# ENHANCING GENERATIVE PERTURBATION MODELS WITH LLM-INFORMED GENE EMBEDDINGS

**Kaspar Märtens, Rory Donovan-Maiye & Jesper Ferkinghoff-Borg**
Digital Science & Innovation, Novo Nordisk
{KQTM,RZDM,JFGB}@novonordisk.com

## ABSTRACT

Genetic perturbations are key to understanding how genes regulate cell behavior, yet the ability to predict responses to these perturbations remains a significant challenge. While numerous generative models have been developed for perturbation data, they typically lack the capability to generalize to perturbations not encountered during training. To alleviate this limitation, we propose a novel methodology that incorporates prior knowledge through embeddings derived from Large Language Models (LLMs), effectively informing our predictive models with a deeper biological context. By leveraging this source of pre-existing information, our models achieve state-of-the-art performance in predicting the outcomes of single-gene perturbations.

## 1 INTRODUCTION

Understanding cellular responses to perturbations, such as genetic modifications or chemical treatments, is a cornerstone in deciphering complex biological systems. By systematically altering cellular conditions, researchers can observe changes in gene expression and cellular behavior, thereby uncovering the mechanisms underlying health and disease. Biotechnological advances have made it possible to perform such perturbations, such as gene knockouts, and obtain the corresponding single-cell transcriptomics (scRNA-seq) readout (Peidli et al., 2024).

While there exist large-scale single-cell datasets containing gene expression profiles from millions of cells across comprehensive projects such as the Human Cell Atlas (Regev et al., 2017), all of this data is *observational* in its nature. Therefore, while such datasets may yield certain insights into gene regulation, the conclusions derived from observational data remain inherently constrained. This is in contrast with only a small number of publicly available perturbation screens, which provide *interventional* data, allowing our analyses to move beyond correlation towards causation. In this paper, we focus on genetic perturbations in particular, where individual genes are knocked out.

From the causal inference perspective, genetic perturbations are a tool to *intervene* on the underlying gene regulatory causal graph, providing us samples from the respective interventional distribution (Tejada-Lapuerta et al., 2023; Chevalley et al., 2023), enabling a more profound understanding of causality within biological networks.

Learning generative perturbation models, i.e. models that characterise conditional distributions $p(\mathbf{x} \mid \mathbf{p})$ where $\mathbf{x} \in \mathbb{R}^D$ is the gene expression vector and $\mathbf{p} \in \{0,1\}^D$ is the perturbation vector, are highly desirable as they would let us predict *unseen* perturbation responses. This can be seen as performing *in silico perturbations*, which could drastically accelerate hypothesis testing and reduce the costs associated with wet-lab experiments. Even when considering *single gene* perturbations, mapping out the entire single-gene-perturbation landscape in a particular cell line of interest, while not entirely impossible, is a significant effort. But once we start to consider this single-gene-perturbation landscape across a range of cell types, it already becomes an intractable problem. This complexity grows quickly further when we start to consider combinatorial perturbations.

Developing machine learning techniques for perturbation data is an active research area (Ji et al., 2021), particularly with recent work focusing on generative modelling. Various deep generative modelling frameworks such as CPA (Lotfollahi et al., 2023) and sVAE+ (Lopez et al., 2023b) have been proposed and shown to work well, however they do not have a mechanism to make predictions

