CREST: CONTRASTIVE METRIC LEARNING FOR PERTURBATION MODELING AND RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

In target and drug discovery, the vast combinatorial space of genetic and chemical perturbations makes exhaustive screening intractable, positioning candidate retrieval – the *in silico* prioritization of promising conditions for physical validation – as a central practical application. Yet current models frequently fail to generalize across unseen cellular contexts, suffering from mode collapse, a behaviour which is encouraged by pointwise reconstruction losses that lack geometric constraints to separate distinct perturbation signatures. To address this, we present CREST (Contrastive Retrieval of Expression State Transitions), a metric learning framework that co-embeds context-conditioned perturbation queries and measured transcriptional responses in a shared space via contrastive alignment. On LINCS L1000 profiles, CREST substantially outperforms all evaluated baselines on candidate retrieval under cell line holdout, and as context diversity scales up to 86 training cell lines, CREST exhibits near-linear gains in retrieval, while capacity-matched pointwise models remain nearly flat. CREST's discriminative space flexibly extends to transcriptional response prediction: retrieving nearest-neighbor reference profiles for unseen queries yields response estimates that outperform dedicated models. Moreover, simply swapping the reference profiles at inference enables the frozen, bulk-trained CREST to transfer zero-shot to single-cell Perturb-seq, outperforming specialized single-cell models across unseen conditions. Overall, CREST establishes contrastive metric learning as a scalable foundation that unifies candidate prioritization and perturbation response prediction within a single geometric framework.

1 INTRODUCTION

Identifying perturbations that induce specific phenotypic responses in cells is a fundamental challenge in target and drug discovery. While recent advances in high-throughput screening enable transcriptomic profiling of large-scale chemical and genetic perturbations (Replogle et al., 2022; Nadig et al., 2025; Zhang et al., 2025), profiling the full combinatorial space of perturbations and cellular contexts remains intractable. This bottleneck has catalyzed the development of predictive models designed to generalize transcriptional responses to unseen contexts, including held-out cell lines, states, or backgrounds, so that *in silico* rankings can prioritize a small set of candidates for physical validation (Rood et al., 2024; Dimitrov et al., 2026).

Recent analyses argue that progress towards designing virtual cell models is currently limited by insufficient coverage of biological contexts (Dibaeinia et al., 2026). Consistent with this view, prominent benchmarking frameworks such as PerturBench (Wu et al., 2025), Systema (Viñas Torné et al., 2025), and the Virtual Cell Challenge (Roohani et al., 2025) find that current models fail to generalize to unseen contexts, underperform simple baselines, and suffer from mode collapse, in which predictions become nearly indistinguishable regardless of perturbations or contexts (Wu et al., 2025; Viñas Torné et al., 2025).

We reason that this behavior stems in part from the dominant training paradigm. Despite architectural differences, many existing methods optimize pointwise reconstruction losses that score each prediction in isolation (Dimitrov et al., 2026). Such objectives impose no geometric constraints to separate distinct perturbation–context signatures, making mode collapse a valid global optimum of the training objective. Indeed, a parallel line of benchmarking focusing on transcriptional response

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.

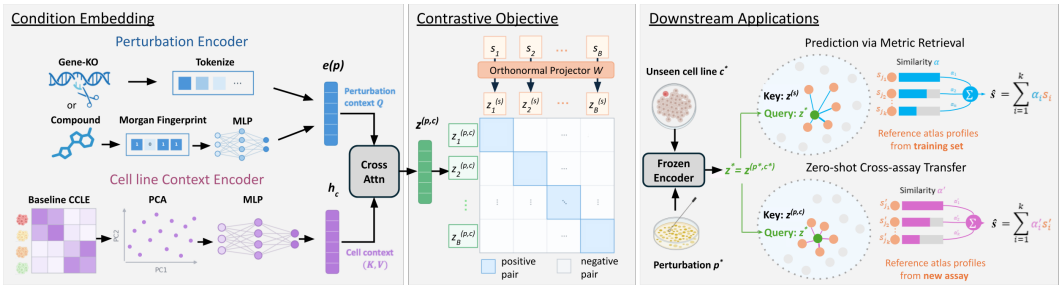


Figure 1: **CREST is a contrastive metric learning framework for perturbation modeling.** A perturbation encoder and a cell line context encoder are fused by a cross-attention conditioner into a query representation $z^{(p,c)}$; an orthonormal projector maps measured responses into the same space; a symmetric contrastive objective aligns query embeddings with response representations; and prediction is softmax-weighted nearest-neighbor retrieval over a reference atlas of measured responses, unifying training, candidate prioritization, and response evaluation within a single geometry.

prediction showed that deep architectures frequently fail to reliably outperform a trivial training set mean baseline on standard reconstruction metrics (Ahlmann-Eltze et al., 2025; Kernfeld et al., 2025; Csendes et al., 2025).

Addressing this limitation suggests moving beyond pointwise regression toward objectives that explicitly incentivize discriminative separation across perturbation profiles. To this end, we introduce CREST (Contrastive Retrieval of Expression State Transitions), a metric learning framework that co-embeds perturbation–context queries and transcriptional responses in a shared space trained with contrastive alignment. A context-conditioned query encoder pulls together matched perturbation–context and response pairs while pushing mismatched pairs apart. The resulting representation space is a metric over perturbation–context pairs, reducing candidate prioritization to a comparison of distances within it.

As a proof-of-concept, we instantiate CREST on post-perturbation LINCS L1000 transcriptional profiles from CMap (Subramanian et al., 2017), which spans dozens of cell lines and unifies genetic and chemical perturbations in a single harmonized resource. Crucially, our contrastive formulation is data-agnostic and extends naturally to higher-resolution scRNA-seq readouts, such as multi-context single-cell atlases.

Our contributions are as follows:

- **A context-conditioned metric learning framework.** We recast perturbation modeling from pointwise regression to metric learning over perturbation–context inputs and transcriptional responses. Contrastive alignment against an interpretable linear response projector overcomes mode collapse and enables similarity-based candidate ranking, substantially outperforming all evaluated baselines on retrieval under cell line holdout (Section 2).

- **Scaling with context diversity.** Across training sets ranging from a small number of cell lines to the full 86-line training panel, CREST exhibits near-linear improvement in retrieval, whereas classical regressors show shallow gains and capacity-matched pointwise models remain nearly flat, indicating that cross-context improvements are driven by the contrastive objective, rather than dataset volume alone (Section 4).

- **Prediction and cross-assay transfer via metric retrieval.** Without a regression decoder, CREST predicts responses for unseen conditions by retrieving nearest neighbors from a reference atlas in the learned metric space. Swapping the atlas at inference time enables zero-shot transfer of a frozen L1000-trained model to single-cell Perturb-seq. CREST outperforms competing methods on genetic and chemical L1000 prediction, and exceeds specialized single-cell and foundation models under zero-shot transfer (Section 4).

2 METHOD

2.1 PROBLEM FORMULATION

We consider perturbation screens measured across multiple cellular contexts. Each observation is a triple (p, c, s), where p denotes a perturbation, such as a genetic knockout or a small molecule treatment, c denotes the cellular context, and $s \in \mathbb{R}^G$ represents the resulting G-dimensional transcriptional response signature. Each cell line is featurized by a context vector derived from its baseline unperturbed expression profile. We study a cell line holdout setting: given a perturbation p and an unseen cellular context c, the goal is to rank and retrieve candidate responses by their compatibility with (p, c) and use the retrieved profiles to predict the transcriptional response $\hat{s}$.

Rather than directly learning a pointwise mapping $(p, c) \mapsto s$, we map perturbation–context conditions and measured responses into a shared d-dimensional metric space. A query encoder maps a condition (p, c) to a query embedding $z^{(p,c)}$, while a response projector maps a measured response profile s to its embedding $z^{(s)} = Ws$ (Section 2.2). Training aligns matched query–response pairs $(z_i^{(p,c)}, z_i^{(s)})$ and separates mismatched pairs through a contrastive objective (Section 2.3). This learned geometry supports two downstream use cases: (1) Response prediction: candidate response profiles s_j from a reference atlas $\mathcal{R} = \{s_j\}_{j=1}^N$ are ranked by the cosine similarity of their projections $z_j^{(s)} = Ws_j$ to the query embedding $z^{(p,c)}$, and a weighted combination of the top-k retrieved profiles is used to estimate the transcriptional response $\hat{s}$ (Section 2.4). (2) Cross-assay generalization: because prediction is retrieval-based rather than tied to a fixed parametric decoder, the same frozen condition encoder can be transferred to a new assay by replacing the reference atlas and retrieving directly between perturbation–context query embeddings $z^{(p,c)}$, enabling zero-shot cross-assay prediction without retraining or fine-tuning (Section 2.5).

2.2 ENCODER ARCHITECTURES

Perturbation encoder. The perturbation encoder maps p to an embedding $e \in \mathbb{R}^d$ using perturbation modality-specific input branches. For a gene knockout (KO), $e(p)$ is a trainable token embedding vector indexed from a learned lookup table, whereas for a small molecule treatment, $e(p)$ is an MLP trained on top of a Morgan fingerprint representation (2048 bits, radius $r = 3$) (Rogers & Hahn, 2010). In each case, $e(p)$ is L^2 normalized prior to context conditioning, yielding $\tilde{e}(p)$.

