n	nested CV	leave-one-out
Norman (all 226)	+0.032 (0.102)	+0.054 (0.104)	+0.056 (0.128)	−0.010 (0.102)	+0.019 (0.101)
Norman (101 singles)	+0.039 (0.207)	+0.055 (0.235)	−0.024 (0.245)	−0.072 (0.242)	−0.049 (0.233)
Replogle	+0.011 (0.074)	+0.036 (0.080)	+0.026 (0.078)	−0.015 (0.073)	−0.001 (0.073)
sci-Plex A549	+0.056 (0.111)	+0.048 (0.132)	+0.060 (0.131)	−0.006 (0.099)	+0.031 (0.108)
sci-Plex K562	+0.052 (0.104)	+0.085 (0.128)	+0.106 (0.145)	−0.048 (0.094)	−0.019 (0.093)
sci-Plex MCF7	+0.024 (0.091)	+0.062 (0.101)	+0.005 (0.097)	−0.055 (0.119)	−0.019 (0.102)

where little signal is left. The fingerprint prior recovers nothing without them either (ε of 0.97 to 1.04); four genes still beat it on K562 and MCF7 (0.69 and 0.77) but not on A549, where neither recovers any signal (1.00 against 0.97). Table 11 gives the concentration of response energy: the histone deacetylase inhibitors carry 36% to 58% of it, and the strongest of them 12 to 30 times the median energy.

The noise floor. Table 13 recomputes the best prior and the four-gene panel under five estimates of the noise floor. On the pooled screens, whose batch noise is near zero, nothing changes. On sci-Plex, where replicate wells differ, the recovered fraction of the signal moves by up to 0.12 for the prior and

Table 10: Optimism (mean of observed minus predicted ε) of Equation (2) for greedy panels on the real responses and on Gaussian responses with the same covariance, noise and numbers of perturbations.

	1 gene		4 genes		16 genes	
Screen	real	Gaussian	real	Gaussian	real	Gaussian
Norman (all 226)	+0.04	−0.01	+0.02	+0.00	+0.04	+0.01
Norman (101 singles)	+0.04	−0.01	+0.05	+0.01	+0.12	+0.01
Replogle	+0.00	+0.00	+0.01	+0.00	+0.02	−0.00
sci-Plex A549	−0.02	−0.03	+0.09	−0.01	+0.11	−0.00
sci-Plex K562	+0.01	+0.00	+0.05	+0.00	+0.15	+0.00
sci-Plex MCF7	−0.03	+0.03	+0.03	+0.01	+0.07	+0.01

Table 11: Response energy $\|R_p - \bar{R}\|^2$ on sci-Plex: histone deacetylase inhibitors (median and largest energy relative to the median of all compounds, share of the total energy, members of the top decile) and the share of the top decile.

			HDAC energy / median		HDAC share	HDAC inhibitors	top-decile share
Line	compounds	HDAC inhibitors	median	max	of energy	in top decile	of energy
sci-Plex A549	165	24	4.7	14.9	44%	13 of 17	46%
sci-Plex K562	181	26	6.9	11.9	36%	15 of 19	39%
sci-Plex MCF7	181	27	10.5	30.3	58%	14 of 19	61%

0.25 for four genes, but four genes beat the prior under every estimate, the measurement equivalents stay within 0.01 genes (they compare errors on the same perturbations), and the calibration of the gene panels keeps a slope of 0.97.

The factor model that guides greedy selection keeps the eigen-directions of $\hat{\Sigma}$ above the noise edge. Table 14 scales the edge by one half and by two (first repeat of the folds): the number of directions kept changes by a factor of two or more, but the selected panels barely change, the four-gene error moves by at most 0.02, and the price of the best prior stays between 0.4 and 0.6 genes on four screens and at 3.4 on Replogle.

The measurement equivalent. Table 15 varies the panel design and the interpolation of Definition 1. Informed designs give similar prices on the four screens where priors are weak, with one exception (greedy information on the Norman singles, whose first gene has the best signal-to-noise ratio rather than the widest reach); random genes are a much weaker currency, and on two sci-Plex lines no random panel of up to 32 genes reaches the prior. Below one gene the equivalent is a monotone function of the ratio between the signal recovered by the prior and by the first greedy gene. It is an exchange rate between two ways of lowering the error, not an additive quantity; whether a prior still helps once genes are measured is tested separately (Table 31).

Measurements in the bits plane. Figure 4 places the gene-panel designs and the few-cell channel in the plane of information and error, computed from training perturbations before any held-out perturbation is measured.

All gene-panel designs. Table 16 reports every gene-panel design and size with the regression decoder. From $k = 2$ on, greedy variance reduction is the best design or within 0.02 of it on every screen; for a single gene, the highest-variance gene is as good or better on five of the six screens. Greedy information is poor at $k = 1$ because it selects the gene with the largest signal-to-noise ratio rather than the gene that explains most of the other genes. The stable genes barely lower the error ($\varepsilon \geq 0.96$), and up to 16 genes their predicted and observed values agree that they carry almost no information about the perturbation.

Crossover cell counts beyond 32 cells. Table 17 extends the raw-mean channel to 64 or 128 cells on the perturbations that have that many half-A cells. The predicted crossover is observed on three screens, and on the K562 subset, where the estimated batch noise of the message exceeds the signal,

Table 12: sci-Plex with and without the top decile of response energy: fold-averaged signal trace and noise floor, four-gene error, and the calibration of Equation (2) over gene panels with $n/q \geq 8$.

Line	compounds	signal trace	noise floor	four genes ε	slope	R^2	bias in ε
sci-Plex A549, all	165	17.6	19.2	0.37	0.99	1.00	+0.029
sci-Plex A549, without top decile	148	6.0	18.9	1.00	0.77	0.77	+0.061
sci-Plex K562, all	181	12.6	15.7	0.40	0.96	0.97	+0.064
sci-Plex K562, without top decile	162	6.6	14.1	0.69	0.88	0.84	+0.120
sci-Plex MCF7, all	181	14.7	5.4	0.25	0.97	1.00	+0.033
sci-Plex MCF7, without top decile	162	4.1	4.5	0.77	0.74	0.87	+0.095

Table 13: Best prior / four greedy genes (ε) under alternative estimates of the target noise floor, and the calibration of Equation (2) on the gene panels ($n/q \geq 8$, all six rows of Table 2).