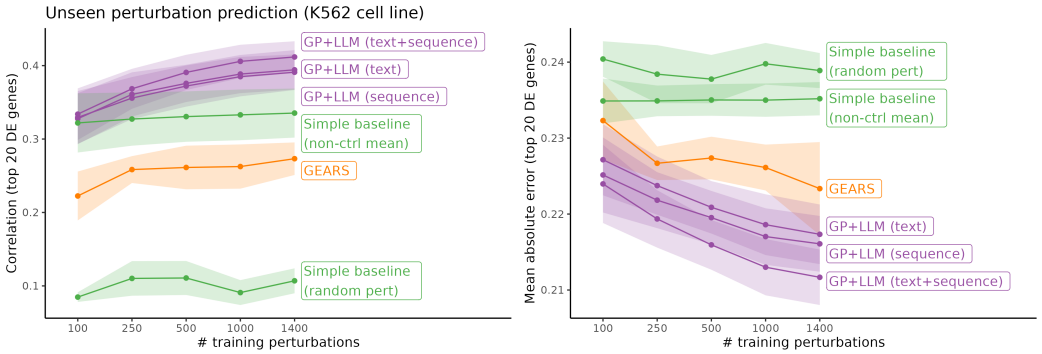


Figure 1: **A simple baseline performs surprisingly well, while LLM-informed GP models achieve state-of-the-art performance.** Performance of various models (simple baselines in green, state-of-the-art perturbation model GEARS in orange, and our proposed LLM-informed models in purple) according to two metrics, Pearson correlation of predicted vs observed differences relative to control cells (higher is better), and mean absolute error relative to control cells (MAE, lower is better). The shaded areas show 1.96 standard error estimates across different data splits. The "random pert" baseline chooses uniformly at random a perturbation from the training set, whereas "non-ctrl mean" estimates the mean expression of all non-control cells in the training set. Our GP+LLM models use either a literature-based text embedding ("text"), a protein sequence based embedding ("sequence"), or both ("text+sequence").

for unseen perturbations. While these models can be fine-tuned once new data becomes available, they lack the capability to make predictions for yet-to-be-seen experiments where new genes would be knocked out. While generalising to an unseen gene knockout is arguably a very challenging task, it would be highly desirable to develop models that have this capability.

A recent graph neural network based method GEARS (Roohani et al., 2023) incorporates prior information about gene-gene relationships based on gene ontology (GO), allowing it to extrapolate to unseen perturbations. To our knowledge, GEARS is the only existing method that has this capability.

Recent advances in Large Language Models, such as GPT-3.5 and GPT-4 (OpenAI et al., 2023), have provided researchers with access to models that have been trained on extensive text corpora, including biomedical text. Chen & Zou (2023) have demonstrated that gene descriptions from the NCBI resource can be embedded using the GPT-3.5 model to produce gene embeddings that capture aspects of biological knowledge. In other domains, analogous models such as protein language models have been created (Elnaggar et al., 2021; Lin et al., 2023). We hypothesise that gene embeddings from such models, either LLMs or dedicated protein language models, could contain valuable information for our task of interest, supplying prior information that would aid in conducting in silico perturbations.

In this work, we focus on predicting *unseen* perturbation responses in the context of single-gene knockouts. Our contributions are as follows:

- We first propose a simple yet effective baseline, which turns out to be competitive with some of the considerably more complex models. Standard practice in the field is to measure predictions against a control group of unperturbed cells; however, we find that the mean expression of such control cells is not zero-centered, which inadvertently provides a baseline predictive signal.
- We propose a class of models that enable to condition on external prior knowledge. Specifically, this knowledge is provided in the form of LLM-based gene embeddings: either gene description text embeddings or protein sequence embeddings.
- We carry out systematic evaluation and demonstrate how our LLM-informed Gaussian Process model sets a new state-of-the-art performance on our benchmarks.

## 2 BACKGROUND

**Deep Generative Models for perturbation data** Various deep generative models have recently been proposed for perturbation data. In general, such models learn to parameterise the conditional densities $p(\mathbf{x} \mid \mathbf{p})$ where $\mathbf{p} \in \{0,1\}^D$ denotes which genes have been perturbed and $\mathbf{x} \in \mathbb{R}^D$ is the post-perturbation gene expression vector (we note that in some formulations, the difference $\Delta := \mathbf{x} - \mathbf{x}_{\text{control}}$ is modelled instead).

Such deep generative models encompass the Compositional Perturbational Autoencoder (Lotfollahi et al., 2023), sVAE+ (Lopez et al., 2023a), and SAMS-VAE (Bereket & Karaletsos, 2023), all of which capture perturbation effects in the latent space. Bunne et al. (2023) propose an optimal transport based approach called CellOT to characterise perturbation effects, in their case these effects are captured as part of the optimal map that transforms unperturbed cells into perturbed ones. However, all of these models exhibit a major limitation – they are not capable of making predictions for an *unseen* perturbation $\mathbf{p}^*$. This is because individual perturbations are effectively treated as discrete categories – while these models learn representations of all the training set perturbations, the respective representations for the test set perturbations are absent.

This limitation has been addressed in GEARS (Roohani et al., 2023) which incorporates prior knowledge about gene-gene relationships in two ways, using a gene co-expression graph as well as a gene-ontology (GO) derived knowledge graph. To our knowledge, GEARS is the only existing method that is equipped with the capability to predict *unseen* single-gene perturbations.

