RETRIEVAL-AUGMENTED GENERATION FOR PREDICTING CELLULAR RESPONSES TO GENE PERTURBATION

Andrea Giuseppe Di Francesco[†]
Sapienza University of Rome, Rome, Italy
ISTI-CNR, Institute of Information Science and Technologies, Pisa, Italy
difrancesco@diag.uniroma1.it

Andrea Rubbi[†]
University of Cambridge, Cambridge, United Kingdom
Wellcome Sanger Institute, Cambridge, United Kingdom
ar2232@cam.ac.uk

Pietro Liò
University of Cambridge, Cambridge, United Kingdom
pl219@cam.ac.uk

[†]Equal contribution.

ABSTRACT

Predicting how cells respond to genetic perturbations is fundamental to understanding gene function, disease mechanisms, and therapeutic development. While recent deep learning approaches have shown promise in modeling single-cell perturbation responses, they struggle to generalize across cell types and perturbation contexts due to limited contextual information during generation. We introduce **PT-RAG** (Perturbation-aware Two-stage Retrieval-Augmented Generation), a novel framework that extends Retrieval-Augmented Generation beyond traditional language model applications to the domain of cellular biology. Unlike standard RAG systems designed for text retrieval with pre-trained LLMs, perturbation retrieval lacks established similarity metrics and requires learning what constitutes relevant context, making differentiable retrieval essential. PT-RAG addresses this through a two-stage pipeline: first, retrieving candidate perturbations K using GenePT embeddings, then adaptively refining the selection through Gumbel-Softmax discrete sampling conditioned on both the cell state and the input perturbation. This cell-type-aware differentiable retrieval enables end-to-end optimization of the retrieval objective jointly with generation. On the Replogle-Nadig single-gene perturbation dataset, we demonstrate that PT-RAG outperforms both STATE and vanilla RAG under identical experimental conditions, with the strongest gains in distributional similarity metrics (W_1, W_2). Notably, vanilla RAG's dramatic failure is itself a key finding: it demonstrates that differentiable, cell-type-aware retrieval is essential in this domain, and that naive retrieval can actively harm performance. Our results establish retrieval-augmented generation as a promising paradigm for modelling cellular responses to gene perturbation. The code to reproduce our experiments is available at https://github.com/difra100/PT-RAG_ICLR.

1 INTRODUCTION

Understanding how cells respond to genetic perturbations is a fundamental challenge in systems biology with profound implications for drug discovery, disease modeling, and gene therapy (Norman et al., 2019). High-throughput Perturb-seq technologies now enable measurement of single-cell transcriptomic responses to thousands of genetic perturbations (Norman et al., 2019), generating

rich datasets that capture the complex landscape of cellular responses. However, the combinatorial explosion of possible perturbations and cell contexts makes comprehensive experimental characterization infeasible, motivating the development of computational methods to predict perturbation responses *in silico*.

Recent advances in deep learning have produced increasingly sophisticated models for perturbation response prediction. Early approaches like scGen (Lotfollahi et al., 2019) learned latent perturbation vectors, while CPA (Lotfollahi et al., 2023) introduced compositional modeling of perturbations and covariates. More recent work has leveraged optimal transport frameworks (Bunne et al., 2023), graph neural networks for gene-gene interactions (Roohani et al., 2024), and transformer architectures that model cell populations as sequences (Adduri et al., 2025). These approaches typically learn to map from a control cell state to a perturbed state conditioned on the perturbation identity (Klein et al., 2025).

Despite this progress, existing methods face a critical limitation: they generate predictions based solely on the control cell state and perturbation identity, without leveraging knowledge about related perturbations that may share similar biological effects. This is particularly problematic for predicting responses to perturbations in novel cell types, where the model has no direct supervision about how that cell type responds to related genetic interventions.

In natural language processing (NLP), Retrieval-Augmented Generation (RAG) (Lewis et al., 2021) has emerged as a powerful paradigm for incorporating external knowledge into generation tasks. RAG systems retrieve relevant documents from a knowledge base and condition the generator on both the input query and retrieved context, substantially improving performance on knowledge-intensive tasks. Despite its success, RAG was extended to few other modalities such as images, audio, and video (Abootorabi et al., 2025). In such settings, retrieval can often rely on off-the-shelf text/image encoders (Song et al., 2020; Radford et al., 2021), and even non-differentiable retrieval yields strong results because the notion of "relevant context" is well-defined.

RAG beyond LLMs. Extending RAG beyond classical domains, where both the notion of relevance and the generator architecture differ fundamentally, remains largely unexplored. In cellular biology, there are no pre-trained retrievers for perturbations, no established similarity metrics between genes, and the "generator" must produce high-dimensional cell distributions rather than text. This raises a critical question: *can RAG improve generation when the retrieval objective itself must be learned?* Recent work on differentiable RAG (Zamani & Bendersky, 2024; Gao et al., 2025) suggests that end-to-end optimization through the retrieval step, using techniques like Gumbel-Softmax (Jang et al., 2017; Maddison et al., 2017), can align retrieval with generation objectives. We argue that in domains like perturbation biology, differentiable retrieval becomes *essential*: without learning what context helps generation, naive retrieval is not known a priori to be beneficial.

We propose **PT-RAG** (Perturbation-aware Two-stage Retrieval-Augmented Generation), a novel framework that brings RAG to single-cell perturbation prediction. To our knowledge, the first application of retrieval-augmented generation to cellular response modeling. Our key insight is that perturbations with similar biological functions should induce similar cellular responses; thus, conditioning the generator on observed responses to related perturbations should improve the generation, especially for unseen cell types. However, naive application of RAG fails in this domain for two fundamental reasons: (1) **No established retrieval metrics**: Unlike NLP where semantic text similarity is well-studied, there is no consensus on how to measure perturbation similarity. One-hot encodings carry no semantic information, and even existing embedding-based similarities (e.g., via GenePT (Chen & Zou, 2023)) capture only functional descriptions, not cellular effects. (2) **Cell-type agnosticism**: Standard retrieval depends only on the query perturbation, providing identical context regardless of whether we predict responses in T-cells, neurons, or hepatocytes. However, in practice the same perturbation can have different effects across cell types and populations.

PT-RAG addresses this through a two-stage differentiable retrieval pipeline. To reduce the pool of perturbations from thousands to just a few, we first retrieve K candidate perturbations based on semantic similarity in GenePT embedding space (Chen & Zou, 2023), namely a foundation model that embeds genes based on their functional descriptions. Second, we employ a Straight-Through Gumbel-Softmax estimator (Jang et al., 2017; Maddison et al., 2017) for a differentiable discrete selection mechanism conditioned on both the cell state and perturbation embedding, adaptively selecting a subset of retrieved perturbations as context. This cell-type-aware retrieval enables the

model to learn which related perturbations are most informative for each specific cellular context, and the differentiable formulation allows end-to-end training.

Our contributions are as follows:

- We introduce PT-RAG, the first retrieval-augmented generation framework for modelling cellular response generation, demonstrating that RAG can be effectively extended beyond classic domains to improve biological sequence generation.
- We effectively deploy a two-stage pipeline combining semantic retrieval (via GenePT embeddings) with Gumbel-Softmax selection conditioned on cell state, enabling end-to-end learning of what context benefits generation, and successfully closing the gap between vanilla RAG, and no RAG baselines.
- We show that **naive retrieval actively hurts performance** in this domain: vanilla RAG with fixed retrieval dramatically underperforms all baselines, itself a key finding that underscores the necessity of differentiable, cell-type-aware retrieval.
- We provide quantitative evidence that PT-RAG learns cell-type-specific retrieval patterns, with only 19% overlap in selected perturbations across cell types for the same query gene.

2 Related Work

Existing perturbation prediction methods (Lotfollahi et al., 2019; 2023; Roohani et al., 2024; Bunne et al., 2023; Adduri et al., 2025; Klein et al., 2025) generate responses based solely on cell state and perturbation identity, without leveraging knowledge from related perturbations. Our work builds on STATE (Adduri et al., 2025), which models cell populations as sequences with distributional losses, but we fundamentally extend it with retrieval-augmented generation to incorporate contextual perturbation information.

Differentiable RAG. While standard RAG (Lewis et al., 2021) uses fixed retrieval, recent work demonstrates benefits of end-to-end optimization (Zamani & Bendersky, 2024; Gao et al., 2025). However, these approaches operate in text domains with well-defined similarity metrics and pretrained generators. PT-RAG is the first to apply differentiable RAG to a domain where: (1) the notion of relevant context must be learned from scratch, (2) context relevance is cell-type-dependent, and (3) both retrieved content (perturbation embeddings) and generator output (cell distributions) are non-textual.

RAG in biology. Existing biological RAG applications (Lin et al., 2024; Yu et al., 2025; Nouri et al., 2025; Jain et al., 2025) either use LLMs to retrieve textual annotations or augment protein encoders, while none of them address generation of cellular responses. PT-RAG bridges this gap, demonstrating that RAG principles can be extended beyond language when paired with learned, task-conditioned retrieval. See Appendix D for detailed discussion.

3 Method

We consider the task of predicting cellular responses to single-gene perturbations. Let $\mathcal{X} \subseteq \mathbb{R}^G$ denote the space of gene expression profiles over G genes. Given a population of control cells $\{x_i^{ctrl}\}_{i=1}^N \subseteq \mathcal{X}$ and a perturbation identifier $p^{pert} \in \mathcal{P}$ (where $\mathcal{P}$ is the set of possible perturbations, e.g., gene knockouts), our goal is to predict the distribution of perturbed cells $\{\hat{x}_i^{pert}\}_{i=1}^N$.

In this section, we first describe the standard **Generation** approach (Adduri et al., 2025) as our baseline, then introduce **Vanilla RAG** as a natural extension that incorporates retrieval-augmented context, and finally present our proposed **PT-RAG** (Perturbation-aware Two-stage Retrieval-Augmented Generation) framework. Figure 1 provides a comprehensive comparison of these three approaches. Throughout the figure, we use visual conventions to indicate trainable components (fire icon), frozen components (ice cube icon), and non-differentiable operations (dotted lines where gradients cannot flow).

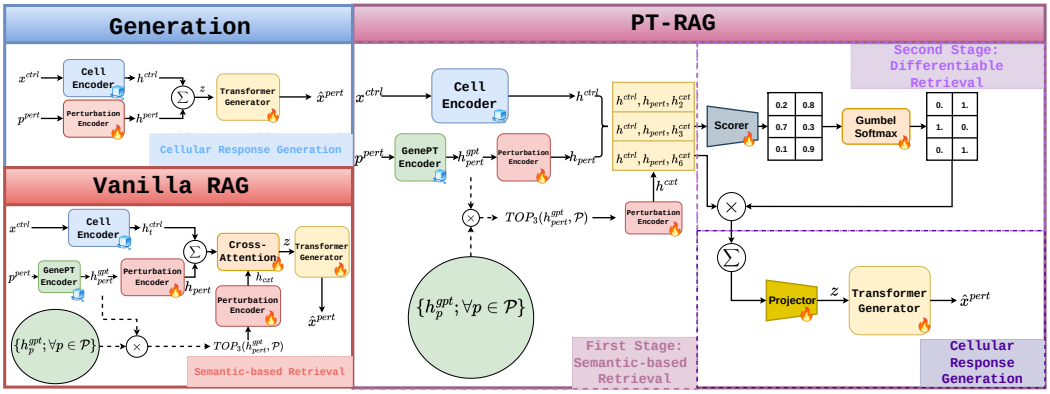

Figure 1: **Comparison of Generation, Vanilla RAG, and PT-RAG architectures.** *Left:* Generation baseline combines cell and perturbation encodings; Vanilla RAG adds non-differentiable retrieval (dotted lines). *Right:* PT-RAG uses two-stage retrieval: (1) semantic similarity for K candidates, (2) differentiable Gumbel-Softmax selection conditioned on h^{ctrl}, h_{pert}, h_k^{cxt}.

