

PERTURBERT: LEARNING GENE CO-VARIATION EMBEDDINGS FROM PERTURBATION SIGNATURES

Anonymous authors

Paper under double-blind review

ABSTRACT

Current foundation models for transcriptomic data are typically trained in a self-supervised manner to predict gene expression within a sample given other genes, thereby learning gene co-variation patterns from *observational* data. However, many translational applications require understanding how gene expression changes in response to *interventions*. We introduce **PerturBERT**, an encoder-only transformer pre-trained with masked-gene modeling on approximately 1M perturbation signatures across 248 cell lines that learns *perturbational* co-variance patterns from gene perturbation responses. PerturBERT tokenizes each signature as a set of (gene, response) pairs and produces gene embeddings contextualized by their response to interventions. PerturBERT gene embeddings achieve state-of-the-art results on a gene embedding benchmark and gene dependency prediction. To our knowledge, PerturBERT is the first transformer explicitly pre-trained on gene perturbation responses, offering representations complementary to models trained on observational gene expression profiles.

1 INTRODUCTION

Large-scale perturbation profiling enables systematic readouts of cellular responses to genetic and chemical interventions across diverse cell types, shifting the field from correlative atlases toward *interventional maps* of cellular function (Zhang et al., 2025; Replogle et al., 2022; Nadig et al., 2025; Rood et al., 2024). While gene2vec (Du et al., 2019) and recent single-cell transformers (Szałata et al., 2024b) such as iSEEEK (Shen et al., 2022) and scGPT (Haotian Cui et al., 2023) have demonstrated the utility of learning gene embeddings from gene co-expression patterns in large-scale datasets, no equivalent method has yet emerged for perturbation datasets.

We hypothesize that gene co-variation under interventions is a strong pre-training signal, and thus learn gene embeddings directly from gene perturbation responses rather than observed expression abundances. We introduce **PerturBERT**, a BERT-style (Devlin et al., 2019) representation learning framework featuring masked-gene modeling on gene perturbation responses. By training on gene perturbation responses, PerturBERT uniquely leverages co-variation from gene perturbational effects to learn gene embeddings that can be repurposed for downstream tasks.

In summary, we (1) formulate perturbation response *co-variation* as a pre-training signal for transcriptomic foundation models; (2) introduce **PerturBERT**, a BERT-style representation learning framework trained on sets of (gene, response) tokens from approximately 1M perturbation signatures; and (3) demonstrate that PerturBERT’s gene embeddings yield state-of-the-art performance on a gene embedding benchmark (Kan-Tor et al., 2024) and gene dependency prediction.

2 RELATED WORK

Gene embedding models. Existing models for learning gene embeddings, including gene2vec and single-cell foundation models (iSEEEK, scGPT), are primarily trained on gene expression abundances of cells and thus capture co-variation with respect to these observational states. These methods differ from each other primarily in how they encode the expression levels of genes (e.g., rank-ordering, value binning, and value projection). While gene embeddings for several of these methods have been shown to be effective for some isolated tasks, such as cell annotation and cross-dataset in-

054
055
056
057
058
059
060
061
062
063
064
065
066
067
068
069
070
071
072
073
074
075
076
077
078
079
080
081
082
083
084
085
086
087
088
089
090
091
092
093
094
095
096
097
098
099
100
101
102
103
104
105
106
107

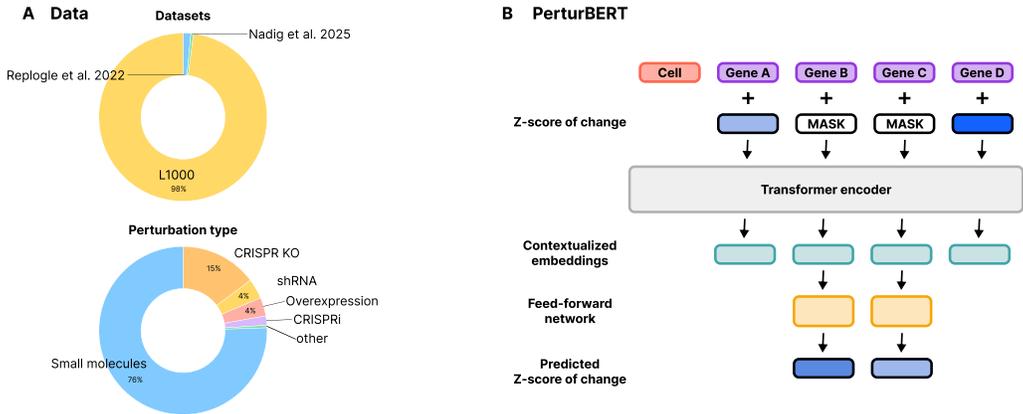


Figure 1: **PerturBERT framework.** (A) The dataset comprises 1M perturbation signatures from L1000 profiles data and CRISPRi genetic screens. (B) PerturBERT is a BERT-style transformer encoder trained via masked-gene modeling. Inputs are tokenized as cell identity and gene-response pairs (gene + binned Z-score). The model learns to reconstruct masked response values, producing gene embeddings that capture co-variation patterns.

tegration, more comprehensive evaluations of their utility report their failures to outperform simple baselines (Liu et al., 2023; Szałata et al., 2024b;a).

Perturbation modeling. While the number of large-scale perturbation datasets has grown in recent years, much of the work around perturbation modeling has focused on forecasting the effects of perturbations rather than representation learning. Prior perturbation prediction methods make use of technical innovations such as latent operators, conditional optimal transport, and flow matching (Lotfollahi et al., 2019; Bunne et al., 2023; Klein et al., 2025), as well as growing gene embedding collections (Littman et al., 2025; Miladinovic et al., 2025). In particular, the Large Perturbation Model (LPM) (Miladinovic et al., 2025) learns gene-level embeddings through perturbation effect prediction; however, it is restricted to genetic perturbations and works solely in a supervised manner. PerturBERT differs in that it leverages self-supervised training on gene-level perturbation responses across various intervention types to learn gene embeddings informed by perturbation effects.