**Gaussian Processes** Gaussian Processes (GPs) offer a probabilistic non-parametric framework for inference over functions, and by extension for non-linear regression (Rasmussen & Williams, 2006). Given some inputs $\mathbf{z} \in \mathbb{R}^D$ (in our case, these could be LLM embeddings), the GP prior is defined via a covariance or kernel function $k(\mathbf{z}, \mathbf{z}')$. A popular choice for the kernel is the RBF kernel $k(\mathbf{z}, \mathbf{z}') = \sigma^2 \exp\left(-0.5 \sum_{j=1}^{D} (z_j - z'_j)^2 / l^2\right)$, where $\sigma^2$ is the kernel variance and $l$ is the lengthscale parameter. Inference for GP regression with inputs $\mathbf{Z} := (\mathbf{z}_1, \ldots, \mathbf{z}_N)$ and outputs $\mathbf{X} := (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ is performed via maximising the log-marginal likelihood

$$\log p(\mathbf{X}|\mathbf{Z}) = \log \mathcal{N}(\mathbf{X} \mid \mathbf{0}, \mathbf{K}_{z,z'} + \sigma^2_{\text{noise}} \mathbf{I}), \tag{1}$$

where $\mathbf{K}_{z,z'}$ is the kernel matrix and $\sigma^2_{\text{noise}}$ is the likelihood noise variance. Here we consider multi-output GP regression, i.e. the case where $\dim(\mathbf{x}) > 1$, assuming independence across output dimensions. The kernel hyperparameters $\sigma^2, l$ and likelihood variance $\sigma^2_{\text{noise}}$ are inferred by maximising the log-marginal likelihood (1).

In this work, we propose to use this multi-output GP model to incorporate LLM-based gene embeddings in a generative framework that enables prediction of unseen gene perturbation responses.

## 3 RESULTS

**Datasets and experimental details** For evaluation, we consider data from 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). After subsetting to the perturbations that were present in both cell lines, we end up with a total of 1774 perturbations. In all our evaluations, we consider 5-fold cross-validation, so in the end, all metrics are calculated on the entire set of 1774 perturbations. In experiments where we consider a gradually increasing number of training perturbations (e.g. along $x$-axis in Figure 1), for every cross-validation split we repeatedly downsample the training set. We use log-transformed gene expression values for 2000 highly variable genes, so $\dim(\mathbf{x}) = 2000$.

**Metrics** Following existing literature, we quantify perturbation prediction performance relative to control cells, i.e. using Pearson correlation and mean absolute error on the differences $\Delta := \mathbf{x} - \mathbf{x}_{\text{control}}$. We calculate both metrics across the top 20 differentially expressed genes[1], as done in (Roohani et al., 2023), as well as across all included genes.

[1] Differentially expressed relative to control cells, resulting in separate gene sets for every perturbation.

### 3.1 COMPETITIVE BASELINE: NON-CONTROL MEAN

**Control cells as a zero-baseline** In general, correlation between observed and predicted log-expression values would be a natural metric to quantify the predictive performance of a model. However, in the context of perturbation data, it is conventional to focus on differences *relative* to control (i.e. unperturbed) cells. In fact, a naive predictor that always predicts the mean control cell expression would achieve correlation close to 1. This is perhaps counter-intuitive, as in general, we would expect a "dummy" model to achieve correlation values close to zero. Therefore, existing literature has adopted an adjusted metric where the focus is on post-perturbation differences relative to control cells, both for observed as well as predicted values (e.g. Bereket & Karaletsos (2023) refer to it as the average treatment effect, and Roohani et al. (2023) call it differential expression relative to unperturbed cells).

**Challenging the zero-baseline assumption** As a result, simply predicting control cell expression will now result in correlation zero, and seemingly, this has brought us back to the typical regime for evaluating models where correlation values above zero indicate that the model has learned something useful in the sense that its predictions are more informative than a random guess. However, it turns out that in the two perturbation datasets considered in this paper, the distribution of the differences $\mathbf{x}_n - \mathbf{x}_{\text{control}}$ across perturbations $n$ is not zero-centered. This can be explained by an observation that for certain genes, the majority of their expression changes across perturbations tend to be unidirectional, i.e. they either go up or down. It turns out that this can be trivially estimated from the training data.

