CellVELA: Contextualizing Cell Foundation Models and LLMs via Retrieved Experimental Evidence

Anonymous authors
Paper under double-blind review

Abstract

Cell foundation models and large language models (LLMs) are increasingly adapted for functional genomics, where predictions often depend on cellular context rather than on genes alone. We study how such context should be transferred to unseen biological settings. Cancer vulnerability provides a stringent testbed: the same gene can be essential in one cancer and dispensable in another, yet standard target-ranking metrics strongly reward genes that are broadly essential. We separate this gene-level prior from context-specific discrimination and show that a fixed ranking can score well while carrying no contextual information. Parameter-efficient fine-tuning of the LLM, including a within-target contrast objective, fits the training contexts but remains near chance on held-out cancer lines across changes in model scale, interface, labels, and sampling. We therefore propose CellVELA, which retrieves matched CRISPR measurements as experimental evidence at inference. This restores held-out discrimination, improves with evidence similarity and quantity, vanishes under random retrieval, and recovers canonical lineage-specific dependencies. The same pattern extends to perturbation-response prediction, where retrieval quality matters more than reader complexity. These results suggest that context-specific functional prediction depends not only on model adaptation, but on preserving the right experimental evidence at inference.

1 Introduction

Cell foundation models learn reusable representations of cellular state from large expression atlases (Cui et al., 2024; Theodoris et al., 2023), while general-purpose LLMs provide flexible interfaces for reasoning over biological knowledge and measurements (Levine et al., 2024; Chen & Zou, 2025). These models are increasingly combined and adapted for functional genomics (Wu et al., 2025; Zhang et al., 2026), where the goal is not simply to describe a cell but to predict what happens when its molecular system is perturbed. Such predictions are inherently context-dependent: the functional effect of the same gene or perturbation changes across cell states, lineages and disease settings. The central question is therefore not only how to adapt these models, but how context-specific functional information should be supplied to them.

Cancer vulnerability prediction is a particularly stringent instance of this problem. Genome-wide CRISPR screens across hundreds of cancer lines (DepMap, 2024; Behan et al., 2019) show that some genes are broadly essential, whereas selective dependencies such as SOX10 in melanoma or ESR1 in breast cancer are confined to particular contexts (Figure 1a). Predicting vulnerability requires knowing not only which genes matter, but *where* each gene matters.

Standard target-ranking evaluation can blur this distinction. Models are scored by the genes they prioritize within each cancer line, but genes that are frequently essential remain good guesses in almost every line. In our benchmark, a single fixed gene ranking recovers two thirds of the conventional top-100 target score while containing no information about which cancer depends on which gene. A model can therefore score well without learning context at all.

CellVELA turns this failure mode into a controlled study of contextual adaptation. We compare three ways to supply context-specific functional information (Figure 1b): through a pretrained cell representation, a frozen cell-foundation-model embedding projected as soft tokens into the LLM;

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.

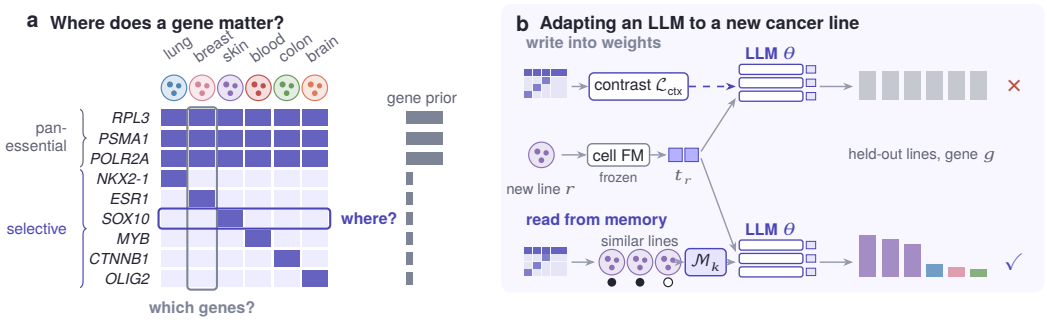

Figure 1: **Functional prediction requires knowing where a target matters.** **(a)** Pan-essential genes are needed in every cancer; selective dependencies, the drug targets, are confined to a lineage. Ranking genes within one cancer (column) is answered by the gene prior; ranking cancers for one gene (row) is not. **(b)** CellVELA reads a new cancer line through a frozen cell foundation model into tokens t_r. The matched screens can be written into the weights θ through a within-target contrast, which does not transfer to held-out lines, or read at inference as experimental memory $\mathcal{M}_k$ retrieved from similar lines, which ranks held-out lines by their dependency.

through the LoRA-adapted weights of the LLM; or through matched experimental evidence retrieved at inference, which we call *experimental memory*. The key comparison uses the same CRISPR screens in two ways: a within-target contrast objective attempts to encode contextual differences in the parameters, with the gene-level prior cancelling by construction, whereas retrieval keeps the measurements of the queried target in molecularly similar training lines directly accessible when a new context is scored. This lets us ask which route actually carries functional context to held-out biological settings. Cancer dependency is the primary testbed; perturbation response (Wu et al., 2025) provides a second functional assay.

The comparison yields a consistent picture. Parameter adaptation learns the contextual contrast on training cancer lines but does not transfer it to held-out lines, across model scale, interface width, label construction and sampling; probes show that the contextual signal reaches the model inputs but is weakened at the answer. Experimental memory restores held-out functional discrimination: it grows with record similarity and number, disappears under random retrieval, and recovers canonical lineage dependencies. The same pattern extends to perturbation response, where changing the retrieved evidence matters more than changing the reader. As for long-tail knowledge in language models (Kandpal et al., 2023; Ovadia et al., 2024; Lewis et al., 2020), context-specific functional adaptation is thus not only a parameter-learning problem but an information-placement problem.

Our contributions are threefold:

- **Evaluation reveals a hidden prior.** We decompose target-ranking performance into a gene-level target term and a context term, and evaluate context by within-target discrimination that is invariant to gene-specific offsets.
- **Post-training does not automatically yield transferable context.** Across model scale, interface width, labels, sampling and a within-target contrast objective, adapted LLMs learn contextual distinctions in the training cancers but do not transfer them to held-out ones; representation probes localize where the signal is lost.
- **Experimental memory changes the information bottleneck.** Measurements from related contexts supplied at inference restore held-out functional discrimination, with gains controlled by record similarity and number; source-matched readers show that the retrieved evidence contributes more than reader complexity.

2 Related Work

Cell foundation models and their evaluation. Transformer models trained on expression atlases supply reusable representations of cellular state (Cui et al., 2024; Theodoris et al., 2023; Hao et al.,

2024), and are adapted with parameter-efficient updates (He et al., 2026). Evaluations have repeatedly found that simple baselines match them (Ahlmann-Eltze et al., 2025; Csendes et al., 2025; Viñas Torné et al., 2026; Kedzierska et al., 2025).

Language models over biological evidence. Language models read cells as ranked gene names (Levine et al., 2024; Rizvi et al., 2025), genes as text (Chen & Zou, 2025), and omics as projected tokens (Sypetkowski et al., 2026); retrieval of measured experiments answers perturbation questions (Wu et al., 2025), and knowledge graphs prioritise cancer targets (Zhang et al., 2026).

Memory and retrieval in language models. Language models absorb frequent facts into their weights but struggle with the long tail (Kandpal et al., 2023), and supplying knowledge at inference often outperforms fine-tuning it in (Lewis et al., 2020; Ovadia et al., 2024). We find the same division in cancer biology: which genes matter is frequent and learned, where they matter is the long tail.