3 DATA AND REPRESENTATION

Training corpus. We curated a dataset of approximately 1M perturbation signatures, combining LINCS L1000 (Subramanian et al., 2017) readouts of chemical, genetic, and biomolecule perturbations with single-cell transcriptomic profiles of genetic perturbations from Perturb-seq screens across diverse cell lines (Replogle et al., 2022). Single-cell data is processed with standard differential expression pipelines (Nadig et al., 2025; Love et al., 2014) and harmonized to match L1000 level 5 conventions. The final training set comprises 940,564 signatures across 248 cell lines with a vocabulary of 932 genes, corresponding to the set of genes measured in the L1000 assays. Extended dataset statistics are in Appendix A.

Tokenization. Each signature is represented as a set of learnable gene embeddings $g \in \mathcal{G}$ and a set of corresponding expression change tokens $b \in \mathcal{B}$ that indicate different levels of up- and down-regulation for each gene. Expression changes were discretized into 17 bins, and each bin was mapped to a unique embedding. The input embedding for each gene in a signature is the element-wise sum of its gene embedding and expression change token embeddings. To disentangle cell line-specific effects from perturbation responses, we condition the model on cell line identity via learnable cell line embeddings. The learnable cell embedding is concatenated to the signature embeddings.

4 MODEL AND OBJECTIVE

Architecture. PerturBERT is an encoder-only transformer with 8 layers and 12 attention heads (57.6M parameters). We do not use positional encodings since perturbation signatures are unordered sets. Full architecture details are provided in Appendix B.

Masked-gene modeling. We randomly mask 40% of (gene, value) pairs during training and train the model to recover the masked expression change tokens from the contextualized embedding (Wettig et al., 2023). Following BERT conventions, masked positions are replaced with a [MASK] token 80% of the time, a random bin 10% of the time, and left unchanged 10% of the time. We optimize cross-entropy loss, where \mathcal{M} denotes the masked index set and \hat{b}_i the predicted logits over expression-change bins at position i : $\mathcal{L}_{\text{MGM}} = \sum_{i \in \mathcal{M}} \text{CE}(\hat{b}_i, b_i)$.

Training. We train for 80 epochs with an effective batch size of 1,600 (200 per device across 8 A100 GPUs) using AdamW (Loshchilov & Hutter, 2018) with learning rate 5×10^{-5} , cosine annealing, and 200 warmup steps. The complete training configuration is detailed in Appendix C.

5 RESULTS

Overview. We evaluate PerturBERT gene embeddings on two categories of tasks: (i) *gene property prediction* using the gene-embedding benchmark of Kan-Tor et al. (2024) (Appendix D), and (ii) *gene dependency prediction* on DepMap Q2 2025 with cell line-based splits (Appendix D) (Tsherniak et al., 2017). Across tasks, we compare against scGPT, GenePT, and gene2vec embeddings, and for dependency prediction, we additionally include a one-hot target-gene baseline.

5.1 GENE PROPERTY PREDICTION

Setup. We follow Kan-Tor et al. (2024), training lightweight predictors on top of frozen gene embeddings for each gene-property task (Appendix D).

Results. PerturBERT gene embeddings match or exceed state-of-the-art performance across a diverse set of gene property prediction tasks (Table 1; Appendix G). PerturBERT achieves the best score or performs within one standard error of the best model on 10 of 15 tasks. PerturBERT shows superior performance on cell cycle-related tasks (CCD Transcript: 0.837 vs. scGPT’s 0.813), pathology prediction (Endometrial: 0.610 vs. scGPT’s 0.593), and tissue specificity tasks. On remaining tasks, PerturBERT typically ranks second, with scGPT achieving the highest scores.

PerturBERT training scale. Where PerturBERT exceeds baselines, absolute margins are modest. However, substantial differences in training data suggest relative data-efficiency of PerturBERT. scGPT, the most competitive baseline, was trained on 33M single-cell transcriptomes with a vocabulary of over 60,000 genes, whereas PerturBERT was trained on approximately 1M primarily bulk perturbation signatures with only 932 genes. Despite the ~ 30 -fold reduction in training samples and ~ 65 -fold reduction in gene vocabulary, PerturBERT achieves competitive or superior performance across most tasks.

5.2 PERTURBERT EMBEDDINGS CAPTURE CELL CYCLE BIOLOGY

PerturBERT recovers cell-cycle structure in its embedding space. On the CCD Transcript benchmark task, PerturBERT outperforms scGPT (ROC-AUC 0.837 vs. 0.813; Table 1). Beyond performance on this task, CCD genes are more similar to each other in embedding space: mean within-CCD cosine similarity is 0.051 versus 0.008 for CCD-to-non-CCD pairs. UMAP and cosine-similarity heatmap visualizations of these relationships are provided in Appendix E.

5.3 GENE DEPENDENCY PREDICTION

Setup. We train LightGBM models to predict gene dependency using (a) cell line gene expression as the cell line representation and (b) a gene embedding as a target-gene representation; these two

Table 1: Gene embedding benchmark results (mean ROC-AUC across cross-validation splits \pm standard error). Included here are 10 of 15 tasks where PerturBERT achieves best score (all results in Appendix G). Bold indicates best, including those within one standard error.