Screen	aggregate (default)	cell sampling only	per-gene clipped	batch noise $\times 0.5$	batch noise $\times 1.5$
Norman (101 singles)	0.93 / 0.59	0.93 / 0.59	0.93 / 0.57	0.93 / 0.59	0.93 / 0.59
Replogle	0.71 / 0.67	0.71 / 0.67	0.70 / 0.66	0.71 / 0.67	0.71 / 0.67
sci-Plex A549	0.71 / 0.37	0.77 / 0.51	0.70 / 0.35	0.74 / 0.45	0.66 / 0.26
sci-Plex K562	0.76 / 0.40	0.80 / 0.51	0.75 / 0.38	0.78 / 0.46	0.73 / 0.33
sci-Plex MCF7	0.70 / 0.25	0.74 / 0.35	0.70 / 0.24	0.72 / 0.30	0.68 / 0.19
calibration (gene panels, $n/q \geq 8$): slope / MAE	0.97 / 0.035	0.97 / 0.031	0.97 / 0.036	0.97 / 0.033	0.97 / 0.038

the raw mean never overtakes the mean response, as Theorem 2b predicts; its error is underpredicted there, because 31 training compounds estimate the batch noise poorly.

A few genes read from a few cells. Table 18 reads the greedy four-gene message from only m cells of the held-out perturbation instead of its half-A pseudobulk, with the regression decoder trained on m-cell messages of the training perturbations (two draws; perturbations with at least 32 half-A cells; first repeat), and compares it with the best zero-shot prior on the same perturbations. With five cells the four genes beat the prior on every screen except Replogle, where the Gene Ontology prior is worth about three genes read from all cells.

Few-cell decoders. Table 19 compares the four decoders of the few-cell channel. The raw mean and per-gene shrinkage have only $O(G)$ parameters: the raw mean is predicted within 7% and shrinkage within 0.03 in ε, but shrinkage gains little. The Bayes decoder with the full plug-in covariance has the lowest error on five of the six screens, but its population formula is optimistic by up to 0.44, because its $G \times G$ gain is estimated from 80 to 1,232 training perturbations. The regression decoder on a projected message usually lies in between, and Equation (2) predicts its error to within 0.11.

Cross-context messages. Table 21 reports all six ordered pairs of sci-Plex cell lines. The misattributed message raises the error to about $2 - \varepsilon$ of the aligned message, as Theorem 2c predicts.

Cross-screen messages for genetic perturbations. Table 20 uses, for the 1,036 Replogle K562 perturbations whose target was also perturbed in the RPE1 essential-gene screen, the RPE1 response of the same target as the message (Appendix B): the linear model on perturbation embeddings from another screen used by Ahlmann-Eltze et al. (2025), which in our taxonomy is a cross-context measurement rather than a prior. Its error falls with q and levels off; at the $q = 64$ that Equation (2) predicts best from training perturbations, it is $\varepsilon = 0.66$ (predicted 0.62), worth 4.7 greedy genes of the K562 response itself, against 0.79 for the Gene Ontology ridge prior and 0.68 for four greedy genes on the same perturbations (the prior reaches 0.71 on all 1,617 perturbations, with more training data; Table 2). The RPE1 response of another target gives 1.55, more than the $2 - \varepsilon = 1.34$ of Theorem 2c, as expected for a least-squares decoder with many coefficients (Appendix A); at $q = 16$ the two are 1.37 and 1.29.

Uncertainty and measurement equivalents. Table 22 gives 95% intervals for the best zero-shot prior, for the paired differences of the greedy one- and four-gene panels to it, and for the measurement equivalent of the prior. Normalized errors are ratios of sums over the (fold, held-out perturbation) records, as in Table 2, and the intervals come from a cluster bootstrap that resamples perturbations. Paired quantities use the folds on which the prior was evaluated (the first repeat for multilayer

Table 14: Greedy panels under a factor-model threshold of 0.5, 1 and 2 times the noise edge (first repeat): directions kept, four-gene ε, and the measurement equivalent of the best prior.

	directions kept			four genes ε			prior worth (genes)		
Screen	0.5×	1×	2×	0.5×	1×	2×	0.5×	1×	2×
Norman (101 singles)	45	31	19	0.61	0.61	0.61	0.6	0.6	0.6
Replogle	110	50	29	0.68	0.68	0.68	3.4	3.4	3.4
sci-Plex A549	31	14	7	0.39	0.40	0.40	0.4	0.4	0.4
sci-Plex K562	39	15	8	0.41	0.40	0.40	0.5	0.5	0.5
sci-Plex MCF7	26	16	9	0.24	0.26	0.26	0.4	0.4	0.4

Table 15: Measurement equivalent of the best prior (measured genes) by panel-selection rule and interpolation.

	panel-selection rule (interpolation in $\log_2(1+k)$)				greedy variance, other interpolation	
Screen	greedy variance	greedy information	top variance	random	linear in k	monotone cubic
Norman (101 singles)	0.6	1.6	0.3	2.3	0.7	0.7
Replogle	3.4	5.2	6.4	26.0	3.5	3.4
sci-Plex A549	0.4	0.5	0.4	>32	0.5	0.3
sci-Plex K562	0.5	0.7	0.6	>32	0.6	0.4
sci-Plex MCF7	0.4	0.4	0.4	5.0	0.5	0.3

perceptrons). Four genes beat the best prior on every screen; one gene beats it on the three chemical screens, is indistinguishable from it on the Norman singles, and is worse than the Gene Ontology prior on Replogle, where the prior is worth 3.4 genes. Without the histone deacetylase inhibitors the sci-Plex priors are worth 0.2 to 0.6 genes. Counted per perturbation, without energy weighting, four genes beat the prior on 58% to 87% of perturbations, and the median per-perturbation error relative to the mean response is 0.69 to 0.92 for four genes against 0.94 to 1.00 for the prior.

Measured genes predicting themselves. The target contains the measured genes, whose half-B values the half-A message predicts directly. Table 23 recomputes the error of greedy panels without the panel genes in the target (first repeat of the folds; the same per-gene noise floor for both columns, so only their difference is interpreted). Excluding them raises ε by 0.03 on average and by about 0.1 at most (sixteen genes on the Norman singles): the gains come from the other genes.

The strongest compound class. On sci-Plex the 24 to 27 histone deacetylase inhibitors per line carry the largest responses. Table 24 repeats the comparison without them (decoders and priors unchanged; only the evaluation set shrinks). The priors then recover 3% to 9% of the signal, one gene 8% to 14%, and four genes 21% to 44%; four genes still beat the prior on every line, with intervals that exclude zero.