3.1 Perturbation Representation with GenePT

Unlike prior work that represents perturbations as one-hot vectors (Adduri et al., 2025), we leverage GenePT (Chen & Zou, 2023) embeddings that capture semantic relationships between genes. For each gene g, GenePT provides an embedding $h_g^{gpt} \in \mathbb{R}^d$ derived from GPT-3.5 encodings of the gene's NCBI description (Brown et al., 2014). This representation enables meaningful similarity computation between perturbations: genes with similar biological functions could be close in the embedding space, facilitating retrieval of functionally related perturbations.

We construct a perturbation database $\mathcal{D} = \{h_p^{gpt}; \forall p \in \mathcal{P}\}$ containing GenePT embeddings for all perturbations in our training set ($|\mathcal{P}| \approx 2009$ perturbations in our experiments).

3.2 Generation Baseline

Following STATE (Adduri et al., 2025), the Generation baseline encodes control cells via a frozen Cell Encoder ($h^{ctrl} = \text{CellEncoder}(x^{ctrl})$) and perturbations via a trainable Perturbation Encoder ($h^{pert} = \text{PertEncoder}(p^{pert})$). These are combined as $z = h^{ctrl} + h^{pert}$ and passed to a Transformer Generator:

$$\hat{x}^{pert} = \text{TransformerGenerator}(z) \tag{1}$$

This approach predicts responses solely from cell state and perturbation identity, without leveraging knowledge about related perturbations, limiting generalization to novel perturbations in unseen cell types.

3.3 Vanilla RAG: A Cell-Type Agnostic Approach

Vanilla RAG extends the Generation baseline by retrieving the top-K perturbations via cosine similarity on GenePT embeddings ($TOP_K(h_{pert}^{gpt}, \mathcal{P})$), then integrating them via Cross-Attention:

$$z = \text{CrossAttention}(q = h^{ctrl} + h_{pert}, k = h_{cxt}, v = h_{cxt}), \tag{2}$$

we consider the sum of h^{ctrl} and h_{pert}, as query, and the context as both key and value. Critically, retrieval is **non-differentiable** and depends *only* on h_{pert}^{gpt}, providing identical context regardless of cell type. This cell-type agnosticism is biologically unrealistic:the same perturbation can have different effects across cell types, and prevents learning which context improves generation.

3.4 PT-RAG: Perturbation-aware Two-stage Retrieval-Augmented Generation

Our proposed PT-RAG framework (Figure 1, right) addresses the limitations of Vanilla RAG through a **two-stage retrieval pipeline**: (1) an initial semantic-based retrieval to narrow down candidates,

followed by (2) a **differentiable**, **cell-type-aware** selection mechanism. This design enables end-to-end optimization of the retrieval objective jointly with generation.

3.4.1 First Stage: Semantic-based Retrieval

Identical to Vanilla RAG, we retrieve the top-K most similar perturbations via cosine similarity on GenePT embeddings:

$$\mathcal{R}_{p^{pert}} = TOP_K(h^{gpt}_{pert}, \mathcal{P}) = \{p_{(1)}, p_{(2)}, \ldots, p_{(K)}\} \tag{3}$$

This non-differentiable stage efficiently prunes the search space from $|\mathcal{P}| \approx 2009$ to K semantically relevant candidates.

3.4.2 Second Stage: Differentiable Retrieval

The second stage (Figure 1 labelled as"Second Stage: Differentiable Retrieval") is the key innovation of PT-RAG. It introduces a **cell-type-aware** selection mechanism that adaptively chooses which retrieved perturbations to include as context, conditioned on the cellular state.

Encoding. We obtain h^{ctrl}, h_{pert}, and context embeddings as in the baseline:

$$h^{cxt}_k = \text{PertEncoder}(h^{gpt}_{p_{(k)}}) \in \mathbb{R}^{d_h}, \quad k = 1, \ldots, K \tag{4}$$

Scoring. For each candidate, we construct a **triplet** $c_k = [h^{ctrl}; h_{pert}; h^{cxt}_k]$ capturing the relationship between cell state, target perturbation, and candidate context. This triplet design enables selection to depend on all three components:

$$s_k = \text{MLP}_{\text{score}}(\text{LayerNorm}(c_k)) \in \mathbb{R}^2 \tag{5}$$

where s_k outputs "`exclude`" and "`include`" logits for candidate k.

Gumbel-Softmax Selection. We apply Straight-Through Gumbel-Softmax (Jang et al., 2017; Maddison et al., 2017) to obtain hard binary decisions while maintaining differentiability:

$$w_k = \text{GumbelSoftmax}(s_k, \tau)[\texttt{include}] \in \{0, 1\}, \tag{6}$$

here τ is a temperature parameter. Unlike Vanilla RAG's fixed retrieval, this enables the model to *learn* what constitutes relevant context for different cell types through end-to-end training. Additional details on how Gumbel-Softmax works are discussed in Appendix A.1.

3.4.3 Cellular Response Generation

Context Aggregation. Each triplet c_k is projected via $h'_k = \text{MLP}_{\text{proj}}(c_k)$, then aggregated using the binary selection weights:

$$z = \sum_{k=1}^{K} w_k \cdot h'_k \tag{7}$$

Since $w_k \in \{0, 1\}$, this aggregates only selected contexts. The representation z is then passed to the Transformer Generator: $\hat{x}^{pert} = \text{TransformerGenerator}(z)$.

3.5 Training Objective

We train PT-RAG with a combination of distributional and sparsity losses:

$$\mathcal{L} = \mathcal{L}_{\text{dist}} + \lambda_{\text{sparse}} \mathcal{L}_{\text{sparse}} \tag{8}$$

Distributional Loss. Following Adduri et al. (2025), we use energy distance: $\mathcal{L}_{\text{dist}} = \text{Energy}(\hat{x}^{pert}, x^{pert})$.

Sparsity Loss. To encourage selective retrieval and prevent mode collapse from selecting all candidates, we add an L_1 penalty: $\mathcal{L}_{\text{sparse}} = \frac{1}{K} \sum_{k=1}^{K} w_k$. We use $\lambda_{\text{sparse}} = 0.1$. We motivate this choice through ablations in Appendix E.4.

4 EXPERIMENTS

We conduct comprehensive experiments to quantitatively assess the advantage of PT-RAG over baseline approaches. Our experimental design aims to answer three key questions: (1) Does retrieval-augmented generation improve perturbation response prediction? (2) Is cell-type-aware, differentiable retrieval essential, or does naive RAG suffice? (3) How does the selection mechanism behave across different cellular contexts?

4.1 DATASET AND EXPERIMENTAL SETUP

We evaluate on the Replogle dataset (Replogle et al., 2022; Nadig et al., 2024), a large-scale Perturb-seq study containing single-cell transcriptomic responses to single-gene perturbations across multiple cell types. The dataset comprises 2,009 unique perturbations (the intersection of Replogle-Nadig perturbations with available GenePT embeddings), with each perturbation measured across populations of cells characterized by 2,000 highly variable genes.

Cross-cell-type evaluation protocol. We adopt a few-shot cross-cell-type generalization task to test whether retrieval can help models adapt to unseen cellular contexts. The dataset spans four cell types: `K562` (chronic myeloid leukemia) (Luchetti et al., 1998), `Jurkat` (T-cell lymphocyte) (Abraham & Weiss, 2004), `RPE1` (retinal pigment epithelial-1) (Vanoni & Nandrot, 2023), and `HepG2` (hepatocellular carcinoma) (Moscato et al., 2015). For each target cell type, we train models to on perturbation responses from the other three cell types, then evaluate prediction of perturbation responses in the held-out target cell type. During training for each target, we provide few-shot examples (30% of perturbations) from the target cell type, while the remaining 70% are held out for validation and test, similarly to (Adduri et al., 2025). We average results across all four cell types.

This protocol directly tests PT-RAG's core hypothesis: when predicting responses in a novel or partially observed cell type, can retrieval of related perturbation responses from training data improve generalization? The limited target cell type data makes this a realistic and challenging scenario.

4.2 EVALUATION METRICS

We evaluate model performance using three categories of metrics (detailed definitions in Appendix B): (1) **Gene-level expression correlations** (Pearson, Spearman on differentially expressed genes (DEGs)) measure biological relevance by quantifying alignment between predicted and observed perturbation effects; (2) **Expression reconstruction accuracy** (MSE, RMSE, MAE) quantifies point-wise prediction errors in gene expression space, and in PCA space ($\text{MSE}_{\text{PCA50}}$); (3) **Distributional similarity** (Wasserstein distances W_1/W_2, Energy distance) assesses whether predicted cell populations capture the true heterogeneity and structure of perturbed cells in low-dimensional PCA space.

4.3 BASELINES

We compare PT-RAG against: (1) **STATE** (Adduri et al., 2025), trained using only cell state and perturbation identity (no retrieval, one-hot encoding); (2) **Vanilla RAG**, which retrieves top-$K = 32$ perturbations via GenePT similarity and integrates them via cross-attention, but without cell-type-aware selection (non-differentiable retrieval); (3) **STATE+GenePT** is the STATE model with GenePT embeddings for gene representation. Complete training configuration and hyperparameters are provided in Appendix C.

4.4 MAIN RESULTS

Table 1 presents our main experimental results across 1635 test perturbations. **Statistical methodology.** To assess statistical significance, we first applied the Shapiro-Wilk test (SHAPIRO & WILK, 1965) to each metric distribution for normality assessment. All metrics across all model groups exhibited significant departures from normality ($p < 0.05$), due to the heterogeneous nature of perturbation responses across diverse genes and cell types. Given this non-normality, we employed the Mann-Whitney U test (also known as Wilcoxon rank-sum test) (Nachar, 2008), a non-parametric alternative to the t-test that compares distributions based on ranks rather than assuming Gaussian

Table 1: **Comparison of PT-RAG with baselines on cross-cell-type generalization.** Results are mean values across 1635 test perturbations from four cell types. Statistical significance relative to PT-RAG (after FDR correction) is indicated: † (FDR-corrected $p < 0.01$), †† (FDR-corrected $p < 0.05$), ††† (FDR-corrected $p < 0.1$). Bold indicates best performance; arrows indicate direction of improvement (↑ higher is better, ↓ lower is better). Results with standard deviations are in Table 2.

Metric	STATE	STATE+GenePT	Vanilla RAG	PT-RAG (Ours)
Gene-level expression correlations				
Pearson DEG ↑	$0.624^{\dagger}$	0.631	$0.396^{\dagger}$	**0.633**
Spearman DEG ↑	$0.403^{\dagger}$	0.411	$0.307^{\dagger}$	**0.412**
Expression reconstruction accuracy				
MSE ↓	0.211	**0.210**	$0.316^{\dagger}$	**0.210**
RMSE ↓	0.458	0.458	$0.562^{\dagger}$	**0.457**
MAE ↓	$0.298^{\dagger}$	0.296	$0.429^{\dagger}$	**0.295**
$\text{MSE}_{\text{PCA50}}$ ↓	8.43	8.42	$12.64^{\dagger}$	**8.39**
Distributional similarity				
W_1 ↓	$35.70^{\dagger}$	$35.53^{\dagger\dagger\dagger}$	$48.48^{\dagger}$	**35.41**
W_2 ↓	$646.1^{\dagger}$	$638.7^{\dagger\dagger}$	$1189.5^{\dagger}$	**633.7**
Energy ↓	$9.41^{\dagger\dagger\dagger}$	9.40	$14.18^{\dagger}$	**9.33**

distributions. To control for multiple comparisons (3 model pairs × 10 metrics = 30 tests), we applied False Discovery Rate (FDR) correction using the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995). Our findings reveal several critical insights:

Vanilla RAG underperforms. Despite incorporating retrieved perturbation context, Vanilla RAG performs substantially worse than STATE without any retrieval (Pearson: 0.293 vs 0.624; Spearman: 0.220 vs 0.403). This failure highlights both that such a retrieval mode might not be helpful for generation or that cell-type-agnostic might be inefficient. Comprehensive ablation studies (Figures 8, 9) demonstrate that even varying $K \in \{2, 5, 10, 32\}$ does not rescue Vanilla RAG: while performance slightly improves with larger K (Pearson reaches 0.351 at $K = 32$), differentiable RAG is key to reach meaningful performance.