Task	Model	ROC-AUC	Task	Model	ROC-AUC
CCD Transcript	PerturBERT	0.837 \pm 0.034	RNA blood cell dist.	PerturBERT	0.780 \pm 0.008
	scGPT	0.813 \pm 0.036		scGPT	0.779 \pm 0.012
	GenePT	0.755 \pm 0.037		gene2vec	0.630 \pm 0.012
	gene2vec	0.695 \pm 0.033		GenePT	0.603 \pm 0.010
NI network	PerturBERT	0.708 \pm 0.066	RNA blood cell spec.	scGPT	0.811 \pm 0.012
	scGPT	0.669 \pm 0.073		PerturBERT	0.804 \pm 0.008
	gene2vec	0.581 \pm 0.076		gene2vec	0.681 \pm 0.019
	GenePT	0.564 \pm 0.073		GenePT	0.676 \pm 0.020
Pathology - Cervical	scGPT	0.589 \pm 0.017	RNA blood lineage dist.	scGPT	0.826 \pm 0.013
	PerturBERT	0.557 \pm 0.028		PerturBERT	0.816 \pm 0.013
	gene2vec	0.512 \pm 0.048		gene2vec	0.699 \pm 0.011
	GenePT	0.508 \pm 0.027		GenePT	0.674 \pm 0.018
Pathology - Endometrial	PerturBERT	0.610 \pm 0.021	RNA blood lineage spec.	scGPT	0.786 \pm 0.019
	scGPT	0.593 \pm 0.024		PerturBERT	0.781 \pm 0.015
	gene2vec	0.500 \pm 0.024		gene2vec	0.714 \pm 0.020
	GenePT	0.464 \pm 0.035		GenePT	0.688 \pm 0.018
Pathology - Pancreatic	scGPT	0.617 \pm 0.026	RNA tissue specificity	PerturBERT	0.723 \pm 0.006
	PerturBERT	0.612 \pm 0.025		scGPT	0.707 \pm 0.021
	GenePT	0.562 \pm 0.025		GenePT	0.634 \pm 0.023
	gene2vec	0.542 \pm 0.029		gene2vec	0.621 \pm 0.011

Table 2: Gene dependency prediction results on DepMap (mean \pm standard error across 20 training seeds). Bold indicates best, including those within one standard error.

Embedding	RMSE \downarrow	MAE \downarrow	R^2 \uparrow
PerturBERT	0.185290 \pm 0.000047	0.125573 \pm 0.000020	0.880468 \pm 0.000060
scGPT	0.185351 \pm 0.000047	0.125546 \pm 0.000023	0.880390 \pm 0.000060
gene2vec	0.185518 \pm 0.000063	0.125680 \pm 0.000022	0.880174 \pm 0.000081
GenePT	0.186387 \pm 0.000051	0.126280 \pm 0.000024	0.879049 \pm 0.000066
one-hot	0.189850 \pm 0.000938	0.127211 \pm 0.000304	0.874455 \pm 0.001248

vectors are concatenated to form model’s model input. The evaluation uses DepMap (Tsherniak et al., 2017) Q2 2025 data with cell line-based splits (Appendix D).

Results. PerturBERT gene embeddings yield competitive performance on cell line dependency prediction (Table 2). PerturBERT achieves the lowest RMSE (0.1853) and highest R^2 (0.8805), matching scGPT within standard error and outperforming gene2vec, GenePT, and one-hot encodings. Again, PerturBERT matches scGPT’s performance despite the substantial disparity in pre-training data scale and gene coverage.

6 DISCUSSION AND CONCLUSION

PerturBERT reframes transcriptomic foundation model pre-training around co-variation in gene perturbation effects, rather than co-expression. This reframing supports the idea that genes that co-vary under perturbations encode distinct and biologically useful structure. PerturBERT matches or exceeds state-of-the-art performance across a diverse downstream tasks evaluating gene embeddings.

The comparison with scGPT illustrates that despite being trained on $\sim 30\times$ fewer samples and covering $\sim 65\times$ fewer genes, PerturBERT achieves comparable or superior results on most tasks. This suggests that perturbation-induced co-variation constitutes a highly informative training signal that may encode functional relationships not captured by co-expression patterns alone.

Importantly, PerturBERT produces reusable gene embeddings that can be probed, fine-tuned, or paired with lightweight heads across downstream applications, in contrast to task-specific perturbation-effect predictors. Hence, we view PerturBERT as complementary to co-expression foundation models, where the distinct training signals may capture orthogonal aspects of gene function, and as progress toward perturbation-centric foundation models for cellular biology.