Cell line context encoder. A cellular context c is featurized from its baseline gene expression profile x_c into a context representation $h_c = \mathrm{MLP}(x_c)$. Baseline gene expression provides the primary bridge across biological contexts, enabling CREST to position unobserved cell lines relative to the screened panel based on underlying transcriptional similarity. Complete preprocessing and architectural specifications are provided in Appendix A.

Gated cross-attention conditioner. To capture context-dependent perturbation effects, the perturbation embedding is conditioned on the cell line context vector h_c (Figure 1). We partition h_c into $m = 64$ sub-vectors $h^{(1)}, \ldots, h^{(m)}$ and project each via a shared linear map U to construct context tokens $T(h_c) = [\,Uh^{(1)}; \ldots; Uh^{(m)}\,] \in \mathbb{R}^{m \times d}$. A cross-attention module (Vaswani et al., 2017) then treats the normalized perturbation embedding $\tilde{e}(p)$ as the query and the context tokens $T(h_c)$ as keys and values. The resulting context shift is scaled by a learned scalar gate g and added back to the base perturbation embedding:

$$z^{(p,c)} \;=\; \tilde{e}(p) \;+\; g \cdot \mathrm{CrossAttn}\big(\,\mathrm{query} = \tilde{e}(p),\;\; \mathrm{key} = \mathrm{value} = T(h_c)\,\big). \tag{1}$$

Initializing $g = 0$ (ReZero residual; Bachlechner et al., 2021) ensures that $z^{(p,c)}$ begins as an unconditioned perturbation embedding $\tilde{e}(p)$, incorporating cellular context only to the extent that it improves contrastive alignment. Under this formulation, $\tilde{e}(p)$ captures the perturbation effect shared across contexts, while the gated cross-attention term isolates the context-dependent variation across cell lines.

Orthonormal response projector. Transcriptional responses are mapped into the shared representation space via a linear projection $z^{(s)} = Ws$, where $W \in \mathbb{R}^{d \times G}$ has orthonormal rows,

enforced through a Householder parametrization (Mhammedi et al., 2017). Geometrically, W orthogonally projects response signatures $s \in \mathbb{R}^G$ onto row(W) before applying an exact isometry into $\mathbb{R}^d$, filtering out variation in null(W) while preserving inner products and norms of the retained signals (Appendix B). The projector also provides an interpretable measure of gene contribution: $(W^\top W)_{g,g} = \|W_{:,g}\|_2^2$ quantifies gene g's rotation-invariant contribution to the learned response subspace.

2.3 Contrastive alignment objective

Given a minibatch $\{(p_i, c_i, s_i)\}_{i=1}^B$, let $z_i^{(p,c)}$ be the query embedding for the ith perturbation–context pair, and let $z_i^{(s)}$ denote the embedding of its corresponding response projection $z_i^{(s)} = Ws_i$. We define the temperature-scaled cosine similarity between query i and response j as $\sigma_{ij} = \cos(z_i^{(p,c)}, z_j^{(s)})/\tau$ for $i, j \in \{1, \ldots, B\}$. The temperature τ is a learnable scalar initialized to $\log(\tau_0) = -2$. The symmetric InfoNCE alignment loss (van den Oord et al., 2018) is

$$\mathcal{L}_{\text{align}} = -\frac{1}{2B}\sum_{i=1}^{B}\left[\log\frac{e^{\sigma_{ii}}}{\sum_{j=1}^{B} e^{\sigma_{ij}}} + \log\frac{e^{\sigma_{ii}}}{\sum_{j=1}^{B} e^{\sigma_{ji}}}\right]. \tag{2}$$

This objective promotes each matched query–response pair $(z_i^{(p,c)}, z_i^{(s)})$ over mismatched pairs in both the query-to-response and response-to-query directions.

2.4 Prediction via metric retrieval

For response prediction, we construct a reference atlas $\mathcal{R} = \{(s_j, z_j^{(s)})\}_{j=1}^N$ of measured training responses s_j and their corresponding response embeddings $z_j^{(s)}$. Given an unseen query (p^*, c^*) with query embedding $z^* = z^{(p^*,c^*)}$, we retrieve the top-k responses by cosine similarity and predict $\hat{s}$ as a softmax-weighted combination of the top-k retrieved responses, with $k = 50$ and temperature $T = 0.1$:

$$\mathcal{N}_k = \underset{j\in\mathcal{R}}{\text{top-}k}\ \cos(z^*, z_j^{(s)}), \qquad \hat{s}(p^*, c^*) = \sum_{j\in\mathcal{N}_k} \frac{\exp\big(\cos(z^*, z_j^{(s)})/T\big)}{\sum_{j'\in\mathcal{N}_k}\exp\big(\cos(z^*, z_{j'}^{(s)})/T\big)} s_j. \tag{3}$$

2.5 Zero-shot cross-assay transfer

Decoupling representation learning from a fixed parametric decoder enables seamless transfer across experimental modalities. While response retrieval within the primary training modality (Section 2.4) uses the projector W, this matrix is inherently bound to a specific G-dimensional gene panel. To transfer a frozen CREST instance to a novel assay or platform $\mathcal{R}'$ (e.g., single-cell Perturb-seq) measured across a different feature space, we bypass W entirely and switch to *query-to-query (condition-to-condition) retrieval* in the shared space:

$$\mathcal{N}_k = \underset{j\in\mathcal{R}'}{\text{top-}k}\ \cos\big(z^*,\, z^{(p_j,c_j)}\big), \qquad \hat{s}(p^*, c^*) = \sum_{j\in\mathcal{N}_k} \frac{\exp\big(\cos(z^*,\, z^{(p_j,c_j)})/T\big)}{\sum_{j'\in\mathcal{N}_k}\exp\big(\cos(z^*,\, z^{(p_{j'},c_{j'})})/T\big)}\, s_j. \tag{4}$$

Operating between perturbation–context query embeddings $z^{(p,c)}$ decouples retrieval from the training response projector W and output gene space G. Because the perturbation encoder $e(p)$ featurizes both genetic knockouts and small molecules, this formulation is inherently modality-agnostic and extends naturally to both gene and compound perturbations. $z^{(p,c)}$ is computed strictly from perturbation–context inputs without observing response signatures, so the retrieved values s_j can belong to an entirely unobserved single-cell readout or gene panel. Adapting CREST to a new measurement platform only requires swapping the reference atlas $\mathcal{R}'$ at inference time, enabling zero-shot cross-assay prediction without retraining or fine-tuning.

3 Related Work

A growing family of models aims to predict transcriptional responses to perturbations across biological contexts, including specialized models such as State (Adduri et al., 2025), Stack (Dong et al.,

2026), HyperMap (Dhaka et al., 2026), TxPert (Wenkel et al., 2025), X-Cell (Wang et al., 2026), and PRESAGE (Littman et al., 2025), as well as tabular foundation models (Palla et al., 2026; Qu et al., 2026). To train these predictors, standard pointwise objectives like mean squared error (MSE) frequently suffer from signal dilution, where sparse and perturbation-specific differentially expressed genes (DEGs) are swamped by uninformative background variation, causing models to default to uninformative mean predictions and mode collapse (Wu et al., 2025). To counter this failure mode, Mejia et al. (2026) proposed DEG-aware objectives such as weighted MSE (wMSE), which scale gene-level errors by differential expression specificity to focus optimization on profile-separating signatures and reduce mode collapse.

Despite these methodological advances, recent benchmark evaluations demonstrate that low reconstruction error on pointwise metrics does not guarantee faithful recovery of context-specific biology or discriminative profile structure (Wu et al., 2025; Viñas Torné et al., 2025). A central challenge in cross-context prediction is isolating how a perturbation's shift varies across cell lines beyond a shared baseline effect. Because pointwise regression objectives evaluate predictions in isolation without comparing representations across conditions, they impose no explicit geometric constraints on embedding alignment. Contrastive learning has gained traction in biological representation learning, with frameworks such as MoCoP (Nguyen et al., 2023), CLOOME (Sanchez-Fernandez et al., 2023), MolPhenix (Fradkin et al., 2024), PhenoCompass (Wu et al., 2026), CellCLIP (Lu et al., 2026), and PertOmni (Liu et al., 2026) aligning molecular, bioimaging, text, and/or transcriptomic states. However, these methods typically align unconditioned entity representations without dynamically incorporating cellular background context into the query. In computer vision and multimodal machine learning, context-conditioned contrastive models address analogous challenges by conditioning query tokens on instance-level context features prior to contrastive alignment (Zhou et al., 2022; Rao et al., 2022). CREST bridges these paradigms: we introduce a context-conditioned query encoder that uses gated cross-attention to dynamically modulate perturbation embeddings with cellular context, enabling direct contrastive alignment against measured transcriptional responses in a shared metric space.