Holding out whole pathways. Table 25 holds out one of the 16 first-level pathways at a time on sci-Plex, so that no compound of the held-out pathway is in training. The fingerprint and annotation priors fall to or below the mean baseline, the four-gene panel loses most of its value, and its training-only prediction, which assumes exchangeable perturbations, stays near its value under random folds. The failure is largest on A549. Under a pathway hold-out a mechanism-of-action label of the held-out class has no training members, so a label-based prior reduces to the mean response: the priors recognize classes they have seen.

Graded coverage of a held-out pathway. Table 26 moves a fraction c of the held-out pathway's compounds into training. A quarter suffices to restore most of the value of the measured genes and the accuracy of Equation (2); the priors recover more slowly. The evaluation set shrinks as c grows, so the rows compare random subsets of the pathways' compounds.

Coverage floor and leverage. Table 27 gives the coverage floor of Proposition 1 (the share of the held-out signal energy outside the span of the training targets, estimated from the two halves) on the same folds, with the four-gene error and its prediction. The floor rises by a factor of 2.5 to 5.4

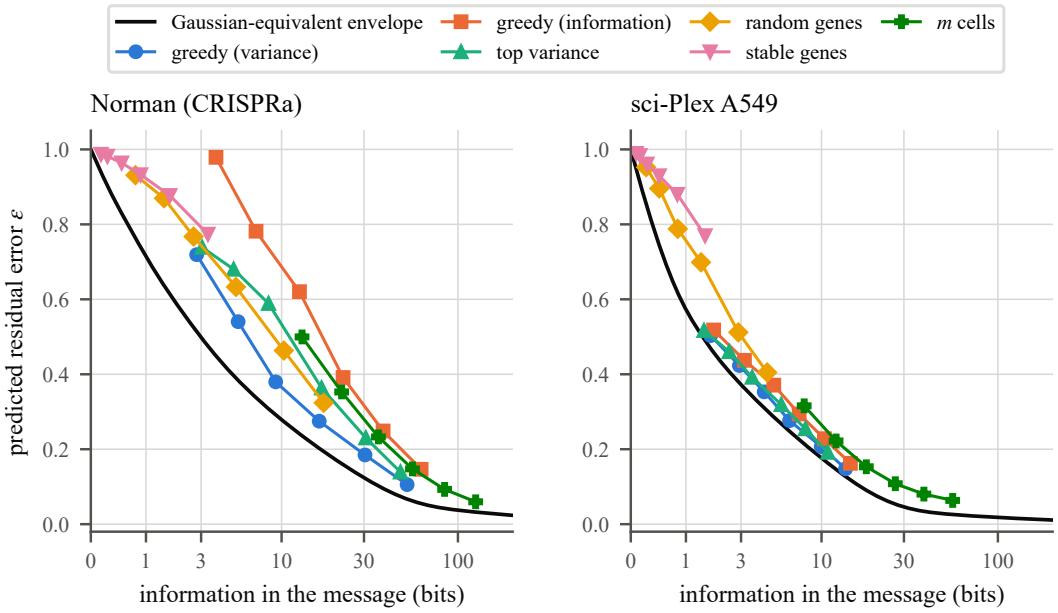


Figure 4: **Measurements in the bits plane, computed from training perturbations.** Predicted residual error of the plug-in Bayes decoder against the Gaussian-equivalent information of the message (Proposition 2), for gene panels of five designs ($k = 1$ to 32) and for m cells. The black curve is the envelope of Proposition 2 for the same covariance estimate; it bounds these points, and it would bound every decoder only if the responses were Gaussian. Held-out errors of the Bayes panels differ from these predictions by 0.02 to 0.04 on average (all screens but the Norman singles). Greedy variance reduction comes closest to the envelope (within 0.03 to 0.09 at four to sixteen genes); greedy information buys bits on directions of small variance; stable genes carry almost none; single cells carry many weak bits, gene panels few strong ones. The m-cell predictions use the full-covariance Bayes decoder, whose held-out errors are higher (Table 19).

when the pathway is withheld and returns near its random-fold value at $c = 0.25$; the four-gene error always lies above it, whereas Equation (2) falls below it at $c = 0$ on A549 and MCF7. The floor needs the held-out response, so it diagnoses a class after its first members are profiled. A training-only diagnostic available after the genes are read, the leverage of the message relative to the training messages, does not separate the shifted perturbations from the others.

Frozen pretrained representations. Table 28 feeds frozen embeddings of large pretrained models into the same nearest-neighbour, ridge and kernel-ridge predictors, on the same folds: ESM-2 (150M parameters) embeddings of the target protein, averaged over residues; Geneformer (V2, 104M parameters, trained on about 104 million single-cell transcriptomes) input embeddings of the target gene; and ChemBERTa (77M parameters) embeddings of the compound's SMILES string, averaged over tokens. None of them is worth more than a fraction of one measured gene. We use the static gene embeddings that the models expose; contextual or in silico perturbation embeddings and fine-tuned models (scGPT, Geneformer) require training on perturbation data that our compute did not allow.

Zero-shot predictors and capacity on identical folds. Table 29 lists the six best zero-shot predictors of each screen together with the mean baseline. The residual correlation r is the cosine similarity between predicted and observed perturbation-specific residuals, averaged over held-out perturbations. Because the MLPs ran on the first repeat of the folds only, Table 30 compares them with the best kernel or linear predictor on exactly those folds, together with the two controls from the default initialization. With early stopping, the default initialization is worse than the mean-response start on every screen. Trained for 1,000 epochs, the networks fit the training perturbations closely; they then match kernel ridge on sci-Plex A549, stay worse on the other chemical lines and are far worse on the genetic screens. No variant improves on the best kernel or linear predictor by more than 0.010. A linear model on perturbation embeddings learned from another screen, the strongest simple baseline

Table 16: Normalized residual error ε [predicted] for all gene-panel designs (regression decoder).