REFERENCES

- 216
217
218 Charlotte Bunne, Andreas Krause, and Marco Cuturi. Supervised Training of Conditional Monge
219 Maps, March 2023. URL <http://arxiv.org/abs/2206.14262>. arXiv:2206.14262 [cs].
- 220 Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep
221 Bidirectional Transformers for Language Understanding. In Jill Burstein, Christy Doran, and
222 Tamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of
223 the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long
224 and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Com-
225 putational Linguistics. doi: 10.18653/v1/N19-1423. URL [https://aclanthology.org/
226 N19-1423/](https://aclanthology.org/N19-1423/).
- 227 Jingcheng Du, Peilin Jia, Yulin Dai, Cui Tao, Zhongming Zhao, and Degui Zhi. Gene2vec: dis-
228 tributed representation of genes based on co-expression. *BMC Genomics*, 20(1):82, February
229 2019. ISSN 1471-2164. doi: 10.1186/s12864-018-5370-x. URL [https://doi.org/10.
230 1186/s12864-018-5370-x](https://doi.org/10.1186/s12864-018-5370-x).
- 231 Haotian Cui, Chloe X. Wang, Hassaan Maan, and Bo Wang. ScGPT: Towards Building a Foundation
232 Model for Single-Cell Multi-omics Using Generative AI. *bioRxiv*, 2023. doi: 10.1101/2023.04.
233 30.538439. S2ID: 88399c4c6574be850918a1bce2dede2c87ef8241.
- 234 Yoav Kan-Tor, Michael Morris Danziger, Eden Zohar, Matan Ninio, and Yishai Shimoni. Does
235 your model understand genes? A benchmark of gene properties for biological and text models,
236 December 2024. URL <http://arxiv.org/abs/2412.04075>. arXiv:2412.04075 [cs].
- 237 Dominik Klein, Jonas Simon Fleck, Daniil Bobrovskiy, Lea Zimmermann, Sören Becker, Alessan-
238 dro Palma, Leander Dony, Alejandro Tejada-Lapuerta, Guillaume Huguet, Hsiu-Chuan Lin,
239 Nadezhda Azbukina, Fátima Sanchís-Calleja, Theo Uscidda, Artur Szalata, Manuel Gander, Aviv
240 Regev, Barbara Treutlein, J. Gray Camp, and Fabian J. Theis. CellFlow enables generative single-
241 cell phenotype modeling with flow matching, April 2025. URL [https://www.biorxiv.
242 org/content/10.1101/2025.04.11.648220v1](https://www.biorxiv.org/content/10.1101/2025.04.11.648220v1). Pages: 2025.04.11.648220 Section:
243 New Results.
- 244 Russell Littman, Jacob Levine, Sepideh Maleki, Yongju Lee, Vladimir Ermakov, Lin Qiu, Alexan-
245 der Wu, Kexin Huang, Romain Lopez, Gabriele Scalia, Tommaso Biancalani, David Rich-
246 mond, Aviv Regev, and Jan-Christian Hütter. Gene-embedding-based prediction and func-
247 tional evaluation of perturbation expression responses with PRESAGE, June 2025. URL
248 <https://www.biorxiv.org/content/10.1101/2025.06.03.657653v1>. Pages:
249 2025.06.03.657653 Section: New Results.
- 250 Tianyu Liu, Kexing Li, Yuge Wang, Hongyu Li, and Hongyu Zhao. Evaluating the Util-
251 ities of Large Language Models in Single-cell Data Analysis, September 2023. URL
252 <https://www.biorxiv.org/content/10.1101/2023.09.08.555192v2>. Pages:
253 2023.09.08.555192 Section: New Results.
- 254 Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. September 2018. URL
255 <https://openreview.net/forum?id=Bkg6RiCqY7>.
- 256 Mohammad Lotfollahi, F. Alexander Wolf, and Fabian J. Theis. scGen predicts single-
257 cell perturbation responses. *Nature Methods*, 16(8):715–721, August 2019. ISSN 1548-
258 7105. doi: 10.1038/s41592-019-0494-8. URL [https://www.nature.com/articles/
259 s41592-019-0494-8](https://www.nature.com/articles/s41592-019-0494-8). Publisher: Nature Publishing Group.
- 260 Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and
261 dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, December 2014.
262 ISSN 1474-760X. doi: 10.1186/s13059-014-0550-8. URL [https://doi.org/10.1186/
263 s13059-014-0550-8](https://doi.org/10.1186/s13059-014-0550-8).
- 264 Djordje Miladinovic, Tobias Höpfe, Mathieu Chevalley, Andreas Georgiou, Lachlan Stuart, Arash
265 Mehrjou, Marcus Bantscheff, Bernhard Schölkopf, and Patrick Schwab. In silico biolog-
266 ical discovery with large perturbation models. *Nature Computational Science*, 5(11):1029–
267 1040, November 2025. ISSN 2662-8457. doi: 10.1038/s43588-025-00870-1. URL <https://doi.org/10.1038/s43588-025-00870-1>.