4 RESULTS

4.1 TRAINING DATA AND EVALUATIONS

We assess CREST's ability to generalize to unobserved cellular contexts by training it on a curated dataset derived from preprocessed LINCS L1000 transcriptional response profiles (Level 5) (Subramanian et al., 2017), featurized as moderated z-score signatures across $G = 10{,}174$ genes, spanning the 978 directly measured landmark genes alongside 9,196 computationally inferred genes. The dataset comprises $130{,}554$ distinct perturbation–context pairs, spanning $5{,}140$ CRISPR gene KO and $9{,}665$ compound perturbations profiled across 91 cell lines. The cell lines are partitioned into 86 training lines, two validation lines, and three held-out test lines (Appendix A). All compound perturbations in the dataset were profiled at a uniform dose and duration ($10\,\mu$M for 24 hours). Because representation learning, candidate retrieval, and transcriptional response estimation share a single underlying metric space, we assess model performance along two complementary axes: (1) retrieval metrics to characterize the discriminative structure of the learned geometry and (2) perturbation prediction metrics to measure transcriptional response fidelity.

Retrieval metrics. To evaluate retrieval accuracy, we query a desired transcriptional response against a pool of candidate conditions to retrieve the true underlying perturbation–context identity that induces the response. Performance is measured using Hits@k, defined as the proportion of queries for which the ground-truth condition ranks within the top k candidates (Appendix C).

We evaluate candidate retrieval across 3 task formulations. Condition Retrieval queries a measured response against $1{,}000$ candidate conditions consisting of 1 true match plus 999 negative test pairs sampled uniformly from held-out profiles. Perturbation Retrieval tests perturbation selection within a specific cell line by querying a measured response against $1{,}000$ candidate perturbations screened in that cell line. Finally, Context retrieval queries a measured response against candidate conditions (c', p) spanning all test cell lines c' to assess whether the model isolates context-specific effects.

Perturbation prediction metrics. To evaluate transcriptional response predictions, we adopt the Pearson Δ correlation, a metric introduced in the Systema framework (Viñas Torné et al., 2025) to explicitly emphasize perturbation-specific effects by removing context-baseline expression shifts. For each test profile, we isolate the perturbation-specific response by subtracting the within-context reference mean and calculate the Pearson correlation between predicted and measured shifts. Pearson Δ correlations are calculated across the full $10{,}174$-gene output space, averaged across perturbations, and macro-averaged over test cell lines.

We also adopt the transposed-rank metric proposed by PerturBench to specifically evaluate mode collapse for perturbation predictions (Wu et al., 2025). For a given perturbation, this metric measures the normalized rank of the predicted response relative to all other ground-truth profiles within the same cell line and perturbation modality. A score of 0 indicates an ideal match, whereas 0.5 corresponds to the expected performance of a random predictor. We refer to this metric as the discrimination rank metric.

Baseline models. We benchmark CREST against: (1) an architecturally matched backbone model trained with mean squared error (MSE) or the DEG-aware weighted MSE (wMSE) objective proposed in Mejia et al. (2026); (2) non-neural regressors (ridge regression (Hoerl & Kennard, 1970), random forest (Breiman, 2001), and XGBoost (Chen & Guestrin, 2016)); (3) a tabular foundation model (TabICL) (Palla et al., 2026); and (4) two context-agnostic reference predictors – the training set per-perturbation mean response (per-perturbation mean) and global training set mean response (global perturbation mean). Full baseline specifications are detailed in Appendix C.

4.2 Contrastive learning enhances candidate retrieval across unseen contexts

To support effective candidate prioritization across unseen cellular contexts, CREST constructs a shared metric space designed to resolve subtle biological variations. We first evaluate whether contrastive metric learning successfully structures the embedding space along the three aforementioned retrieval axes (Table 1). CREST's performance across all three tasks substantially exceeds random expectation, with condition and perturbation retrieval accuracy surpassing random chance by over $50\times$. Crucially, these strong non-trivial hit rates confirm that CREST's contrastive metric learning avoids mode collapse, successfully mapping distinct perturbation–context pairs into separable regions of the shared embedding space.

Model	Condition Retrieval@10 (↑)	Perturbation Retrieval@10 (↑)	Context Retrieval@1 (↑)
CREST	$\mathbf{0.608}_{\pm 0.009}$	$\mathbf{0.579}_{\pm 0.001}$	$\mathbf{0.564}_{\pm 0.051}$
Matched backbone (MSE)	$0.046_{\pm 0.003}$	$0.045_{\pm 0.004}$	$0.344_{\pm 0.004}$
Matched backbone (wMSE)	$0.069_{\pm 0.015}$	$0.069_{\pm 0.016}$	$0.337_{\pm 0.017}$
Ridge regression	0.099	0.101	0.345
Random forest	$0.164_{\pm 0.001}$	$0.163_{\pm 0.001}$	$0.329_{\pm 0.005}$
XGBoost	$0.118_{\pm 0.002}$	$0.119_{\pm 0.001}$	$0.333_{\pm 0.006}$
TabICL	$0.109_{\pm 0.001}$	$0.110_{\pm 0.002}$	$0.322_{\pm 0.003}$
Random predictor	0.010	0.010	0.333

Table 1: **Candidate retrieval accuracy on the LINCS L1000 cell-line holdout split** *Condition Retrieval@10* matches a response profile to its perturbation–context pair among $1{,}000$ candidate conditions. *Perturbation Retrieval@10* identifies the correct perturbation within the query cell line among $1{,}000$ candidates. *Context Retrieval@1* identifies which of the three held-out cell lines produced a given perturbation. Values are the mean $\pm$ SD (standard deviation) over three seeds.

In comparison, the architecturally identical variants trained with pointwise losses (MSE and wMSE) and remaining regression baselines exhibit markedly lower accuracy on this task. These models only marginally exceed random chance on condition and perturbation retrieval and drop to near-chance levels on context retrieval. Across both the condition and perturbation retrieval tasks, CREST outperforms the next-best baseline by at least $3.5\times$. These trends hold across both perturbation modalities, with particularly pronounced gains on gene KOs (Appendix D). This behavior highlights the structural limitation of pointwise objectives: because they evaluate predictions in isolation, they fail to

impose geometric constraints on representation alignment across conditions, leading to uninformative or collapsed geometries. Contrastive alignment directly overcomes this bottleneck, explicitly organizing the embedding space to preserve fine-grained signatures. Indeed, gene set enrichment analysis of the response projector weights, W, confirms that CREST structures this space around profile-separating biological mechanisms, prioritizing specific phenotypic processes (e.g., cell-cycle progression, kinase activity regulation), while down-weighting common-mode processes (e.g., cytoplasmic translation) (Section 2.2, Appendix E). This weighting of signatures translates directly into the learned geometry: perturbations sharing a mechanism of action (MoA) align into significantly tighter, more coherent clusters across contexts, confirming that CREST structures its embedding space around functional biology rather than context-specific artifacts (Figure 2, Appendix F).

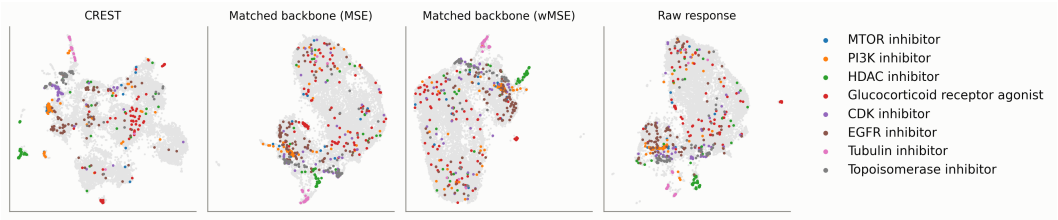


Figure 2: **UMAP visualization of compound response embeddings.** UMAP projections of test set compound responses across evaluated representations. Colored points highlight a representative subset of mechanism of action (MoA) classes.

4.3 Metric retrieval enables robust cross-context transcriptional response prediction

We next evaluate whether CREST's discriminative metric space enables accurate transcriptional response predictions when paired with non-parametric metric retrieval (Section 2.4, Table 2). CREST achieves substantial gains over competing approaches on the Pearson Δ correlation metric, outperforming all baseline models except random forests and XGBoost by at least 16% on genetic knockouts and 43% on compound predictions. These improvements are preserved when restricting the evaluation to the 978 directly measured L1000 landmark genes (Appendix G) and across each cell line (Appendix H).

The discrimination rank metric, which assesses mode collapse, reveals an even starker contrast. CREST's discrimination rank scores are 85% and 69% lower than the next-best baseline model for gene KO and compound predictions, respectively. Moreover, CREST is the only method to consistently outperform the per-perturbation mean baseline, both on the Pearson Δ and the discrimination rank metric. Every baseline model fails for at least one perturbation modality, highlighting their failure to capture meaningful context-specific variation.

These results illustrate a fundamental strength of the underlying architecture: through contrastive metric learning, CREST builds a rich, discriminative representation space capable of separating subtle context-dependent responses. In doing so, it effectively prevents representation collapse, enabling robust generalization of fine-grained signatures across unobserved cellular contexts. Finally, CREST can leverage this geometry through non-parametric reference atlas lookup to produce high-fidelity transcriptional predictions despite never optimizing for expression reconstruction.