**Competitive baseline** Figure 1 displays two such baselines: one shows performance for a randomly chosen perturbation from a training set, and the other estimates the mean of all the non-control cells in the training set. Both of these baselines achieve above zero correlation, but the former performs particularly well. Interestingly, the non-control mean seems to perform better according to the correlation metric than the prediction error MAE, whereas the relative ordering of the rest of methods remains consistent. Notably, even with relatively small training sets consisting of 100 perturbations, this baseline establishes a strong baseline performance.

This observation suggests the possibility of gene clusters that exhibit similar transcriptional responses to knockout events. Furthermore, we speculate that the strength of this baseline effect could be amplified due to a selection bias in the perturbations included in the original experiment. Specifically, the researchers focused on *essential genes* which may have a more pronounced and consistent impact on cell behavior when perturbed.

### 3.2 LLM-INFORMED PERTURBATION MODELS IMPROVE STATE-OF-THE-ART

**Prior knowledge** In order to construct generative models with an ability to predict unseen gene perturbations, we would need to incorporate some notion of prior knowledge. Fundamentally, using a generative model to sample from $p(\mathbf{x} \mid \mathbf{p}^*)$ for unseen perturbations $\mathbf{p}^*$ is a challenging extrapolation task. To successfully tackle this, it would be important to incorporate some form of prior knowledge on how $\mathbf{p}^*$ relates to the training perturbations $\mathbf{p}_1, \ldots, \mathbf{p}_N$, e.g. via some notion of distance. Without this, $\mathbf{p}^*$ would have to be treated like just a new discrete label category.

**LLM-informed gene embeddings** Motivated by the overall success of LLMs, and their ability to produce embeddings that have been shown to be broadly useful for various downstream tasks, we sought to investigate whether gene embeddings extracted from language models provide a useful source of information to guide this extrapolation task. Specifically, we consider two sets of LLM-informed gene embeddings:

1. **Gene descriptions (text)**: Text embeddings of NCBI gene descriptions obtained using GPT-3.5, as proposed by Chen & Zou (2023)[2]
2. **Protein sequence**: Sequence embeddings from a protein language model ProtT5 (Elnaggar et al., 2021) available from UniProt[3].

[2]Extracted from https://github.com/yiqunchen/GenePT

[3]Embeddings available from https://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/embeddings/

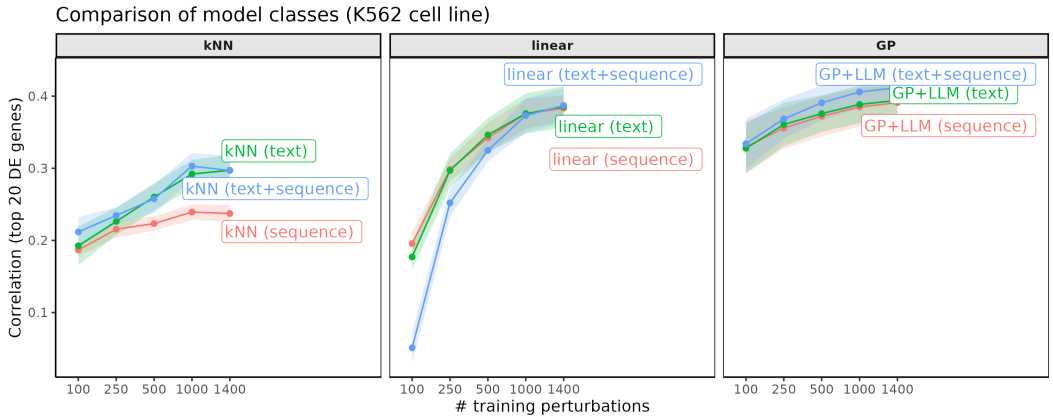


Figure 2: **Comparison of model classes (kNN, linear, and GP)** for multi-output regression when applied to the same set of input embeddings (sequence, text, and both). Here, kNN has a fixed hyperparameter $k = 1$, corresponding to an arguably highly interpretable model, effectively performing nearest-neighbour search in the embedding space. Linear models become competitive with GPs at 1400 perturbations, but underperform at smaller sample sizes. GPs perform best throughout.

**Model classes** In principle, these embeddings can be used as an input to any machine learning model. Here, we consider three model classes: k-nearest neighbors (kNN), linear regression, and Gaussian Process regression. We note that as the outputs are high-dimensional gene expression vectors, we use multi-output adaptations of all these (kNN, linear, GP) models. To reduce the dimensionality of embeddings, we use the top 50 principal components of each embedding (text and/or sequence). For the GP model, we use an RBF kernel with a shared lengthscale parameter, and we assume independence across output dimensions.

