LANGPERT: LLM-DRIVEN CONTEXTUAL SYNTHESIS FOR UNSEEN PERTURBATION PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting cellular responses to previously unseen genetic perturbations remains a fundamental challenge in computational biology, with broad applications in understanding gene function, disease mechanisms, and therapeutic development. Despite advances in computational approaches, developing models that generalise effectively to novel perturbations continues to be difficult. Large Language Models (LLMs) have shown promise in biological applications by synthesizing scientific knowledge, but their direct application to high-dimensional gene expression data has been impractical due to numerical limitations. We propose LangPert, a novel hybrid framework that leverages LLMs to guide a downstream k-nearest neighbors (kNN) aggregator, combining biological reasoning with efficient numerical inference. We demonstrate that LangPert achieves state-of-the-art performance on single-gene perturbation prediction tasks across multiple datasets.

1 INTRODUCTION

Understanding cellular responses to perturbations, in particular gene knockouts, is a cornerstone in deciphering complex biological systems. By systematically altering cellular components via genetic perturbations, researchers can observe cellular behavior changes in genome-wide gene expression vectors, thereby uncovering the genetic mechanisms underlying health and disease. Mapping out even the single-gene-perturbation landscape in a particular cell line requires significant experimental effort. This challenge scales with the number of cell types under investigation, and grows exponentially when considering combinatorial perturbations. This motivates the development of computational approaches that could reduce the need for exhaustive experimental testing by predicting the results of unseen perturbations, particularly important when trying to deconvolute the multicellular functional impact of the thousands of genetic variants associated with complex metabolic disorders like Type 2 Diabetes and Obesity.

The ability to computationally predict the results of unseen genetic perturbations would dramatically accelerate biological discovery while reducing experimental costs. This challenge has recently attracted significant attention, with researchers developing various approaches to leverage prior biological knowledge. These range from transformer-based foundation models, such as scGPT (Cui et al., 2024) and scFoundation (Hao et al., 2024), pre-trained on large-scale cell atlases to methods that explicitly incorporate structured knowledge like gene-gene relationships and ontologies (Roohani et al., 2023). However, despite the sophistication of these approaches, recent studies have revealed a surprising finding: seemingly simple baselines, such as predicting the mean expression response, often outperform more complex deep learning methods (Ahlmann-Eltze et al., 2025; Kernfeld et al., 2024; Wong et al., 2025).

These findings motivate the exploration of alternative approaches that can better leverage biological knowledge while maintaining the ability to handle high-dimensional gene expression data. Particularly promising are methods that can incorporate the vast amount of unstructured biological knowledge present in the scientific literature, which contains detailed and interrelated (if unstructured) information about gene functions, interactions, and regulatory mechanisms that could potentially inform and enhance current perturbation prediction methods.

Large Language Models (LLMs) have recently demonstrated remarkable success in scientific applications, particularly in assisting with data analysis, literature mining, and complex reasoning tasks (Guo et al., 2025; Gao et al., 2024). Their ability to synthesize knowledge from vast scientific corpora and

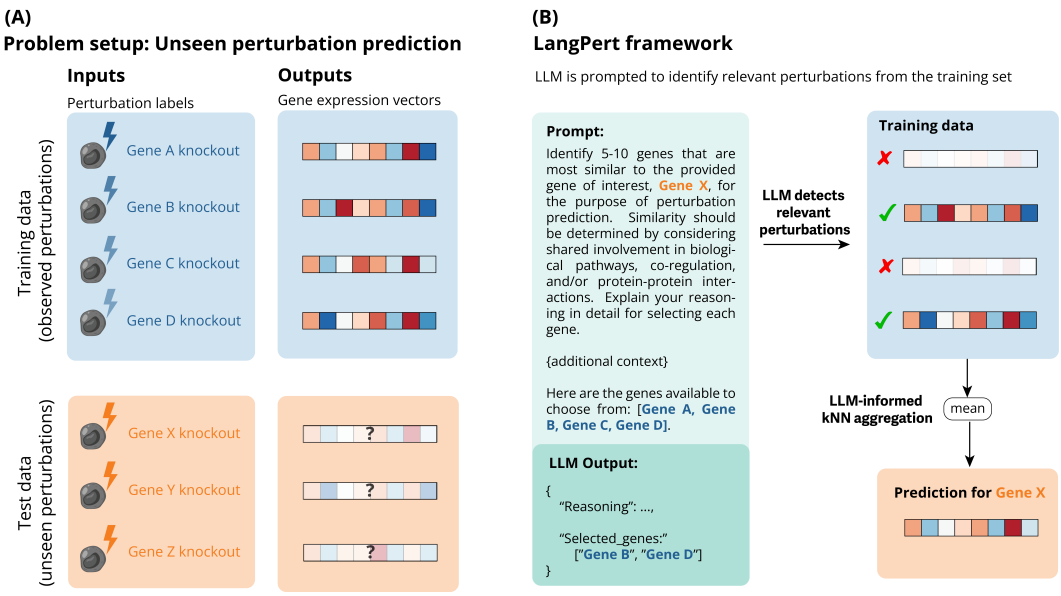

Figure 1: **(A) The task of unseen perturbation outcome prediction illustrated.** The training set consists of pairs $\{(\mathbf{x}_n, \mathbf{y}_n)\}$, where inputs $\mathbf{x}_n$ are discrete perturbation labels and outputs $\mathbf{y}_n$ are high-dimensional gene expression vectors. At test time, the goal is to predict outcomes $\mathbf{y}_*$ corresponding to unseen perturbation labels $\mathbf{x}_*$. **(B) Our proposed LLM-based LangPert framework.** Instead of using the LLM to directly predict high-dimensional $\mathbf{y}_*$, the LLM is tasked with finding a relevant small subset from the training perturbation labels $\{(\mathbf{x}_n)\}$ for every $\mathbf{x}_*$ from the held out test set. We use the LLM output to aggregate the corresponding subset of training set expression vectors, effectively resulting in an LLM-informed contextual nearest neighbour prediction.

perform step-by-step reasoning has shown promise for various scientific domains. However, their ability to handle high-dimensional numerical data remains limited due to tokenization constraints and fundamental challenges in numerical computation (Gambardella et al., 2024; Johnson and Hyland-Wood, 2024) with particular difficulties in generating precise continuous values. This limitation is particularly acute in transcriptional response modeling, where predictions are high-dimensional and must capture complex, noisy patterns across thousands of genes.

To leverage the impressive knowledge synthesis capabilities of LLMs while overcoming their limitations with high-dimensional data, we propose a novel framework for predicting cellular outcomes to unseen genetic perturbations. Instead of directly interacting with expression data, we engineer a system where the LLM guides a downstream k-nearest neighbors (kNN) aggregator, combining contextual biological insights with efficient numerical computation, as illustrated in Figure 1. We demonstrate that this hybrid LLM-kNN framework achieves state-of-the-art performance on single-gene perturbation prediction tasks.