Answer priors in benchmarks. Input-blind baselines expose how much of a benchmark score is a property of the answer distribution (Gururangan et al., 2018; Poliak et al., 2018), metrics can manufacture apparent ability (Schaeffer et al., 2023), and models learn what training rewards (Geirhos et al., 2020). In functional genomics that prior is a biological fact, the genes essential in nearly every context (Hart et al., 2015; Behan et al., 2019; Vinceti et al., 2021).

Where information lives in composed models. Prior work studies pretrained cellular representations, parameter-efficient adaptation and retrieval-augmented prediction largely as separate design choices. CellVELA treats their roles as an information-placement problem: it separates four components, a gene-level prior, a cellular representation, adapted parameters and experimental memory, and measures which of them carries context that transfers to held-out biological contexts.

3 WHERE CONTEXT-SPECIFIC INFORMATION CAN ENTER

3.1 SEPARATING THE GENE PRIOR FROM CONTEXT

Let $\mathcal{R}$ be cancer lines with expression profiles $x_r \in \mathbb{R}^G$ and CRISPR gene effects D_{rg}, split into training lines $\mathcal{R}_{\mathrm{tr}}$ and held-out lines $\mathcal{R}_{\mathrm{te}}$. The reference set Y_r holds the K genes with the most negative effect in line r ($K = 100$), and $y_{rg} = \mathbf{1}[g \in Y_r]$; a method predicts a set P_r of K genes, $P_{rg} = \mathbf{1}[g \in P_r]$. *Target prioritization* ranks genes within a line; *functional discrimination* ranks lines for a gene. The average overlap splits into a term shared by all lines and a context term,

$$\frac{1}{n}\sum_r |P_r \cap Y_r| = \underbrace{\sum_g \bar{p}_g\,\bar{y}_g}_{\text{target term}} + \underbrace{\sum_g \mathrm{Cov}_r\big(P_{rg}, y_{rg}\big)}_{\text{context term}}, \tag{1}$$

where $\bar{p}_g, \bar{y}_g$ are means and Cov_r the covariance over the n held-out lines, so a single list shared by all lines can earn the whole target term but none of the context term (Appendix A). We score functional discrimination by the *context AUC*, a within-target AUC averaged over the evaluable genes $\mathcal{G}_e$, those with at least four dependent held-out lines $\mathcal{R}_g^+ = \{r : y_{rg} = 1\}$ and four non-dependent ones $\mathcal{R}_g^-$,

$$\mathrm{A}(s) = \frac{1}{|\mathcal{G}_e|}\sum_{g\in\mathcal{G}_e} \Pr_{r\in\mathcal{R}_g^+,\, r'\in\mathcal{R}_g^-}\big[s(r,g) > s(r',g)\big], \tag{2}$$

with ties counted as one half; the context ρ is the analogous within-gene Spearman correlation with $-D_{rg}$.

Proposition 1. *Write $s(r,g) = b_g + h(r,g)$. If $h \equiv 0$, then $\mathrm{A}(s) = 1/2$ exactly; in general $\mathrm{A}(s)$ depends on h only and is invariant to any strictly increasing transform of $s(\cdot, g)$ for each g.*

3.2 ONE SCORING FUNCTION, THREE PLACES FOR CONTEXT

Every system in this paper is an instance of one scoring function of a context representation t_r, a target g and an experimental memory $\mathcal{M}_k(r,g)$,

$$s(r,g) = f_\theta\big(t_r,\ g,\ \mathcal{M}_k(r,g)\big), \tag{3}$$

so context-specific information can enter through the representation, the weights θ or the memory. *Representation.* A frozen cell foundation model E and two trained projectors map a line to tokens that enter a LoRA-adapted (Hu et al., 2022) language model as soft prompts (Lester et al., 2021),

$$t_r = \big[\pi_x(x_r);\ \pi_z(E(x_r))\big] \in \mathbb{R}^{2\times d}. \tag{4}$$

Memory. With $\tilde{x}_r$ the expression profile standardised over training lines, $\sigma(r, r') = \cos(\tilde{x}_r, \tilde{x}_{r'})$ and $\mathcal{R}^g_{\mathrm{tr}}$ the training lines in which g was screened, the memory holds the measured labels of g in the k most similar of them,

$$\mathcal{M}_k(r, g) = \big\{\big(\sigma(r, r'),\ y_{r'g}\big) : r' \in \underset{r'\in\mathcal{R}^g_{\mathrm{tr}}\setminus\{r\}}{\text{top-}k}\ \sigma(r, r')\big\}, \tag{5}$$

written into the prompt as text. The model scores the gene by the log-odds $s_\theta(r, g) = \log p_\theta(\textsc{Yes} \mid \cdot) - \log p_\theta(\textsc{No} \mid \cdot)$ (Eq. 10) and never sees a line name. For perturbation response the same construction applies with perturbations as contexts: the memory for a perturbation u holds the measured responses of the k training perturbations whose gene representations are most similar to that of u (Appendix F). *Weights.* Matched screens pair, for a target g, a training line that depends on it with one that does not, $\mathcal{P} = \{(r, r', g) : y_{rg} = 1,\ y_{r'g} = 0,\ D_{rg} < D_{r'g} - \delta\}$, and a contrast term writes these functional differences into θ; over a sample $\mathcal{B} \subset \mathcal{P}$,

$$\mathcal{L}_{\mathrm{ctx}} = \frac{1}{|\mathcal{B}|} \sum_{(r,r',g)\in\mathcal{B}} \log\Big(1 + \exp\big\{-\big(s_\theta(r, g) - s_\theta(r', g)\big)\big\}\Big), \tag{6}$$

added to answer supervision as $\mathcal{L} = \mathcal{L}_{\mathrm{ans}} + \lambda\mathcal{L}_{\mathrm{ctx}}$. Any gene-level offset b_g cancels in $s_\theta(r, g) - s_\theta(r', g)$, so, as in Proposition 1, no gene prior can reduce $\mathcal{L}_{\mathrm{ctx}}$. The two routes differ only in where the functional difference goes: $\mathcal{L}_{\mathrm{ctx}}$ writes it into θ, $\mathcal{M}_k$ supplies it at inference. CellVELA reads memory with $\lambda = 0$; $\lambda = 0.1$ is the contrast variant (Algorithm 1).

Controls. Each comparison changes one component of Eq. (3). Fine-tuning without records sets $\mathcal{M}_k = \varnothing$; a similarity vote reads the same $\mathcal{M}_k$ without a learned reader, as a nearest-neighbour kernel estimate over contexts; random records break the coupling between memory and context; fitted linear and tree readers read the same records for perturbation response. As a reference with more supervision, a per-gene ridge is fitted to continuous effects from all training lines. A width-matched Gaussian random projection of the expression profile tests what pretraining adds to E, and linear probes along the model locate where the context signal is lost. Appendix B defines each.

4 Experiments

Setup. *Cancer dependency*: DepMap 24Q4 CRISPR screens (DepMap, 2024; Dempster et al., 2021) with the Target-QA split (Zhang et al., 2026) of 297 training and 63 held-out lines and $K = 100$; the selective endpoint removes common essential genes, leaving the dependencies that could be drug targets. *Perturbation response*: PerturbQA (Wu et al., 2025), four cell lines, differential expression and its direction, with test perturbations removed from every training set and record bank. *Baselines*: context-free lists; per-gene ridge regressions on expression and on frozen Geneformer-V2 and scGPT representations; Qwen3 (Yang et al., 2025) from 0.6B to 8B and SmolLM3-3B (Bakouch et al., 2025) with LoRA. CellVELA uses Qwen3-4B, $k = 6$ records and three seeds; each fine-tuned model is trained on the labels of the endpoint it is scored on (Appendix B).