STATE+GenePT shows modest gains. Replacing one-hot encodings with GenePT embeddings provides small but consistent improvements over STATE (Pearson: 0.631 vs 0.624; Spearman: 0.411 vs 0.403), demonstrating the value of semantic gene representations. However, this model still benefits from the use of learned context selection, specifically in the Wasserstein distances.

PT-RAG achieves best overall performance. With both cell-type-aware selection and sparsity regularization, PT-RAG demonstrates statistically significant improvements over STATE in gene-level correlations (Pearson: 0.633 vs 0.624; Spearman: 0.412 vs 0.403), reconstruction accuracy (MAE: 0.295 vs 0.298), and distributional similarity (W_1: 35.41 vs 35.70; W_2: 633.7 vs 646.1), while Energy distance shows a marginally significant improvement.

4.5 Why Cell-Type-Aware Retrieval Matters

Vanilla RAG retrieves the same semantically similar perturbations for a query gene regardless of the cellular context, which can introduce biologically irrelevant information when gene function or pathway activity differs across cell types. It also exploits only the GenePT information, which relies on genes description, which totally neglects the generator's objective. In contrast, PT-RAG conditions retrieval on both the query perturbation and the target cell state, enabling the model to preferentially select context that improves generation quality. Because this selection is differentiable and trained end-to-end, retrieved perturbations that do not reduce prediction error are progressively downweighted during training. Consistent with the performance gaps in Table 1, these results indicate that task-conditioned, cell-type-aware retrieval is necessary for retrieval augmentation to provide measurable benefit in cross-cell-type perturbation prediction.

In Appendix E.4, we provide additional experiments on the sparsity weight λ_{sparse}, where we highlight how its introduction is not sensitive, but it is necessary to have it different from 0. We provide experiments on different K, both for PT-RAG and vanilla RAG.

4.6 Quantitative Analysis: Cell-Type-Specific Retrieval Patterns

A central claim of PT-RAG is that its differentiable retrieval mechanism is *cell-type-aware*: the same query perturbation should trigger selection of different contextual perturbations depending on the target cell type. To quantitatively validate this hypothesis, we analyze the overlap of retrieved perturbations across cell types, and among same query perturbations using Jaccard similarity.

Jaccard similarity analysis. For each of the 33 genes (perturbations) common to all four cell lines in our test set, we extract the top-10 most frequently selected perturbations by PT-RAG's Gumbel-Softmax mechanism in each cell type (`K562`, `Jurkat`, `HepG2`, `RPE1`). For each pair of cell types, we compute the Jaccard similarity index between their top-10 sets as $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$. where A and B are the sets of top-10 retrieved perturbations for a given gene in two different cell types, occurring within a cell population. We average these Jaccard scores across all 33 genes to produce the cell type similarity matrix shown in Figure 2.

Low overlap confirms cell-type-specific retrieval. The heatmap reveals low Jaccard similarities for all off-diagonal pairs (range: 0.185–0.196, mean: 0.191), indicating that *only ∼19% of retrieved perturbations overlap* between different cell types for the same query gene. We believe this result confirms that PT-RAG does not retrieve the context focusing completely on the query perturbations suggesting that cellular context fundamentally alters which perturbations are relevant to the generation objective.

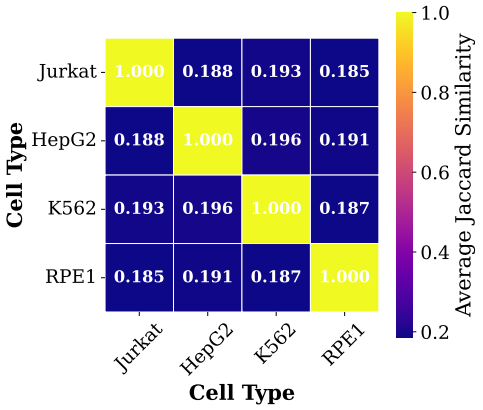

Figure 2: Jaccard similarity matrix across cell types.

Biological interpretation: Functional coherence with cell-specific selection. To understand what drives these differences, we examine specific examples.

Let us consider `WARS` (tryptophanyl-tRNA synthetase): across all cell types, PT-RAG retrieves other aminoacyl-tRNA synthetases, maintaining functional coherence with the query gene's role in translation. However, the specific synthetases differ markedly (additional examples in Appendix E.1): Jurkat selects EARS2, DARS, VARS; HepG2 selects SARS2, GART, TARS; K562 selects FARSB, KARS, FARS2; RPE1 selects KARS, GART, TARS, QARS. This is consistent with known functional relationships in amino acid metabolism: different cell types are known to rely on different tRNA charging pathways based on their protein synthesis demands, though we note these observations are suggestive rather than a rigorous biological validation.

Implications. This quantitative analysis provides strong evidence that PT-RAG's cell-type-aware retrieval is not merely a theoretical feature but is actively learned and utilized by the model. The low Jaccard similarities confirm that naive, cell-agnostic retrieval, as in Vanilla RAG, cannot capture the context-dependent relevance of perturbations. This validates our architectural choice of conditioning the retrieval scorer on $[h^{ctrl}; h_{pert}; h_k^{cxt}]$ enabling the model to tailor context selection to the specific cellular background.

5 Conclusion

We introduced PT-RAG, the first retrieval-augmented generation framework for predicting cellular responses to genetic perturbations. Our work demonstrates a fundamental insight: naive RAG, which succeeds in language domains with well-defined retrieval metrics can't be directly applied to perturbation biology, where (1) the notion of relevant context is not predefined, and (2) context rele-

vance depends critically on the target cell type. PT-RAG addresses these challenges through a two-stage differentiable retrieval pipeline that first narrows candidates via semantic similarity (GenePT embeddings), then adaptively selects cell-type-aware context via Gumbel-Softmax sampling conditioned on both cellular state and perturbation identity. Our experiments across 1635 test perturbations demonstrate that Vanilla RAG severely degrades performance (W_2: 1189.5 vs STATE's 646.1), which is a key finding itself highlighting the necessity of differentiable retrieval. PT-RAG achieves statistically significant improvements over STATE across multiple metrics; improvements over STATE+GenePT are more modest and concentrated in Wasserstein distances, reflecting that GenePT embeddings already capture functional gene similarity, with PT-RAG's marginal gain stemming from cell-type-specific context selection rather than semantic encoding. Our Jaccard score analysis reveals only $\sim$19% overlap in retrieved perturbations across cell types for identical genes(perturbations), confirming that PT-RAG effectively learns to retrieve context in a cell-aware manner, an ability enabled by jointly optimizing retrieval and generation rather than relying on fixed, non-differentiable context selection.

Limitations and future work. PT-RAG incurs approximately 1.7$\times$ more FLOPs per batch than the baselines due to the scoring and Gumbel-Softmax mechanism (see Table 7 in the Appendix for full details), a cost that should be weighed against performance gains in deployment settings. This work focuses on single-gene perturbations; extending PT-RAG to combinatorial perturbations (multiple simultaneous knockouts), chemical compounds, or CRISPR activation/interference are important next steps. We also plan to explore richer retrieval mechanisms, including GraphRAG approaches leveraging gene regulatory network structure, and multi-modal retrieval combining sequence, structure, and functional annotations. Finally, systematic biological analysis of learned selection patterns could reveal novel gene functional relationships that complement existing pathway databases.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their thoughtful feedback.

REFERENCES

Mohammad Mahdi Abootorabi, Amirhosein Zobeiri, Mahdi Dehghani, Mohammadali Mohammadkhani, Bardia Mohammadi, Omid Ghahroodi, Mahdieh Soleymani Baghshah, and Ehsaneddin Asgari. Ask in any modality: A comprehensive survey on multimodal retrieval-augmented generation, 2025. URL https://arxiv.org/abs/2502.08826.

Robert Abraham and Arthur Weiss. Jurkat t cells and development of the t-cell receptor paradigm. *Nature reviews. Immunology*, 4:301–8, 05 2004. doi: 10.1038/nri1330.

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 Tirukovular, 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. URL https://www.biorxiv.org/content/10.1101/2025.06.26.661135v1. Preprint.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J. Royal Statist. Soc., Series B*, 57:289 – 300, 11 1995. doi: 10.2307/2346101.

Garth Brown, Vichet Hem, Kenneth Katz, Michael Ovetsky, Craig Wallin, Olga Ermolaeva, Igor Tolstoy, Tatiana Tatusova, Kim Pruitt, Donna Maglott, and Terence Murphy. Gene: A gene-centered information resource at ncbi. *Nucleic acids research*, 43, 10 2014. doi: 10.1093/nar/gku1055.

Charlotte Bunne, Stefan G Stark, Gabriele Gut, Jacobo Saez-Rodriguez Del Castillo, Mitchell Levesque, Kjong-Van Lehmann, Lucas Pelkmans, Andreas Krause, and Gunnar Rätsch. Learning single-cell perturbation responses using neural optimal transport. *Nature Methods*, 20(11): 1759–1768, 2023.

Yiqun T. Chen and James Zou. Genept: A simple but hard-to-beat foundation model for genes and cells built from chatgpt. *bioRxiv*, 2023. doi: 10.1101/2023.10.16.562533. URL https://www.biorxiv.org/content/early/2023/10/19/2023.10.16.562533.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, volume 26, 2013.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization, 2025. URL https://arxiv.org/abs/2404.16130.

Guangze Gao, Zixuan Li, Chunfeng Yuan, Jiawei Li, Wu Jianzhuo, Yuehao Zhang, Xiaolong Jin, Bing Li, and Weiming Hu. D-RAG: Differentiable retrieval-augmented generation for knowledge graph question answering. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 35398–35417, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1793. URL https://aclanthology.org/2025.emnlp-main.1793/.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry, 2017. URL https://arxiv.org/abs/1704.01212.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander J Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. Lightrag: Simple and fast retrieval-augmented generation, 2025. URL https://arxiv.org/abs/2410.05779.

Sarthak Jain, Joel Beazer, Jeffrey A. Ruffolo, Aadyot Bhatnagar, and Ali Madani. E1: Retrieval-augmented protein encoder models. *bioRxiv*, 2025. doi: 10.1101/2025.11.12.688125. URL https://www.biorxiv.org/content/early/2025/11/13/2025.11.12.688125.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax, 2017. URL https://arxiv.org/abs/1611.01144.

Dominik Klein, Jonas Simon Fleck, Daniil Bobrowskiy, Lea Zimmermann, Sören Becker, Alessandro Palma, Leander Dony, Alejandro Tejada-Lapuerta, Guillaume Huguet, Hsiu-Chuan Lin, Nadezhda Azbukina, Fátima Sanchís-Calleja, Theo Uscidda, Artur Szalata, Manuel Gander, Aviv Regev, Barbara Treutlein, J. Gray Camp, and Fabian J. Theis. Cellflow enables generative single-cell phenotype modeling with flow matching. *bioRxiv*, 2025. doi: 10.1101/2025.04.11.648220. URL https://www.biorxiv.org/content/10.1101/2025.04.11.648220v1. Preprint.

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, 2021. URL https://arxiv.org/abs/2005.11401.