4.4 Retrieval quality and predictive accuracy scale with context diversity

To evaluate how CREST and the baseline methods scale with context diversity, we downsample training set cell lines across five training set fractions and evaluate performance on the same evaluation tasks as above (Figure 3, Appendix I). CREST consistently outperforms all competing paradigms across every training fraction on both zero-shot Condition Retrieval@10 and perturbation prediction via discrimination rank. Crucially, CREST achieves a substantially steeper scaling trajectory than any baseline; on Condition Retrieval@10, CREST scales near-linearly from 0.104 to 0.608, whereas classical models display shallow, decelerating gains and capacity-matched regression backbones remain virtually flat. CREST also exhibits strong sample efficiency in low-data regimes: at just 7.6% context fraction, its discrimination rank (0.335) already surpasses full-dataset

Model	Pearson Δ correlation (↑)		Discrimination rank (↓)	
	Gene-KO	Compound	Gene-KO	Compound
CREST	$\mathbf{0.205}_{\pm 0.005}$	$\mathbf{0.191}_{\pm 0.003}$	$\mathbf{0.038}_{\pm 0.001}$	$\mathbf{0.112}_{\pm 0.002}$
Matched backbone (MSE)	$0.117_{\pm 0.010}$	$0.131_{\pm 0.006}$	$0.392_{\pm 0.014}$	$0.410_{\pm 0.014}$
Matched backbone (wMSE)	$0.146_{\pm 0.016}$	$0.133_{\pm 0.007}$	$0.343_{\pm 0.025}$	$0.367_{\pm 0.011}$
Ridge regression	0.176	0.044	0.252	0.459
Random forest	$0.189_{\pm 0.001}$	$0.157_{\pm 0.001}$	$0.277_{\pm 0.000}$	$0.365_{\pm 0.002}$
XGBoost	$0.190_{\pm 0.003}$	$0.093_{\pm 0.004}$	$0.282_{\pm 0.001}$	$0.455_{\pm 0.002}$
TabICL	$0.136_{\pm 0.002}$	$0.105_{\pm 0.006}$	$0.328_{\pm 0.005}$	$0.396_{\pm 0.005}$
Per-perturbation mean	0.171	0.167	0.224	0.269
Global mean	−0.007	0.019	0.500	0.500
Oracle	0.551	0.540	0	0

Table 2: **Cross-context transcriptional response prediction on the LINCS L1000 cell line hold-out split.** Pearson Δ correlation and discrimination rank by perturbation modality. Values are the mean ± SD over three seeds; ridge regression and oracle results are deterministic.

matched backbone models (MSE, 0.395; wMSE, 0.346) and approaches the full-dataset version of the next-best predictor (random forest, 0.315). Similar trends are observed across the other retrieval and perturbation prediction tasks from above (Appendix I)

This scaling behavior further highlights the distinction between contrastive metric learning and conventional regression pipelines: expanding context diversity allows CREST to learn increasingly refined geometric boundaries between distinct perturbation–context states. Conversely, regression-trained baselines exhibit virtually flat scaling curves, failing to convert additional context diversity into discriminative latent structure. This demonstrates that cross-context generalization gains are driven primarily by the contrastive alignment objective rather than dataset volume.

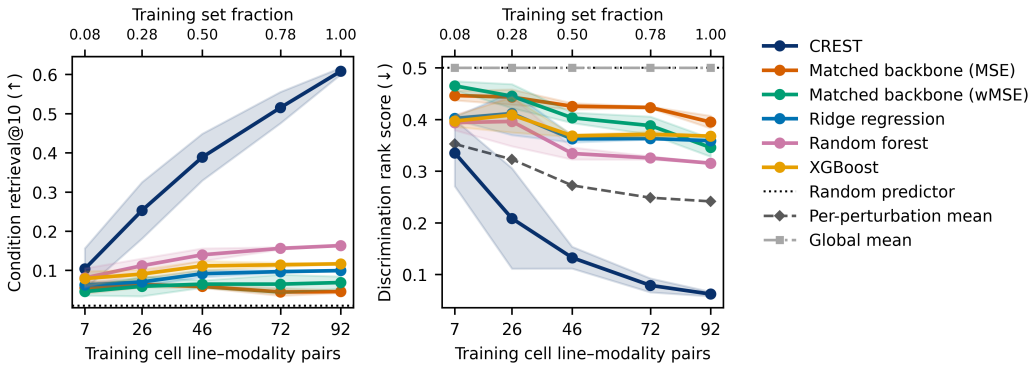


Figure 3: **Data scaling on retrieval and perturbation prediction performance.** Panels show Condition Retrieval@10 and discrimination rank score as a function of training context diversity and size. Points show the mean over three seeds, and shaded bands show ± 1 SD across seeds.

4.5 Zero-shot cross-assay transfer to single-cell Perturb-seq

A key advantage of decoupling representation learning from parametric decoders is the capacity for zero-shot cross-platform generalization (Section 2.5). To test this, we evaluate the frozen, bulk L1000-trained CREST model on the single-cell Replogle RPE-1 Perturb-seq benchmark curated by Ahlmann-Eltze et al. (2025). Notably, RPE-1 is absent from the LINCS L1000 training set. Transcriptional responses are predicted for unseen gene KOs using our non-parametric metric retrieval framework, querying candidate perturbation–context embeddings to retrieve nearest-neighbor response profiles directly from the RPE-1 training set without fine-tuning any parameters. We compare its performance against pre-computed predictions from published model outputs, re-scored under our unified metric suite (details in Appendix J).

CREST achieves the highest Pearson Δ correlation, outperforming the linear prediction model (LPM) from Ahlmann-Eltze et al. (2025) as well as published single-cell foundation models over the subset of perturbations covered by all models (Table 3). While discrimination rank scores reflect the challenge of zero-shot predictions across all architectures, CREST nonetheless yields the top performance. This advantage is preserved when expanding evaluation to a broader perturbation panel (Appendix J). The sole exception occurs on discrimination rank for the self-trained LPM, which benefits from direct in-domain training on the target RPE-1 dataset. These findings confirm that contrastive metric alignment captures core biological effects that can generalize across distinct cellular contexts and experimental modalities.

Model	Pearson Δ correlation ($\uparrow$)	Discrimination rank ($\downarrow$)
CREST	$\mathbf{0.185}_{\pm 0.003}$	$\mathbf{0.413}_{\pm 0.021}$
Linear prediction model (Ahlmann-Eltze et al., 2025)	$0.142_{\pm 0.025}$	$0.454_{\pm 0.015}$
GEARS (Roohani et al., 2024)	$0.053_{\pm 0.126}$	$0.464_{\pm 0.022}$
scGPT (Cui et al., 2024)	$0.125_{\pm 0.047}$	$0.446_{\pm 0.009}$
Geneformer (Theodoris et al., 2023)	$0.070_{\pm 0.014}$	$0.488_{\pm 0.000}$
CPA (Lotfollahi et al., 2023)	$0.036_{\pm 0.014}$	$0.499_{\pm 0.001}$
UCE (Rosen et al., 2026)	$0.009_{\pm 0.022}$	$0.500_{\pm 0.001}$
scBERT (Yang et al., 2022)	$-0.004_{\pm 0.015}$	$0.500_{\pm 0.000}$
Mean baseline	$-0.005_{\pm 0.016}$	$0.500_{\pm 0.000}$

Table 3: **Zero-shot L1000 cross-assay transfer to Perturb-seq.** Evaluation of a frozen, L1000-trained CREST against published baseline predictions on unseen perturbations conditions following the Ahlmann-Eltze et al. (2025) RPE-1 Perturb-seq benchmark task. Results are aggregated over the set of test perturbation conditions shared across all compared methods. Values are the mean $\pm$ SD across two train-test splits.

5 Discussion and Conclusion

Predicting cellular responses across unobserved combinations of perturbations and cellular contexts is central to accelerating target and drug discovery. CREST demonstrates that reformulating this problem as contrastive metric learning yields a shared geometry for candidate prioritization and perturbation response prediction. Empirically, this framing overcomes mode collapse under cell line holdout, substantially improves retrieval over capacity-matched pointwise models, and is the only evaluated method that consistently exceeds the context-blind per-perturbation mean on both Pearson Δ and discrimination rank, exhibiting near-linear performance improvements with increasing context diversity.

While CREST establishes a flexible, principled framework for perturbation modeling, several limitations point to clear next steps. First, CREST currently featurizes context from baseline CCLE expression. Richer descriptors such as mutational status, epigenetic state, and proteomic or metabolomic readouts could better capture effect modifiers beyond transcriptome similarity. Second, gene KOs are represented as learned token embeddings and compounds as Morgan fingerprints; incorporating biological priors into these encoders (e.g., aggregated gene embeddings (Littman et al., 2025; Wenkel et al., 2025), pre-trained compound representations) may improve generalization to rare or out-of-vocabulary perturbations. Third, a gap remains to the empirical oracle retrieval ceiling, and non-parametric prediction is bounded by atlas coverage: conditions lacking representative neighbors in the reference database cannot extrapolate unobserved transcriptional modes. Hybrid architectures that pair contrastive metric learning with parametric decoders are a natural solution. Finally, while zero-shot transfer from bulk data shows promising performance, fine-tuning or training CREST directly on expanding single-cell Perturb-seq atlases represents an immediate path toward enhancing resolution.