- 270 //www.nature.com/articles/s43588-025-00870-1. Publisher: Nature Publishing
271 Group.
- 272
- 273 Ajay Nadig, Joseph M. Replogle, Angela N. Pogson, Mukundh Murthy, Steven A. McCarroll,
274 Jonathan S. Weissman, Elise B. Robinson, and Luke J. O'Connor. Transcriptome-wide analy-
275 sis of differential expression in perturbation atlases. *Nature Genetics*, 57(5):1228–1237, May
276 2025. ISSN 1546-1718. doi: 10.1038/s41588-025-02169-3. URL <https://www.nature.com/articles/s41588-025-02169-3>. Publisher: Nature Publishing Group.
- 277
- 278 F Pontén, K Jirstrom, and M Uhlen. The Human Protein Atlas—a tool for pathology. *The Journal*
279 *of Pathology*, 216(4):387–393, 2008. ISSN 1096-9896. doi: 10.1002/path.2440. URL
280 <https://onlinelibrary.wiley.com/doi/abs/10.1002/path.2440>. eprint:
281 <https://pathsocjournals.onlinelibrary.wiley.com/doi/pdf/10.1002/path.2440>.
- 282 Joseph M. Replogle, Reuben A. Saunders, Angela N. Pogson, Jeffrey A. Hussmann, Alexander
283 Lenail, Alina Guna, Lauren Mascibroda, Eric J. Wagner, Karen Adelman, Gila Lithwick-Yanai,
284 Nika Iremadze, Florian Oberstrass, Doron Lipson, Jessica L. Bonnar, Marco Jost, Thomas M.
285 Norman, and Jonathan S. Weissman. Mapping information-rich genotype-phenotype landscapes
286 with genome-scale Perturb-seq. *Cell*, 185(14):2559–2575.e28, July 2022. ISSN 1097-4172. doi:
287 10.1016/j.cell.2022.05.013.
- 288 Jennifer E. Rood, Anna Hupalowska, and Aviv Regev. Toward a foundation model of causal cell
289 and tissue biology with a Perturbation Cell and Tissue Atlas. *Cell*, 187(17):4520–4545, August
290 2024. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2024.07.035. URL [https://www.cell.com/cell/abstract/S0092-8674\(24\)00829-8](https://www.cell.com/cell/abstract/S0092-8674(24)00829-8). Publisher: Elsevier.
- 291
- 292 Hongru Shen, Xilin Shen, Mengyao Feng, Dan Wu, Chao Zhang, Yichen Yang, Meng Yang, Jiani
293 Hu, Jilei Liu, Wei Wang, Yang Li, Qiang Zhang, Jilong Yang, Kexin Chen, and Xiangchun Li.
294 A universal approach for integrating super large-scale single-cell transcriptomes by exploring
295 gene rankings. *Briefings in Bioinformatics*, 23(2):bbab573, March 2022. ISSN 1477-4054. doi:
296 10.1093/bib/bbab573. URL <https://doi.org/10.1093/bib/bbab573>.
- 297
- 298 Aravind Subramanian, Rajiv Narayan, Steven M. Corsello, David D. Peck, Ted E. Natoli, Xi-
299 aodong Lu, Joshua Gould, John F. Davis, Andrew A. Tubelli, Jacob K. Asiedu, David L.
300 Lahr, Jodi E. Hirschman, Zihan Liu, Melanie Donahue, Bina Julian, Mariya Khan, David Wad-
301 den, Ian C. Smith, Daniel Lam, Arthur Liberzon, Courtney Toder, Mukta Bagul, Marek Orze-
302 chowski, Oana M. Enache, Federica Piccioni, Sarah A. Johnson, Nicholas J. Lyons, Alice H.
303 Berger, Alykhan F. Shamji, Angela N. Brooks, Anita Vrcic, Corey Flynn, Jacqueline Rosains,
304 David Y. Takeda, Roger Hu, Desiree Davison, Justin Lamb, Kristin Ardlie, Larson Hogstrom,
305 Peyton Greenside, Nathanael S. Gray, Paul A. Clemons, Serena Silver, Xiaoyun Wu, Wen-
306 Ning Zhao, Willis Read-Button, Xiaohua Wu, Stephen J. Haggarty, Lucienne V. Ronco, Jesse S.
307 Boehm, Stuart L. Schreiber, John G. Doench, Joshua A. Bittker, David E. Root, Bang Wong,
308 and Todd R. Golub. A Next Generation Connectivity Map: L1000 Platform and the First
309 1,000,000 Profiles. *Cell*, 171(6):1437–1452.e17, November 2017. ISSN 0092-8674, 1097-
310 4172. doi: 10.1016/j.cell.2017.10.049. URL [https://www.cell.com/cell/abstract/S0092-8674\(17\)31309-0](https://www.cell.com/cell/abstract/S0092-8674(17)31309-0). Publisher: Elsevier.
- 311
- 312 Artur Szalata, Andrew Benz, Robrecht Cannoodt, Mauricio Cortes, Jason Fong, Sunil Kuppasani,
313 Richard Lieberman, Tianyu Liu, Javier A. Mas-Rosario, Rico Meinl, Jalil Nourisa, Jared Tumieli,
314 Tin M. Tunjic, Mengbo Wang, Noah Weber, Hongyu Zhao, Benedict Anchang, Fabian J.
315 Theis, Malte D. Luecken, and Daniel B. Burkhardt. A benchmark for prediction of transcrip-
316 tomic responses to chemical perturbations across cell types. *Advances in Neural Information*
317 *Processing Systems*, 37:20566–20616, December 2024a. doi: 10.52202/079017-0650.
318 URL [https://proceedings.neurips.cc/paper_files/paper/2024/
hash/24c4d51f3ef48dd2dbab78243ecb26a1-Abstract-Datasets_and_](https://proceedings.neurips.cc/paper_files/paper/2024/hash/24c4d51f3ef48dd2dbab78243ecb26a1-Abstract-Datasets_and_Benchmarks_Track.html)
319 [Benchmarks_Track.html](https://proceedings.neurips.cc/paper_files/paper/2024/hash/24c4d51f3ef48dd2dbab78243ecb26a1-Abstract-Datasets_and_Benchmarks_Track.html).
- 320
- 321 Artur Szalata, Karin Hrovatin, Sören Becker, Alejandro Tejada-Lapuerta, Haotian Cui, Bo Wang,
322 and Fabian J. Theis. Transformers in single-cell omics: a review and new perspectives. *Nature*
323 *Methods*, 21(8):1430–1443, August 2024b. ISSN 1548-7105. doi: 10.1038/s41592-024-02353-z.
URL <https://www.nature.com/articles/s41592-024-02353-z>. Publisher: Nature Publishing Group.

Aviad Tsherniak, Francisca Vazquez, Phil G. Montgomery, Barbara A. Weir, Gregory Kryukov, Glenn S. Cowley, Stanley Gill, William F. Harrington, Sasha Pantel, John M. Krill-Burger, Robin M. Meyers, Levi Ali, Amy Goodale, Yenarae Lee, Guozhi Jiang, Jessica Hsiao, William F. J. Gerath, Sara Howell, Erin Merkel, Mahmoud Ghandi, Levi A. Garraway, David E. Root, Todd R. Golub, Jesse S. Boehm, and William C. Hahn. Defining a Cancer Dependency Map. *Cell*, 170(3):564–576.e16, July 2017. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2017.06.010. URL [https://www.cell.com/cell/abstract/S0092-8674\(17\)30651-7](https://www.cell.com/cell/abstract/S0092-8674(17)30651-7). Publisher: Elsevier.

Alexander Wettig, Tianyu Gao, Zexuan Zhong, and Danqi Chen. Should You Mask 15% in Masked Language Modeling? In Andreas Vlachos and Isabelle Augenstein (eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2985–3000, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.217. URL <https://aclanthology.org/2023.eacl-main.217/>.