We first study contextual adaptation on cancer dependency, where the same target is measured across many cellular contexts, and then test whether the conclusions transfer to perturbation-response prediction (Section 4.4).

4.1 Target ranking can succeed without context

A fixed list of the genes most often essential in training ties the best readout of the expression profile on Rec.@100 at the standard endpoint and beats every fine-tuned model (Table 1). By Eq. (1), almost all of this score is the target term. The same list has context AUC exactly one half and ranks no gene's dependent lines above its non-dependent ones. A lineage prior already carries coarse context and sets a strong reference; CellVELA receives no line identity or lineage label and derives context from molecular similarity and measured records.

Method	Uses LLM	Uses Memory	**Standard endpoint** Rec.@100 ↑	**Standard endpoint** Context AUC ↑	**Selective endpoint** (common essentials removed) Rec.@100 ↑	**Selective endpoint** Context AUC ↑	**Selective endpoint** Context ρ ↑	$> \frac{1}{2}$ (%)
Context-free priors								
Gene prior	–	–	**67.5** (.7)	0.500 (.000)	32.5 (.7)	0.500 (.000)	0.000 (.000)	0
Lineage prior	–	–	66.3 (.7)	<u>0.554</u> (.008)	32.6 (.9)	0.576 (.005)	0.105 (.004)	53
Parameter adaptation: context written into LoRA weights								
Fine-tuned LLM	✓	–	57.1 (.6)	0.501 (.006)	25.2 (.6)	0.509 (.004)	0.010 (.002)	52
+ contrast loss	✓	–	51.4 (.5)	0.507 (.006)	24.9 (.6)	0.500 (.004)	0.012 (.002)	48
Experimental memory: records retrieved at inference								
Random records (k=50)	–	✓	67.1 (.7)	0.498 (.005)	31.4 (.7)	0.499 (.003)	0.003 (.002)	41
Similarity vote (k=6)	–	✓	64.3 (.6)	0.533 (.006)	32.5 (.8)	0.570 (.004)	0.123 (.005)	44
Similarity vote (k=20)	–	✓	66.7 (.7)	0.551 (.008)	<u>35.3</u> (.8)	<u>0.595</u> (.005)	<u>0.143</u> (.005)	54
Similarity vote (k=50)	–	✓	<u>67.2</u> (.7)	**0.559** (.008)	**35.9** (.7)	**0.604** (.005)	**0.152** (.005)	60
CellVELA (k=6)	✓	✓	64.2 (.6)	0.531 (.008)	32.4 (.8)	0.572 (.005)	0.088 (.004)	**64**
+ contrast loss	✓	✓	54.0 (.8)	0.506 (.007)	32.4 (.8)	0.548 (.005)	0.067 (.003)	<u>61</u>
Reference: per-gene ridge on continuous effects								
Ridge, expression	–	–	67.5 (.8)	0.598 (.009)	36.2 (.8)	0.669 (.006)	0.232 (.004)	80
Ridge, Geneformer-V2	–	–	67.0 (.8)	0.586 (.008)	35.5 (.8)	0.647 (.006)	0.211 (.004)	77
Ridge, scGPT	–	–	67.1 (.8)	0.575 (.009)	33.6 (.7)	0.627 (.006)	0.178 (.004)	73

Table 1: **Held-out cancer lines.** DepMap 24Q4, 63 held-out lines. Rec.@100: reference targets in the top 100 of each line (%). Context AUC and ρ: per-gene discrimination and rank correlation between score and measured dependency across held-out lines (Eq. 2), averaged over genes; a context-free score gives exactly 0.5 and 0. $> \frac{1}{2}$: share of genes with context AUC above one half. LLM: a fine-tuned language model scores the query; Memory: measured records of the target in the k most similar (or random) training lines are read at inference. Mean (s.e.), s.e. across lines for Rec.@100 and across genes otherwise. LLM rows: Qwen3-4B with LoRA, seed means (two or three seeds); CellVELA with k=20 is in Figure 3b. Reference rows use continuous effects from every training screen and are not ranked. **Bold**, best; <u>underlined</u>, second.

4.2 Parameter adaptation fits context but does not transfer it

Writing contextual supervision into the model weights is sufficient to fit the training cancers, but not to transfer the same distinction to new ones. Without records, LoRA-adapted LLMs remain near chance on held-out lines on every functional discrimination metric, including when trained with a within-target contrast that removes the gene prior by construction (Table 1). They do learn the comparison: on training lines the contrast term of Eq. (6) raises context AUC well above chance, but the gain does not transfer (Figure 2a). Twelve interventions on the label set, a gene-level offset, the width of the cell interface and the backbone, from 0.6B to 8B parameters, all leave held-out context AUC within ± 0.03 of the unchanged recipe (Figure 2b; Appendix D). Denser supervision does not resolve it either: a sampler that concentrates training on pairs of lines that differ for the same target raises context AUC on training lines but leaves held-out lines at chance. This is not a matter of exposure alone: per epoch the median target appears as a non-dependency in two distinct training lines under the balanced sampler and in eight under the pair-weighted one, and the fourfold increase still leaves held-out lines at chance (Appendix D). Two analyses locate the loss. A frozen cell representation carries less context than a random projection of the same expression at every training budget (Figure 2d), and linear probes find the signal in the cell tokens but much weaker at the answer position (Figure 2c): the scores move with the line without following its measured dependency.

4.3 Experimental memory restores held-out context

The same functional measurements become useful when they remain available at inference. Reading six records, CellVELA moves from chance to the level of a similarity vote over the same records on held-out lines, with a paired gain in each of three seeds (Table 1). With twenty records it rises further and stays level with the vote over the same twenty (Figure 3b), and its per-gene ranking agrees more closely with the vote (rank correlation 0.82, against 0.66 with six): the more evidence it is given, the more closely the adapted model follows the explicit estimator of Eq. (9). Adding the contrast term to CellVELA lowers it: once the functional difference is supplied, asking the model

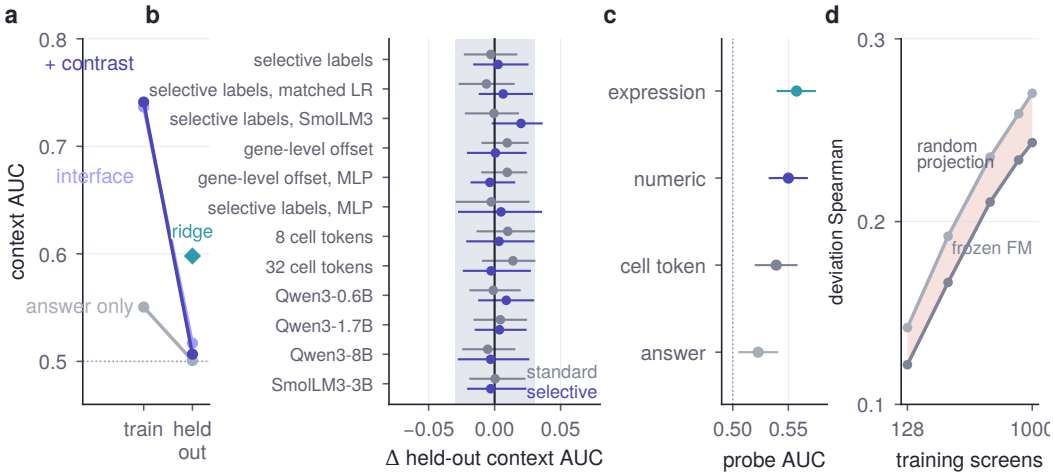

Figure 2: **Parameter adaptation fits contextual supervision but does not transfer it.** **(a)** Context AUC on training and held-out lines. **(b)** Twelve interventions on fine-tuning without records: change in held-out context AUC against the unchanged recipe, pooled over seeds, with 95% intervals over genes and lines; shaded, ± 0.03. **(c)** Linear probes along the model, scored by held-out context AUC. **(d)** A frozen cell foundation model against a width-matched random projection of the same expression, as training screens grow.

