LANGPERT: LLM-DRIVEN CONTEXTUAL SYNTHESIS FOR UNSEEN PERTURBATION PREDICTION

Kaspar Märtens, Marc Boubnovski Martell, Cesar A. Prada-Medina & Rory Donovan-Maiye
AI & Digital Innovation, Novo Nordisk
{KQTM,MBVK,CPDM,RZDM}@novonordisk.com

ABSTRACT

Systematic genetic perturbation provides critical insights into cell functioning, yet predicting their cellular effects remains a major challenge. Despite advances in computational approaches, accurately modelling cellular responses to unseen 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

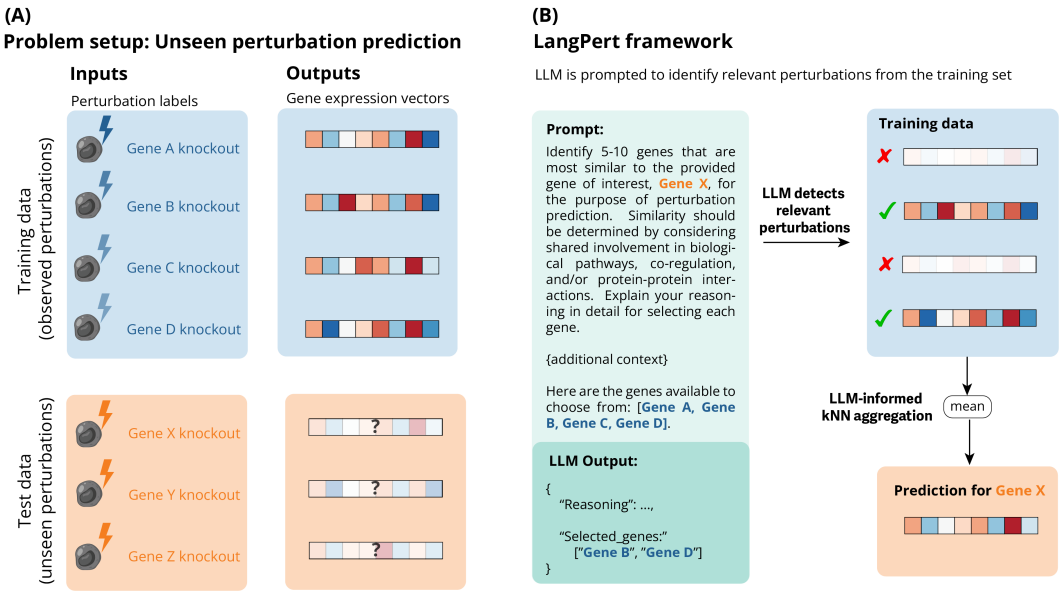

Figure 1: **(A) The task of unseen perturbation outcome prediction illustrated.** The training set consists of pairs $\{(\mathbf{x}_n, \mathbf{y}_n)\}$, where inputs $\mathbf{x}_n$ are discrete perturbation labels and outputs $\mathbf{y}_n$ are high-dimensional gene expression vectors. At test time, 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.

tasks (Guo et al., 2025; Gao et al., 2024). Their ability to synthesize knowledge from vast scientific corpora and 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 & 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 & 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 & 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 & 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 & 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.

While LLMs can extract meaningful biological relationships from text, their effectiveness in high-dimensional perturbation modeling is constrained by tokenization limits and numerical precision issues. These limitations motivate our hybrid approach, which integrates LLM-driven biological reasoning with computational models capable of 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 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

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

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 $\Delta_n :=_n -_{\text{control}}$. 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 v2. Later, in Section 4.3, we examine how performance varies across different LLMs and explore a more advanced prompting strategy incorporating a self-refinement step.

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.

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 cell line benchmarks.

4.3 Impact of LLM Design Choices

The LLM component in LangPert framework has two main elements: the choice of the LLM itself and the prompting strategy. In this section, we evaluate how different LLM architectures affect performance and examine the role of a refinement-based prompting approach in improving the selection of relevant gene perturbations.

To assess the significance of LLM selection, we conducted a comparative analysis across three frontier models: OpenAI o1, OpenAI o3-mini, and Claude 3.5 Sonnet v2. Our evaluation considered both overall predictive performance and the consistency of selected genes across models.