**Comparisons and interpretation of results** Figure 2 shows the performance of all model classes, for "text", "sequence" and combined inputs. For kNN, here we show the special case with a fixed $k = 1$, as this is a highly interpretable configuration that provides some insight into the gene embedding landscape, as predictions are based on the nearest neighbour search in the embedding space. For this $k = 1$ case, the direct relevant baseline is "random perturbation" from Figure 1, and indeed all kNN versions outperform this baseline[4]. Overall, Gaussian Processes are the best performing models throughout. In the rest of the comparisons, we show the results for the GP model, referred to as "GP+LLM".

Table 1: **Model performance metrics (1400 training perturbations, K562 cell line)**. Various metrics from left to right: Pearson correlation across top 20 DE genes, Pearson correlation across all genes, MAE across top 20 DE genes, MAE on all genes.

| Model | Cor top 20 | Cor | MAE top20 | MAE |
|---|---|---|---|---|
| Simple baseline (random pert) | 0.107 ± 0.02 | 0.073 ± 0.01 | 0.239 ± 0.00 | 0.091 ± 0.00 |
| Simple baseline (non-ctrl mean) | 0.335 ± 0.03 | 0.264 ± 0.01 | 0.235 ± 0.00 | 0.064 ± 0.00 |
| GEARS | 0.273 ± 0.02 | 0.202 ± 0.01 | 0.223 ± 0.01 | 0.068 ± 0.00 |
| GP+LLM (text) | 0.394 ± 0.03 | 0.284 ± 0.01 | 0.217 ± 0.00 | 0.062 ± 0.00 |
| GP+LLM (sequence) | 0.391 ± 0.02 | 0.287 ± 0.01 | 0.216 ± 0.00 | **0.061** ± 0.00 |
| GP+LLM (text + sequence) | **0.412** ± 0.02 | **0.294** ± 0.01 | **0.212** ± 0.00 | **0.061** ± 0.00 |

As we would expect the two information sources ("text" and "sequence") to provide complementary information about the gene landscape, it is interesting to see that they are comparably informative in terms of predictive performance. Further, combining the two leads to a further small performance boost. In Figures 1 and Supplementary S1 (for K562 and RPE1 cell lines respectively) we compare

[4]We note that the performance of kNN could be improved further by performing hyperparameter search for $k$, but in our experience it did not outperform the GP model. The latter has a further advantage that its kernel hyperparameters can be optimised with gradient descent.

our proposed models – GP+LLM(text), GP+LLM(sequence), and GP+LLM(text+sequence) – with the two baselines and state-of-the-art GEARS model. In Tables 1 and Supplementary Table S1, we also provide further metrics, covering evaluation on both the top 20 differentially expressed genes as well as all 2000 highly variable genes. Across all these metrics, GP+LLM models outperform the other baselines, and in particular the GP+LLM(text+sequence) model performing the best.

Finally, alongside overall average performance, we wanted to shed light into performance on the level of individual perturbations – afterall, making predictions for individual gene knockouts is closer to the actual real-world usecase. Therefore in Figure 3, where every dot corresponds to one perturbation (for K562 cell line on the left, and RPE1 on the right), we compare the predictive performance of GEARS with our best GP+LLM(text+sequence) model.

Results in Figure 3 indicate that there exist perturbations which are better predicted by our proposed model (in blue) as well as some which are better predicted by GEARS (in red). In future work, it would be valuable to further investigate whether there exist any patterns where one model is systematically better (or worse) than the other. However, overall, in 78% and 80% of the cases (K562 and RPE1 respectively) in Figure 3, GP+LLM outperforms GEARS.