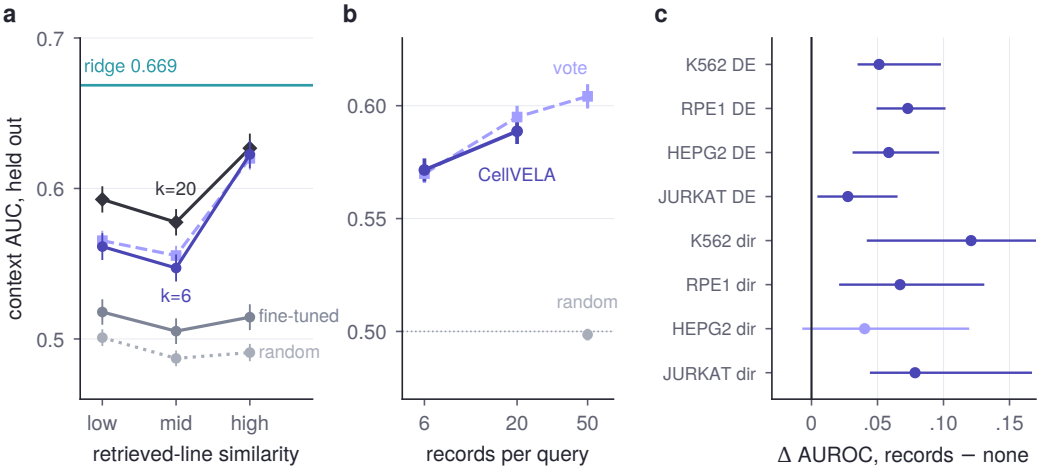

Figure 3: **Experimental memory restores the context that weight adaptation fails to transfer.** **(a)** Held-out context AUC by the similarity of the retrieved lines: CellVELA with six and twenty records, the similarity vote over the same six records (dashed), the fine-tuned model without records, and random records. **(b)** Context AUC against the number of records, for CellVELA and the vote; random records at $k = 50$. **(c)** Perturbation response: CellVELA with six records minus the same model without records, per setting, three seeds with joint seed and perturbation intervals.

to also learn it does not help. The records carry context rather than prior. Across held-out lines grouped by the similarity of their retrieved lines, CellVELA tracks the vote and is highest where the retrieved lines are most similar, while the fine-tuned model and random records stay flat (Figure 3a); per gene, records beat random records for most targets (Wilcoxon $p < 10^{-39}$). The per-gene ridge, fitted to continuous effects, remains higher and already contains what the records add. Proposition 2

Retrieval key	Reader	K562	RPE1	HepG2	Jurkat	All eight ↑
None	Zero-shot LLM	0.519	0.493	0.522	0.502	0.509
	Fine-tuned LLM	0.534	0.555	0.565	0.559	0.553
scGPT-human tokens	Linear	0.563	0.571	0.533	0.562	0.557
	Trees	0.558	0.566	0.547	0.575	0.561
	CellVELA	0.593	0.600	0.599	0.603	0.599
GenePT text	Linear	<u>0.622</u>	0.597	0.599	0.607	0.606
	Trees	0.606	0.617	0.601	0.606	0.607
	CellVELA	0.610	0.625	0.605	<u>0.614</u>	0.614
scGPT + GenePT	Linear	0.614	0.615	0.613	0.603	0.611
	Trees	**0.626**	<u>0.627</u>	**0.620**	**0.616**	**0.622**
	Zero-shot LLM	0.567	0.570	0.569	0.568	0.568
	CellVELA	0.620	0.625	0.615	0.612	<u>0.618</u>
Qwen3-Emb. text	Linear	0.606	0.625	0.593	**0.616**	0.610
	Trees	0.611	**0.632**	<u>0.616</u>	0.606	0.616
	CellVELA	0.613	0.625	0.610	0.613	0.615
Same perturbation, other line	Copy rule	0.597	0.633	0.629	0.645	0.626

Table 2: **Memory or reader? Perturbation response on PerturbQA.** Response-gene macro AUROC per cell line (mean of differential expression and its direction) and over all eight settings. Within a group every reader receives the same six records, retrieved by the gene representation on the left. Zero-shot LLM: Qwen3-8B without fine-tuning; fine-tuned LLM: Qwen3-4B with LoRA; CellVELA: the fine-tuned LLM reading records. Copy rule: the label of the same perturbation measured in another cell line. **Bold**, best; <u>underlined</u>, second, excluding the copy rule.

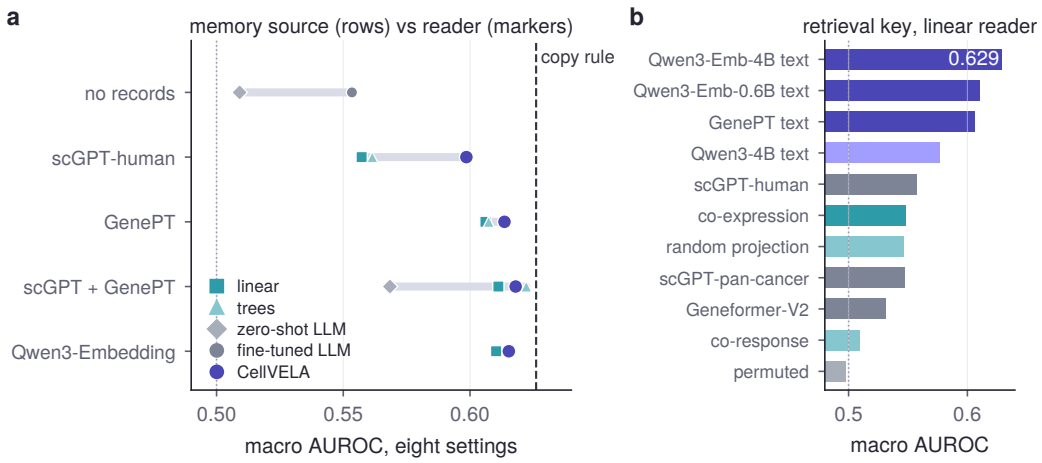

Figure 4: **Memory choice matters more than reader complexity.** PerturbQA macro AUROC over eight settings. **(a)** Table 2 as a dot plot: each row fixes the retrieval key, markers are readers of the same records; dashed, the copy rule. **(b)** Every gene representation as retrieval key under one linear reader.

(Appendix A) explains the random-record control: without similarity, a record vote is a gene prior in expectation.