First, examining overall performance (Column "Single-pass" in Table 3), we observed robust performance across all models. All evaluated LLMs outperformed the best GP+LLM baseline in both

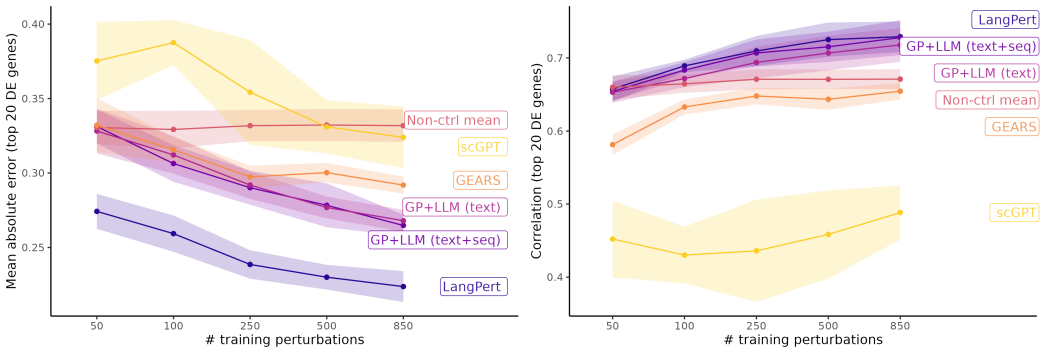

Figure 2: Performance comparison across models (scGPT, GEARS, non-control mean baseline, GP+LLM, and LangPert) evaluated using mean absolute error (MAE, lower is better) and Pearson correlation between predicted and observed differences from control cells (higher is better). Shaded regions indicate ±1.96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

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

	100 training perturbations			**850 training perturbatons**		
Model	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.

	50 training perturbations			**100 training perturbations**		
Model	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**

MAE and MSE metrics, with only the o3-mini model showing slightly lower performance on the correlation metric relative to the GP+LLM baseline. This consistent strong performance across different models suggests that the LangPert framework is reasonably robust to model choice.

When analyzing the overlap in model responses, we found considerable variation in gene selection across different LLMs, yet they produced similarly predictive gene sets. For example, when comparing Claude 3.5 Sonnet v2 and OpenAI o3-mini inferred gene lists, the average overlap is 2.5 genes (Fig 3 left), when an average total number of listed genes is 6.2. When focusing on the overlap across all three considered models, the average overlap is 1.8 genes (Fig 3 right). This variation, despite similar predictive performance, suggests that different models may identify distinct but sim-

Table 3: Performance comparison of language models using two strategies: (1) a single LLM call ("Single-pass") and (2) an additional call for self-critique ("Refinement-based strategy").

	Single pass			Refinement-based strategy		
Model	MAE ↓	MSE ↓	Correlation ↑	MAE ↓	MSE ↓	Correlation ↑
OpenAI o1	**0.220 ± 0.004**	**0.105 ± 0.003**	**0.731 ± 0.007**	0.220 ± 0.004	0.105 ± 0.003	0.731 ± 0.007
Claude 3.5 Sonnet	0.229 ± 0.004	0.111 ± 0.004	**0.731 ± 0.007**	0.229 ± 0.004	0.111 ± 0.004	0.731 ± 0.007
OpenAI o3-mini	0.235 ± 0.004	0.116 ± 0.004	0.704 ± 0.007	**0.215 ± 0.004**	**0.101 ± 0.003**	**0.739 ± 0.007**

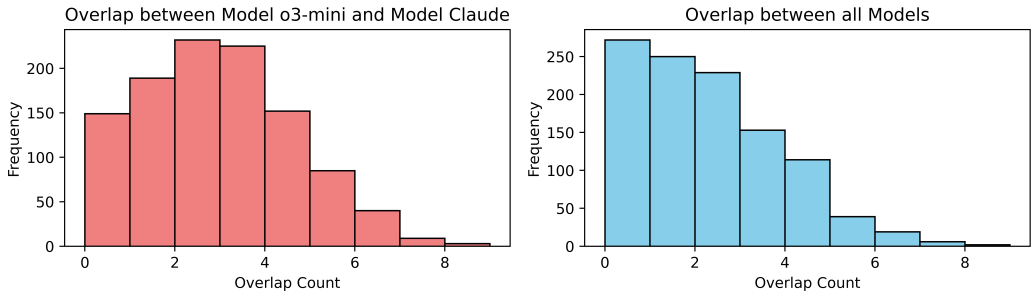

Figure 3: Gene selection overlap across LLM models. (Left): Distribution of overlapping gene counts between o3-mini and Claude 3.5 Sonnet v2. (Right): Distribution of intersecting gene counts across all three models.

Question: What are the most relevant genes for the purpose of predicting SMG5 perturbation?