Xinyi Lin, Gelei Deng, Yuekang Li, Jingquan Ge, Joshua Wing Kei Ho, and Yi Liu. Generag: Enhancing large language models with gene-related task by retrieval-augmented generation. *bioRxiv*, 2024. doi: 10.1101/2024.06.24.600176. URL https://www.biorxiv.org/content/early/2024/06/28/2024.06.24.600176.

Mohammad Lotfollahi, F Alexander Wolf, and Fabian J Theis. scgen predicts single-cell perturbation responses. *Nature methods*, 16(8):715–721, 2019.

Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L Ibarra, Sanjay R Srivatsan, Mohsen Brahim, Junyue Cao, Fabian J Theis, et al. Compositional perturbation autoencoder for single-cell response modeling. *Molecular Systems Biology*, 19(6):e11517, 2023.

Francesca Luchetti, A Gregorini, Stefano Papa, Sabrina Burattini, Barbara Canonico, Maria Valentini, and E Falcieri. The k562 chronic myeloid leukemia cell line undergoes apoptosis in response to interferon-α. *Haematologica*, 83:974–80, 12 1998.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables, 2017. URL https://arxiv.org/abs/1611.00712.

Stefania Moscato, Francesca Ronca, Daniela Campani, and Serena Danti. Poly(vinyl alcohol)/gelatin hydrogels cultured with hepg2 cells as a 3d model of hepatocellular carcinoma: A morphological study. *Journal of Functional Biomaterials (JFB)*, 6:16–32, 01 2015. doi: 10.3390/jfb6010016.

Nadim Nachar. The mann-whitney u: A test for assessing whether two independent samples come from the same distribution. *Tutorials in Quantitative Methods for Psychology*, 4, 03 2008. doi: 10.20982/tqmp.04.1.p013.

Ajay Nadig, Joseph M. Replogle, Angela N. Pogson, Steven A McCarroll, Jonathan S. Weissman, Elise B. Robinson, and Luke J. O'Connor. Transcriptome-wide characterization of genetic perturbations. *bioRxiv*, 2024. doi: 10.1101/2024.07.03.601903. URL https://www.biorxiv.org/content/early/2024/07/03/2024.07.03.601903.

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

Nima Nouri, Ronen Artzi, and Virginia Savova. An agentic ai framework for ingestion and standardization of single-cell rna-seq data analysis. *bioRxiv*, 2025. doi: 10.1101/2025.07.31.667880. URL https://www.biorxiv.org/content/early/2025/08/01/2025.07.31.667880.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

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. ISSN 0092-8674. doi: https://doi.org/10.1016/j.cell.2022.05.013. URL https://www.sciencedirect.com/science/article/pii/S0092867422005979.

Maria L Rizzo and Gábor J Székely. Energy distance. *Wiley Interdisciplinary Reviews: Computational Statistics*, 8(1):27–38, 2016.

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

Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000. doi: 10.1023/A:1026543900054. URL https://doi.org/10.1023/A:1026543900054.

Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berber, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

S. S. SHAPIRO and M. B. WILK. An analysis of variance test for normality (complete samples)†. *Biometrika*, 52(3-4):591–611, 12 1965. ISSN 0006-3444. doi: 10.1093/biomet/52.3-4.591. URL https://doi.org/10.1093/biomet/52.3-4.591.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding, 2020. URL https://arxiv.org/abs/2004.09297.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. URL https://arxiv.org/abs/2302.13971.

Elora Vanoni and Emeline Nandrot. The retinal pigment epithelium: Cells that know the beat! *Advances in experimental medicine and biology*, 1415:539–545, 07 2023. doi: 10.1007/978-3-031-27681-1_79.

Zhiyin Yu, Chao Zheng, Chong Chen, Xian-Sheng Hua, and Xiao Luo. scRAG: Hybrid retrieval-augmented generation for LLM-based cross-tissue single-cell annotation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 954–970, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.53. URL https://aclanthology.org/2025.findings-acl.53/.

Hamed Zamani and Michael Bendersky. Stochastic rag: End-to-end retrieval-augmented generation through expected utility maximization. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, pp. 2641–2646, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704314. doi: 10.1145/3626772.3657923. URL https://doi.org/10.1145/3626772.3657923.

Appendix Overview

This appendix provides additional technical details, extended experimental results, and comprehensive supplementary material to support the main paper. The appendix is organized as follows:

- **Section A (Additional Preliminaries):** Detailed mathematical background on the Gumbel distribution and Gumbel-Softmax mechanism, including the straight-through estimator used for differentiable discrete selection in PT-RAG.
- **Section B (Evaluation Metrics):** Comprehensive definitions of all evaluation metrics used in our experiments, including gene-level expression correlations, reconstruction accuracy measures, and distributional similarity metrics in PCA space.
- **Section C (Implementation Details):** Complete architectural specifications, training configurations, hyperparameters, and computational resource requirements for reproducing our experiments.
- **Section D (Extended Related Works):** Expanded discussion of related work in single-cell perturbation response modeling and retrieval-augmented generation, providing broader context for PT-RAG's contributions.
- **Section E (Extended Results):** Additional experimental results including cell-type-specific retrieval examples with biological interpretation, complete cross-cell-type evaluation with standard deviations, per-cell-type disaggregated results, and comprehensive sensitivity analyses.
- **Section F (Additional Details on Statistical Tests):** Complete statistical test results with FDR correction, detailed interpretation of effect sizes in the context of perturbation prediction, and justification for evaluating improvements in cross-cell-type generalization tasks.
- **Section G (LLM usage)**.

A Additional Preliminaries

A.1 The Gumbel Distribution and the Gumbel-Softmax Temperature

The Gumbel distribution is widely used to model the maximum (or minimum) of a set of random variables. Its probability density function is asymmetric and has heavy tails, making it suitable for representing rare or extreme events. When applied to logits or scores corresponding to discrete

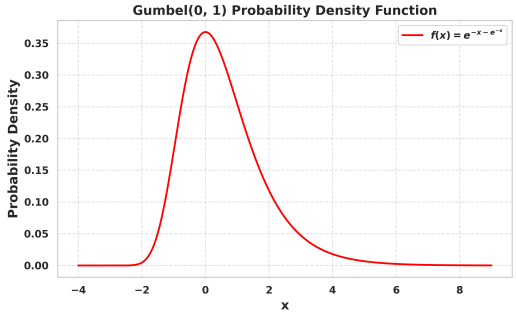

Figure 3: The pdf $f(x) = e^{-x-e^{-x}}$ of Gumbel(0, 1).

choices, the Gumbel-Softmax estimator transforms these into a probability distribution over the available options.

The probability density function of a variable $X \sim \text{Gumbel}(0, 1)$ is defined as:

$$f(x) = e^{-x-e^{-x}}, \tag{9}$$

and is shown in Figure 3.

The Gumbel-Softmax Trick. Given a categorical distribution with logits $\pi = (\pi_1, \ldots, \pi_n)$, the Gumbel-Softmax (Jang et al., 2017; Maddison et al., 2017) provides a continuous, differentiable approximation to sampling from this distribution. To sample, we first draw i.i.d. Gumbel noise $g_i \sim \text{Gumbel}(0, 1)$ for each class i, then compute:

$$y_i = \frac{\exp((\log \pi_i + g_i)/\tau)}{\sum_{j=1}^{n} \exp((\log \pi_j + g_j)/\tau)} \tag{10}$$

where $\tau > 0$ is a temperature parameter. As $\tau \to 0$, the distribution approaches a categorical distribution (one-hot), while higher τ produces softer probability distributions.

Straight-Through Estimator. For discrete decisions, we use the straight-through Gumbel-Softmax estimator. In the forward pass, we apply $\arg\max$ to obtain a hard one-hot vector (equivalent to $\tau \to 0$). In the backward pass, we compute gradients as if we had used the soft Gumbel-Softmax probabilities with finite τ. This allows gradients to flow through discrete choices while maintaining hard selections at inference time.

Formally, let $\mathbf{y}^{\text{soft}} = \text{GumbelSoftmax}(\pi, \tau)$ be the soft probabilities and $\mathbf{y}^{\text{hard}} = \text{one\_hot}(\arg\max_i y_i^{\text{soft}})$ be the hard selection. The straight-through estimator computes:

$$\mathbf{y} = \mathbf{y}^{\text{hard}} + (\mathbf{y}^{\text{soft}} - \text{sg}(\mathbf{y}^{\text{soft}})) \tag{11}$$

where $\text{sg}(\cdot)$ denotes stop-gradient. In the forward pass, this evaluates to $\mathbf{y}^{\text{hard}}$. In the backward pass, gradients flow through $\mathbf{y}^{\text{soft}}$, enabling end-to-end training of models with discrete selection.

Why This Enables Differentiable Choice. Without the straight-through trick, the $\arg\max$ operation has zero gradients almost everywhere (undefined at ties), preventing gradient-based optimization. The Gumbel-Softmax approximates the discrete sampling distribution with a continuous one, and the straight-through estimator allows us to use hard decisions (necessary for actual discrete selections) while still propagating useful gradients that guide the scoring function toward better choices. This is essential for PT-RAG: we need to actually retrieve a discrete subset of perturbations (hard selection), but we must also train the scoring MLP to learn which perturbations are relevant for each cell type (requires gradients).

B Evaluation Metrics

B.1 Gene-Level Expression Correlations

To evaluate the biological relevance of predictions, we focus on differentially expressed genes (DEGs)—those genes whose expression changes significantly under perturbation. We identify

DEGs using Welch's t-test comparing control and perturbed populations with a significance threshold of $p < 0.05$, then compute:

- **Pearson correlation**: Measures the linear relationship between predicted and observed expression changes across DEGs. High values indicate the model correctly predicts both the direction and relative magnitude of expression changes.
- **Spearman correlation**: Captures the rank-order correlation, assessing whether the model correctly identifies which genes are most strongly up- or down-regulated, regardless of exact magnitudes.

These correlation metrics are critical for downstream biological interpretation, as they indicate whether predicted perturbation effects align with true molecular responses.

B.2 Expression Reconstruction Accuracy

We measure point-wise prediction errors in the original gene expression space:

- **Mean Squared Error (MSE)**: $\frac{1}{N \cdot G} \sum_{i=1}^{N} \sum_{g=1}^{G} (\hat{x}_{i,g} - x_{i,g})^2$, where N is the number of cells and G is the number of genes.
- **Root Mean Squared Error (RMSE)**: $\sqrt{\text{MSE}}$, providing error in the original scale.
- **Mean Absolute Error (MAE)**: $\frac{1}{N \cdot G} \sum_{i=1}^{N} \sum_{g=1}^{G} |\hat{x}_{i,g} - x_{i,g}|$, which is less sensitive to outliers than MSE.
- **$\text{MSE}_{\text{PCA50}}$**: Mean squared error in PCA space, measuring overall distributional shift.

Lower values indicate more accurate reconstruction of the perturbed cell state.

B.3 Distributional Similarity in Low-Dimensional Space

Cell populations are inherently high-dimensional and exhibit complex distributional structure. To address this, we project both predicted and true cell populations to a 50-dimensional PCA space (fit on the training data) and compute distributional metrics:

- **Wasserstein distances (W_1, W_2)** (Rubner et al., 2000; Cuturi, 2013): Optimal transport distances that measure the cost of transforming the predicted distribution into the true distribution. W_1 is the 1-Wasserstein (earth mover's distance), while W_2 is the 2-Wasserstein distance. These are sensitive to both the shape and support of distributions.
- **Maximum Mean Discrepancy (MMD)** (Gretton et al., 2012): A kernel-based metric that compares distributions without assuming a specific parametric form. We use an RBF kernel with bandwidth selected via the median heuristic.
- **Energy distance** (Rizzo & Székely, 2016): A metric based on pairwise distances between and within distributions, defined as $\mathcal{E}(P, Q) = 2\mathbb{E}[\|X - Y\|] - \mathbb{E}[\|X - X'\|] - \mathbb{E}[\|Y - Y'\|]$ where $X, X' \sim P$ and $Y, Y' \sim Q$. This is our primary training objective.