2 Background

2.1 Existing methods for perturbation data

VAE-based methods Variational Autoencoders (VAEs) (Kingma and Welling, 2014; Rezende et al., 2014) have been widely employed as a (conditional) generative model for single-cell perturbation data. Various adaptations, including the Compositional Perturbational Autoencoder (Lotfollahi et al., 2023), sVAE+ (Lopez et al., 2023) and SAMS-VAE (Bereket and Karaletsos, 2023), all capture perturbation effects in the latent space. However, such VAE-based approaches fundamentally lack a mechanism to generalize to *unseen* single-gene perturbations, as they are inherently limited to modeling only the perturbations observed during training. This is because these methods effectively represent perturbations as distinct categorical conditions—while they learn representations of all training-set perturbations, the respective representations for test-set perturbations are absent. The

one-hot or categorical encoding of perturbations does not provide a natural way to infer relationships between observed and unseen conditions, preventing these models from extrapolating beyond the training set.

GEARS Roohani et al. (2023) proposed a graph neural-network based model called GEARS for perturbation response modelling. Unlike VAE-based approaches, GEARS explicitly incorporates structured biological prior knowledge, allowing it to generalize to perturbations involving genes that have not been experimentally tested in the prediction task of interest. Specifically, GEARS incorporates information about gene-gene relationships in two ways, using a gene co-expression graph as well as a gene ontology (GO) knowledge graph.

Single-cell foundation models The success of transformer-based foundation models has spurred their adaptation to single-cell biology. Models such as GeneFormer (Theodoris et al., 2023), scGPT (Cui et al., 2024), and scFoundation (Hao et al., 2024) are pre-trained on large-scale single-cell atlases to learn gene expression patterns. While these models have shown promise in various single-cell analysis tasks, careful evaluations have questioned their fundamental capabilities compared to simpler approaches (Boiarsky et al., 2024; Kedzierska et al., 2023). Particularly in the context of perturbation prediction, where models are fine-tuned to predict responses to unseen genetic interventions, these sophisticated approaches often fail to outperform simple baselines such as mean prediction (Ahlmann-Eltze et al., 2025; Kernfeld et al., 2024).

LLM-informed gene embeddings LLMs have been widely applied across scientific domains, including biology (Lee et al., 2020). In the context of gene-level biological knowledge, recent methods such as GenePT (Chen and Zou, 2024) have taken a novel approach: instead of training foundation models on gene expression data, they leverage LLMs' understanding of scientific literature to generate gene embeddings. These embeddings, derived from NCBI text descriptions of genes, have shown promising results in observational single-cell analysis tasks. Building on this idea, Märtens et al. (2024) extended the approach to interventional settings, developing a GP+LLM model that combines a Gaussian Process with literature-derived embeddings as well as protein language model embeddings to predict perturbation outcomes, demonstrating that LLMs can effectively encode biologically relevant prior knowledge.

2.2 Leveraging LLMs for Biological Knowledge

Large Language Models have emerged as powerful tools for synthesizing biological knowledge from scientific literature, offering new approaches to understanding gene functions and relationships. While methods like GP+LLM (Märtens et al., 2024) have shown promise by leveraging LLM-derived embeddings in predictive models, these embeddings are inherently *static*, as they are extracted from a fixed body of literature (e.g., NCBI abstracts in (Chen and Zou, 2024)) as condensed into an LLM at a specific training data corpus cutoff-date. This limits their adaptability when reasoning about unseen perturbations, where context-dependent interactions may play a crucial role.

Recently, Wu et al. (2024) introduced PerturbQA, a benchmark that represents perturbation data as "(perturbation, gene, outcome)" triplets where outcome is a binary variable. They consider two types of outcomes: whether a particular gene is differentially expressed, and for differentially expressed genes, whether they are up- or down-regulated. This classification setup makes the data more amenable for use with LLMs. In contrast, our work aims to predict high-dimensional gene expression vectors directly, an approach that presents different technical challenges for LLM implementation.

This direct application of LLMs to high-dimensional perturbation modeling is constrained by tokenization limits and numerical precision issues. These challenges motivate our hybrid approach, which integrates LLM-driven biological reasoning with computational models designed for handling high-dimensional data – a paradigm we explore in LangPert.

3 LangPert: A Hybrid LLM-kNN framework

We propose LangPert, a framework that leverages LLMs' ability to reason about biological mechanisms while ameliorating their limitations in handling high-dimensional data. Instead of using

LLMs to generate static embeddings or to directly predict expression values, LangPert employs an LLM to identify biologically relevant training examples that can inform predictions for unseen perturbations. These LLM-selected examples then guide a downstream aggregation function (here we employ a k-nearest neighbors (kNN) scheme) that performs the actual numerical computations in the high-dimensional expression space.

Problem formulation As illustrated in Figure 1, the task of unseen perturbation response prediction presents a challenging supervised learning challenge, where given pairs $\{(\mathbf{x}_n, \mathbf{y}_n)\}$, inputs $\mathbf{x}_n$ are discrete perturbation labels and outputs $\mathbf{y}_n \in \mathbb{R}^D$ are high-dimensional numeric readouts corresponding to those labels. The goal is to predict responses $\mathbf{y}_*$ for test inputs $\mathbf{x}_*$ which are distinct from those in the training set. This makes the task fundamentally different from standard supervised learning, as it requires extrapolation to entirely new perturbations, which would fall outside of the one-hot representational space of the training data perturbation categories.

Naive application of LLMs One approach to utilise LLMs for this problem is via in-context learning (ICL), where $(\mathbf{x}_n, \mathbf{y}_n)$ pairs are given to the LLM as part of the input prompt, alongside with new inputs $\mathbf{x}_*$. However, this approach is problematic due to the high dimensionality of gene expression vectors. Therefore, we propose a strategy to remedy this challenge inherent in a naive / brute force application of LLMs in this doamin.

LangPert adaptation for high-dimensional outcomes Specifically, we propose to only show the model training inputs $\{\mathbf{x}_n\}$ alongside with a test input $\mathbf{x}_*$. As these inputs correspond to perturbation labels – for genetic perturbations these would be gene names – LLMs have demonstrated strong capabilities in reasoning about biological relationships and identifying functionally related genes through their training on scientific literature (Hu et al., 2025). We leverage these established capabilities, using the LLM's comprehensive knowledge of biological systems and its proven few-shot learning abilities (Brown et al., 2020) to identify genes from the training set that are functionally relevant for the prediction target $\mathbf{x}_*$. This biological reasoning can be further enhanced through relevant context included in the prompt by a human user.

Contextual aggregation Given the LLM output of a subset of gene perturbation labels relevant to the input prompt, the final output of an unseen perturbation prediction pipeline can be made via an aggregation / reduction of the gene expression vectors of those relevant genes. That is, given a relevant subset $\mathcal{G}$ where $\{(\mathbf{x}_n, \mathbf{y}_n)\}$ for $n \in \mathcal{G}$, we propose to make the prediction $\frac{1}{|\mathcal{G}|}\sum_{n\in\mathcal{G}} \mathbf{y}_n$ for the unseen output. This can be interpreted as a k nearest neigbour predictor, where the relevant neighbours are identified by the LLM and then averaged. In principle, different aggregation techniques can be adapted here, for examaple weighting the inputs, using nonparametric (e.g. median) reductions, or even passing the subset data $\{(\mathbf{x}_n, \mathbf{y}_n)\}$ to a small tabular prediction framework. In this work, we found it sufficient to use simple averaging to achienve SOTA results, but future work investigating more principled aggregation/reduction strategies may yield even stronger predictions.