Yiming Zhang, Ting Zhang, Gaoxia Yang, Zhenzhong Pan, Min Tang, Yue Wen, Ping He, Yuan Wang, and Ran Zhou. PerturbAtlas: a comprehensive atlas of public genetic perturbation bulk RNA-seq datasets. *Nucleic Acids Research*, 53(D1):D1112–D1119, January 2025. ISSN 1362-4962. doi: 10.1093/nar/gkae851. URL <https://doi.org/10.1093/nar/gkae851>.

A DATASET DETAILS

Training corpus statistics. The PerturBERT training corpus comprises perturbation signatures from LINCS L1000 (level 5, LFC/LFC standard error) and harmonized single-cell Perturb-seq data. Table 3 summarizes the dataset.

Table 3: Training corpus statistics.

Property	Value
Number of genes	932
Number of cell lines	248
Training samples	940,564
Validation samples	95,665
Test samples	95,665

Tokenization scheme. Expression responses are discretized into 17 bins (8 in each direction plus a central no-change bin) representing different magnitudes of up- and down-regulation. An additional [MASK] token is used during training.

Gene vocabulary. The 932-gene vocabulary was selected for training based on overlap across the L1000 and Perturb-seq datasets.

B MODEL ARCHITECTURE

Table 4 provides complete architectural specifications for PerturBERT.

C TRAINING CONFIGURATION

Pre-training objective. PerturBERT is trained with masked bin modeling, where the model predicts the expression bin of masked gene positions. Table 5 details the masking strategy.

Optimization. Table 6 summarizes optimization hyperparameters.

Training details. Table 7 provides the complete training configuration.

378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431

Table 4: PerturBERT architecture details.

Parameter	Value
<i>Transformer Configuration</i>	
Number of layers	8
Number of attention heads	12
Hidden size	768
Intermediate (FFN) size	3,072
Activation function	SiLU (Swish)
Attention mechanism	Scaled dot-product (SDPA)
Position embedding type	None (set-based input)
Hidden dropout probability	0.1
Attention dropout probability	0.1
Layer normalization ϵ	1×10^{-12}
Initializer range	0.02
<i>Vocabulary Sizes</i>	
Gene vocabulary	932
Bin vocabulary	17 (+1 [MASK] token)
Cell line vocabulary	248
<i>Embedding Dimensions</i>	
Gene embeddings	768
Bin embeddings	768
Cell line embeddings	768
Total parameters	57,623,048

Table 5: Masking strategy for training

Token Replacement	Training (40% masked)
Replaced with [MASK]	80% of masked
Replaced with random bin	10% of masked
Unchanged	10% of masked
Excluded from loss	Non-masked positions

Hardware. Training was performed on 8 NVIDIA A100-SXM4-80GB GPUs with CUDA 12.8, 96 logical CPU cores, and 1.12 TB system memory.

D DOWNSTREAM EVALUATION DETAILS

Gene embedding benchmark. Binary and categorical classification tasks use logistic regression with L2 penalty, a maximum of 5,000 iterations, and 20-fold stratified cross-validation with shuffling. Fold count is dynamically adjusted based on minimum class size.

Gene dependency prediction. We predict gene essentiality using DepMap Q2 2025 data with cell line-based splits: 80% training (1,347 lines), 10% validation (168 lines), 10% test (168 lines). Cell line expression profiles (log TPM) are reduced via PCA retaining 90% variance. Gene embeddings are similarly reduced to 50 dimensions via PCA. Gene symbols serve as categorical features for LightGBM. LightGBM regressor uses RMSE objective, 10,000 estimators, bagging fraction 0.8, feature fraction 0.8, and early stopping at 100 rounds. Results are averaged across 20 random seeds.

E CELL CYCLE STRUCTURE IN PERTURBERT EMBEDDINGS

Setup. We analyze genes from the CCD Transcript task in the gene embedding benchmark and use CCD annotations from the Human Protein Atlas (Pontén et al., 2008). We embed genes using the frozen PerturBERT gene embeddings.

432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485

Table 6: Optimization configuration.

Parameter	Value
Optimizer	AdamW
Learning rate	5×10^{-5}
Adam β_1	0.9
Adam β_2	0.999
Adam ϵ	1×10^{-8}
Weight decay	0.05
Learning rate schedule	Cosine annealing
Warmup steps	200
Gradient clipping (max norm)	1.0

Table 7: Training configuration.

Parameter	Value
Number of epochs	80
Batch size per device	200
Number of GPUs	8
Mixed precision	bfloat16
Torch compile	True (Inductor backend)
Total training time	~21 hours

Visualization and similarity analysis. We compute a 2D UMAP projection of gene embeddings for visualization and compute cosine similarities between all gene pairs. We summarize separation by comparing the distribution of within-CCD cosine similarities to CCD-to-non-CCD cosine similarities.

Results. We visualize the results in Appendix Figure 2. We find that mean within-CCD cosine similarity is 0.051 versus 0.008 for CCD-to-non-CCD pairs.

F BASELINE EMBEDDING COVERAGE

We use published gene embeddings from gene2vec, scGPT, and GenePT (collected in PRESAGE (Littman et al., 2025)). For fair comparison, we restrict evaluation to the 924 genes shared across PerturbBERT, scGPT, and K562 genome-wide Perturb-seq embeddings. Table 8 shows coverage statistics for baseline embeddings.

Table 8: Gene coverage across embedding methods.

Embedding	Total Genes	In L1000	Missing
gene2vec	19,346	930	2
scGPT	60,697	928	4
GenePT	93,799	926	6

G FULL BENCHMARK RESULTS

Table 9 reports complete results across all gene property prediction tasks from the gene embedding benchmark, including ROC-AUC, accuracy, precision, and recall, averaged over k cross-validation splits (see Table caption for task-specific n and k). Overall, PerturbBERT is competitive with scGPT and other embedding baselines across most tasks.