These distributional metrics assess whether the predicted cell population captures the true heterogeneity and structure of perturbed cells, going beyond average accuracy to evaluate population-level fidelity—a key consideration for single-cell data where cell-to-cell variability carries biological information.

C Implementation Details

C.1 Model Architecture

All models use a Llama backbone (Touvron et al., 2023) with sequence (i.e. cell population size) of 64 cells, and batch size of 64.

Cell Encoder. We use a pretrained cell encoder[1] that maps gene expression profiles ($G = 2000$ genes) to 128-dimensional latent representations. This encoder is kept *frozen* throughout training to preserve its learned representations of cellular states, which were pretrained on large-scale single-cell datasets.

Perturbation Encoder. A trainable single-layer MLP that maps GenePT embeddings (1536-dim) to the model's latent space (128-dim).

PT-RAG-specific components:

- **Scoring MLP** ($\text{MLP}_{\text{score}}$): Two-layer network with hidden dimension 128, mapping concatenated triplets $[h^{ctrl}; h_{pert}; h_k^{cxt}] \in \mathbb{R}^{384}$ to binary logits $\mathbb{R}^2$.
- **Projection MLP** (MLP_{proj}): one-layer network with hidden dimension 128, mapping triplets to retrieval representations.

C.2 Training Configuration

Optimizer and learning rate. We use the Adam optimizer with learning rate 10^{-3}, and weight decay 0.0005.

Training schedule. All models are trained for a maximum of 50,000 steps with validation every 2,000 steps.

C.3 PT-RAG-Specific Hyperparameters

Retrieval configuration:

- $K = 32$: Number of candidate perturbations retrieved via GenePT similarity in the first stage. This balances computational efficiency (avoiding scoring all $\sim$2009 perturbations) with sufficient context diversity.
- Cosine similarity threshold: None (we always retrieve exactly top-K).

Gumbel-Softmax parameters:

- Temperature $\tau = 0.5$: Controls the sharpness of the soft distribution. Lower values produce sharper (more discrete-like) distributions.
- Straight-through estimator: Used for hard binary decisions in forward pass while maintaining differentiability in backward pass.

Loss configuration:

- Distributional loss: Energy distance (primary training objective).
- Sparsity regularization weight: $\lambda_{\text{sparse}} = 0.1$.

C.4 GenePT Embeddings and Perturbation Database

GenePT Embeddings. We use the publicly available GenePT embeddings (Chen & Zou, 2023) computed from GPT-3.5 encodings of NCBI gene descriptions. Each gene is represented as a 1536-dimensional vector. We normalize embeddings to unit norm before computing cosine similarity for retrieval: $\text{sim}(g_i, g_j) = \frac{h_{g_i}^{gpt} \cdot h_{g_j}^{gpt}}{\|h_{g_i}^{gpt}\| \|h_{g_j}^{gpt}\|}$.

Perturbation Database. Our database contains GenePT embeddings for all 2,009 perturbations in the Replogle training set. During retrieval, we exclude the query perturbation itself to avoid information leakage (i.e., if predicting response to perturbation p, we retrieve from $\mathcal{P} \setminus \{p\}$).

[1] https://huggingface.co/arcinstitute/SE-600M

C.5 Computational Resources

All experiments were conducted on NVIDIA A100 GPUs (40GB memory). Training PT-RAG to convergence (typically 30,000-40,000 steps) takes approximately 8-10 hours per target cell type.

D Extended Related Works

D.1 Single-Cell Perturbation Response Modeling

Computational prediction of single-cell perturbation responses has advanced rapidly in recent years. Early methods focused on learning perturbation-specific latent shifts: scGen (Lotfollahi et al., 2019) learned additive perturbation vectors in a VAE latent space, while CPA (Lotfollahi et al., 2023) introduced compositional modeling of multiple perturbations and covariates. These approaches treat perturbation prediction as learning a mapping in latent space but do not explicitly model cell-cell interactions within populations.

A complementary line of work leverages gene-gene interaction networks. GEARS (Roohani et al., 2024) constructs gene regulatory graphs and uses graph neural networks (Gilmer et al., 2017) to propagate perturbation effects, enabling prediction of combinatorial perturbation outcomes. However, such approaches require pre-defined interaction networks and may not capture context-dependent relationships.

More recently, optimal transport and flow-based methods have gained prominence. Bunne et al. (2023) introduced CellOT, which learns neural optimal transport maps between control and perturbed distributions. Schiebinger et al. (2019) applied optimal transport to temporal trajectory inference. The STATE framework (Adduri et al., 2025) models cell populations as sequences processed by transformer architectures, using distributional losses (Sinkhorn divergence, energy distance) to match predicted and observed perturbed populations. CellFlow (Klein et al., 2025) extends this with conditional flow matching for trajectory prediction. Our work builds on the STATE architecture, as we study how cell population evolves with perturbation, but we recognize and empirically prove the importance of relevant context in terms of 'similar' perturbations, and we propose and effective pipeline based on differentiable retrieval augmented generation that accounts for context augmentation.

D.2 Retrieval-Augmented Generation

RAG (Lewis et al., 2021) combines parametric knowledge stored in neural network weights with non-parametric knowledge retrieved from external databases. The standard RAG pipeline retrieves relevant documents based on query similarity and conditions the generator on the concatenation of query and retrieved content. This approach has proven highly effective for knowledge-intensive NLP tasks.

A key limitation of standard RAG is that retrieval is typically non-differentiable, preventing end-to-end optimization. Several recent works address this. Stochastic RAG (Zamani & Bendersky, 2024) formulates retrieval as sampling without replacement and uses Gumbel-Top-K for differentiable approximation. D-RAG (Gao et al., 2025) applies differentiable retrieval to knowledge graph question answering, demonstrating that joint optimization of retriever and generator improves performance. Our work adapts these ideas to the single-cell domain, where the "documents" are perturbation responses and the "query" includes both the target perturbation and cellular context.

RAG has been applied across diverse domains beyond text, including images, video, audio (Abootorabi et al., 2025), and graphs (Edge et al., 2025; Guo et al., 2025). Within single-cell biology, existing RAG applications have focused on text-based question answering: Lin et al. (2024) and Yu et al. (2025) developed Q&A pipelines that retrieve relevant literature or annotations, while Nouri et al. (2025) employed Agentic RAG to deploy scRNA-seq processing steps to transform raw count matrices into analysis-ready single-cell profiles. Critically, all these approaches use LLMs as generators and retrieve textual context.

More related to our setting, retrieval-augmented encoding has recently been applied to protein representations (Jain et al., 2025), demonstrating that RAG can benefit biological domains without LLMs. However, no existing work has applied RAG to cellular responses generation, where both

the retrieved context (perturbation embeddings) and the generator output (cell distributions) differ fundamentally from text. PT-RAG fills this gap as the first RAG framework for cellular response modeling.

E Extended Results

E.1 Cell-Type-Specific Retrieval Examples

To illustrate the cell-type-aware retrieval behavior quantified in Section 4.6, we present detailed examples of the top-5 most frequently selected perturbations by PT-RAG's differentiable retrieval mechanism for specific genes across different cell types.

Example 1: WARS (Tryptophanyl-tRNA Synthetase). WARS catalyzes the attachment of tryptophan to its cognate tRNA, a fundamental step in protein translation. Across all four cell types, PT-RAG consistently retrieves other aminoacyl-tRNA synthetases, demonstrating functional coherence. However, the specific synthetases differ:

- **Jurkat:** EARS2, DARS, SEPSECS, CTU2, VARS
- **HepG2:** SARS2, GART, KARS, EARS2, TARS
- **K562:** FARSB, KARS, FARS2, EPRS, DARS
- **RPE1:** KARS, GART, TARS, QARS, AARS2

This pattern reflects cell-type-specific amino acid metabolism: T-lymphocytes (Jurkat) and myeloid cells (K562) may have different protein synthesis demands than hepatocytes (HepG2) or epithelial cells (RPE1), leading to differential reliance on specific tRNA charging pathways.

Example 2: DDX27 (DEAD-Box RNA Helicase). DDX27 is involved in ribosome biogenesis and RNA processing. PT-RAG retrieves other DDX/DHX family helicases, but with distinct cell-type patterns:

- **Jurkat:** DHX15, GEMIN4, DHX16, DDX19A, DDX23
- **HepG2:** EDC4, DDX56, DHX36, GEMIN4, DDX1
- **K562:** DDX11, DDX1, DDX6, DHX36, DHX33
- **RPE1:** EDC4, DDX10, DDX56, DHX29, DDX19A

The divergence suggests that different cell types prioritize distinct RNA processing pathways, with some overlap (e.g., GEMIN4 appears in Jurkat and HepG2) but substantial cell-specific selection.

Example 3: MRPL39 (Mitochondrial Ribosomal Protein L39). MRPL39 is a component of the large subunit of the mitochondrial ribosome. PT-RAG retrieves other mitochondrial ribosomal proteins (MRPL family):

- **Jurkat:** MRPL23, MRPL50, MRPL51, MRPL37, BRIP1
- **HepG2:** RPL13, MRPL20, MRPL17, MRPL4, MRPL33
- **K562:** MRPL36, MRPL20, MRPL51, MRPL54, MRPL4
- **RPE1:** MRPL10, MRPL3, MRPL37, MRPL51, MRPL35

HepG2 notably includes RPL13 (a cytoplasmic ribosomal protein), possibly reflecting the high metabolic and protein synthesis demands of hepatocytes, which coordinate both mitochondrial and cytoplasmic translation.

Example 4: RPS2 (Ribosomal Protein S2). RPS2 is a component of the small ribosomal subunit. PT-RAG retrieves related ribosomal proteins with cell-type variation:

- **Jurkat:** RPS17, RPS28, RPL35, RPS3A, RPS15
- **HepG2:** RPL35, RPS10, RPS16, RPS17, RPL4
- **K562:** RPS24, RPL30, RPS17, RPS9, RPL4

- **RPE1:** RPS3A, RPS5, RPS28, RPL9, RPS10

The mixture of large (RPL) and small (RPS) subunit proteins varies by cell type, with only partial overlap (e.g., RPS17 appears in Jurkat, HepG2, and K562; RPL4 in HepG2 and K562). This potentially reflects different ribosome assembly or stoichiometry requirements across cellular contexts.

Example 5: TPRKB (TP53RK Binding Protein). TPRKB is involved in regulation of transcription and kinase activity. PT-RAG retrieves kinases and regulatory proteins:

- **Jurkat:** PGAM5, ING3, PTK2, PKMYT1, DBF4
- **HepG2:** TTK, PLK1, ING3, PGAM5, GAK
- **K562:** DBF4, PLK1, BRK1, PTK2, PRKRA
- **RPE1:** CDK2, GTF3C4, GAK, BRK1, PLK1

The selection of different kinases (TTK, PLK1, CDK2, PKMYT1) across cell types suggests context-dependent cell cycle regulation, with proliferative cell types (K562, RPE1) favoring cell division-related kinases.

Summary. These examples demonstrate that PT-RAG's differentiable retrieval learns biologically meaningful, cell-type-specific patterns. The model maintains functional coherence (retrieving genes from the same family or pathway), while adapting the specific selection to cellular context. This behavior validates our core hypothesis that cell-type-aware retrieval is essential for effective perturbation prediction.

E.2 Cross-Cell-Type Evaluation with Standard Deviations

Table 2 presents the complete cross-cell-type evaluation results including standard deviations, providing detailed variability information across the 1635 test perturbations from four cell types (`HepG2`, `RPE1`, `Jurkat`, `K562`). Here we can assess the standard deviation information.