Screen	design	$k=1$	$k=2$	$k=4$	$k=8$	$k=16$	$k=32$
Norman (all 226)	greedy variance	0.77 [0.73]	0.58 [0.56]	0.42 [0.39]	0.34 [0.30]	0.28 [0.24]	0.25 [0.18]
	greedy information	0.99 [0.99]	0.81 [0.82]	0.66 [0.65]	0.46 [0.44]	0.35 [0.33]	0.32 [0.25]
	top variance	0.75 [0.74]	0.70 [0.69]	0.61 [0.59]	0.41 [0.39]	0.32 [0.29]	0.29 [0.22]
	top mean response	0.75 [0.74]	0.67 [0.66]	0.54 [0.53]	0.40 [0.39]	0.36 [0.32]	0.35 [0.28]
	random (10 draws)	0.93 [0.93]	0.88 [0.88]	0.80 [0.78]	0.70 [0.66]	0.61 [0.53]	0.55 [0.43]
	stable genes	0.99 [0.99]	0.99 [0.98]	1.01 [0.97]	1.01 [0.96]	1.03 [0.92]	1.04 [0.88]
Norman (101 singles)	greedy variance	0.88 [0.80]	0.67 [0.62]	0.59 [0.51]	0.53 [0.42]	0.53 [0.36]	0.68 [0.36]
	greedy information	1.02 [0.97]	0.98 [0.86]	0.79 [0.68]	0.70 [0.62]	0.62 [0.53]	0.73 [0.50]
	top variance	0.81 [0.81]	0.78 [0.76]	0.76 [0.70]	0.62 [0.53]	0.62 [0.47]	0.77 [0.44]
	top mean response	0.84 [0.83]	0.72 [0.71]	0.69 [0.64]	0.64 [0.56]	0.67 [0.52]	0.82 [0.59]
	random (10 draws)	0.97 [0.95]	0.96 [0.91]	0.91 [0.86]	0.87 [0.78]	0.91 [0.72]	1.19 [0.78]
	stable genes	1.01 [1.00]	1.02 [1.00]	1.03 [1.01]	1.09 [1.03]	1.22 [1.10]	1.82 [1.37]
Replogle	greedy variance	0.91 [0.89]	0.80 [0.79]	0.67 [0.66]	0.55 [0.54]	0.46 [0.44]	0.39 [0.36]
	greedy information	0.94 [0.94]	0.87 [0.86]	0.76 [0.75]	0.61 [0.60]	0.48 [0.45]	0.41 [0.37]
	top variance	0.89 [0.89]	0.83 [0.83]	0.76 [0.75]	0.68 [0.66]	0.50 [0.48]	0.41 [0.38]
	top mean response	0.89 [0.89]	0.86 [0.85]	0.76 [0.75]	0.67 [0.66]	0.53 [0.51]	0.46 [0.43]
	random (10 draws)	0.97 [0.97]	0.96 [0.96]	0.91 [0.91]	0.86 [0.86]	0.77 [0.75]	0.68 [0.65]
	stable genes	1.00 [1.00]	1.00 [1.00]	1.00 [1.00]	0.99 [0.99]	0.98 [0.98]	0.96 [0.95]
sci-Plex A549	greedy variance	0.44 [0.43]	0.38 [0.36]	0.37 [0.32]	0.35 [0.30]	0.40 [0.34]	0.61 [0.51]
	greedy information	0.51 [0.49]	0.40 [0.40]	0.37 [0.36]	0.38 [0.33]	0.41 [0.36]	0.61 [0.51]
	top variance	0.41 [0.42]	0.40 [0.39]	0.40 [0.36]	0.37 [0.34]	0.44 [0.39]	0.66 [0.55]
	top mean response	0.73 [0.73]	0.69 [0.67]	0.57 [0.53]	0.41 [0.38]	0.44 [0.39]	0.68 [0.58]
	random (10 draws)	0.97 [0.95]	0.94 [0.92]	0.87 [0.85]	0.79 [0.76]	0.77 [0.71]	0.95 [0.83]
	stable genes	1.02 [1.00]	1.04 [1.01]	1.07 [1.03]	1.11 [1.07]	1.20 [1.18]	1.47 [1.40]
sci-Plex K562	greedy variance	0.57 [0.55]	0.43 [0.40]	0.40 [0.34]	0.36 [0.28]	0.43 [0.28]	0.63 [0.37]
	greedy information	0.70 [0.64]	0.48 [0.43]	0.42 [0.35]	0.40 [0.31]	0.44 [0.30]	0.65 [0.40]
	top variance	0.62 [0.61]	0.57 [0.55]	0.47 [0.42]	0.38 [0.31]	0.43 [0.30]	0.68 [0.41]
	top mean response	0.66 [0.65]	0.62 [0.60]	0.56 [0.52]	0.57 [0.51]	0.52 [0.42]	0.65 [0.48]
	random (10 draws)	0.96 [0.95]	0.91 [0.90]	0.86 [0.82]	0.79 [0.72]	0.79 [0.65]	0.93 [0.68]
	stable genes	1.02 [1.00]	1.02 [1.00]	1.06 [1.01]	1.12 [1.03]	1.27 [1.09]	1.60 [1.23]
sci-Plex MCF7	greedy variance	0.37 [0.36]	0.30 [0.27]	0.25 [0.21]	0.22 [0.16]	0.19 [0.11]	0.25 [0.12]
	greedy information	0.37 [0.37]	0.33 [0.31]	0.29 [0.25]	0.26 [0.19]	0.20 [0.13]	0.24 [0.12]
	top variance	0.35 [0.35]	0.34 [0.33]	0.24 [0.23]	0.20 [0.17]	0.21 [0.14]	0.27 [0.15]
	top mean response	0.70 [0.68]	0.45 [0.43]	0.31 [0.30]	0.24 [0.20]	0.22 [0.16]	0.26 [0.16]
	random (10 draws)	0.90 [0.90]	0.86 [0.83]	0.75 [0.70]	0.61 [0.56]	0.52 [0.45]	0.50 [0.39]
	stable genes	1.00 [0.99]	1.00 [0.98]	1.00 [0.98]	1.02 [0.99]	1.05 [1.01]	1.23 [1.09]

Table 17: Raw mean of m cells on perturbations with many half-A cells: predicted and observed crossover with the mean response (observed by log-linear interpolation between the bracketing budgets), ε at the largest budget, and the largest relative error of the predicted ε over the budgets.