In summary, we have proposed a hybrid LLM-kNN framework, where predictions take the following form

$$\mathbf{y}_* = \frac{1}{\sum_n w_n} \sum_n w_n \mathbf{y}_n, \quad \text{where} \quad w_n = \text{LLM}(\mathbf{x}_*, \{\mathbf{x}_n\}, \text{context}) \in \{0, 1\}$$

where the LLM sees the test perturbation label $\mathbf{x}_*$, all training labels $\{\mathbf{x}_n\}$ and potentially additional information presented in the prompt. The number of chosen training perturbations, i.e. the number of nearest neighbours $k := \sum_n w_n$ can either be specified in the prompt or remain unspecified, giving the LLM flexibility to choose.

4 Results

4.1 Experimental Setup

Datasets For evaluation, we consider data from large-scale Perturb-seq screens across two cell lines: the leukemia cell line (K562) and the retinal pigment epithelial (RPE1) cell line from (Replogle

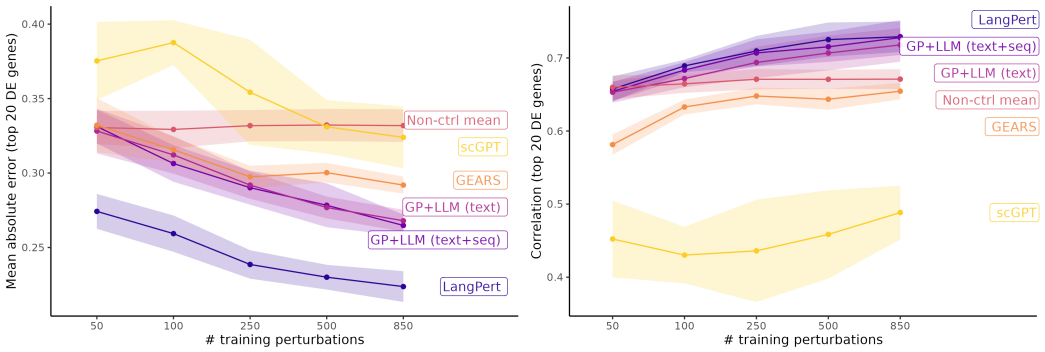

Figure 2: Performance comparison across models (scGPT, GEARS, non-control mean baseline, GP+LLM, and LangPert) evaluated using mean absolute error (MAE, lower is better) and Pearson correlation between predicted and observed differences from control cells (higher is better). Shaded regions indicate ±1.96 standard errors across data splits. Model performance is shown as a function of training set size, varying from 50 to 850 perturbations.

et al., 2022). We use the version of the data curated by Roohani et al. (2023)[1], with a total of 1092 perturbations in the K562 cell line and a total of 1543 perturbations in the RPE1 cell line.

Experimental details and metrics In all our evaluations, we assess performance in a 5-fold cross-validation setting, so in the end, all metrics are calculated on the entire set of 1092 perturbations in K562 cell line, and 1543 perturbations in RPE1. Predictions on every held-out cross-validation fold are made independently to avoid any data leakage. In experiments where we consider a gradually increasing number of training perturbations (e.g. along x-axis in Figure 2), for every cross-validation split we repeatedly downsample the training set.

Following practices from literature, we quantify perturbation prediction performance relative to control cells, i.e. using Pearson correlation, mean absolute error (MAE), and mean squared error (MSE) on the differences $\Delta_n := \mathbf{y}_n - \mathbf{y}_{\text{control}}$. Following Roohani et al. (2023), we calculate both metrics across the top 20 differentially expressed genes relative to control cells, resulting in gene sets that are specific for every perturbation.

Baselines As discussed in the Background section 2.1, we consider existing methods which have a capability to generalise to unseen single-gene perturbations. Specifically, we consider the graph neural network approach GEARS, fine-tuning a single-cell foundation model scGPT, and two versions of the GP+LLM model (one using NCBI text embeddings as input, the other combining text embeddings and protein sequence embeddings). We also consider a non-control mean baseline that has been shown to be surprisingly effective (Kernfeld et al., 2024; Märtens et al., 2024).

Choice of LLM A crucial component of LangPert is its LLM engine, meaning the choice of LLM can significantly impact its behavior and performance. For all comparisons in Section 4.2, we use Claude 3.5 Sonnet. Later, in Section 4.4, we examine how performance varies across different LLMs, ranging from small open-weight LLMs with 8 billion parameters to frontier models.

4.2 Performance Comparison

Results on K562 cell line Figure 2 shows the performance metrics (MAE and correlation) on the K562 cell line dataset across a varying number of training perturbations. The ordering of existing baselines is aligned with what has been reported in literature: the fine-tuned scGPT is the lowest performing model, followed by GEARS. GEARS outperforms the non-control mean baseline according to the MAE metric, but slightly underperforms in correlation. The GP+LLM models outperform both scGPT and GEARS.

[1]Available in https://github.com/snap-stanford/GEARS

Table 1: **Results on K562 cell line dataset:** Performance comparison of different models at different training data sizes. Values shown as mean ± standard error.

Model	100 training perturbations			850 training perturbatons		
	MAE ↓	MSE ↓	Correlation ↑	MAE ↓	MSE ↓	Correlation ↑
scGPT	0.388 ± 0.016	0.234 ± 0.014	0.430 ± 0.021	0.324 ± 0.008	0.195 ± 0.012	0.488 ± 0.016
GEARS	0.316 ± 0.004	0.164 ± 0.004	0.633 ± 0.005	0.292 ± 0.002	0.147 ± 0.003	0.655 ± 0.007
Non-ctrl mean	0.329 ± 0.006	0.175 ± 0.008	0.665 ± 0.006	0.332 ± 0.006	0.176 ± 0.007	0.671 ± 0.007
GP+LLM (text)	0.312 ± 0.006	0.162 ± 0.008	0.672 ± 0.006	0.268 ± 0.004	0.132 ± 0.006	0.718 ± 0.012
GP+LLM (text+seq)	0.306 ± 0.006	0.158 ± 0.008	0.683 ± 0.006	0.265 ± 0.004	0.130 ± 0.005	0.728 ± 0.012
LangPert	**0.259 ± 0.006**	**0.132 ± 0.005**	**0.689 ± 0.004**	**0.224 ± 0.005**	**0.108 ± 0.005**	**0.731 ± 0.011**

Table 2: **Results on RPE1 cell line dataset:** Performance comparison of different models at different training data sizes. Values shown as mean ± standard error.