Looking forward, as multi-context perturbation atlases grow, contrastive metric learners such as CREST offer a flexible framework for indexing, querying, and transferring knowledge across expansive single-cell repositories, paving the way for high-throughput candidate prioritization and cross-assay perturbation modeling at scale.

AI USE STATEMENT

In this work, we used generative AI tools to propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, clean and reformat datasets, support qualitative and thematic data analysis, and interpret results. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, or assist in the writing of proofs. The following tasks were not applicable to this work: generate synthetic data sets, assist with translation. Additionally, we used generative AI tools to create or modify scientific figures or images, suggest experimental parameters, create or edit software code, draft parts of a research paper, discover research topics or identify gaps, brainstorm, source/search for information, edit a research paper to improve readability, identify relevant literature, propose a title or keywords for a research paper.

We have reviewed all AI-assisted work. All references were cross-referenced against original DOIs and publisher databases to ensure non-hallucinated citations. AI-assisted edits, LaTeX math, and rendered tables were audited line-by-line for scientific accuracy, original framing, and notation consistency. Mathematical equations and LaTeX rendering suggestions were manually verified for mathematical correctness and notation consistency across sections. LaTeX table formatting and plotting scripts produced with AI assistance were audited by cross-checking rendered figures and table values against raw output files. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Abhinav K. Adduri, Dhruv Gautam, Beatrice Bevilacqua, Alishba Imran, Rohan Shah, Mohsen Naghipourfar, Noam Teyssier, Rajesh Ilango, Sanjay Nagaraj, Mingze Dong, Chiara Ricci-Tam, Christopher Carpenter, Vishvak Subramanyam, Aidan Winters, Sravya Tirukkovular, Jeremy Sullivan, Brian S. Plosky, Basak Eraslan, Nicholas D. Youngblut, Jure Leskovec, Luke A. Gilbert, Silvana Konermann, Patrick D. Hsu, Alexander Dobin, Dave P. Burke, Hani Goodarzi, and Yusuf H. Roohani. Predicting cellular responses to perturbation across diverse contexts with State. *bioRxiv*, 2025. doi: 10.1101/2025.06.26.661135.

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.

Suzi A Aleksander, James P Balhoff, Seth Carbon, J Michael Cherry, Dustin Ebert, Marc Feuermann, Pascale Gaudet, Nomi L Harris, David P Hill, Patrick Kalita, et al. The gene ontology knowledgebase in 2026. *Nucleic Acids Research*, 54(D1):D1779–D1792, 2026. doi: 10.1093/nar/gkaf1292.

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.

Thomas Bachlechner, Bodhisattwa Prasad Majumder, Henry Mao, Gary Cottrell, and Julian McAuley. ReZero is all you need: Fast convergence at large depth. In Cassio de Campos and Marloes H. Maathuis (eds.), *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pp. 1352–1361. PMLR, 2021. URL https://proceedings.mlr.press/v161/bachlechner21a.html.

Jordi Barretina, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A. Margolin, Sungjoon Kim, Christopher J. Wilson, Joseph Lehár, Gregory V. Kryukov, Dmitriy Sonkin, Anupama Reddy, Manway Liu, Lauren Murray, Michael F. Berger, John E. Monahan, Paula Morais, Jodi Meltzer, Adam Korejwa, Judit Jané-Valbuena, Felipa A. Mapa, Joseph Thibault, Eva Bric-Furlong, Pichai Raman, Aaron Shipway, Ingo H. Engels, Jill Cheng, Guoying K. Yu, Jianjun Yu, Peter Aspesi, Melanie de Silva, Kalpana Jagtap, Michael D. Jones, Li Wang, Charles Hatton, Emanuele Palescandolo, Supriya Gupta, Scott Mahan, Carrie Sougnez, Robert C. Onofrio, Ted Liefeld, Laura MacConaill, Wendy Winckler, Michael Reich, Nanxin Li, Jill P. Mesirov, Stacey B. Gabriel, Gad Getz, Kristin Ardlie, Vivien Chan, Vic E. Myer, Barbara L. Weber, Jeff

Porter, Markus Warmuth, Peter Finan, Jennifer L. Harris, Matthew Meyerson, Todd R. Golub, Michael P. Morrissey, William R. Sellers, Robert Schlegel, and Levi A. Garraway. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483 (7391):603–607, 2012. doi: 10.1038/nature11003.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 785–794. ACM, 2016. doi: 10.1145/2939672.2939785.

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

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

Bhavya Dhaka, Jiahao Gao, and Trey Ideker. HyperMap: An efficient framework for transferring perturbation responses across diverse biological contexts. *bioRxiv*, 2026. doi: 10.64898/2026.04.23.720505.

Payam Dibaeinia, Sudarshan Babu, Mei Knudson, Ali ElSheikh, Yibo Wen, Han Liu, Jason Perera, and Aly A. Khan. Virtual cells need context, not just scale. *bioRxiv*, 2026.

Daniel Dimitrov, Stefan Schrod, Martin Rohbeck, and Oliver Stegle. Interpretation, extrapolation and perturbation of single cells. *Nature Reviews Genetics*, 27(5):349–370, 2026. doi: 10.1038/s41576-025-00920-4.

Mingze Dong, Abhinav Adduri, Dhruv Gautam, Lujing Wu, Courtney Kernick, Mary Margaret Coons, Yi-Chen Chih, Christopher Carpenter, Rohan Shah, Chiara Ricci-Tam, Po-Yuan Tung, Nianzhen Li, Alexander Dobin, Yuval Kluger, Dave Burke, Theodore Roth, and Yusuf Husein Roohani. Stack: In-context learning of single-cell biology. *bioRxiv*, 2026. doi: 10.64898/2026.01.09.698608. URL https://doi.org/10.64898/2026.01.09.698608.

Zhuoqing Fang, Xinyuan Liu, and Gary Peltz. GSEApy: a comprehensive package for performing gene set enrichment analysis in Python. *Bioinformatics*, 39(1):btac757, 2023.

Philip Fradkin, Puria Azadi, Karush Suri, Frederik Wenkel, Ali Bashashati, Maciej Sypetkowski, and Dominique Beaini. How molecules impact cells: Unlocking contrastive phenomolecular retrieval. *Advances in Neural Information Processing Systems*, 37:110667–110701, 2024.

Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Eric Kernfeld, Yunxiao Yang, Joshua S Weinstock, Alexis Battle, and Patrick Cahan. A comparison of computational methods for expression forecasting. *Genome biology*, 26(1):388, 2025. doi: 10.1186/s13059-025-03840-y.

Russell Littman, Jacob Levine, Sepideh Maleki, Yongju Lee, Vladimir Ermakov, Lin Qiu, Alexander Wu, Kexin Huang, Romain Lopez, Gabriele Scalia, Tommaso Biancalani, David Richmond, Aviv Regev, and Jan-Christian Hütter. Gene-embedding-based prediction and functional evaluation of perturbation expression responses with PRESAGE. *bioRxiv*, 2025. doi: 10.1101/2025.06.03.657653. URL https://www.biorxiv.org/content/10.1101/2025.06.03.657653v1.

Tianyu Liu, Artur Szałata, Wenxin Long, Fabian J Theis, Lingzhou Xue, and Hongyu Zhao. Learning perturbation effects through contrastive alignment of transcriptomics and textual embeddings. In *ICLR 2026 Workshop on Machine Learning for Genomics Explorations*, 2026.

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):MSB202211517, 2023.

Mingyu Lu, Ethan Weinberger, Chanwoo Kim, and Su-In Lee. Cellclip-learning perturbation effects in cell painting via text-guided contrastive learning. *Advances in Neural Information Processing Systems*, 38:124505–124537, 2026.

Gabriel Mateo Mejia, Henry E Miller, Francis JA Leblanc, Bo Wang, Brendan Swain, and Lucas Paulo de Lima Camillo. Needles in the haystack: Addressing signal dilution improves scRNA-seq perturbation response modeling and evaluation. In *Forty-third International Conference on Machine Learning*, 2026.

Zakaria Mhammedi, Andrew Hellicar, Ashfaqur Rahman, and James Bailey. Efficient orthogonal parametrisation of recurrent neural networks using Householder reflections. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2401–2409. PMLR, 06–11 Aug 2017.

Ajay Nadig, Joseph M. Replogle, Angela N. Pogson, Mukundh Murthy, Steven A. McCarroll, Jonathan S. Weissman, Elise B. Robinson, and Luke J. O'Connor. Transcriptome-wide analysis of differential expression in perturbation atlases. *Nature Genetics*, 57:1228–1237, 2025. doi: 10.1038/s41588-025-02169-3.

Cuong Q Nguyen, Dante Pertusi, and Kim M Branson. Molecule-morphology contrastive pretraining for transferable molecular representation. *bioRxiv*, pp. 2023–05, 2023.

Giovanni Palla, Alexander Hillsley, Yang-Joon Kim, and Loic A Royer. Tabular foundation models are competitive cellular perturbation predictors across biological scales. *bioRxiv*, pp. 2026–06, 2026. doi: 10.64898/2026.06.28.735106.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICLv2: A better, faster, scalable, and open tabular foundation model. *arXiv preprint arXiv:2602.11139*, 2026.

Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 18061–18070. IEEE, 2022.

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

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. doi: 10.1021/ci100050t.

Jennifer E Rood, Anna Hupalowska, and Aviv Regev. Toward a foundation model of causal cell and tissue biology with a perturbation cell and tissue atlas. *Cell*, 187(17):4520–4545, 2024.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nature biotechnology*, 42(6):927–935, 2024.

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, 2025.

Yanay Rosen, Yusuf Roohani, Ayush Agrawal, Leon Samotorčan, Tabula Sapiens Consortium, Stephen R Quake, and Jure Leskovec. Universal cell embedding provides a foundation model for cell biology. *Nature*, pp. 1–9, 2026.

Ana Sanchez-Fernandez, Elisabeth Rumetshofer, Sepp Hochreiter, and Günter Klambauer. Cloome: contrastive learning unlocks bioimaging databases for queries with chemical structures. *Nature Communications*, 14(1):7339, 2023.

Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the national academy of sciences*, 102(43):15545–15550, 2005.

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, David L. Lahr, Jodi E. Hirschman, Zihan Liu, Melanie Donahue, Bina Julian, Mariya Khan, David Wadden, Ian C. Smith, Daniel Lam, Arthur Liberzon, Courtney Toder, Mukta Bagul, Marek Orzechowski, Oana M. Enache, Federica Piccioni, Sarah A. Johnson, Nicholas J. Lyons, Alice H. Berger, Alykhan F. Shamji, Angela N. Brooks, Anita Vrcic, Corey Flynn, Jacqueline Rosains, David Y. Takeda, Roger Hu, Desiree Davison, Justin Lamb, Kristin Ardlie, Larson Hogstrom, Peyton Greenside, Nathanael S. Gray, Paul A. Clemons, Serena Silver, Xiaoyun Wu, Wen-Ning Zhao, Willis Read-Button, Xiaohua Wu, Stephen J. Haggarty, Lucienne V. Ronco, Jesse S. Boehm, Stuart L. Schreiber, John G. Doench, Joshua A. Bittker, David E. Root, Bang Wong, and Todd R. Golub. 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.

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, et al. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

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*, 2025. doi: 10.1038/s41587-025-02777-8.

C. Wang, M. Karimzadeh, N. G. Ravindra, L. R. Bounds, N. Alerasool, A. C. Huang, S. Ma, D. R. Gulbranson, H. Cui, Y. Lee, A. Arjavalingam, E. J. MacKrell, M. S. Wilken, J. Chen, B. W. Herken, J. A. Weber, M. M. Onesto, B. Gonzalez-Teran, N. F. Leung, S. Y. Shi, B. J. Smith, S. K. Lam, A. Barner, P. Wright, E. M. Rumsey, S. Kim, R. V. Sit, A. J. Litterman, C. Chu, and B. Wang. X-Cell: Scaling causal perturbation prediction across diverse cellular contexts via diffusion language models. *bioRxiv*, 2026. doi: 10.64898/2026.03.18.712807. URL https://www.biorxiv.org/content/10.64898/2026.03.18.712807v1. Preprint.

Frederik Wenkel, Wilson Tu, Cassandra Masschelein, Hamed Shirzad, Cian Eastwood, Shawn T. Whitfield, Ihab Bendidi, Craig Russell, Liam Hodgson, Yassir El Mesbahi, Jiarui Ding, Marta M. Fay, Berton Earnshaw, Emmanuel Noutahi, and Alisandra K. Denton. TxPert: Leveraging biochemical relationships for out-of-distribution transcriptomic perturbation prediction. *arXiv preprint arXiv:2505.14919*, 2025.

Alexander P Wu, Heming Yao, Burkhard Hoeckendorf, Garrett Gaskins, Nont Kosaisawe, Ziqing Liu, Phil Hanslovsky, Oleg Mayba, Nicholas Skelton, Gabriele Scalia, et al. Virtual phenotypic screening discovers novel scaffolds inhibiting the pi3k/mtor pathway. *bioRxiv*, pp. 2026–06, 2026.

Yan Wu, Esther Wershof, Sebastian 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. *Advances in Neural Information Processing Systems*, 38, 2025. doi: 10.52202/085713-3225.

Fan Yang, Wenchuan Wang, Fang Wang, Yuan Fang, Duyu Tang, Junzhou Huang, Hui Lu, and Jianhua Yao. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nature machine intelligence*, 4(10):852–866, 2022.

Jesse Zhang, Airol A Ubas, Richard De Borja, Valentine Svensson, Nicole Thomas, Neha Thakar, Ian Lai, Aidan Winters, Umair Khan, Matthew G Jones, et al. Tahoe-100m: A giga-scale single-cell perturbation atlas for context-dependent gene function and cellular modeling. *BioRxiv*, pp. 2025–02, 2025.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 16795–16804. IEEE, 2022.

A Model and Training Details

Cellular context featurization and processing. Across all models, cellular context c is featurized using baseline gene expression profiles x_c from Cancer Cell Line Encyclopedia (CCLE) (Barretina et al., 2012). To reduce dimensionality while preventing data leakage across splits, principal components analysis (PCA) is applied exclusively on the training set CCLE profiles, and we project all context profiles onto the first 512 principal components (PCs).

Model and training configuration. Table S1 details the hyperparameter configuration of CREST and its training pipeline. These settings are held constant across all reported random seeds.

Component	Setting
Embedding dimension d	128
Cell line context	Top 512 PCs of CCLE expression profiles (PCA fit on training set cell lines)
encoder MLP	$512 \to 256 \to 128 \to 128$
block operations	LayerNorm, GELU, Dropout
Conditioner	Gated cross-attention
context tokens m	64
attention heads	4
query normalization	LayerNorm
gate	Scalar, zero-initialization (ReZero)
Response projector	Householder-orthonormal
initialization	PCA of training set responses
Objective	Symmetric InfoNCE
temperature	Learned, $\log \tau_0 = -2.0$
Optimizer / learning rate	Adam / 1×10^{-3}
Batch size	8192
Early stopping	Validation set Condition Retrieval@1, patience of 15 epochs
Retrieval readout	Softmax-kNN, k=50, $T = 0.1$

Table S1: **CREST implementation and training configuration details.**

B Response Projector Geometry

Because $W \in \mathbb{R}^{d \times G}$ has orthonormal rows, it satisfies $WW^\top = I_d$. The symmetric matrix $P = W^\top W \in \mathbb{R}^{G \times G}$ defines an orthogonal projection operator onto the d-dimensional subspace spanned by the rows of W, denoted $\mathcal{S} = \text{row}(W)$. Any vector $s \in \mathcal{S}$ can be expressed as $s = W^\top x$ for some $x \in \mathbb{R}^d$, from which it follows that:

$$W^\top W s = W^\top W (W^\top x) = W^\top (W W^\top) x = W^\top x = s. \tag{5}$$

Consequently, for any two vectors $s, s' \in \mathcal{S}$, the inner product of their projections in $\mathbb{R}^d$ equals their original inner product in $\mathbb{R}^G$:

$$\langle Ws, Ws' \rangle = (Ws)^\top (Ws') = s^\top (W^\top W s') = s^\top s' = \langle s, s' \rangle. \tag{6}$$

Setting $s' = s$ establishes norm preservation $\|Ws\|_2 = \|s\|_2$, demonstrating that W acts as an isometry on $\text{row}(W)$.

C Baseline Details

Matched backbone regressors (MSE, wMSE). These baselines share CREST's exact architecture, replacing the non-parametric contrastive loss with a parametric regression loss on decoded expression profiles $\hat{s} = W^\top z$. Standard MSE minimizes unweighted reconstruction error: $\|W^\top z - s\|_2^2$. Following Mejia et al. (2026), the wMSE objective scales gene-level errors by a leave-one-out vs.-rest specificity score:

$$\mathcal{L}_{\text{wMSE}} = \left\| \sqrt{w} \odot (W^\top z - s) \right\|_2^2, \quad \text{where} \quad w = \text{minmax}\left(\left| \frac{\bar{s}_p - \mu_{\text{rest}}}{\sigma_{\text{rest}}} \right| \right)^2. \tag{7}$$

Here, $\bar{s}_p$ represents the mean expression profile for perturbation p, while μ_{rest} and σ_{rest} are the mean and standard deviation across all remaining perturbations. This objective focuses optimization on genes that distinguish a perturbation from background variation rather than broadly responsive, non-specific targets.

Non-neural regressors and tabular foundation model baselines (Ridge, Random Forest, XGBoost, TabICL). Non-neural regressors (Hoerl & Kennard, 1970; Breiman, 2001; Chen & Guestrin, 2016) and the tabular foundation model TabICL (Palla et al., 2026) map concatenated perturbation and cellular context vectors directly to target gene expression shifts without a shared embedding space. Across all evaluated models—including CREST, matched backbones, classical regressors, and TabICL—cellular context is identically featurized using the top 512 principal components of baseline CCLE gene expression profiles. Compounds are represented using 2048-bit Morgan fingerprints (radius $r = 3$). Gene knockout (KO) perturbations are featurized by first averaging their L1000 training expression profiles per knockout and then projecting these averaged profiles onto their top 256 principal components, aligning baseline data exposure with CREST's perturbation token learning. To handle the high-dimensionality of the transcriptional response output space, we follow the procedure used by Palla et al. (2026), in which PCA is applied to the training set response profiles, the response profiles are projected onto the first 128 principal components, and a separate model is fit for each PCA component. Models are trained independently on the genetic and compound perturbations.