Screen	cells $\geq$	perturbations	$m_\times$ predicted	crossover observed	ε at largest m [pred.]	max. rel. error
Norman	64	212	23	24 (16 to 24)	0.40 [0.38]	5%
Replogle	128	222	41	44 (32 to 48)	0.36 [0.34]	5%
sci-Plex MCF7	64	146	41	41 (32 to 48)	0.70 [0.68]	3%
sci-Plex K562	128	39	never	none up to 128	3.93 [2.98]	24%

of Ahlmann-Eltze et al. (2025), cannot be run on the Norman singles (only 4 of their 101 target genes are in the Replogle essential-gene screen); it is run on Replogle with the RPE1 screen as a cross-screen message (Table 20), and for compounds its analogue, the same compound's response in another cell line, is our cross-context channel (Table 21). GEARS (Norman singles, five folds) reaches $\varepsilon = 1.56$ with $r = 0.04$ at its reduced budget; its validation error stopped improving within six epochs, but more cells per perturbation, a larger model or tuning may help, so it illustrates rather than bounds what graph models achieve.

Priors and measured genes together. Table 31 adds the prior's prediction of each gene as a regressor to the k-gene message (per-gene least squares; the prior's predictions for training perturbations are exact leave-one-out, so the decoder never sees an in-sample prior; first repeat). On sci-Plex and on the Norman singles the prior adds at most 0.02 to four genes: the two carry the same information. On Replogle they are complementary: Gene Ontology and four genes together reach 0.53. The measurement equivalent is therefore an exchange rate, not a decomposition of information.

Table 18: Normalized residual error ε of the greedy four-gene message read from m cells, and of the best zero-shot prior, on the perturbations with at least 32 half-A cells.

Screen	best prior	4 genes, 5 cells	4 genes, 10 cells	4 genes, 20 cells	4 genes, all cells
Norman (101 singles)	0.93	0.76	0.71	0.65	0.61
Replogle	0.70	0.88	0.83	0.76	0.67
sci-Plex A549	0.80	0.63	0.57	0.51	0.47
sci-Plex K562	0.79	0.72	0.65	0.54	0.42
sci-Plex MCF7	0.72	0.35	0.33	0.27	0.23

Table 19: Normalized residual error ε [predicted] of the few-cell decoders on the perturbations with at least 32 cells in half A.

Screen	decoder	$m = 1$	$m = 2$	$m = 4$	$m = 8$	$m = 16$	$m = 32$
Norman (all 226)	raw mean	21.52 [20.96]	10.87 [10.50]	5.41 [5.27]	2.75 [2.66]	1.40 [1.35]	0.72 [0.69]
	per-gene shrinkage	0.84 [0.84]	0.75 [0.75]	0.64 [0.64]	0.51 [0.52]	0.39 [0.40]	0.28 [0.29]
	Bayes (full covariance)	0.66 [0.51]	0.53 [0.37]	0.39 [0.25]	0.29 [0.16]	0.22 [0.11]	0.17 [0.08]
	regression ($q = 8$)	0.72 [0.68]	0.63 [0.56]	0.48 [0.44]	0.40 [0.35]	0.34 [0.30]	0.30 [0.26]
Norman (101 singles)	raw mean	32.20 [30.21]	15.76 [15.13]	8.00 [7.59]	4.04 [3.82]	2.05 [1.94]	1.06 [1.00]
	per-gene shrinkage	0.91 [0.88]	0.82 [0.81]	0.74 [0.72]	0.62 [0.60]	0.49 [0.48]	0.37 [0.37]
	Bayes (full covariance)	0.97 [0.55]	0.82 [0.40]	0.71 [0.27]	0.57 [0.18]	0.47 [0.12]	0.40 [0.09]
	regression ($q = 8$)	0.91 [0.82]	0.81 [0.72]	0.69 [0.59]	0.58 [0.49]	0.52 [0.42]	0.46 [0.38]
Replogle	raw mean	66.68 [62.94]	33.10 [31.49]	16.59 [15.77]	8.32 [7.91]	4.19 [3.98]	2.13 [2.01]
	per-gene shrinkage	0.97 [0.97]	0.93 [0.93]	0.88 [0.88]	0.80 [0.80]	0.68 [0.69]	0.55 [0.55]
	Bayes (full covariance)	0.84 [0.75]	0.73 [0.63]	0.61 [0.48]	0.48 [0.35]	0.37 [0.24]	0.28 [0.17]
	regression ($q = 8$)	0.87 [0.83]	0.79 [0.74]	0.69 [0.65]	0.58 [0.54]	0.50 [0.46]	0.44 [0.40]
sci-Plex A549	raw mean	56.62 [54.84]	28.35 [27.63]	14.40 [14.02]	7.34 [7.22]	3.82 [3.81]	2.06 [2.11]
	per-gene shrinkage	0.93 [0.94]	0.89 [0.89]	0.83 [0.83]	0.75 [0.75]	0.65 [0.66]	0.55 [0.56]
	Bayes (full covariance)	0.69 [0.31]	0.59 [0.21]	0.49 [0.14]	0.40 [0.10]	0.35 [0.07]	0.32 [0.05]
	regression ($q = 8$)	0.74 [0.65]	0.61 [0.53]	0.54 [0.44]	0.48 [0.38]	0.43 [0.33]	0.41 [0.30]
sci-Plex K562	raw mean	68.41 [65.19]	34.01 [32.74]	17.10 [16.52]	8.58 [8.40]	4.32 [4.35]	2.20 [2.32]
	per-gene shrinkage	0.96 [0.96]	0.93 [0.93]	0.88 [0.88]	0.79 [0.81]	0.69 [0.71]	0.57 [0.60]
	Bayes (full covariance)	0.66 [0.44]	0.54 [0.32]	0.44 [0.22]	0.35 [0.14]	0.29 [0.09]	0.24 [0.06]
	regression ($q = 8$)	0.73 [0.65]	0.60 [0.54]	0.50 [0.42]	0.41 [0.32]	0.34 [0.24]	0.28 [0.19]
sci-Plex MCF7	raw mean	31.63 [31.42]	15.95 [15.79]	8.05 [7.98]	4.11 [4.07]	2.14 [2.12]	1.15 [1.14]
	per-gene shrinkage	0.83 [0.83]	0.76 [0.75]	0.66 [0.67]	0.55 [0.56]	0.45 [0.46]	0.35 [0.37]
	Bayes (full covariance)	0.34 [0.27]	0.26 [0.20]	0.21 [0.14]	0.16 [0.10]	0.13 [0.07]	0.10 [0.05]
	regression ($q = 8$)	0.38 [0.36]	0.31 [0.27]	0.24 [0.20]	0.18 [0.14]	0.14 [0.10]	0.12 [0.08]

A standard per-perturbation metric. Table 32 reports the Pearson correlation between predicted and observed change from control, over all modelled genes and over the 20 genes with the largest observed change, averaged over perturbations. The mean response already scores 0.27 to 0.48 on all genes, because it captures the change shared by all perturbations; four genes add 0.07 to 0.15 on all genes and 0.05 to 0.23 on the top 20. GEARS scores below the mean response on all genes and equal to it on the top 20, consistent with its ε above 1 and with Ahlmann-Eltze et al. (2025).