Model	50 training perturbations			100 training perturbations		
	MAE ↓	MSE ↓	Correlation ↑	MAE ↓	MSE ↓	Correlation ↑
scGPT	0.452 ± 0.008	0.346 ± 0.016	0.627 ± 0.009	0.451 ± 0.009	0.354 ± 0.006	0.642 ± 0.003
GEARS	0.471 ± 0.013	0.343 ± 0.019	0.670 ± 0.007	0.435 ± 0.022	0.307 ± 0.026	0.706 ± 0.005
Non-ctrl mean	0.427 ± 0.010	0.303 ± 0.015	**0.737 ± 0.006**	0.430 ± 0.006	0.305 ± 0.013	**0.738 ± 0.005**
GP+LLM (text)	0.414 ± 0.011	0.289 ± 0.014	0.721 ± 0.002	0.408 ± 0.005	0.282 ± 0.009	0.725 ± 0.002
GP+LLM (text+seq)	0.417 ± 0.013	0.294 ± 0.016	0.717 ± 0.004	0.403 ± 0.005	0.277 ± 0.009	0.723 ± 0.004
LangPert	**0.368 ± 0.008**	**0.249 ± 0.012**	0.726 ± 0.005	**0.361 ± 0.008**	**0.239 ± 0.013**	**0.737 ± 0.008**
	250 training perturbations			**1170 training perturbations**		
	MAE ↓	MSE ↓	Correlation ↑	MAE ↓	MSE ↓	Correlation ↑
scGPT	0.436 ± 0.024	0.329 ± 0.030	0.641 ± 0.009	0.449 ± 0.038	0.349 ± 0.040	0.603 ± 0.040
GEARS	0.430 ± 0.011	0.291 ± 0.015	0.720 ± 0.007	0.405 ± 0.015	0.266 ± 0.015	0.710 ± 0.011
Non-ctrl mean	0.432 ± 0.005	0.306 ± 0.011	0.741 ± 0.004	0.434 ± 0.005	0.308 ± 0.011	0.743 ± 0.005
GP+LLM (text)	0.400 ± 0.004	0.271 ± 0.010	0.735 ± 0.003	0.371 ± 0.004	0.238 ± 0.009	0.759 ± 0.005
GP+LLM (text+seq)	0.397 ± 0.003	0.269 ± 0.009	0.732 ± 0.003	0.364 ± 0.004	0.233 ± 0.009	0.760 ± 0.005
LangPert	**0.344 ± 0.005**	**0.218 ± 0.010**	**0.753 ± 0.009**	**0.318 ± 0.004**	**0.192 ± 0.006**	**0.772 ± 0.005**

Our proposed LangPert significantly outperforms all existing models according to the MAE and MSE metrics (see Table 1 for numerical values), and also achieves a slightly higher correlation value. For example, in the scenario with 850 training perturbations, LangPert achieves MAE of $0.224(\pm 0.005)$ which is a substantial improvement over the second best GP+LLM's $0.265(\pm 0.004)$.

Results on RPE1 cell line We conducted a similar experiment on the RPE1 cell line, with results summarized in Table 2. The ranking of methods remains consistent with previous benchmarks, with LangPert achieving the best results in MAE and MSE metrics. For the correlation metric, at the smallest sample size (50 training perturbations), the non-control mean achieves the highest correlation (0.737 ± 0.006), slightly surpassing LangPert (0.726 ± 0.005). However, as the sample size increases, LangPert outperforms all baselines. At the largest sample size (1170 perturbations), LangPert achieves an MAE of 0.318 ± 0.004 (compared to the second-best LLM+GP at 0.364 ± 0.004) and a correlation of 0.772 ± 0.005 (vs. LLM+GP's 0.760 ± 0.005).

Overall, LangPert sets a new state-of-the-art performance on both the K562 and RPE1 benchmarks.

4.3 LangPert's Context-Dependent Reasoning

To understand the source of LangPert's performance gains, we analysed the model's reasoning traces to examine how it selects relevant genes for each perturbation. A key challenge in kNN-based approaches is determining which similarity criterion to use: genes can be related through shared pathways, protein complexes, subcellular localization, or functional processes. Our analysis revealed that rather than applying a single similarity metric uniformly, LangPert adaptively identifies the most relevant organising principle for each gene based on its specific biological context. The model selects different organising principles based on each gene's functional role:

- **MTOR → Signaling pathway hierarchy**: selects RPTOR (direct complex partner), EIF4E/EIF4G1 (downstream effectors), PPP2R1A (feedback regulator)
- **EIF3E → Protein complex membership**: selects other eIF3 subunits (EIF3A, EIF3B, EIF3I, EIF3F) within the translation initiation complex

- **PSMD11 → Functional machinery**: selects other proteasome subunits (PSMD1, PSMC4, PSMC6, PSMD3, PSMA3) required for protein degradation

This context-dependent reasoning allows LangPert to capture synergistic interactions within biological systems by adaptively choosing the right biological lens for each gene—pathway hierarchies for regulators, protein complexes for structural components, or functional machinery for metabolic genes. This contrasts with static database methods like GEARS, which apply Gene Ontology uniformly across genes, rather than selecting the most relevant similarity framework for each perturbation.

4.4 Comparative Performance of LLM Variants

The LLM component in the LangPert framework has two main elements: the choice of the LLM itself and the prompting strategy. In this section, we evaluate how different LLM variants in the LangPert framework affect its performance.

We compare a range of models, from small open LLMs such as Llama 3.1 8b (Grattafiori et al., 2024) and TxGemma-9b (a Gemma variant pre-trained by its creators for therapeutic applications) (Wang et al., 2025), to a medium-sized Llama 3.3 70b model and the larger DeepSeek R1 model (DeepSeek-AI et al., 2025). We also evaluate frontier models Claude Sonnet 3.5 and OpenAI o1.

The LLM comparison is illustrated in Figure 3, where the blue shaded area represents the performance coverage across the LLM variants. Here we show two metrics, correlation and MAE, with the non-control mean baseline dividing the plot into four quadrants. Methods that outperform the mean in both metrics appear in the green area. Existing non-LLM baselines are shown in red (scGPT, GEARS, GP+LLM). While all methods outperform the mean baseline in MAE, not all of them do in correlation.

Among LLM variants, the smaller TxGemma-9b and Llama 8b models show the lowest performance (achieving MAE values of 0.315 and 0.284, respectively), yet still outperform scGPT. Next in the performance hierarchy, LangPert using either DeepSeek-R1 or Llama 3.3 70b models (with MAE values of 0.254 and 0.247, respectively) outperforms GEARS. Finally, the frontier models Claude

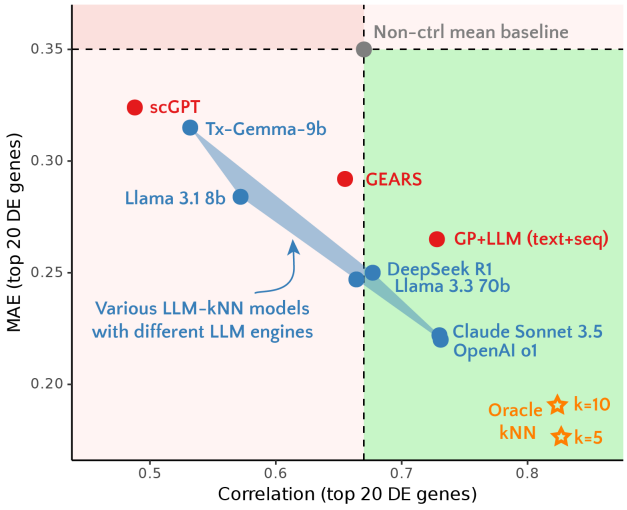

Figure 3: Performance comparison of various versions of LangPert across different LLM backends (blue shaded area) to existing methods (in red), according to two metrics, correlation (x-axis) and MAE (y-axis) on the full K562 cell line dataset. The non-control mean baseline divides the area into four quadrants: Ideally, strong performance lies in the green shaded area (i.e. a smaller MAE and higher correlation). Among all LLMs, the small open models (Llama 8b and TxGemma-9b) are the least performant, and frontier models (Claude Sonnet 3.5 and OpenAI o1) are best performing. For reference, we have added an Oracle kNN method (in orange, for $k = 5$ and $k = 10$) to indicate an approximate upper bound on the performance of a kNN-based approach which has access to the ground truth.