Context-blind reference baselines. The global mean baseline predicts the unweighted average training response across all queries. The per-perturbation mean baseline predicts the mean response of the queried perturbation averaged across training cell lines, ignoring query context c^*.

Retrieval evaluations for baselines. Because the non-neural regressors and TabICL baselines lack a joint metric space, their retrieval metrics are computed directly in *expression space* using predicted transcriptional response profiles $\hat{s}$. For the matched backbone models (MSE, wMSE), evaluations were performed in the joint embedding space using $\cos(Ws, z)$. Because the Householder parametrization enforces orthonormal rows ($WW^\top = I$), the linear transformation W acts as an isometry on the response subspace, rendering cosine similarity ranking metrics identical whether evaluated in embedding space or decoded expression space.

Oracle for transcriptional response prediction. The oracle baseline defines an empirical performance ceiling for retrieval-based prediction by evaluating an ideal retriever with access to ground-truth test profiles. For each test profile s_{test} in a cellular context with reference mean $\bar{s}_{\text{ref}}$, the oracle computes the target shift $d_{\text{test}} = s_{\text{test}} - \bar{s}_{\text{ref}}$. Candidates consist of all training profiles $s_{\text{train},j}$ across the reference atlas cell lines, shifted with respect to the test context reference: $d_{\text{train},j} = s_{\text{train},j} - \bar{s}_{\text{ref}}$. Candidates are ranked by Pearson Δ correlation to d_{test} over the target gene set to align with the evaluation metric. The oracle prediction $\hat{s}$ is the unweighted average of the raw training profiles corresponding to the top-k candidates. We set $k = 10$, as averaging across multiple shift-matched training profiles reduces profile-level noise and provides a robust performance ceiling for retrieval-based prediction. Predictions are then evaluated against ground truth using standard Pearson Δ correlation.

D Retrieval Accuracy Stratified by Perturbation Modality

Table S2 reports candidate retrieval accuracy by perturbation modality.

Model	Gene-KO			Compound		
	Condition Ret@10	Perturbation Ret@10	Context Ret@1	Condition Ret@10	Perturbation Ret@10	Context Ret@1
CREST	$\mathbf{0.643}_{\pm 0.006}$	$\mathbf{0.614}_{\pm 0.004}$	$\mathbf{0.568}_{\pm 0.036}$	$\mathbf{0.456}_{\pm 0.011}$	$\mathbf{0.427}_{\pm 0.003}$	$\mathbf{0.559}_{\pm 0.093}$
Matched backbone (MSE)	$0.041_{\pm 0.005}$	$0.041_{\pm 0.004}$	$0.348_{\pm 0.003}$	$0.039_{\pm 0.008}$	$0.039_{\pm 0.008}$	$0.340_{\pm 0.005}$
Matched backbone (wMSE)	$0.046_{\pm 0.008}$	$0.048_{\pm 0.009}$	$0.342_{\pm 0.028}$	$0.066_{\pm 0.018}$	$0.066_{\pm 0.017}$	$0.331_{\pm 0.009}$
Ridge regression	0.130	0.130	0.354	0.050	0.051	0.335
Random forest	$0.125_{\pm 0.001}$	$0.127_{\pm 0.000}$	$0.322_{\pm 0.004}$	$0.149_{\pm 0.002}$	$0.148_{\pm 0.001}$	$0.338_{\pm 0.006}$
XGBoost	$0.121_{\pm 0.003}$	$0.125_{\pm 0.002}$	$0.323_{\pm 0.010}$	$0.077_{\pm 0.001}$	$0.079_{\pm 0.001}$	$0.343_{\pm 0.007}$
TabICL	$0.091_{\pm 0.001}$	$0.095_{\pm 0.001}$	$0.311_{\pm 0.005}$	$0.101_{\pm 0.002}$	$0.100_{\pm 0.002}$	$0.334_{\pm 0.001}$
Random predictor	0.010	0.010	0.333	0.010	0.010	0.333

Table S2: **Candidate retrieval accuracy on the LINCS L1000 cell-line holdout split, stratified by perturbation modality.** Values are the mean ± SD over three seeds.

E Response Projector Weight Analysis

To evaluate the response projector $W \in \mathbb{R}^{128\times 10174}$ across seeds, output genes were ranked by a consensus importance score, calculated as their mean column norm $\|W_{:,g}\|_2$ across seeds, reflecting their rotation-invariant contribution $(W^\top W)_{g,g}$ to the 128-dimensional subspace. Based on this consensus gene ranking, pathway enrichment was performed against Gene Ontology (GO) Biological Processes (Ashburner et al., 2000; Aleksander et al., 2026) using Preranked GSEA in GSEApy (Fang et al., 2023; Subramanian et al., 2005) (default parameters; 10,174 readout genes as background).

Genes with the highest weight importance in the response projector W are predominantly enriched for perturbation-driven signaling, led by positive regulation of protein Ser/Thr kinase activity (NES $= 3.21$) and mitotic cell-cycle phase transition (NES $= 2.98$). Conversely, widespread common-mode functions are de-emphasized, including cytoplasmic translation (NES $= -2.92$, FDR $q = 0.001$).

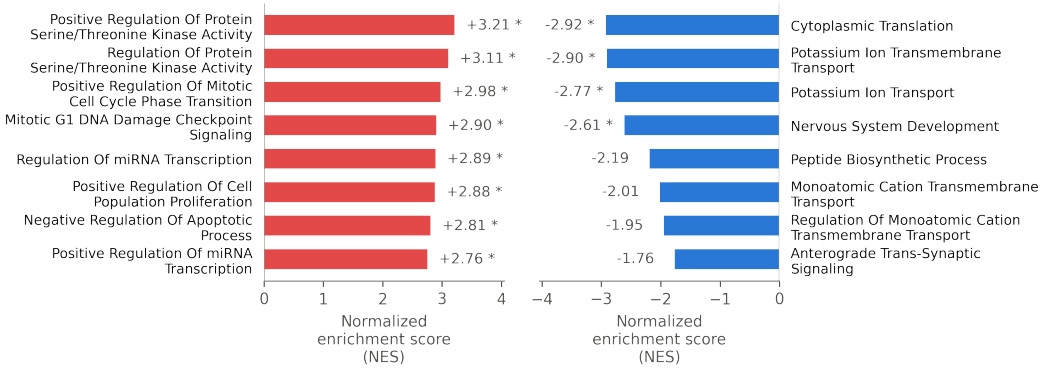


Figure S1: **Gene Ontology enrichment of response projector weights.** Top 8 enriched Gene Ontology (GO) Biological Process terms for genes with the highest (left) and lowest (right) consensus importance scores in the response projector W. Bar lengths indicate Normalized Enrichment Scores (NES) from Preranked GSEA. Asterisks (*) denote FDR $q < 0.05$.

F Cross-Context Biological Coherence of Response Projections

To evaluate whether CREST's learned response projector captures meaningful biology, we analyze the cross-cell-line similarity of compound perturbations sharing the same mechanism of action (MoA) in the test set. We expect that if a model organizes representations around profile-separating biological mechanisms, compounds within the same functional class should exhibit higher mutual similarity across distinct cell lines than expected by chance. We evaluate four response representations: CREST's response projections $z^{(s)} = Ws$, matched-backbone response projections trained with MSE and wMSE losses, and unprojected raw response profiles s. Evaluating projec-

tions through the response matrix W allows direct comparison between CREST and architecturally matched regression backbones across three independent training seeds.

For an MoA class S comprising k compounds, cross-cell-line coherence measures the average pairwise cosine similarity between *different* compounds measured in *different* cell lines. Specifically, for every pair of distinct compounds ($i \neq j$) across all pairs of distinct cell lines ($a < b$), we compute the cosine similarity of their response projections and average across all $3k(k-1)$ valid pairs: $C(S) = \frac{1}{3k(k-1)} \sum_{a<b} \sum_{i \neq j} \cos\left(z_{i,a}^{(s)}, z_{j,b}^{(s)}\right)$. Self-comparison pairs ($i = j$) are explicitly excluded so that an MoA class scores highly only when distinct drugs sharing a mechanism resemble one another across cell lines rather than when a single drug merely resembles itself.

To assess whether an observed coherence score $C(S)$ is statistically significant, we construct an empirical null distribution for each class size k by drawing $N_{\text{null}} = 50{,}000$ random k-compound sets from the annotated compound universe. Empirical p-values are computed as the proportion of null draws yielding a coherence score at least as extreme as $C(S)$, incorporating a standard pseudocount correction: $p = \frac{1 + \sum_{m=1}^{N_{\text{null}}} \mathbb{1}\left(C\left(S_{\text{null}}^{(m)}\right) \geq C(S)\right)}{N_{\text{null}}+1}$, where $\mathbb{1}(\cdot)$ is the indicator function evaluating to 1 when the condition holds and 0 otherwise.