Table 2: **Complete cross-cell-type generalization results with standard deviations.** Results are averaged over four target cell types, with mean ± std computed across 1635 test perturbations (375 from `HepG2`, 416 from `RPE1`, 443 from `Jurkat`, 401 from `K562`). Bold indicates best performance; arrows indicate direction of improvement (↑ higher is better, ↓ lower is better).

Metric	STATE	STATE+GenePT	Vanilla RAG	PT-RAG (Ours)
Gene-level expression correlations				
Pearson DEG ↑	0.624 ± 0.048	0.631 ± 0.051	0.396 ± 0.063	$\mathbf{0.633 \pm 0.048}$
Spearman DEG ↑	0.403 ± 0.046	0.411 ± 0.052	0.307 ± 0.041	$\mathbf{0.412 \pm 0.051}$
Expression reconstruction accuracy				
MSE ↓	0.211 ± 0.023	$\mathbf{0.210 \pm 0.022}$	0.316 ± 0.026	$\mathbf{0.210 \pm 0.022}$
RMSE ↓	0.458 ± 0.025	0.458 ± 0.024	0.562 ± 0.023	$\mathbf{0.457 \pm 0.024}$
MAE ↓	0.298 ± 0.020	0.296 ± 0.017	0.429 ± 0.021	$\mathbf{0.295 \pm 0.018}$
$\text{MSE}_{\text{PCA50}}$ ↓	8.43 ± 0.93	8.42 ± 0.88	12.64 ± 1.04	$\mathbf{8.39 \pm 0.87}$
Distributional similarity (PCA space)				
W_1 ↓	35.70 ± 1.76	35.53 ± 1.68	48.48 ± 1.71	$\mathbf{35.41 \pm 1.62}$
W_2 ↓	646.1 ± 63.6	638.7 ± 60.6	1189.5 ± 83.3	$\mathbf{633.7 \pm 58.4}$
Energy ↓	9.41 ± 1.18	9.40 ± 1.15	14.18 ± 1.17	$\mathbf{9.33 \pm 1.15}$
MMD ↓	0.0142 ± 0.0079	0.0142 ± 0.0079	0.0142 ± 0.0079	0.0142 ± 0.0079

E.3 Per-Cell-Type Results

We also report the same baselines and metrics of Table 1, but before aggregating by cell type. The metrics are computed independently for each cell-perturbation pair within each cell type. PT-RAG demonstrates superior or competitive performance across most cell types, with particularly strong results in `HepG2`, `RPE1`, and `Jurkat` cell lines.

Table 3: Comparison of PT-RAG with baselines on `Hepg2` cell-line. Results are averaged over 375 cell-perturbation samples. Bold indicates best performance; arrows indicate direction of improvement (↑ higher is better, ↓ lower is better).

Metric	STATE	STATE+GenePT	Vanilla RAG	PT-RAG (Ours)
Gene-level expression correlations				
Pearson DEG ↑	0.596 ± 0.051	0.595 ± 0.055	0.351 ± 0.054	$\mathbf{0.604 \pm 0.049}$
Spearman DEG ↑	0.398 ± 0.044	0.394 ± 0.058	0.289 ± 0.036	$\mathbf{0.401 \pm 0.049}$
Expression reconstruction accuracy				
MSE ↓	0.222 ± 0.008	0.221 ± 0.008	0.334 ± 0.019	$\mathbf{0.221 \pm 0.008}$
RMSE ↓	0.471 ± 0.009	0.470 ± 0.009	0.577 ± 0.016	$\mathbf{0.470 \pm 0.008}$
MAE ↓	0.313 ± 0.009	0.305 ± 0.008	0.442 ± 0.013	$\mathbf{0.305 \pm 0.007}$
$\text{MSE}_{\text{PCA50}}$ ↓	8.89 ± 0.32	8.86 ± 0.32	13.35 ± 0.77	$\mathbf{8.83 \pm 0.31}$
Distributional similarity (PCA space)				
W_1 ↓	36.30 ± 1.34	35.53 ± 1.20	48.68 ± 1.31	$\mathbf{35.41 \pm 1.18}$
W_2 ↓	668.5 ± 49.5	637.7 ± 43.3	1198.5 ± 64.5	$\mathbf{631.8 \pm 42.3}$
Energy ↓	9.68 ± 0.69	9.54 ± 0.67	14.25 ± 0.90	$\mathbf{9.48 \pm 0.67}$
MMD ↓	0.0189 ± 0.0079	0.0189 ± 0.0079	0.0189 ± 0.0079	0.0189 ± 0.0079

Table 4: Comparison of PT-RAG with baselines on `Rpe1` cell-line. Results are averaged over 416 cell-perturbation samples. Bold indicates best performance; arrows indicate direction of improvement (↑ higher is better, ↓ lower is better).

Metric	STATE	STATE+GenePT	Vanilla RAG	PT-RAG (Ours)
Gene-level expression correlations				
Pearson DEG ↑	0.621 ± 0.043	0.633 ± 0.043	0.450 ± 0.052	$\mathbf{0.640 \pm 0.042}$
Spearman DEG ↑	0.419 ± 0.035	0.432 ± 0.033	0.342 ± 0.027	$\mathbf{0.438 \pm 0.033}$
Expression reconstruction accuracy				
MSE ↓	0.239 ± 0.009	0.237 ± 0.008	0.332 ± 0.020	$\mathbf{0.236 \pm 0.008}$
RMSE ↓	0.489 ± 0.009	0.487 ± 0.008	0.576 ± 0.017	$\mathbf{0.486 \pm 0.008}$
MAE ↓	0.320 ± 0.009	0.317 ± 0.008	0.449 ± 0.015	$\mathbf{0.316 \pm 0.008}$
$\text{MSE}_{\text{PCA50}}$ ↓	9.56 ± 0.35	9.48 ± 0.32	13.29 ± 0.80	$\mathbf{9.44 \pm 0.33}$
Distributional similarity (PCA space)				
W_1 ↓	36.75 ± 1.25	36.62 ± 1.24	49.31 ± 1.46	$\mathbf{36.53 \pm 1.22}$
W_2 ↓	682.5 ± 47.1	676.3 ± 46.4	1227.5 ± 72.9	$\mathbf{672.8 \pm 45.8}$
Energy ↓	10.17 ± 0.84	10.14 ± 0.84	14.56 ± 0.81	$\mathbf{10.08 \pm 0.84}$
MMD ↓	0.0145 ± 0.0079	0.0145 ± 0.0079	0.0145 ± 0.0079	0.0145 ± 0.0079

E.4 Sensitivity studies

We conduct comprehensive sensitivity analyses to understand the impact of the sparsity regularization parameter λ_{sparse} on both retrieval behavior and model performance. The sparsity loss encourages the model to retrieve fewer perturbations by penalizing the number of selected contexts, which is crucial for computational efficiency and avoiding noisy retrieval contexts.

Sparsity regularization analysis. Figures 4 and 5 present our analysis of different λ_{sparse} values: 0 (no sparsity loss), 0.01, 0.10, and 1.00, evaluated on the HepG2 cell type with retrieval parameter $K = 32$. The results reveal a critical finding: without sparsity regularization ($\lambda_{\text{sparse}} = 0$), the model retrieves nearly all available perturbations (31.949 on average), leading to substantially degraded performance across all metrics.

Specifically, when $\lambda_{\text{sparse}} = 0$, we observe poor gene-level expression correlations with Pearson DEG correlation of only 0.134 and negative Spearman DEG correlation of -0.025. The reconstruction accuracy is also severely impaired, with high RMSE (0.567), MSE (0.322), and MAE (0.362) values. Distributional similarity metrics in PCA space show similar degradation, with W2 distance reaching 651.744 and Energy distance of 11.974.

Table 5: Comparison of PT-RAG with baselines on `Jurkat` cell-line. Results are averaged over 443 cell-perturbation samples. Bold indicates best performance; arrows indicate direction of improvement (↑ higher is better, ↓ lower is better).

Metric	STATE	STATE+GenePT	Vanilla RAG	PT-RAG (Ours)
Gene-level expression correlations				
Pearson DEG ↑	0.636 ± 0.048	$\mathbf{0.652 \pm 0.049}$	0.398 ± 0.056	0.649 ± 0.048
Spearman DEG ↑	0.392 ± 0.056	0.424 ± 0.058	0.318 ± 0.033	0.424 ± 0.054
Expression reconstruction accuracy				
MSE ↓	0.184 ± 0.010	0.184 ± 0.010	0.287 ± 0.016	0.184 ± 0.010
RMSE ↓	0.428 ± 0.012	0.429 ± 0.012	0.535 ± 0.014	$\mathbf{0.429 \pm 0.012}$
MAE ↓	0.282 ± 0.008	0.279 ± 0.008	0.407 ± 0.011	$\mathbf{0.277 \pm 0.008}$
$\text{MSE}_{\text{PCA50}}$ ↓	$\mathbf{7.35 \pm 0.41}$	7.36 ± 0.41	11.46 ± 0.62	7.36 ± 0.41
Distributional similarity (PCA space)				
W_1 ↓	35.94 ± 1.52	36.10 ± 1.50	49.06 ± 1.50	$\mathbf{35.82 \pm 1.48}$
W_2 ↓	654.4 ± 56.7	660.1 ± 56.1	1219.6 ± 75.4	$\mathbf{649.9 \pm 55.1}$
Energy ↓	8.15 ± 1.01	8.23 ± 1.02	13.12 ± 1.10	$\mathbf{8.14 \pm 1.01}$
MMD ↓	0.0126 ± 0.0067	0.0126 ± 0.0067	0.0126 ± 0.0067	0.0126 ± 0.0067

Table 6: Comparison of PT-RAG with baselines on `K562` cell-line. Results are averaged over 401 cell-perturbation samples. Bold indicates best performance; arrows indicate direction of improvement (↑ higher is better, ↓ lower is better).

Metric	STATE	STATE+GenePT	Vanilla RAG	PT-RAG (Ours)
Gene-level expression correlations				
Pearson DEG ↑	$\mathbf{0.639 \pm 0.037}$	0.638 ± 0.037	0.381 ± 0.046	0.633 ± 0.041
Spearman DEG ↑	$\mathbf{0.406 \pm 0.040}$	0.391 ± 0.044	0.274 ± 0.029	0.383 ± 0.047
Expression reconstruction accuracy				
MSE ↓	0.200 ± 0.008	$\mathbf{0.202 \pm 0.008}$	0.315 ± 0.015	0.201 ± 0.008
RMSE ↓	$\mathbf{0.448 \pm 0.009}$	0.449 ± 0.009	0.561 ± 0.013	0.448 ± 0.009
MAE ↓	$\mathbf{0.280 \pm 0.007}$	0.283 ± 0.007	0.419 ± 0.009	0.284 ± 0.007
$\text{MSE}_{\text{PCA50}}$ ↓	$\mathbf{8.02 \pm 0.33}$	8.07 ± 0.32	12.61 ± 0.59	8.04 ± 0.33
Distributional similarity (PCA space)				
W_1 ↓	33.77 ± 1.21	$\mathbf{33.76 \pm 1.18}$	46.77 ± 1.23	33.80 ± 1.15
W_2 ↓	578.1 ± 42.4	$\mathbf{576.9 \pm 41.3}$	1108.5 ± 58.9	577.1 ± 40.2
Energy ↓	$\mathbf{9.73 \pm 0.91}$	9.80 ± 0.92	14.89 ± 0.97	9.74 ± 0.91
MMD ↓	0.0113 ± 0.0069	0.0113 ± 0.0069	0.0113 ± 0.0069	0.0113 ± 0.0069

In contrast, introducing sparsity regularization with $\lambda_{\text{sparse}} \in \{0.01, 0.10, 1.00\}$ significantly improves performance while reducing the number of retrieved perturbations to 12.536, 6.612, and 4.915 respectively. Remarkably, all performance metrics stabilize across these non-zero sparsity values: Pearson DEG correlations consistently range from 0.594-0.604, Spearman DEG correlations from 0.386-0.401, and reconstruction errors (RMSE: 0.469-0.470, MSE: 0.220-0.221, MAE: 0.305) remain nearly identical.