Claude Sonnet 3.5 output for query gene SMG5:

Reasoning: SMG5 is a critical component of the nonsense-mediated mRNA decay (NMD) pathway, which degrades mRNAs containing premature termination codons. In analyzing potential functional similarities, I focused on several key aspects:

1. Core NMD machinery components and regulators that would disrupt similar RNA surveillance mechanisms
2. Genes involved in mRNA quality control and processing that functionally overlap with NMD
3. Translation termination factors and related proteins, as SMG5 interacts with release factors
4. Proteins involved in mRNA degradation pathways

In K562 cells specifically: (1) The NMD pathway is particularly important due to high transcriptional activity in these cancer cells, (2) Many aberrant transcripts need surveillance due to genomic instability, (3) Translation regulation is critical for K562 survival. Key predicted similar phenotypes would come from disrupting: (1) Core NMD factors, (2) General mRNA decay machinery, (3) Translation termination components, (4) Related quality control pathways.

Inferred kNN gene set: [UPF2, UPF1, DIS3, EXOSC10, RBM8A, RNPS1]

OpenAI o1 output for query gene SMG5:

Reasoning: SMG5 encodes a core component of the nonsense-mediated mRNA decay (NMD) pathway, where it helps recruit phosphatase activity to UPF1, leading to the degradation of aberrant transcripts. In K562 cells (a leukemic line with high transcription/translation activity), knocking out other critical NMD factors or key Exon Junction Complex (EJC) components should yield phenotypes most similar to SMG5 loss, because these are all essential for detecting and degrading faulty mRNAs. First among these is UPF1, the central NMD helicase that physically interacts with SMG5. UPF2 is another core NMD scaffold protein downstream of UPF1. RBM8A (Y14) and MAGOH are core EJC members required for coupling exon junction marking to transcript decay. RNPS1 also associates with EJC/NMD surveillance. Finally, DDX6, while more broadly involved in mRNA decapping and turnover, is still closely tied to mRNA quality-control outcomes and thus expected to share certain knockout effects with SMG5 in K562.

Inferred kNN gene set: [UPF1, UPF2, RBM8A, MAGOH, RNPS1, DDX6]

Figure 4: Comparison of LLM-based inference for identifying genes with similar perturbation outcomes to SMG5 in K562 cells. Both Claude and OpenAI o1 identified key nonsense-mediated mRNA decay (NMD) pathway components and associated factors.

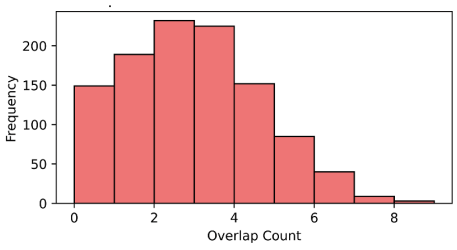

Figure 5: Overlap in inferred gene sets between the two frontier models: Claude Sonnet 3.5 and OpenAI o1.

Sonnet 3.5 and OpenAI o1 outperform all other approaches. Interestingly, both frontier models achieve near-identical performance on this task.

A natural question is whether future improvements in LLM capabilities could lead to further performance gains in our framework. To establish an approximate upper bound on what LLM-kNN approaches can achieve, we implemented an Oracle kNN model that has access to the ground truth (unseen) perturbation outcome. We added the performance of this Oracle model with $k = 5$ and $k = 10$ to Figure 3 as a reference point ("Oracle kNN" in orange).

Given the near-identical performance of the two frontier models, we investigated whether their inferred kNN gene sets were also identical. Figure 5 shows the overlap distribution between the two models, revealing considerable variation in gene set selection despite resulting in similarly predictive outcomes. This suggests that different models identify distinct but equally relevant gene relationships, potentially corresponding to alternative biological hypotheses with similar predictive power.

In Figure 4, we illustrate this phenomenon with outputs for a selected query perturbation, SMG5, comparing the two frontier models. Both Claude Sonnet 3.5 and OpenAI o1 successfully identified key nonsense-mediated mRNA decay (NMD) pathway components and associated factors. Consistent with our overlap analysis, they inferred partially overlapping but not identical gene sets. These observations suggest promising directions for future work, including exploring ensembling strategies or methods to integrate insights from multiple LLMs to enhance both interpretability and robustness.

5 Discussion

We introduce LangPert, a novel hybrid framework that leverages Large Language Models to guide k-nearest neighbour predictions for unseen genetic perturbations. LangPert achieves state-of-the-art performance across multiple datasets and metrics, particularly as measured by mean absolute error of predictions across the entire spectrum of low-to-high data regimes. Our analysis reveals that this performance stems from LangPert's ability to adaptively select different biological similarity frameworks based on each gene's functional context.

LangPert's key innovation lies in its ability to harness LLMs' biological reasoning capabilities without being constrained by their numerical limitations. By using LLMs to identify relevant training examples and employing kNN for aggregation, our approach effectively bridges the gap between knowledge-driven and data-driven methodologies. This hybrid strategy outperforms existing methods such as GEARS, scGPT, and GP+LLM.

The superior performance of LangPert has significant implications for perturbation biology. Improved predictive models could dramatically reduce the need for exhaustive experimental testing, accelerating biological discovery and potentially informing more efficient experimental design strategies. Furthermore, LangPert's flexible framework allows for the incorporation of different LLMs, prompting strategies, and aggregation methods, suggesting ample room for further optimization and adaptation to various biological contexts.

Despite its promising results, LangPert is not without limitations. Its performance depends on the quality and up-to-date nature of the LLM's knowledge, and potential biases in LLM training data could influence predictions. Future work should explore more sophisticated aggregation methods beyond averaging, incorporate uncertainty quantification, and extend the approach to multi-gene perturbations or other types of biological interventions. Additionally, investigating the impact of different LLMs as well as prompting strategies could further enhance the model's capabilities.

The success of LangPert in integrating LLMs with traditional machine learning techniques for high-dimensional biological data suggests potential applications beyond perturbation biology. This approach could be adapted to other scientific domains characterized by high-dimensional outcomes and rich contextual knowledge, such as multi-objective molecular property prediction.

In conclusion, LangPert represents a significant advance in our ability to predict cellular responses to unseen genetic perturbations. By effectively combining the strengths of LLMs and traditional machine learning approaches, it opens new avenues for accelerating biological discovery and deepening our understanding of complex cellular systems. As we continue to refine and expand this approach, we anticipate its impact to grow, potentially transforming how we approach predictive modeling in biology and beyond.

REFERENCES

C. Ahlmann-Eltze, W. Huber, and S. Anders. Deep learning-based predictions of gene perturbation effects do not yet outperform simple linear baselines, Feb. 2025. URL https://www.biorxiv.org/content/10.1101/2024.09.16.613342v5. Pages: 2024.09.16.613342 Section: New Results.

M. Bereket and T. Karaletsos. Modelling Cellular Perturbations with the Sparse Additive Mechanism Shift Variational Autoencoder. Nov. 2023. URL https://openreview.net/forum?id=DzaCE00jGV.

R. Boiarsky, N. M. Singh, A. Buendia, A. P. Amini, G. Getz, and D. Sontag. Deeper evaluation of a single-cell foundation model. *Nature Machine Intelligence*, 6(12):1443–1446, Dec. 2024. ISSN 2522-5839. doi: 10.1038/s42256-024-00949-w. URL https://www.nature.com/articles/s42256-024-00949-w. Publisher: Nature Publishing Group.