4.4 Memory choice matters more than reader complexity

On PerturbQA, six records raise the adapted reader in all eight settings, seven with intervals excluding zero (Figure 3c); on the development split, removing or reversing them at inference erases or inverts the gain. Across memory sources and readers (Table 2, Figure 4a), performance varies more with the retrieved evidence than with the reader. The largest source-matched gap between readers occurs with records retrieved by scGPT gene tokens, but reader-specific advantages are not uniform

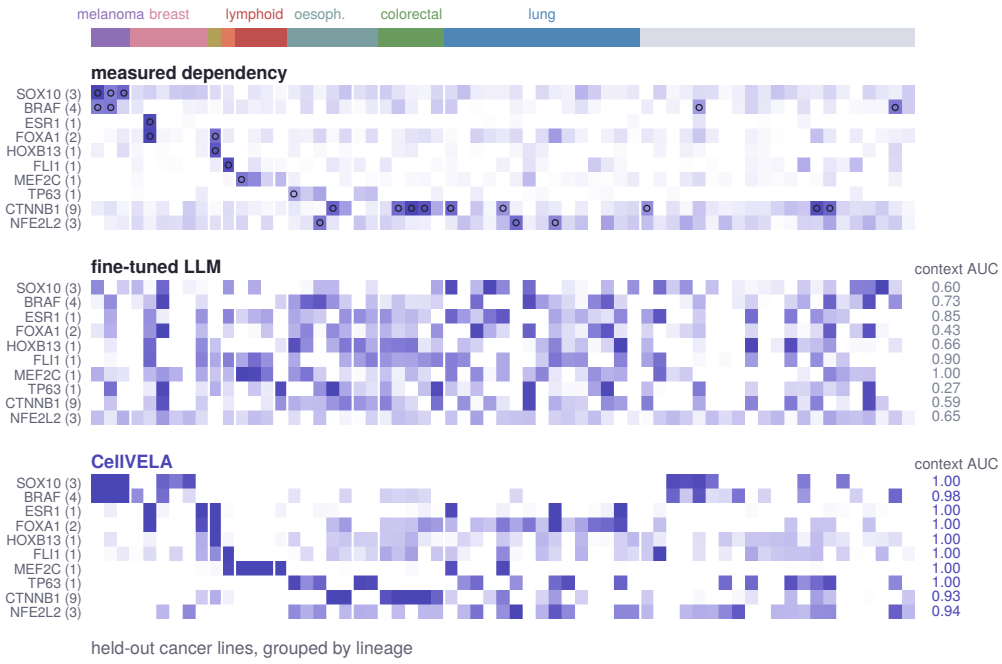

Figure 5: **Retrieved experimental memory recovers lineage-specific dependency structure.** Ten canonical selective dependencies (rows; number of dependent held-out lines in parentheses) across the 63 held-out lines (columns), grouped by lineage. Top: measured CRISPR dependency, dependent lines circled. Middle and bottom: scores of the fine-tuned LLM without records and of CellVELA with twenty records, standardised within each gene; right, per-gene context AUC.

across sources. Which representation retrieves the memory is the main lever: text embeddings of gene descriptions retrieve better records than the gene tokens of cell foundation models, which carry no more than co-expression or a random projection of it (Figure 4b).

4.5 Retrieved context recovers lineage structure

Per gene, CellVELA with twenty records improves on the fine-tuned model for 62% of the $1{,}444$ selective targets, raising mean context AUC from 0.510 to 0.588 (Appendix G). The structure of these gains is biological (Figure 5). CellVELA places the dependent lines of these canonical examples near the top, reproducing the lineage blocks of the measurement: SOX10 and BRAF in melanoma, ESR1 and FOXA1 in breast cancer, FLI1 in Ewing sarcoma, MEF2C in lymphoid cancers, TP63 in squamous and CTNNB1 in colorectal cancers. Several of these genes have a single dependent held-out line, shown in parentheses. The fine-tuned model's scores vary across lines without following lineage, and they rank some of these dependencies below chance: TP63 and FOXA1 score 0.27 and 0.43 without records and 1.00 with them, and across the ten genes CellVELA reaches 0.93 to 1.00. Records thus do more than raise an average; they carry the lineage structure of the screen to lines the model was never trained on.

5 Discussion

Our results recast functional adaptation as an information-placement problem. A composed system can carry new knowledge in its representation, its adapted parameters or explicit external memory, and these locations are not interchangeable. Three design principles follow.

Score context apart from the prior. Eq. (1) splits any set score into a target term and a context term, and only the second asks where a gene matters. Context AUC isolates it and is invariant to gene-level offsets (Proposition 1), so a gene prior, a random-record control and the model can be compared on one axis. On the standard endpoint the fixed gene list ties the best Rec.@100 of any method; without this split, that list would look like a strong context model.

Preserve long-tail context as retrievable memory. Functional distinctions that recur broadly across contexts can be reflected in model parameters, but the sparse relations that distinguish where a target matters are not reliably transferred by the adaptation procedures tested here; what survives is mostly the gene prior, much as rarely seen facts are poorly stored by language models (Kandpal et al., 2023). A record vote is a kernel regression over contexts (Eq. 9) and separates lines only through similarity; the adapted reader, given the same records, matches it (Table 1). Keeping the measurements externally accessible avoids forcing them through a fixed parametric bottleneck and lets the memory update without retraining when new lines are screened.

Choose the memory before the reader. On perturbation response, changing which gene representation retrieves the records moves every reader more than changing the reader does (Table 2), and text embeddings of gene function retrieve better than cell-model gene tokens. The language model is useful as a single interface that reads cell-model tokens, records and gene text together and moves between assays without redesign; it is not where the context signal comes from. A per-gene ridge on continuous effects remains stronger on dependency, and closing that gap with richer records is a natural next step.

Limitations. Claims rest on cell lines and CRISPR fitness rather than therapeutic benefit; records require that related contexts have been screened; frozen cell models read bulk profiles outside their single-cell pretraining; closed frontier models were not evaluated because Eq. (10) needs token log-probabilities.

6 Conclusion

We studied context-specific functional adaptation as a question of where information enters a composed foundation-model system. Conventional target ranking strongly rewards a context-free gene prior; once that prior is removed from evaluation, LoRA adaptation learns contextual distinctions in the training cancers but does not transfer them to held-out ones across the model, interface, supervision and sampling variants we tested. Keeping the same functional measurements available as experimental memory changes the outcome: retrieved records restore held-out discrimination, scale with record similarity and number, recover lineage-specific dependencies and improve perturbation-response prediction, and source-matched controls show that memory selection matters more than reader complexity. For context-dependent scientific prediction, adaptation is therefore not only about what to encode in parameters, but also about what evidence should remain explicitly retrievable at inference.

AI USE STATEMENT

We used large language models to aid or polish writing, for retrieval and discovery of related work, and for research ideation and execution. The authors take full responsibility for all content.

REFERENCES

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

Elie Bakouch, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Lewis Tunstall, Carlos Miguel Patiño, Edward Beeching, Aymeric Roucher, Aksel Joonas Reedi, Quentin Gallouédec, Kashif Rasul, Nathan Habib, Clémentine Fourrier, Hynek Kydlicek, Guilherme Penedo, Hugo Larcher, Mathieu Morlon, Vaibhav Srivastav, Joshua Lochner, Xuan-Son Nguyen, Colin Raffel, Leandro von Werra, and Thomas Wolf. SmolLM3: smol, multilingual, long-context reasoner. Hugging Face blog, July 2025. URL https://huggingface.co/blog/smollm3.