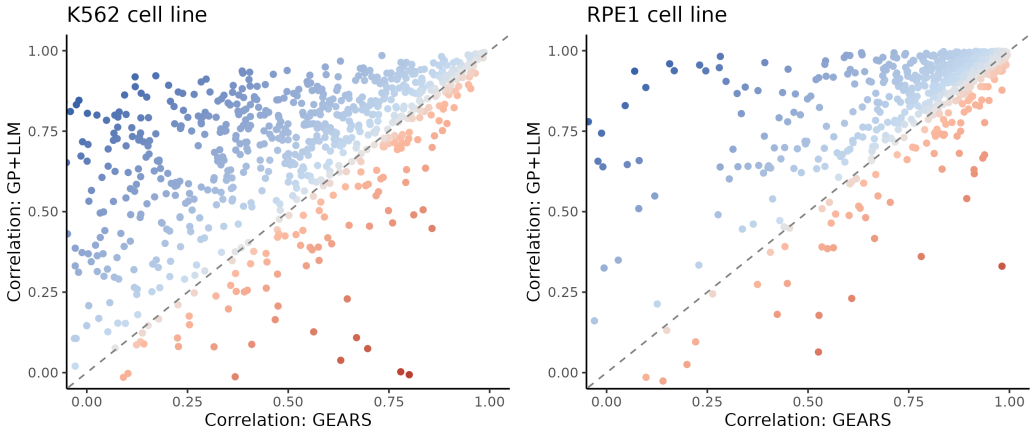


Figure 3: **GEARS vs GP+LLM model: shedding light beyond overall average performance.** For individual perturbations (here shown for top 50% of perturbations with strong effect sizes, every perturbation is shown as one dot), comparing the GEARS and our GP+LLM(text+sequence) model based on the correlation metric on top 20 genes. For dots in the upper triangle, GP+LLM outperforms GEARS (shown in blue), and vice versa in the bottom triangle (shown in red). We note that in 78% and 80% of the perturbations respectively for K562 and RPE1, GP+LLM outperforms GEARS.

## 4 Discussion

In this paper, we proposed to make use of external prior information in the form of gene embeddings with the goal to enchance the out-of-distribution generative capabilities of perturbation models. By embedding natural text and/or protein sequences, combined with a Gaussian Process on those embeddings, we achieved state-of-the-art performance on the task of predicting the outcomes of unseen single-gene perturbations.

In future work, it would be valuable to investigate whether LLM-informed embeddings also provide a useful prior to predict double-gene perturbation outcomes. This could potentially extend the applicability of our models to more complex genetic interventions. Additionally, it would be beneficial to explore and incorporate a broader array of data sources that can inform and improve the gene embeddings for predictive modeling.

## REFERENCES


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.

Charlotte Bunne, Stefan G. Stark, Gabriele Gut, Jacobo Sarabia Del Castillo, Mitch Levesque, Kjong-Van Lehmann, Lucas Pelkmans, Andreas Krause, and Gunnar Rätsch. Learning single-cell perturbation responses using neural optimal transport. *Nature Methods*, 20(11):1759–1768, 2023. URL https://www.nature.com/articles/s41592-023-01969-x. Publisher: Nature Publishing Group US New York.

Yiqun T. Chen and James Zou. GenePT: A Simple But Hard-to-Beat Foundation Model for Genes and Cells Built From ChatGPT. *bioRxiv*, pp. 2023–10, 2023. URL https://www.biorxiv.org/content/10.1101/2023.10.16.562533.abstract. Publisher: Cold Spring Harbor Laboratory.

Mathieu Chevalley, Jacob Sackett-Sanders, Yusuf Roohani, Pascal Notin, Artemy Bakulin, Dariusz Brzezinski, Kaiwen Deng, Yuanfang Guan, Justin Hong, Michael Ibrahim, Wojciech Kotlowski, Marcin Kowiel, Panagiotis Misiakos, Achille Nazaret, Markus Püschel, Chris Wendler, Arash Mehrjou, and Patrick Schwab. The CausalBench challenge: A machine learning contest for gene network inference from single-cell perturbation data, August 2023. URL http://arxiv.org/abs/2308.15395. arXiv:2308.15395 [cs, q-bio].

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Learning. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, May 2021. doi: 10.1101/2020.07.12.199554. URL https://www.biorxiv.org/content/10.1101/2020.07.12.199554v3. Pages: 2020.07.12.199554 Section: New Results.

Yuge Ji, Mohammad Lotfollahi, F. Alexander Wolf, and Fabian J. Theis. Machine learning for perturbational single-cell omics. *Cell Systems*, 12(6):522–537, 2021. URL https://www.cell.com/cell-systems/pdf/S2405-4712(21)00202-7.pdf. Publisher: Elsevier.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, March 2023. doi: 10.1126/science.ade2574. URL https://www.science.org/doi/full/10.1126/science.ade2574. Publisher: American Association for the Advancement of Science.

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 2023a. URL https://openreview.net/forum?id=IOWJsPJ2xGd.