T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language Models are Few-Shot Learners, July 2020. URL http://arxiv.org/abs/2005.14165. arXiv:2005.14165 [cs].

Y. Chen and J. Zou. Simple and effective embedding model for single-cell biology built from chatgpt. *Nature Biomedical Engineering*, pages 1–11, 2024.

H. Cui, C. Wang, H. Maan, K. Pang, F. Luo, N. Duan, and B. Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, pages 1–11, 2024.

DeepSeek-AI, D. Guo, D. Yang, H. Zhang, J. Song, A. Liu, B. Wang, B. Wu, B. Feng, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, D. Dai, D. Chen, D. Ji, E. Li, F. Lin, F. Dai, F. Luo, G. Hao, G. Li, H. Zhang, H. Bao, H. Xu, H. Wang, H. Ding, H. Xin, H. Gao, H. Qu, H. Li, J. Guo, J. Li, J. Wang, J. Chen, J. Yuan, J. Qiu, J. Li, J. L. Cai, J. Ni, J. Liang, J. Chen, K. Dong, K. Hu, K. Gao, K. Guan, K. Huang, K. Yu, L. Wang, L. Zhang, L. Zhao, L. Wang, L. Zhang, L. Xu, L. Xia, M. Zhang, M. Zhang, M. Tang, M. Li, M. Wang, M. Li, N. Tian, P. Huang, P. Zhang, Q. Wang, Q. Chen, Q. Du, R. Ge, R. Zhang, R. Pan, R. Wang, R. J. Chen, R. L. Jin, R. Chen, S. Lu, S. Zhou, S. Chen, S. Ye, S. Wang, S. Yu, S. Zhou, S. Pan, S. S. Li, S. Zhou, S. Wu, S. Ye, T. Yun, T. Pei, T. Sun, T. Wang, W. Zeng, W. Zhao, W. Liu, W. Liang, W. Gao, W. Yu, W. Zhang, W. L. Xiao, W. An, X. Liu, X. Wang, X. Chen, X. Nie, X. Cheng, X. Liu, X. Xie, X. Liu, X. Yang, X. Li, X. Su, X. Lin, X. Q. Li, X. Jin, X. Shen, X. Chen, X. Sun, X. Wang, X. Song, X. Zhou, X. Wang, X. Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. Zhang, Y. Xu, Y. Li, Y. Zhao, Y. Sun, Y. Wang, Y. Yu, Y. Zhang, Y. Shi, Y. Xiong, Y. He, Y. Piao, Y. Wang, Y. Tan, Y. Ma, Y. Liu, Y. Guo, Y. Ou, Y. Wang, Y. Gong, Y. Zou, Y. He, Y. Xiong, Y. Luo, Y. You, Y. Liu, Y. Zhou, Y. X. Zhu, Y. Xu, Y. Huang, Y. Li, Y. Zheng, Y. Zhu, Y. Ma, Y. Tang, Y. Zha, Y. Yan, Z. Z. Ren, Z. Ren, Z. Sha, Z. Fu, Z. Xu, Z. Xie, Z. Zhang, Z. Hao, Z. Ma, Z. Yan, Z. Wu, Z. Gu, Z. Zhu, Z. Liu, Z. Li, Z. Xie, Z. Song, Z. Pan, Z. Huang, Z. Xu, Z. Zhang, and Z. Zhang. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, Jan. 2025. URL http://arxiv.org/abs/2501.12948. arXiv:2501.12948 [cs].

A. Gambardella, Y. Iwasawa, and Y. Matsuo. Language Models Do Hard Arithmetic Tasks Easily and Hardly Do Easy Arithmetic Tasks, June 2024. URL http://arxiv.org/abs/2406.02356. arXiv:2406.02356 [cs].

S. Gao, A. Fang, Y. Huang, V. Giunchiglia, A. Noori, J. R. Schwarz, Y. Ektefaie, J. Kondic, and M. Zitnik. Empowering biomedical discovery with AI agents. *Cell*, 187(22):6125–6151, Oct. 2024. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2024.09.022. URL https://www.cell.com/cell/abstract/S0092-8674(24)01070-5. Publisher: Elsevier.

A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Yang, A. Mitra, A. Sravankumar, A. Korenev, A. Hinsvark, A. Rao, A. Zhang, A. Rodriguez, A. Gregerson, A. Spataru,