Decoder capacity at a fixed message. Table 33 fixes the message (the greedy k-gene panel) and varies the decoder. Nonlinear decoders improve ε by at most 0.06 and are often worse than the linear decoder, sometimes much worse, whereas adding measured genes changes it by tenths: at a fixed message, capacity buys little.

Components of the zero-bit error. Table 34 decomposes the zero-bit distortion of each screen into the signal trace, the cell-sampling noise of the target and its aggregate batch noise (Appendix B). The normalized error ε uses these values; the regression-decoder predictions themselves do not depend on them. A predicted ε is training-only: its denominator is the predicted zero-bit distortion (Equation (2) with $q = 0$) minus the training estimate of $\operatorname{tr} N_B$, not the observed D_0 (Table 8, which compares distortions across training sizes, puts all columns on the observed full-size scale instead). The cost of estimating the mean response from n training perturbations, the factor $1 + 1/n$ of Theorem 3, is 1.2% of the zero-bit error on the Norman singles ($n = 81$) and below 0.1% on Replogle ($n = 1{,}294$).

Batch-disjoint versus random halves. With random rather than replicate-disjoint halves on sci-Plex A549 (the same 165 compounds, same folds), well effects shared by the two halves count as signal: the estimated signal trace grows by 11% (19.6 against 17.6). The inflation falls mostly on the

Table 20: Cross-screen messages on Replogle K562 (1,036 perturbations whose target is in both screens; two repeats of five folds): ε [predicted] of the RPE1 response of the same target projected on q leading directions, of a misattributed RPE1 message, and of the Gene Ontology ridge prior, with greedy-panel measurement equivalents.

Message or prior	ε [pred.]	misattributed	worth (genes)
Gene Ontology ridge prior	0.79		2.4
RPE1 response of the same target, $q = 1$	0.91 [0.91]	1.09	1.0
RPE1 response of the same target, $q = 2$	0.86 [0.86]	1.15	1.5
RPE1 response of the same target, $q = 4$	0.82 [0.81]	1.21	2.1
RPE1 response of the same target, $q = 8$	0.76 [0.75]	1.29	2.8
RPE1 response of the same target, $q = 16$	0.71 [0.68]	1.37	3.5
RPE1 response of the same target, $q = 32$	0.67 [0.63]	1.46	4.3
RPE1 response of the same target, $q = 64$	0.66 [0.62]	1.55	4.7

Table 21: Cross-context messages: normalized residual error ε [predicted] in the target line when the message is the compound's response in the source line projected on q leading directions, and when it is the response of another compound (misattributed).

Source → target	$q = 1$	$q = 2$	$q = 4$	$q = 8$	$q = 16$	misattributed ($q = 1$)
A549 → K562	0.57 [0.56]	0.57 [0.56]	0.52 [0.48]	0.50 [0.45]	0.49 [0.42]	1.46 [1.47]
A549 → MCF7	0.36 [0.34]	0.35 [0.34]	0.32 [0.30]	0.27 [0.23]	0.24 [0.19]	1.57 [1.67]
K562 → A549	0.45 [0.45]	0.43 [0.42]	0.42 [0.37]	0.36 [0.29]	0.31 [0.25]	1.54 [1.57]
K562 → MCF7	0.42 [0.41]	0.40 [0.38]	0.38 [0.34]	0.30 [0.25]	0.28 [0.22]	1.61 [1.60]
MCF7 → A549	0.39 [0.40]	0.37 [0.37]	0.33 [0.33]	0.33 [0.31]	0.28 [0.22]	1.54 [1.63]
MCF7 → K562	0.55 [0.54]	0.53 [0.50]	0.44 [0.40]	0.43 [0.37]	0.45 [0.37]	1.53 [1.49]

few-cell channel, because single cells carry the batch signature of their well: one decoded cell reaches $\varepsilon = 0.53$ with random halves against 0.74 with replicate-disjoint halves, whereas one measured gene reaches 0.48 against 0.44. The predictions remain calibrated under both protocols (slope 0.94 and 0.97). All main results use batch-disjoint halves, the conservative choice for the cell channel.

Measured responses in the training set. For Norman two-gene perturbations whose single-gene responses are in the training set, ridge regression on summed gene features reaches $\varepsilon = 0.26$, the additive model 0.41 and nearest neighbours 0.46 to 0.52: the same models that stay near $\varepsilon = 1$ for unseen genes recover most of the signal once the training set contains measured responses of both genes. Because a two-gene response is largely additive in its singles, this is a positive control for (P1) rather than a test of generalization to new genes. Ridge beats the additive model, unlike in Ahlmann-Eltze et al. (2025); the additive model sums two noisy single-gene responses, and energy weighting and our gene universe differ from theirs, but we did not isolate the cause.

Table 22: Best zero-shot prior (ε), paired differences of the greedy one- and four-gene panels to it (negative: the measured genes are better), and the measurement equivalent of the prior (the number of greedily chosen measured genes with the same error), with 95% cluster-bootstrap intervals over perturbations.

Screen	best prior	1 gene − prior	4 genes − prior	prior worth (genes)
Norman (101 singles)	0.93 [0.85, 1.00]	−0.03 [−0.11, +0.04]	−0.32 [−0.41, −0.23]	0.6 [0.0, 1.2]
Replogle	0.71 [0.68, 0.73]	+0.20 [+0.17, +0.23]	−0.03 [−0.06, −0.01]	3.4 [2.9, 3.9]
sci-Plex A549	0.71 [0.61, 0.84]	−0.27 [−0.37, −0.16]	−0.34 [−0.44, −0.23]	0.4 [0.3, 0.6]
sci-Plex K562	0.76 [0.67, 0.85]	−0.19 [−0.31, −0.08]	−0.36 [−0.46, −0.26]	0.5 [0.3, 0.7]
sci-Plex MCF7	0.70 [0.62, 0.82]	−0.33 [−0.43, −0.23]	−0.46 [−0.54, −0.37]	0.4 [0.2, 0.5]

Table 23: Normalized residual error ε of greedy panels on all genes and on the genes outside the panel.