Fiona M. Behan, Francesco Iorio, Gabriele Picco, Emanuel Gonçalves, Charlotte M. Beaver, Giorgia Migliardi, Rita Santos, Yanhua Rao, Francesco Sassi, Marika Pinnelli, Rizwan Ansari, Sarah Harper, David Adam Jackson, Rebecca McRae, Rachel Pooley, Piers Wilkinson, Dieudonne van der Meer, David Dow, Carolyn Buser-Doepner, Andrea Bertotti, Livio Trusolino, Euan A. Stronach, Julio Saez-Rodriguez, Kosuke Yusa, and Mathew J. Garnett. Prioritization of cancer therapeutic targets using CRISPR–Cas9 screens. *Nature*, 568(7753):511–516, 2019. doi: 10.1038/s41586-019-1103-9.

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.

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

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

Joshua M. Dempster, Isabella Boyle, Francisca Vazquez, David E. Root, Jesse S. Boehm, William C. Hahn, Aviad Tsherniak, and James M. McFarland. Chronos: a cell population dynamics model of CRISPR experiments that improves inference of gene fitness effects. *Genome Biology*, 22(1): 343, 2021. doi: 10.1186/s13059-021-02540-7.

Broad DepMap. DepMap 24Q4 public. Figshare+ dataset, 2024. URL https://plus.figshare.com/articles/dataset/DepMap_24Q4_Public/27993248/1.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 107–112. Association for Computational Linguistics, 2018. doi: 10.18653/v1/N18-2017. URL https://aclanthology.org/N18-2017/.

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

Traver Hart, Megha Chandrashekhar, Michael Aregger, Zachary Steinhart, Kevin R. Brown, Graham MacLeod, Monika Mis, Michal Zimmermann, Amelie Fradet-Turcotte, Song Sun, Patricia Mero, Peter Dirks, Sachdev Sidhu, Frederick P. Roth, Olivia S. Rissland, Daniel Durocher, Stephane Angers, and Jason Moffat. High-resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities. *Cell*, 163(6):1515–1526, 2015. doi: 10.1016/j.cell.2015.11.015.

Fei He, Ruixin Fei, Jordan E. Krull, Yang Yu, Xinyu Zhang, Xianyu Wang, Hao Cheng, Mingyue Gao, Li Su, Yibo Chen, Jinpu Li, Baichuan Jin, Yuzhou Chang, Anjun Ma, Qin Ma, and Dong Xu. Harnessing the power of single-cell large language models with parameter-efficient fine-tuning using scPEFT. *Nature Machine Intelligence*, 8(1):118–133, 2026. doi: 10.1038/s42256-025-01170-z.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 15696–15707. PMLR, 2023. URL https://proceedings.mlr.press/v202/kandpal23a.html.

Kasia Z. Kedzierska, Lorin Crawford, Ava P. Amini, and Alex X. Lu. Zero-shot evaluation reveals limitations of single-cell foundation models. *Genome Biology*, 26(1):101, 2025. doi: 10.1186/s13059-025-03574-x.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3045–3059. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.243. URL https://aclanthology.org/2021.emnlp-main.243/.

Daniel Levine, Syed A. Rizvi, Sacha Lévy, Nazreen Pallikkavaliyaveetil, David Zhang, Xingyu Chen, Sina Ghadermarzi, Ruiming Wu, Zihe Zheng, Ivan Vrkic, Anna Zhong, Daphne Raskin, Insu Han, Antonio Henrique De Oliveira Fonseca, Josue Ortega Caro, Amin Karbasi, Rahul Madhav Dhodapkar, and David van Dijk. Cell2Sentence: Teaching large language models the language of biology. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 27299–27325. PMLR, 2024. URL https://proceedings.mlr.press/v235/levine24a.html.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. Fine-tuning or retrieval? comparing knowledge injection in LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 237–250. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.15. URL https://aclanthology.org/2024.emnlp-main.15/.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pp. 180–191. Association for Computational Linguistics, 2018. doi: 10.18653/v1/S18-2023. URL https://aclanthology.org/S18-2023/.

Syed Asad Rizvi, Daniel Levine, Aakash Patel, Shiyang Zhang, Eric Wang, Curtis Jamison Perry, Ivan Vrkic, Nicole Mayerli Constante, Zirui Fu, Sizhuang He, David Zhang, Cerise Tang,

Zhuoyang Lyu, Rayyan Darji, Chang Li, Emily Sun, David Jeong, Lawrence Zhao, Jennifer Kwan, David Braun, Brian Hafler, Hattie Chung, Rahul M. Dhodapkar, Paul Jaeger, Bryan Perozzi, Jeffrey Ishizuka, Shekoofeh Azizi, and David van Dijk. Scaling large language models for next-generation single-cell analysis. bioRxiv preprint 2025.04.14.648850, 2025.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems*, volume 36, pp. 55565–55581. Curran Associates, Inc., 2023. doi: 10.52202/075280-2425. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/adc98a266f45005c403b8311ca7e8bd7-Abstract-Conference.html.

Maciej Sypetkowski, Joanna Krawczyk, Łukasz Smoliński, Remigiusz Kinas, Przemysław Pietrzak, Tomasz Jetka, and Rafał Powalski. OmicsLM: A multimodal large language model for multi-sample omics reasoning. arXiv preprint arXiv:2605.06728, 2026. URL https://arxiv.org/abs/2605.06728.

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.

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.

Alessandro Vinceti, Emre Karakoc, Clare Pacini, Umberto Perron, Riccardo Roberto De Lucia, Mathew J. Garnett, and Francesco Iorio. CoRe: a robustly benchmarked R package for identifying core-fitness genes in genome-wide pooled CRISPR-Cas9 screens. *BMC Genomics*, 22(1):828, 2021. doi: 10.1186/s12864-021-08129-5.

Menghua Wu, Russell Littman, Jacob Levine, Lin Qiu, Tommaso Biancalani, David Richmond, and Jan-Christian Huetter. Contextualizing biological perturbation experiments through language. In *International Conference on Learning Representations*, pp. 73598–73624, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/b6e67ae290635d0874c4cb43ba2a2cfb-Paper-Conference.pdf.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. arXiv preprint arXiv:2505.09388, 2025. URL https://arxiv.org/abs/2505.09388.

Heming Zhang, Di Huang, Wenyu Li, Michael A. Province, Yixin Chen, Philip Payne, and Fuhai Li. GALAX: Graph-augmented language model for explainable reinforcement-guided subgraph reasoning in precision medicine. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ADFXCeYXvR.

APPENDIX

A Theory B Model, training and controls C Metrics and data D Supervision E Memory controls F Perturbation response G Case selection

A THEORY

Notation. Held-out lines $r = 1, \dots, n$ and genes g; $y_{rg} \in \{0, 1\}$ marks a reference target ($Y_r = \{g : y_{rg} = 1\}$), $P_{rg} \in \{0, 1\}$ a predicted one ($P_r = \{g : P_{rg} = 1\}$), and bars denote means over lines, $\bar{p}_g = \frac{1}{n}\sum_r P_{rg}$ and $\bar{y}_g = \frac{1}{n}\sum_r y_{rg}$. For gene g, $\mathcal{R}_g^+ = \{r : y_{rg} = 1\}$ and $\mathcal{R}_g^- = \{r : y_{rg} = 0\}$.