B. Roziere, B. Biron, B. Tang, B. Chern, C. Caucheteux, C. Nayak, C. Bi, C. Marra, C. McConnell, C. Keller, C. Touret, C. Wu, C. Wong, C. C. Ferrer, C. Nikolaidis, D. Allonsius, D. Song, D. Pintz, D. Livshits, D. Wyatt, D. Esiobu, D. Choudhary, D. Mahajan, D. Garcia-Olano, D. Perino, D. Hupkes, E. Lakomkin, E. AlBadawy, E. Lobanova, E. Dinan, E. M. Smith, F. Radenovic, F. Guzmán, F. Zhang, G. Synnaeve, G. Lee, G. L. Anderson, G. Thattai, G. Nail, G. Mialon, G. Pang, G. Cucurell, H. Nguyen, H. Korevaar, H. Xu, H. Touvron, I. Zarov, I. A. Ibarra, I. Kloumann, I. Misra, I. Evtimov, J. Zhang, J. Copet, J. Lee, J. Geffert, J. Vranes, J. Park, J. Mahadeokar, J. Shah, J. v. d. Linde, J. Billock, J. Hong, J. Lee, J. Fu, J. Chi, J. Huang, J. Liu, J. Wang, J. Yu, J. Bitton, J. Spisak, J. Park, J. Rocca, J. Johnstun, J. Saxe, J. Jia, K. V. Alwala, K. Prasad, K. Upasani, K. Plawiak, K. Li, K. Heafield, K. Stone, K. El-Arini, K. Iyer, K. Malik, K. Chiu, K. Bhalla, K. Lakhotia, L. Rantala-Yeary, L. v. d. Maaten, L. Chen, L. Tan, L. Jenkins, L. Martin, L. Madaan, L. Malo, L. Blecher, L. Landzaat, L. d. Oliveira, M. Muzzi, M. Pasupuleti, M. Singh, M. Paluri, M. Kardas, M. Tsimpoukelli, M. Oldham, M. Rita, M. Pavlova, M. Kambadur, M. Lewis, M. Si, M. K. Singh, M. Hassan, N. Goyal, N. Torabi, N. Bashlykov, N. Bogoychev, N. Chatterji, N. Zhang, O. Duchenne, O. Çelebi, P. Alrassy, P. Zhang, P. Li, P. Vasic, P. Weng, P. Bhargava, P. Dubal, P. Krishnan, P. S. Koura, P. Xu, Q. He, Q. Dong, R. Srinivasan, R. Ganapathy, R. Calderer, R. S. Cabral, R. Stojnic, R. Raileanu, R. Maheswari, R. Girdhar, R. Patel, R. Sauvestre, R. Polidoro, R. Sumbaly, R. Taylor, R. Silva, R. Hou, R. Wang, S. Hosseini, S. Chennabasappa, S. Singh, S. Bell, S. S. Kim, S. Edunov, S. Nie, S. Narang, S. Raparthy, S. Shen, S. Wan, S. Bhosale, S. Zhang, S. Vandenhende, S. Batra, S. Whitman, S. Sootla, S. Collot, S. Gururangan, S. Borodinsky, T. Herman, T. Fowler, T. Sheasha, T. Georgiou, T. Scialom, T. Speckbacher, T. Mihaylov, T. Xiao, U. Karn, V. Goswami, V. Gupta, V. Ramanathan, V. Kerkez, V. Gonguet, V. Do, V. Vogeti, V. Albiero, V. Petrovic, W. Chu, W. Xiong, W. Fu, W. Meers, X. Martinet, X. Wang, X. Wang, X. E. Tan, X. Xia, X. Xie, X. Jia, X. Wang, Y. Goldschlag, Y. Gaur, Y. Babaei, Y. Wen, Y. Song, Y. Zhang, Y. Li, Y. Mao, Z. D. Coudert, Z. Yan, Z. Chen, Z. Papakipos, A. Singh, A. Srivastava, A. Jain, A. Kelsey, A. Shajnfeld, A. Gangidi, A. Victoria, A. Goldstand, A. Menon, A. Sharma, A. Boesenberg, A. Baevski, A. Feinstein, A. Kallet, A. Sangani, A. Teo, A. Yunus, A. Lupu, A. Alvarado, A. Caples, A. Gu, A. Ho, A. Poulton, A. Ryan, A. Ramchandani, A. Dong, A. Franco, A. Goyal, A. Saraf, A. Chowdhury, A. Gabriel, A. Bharambe, A. Eisenman, A. Yazdan, B. James, B. Maurer, B. Leonhardi, B. Huang, B. Loyd, B. D. Paola, B. Paranjape, B. Liu, B. Wu, B. Ni, B. Hancock, B. Wasti, B. Spence, B. Stojkovic, B. Gamido, B. Montalvo, C. Parker, C. Burton, C. Mejia, C. Liu, C. Wang, C. Kim, C. Zhou, C. Hu, C.-H. Chu, C. Cai, C. Tindal, C. Feichtenhofer, C. Gao, D. Civin, D. Beaty, D. Kreymer, D. Li, D. Adkins, D. Xu, D. Testuggine, D. David, D. Parikh, D. Liskovich, D. Foss, D. Wang, D. Le, D. Holland, E. Dowling, E. Jamil, E. Montgomery, E. Presani, E. Hahn, E. Wood, E.-T. Le, E. Brinkman, E. Arcaute, E. Dunbar, E. Smothers, F. Sun, F. Kreuk, F. Tian, F. Kokkinos, F. Ozgenel, F. Caggioni, F. Kanayet, F. Seide, G. M. Florez, G. Schwarz, G. Badeer, G. Swee, G. Halpern, G. Herman, G. Sizov, Guangyi, Zhang, G. Lakshminarayanan, H. Inan, H. Shojanazeri, H. Zou, H. Wang, H. Zha, H. Habeeb, H. Rudolph, H. Suk, H. Aspegren, H. Goldman, H. Zhan, I. Damlaj, I. Molybog, I. Tufanov, I. Leontiadis, I.-E. Veliche, I. Gat, J. Weissman, J. Geboski, J. Kohli, J. Lam, J. Asher, J.-B. Gaya, J. Marcus, J. Tang, J. Chan, J. Zhen, J. Reizenstein, J. Teboul, J. Zhong, J. Jin, J. Yang, J. Cummings, J. Carvill, J. Shepard, J. McPhie, J. Torres, J. Ginsburg, J. Wang, K. Wu, K. H. U, K. Saxena, K. Khandelwal, K. Zand, K. Matosich, K. Veeraraghavan, K. Michelena, K. Li, K. Jagadeesh, K. Huang, K. Chawla, K. Huang, L. Chen, L. Garg, L. A, L. Silva, L. Bell, L. Zhang, L. Guo, L. Yu, L. Moshkovich, L. Wehrstedt, M. Khabsa, M. Avalani, M. Bhatt, M. Mankus, M. Hasson, M. Lennie, M. Reso, M. Groshev, M. Naumov, M. Lathi, M. Keneally, M. Liu, M. L. Seltzer, M. Valko, M. Restrepo, M. Patel, M. Vyatskov, M. Samvelyan, M. Clark, M. Macey, M. Wang, M. J. Hermoso, M. Metanat, M. Rastegari, M. Bansal, N. Santhanam, N. Parks, N. White, N. Bawa, N. Singhal, N. Egebo, N. Usunier, N. Mehta, N. P. Laptev, N. Dong, N. Cheng, O. Chernoguz, O. Hart, O. Salpekar, O. Kalinli, P. Kent, P. Parekh, P. Saab, P. Balaji, P. Rittner, P. Bontrager, P. Roux, P. Dollar, P. Zvyagina, P. Ratanchandani, P. Yuvraj, Q. Liang, R. Alao, R. Rodriguez, R. Ayub, R. Murthy, R. Nayani, R. Mitra, R. Parthasarathy, R. Li, R. Hogan, R. Battey, R. Wang, R. Howes, R. Rinott, S. Mehta, S. Siby, S. J. Bondu, S. Datta, S. Chugh, S. Hunt, S. Dhillon, S. Sidorov, S. Pan, S. Mahajan, S. Verma, S. Yamamoto, S. Ramaswamy, S. Lindsay, S. Lindsay, S. Feng, S. Lin, S. C. Zha, S. Patil, S. Shankar, S. Zhang, S. Zhang, S. Wang, S. Agarwal, S. Sajuyigbe, S. Chintala, S. Max, S. Chen, S. Kehoe, S. Satterfield, S. Govindaprasad, S. Gupta, S. Deng, S. Cho, S. Virk, S. Subramanian, S. Choudhury, S. Goldman, T. Remez, T. Glaser, T. Best, T. Koehler, T. Robinson, T. Li, T. Zhang, T. Matthews, T. Chou, T. Shaked, V. Vontimitta, V. Ajayi, V. Montanez, V. Mohan, V. S. Kumar, V. Mangla, V. Ionescu, V. Poenaru, V. T. Mihailescu, V. Ivanov, W. Li, W. Wang, W. Jiang, W. Bouaziz, W. Constable,

X. Tang, X. Wu, X. Wang, X. Wu, X. Gao, Y. Kleinman, Y. Chen, Y. Hu, Y. Jia, Y. Qi, Y. Li, Y. Zhang, Y. Zhang, Y. Adi, Y. Nam, Yu, Wang, Y. Zhao, Y. Hao, Y. Qian, Y. Li, Y. He, Z. Rait, Z. DeVito, Z. Rosnbrick, Z. Wen, Z. Yang, Z. Zhao, and Z. Ma. The Llama 3 Herd of Models, Nov. 2024. URL http://arxiv.org/abs/2407.21783. arXiv:2407.21783 [cs].

D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

M. Hao, J. Gong, X. Zeng, C. Liu, Y. Guo, X. Cheng, T. Wang, J. Ma, X. Zhang, and L. Song. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, pages 1–11, 2024. URL https://idp.nature.com/authorize/casa?redirect_uri=https://www.nature.com/articles/s41592-024-02305-7&casa_token=pRbl5O0caJkAAAAA:FZirFybc1gGR1gHqgRiUyMsPm_-3B8fMttHLux3Lc7AjTWfp_MSZUTYOfd4TsubiDWqwvphBXOUuBdKa3Q. Publisher: Nature Publishing Group US New York.