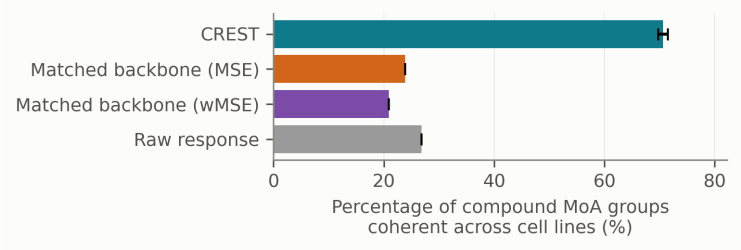


Figure S2: **Cross-cell-line coherence of compound MoAs.** Percentage of mechanism of action (MoA) classes exhibiting statistically significant pairwise cross-cell-line cosine similarity (FDR $q < 0.05$) across evaluated response representations. Error bars indicate ± 1 SD across three independent training seeds.

G Transcriptional Response Prediction Evaluation on L1000 Landmark Genes

Table S3 extends on Table 2 and reports cross-context transcriptional response prediction performance evaluated on only the L1000 landmark genes.

	Pearson Δ correlation(↑)	
Model	Gene-KO	Compound
CREST (retrieval)	$\mathbf{0.227}_{\pm 0.004}$	$\mathbf{0.223}_{\pm 0.003}$
Matched backbone (MSE)	$0.113_{\pm 0.009}$	$0.140_{\pm 0.011}$
Matched backbone (wMSE)	$0.130_{\pm 0.019}$	$0.116_{\pm 0.013}$
Ridge regression	0.201	0.072
Random Forest	$0.203_{\pm 0.000}$	$0.182_{\pm 0.001}$
XGBoost	$0.201_{\pm 0.002}$	$0.113_{\pm 0.001}$
TabICL	$0.156_{\pm 0.002}$	$0.133_{\pm 0.005}$
Per-perturbation mean	0.202	0.205
Global perturbation mean	−0.003	0.047
Oracle	0.494	0.491

Table S3: **Cross-context transcriptional response prediction on the LINCS L1000 cell line hold-out split split (L1000 landmark genes only).** Pearson Δ correlation (the per-profile Pearson correlation of predicted vs. measured perturbation-specific shift over the L1000 landmark genes), split by perturbation modality. Values are the mean ± sd over three seeds for seeded methods; ridge regression and oracle results are deterministic.

H Per-Cell Line Transcriptional Response Prediction Breakdown

Table S4 details Pearson Δ correlation performance broken down by individual test cell lines (A375, A549, MCF7). CREST consistently outperforms the context-blind per-perturbation mean baseline across all three test cell lines and both perturbation modalities (CRISPR gene KOs and small molecules). This uniform advantage confirms that overall performance gains reflect consistent cross-context generalization across diverse cellular contexts rather than single-cell line artifacts.

	A375		A549		MCF7	
Model	Gene KO	Compound	Gene KO	Compound	Gene KO	Compound
CREST	$\mathbf{0.218}_{\pm 0.006}$	$\mathbf{0.231}_{\pm 0.003}$	$\mathbf{0.244}_{\pm 0.006}$	$\mathbf{0.159}_{\pm 0.006}$	$\mathbf{0.151}_{\pm 0.004}$	$\mathbf{0.181}_{\pm 0.005}$
Matched backbone (MSE)	$0.133_{\pm 0.017}$	$0.170_{\pm 0.008}$	$0.135_{\pm 0.009}$	$0.102_{\pm 0.009}$	$0.084_{\pm 0.010}$	$0.120_{\pm 0.019}$
Matched backbone (wMSE)	$0.164_{\pm 0.018}$	$0.164_{\pm 0.005}$	$0.168_{\pm 0.019}$	$0.108_{\pm 0.006}$	$0.105_{\pm 0.011}$	$0.126_{\pm 0.012}$
Ridge regression	0.187	0.057	0.203	0.019	0.139	0.056
Random forest	$0.199_{\pm 0.001}$	$0.206_{\pm 0.001}$	$0.224_{\pm 0.001}$	$0.110_{\pm 0.005}$	$0.145_{\pm 0.001}$	$0.154_{\pm 0.001}$
XGBoost	$0.203_{\pm 0.001}$	$0.144_{\pm 0.010}$	$0.224_{\pm 0.006}$	$0.042_{\pm 0.005}$	$0.144_{\pm 0.001}$	$0.094_{\pm 0.005}$
TabICL	$0.148_{\pm 0.002}$	$0.130_{\pm 0.011}$	$0.155_{\pm 0.004}$	$0.078_{\pm 0.002}$	$0.107_{\pm 0.001}$	$0.108_{\pm 0.006}$
Per-perturbation mean	0.182	0.195	0.194	0.147	0.136	0.160
Global perturbation mean	−0.007	0.052	−0.003	0.006	−0.012	−0.001
Oracle	0.557	0.540	0.555	0.527	0.539	0.551

Table S4: **Per-cell line cross-context transcriptional response prediction on the LINCS L1000 cell line holdout split.** Pearson Δ correlation by held-out test cell line (A375, A549, MCF7) and perturbation modality. Values are the mean ± SD over three seeds; ridge regression and oracle results are deterministic.

I Retrieval Quality and Predictive Accuracy Scale with Context Diversity (Additional Evaluations)

Figure S3 extends the training context scaling analysis of Section 4.4 on the remaining evaluation tasks.

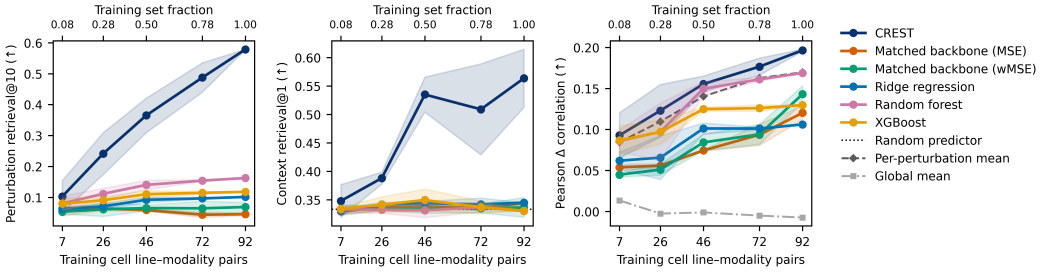


Figure S3: **Data scaling on retrieval and perturbation prediction performance.** Panels show Perturbation Retrieval@10, Context Retrieval@1, and Pearson Δ correlation as a function of training context diversity and size. Points show the mean over three seeds, and shaded bands show ± 1 SD across seeds.

J Zero-Shot Perturb-seq Transfer Details

We benchmark CREST against published predictions from Ahlmann-Eltze et al. (2025) (Zenodo `10.5281/zenodo.14832393`), covering GEARS, CPA, scGPT, Geneformer, UCE, scBERT, linear baselines, and a mean baseline evaluated on Replogle RPE-1 Perturb-seq data across two independent data seeds. Each seed split comprises 1,036 training, 115 validation, and 384 test perturbations evaluated over a fixed 5,000 highly variable gene panel. All models are scored on

identical held-out test sets against shared ground-truth profiles. To prevent data leakage, CREST's reference atlas is built strictly from each seed's training and validation perturbations.

Without retraining on the target dataset, we embed held-out RPE-1 perturbations using our frozen contrastive encoder and predict response profiles via softmax-weighted k-NN retrieval ($k = 50$, $T = 0.1$) against each seed's training bank. To accommodate varying perturbation overlap across models and our training vocabulary, we report two coverage tiers. Primary rankings in the main text rely on an intersection subset of 37 perturbations predicted across all methods. Here, we present an expanded subset of 80–81 perturbations that comprise the intersection of perturbations covered by CREST, GEARS, CPA, Geneformer, the linear prediction model, and the mean baseline.

Model	Pearson Δ correlation (↑)	Discrimination rank (↓)
CREST	$\mathbf{0.202}_{\pm 0.018}$	$\mathbf{0.392}_{\pm 0.010}$
Linear prediction model (Ahlmann-Eltze et al., 2025)	$0.153_{\pm 0.005}$	$0.189_{\pm 0.016}$
GEARS (Roohani et al., 2024)	$0.035_{\pm 0.121}$	$0.450_{\pm 0.018}$
Geneformer (Theodoris et al., 2023)	$0.003_{\pm 0.015}$	$0.493_{\pm 0.001}$
CPA (Lotfollahi et al., 2023)	$0.049_{\pm 0.010}$	$0.500_{\pm 0.001}$
Mean baseline	$-0.036_{\pm 0.007}$	$0.500_{\pm 0.000}$

Table S5: **Zero-shot L1000 cross-assay transfer to Perturb-seq.** Evaluation of a frozen, L1000-trained CREST without fine-tuning against published baseline predictions on unseen perturbations conditions following the Ahlmann-Eltze et al. (2025) RPE-1 Perturb-seq benchmark task. Results are aggregated over the set of test perturbation conditions shared across CREST , GEARS, Geneformer, and CPA; the published scGPT, UCE, and scBERT predictions featured a smaller subset of test perturbations. Values are the mean $\pm$ SD across two train-test splits.