Decomposition (Eq. 1). Since $|P_r \cap Y_r| = \sum_g P_{rg} y_{rg}$, exchanging the sums and writing each product about its mean gives, for every g,

$$\frac{1}{n}\sum_r P_{rg} y_{rg} = \bar{p}_g \bar{y}_g + \frac{1}{n}\sum_r (P_{rg} - \bar{p}_g)(y_{rg} - \bar{y}_g) = \bar{p}_g \bar{y}_g + \mathrm{Cov}_r(P_{rg}, y_{rg}), \tag{7}$$

with the population covariance over lines. Summing over g gives Eq. (1). If the predicted set is the same for every line, $P_{rg} = p_g$, each covariance is zero and the score equals the target term. □

Proposition 1 and invariance. Let $\mathrm{A}_g(s) = \frac{1}{|\mathcal{R}_g^+||\mathcal{R}_g^-|}\sum_{r\in\mathcal{R}_g^+}\sum_{r'\in\mathcal{R}_g^-}\big(\mathbf{1}[s(r,g) > s(r',g)] + \frac{1}{2}\mathbf{1}[s(r,g) = s(r',g)]\big)$ and $\mathrm{A}(s)$ its mean over evaluable genes. Every comparison involves two scores of the same gene, so for any strictly increasing ϕ_g and any offsets b_g,

$$\mathrm{A}\big(\{\phi_g \circ s(\cdot, g)\}_g\big) = \mathrm{A}(s), \qquad \mathrm{A}(b_g + h) = \mathrm{A}(h). \tag{8}$$

If $h \equiv 0$, every comparison is a tie and contributes one half, so $\mathrm{A} = 1/2$ exactly. The same argument gives $\rho_g = 0$ by convention for a constant score, and $\mathrm{Cov}_r(P_{rg}, y_{rg}) = 0$ for any top-K set computed from b_g alone. □

Memory as kernel regression over contexts. The similarity vote reads $\mathcal{M}_k(r, g)$ as

$$\hat{y}_{rg} = \sum_{r'\in\mathcal{N}_k(r,g)} \frac{w(r,r')}{\sum_{r''\in\mathcal{N}_k(r,g)} w(r,r'')}\, y_{r'g}, \qquad w(r,r') = \max\{\sigma(r,r'), \epsilon\}, \tag{9}$$

where $\mathcal{N}_k(r, g)$ are the lines in $\mathcal{M}_k(r, g)$ and $\epsilon = 10^{-6}$. This is a weighted k-nearest-neighbour estimate of $\Pr(y_{rg} = 1 \mid x_r)$ in the metric induced by σ; it separates two held-out lines only when their neighbourhoods differ in labels.

Proposition 2. *Let $\mathcal{N}_k(r, g)$ be a uniformly random k-subset of the n_g training lines that measure g, drawn independently of r, and let the vote use equal weights. Then $\mathbb{E}[\hat{y}_{rg} \mid r] = \bar{y}_g^{\mathrm{tr}}$, the training frequency of g, for every held-out r, and $\mathrm{Var}[\hat{y}_{rg} \mid r] = \frac{\bar{y}_g^{\mathrm{tr}}(1-\bar{y}_g^{\mathrm{tr}})}{k} \cdot \frac{n_g - k}{n_g - 1}$.*

Proof. $\hat{y}_{rg}$ is the mean of k labels drawn without replacement from a fixed population with frequency $\bar{y}_g^{\mathrm{tr}}$; its mean and variance are those of the hypergeometric proportion, and neither depends on r. □

By Proposition 1, the expected random-record score is a gene prior with context AUC one half; any realised deviation is sampling noise that is independent of the held-out labels. Empirically, the random-record vote stays within 0.015 of one half at every k on both endpoints (Appendix E). Context can therefore enter the vote only through $\sigma(r, \cdot)$, and a reader of $\mathcal{M}_k$ can recover it only to the extent that the records it is given are informative.

B MODEL, TRAINING AND CONTROLS

Scoring and objective. With the prompt $(t_r, \mathcal{M}_k(r,g), g)$,

$$s_\theta(r,g) = \log p_\theta(\textsc{Yes} \mid t_r, \mathcal{M}_k(r,g), g) - \log p_\theta(\textsc{No} \mid t_r, \mathcal{M}_k(r,g), g), \tag{10}$$

$$\mathcal{L}_{\text{ans}} = -\frac{1}{|\mathcal{Q}|} \sum_{(r,g)\in\mathcal{Q}} \Big[y_{rg} \log \varsigma\big(s_\theta(r,g)\big) + (1-y_{rg}) \log\big(1-\varsigma(s_\theta(r,g))\big)\Big], \tag{11}$$

$$\mathcal{L} = \mathcal{L}_{\text{ans}} + \lambda\, \mathcal{L}_{\text{ctx}}, \tag{12}$$

with ς the logistic function, $\mathcal{Q}$ the training queries of a batch (training lines only), and $\mathcal{L}_{\text{ctx}}$ from Eq. (6) on a sample $\mathcal{B}$ of contrast pairs. Since $\varsigma(s_\theta) = p_\theta(\textsc{Yes} \mid \cdot)/(p_\theta(\textsc{Yes} \mid \cdot) + p_\theta(\textsc{No} \mid \cdot))$, $\mathcal{L}_{\text{ans}}$ is the cross-entropy of the answer restricted to the two answer tokens, and $\mathcal{L}_{\text{ctx}}$ is a pairwise logistic (RankNet) loss on the same scores.

Algorithm 1 CellVELA training and inference

Require: training lines $\mathcal{R}_{\text{tr}}$ with x_r, D_{rg}, y_{rg}; frozen E; language model f_θ; k, λ, δ, M

procedure RETRIEVEMEMORY(r, g)
  $\mathcal{C} \leftarrow \mathcal{R}_{\text{tr}}^{g} \setminus \{r\}$ ▷ training lines that screened g
  return $\big[(\sigma(r,r'), y_{r'g})\big]$ for the k lines of $\mathcal{C}$ with largest $\sigma(r,r')$, in decreasing order
end procedure
procedure SAMPLECONTRASTPAIRS
  return $\mathcal{B}$: M triples drawn uniformly from $\mathcal{P} = \{(r,r',g) : y_{rg}=1,\ y_{r'g}=0,\ D_{rg} < D_{r'g} - \delta\}$
end procedure
Standardise x_r over $\mathcal{R}_{\text{tr}}$ and normalise; $\sigma(r,r') \leftarrow \tilde{x}_r^\top \tilde{x}_{r'}$
for each epoch and batch of training queries (r,g) **do**
  $t_r \leftarrow [\pi_x(x_r);\ \pi_z(E(x_r))]$; $\mathcal{M} \leftarrow$ RETRIEVEMEMORY(r,g) ▷ Eqs. 4, 5
  $s_\theta(r,g) \leftarrow$ log-odds of YES given $(t_r, \mathcal{M}, g)$ ▷ Eq. 10
  if $\lambda > 0$: $\mathcal{B} \leftarrow$ SAMPLECONTRASTPAIRS, add $\lambda\, \mathcal{L}_{\text{ctx}}(\mathcal{B})$ ▷ Eq. 6
  update π_x, π_z and the LoRA weights of f_θ on $\mathcal{L}$ ▷ Eq. 12
end for
Inference: for a held-out line r, $\mathcal{M} \leftarrow$ RETRIEVEMEMORY(r,g) and score $s_\theta(r,g)$

Prompt anatomy. The prompt has five parts, in this order:

two soft cell tokens t_r | question | target gene g | k measured records $\mathcal{M}_k(r,g)$ | `Answer:`

For example, with $k = 6$: "*Gene: ACSL3. In the six most similar training cell lines (expression similarity), ACSL3 was a top-100 dependency in 3 of 6: line1 yes (0.34), line2 no (0.31), line3 yes (0.29), line4 no (0.29), line5 yes (0.28), line6 no (0.27).*" With twenty records, similarities are printed for the five nearest lines.