- **OpenAI o3-mini:**
  - **Single pass:** 'UPF1', 'DDX6', 'PPP4C', 'CNOT3', 'PNN'
  - **Refinement-based:** 'UPF1', 'UPF2', 'SMG7', 'DDX6', 'CNOT3'
  - **Reasoning (Single pass):** I selected UPF1 because it is a core component of the nonsense-mediated decay (NMD) pathway and directly interacts with SMG5. The additional genes, DDX6, PPP4C, CNOT3, and PNN, are involved in RNA metabolism—through mRNA decapping, dephosphorylation, and processing—functions that are closely linked to NMD and are relevant in the context of perturbation prediction in K562 cells.
  - **Reasoning (Refinement-based):** UPF1, UPF2, and SMG7 are core components of the nonsense-mediated decay pathway and share functional relationships with SMG5, making them highly relevant for perturbation prediction in chronic myeloid leukemia. DDX6 and CNOT3 are key regulators in mRNA decay and gene expression, likely resulting in expression changes similar to those induced by SMG5 perturbation.
- **OpenAI o1:**
  - **Single pass:** 'UPF1', 'UPF2', 'RBM8A', 'RNPS1', 'MAGOH'
  - **Refinement-based:** 'UPF1', 'UPF2', 'RBM8A', 'RNPS1', 'MAGOH'
  - **Reasoning (Single pass):** These genes participate in the nonsense-mediated mRNA decay pathway alongside SMG5. Their shared involvement in recognizing and degrading aberrant transcripts suggests similar perturbation effects. Additionally, literature supports interactions among these factors, reinforcing their functional similarity in K562 cells.
  - **Reason (Refinement-based):** These genes are central components of the nonsense-mediated decay (NMD) pathway, closely linked to SMG5's function. Their co-regulation and interactions ensure that perturbations resemble SMG5-related cellular effects, which is crucial for chronic myeloid leukemia research.
- **Claude:**
  - **Single pass:** 'UPF1', 'UPF2', 'RNPS1', 'RBM8A', 'MAGOH', 'SRSF1', 'SRSF2'
  - **Refinement-based:** 'UPF1', 'UPF2', 'MAGOH', 'RBM8A', 'RNPS1'
  - **Reasoning (Single pass):** Selected genes are involved in the nonsense-mediated mRNA decay (NMD) pathway, where SMG5 plays a crucial role. UPF1 and UPF2 are core NMD factors that directly interact with SMG5. RNPS1, RBM8A, and MAGOH are part of the exon junction complex (EJC) which is essential for NMD. SRSF1 and SRSF2 are splicing factors that can influence NMD efficiency and interact with the NMD machinery. These genes are particularly relevant in K562 cells where proper mRNA quality control is essential for cellular homeostasis.
  - **Reason (Refinement-based):** These genes form a core component of the nonsense-mediated mRNA decay (NMD) pathway, which is closely related to SMG5's function. UPF1 and UPF2 are central NMD factors, while MAGOH, RBM8A, and RNPS1 are part of the exon junction complex (EJC) that interacts with the NMD machinery. SRSF1 and SRSF2 were removed as they are more generally involved in splicing regulation rather than specifically in the NMD pathway like SMG5.

Figure 4: Example output for SMG5 gene perturbation across three LLMs, comparing a single LLM call ('Single-pass') with an additional self-critique call ('Refinement-based strategy'). Full prompt examples are provided in Supplementary Material A.

ilarly relevant gene relationships. Future work may explore ensembling strategies or methods to integrate insights from multiple LLMs to enhance interpretability and robustness.

Refinement-based approaches have become widely adopted in LLM applications (Gao et al., 2024), with recent work formalizing and analyzing their effectiveness (Madaan et al., 2024). Building on these insights, we investigated a refinement-based approach in LangPert. This strategy involves a secondary call to the LLM, where the initial gene selection is presented back to the model for critical evaluation and potential refinement. Examples of both single-pass and refinement-based prompting are provided in Supplementary Material A. As shown in Table 3, the impact of refinement varies across models. The o3-mini model achieves optimal results when employing the refinement-based approach, while the o1 model maintains its strong performance with or without refinement. Figure 4 illustrates examples of how the refinement process can lead to modifications in the initially selected gene list, demonstrating the model's ability to reconsider its selections in light of its first response.

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 model demonstrates the potential of combining LLMs' ever-growing biological knowledge with efficient numerical computation to address challenges in predicting unseen genetic perturbations.

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

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

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

Rebecca Boiarsky, Nalini M. Singh, Alejandro Buendia, Ava P. Amini, Gad Getz, and David Sontag. Deeper evaluation of a single-cell foundation model. *Nature Machine Intelligence*, 6(12):1443–1446, December 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.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners, July 2020. URL http://arxiv.org/abs/2005.14165. arXiv:2005.14165 [cs].

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

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, pp. 1–11, 2024.

Andrew Gambardella, Yusuke Iwasawa, and Yutaka 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].

Shanghua Gao, Ada Fang, Yepeng Huang, Valentina Giunchiglia, Ayush Noori, Jonathan Richard Schwarz, Yasha Ektefaie, Jovana Kondic, and Marinka Zitnik. Empowering biomedical discovery with AI agents. *Cell*, 187(22):6125–6151, October 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.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, and Le Song. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, pp. 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.