	1 gene		4 genes		16 genes	
Screen	all genes	panel excluded	all genes	panel excluded	all genes	panel excluded
Norman (all 226)	0.78	0.81	0.41	0.46	0.27	0.34
Norman (101 singles)	0.90	0.92	0.60	0.66	0.51	0.61
Replogle	0.91	0.92	0.67	0.69	0.44	0.49
sci-Plex A549	0.41	0.42	0.38	0.40	0.41	0.44
sci-Plex K562	0.54	0.55	0.38	0.38	0.42	0.44
sci-Plex MCF7	0.36	0.38	0.25	0.27	0.19	0.22

Table 24: Normalized residual error ε on sci-Plex without the histone deacetylase inhibitors, with 95% cluster-bootstrap intervals.

Screen	P kept	best prior	1 gene	4 genes	4 genes − prior
sci-Plex A549	141	0.97 [0.87, 1.07]	0.92 [0.86, 0.96]	0.79 [0.67, 0.89]	−0.18 [−0.36, −0.03]
sci-Plex K562	155	0.91 [0.82, 0.99]	0.87 [0.79, 0.92]	0.56 [0.40, 0.67]	−0.35 [−0.51, −0.24]
sci-Plex MCF7	154	0.96 [0.86, 1.06]	0.86 [0.76, 0.93]	0.56 [0.44, 0.66]	−0.40 [−0.54, −0.28]

Table 25: Random folds against leave-one-pathway-out folds on sci-Plex: normalized residual error ε of the best zero-shot prior and of the greedy four-gene panel (regression decoder), and the four-gene error predicted from training perturbations.

	best zero-shot prior		4 genes		4 genes, predicted	
Screen	random folds	pathway out	random folds	pathway out	random folds	pathway out
sci-Plex A549	0.71	1.02	0.37	0.93	0.32	0.36
sci-Plex K562	0.76	1.04	0.40	0.70	0.34	0.36
sci-Plex MCF7	0.70	1.02	0.25	0.68	0.21	0.23

Table 26: Pathway held out with a fraction c of its compounds moved into training: normalized residual error ε [predicted] of the greedy one- and four-gene panels, of the fingerprint and class-mean priors, and the optimism of the four-gene prediction.

Line, c	held out	1 gene [pred.]	4 genes [pred.]	kernel ridge	class mean	4 genes − pred.
sci-Plex A549, 0.00	165	0.97 [0.47]	0.93 [0.36]	1.02	1.07	+0.57
sci-Plex A549, 0.25	125	0.50 [0.46]	0.47 [0.34]	0.84	0.89	+0.13
sci-Plex A549, 0.50	82	0.43 [0.46]	0.38 [0.34]	0.77	0.85	+0.04
sci-Plex A549, 0.75	42	0.50 [0.45]	0.43 [0.33]	0.66	0.77	+0.09
sci-Plex K562, 0.00	181	0.95 [0.55]	0.70 [0.36]	1.04	1.07	+0.35
sci-Plex K562, 0.25	136	0.87 [0.57]	0.43 [0.35]	0.95	0.93	+0.08
sci-Plex K562, 0.50	91	0.73 [0.57]	0.39 [0.34]	0.79	0.90	+0.05
sci-Plex K562, 0.75	47	0.61 [0.56]	0.38 [0.34]	0.80	0.83	+0.04
sci-Plex MCF7, 0.00	181	0.84 [0.39]	0.68 [0.23]	1.03	1.02	+0.45
sci-Plex MCF7, 0.25	134	0.45 [0.39]	0.24 [0.22]	0.88	0.86	+0.02
sci-Plex MCF7, 0.50	89	0.45 [0.39]	0.20 [0.22]	0.87	0.92	−0.02
sci-Plex MCF7, 0.75	49	0.35 [0.37]	0.20 [0.21]	0.69	0.74	0.00

Table 27: Coverage floor (share of held-out signal energy outside the span of the training targets), four-gene error and its prediction, and the leverage diagnostic (share of held-out messages above the 95th percentile of training leverages; median ratio of held-out to mean training leverage), under random folds and one-pathway hold-out with coverage c.

Line	folds	coverage floor	four genes	four genes, pred.	flagged by leverage	median leverage ratio
sci-Plex A549	random	0.11	0.40	0.32	8%	0.34
sci-Plex A549	$c = 0$	0.47	0.93	0.36	8%	0.36
sci-Plex A549	$c = 0.25$	0.16	0.47	0.34	8%	0.34
sci-Plex A549	$c = 0.5$	0.12	0.38	0.34	9%	0.37
sci-Plex A549	$c = 0.75$	0.12	0.43	0.33	5%	0.40
sci-Plex K562	random	0.13	0.40	0.34	4%	0.40
sci-Plex K562	$c = 0$	0.33	0.70	0.36	12%	0.43
sci-Plex K562	$c = 0.25$	0.16	0.43	0.35	11%	0.41
sci-Plex K562	$c = 0.5$	0.14	0.39	0.34	9%	0.39
sci-Plex K562	$c = 0.75$	0.12	0.38	0.34	9%	0.35
sci-Plex MCF7	random	0.07	0.26	0.21	6%	0.39
sci-Plex MCF7	$c = 0$	0.38	0.68	0.23	10%	0.42
sci-Plex MCF7	$c = 0.25$	0.08	0.24	0.22	10%	0.37
sci-Plex MCF7	$c = 0.5$	0.06	0.20	0.22	10%	0.34
sci-Plex MCF7	$c = 0.75$	0.06	0.20	0.21	8%	0.39

Table 28: Zero-shot predictors on frozen pretrained representations: best model per representation, coverage of the perturbations, normalized residual error ε and measurement equivalent with 95% cluster-bootstrap intervals, and residual correlation r.

Screen	representation (best model)	coverage	ε	residual r	prior worth (genes)
Norman (101 singles)	ESM-2 protein (kernel ridge)	100%	0.98 [0.95, 1.02]	0.15	0.1 [0.0, 0.3]
	Geneformer gene (kernel ridge)	100%	1.00 [1.00, 1.00]	0.00	0.0 [0.0, 0.0]
Replogle	ESM-2 protein (kernel ridge)	99%	0.93 [0.91, 0.94]	0.16	0.7 [0.6, 0.9]
	Geneformer gene (ridge)	100%	0.93 [0.92, 0.95]	0.14	0.6 [0.5, 0.8]
sci-Plex A549	ChemBERTa molecule (kernel ridge)	100%	0.89 [0.83, 0.97]	0.11	0.1 [0.0, 0.2]
sci-Plex K562	ChemBERTa molecule (kernel ridge)	100%	0.90 [0.85, 0.96]	0.15	0.2 [0.1, 0.3]
sci-Plex MCF7	ChemBERTa molecule (kernel ridge)	100%	0.88 [0.81, 1.00]	0.16	0.1 [0.0, 0.2]

Table 29: Best zero-shot predictors (ε and residual correlation r). MLPs were run on the first repeat of the folds; the other models on all folds.