Readers, references and controls. Fitted readers of the same memory use features $\phi(\mathcal{M}) = [\,y_i,\ \sigma_i,\ y_i\sigma_i,\ m_i\,]_{i=1}^{k}$, with m_i marking a present record, in a logistic model or gradient-boosted trees. The reference ridge and the representation control are

$$\hat{D}_{\cdot g} = \mu_g \mathbf{1} + Z\big(Z_{\text{tr}}^\top Z_{\text{tr}} + \alpha I\big)^{-1} Z_{\text{tr}}^\top \big(D_{\text{tr},g} - \mu_g \mathbf{1}\big), \tag{13}$$

$$z_r^{\text{RP}} = \Omega \tilde{x}_r, \qquad \Omega \in \mathbb{R}^{d_E \times G}, \quad \Omega_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, 1/d_E), \tag{14}$$

with Z the standardised expression or a frozen cell representation (columns centred on training lines), μ_g the mean training effect of g, α chosen on validation lines and $d_E = \dim E(x_r)$; the ridge scores $s(r,g) = -\hat{D}_{rg}$. The random projection replaces $E(x_r)$ by z_r^{RP} at equal width. A linear probe $\varsigma(v_\ell^\top h_\ell(r,g))$ is trained on training lines at each stage ℓ of the model and scored by held-out context AUC.

Table B.1: Training settings.

Language model	Qwen3-4B-Base; LoRA rank 8 on all projections
Cell representation E	scGPT-human, frozen; projectors π_x, π_z to two tokens
Learning rate	3×10^{-5} (LoRA), 10^{-4} (projectors)
Training	10 epochs, about 5,500 steps; three seeds
Memory	$k \in \{6, 20\}$; cosine on standardised expression; training lines only
Contrast	$\lambda = 0.1$; $M = 5{,}800$ pairs

C METRICS AND DATA

For a held-out line r with reference set Y_r of $K = 100$ targets and a gene g with effects D_{rg},

$$\mathrm{Rec@100}(r) = \tfrac{100}{K}\,\big|\,\mathrm{top}_{100}\; s(r,\cdot) \cap Y_r\big|\;(\%), \qquad \rho_g = \mathrm{Spearman}_r\big(s(r,g),\,-D_{rg}\big), \tag{15}$$

and A_g as in Appendix A. Context AUC and ρ are averaged over genes with at least four held-out lines in each of $\mathcal{R}_g^+$ and $\mathcal{R}_g^-$ (446 genes on the standard endpoint, 1,444 on the selective one). Standard errors are taken across lines for Rec.@100 and across genes otherwise; for seeded models the score is the seed mean. DepMap 24Q4 provides 297 training and 63 held-out lines with no shared patient; the selective endpoint removes DepMap common essential genes before Y_r is formed. Algorithm 2 summarises the protocol.

Algorithm 2 Evaluation on held-out contexts

Require: scores $s(r,g)$ and effects D_{rg} on held-out lines; reference size K; minimum class size 4

1: **for** each held-out line r **do** ▷ target term and context term together
2: $\quad\mathrm{Rec@100}(r) \leftarrow |\,\mathrm{top}_{100}\, s(r,\cdot) \cap Y_r|$
3: **end for**
4: **for** each gene g with $|\mathcal{R}_g^+| \geq 4$ and $|\mathcal{R}_g^-| \geq 4$ **do** ▷ context term only
5: $\quad\mathrm{A}_g \leftarrow$ pairwise AUC of $s(\cdot,g)$ over $\mathcal{R}_g^+ \times \mathcal{R}_g^-$, ties one half
6: $\quad\rho_g \leftarrow \mathrm{Spearman}(s(\cdot,g), -D_{\cdot g})$, or 0 if $s(\cdot,g)$ is constant
7: **end for**
8: **return** means and standard errors of Rec@100 over lines and of A_g, ρ_g over genes; share of genes with $\mathrm{A}_g > \frac{1}{2}$
9: **Controls:** repeat with the gene prior ($s = \bar{y}_g^{\mathrm{tr}}$), with random records, and with the same records under each reader

D PARAMETRIC ADAPTATION

Figure 2b lists the twelve interventions on fine-tuning without records; none moves held-out context AUC by more than 0.03 on either endpoint. Supervision per target is sparse by construction: in one epoch (seed 0) the balanced sampler shows the median target as a dependency in six distinct training lines and as a non-dependency in two, whereas a per-target ridge is fitted on every training line that screened the target (median 246 non-dependent lines). A pair-weighted sampler that draws dependent and non-dependent lines of the same target together raises the median to eight non-dependent lines per epoch and raises training context AUC, but leaves held-out lines at chance. Adding the contrast loss changes held-out context AUC by at most 0.010 under either sampler.

E EXPERIMENTAL-MEMORY CONTROLS

Without a language model, the similarity vote of Eq. (9) has context AUC above one half at every $k \in \{6, 20, 50\}$ and on both endpoints (Table 1), rising with k from 0.570 to 0.604 on the selective endpoint. A vote over the same number of randomly drawn training lines stays within 0.015 of one half at every k and on both endpoints, as Proposition 2 predicts, while its Rec.@100 stays close to the gene prior: random records transmit the target term but not the context term of Eq. (1).

F PERTURBATION RESPONSE

A PerturbQA query is a cell line c, a perturbed gene u and a response gene v, with label y_{cuv} for differential expression or its direction. Memory for (c,u,v) holds the labels $y_{cu'v}$ of the k training perturbations u' most similar to u under a gene representation γ, and performance is the macro AUROC over response genes,

$$\mathrm{AUROC}_{\mathrm{macro}} = \frac{1}{|\mathcal{V}_e|}\sum_{v\in\mathcal{V}_e} \mathrm{AUC}\big(\{s(c,u,v)\}_u,\;\{y_{cuv}\}_u\big), \tag{16}$$

over response genes with both labels; test perturbations are removed from every training set and memory bank. Table 2 gives the per-cell-line values and Figure 4b compares gene representations as retrieval keys under one linear reader.

G CASE SELECTION

The ten genes in Figure 5 are well-known lineage dependencies from the cancer literature, restricted to genes with at least one dependent held-out line. They were chosen after inspecting per-gene results of the similarity vote and before CellVELA's twenty-record predictions were scored; they illustrate rather than estimate performance, which Table 1 reports over all $1{,}444$ selective-endpoint genes. Table G.1 lists their lineage, the number of dependent held-out lines (Dep.) and the context AUC of each method.

Table G.1: Canonical lineage dependencies on held-out cancer lines. Context AUC on the selective endpoint; CellVELA and the vote read twenty records. With a single dependent line, context AUC is the rank of that line among all held-out lines.

Gene	Lineage	Dep.	Fine-tuned	CellVELA	Vote
SOX10	melanoma	3	0.600	1.000	1.000
BRAF	melanoma	4	0.729	0.983	0.970
ESR1	breast	1	0.855	1.000	1.000
FOXA1	breast	2	0.426	1.000	1.000
HOXB13	prostate	1	0.661	1.000	1.000
FLI1	Ewing sarcoma	1	0.903	1.000	1.000
MEF2C	lymphoid	1	1.000	1.000	1.000
TP63	oesophageal	1	0.274	1.000	1.000
CTNNB1	colorectal	9	0.593	0.934	0.926
NFE2L2	lung	3	0.650	0.939	0.928
All genes			0.510	0.588	0.595