Romain Lopez, Nataša Tagasovska, Stephen Ra, Kyunghyn Cho, Jonathan K. Pritchard, and Aviv Regev. Learning Causal Representations of Single Cells via Sparse Mechanism Shift Modeling. Conference on Causal Learning and Reasoning, February 2023b. doi: 10.48550/arXiv.2211.03553. URL http://arxiv.org/abs/2211.03553.

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.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and others. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Stefan Peidli, Tessa D. Green, Ciyue Shen, Torsten Gross, Joseph Min, Samuele Garda, Bo Yuan, Linus J. Schumacher, Jake P. Taylor-King, Debora S. Marks, Augustin Luna, Nils Blüthgen, and Chris Sander. scPerturb: harmonized single-cell perturbation data. *Nature Methods*, pp. 1–10, January 2024. ISSN 1548-7105. doi: 10.1038/s41592-023-02144-y. URL https://www.nature.com/articles/s41592-023-02144-y. Publisher: Nature Publishing Group.

Carl Edward Rasmussen and Christopher Williams. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006. URL http://newton.cam.ac.uk/files/seminar/20070809140015001-150844.pdf.

Aviv Regev, Sarah A. Teichmann, Eric S. Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, and Menna Clatworthy. The human cell atlas. *elife*, 6:e27041, 2017. URL https://elifesciences.org/articles/27041. Publisher: eLife Sciences Publications, Ltd.

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.

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.

Alejandro Tejada-Lapuerta, Paul Bertin, Stefan Bauer, Hananeh Aliee, Yoshua Bengio, and Fabian J. Theis. Causal machine learning for single-cell genomics, October 2023. URL http://arxiv.org/abs/2310.14935. arXiv:2310.14935 [cs, q-bio].

## SUPPLEMENTARY FIGURES

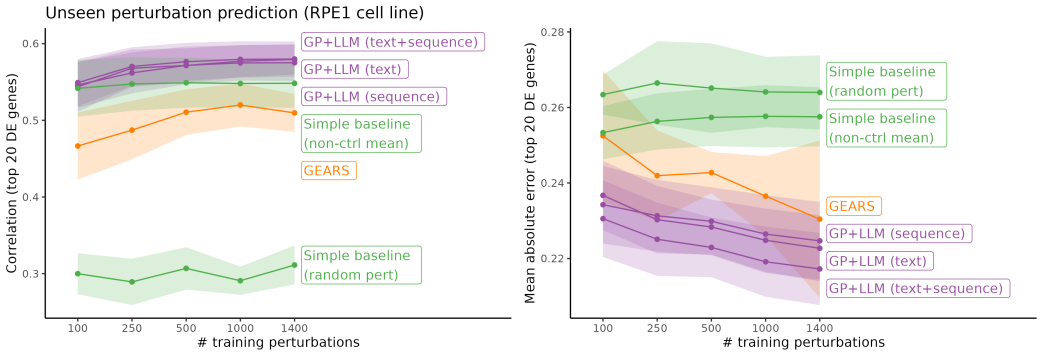


Figure S1: Analogously to Figure 1, here is the performance of various models on the RPE1 cell line, according to Pearson correlation (left panel) and MAE (right panel).

## SUPPLEMENTARY TABLES

Table S1: **Model Performance Metrics (1400 perturbations, RPE1 cell line).**

| Model | Cor top20 | Cor | MAE top20 | MAE |
|---|---|---|---|---|
| Simple baseline (random pert) | 0.311 ± 0.03 | 0.245 ± 0.02 | 0.264 ± 0.01 | 0.123 ± 0.00 |
| Simple baseline (non-ctrl mean) | **0.548** ± 0.03 | 0.479 ± 0.02 | 0.258 ± 0.01 | 0.088 ± 0.00 |
| GEARS | 0.510 ± 0.02 | 0.434 ± 0.03 | 0.230 ± 0.02 | 0.090 ± 0.01 |
| GP+LLM (text) | **0.579** ± 0.02 | 0.480 ± 0.01 | 0.223 ± 0.01 | 0.081 ± 0.00 |
| GP+LLM (sequence) | **0.575** ± 0.02 | 0.481 ± 0.01 | 0.225 ± 0.01 | 0.080 ± 0.00 |
| GP+LLM (text + sequence) | **0.580** ± 0.02 | **0.483** ± 0.01 | **0.217** ± 0.01 | **0.079** ± 0.00 |