Screen	predictor (side information: model)	ε	residual r
Norman (101 singles)	GO: MLP 256×2	0.93	0.18
	GO: kernel ridge	0.95	0.21
	GO: 20-NN	0.96	0.17
	GenePT: kernel ridge	0.97	0.15
	GO: MLP 1024×3	0.97	0.17
	GO: MLP 256×1	0.97	0.16
	mean	1.00	0.00
Replogle	GO: ridge	0.71	0.29
	GO: kernel ridge	0.72	0.28
	GO: 20-NN	0.72	0.28
	GO: MLP 256×2	0.76	0.26
	GO: MLP 1024×3	0.76	0.26
	GO: MLP 64×1	0.78	0.24
	mean	1.00	0.00
sci-Plex A549	Morgan: kernel ridge	0.71	0.17
	target: group mean	0.76	0.19
	Morgan: MLP 1024×3	0.76	0.13
	Morgan: 20-NN	0.77	0.13
	pathway (level 2): group mean	0.78	0.21
	Morgan: MLP 256×2	0.80	0.09
	mean	1.00	0.00
sci-Plex K562	Morgan: kernel ridge	0.76	0.19
	pathway (level 2): group mean	0.85	0.20
	target: group mean	0.86	0.17
	Morgan: MLP 1024×3	0.87	0.14
	Morgan: 20-NN	0.87	0.10
	Morgan: MLP 1024×2	0.88	0.13
	mean	1.00	0.00
sci-Plex MCF7	Morgan: kernel ridge	0.70	0.25
	target: group mean	0.78	0.22
	Morgan: ridge	0.80	0.23
	Morgan: 20-NN	0.81	0.10
	Morgan: MLP 1024×3	0.82	0.17
	pathway (level 2): group mean	0.83	0.22
	mean	1.00	0.00

Table 30: Zero-shot capacity on identical folds (first repeat): ε of the best kernel or linear predictor and of the best of six MLPs on the same side information, with the original mean-response start and early stopping, and from the default initialization with early stopping or trained for 1,000 epochs.

	best kernel or linear	best of six MLPs (1.5×10^5 to 7×10^6 parameters)		
Screen	predictor (first repeat)	mean-response start, early stop	default start, early stop	default start, 1,000 epochs
Norman (101 singles)	0.94 (GO kernel ridge)	0.93	0.98	1.08
Replogle	0.71 (GO ridge)	0.76	0.80	0.99
sci-Plex A549	0.70 (Morgan kernel ridge)	0.76	0.80	0.70
sci-Plex K562	0.76 (Morgan kernel ridge)	0.87	0.89	0.82
sci-Plex MCF7	0.72 (Morgan kernel ridge)	0.82	0.87	0.78

Table 31: Priors and measured genes together (first repeat, ε): the best prior (kernel ridge or ridge on its side information), greedy one- and four-gene panels, and panels plus the prior's prediction.

Screen	prior	1 gene	prior + 1 gene	4 genes	prior + 4 genes
Norman (101 singles)	0.96	0.90	0.88	0.61	0.62
Replogle	0.71	0.91	0.65	0.68	0.53
sci-Plex A549	0.71	0.42	0.40	0.40	0.38
sci-Plex K562	0.78	0.56	0.52	0.40	0.40
sci-Plex MCF7	0.74	0.37	0.34	0.26	0.25

Table 32: Pearson correlation between predicted and observed response (change from control), all genes / 20 genes with the largest observed change, averaged over perturbations (first repeat).

Screen	mean response	best prior	1 gene	4 genes	GEARS
Norman (101 singles)	0.48 / 0.51	0.47 / 0.57	0.50 / 0.54	0.63 / 0.74	0.37 / 0.51
Replogle	0.27 / 0.45	0.35 / 0.58	0.30 / 0.50	0.36 / 0.60	n/a
sci-Plex A549	0.33 / 0.48	0.34 / 0.51	0.38 / 0.52	0.40 / 0.53	n/a
sci-Plex K562	0.31 / 0.52	0.31 / 0.53	0.39 / 0.56	0.42 / 0.59	n/a
sci-Plex MCF7	0.36 / 0.56	0.34 / 0.54	0.45 / 0.59	0.50 / 0.64	n/a

Table 33: Normalized residual error ε for a fixed k-gene message and decoders of increasing capacity (first repeat of the folds; RBF: kernel ridge with kernel $\exp(-s\|z - z'\|^2)$ on standardized messages).

Screen	k	linear (OLS)	RBF (s=0.1)	RBF (s=0.5)	MLP 64×1	MLP 256×2	MLP 1024×3
Norman (all 226)	1	0.79	0.75	0.75	0.76	0.76	0.77
	4	0.43	0.39	0.44	0.42	0.44	0.46
	16	0.28	0.32	0.62	0.30	0.31	0.31
sci-Plex A549	1	0.42	0.41	0.43	0.43	0.44	0.44
	4	0.40	0.34	0.47	0.42	0.42	0.39
	16	0.43	0.50	0.92	0.39	0.42	0.39
sci-Plex MCF7	1	0.37	0.37	0.39	0.36	0.38	0.37
	4	0.26	0.23	0.32	0.25	0.27	0.29
	16	0.20	0.26	0.67	0.18	0.18	0.17

Table 34: Fold-averaged components of the zero-bit distortion (n: training perturbations per fold). The components are training estimates; D_0 is the observed held-out distortion of the mean response, which they match to within about 1%.

Screen	P	n	$\mathrm{tr}\,\hat{\Sigma}$ (signal)	cell noise	batch noise	D_0
Norman (all 226)	226	181	12.46	2.34	0.00	14.74
Norman (101 singles)	101	81	7.70	1.36	0.00	9.01
Replogle	1617	1294	4.71	5.15	0.00	9.77
sci-Plex A549	165	132	17.64	13.94	5.22	36.99
sci-Plex K562	181	145	12.55	13.06	2.63	28.44
sci-Plex MCF7	181	145	14.69	3.19	2.21	20.33