This sensitivity analyses demonstrate that the sparsity regularization gives to the differentiable RAG the capabilities to improve performance, and it is fundamental to avoid mode collapse and retrieving always all the perturbations. The robustness of performance across different non-zero λ_{sparse} values (0.01, 0.10, 1.00) indicates that the model is not overly sensitive to the exact choice of this hyperparameter, provided it is sufficiently large to encourage meaningful sparsity in the retrieval process.

Retrieval size analysis. To further investigate the robustness of our approach to hyperparameter choices, we compare PT-RAG performance with two different retrieval sizes: $K = 16$ and $K = 32$, across the same sparsity regularization values. Figures 6 and 7 present this comparative analysis on the `HepG2` cell type.

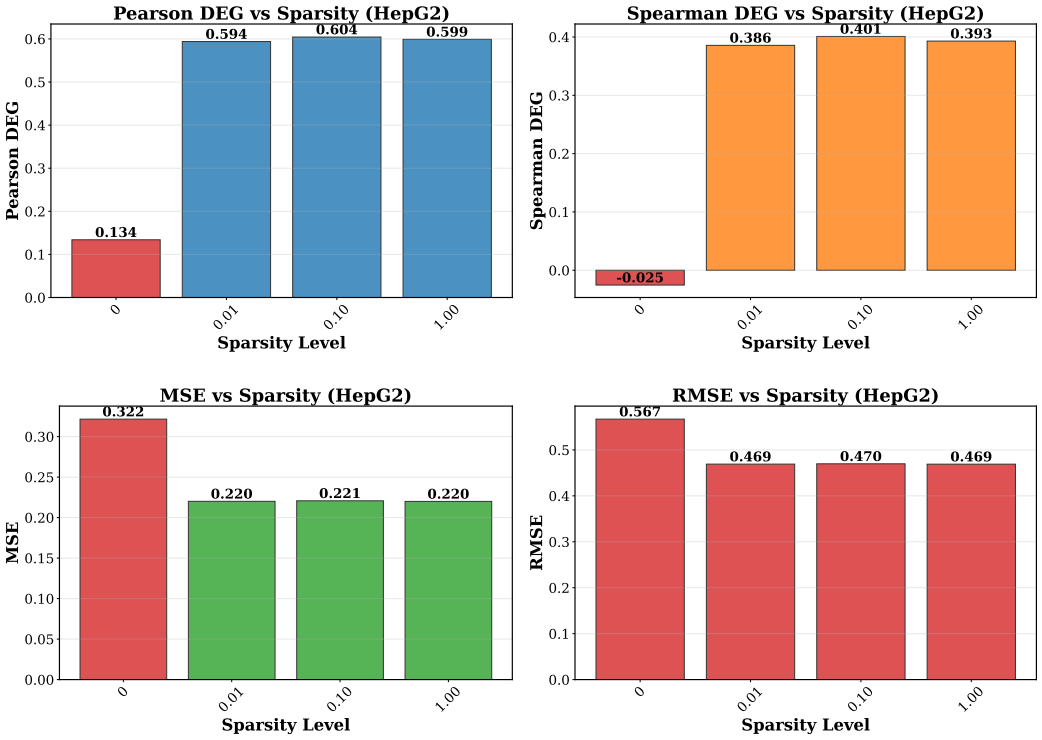

Figure 4: **Sensitivity analysis of sparsity regularization parameter λ_{sparse} - Core metrics.** Performance metrics across different sparsity levels on `HepG2` cell type with $K = 32$, focusing on gene-level correlations and reconstruction accuracy.

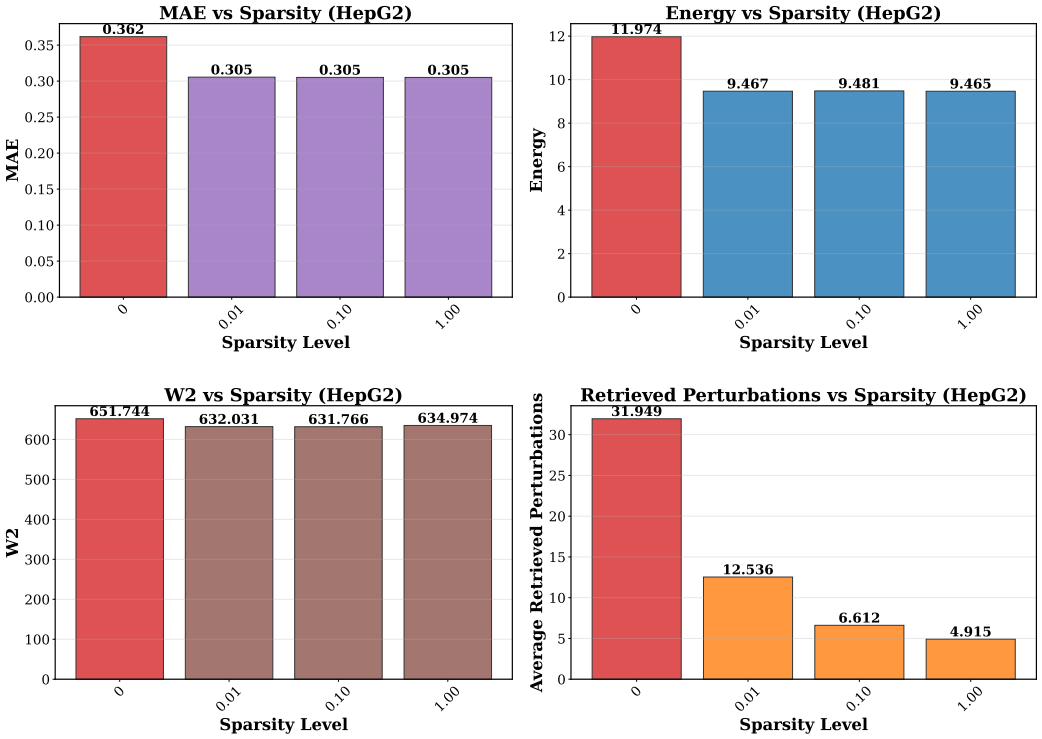

Figure 5: **Sensitivity analysis of sparsity regularization parameter λ_{sparse} - Distributional metrics and retrieval behavior.** Additional performance metrics and retrieval patterns across different sparsity levels.

The results demonstrate that PT-RAG exhibits remarkable stability across different K values, with only modest variations between $K = 16$ and $K = 32$. For gene-level expression correlations, both configurations show comparable performance: Pearson DEG correlations range from 0.578-0.604 for $K = 16$ and 0.594-0.604 for $K = 32$, while Spearman DEG correlations span 0.369-0.401 for $K = 16$ and 0.386-0.401 for $K = 32$. The reconstruction accuracy metrics (MSE and MAE) show similarly tight performance bands, with MSE values around 0.220-0.221 and MAE values consistently at 0.305-0.306 for both K settings.

Interestingly, the distributional similarity metrics reveal slight variations that depend on the sparsity level. For W2 distance in PCA space, $K = 32$ shows marginally better performance at lower sparsity levels ($\lambda_{\text{sparse}} = 0.01, 0.10$), while both configurations converge at higher sparsity ($\lambda_{\text{sparse}} = 1.00$). The Energy distance metric shows the opposite trend, with $K = 16$ performing slightly better at lower sparsity levels.

These findings suggest that the choice between $K = 16$ and $K = 32$ has minimal impact on overall model performance, indicating that PT-RAG is robust to the specific retrieval size within this range. This robustness is practically valuable, as it allows practitioners to choose K based on computational constraints without significantly sacrificing performance quality.

Vanilla RAG retrieval size ablation: The critical role of differentiable selection. To understand whether the failure of Vanilla RAG (reported in the main results) is primarily due to insufficient retrieval size K or fundamentally stems from its non-differentiable, cell-type-agnostic design, we conduct an extensive ablation study evaluating Vanilla RAG with varying $K \in \{2, 5, 10, 32\}$ on the `HepG2` cell type. Figures 8 and 9 present this comprehensive comparison, where we assess Vanilla RAG's performance across different retrieval sizes and against PT-RAG with $K = 32$.

The results reveal that increasing K for Vanilla RAG does provide modest improvements, the gains are inconsistent across metrics and plateau well below PT-RAG's performance. Specifically, for gene-level expression correlations, Pearson DEG improves from 0.293 at $K = 2$ to 0.351 at $K = 32$, and falls 42% short of PT-RAG's 0.604. Spearman DEG shows a similar non-monotonic pattern: starting at 0.220 ($K = 2$), dipping to 0.170 ($K = 5$), recovering to 0.189 ($K = 10$), and reaching 0.289 ($K = 32$), but remaining 28% below PT-RAG's 0.401.

Reconstruction accuracy metrics exhibit even more striking patterns. RMSE shows an initial degradation as K increases from 2 to 5 (0.608 $\rightarrow$ 0.631), then gradually improves to 0.577 at $K = 32$, but still lags behind PT-RAG which achieves 0.470. Similarly, MSE follows a non-monotonic trajectory: starting at 0.370 ($K = 2$), peaking at 0.399 ($K = 5$), then declining to 0.334 ($K = 32$), with PT-RAG reaching 0.221.

The distributional similarity metrics in PCA space tell a consistent story..

Why differentiable retrieval dominates. These ablation results provide strong empirical evidence for our central thesis: the failure of Vanilla RAG is not due to insufficient retrieval, as even with $K = 32$ (matching PT-RAG's retrieval pool), Vanilla RAG underperforms. Instead, the fundamental limitation lies in its cell-type-agnostic, non-differentiable retrieval mechanism. Without conditioning on cellular context and without gradient-based optimization of which perturbations to include, Vanilla RAG cannot learn to filter relevant from irrelevant context. Increasing K provides more candidate perturbations but simultaneously introduces more noise, leading to the non-monotonic performance curves observed.

In contrast, PT-RAG's differentiable selection mechanism actively learns to identify which retrieved perturbations are informative for each cell type, effectively leveraging the larger retrieval pool while filtering out misleading context. This explains why PT-RAG with $K = 32$ consistently outperforms all Vanilla RAG configurations across all metrics: it's not about retrieving more perturbations, but about *learning which perturbations to use*.

E.5 Computational Costs

Table 7 compares the computational requirements across all evaluated models.

