A174 · Do Perturbation Models Need to See the Perturbation?

Presenter: Danqi Liao — WindMirror

Authors: Danqi Liao, Ann Yu

Session: Day 2 · Fri Oct 2, 2026 · 2:15–4:15 PM

Perturbation response prediction is a central challenge in biology and drug discovery. While reproducing scGPT, a widely used foundation model for perturbation-response prediction, we identified a gene-subsampling issue in its implementation. When the number of genes exceeds the model's maximum sequence length, genes are randomly subsampled, which can remove the perturbed gene from the input. With approximately 5,000 genes and a maximum sequence length of 1,536, this occurs in roughly 70% of sampled training inputs. As a result, the input representation becomes indistinguishable from that of an unperturbed cell, even though the training target remains a perturbed expression profile. In effect, the model often does not see the perturbation during training.

We corrected this behavior by always retaining perturbed genes during subsampling and evaluated the train-mean baseline, original scGPT, and corrected implementation on Adamson, Norman, and Replogle K562. Across five independent runs with different data splits and random seeds, preserving the perturbation signal produced little change across most metrics. Performance was often unchanged or slightly lower, although some measurements improved; for example, Pearson correlation on expression changes across all genes increased 15% on Replogle K562.

These results reveal a surprising disconnect between explicit perturbation conditioning and measured predictive performance. They suggest that current models may rely heavily on shared or average expression structure, that commonly used metrics may be insensitive to perturbation-specific effects, or both. More broadly, our findings highlight the importance of auditing evaluation metrics, datasets, and training-data preparation in addition to modeling choices.

Do Perturbation Models Need to See the Perturbation?

Danqi Liao*, Ann Yu

*Presenting author. WindMirror
danqiliao@windmirror.ai, annyu@windmirror.ai

Perturbation response prediction is a central challenge in biology and drug discovery. Perturbation models are designed to predict transcriptional responses conditioned on the identity of an applied perturbation. While reproducing scGPT [1], one of the most widely used foundation models for perturbation-response prediction, we identified a gene-subsampling issue in its implementation. Specifically, scGPT takes as input genes, gene expression values, and a perturbation mask in which perturbed genes are assigned a value of 1 and unperturbed genes a value of 0. When the number of genes exceeds the model's maximum sequence length, genes are randomly subsampled, which can remove the perturbed gene from the input. With approximately 5,000 genes and a maximum sequence length of 1,536, this occurs in roughly 70% of sampled training inputs. As a result, the input representation becomes indistinguishable from that of an unperturbed cell in most cases, even though the corresponding training target remains the expression values of a perturbed cell. In effect, the perturbation model does not see the perturbation during training.

We fixed this bug by always retaining the perturbed genes during subsampling and evaluated the train mean baseline, original scGPT and corrected implementations on Adamson [2], Norman [3], and Replogle K562 datasets [4]. In addition to standard Pearson correlation metrics on expression changes, we evaluated perturbation distance correlation and direction accuracy. Pearson on perturbation distance measures the correlation between ground truth perturbation distances and the predicted distance, with perturbation distance defined as the L_2 distance between perturbation mean expressions. Across five independent runs with different data splits and random seeds, preserving the perturbation signal produced little change across most metrics (Figure 1). Performance was often approximately unchanged or slightly lower, although some measurements improved; for example, Pearson delta across all genes increased 15% on Replogle K562.

These results reveal a surprising disconnect between explicit perturbation conditioning and measured predictive performance. They suggest that current models may rely heavily on shared or average expression structure, that commonly used metrics may be insensitive to perturbation-specific effects, or both. More broadly, our findings highlight the importance of auditing evaluation metrics, datasets, and training-data preparation, in addition to modeling choices, when evaluating models for perturbation response prediction.

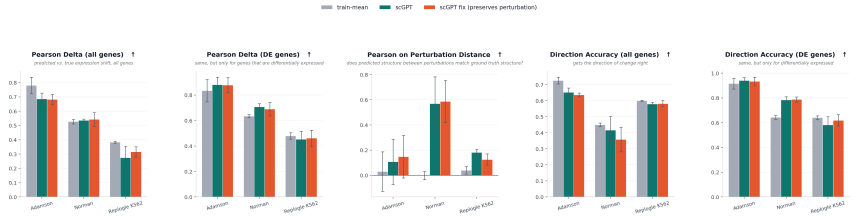

Figure 1: Mean performance across five independent runs with different data splits and random seeds. Preserving the perturbation signal has little effect on most standard benchmark metrics.

References

[1] Cui, Haotian, et al. "scGPT: toward building a foundation model for single-cell multi-omics using generative AI." Nature methods 21.8 (2024): 1470-1480.

[2] Adamson, Britt, et al. "A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response." Cell 167.7 (2016): 1867-1882.

[3] Norman, Thomas M., et al. "Exploring genetic interaction manifolds constructed from rich single-cell phenotypes." Science 365.6455 (2019): 786-793.

[4] Replogle, Joseph M., et al. "Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq." Cell 185.14 (2022): 2559-2575.