M. Hu, S. Alkhairy, I. Lee, R. T. Pillich, D. Fong, K. Smith, R. Bachelder, T. Ideker, and D. Pratt. Evaluation of large language models for discovery of gene set function. *Nature Methods*, 22 (1):82–91, Jan. 2025. ISSN 1548-7105. doi: 10.1038/s41592-024-02525-x. URL https://www.nature.com/articles/s41592-024-02525-x. Publisher: Nature Publishing Group.

S. Johnson and D. Hyland-Wood. A Primer on Large Language Models and their Limitations, Dec. 2024. URL http://arxiv.org/abs/2412.04503. arXiv:2412.04503 [cs].

K. Z. Kedzierska, L. Crawford, A. P. Amini, and A. X. Lu. Assessing the limits of zero-shot foundation models in single-cell biology, Nov. 2023. URL https://www.biorxiv.org/content/10.1101/2023.10.16.561085v2. Pages: 2023.10.16.561085 Section: New Results.

E. Kernfeld, Y. Yang, J. S. Weinstock, A. Battle, and P. Cahan. A systematic comparison of computational methods for expression forecasting, Oct. 2024. URL https://www.biorxiv.org/content/10.1101/2023.07.28.551039v2. Pages: 2023.07.28.551039 Section: New Results.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

R. Lopez, N. Tagasovska, S. Ra, K. Cho, J. Pritchard, and A. Regev. Learning Causal Representations of Single Cells via Sparse Mechanism Shift Modeling. Mar. 2023. URL https://openreview.net/forum?id=IOWJsPJ2xGd.

M. Lotfollahi, A. Klimovskaia Susmelj, C. De Donno, L. Hetzel, Y. Ji, I. L. Ibarra, S. R. Srivatsan, M. Naghipourfar, R. M. Daza, B. Martin, J. Shendure, J. L. McFaline-Figueroa, P. Boyeau, F. A. Wolf, N. Yakubova, S. Günnemann, C. Trapnell, D. Lopez-Paz, and F. J. Theis. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, 19(6):e11517, June 2023. ISSN 1744-4292. doi: 10.15252/msb.202211517. URL https://www.embopress.org/doi/full/10.15252/msb.202211517. Publisher: John Wiley & Sons, Ltd.

K. Märtens, R. Donovan-Maiye, and J. Ferkinghoff-Borg. Enhancing generative perturbation models with llm-informed gene embeddings. In *ICLR 2024 Workshop on Machine Learning for Genomics Explorations*, 2024.

J. M. Replogle, R. A. Saunders, A. N. Pogson, J. A. Hussmann, A. Lenail, A. Guna, L. Mascibroda, E. J. Wagner, K. Adelman, and G. Lithwick-Yanai. Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell*, 185(14):2559–2575, 2022. URL https://www.cell.com/cell/pdf/S0092-8674(22)00597-9.pdf. Publisher: Elsevier.

D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

Y. Roohani, K. Huang, and J. Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nature Biotechnology*, pages 1–9, Aug. 2023. ISSN 1546-1696. doi: 10.1038/s41587-023-01905-6. URL https://www.nature.com/articles/s41587-023-01905-6. Publisher: Nature Publishing Group.

C. V. Theodoris, L. Xiao, A. Chopra, M. D. Chaffin, Z. R. Al Sayed, M. C. Hill, H. Mantineo, E. M. Brydon, Z. Zeng, and X. S. Liu. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023. URL https://idp.nature.com/authorize/casa?redirect_uri=https://www.nature.com/articles/s41586-023-06139-9&casa_token=T6Kd54XSnhwAAAAA:QHLXOg37bjq_ki4B_rC-zG033wKzhCnWlP5_RI1kq00de9-AGVMDZWj4-_4KC8mKZyhkDkIs8z8Jn1h-Bg. Publisher: Nature Publishing Group UK London.

E. Wang, S. Schmidgall, P. F. Jaeger, F. Zhang, R. Pilgrim, Y. Matias, J. Barral, D. Fleet, and S. Azizi. TxGemma: Efficient and Agentic LLMs for Therapeutics, Apr. 2025. URL http://arxiv.org/abs/2504.06196. arXiv:2504.06196 [cs].

D. R. Wong, A. S. Hill, and R. Moccia. Simple controls exceed best deep learning algorithms and reveal foundation model effectiveness for predicting genetic perturbations, Jan. 2025. URL https://www.biorxiv.org/content/10.1101/2025.01.06.631555v2. Pages: 2025.01.06.631555 Section: New Results.

M. Wu, R. Littman, J. Levine, L. Qiu, T. Biancalani, D. Richmond, and J.-C. Huetter. Contextualizing biological perturbation experiments through language. Oct. 2024. URL https://openreview.net/forum?id=5WEpbilssv.

SUPPLEMENTARY MATERIAL

A PROMPTS

The prompts below aim to identify approximately 5-10 genes from a provided list that closely resemble the specified gene based on shared involvement in specific biological pathways, co-regulation, or protein-protein interactions in the context of a given cell line (K562 or RPE1).

Prompt for K562 cell line

```
Instruction: Analyze the gene {gene} and identify 5-10 most similar
genes from the provided list. Rank them by similarity (most similar
first).
Consider similarity based on:
• Shared biological pathways and functions
• Protein-protein interactions and complex formation
• Co-regulation patterns and co-expression relationships
• Similar effects when knocked out (predicted knockout outcomes)
Context: Analysis should focus on the K562 cell line (chronic myeloid
leukemia model). Consider cancer-relevant pathways including ribosome
biogenesis, transcriptional regulation, mitochondrial function, stress
responses, and known genetic dependencies specific to K562 survival.
Available genes: {list_of_genes}
Format your response as JSON with two parts:
1. "reasoning": Explain your analysis, discussing potential
connections between {gene} and relevant genes. Include both general
biological knowledge and cell-specific considerations where relevant.
2. "kNN": List the most similar genes in order of similarity (5-10
genes)
Example response format:
{ "reasoning": "Gene X is involved in pathway Y which
directly interacts with gene Z...", "kNN": ["MostSimilarGene",
"SecondMostSimilar", "ThirdMostSimilar"] }
Note: Prioritize confidence over hitting exactly 5-10 genes. Provide
ONLY ONE JSON response.
```

Prompt for RPE1 cell line

```
Instruction: Analyze the gene {gene} and identify 5-10 most similar
genes from the provided list. Rank them by similarity (most similar
first).
Consider similarity based on:
• Shared biological pathways and functions
• Protein-protein interactions and complex formation
• Co-regulation patterns and co-expression relationships
• Similar effects when knocked out (predicted knockout outcomes)
Context: Analysis should focus on the RPE1 cell line (near-diploid,
non-transformed human retinal pigment epithelial cells). Consider
pathways relevant to RPE1 biology including epithelial polarity, cilia
signaling, cell cycle control, DNA repair, oxidative stress response,
and retinal metabolism.
Available genes: {list_of_genes}
Format your response as JSON with two parts: 1. "reasoning": Explain
your analysis, discussing potential connections between {gene} and
relevant genes. Include both general biological knowledge and
RPE1-specific considerations where relevant. 2. "kNN": List the most
similar genes in order of similarity (5-10 genes)
Example response format: { "reasoning": "Gene X is involved in pathway
Y which directly interacts with gene Z...", "kNN": ["MostSimilarGene",
"SecondMostSimilar", "ThirdMostSimilar"] }
Note: Prioritize confidence over hitting exactly 5-10 genes. Provide
ONLY ONE JSON response.
```