All models have comparable parameter counts ($\sim$20–21M parameters). PT-RAG requires approximately 1.7$\times$ more FLOPs per batch (2.86B vs. 1.67B) compared to baselines due to triplet construction, scoring, and Gumbel-Softmax sampling for $K = 32$ candidates. This computational overhead

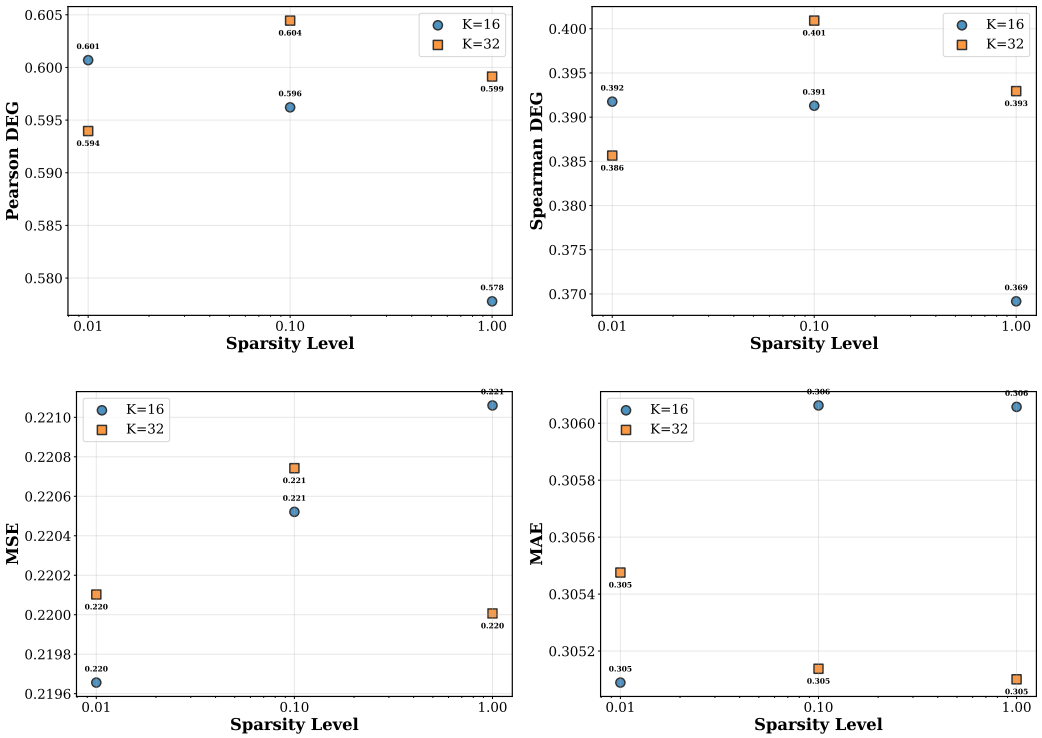

Figure 6: Comparison of retrieval sizes $K = 16$ vs $K = 32$ across sparsity regularization levels - Core metrics. Performance comparison on `HepG2` cell type.

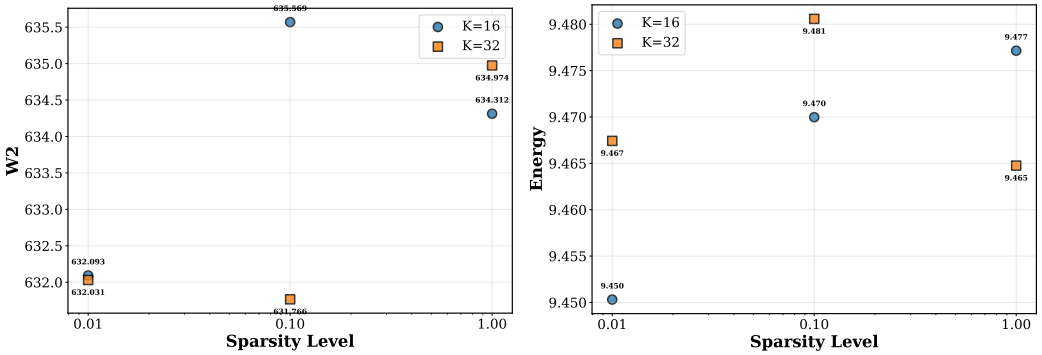

Figure 7: Comparison of retrieval sizes $K = 16$ vs $K = 32$ across sparsity regularization levels - Core metrics. Performance comparison on `HepG2` cell type.

Table 7: **Computational costs comparison.** We report model parameters, average FLOPs per batch (64 cells), and average FLOPs per cell for all methods. PT-RAG incurs higher computational cost due to the Gumbel-Softmax selection mechanism and scoring MLP, but remains tractable.

Model	Parameters	FLOPs/Batch	FLOPs/Cell
STATE	20,936,480	1.67B	34.9M
STATE+GenePT	20,874,016	1.67B	34.8M
Vanilla RAG	21,120,288	1.67B	34.8M
PT-RAG (Ours)	20,973,602	2.86B	59.9M

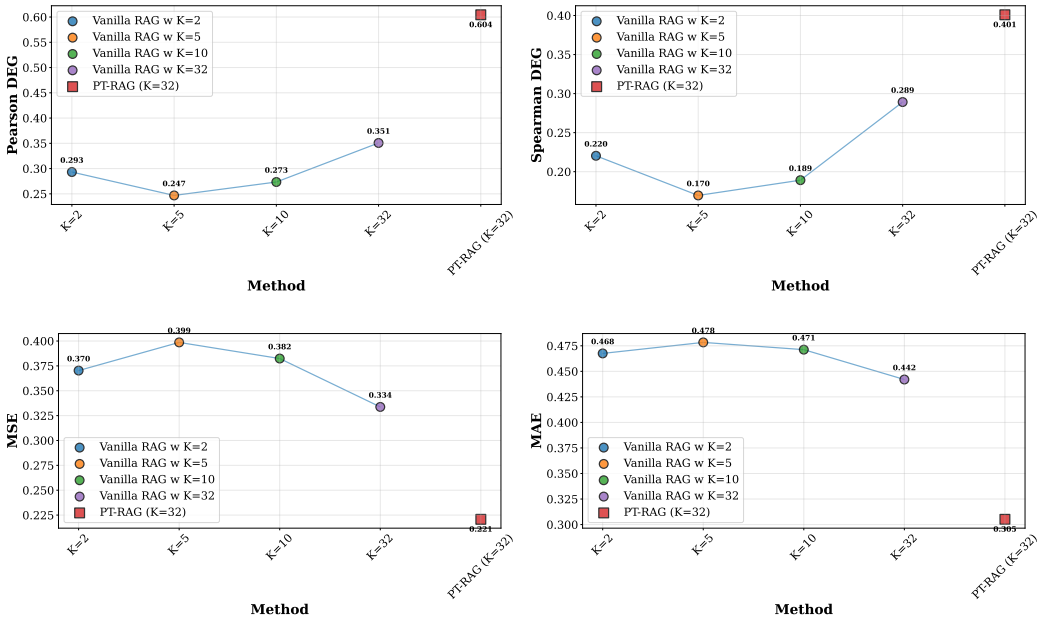

Figure 8: Performance comparison of Vanilla RAG with varying retrieval sizes $K \in \{2, 5, 10, 32\}$ against PT-RAG ($K = 32$) on `HepG2` cell type.

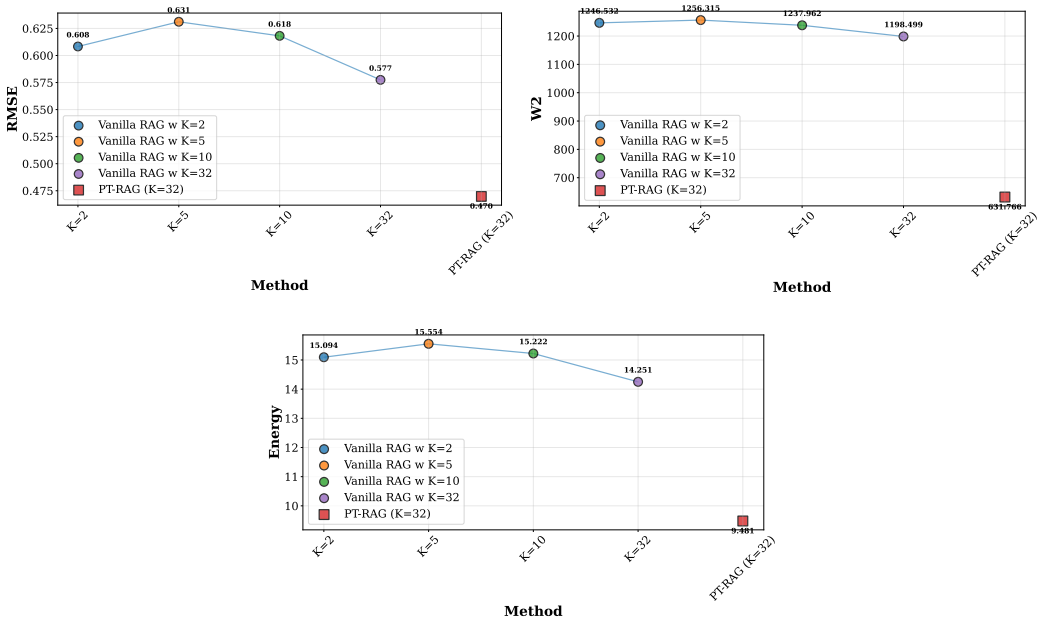

Figure 9: Performance comparison of Vanilla RAG with varying retrieval sizes $K \in \{2, 5, 10, 32\}$ against PT-RAG ($K = 32$) on `HepG2` cell type.

Table 8: **Statistical significance comparison of PT-RAG against baselines.** FDR-corrected p-values (p_{FDR}) from Mann-Whitney U tests with Benjamini-Hochberg correction. Significance levels: † ($p_{FDR} < 0.01$), †† ($p_{FDR} < 0.05$), † † † ($p_{FDR} < 0.1$). All comparisons based on $n = 1635$ test perturbations.

Metric	PT-RAG vs. STATE	PT-RAG vs. STATE+GenePT
Gene-level expression correlations		
Pearson DEG	$2.44 \times 10^{-8\dagger}$	0.590
Spearman DEG	$4.89 \times 10^{-10\dagger}$	0.785
Expression reconstruction accuracy		
MSE	0.577	0.590
RMSE	0.577	0.590
MAE	$9.33 \times 10^{-5\dagger}$	0.590
MSE_{PCA50}	0.577	0.590
Distributional similarity (PCA space)		
W_1	$3.41 \times 10^{-6\dagger}$	$0.082^{\dagger\dagger\dagger}$
W_2	$5.47 \times 10^{-8\dagger}$	$0.041^{\dagger\dagger}$
Energy	$0.082^{\dagger\dagger\dagger}$	0.186

is justified by PT-RAG's substantial performance gains across all metrics, while the absolute cost (60M FLOPs per cell) remains highly tractable on modern hardware.

F Additional Details on Statistical Tests

To ensure full transparency and reproducibility of our statistical analysis, we provide complete details of the Mann-Whitney U tests with FDR correction. Table 8 reports the FDR-corrected p-values (p_{FDR}) from Benjamini-Hochberg correction for PT-RAG comparisons against both STATE baselines.

PT-RAG vs. STATE. Our primary comparison reveals five metrics with highly significant FDR-corrected improvements ($p_{FDR} < 0.001$): Pearson DEG ($p_{FDR} = 2.44 \times 10^{-8}$), Spearman DEG ($p_{FDR} = 4.89 \times 10^{-10}$), MAE ($p_{FDR} = 9.33 \times 10^{-5}$), W1 ($p_{FDR} = 3.41 \times 10^{-6}$), and W2 ($p_{FDR} = 5.47 \times 10^{-8}$). Energy distance shows marginal significance with $p_{FDR} = 0.082$. The W2 metric achieves the most significant improvement, reflecting PT-RAG's enhanced modeling of cell population distributions in low-dimensional space.

PT-RAG vs. STATE+GenePT. Comparison against the GenePT-enhanced baseline yields one significant improvement: W2 distance ($p_{FDR} = 0.041$), and one marginal improvement: W1 distance ($p_{FDR} = 0.082$). The smaller number of significant improvements relative to STATE reflects that GenePT embeddings already capture functional gene relationships, leaving less room for retrieval to add value. Nonetheless, the significant W2 improvement demonstrates that PT-RAG's cell-type-aware retrieval provides complementary benefits beyond semantic embeddings alone.

G Use of Large Language Models

Large Language Models were used as writing assistants to improve clarity and presentation of the manuscript and software documentation. All scientific ideas, methods, experiments, and conclusions are solely those of the authors, and all LLM-assisted content was reviewed and edited for correctness prior to inclusion.