Mengzhou Hu, Sahar Alkhairy, Ingoo Lee, Rudolf T. Pillich, Dylan Fong, Kevin Smith, Robin Bachelder, Trey Ideker, and Dexter Pratt. Evaluation of large language models for discovery of gene set function. *Nature Methods*, 22(1):82–91, January 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.

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

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

Eric Kernfeld, Yunxiao Yang, Joshua S. Weinstock, Alexis Battle, and Patrick Cahan. A systematic comparison of computational methods for expression forecasting, October 2024. URL https://www.biorxiv.org/content/10.1101/2023.07.28.551039v2. Pages: 2023.07.28.551039 Section: New Results.

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

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

Romain Lopez, Natasa Tagasovska, Stephen Ra, Kyunghyun Cho, Jonathan Pritchard, and Aviv Regev. Learning Causal Representations of Single Cells via Sparse Mechanism Shift Modeling. March 2023. URL https://openreview.net/forum?id=IOWJsPJ2xGd.

Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L Ibarra, Sanjay R Srivatsan, Mohsen Naghipourfar, Riza M Daza, Beth Martin, Jay Shendure, Jose L McFaline-Figueroa, Pierre Boyeau, F Alexander Wolf, Nafissa Yakubova, Stephan Günnemann, Cole Trapnell, David Lopez-Paz, and Fabian 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.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, and Yiming Yang. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91edff07232fb1b55a505a9e9f6c0ff3-Abstract-Conference.html.

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

Joseph M. Replogle, Reuben A. Saunders, Angela N. Pogson, Jeffrey A. Hussmann, Alexander Lenail, Alina Guna, Lauren Mascibroda, Eric J. Wagner, Karen Adelman, and Gila 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.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nature Biotechnology*, pp. 1–9, August 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.

Christina V. Theodoris, Ling Xiao, Anant Chopra, Mark D. Chaffin, Zeina R. Al Sayed, Matthew C. Hill, Helene Mantineo, Elizabeth M. Brydon, Zexian Zeng, and X. Shirley 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.

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

SUPPLEMENTARY MATERIAL

A PROMPTS

This section provides the prompts used in the paper, the bold text in the prompts are variables.

A.1 SINGLE PASS GENE SELECTION

The prompt aims 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 K562 cell line. These selected genes should be pertinent for perturbation prediction, implying that their removal or inactivation would likely lead to similar alterations in gene expression as the specified gene. The task also requires ranking these genes in descending order of similarity, with the most similar gene listed first.

Given a gene of interest **gene**, choose around 5-10 genes from the list that are most similar to gene **gene** based on shared involvement in specific biological pathways, co-regulation, or protein-protein interactions. These genes should be relevant for perturbation prediction, meaning their knockout effect is likely to result in similar changes in gene expression as the knockout of gene **gene**. Rank the genes in order of decreasing similarity, with the most similar gene first. Consider data from relevant databases or literature to assess the similarity between genes.
Focus on the context of the K562 cell line, a model for chronic myeloid leukemia. Consider the role of genes in pathways relevant to cancer biology, including, but not limited to, ribosome biogenesis, transcriptional regulation, mitochondrial function, and stress responses.
Here is the list of genes available to choose from: **gene_train_list**
Provide your response as LIST:
Note: You may choose NO genes if NOT CONFIDENT in the similarity of others. Equally, when there are many genes involved in the same pathway, feel free to include more relevant genes in the list.
OUTPUT JSON FORMAT

A.2 REFINEMENT-BASED GENE SELECTION

The task involves a detailed examination of the generated gene list, emphasizing its suitability for perturbation prediction and its alignment with the specified gene of interest. As an expert, the objective is to evaluate and potentially refine the gene list to guarantee its relevance and suitability within the domain of gene perturbation prediction, especially within the context of the K562 cell line, which serves as a model for chronic myeloid leukemia.

As an expert in gene perturbation prediction for the K562 cell line, a model for chronic myeloid leukemia, your task is to carefully review and, if necessary, alter the following LIST **single_pass_gene_list** based on their relevance to perturbation prediction and similarity to the **gene** gene of interest.
Gene List for Considering: **gene_train_list**
Consider the biological pathways, co-regulation, and protein-protein interactions of each gene. Ensure that the listed genes are highly relevant for perturbation prediction and are likely to result in similar changes in gene expression as the gene of interest when perturbed. You may replace or remove genes as needed to optimize the list for perturbation prediction.
Please make any necessary alterations to the gene list to improve its relevance for perturbation prediction in the context of chronic myeloid leukemia.
Once you have reviewed and made any alterations, provide the updated gene LIST of genes:
OUTPUT JSON FORMAT