486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539

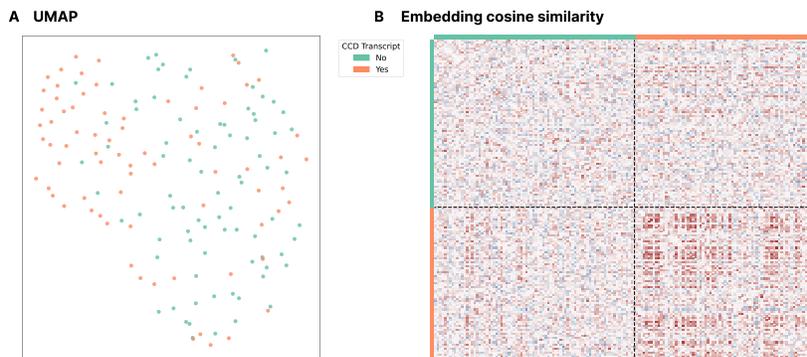


Figure 2: **PerturBERT embeddings capture cell cycle biology.** (A) UMAP visualization of PerturBERT gene embeddings for genes in the CCD Transcript prediction task, colored by cell-cycle dependence annotation. (B) Cosine similarity heatmap of gene embeddings, ordered by CCD status. CCD annotations were derived from the Human Protein Atlas (Pontén et al., 2008).

H FUTURE DIRECTIONS

Extensions include combining co-variation with co-expression pre-training in multi-task objectives, expanding gene vocabulary through additional perturbation platforms, cross-modal alignment with morphology and proteomics, attention variants for whole-transcriptome inputs, and more extensive evaluation across biological domains, including some to highlight differences compared to co-expression derived signatures and their complementarity.

I LIMITATIONS

Gene coverage is constrained by the L1000 platform, and evaluation is restricted to the intersection of available genes across embedding methods (924 genes). Gene coverage can be arbitrarily enlarged through the addition of perturbation datasets with more genes. Masked-gene modeling on discretized targets may discard subtle quantitative effects. We did not model full single-cell distributions, focusing instead on bulk and pseudobulk differential expression. While PerturBERT matches or exceeds state-of-the-art on many tasks, performance gains where they occur are modest in absolute terms.

J USE OF LARGE LANGUAGE MODELS

Large language models (LLMs) were used solely to refine the text and improve readability. They did not contribute substantially to the core research ideas, the study design, or the overall direction of the paper.

Table 9: Complete gene embedding benchmark results. Best ROC-AUC per task in bold, including those within one standard error of the best. n : sample size, k : CV splits.

Task	n	k	Class Size	Class Name	Model	ROC-AUC	Acc.	Prec.	Recall
CCD Protein	106	17	53	Yes	gene2vec	0.804 ± 0.040	0.724 ± 0.052	0.745 ± 0.055	0.706 ± 0.065
				No	GenePT	0.681 ± 0.056	0.578 ± 0.049	0.566 ± 0.056	0.613 ± 0.071
			53	PerturBERT	0.667 ± 0.047	0.618 ± 0.038	0.639 ± 0.051	0.686 ± 0.065	
				scGPT	0.652 ± 0.047	0.590 ± 0.042	0.619 ± 0.050	0.667 ± 0.047	
CCD Transcript	156	20	82	No	PerturBERT	0.837 ± 0.034	0.756 ± 0.025	0.737 ± 0.028	0.913 ± 0.045
				Yes	scGPT	0.813 ± 0.036	0.715 ± 0.040	0.747 ± 0.046	0.753 ± 0.051
			74	GenePT	0.755 ± 0.037	0.682 ± 0.032	0.670 ± 0.033	0.855 ± 0.044	
				gene2vec	0.695 ± 0.033	0.619 ± 0.030	0.664 ± 0.036	0.658 ± 0.049	
Gene2Gene	1623	20	1065	1	scGPT	0.823 ± 0.010	0.809 ± 0.009	0.767 ± 0.019	0.651 ± 0.012
				0	GenePT	0.801 ± 0.007	0.703 ± 0.006	0.729 ± 0.024	0.219 ± 0.016
			558	PerturBERT	0.801 ± 0.011	0.708 ± 0.007	0.788 ± 0.041	0.219 ± 0.015	
				gene2vec	0.782 ± 0.011	0.734 ± 0.010	0.650 ± 0.019	0.495 ± 0.020	
N1 network	69	10	39	nontarget	PerturBERT	0.708 ± 0.066	0.564 ± 0.023	0.100 ± 0.100	0.033 ± 0.033
				activated	scGPT	0.669 ± 0.073	0.633 ± 0.062	0.700 ± 0.093	0.533 ± 0.074
			30	gene2vec	0.581 ± 0.076	0.550 ± 0.058	0.517 ± 0.123	0.333 ± 0.070	
				GenePT	0.564 ± 0.073	0.564 ± 0.007	0.000 ± 0.000	0.000 ± 0.000	
TF vs non-TF	177	20	101	Non-TF	GenePT	0.852 ± 0.033	0.669 ± 0.023	0.641 ± 0.019	0.990 ± 0.010
				TF	gene2vec	0.739 ± 0.035	0.671 ± 0.032	0.706 ± 0.030	0.732 ± 0.042
			76	scGPT	0.635 ± 0.048	0.617 ± 0.042	0.656 ± 0.043	0.713 ± 0.053	
				PerturBERT	0.549 ± 0.038	0.572 ± 0.018	0.576 ± 0.012	0.950 ± 0.020	
Pathology - Cervical	923	8	154	unprognostic	scGPT	0.589 ± 0.017	0.905 ± 0.004	0.863 ± 0.004	0.905 ± 0.004
				prog. favorable	PerturBERT	0.557 ± 0.028	0.927 ± 0.002	0.860 ± 0.003	0.927 ± 0.002
			3	prog. unfavorable	gene2vec	0.512 ± 0.048	0.923 ± 0.003	0.860 ± 0.003	0.923 ± 0.003
				GenePT	0.508 ± 0.027	0.927 ± 0.002	0.860 ± 0.003	0.927 ± 0.002	
Pathology - Endometrial	923	17	151	unprognostic	PerturBERT	0.610 ± 0.021	0.874 ± 0.002	0.765 ± 0.003	0.874 ± 0.002
				prog. unfavorable	scGPT	0.593 ± 0.024	0.821 ± 0.010	0.785 ± 0.009	0.821 ± 0.010
			6	prog. favorable	gene2vec	0.500 ± 0.024	0.868 ± 0.003	0.766 ± 0.003	0.868 ± 0.003
				GenePT	0.464 ± 0.035	0.874 ± 0.002	0.765 ± 0.003	0.874 ± 0.002	
Pathology - Pancreatic	923	15	140	unprognostic	scGPT	0.617 ± 0.026	0.823 ± 0.009	0.794 ± 0.012	0.823 ± 0.009
				prog. unfavorable	PerturBERT	0.612 ± 0.025	0.868 ± 0.001	0.753 ± 0.002	0.868 ± 0.001
			9	prog. favorable	GenePT	0.562 ± 0.025	0.868 ± 0.001	0.753 ± 0.002	0.868 ± 0.001
				gene2vec	0.542 ± 0.029	0.864 ± 0.003	0.757 ± 0.005	0.864 ± 0.003	
Pathology - Renal	923	20	92	unprognostic	scGPT	0.640 ± 0.018	0.500 ± 0.017	0.501 ± 0.017	0.500 ± 0.017
				prog. unfavorable	PerturBERT	0.578 ± 0.010	0.513 ± 0.009	0.424 ± 0.030	0.513 ± 0.009
			27	prog. favorable	GenePT	0.555 ± 0.013	0.513 ± 0.004	0.299 ± 0.022	0.513 ± 0.004
				gene2vec	0.533 ± 0.014	0.471 ± 0.011	0.436 ± 0.021	0.471 ± 0.011	
Pathology - Urothelial	923	9	161	unprognostic	scGPT	0.643 ± 0.025	0.920 ± 0.003	0.864 ± 0.004	0.920 ± 0.003
				prog. favorable	PerturBERT	0.608 ± 0.048	0.927 ± 0.002	0.860 ± 0.003	0.927 ± 0.002
			3	prog. unfavorable	gene2vec	0.554 ± 0.039	0.924 ± 0.002	0.860 ± 0.003	0.924 ± 0.002
				GenePT	0.391 ± 0.024	0.927 ± 0.002	0.860 ± 0.003	0.927 ± 0.002	
RNA blood cell dist.	924	12	121	Detected in all	PerturBERT	0.780 ± 0.008	0.594 ± 0.007	0.460 ± 0.010	0.594 ± 0.007
				Detected in many	scGPT	0.779 ± 0.012	0.592 ± 0.012	0.568 ± 0.010	0.592 ± 0.012
			9	Detected in some	gene2vec	0.630 ± 0.012	0.548 ± 0.011	0.458 ± 0.015	0.548 ± 0.011
				Detected in single	GenePT	0.603 ± 0.010	0.556 ± 0.002	0.333 ± 0.022	0.556 ± 0.002
				3	Not detected				
RNA blood cell spec.	924	12	143	Low specificity	scGPT	0.811 ± 0.012	0.698 ± 0.014	0.666 ± 0.016	0.698 ± 0.014
				Enhanced	PerturBERT	0.804 ± 0.008	0.709 ± 0.004	0.533 ± 0.014	0.709 ± 0.004
			4	Enriched	gene2vec	0.681 ± 0.019	0.688 ± 0.004	0.555 ± 0.012	0.688 ± 0.004
				Not detected	GenePT	0.676 ± 0.020	0.708 ± 0.002	0.501 ± 0.003	0.708 ± 0.002
				2	Group enriched				
RNA blood lineage dist.	924	16	143	Detected in all	scGPT	0.826 ± 0.013	0.688 ± 0.013	0.669 ± 0.012	0.688 ± 0.013
				Detected in many	PerturBERT	0.816 ± 0.013	0.717 ± 0.005	0.575 ± 0.015	0.717 ± 0.005
			5	Detected in some	gene2vec	0.699 ± 0.011	0.693 ± 0.010	0.583 ± 0.017	0.693 ± 0.010
				Detected in single	GenePT	0.674 ± 0.018	0.704 ± 0.003	0.495 ± 0.004	0.704 ± 0.003
				3	Not detected				
RNA blood lineage spec.	924	16	156	Low specificity	scGPT	0.786 ± 0.019	0.733 ± 0.018	0.723 ± 0.018	0.733 ± 0.018
				Lineage enriched	PerturBERT	0.781 ± 0.015	0.767 ± 0.003	0.589 ± 0.005	0.767 ± 0.003
			5	Group enriched	gene2vec	0.714 ± 0.020	0.753 ± 0.006	0.634 ± 0.013	0.753 ± 0.006
				Not detected	GenePT	0.688 ± 0.018	0.767 ± 0.003	0.589 ± 0.005	0.767 ± 0.003
RNA tissue specificity	924	7	126	Low specificity	PerturBERT	0.723 ± 0.006	0.665 ± 0.011	0.586 ± 0.021	0.665 ± 0.011
				Tissue enhanced	scGPT	0.707 ± 0.021	0.635 ± 0.018	0.613 ± 0.015	0.635 ± 0.018
			4	Group enriched	GenePT	0.634 ± 0.023	0.644 ± 0.003	0.559 ± 0.057	0.644 ± 0.003
				Tissue enriched	gene2vec	0.621 ± 0.011	0.623 ± 0.006	0.555 ± 0.008	0.623 ± 0.006
				3					