# Prospective Evaluation of Foundation Model Performance in Precision Medicine

*This manuscript (permalink) was automatically generated from greenelab/fm-pm-eval-manuscript@7fec83f on June 20, 2026.*

## Authors


- **Lucas A. Gillenwater** ✉
  0000-0002-6995-0130 · lagillenwater
  Department of Biomedical Informatics, University of Colorado School of Medicine, Aurora, CO, USA · Funded by R01 HD109765

- **Casey S. Greene** ✉
  0000-0001-8713-9213 · cgreene · GreeneScientist
  Department of Biomedical Informatics, University of Colorado School of Medicine, Aurora, CO, USA; Center for Health AI, University of Colorado School of Medicine, Aurora, CO, USA

✉ — Correspondence possible via GitHub Issues or email to Lucas A. Gillenwater <lucas.gillenwater@cuanschutz.edu>, Casey S. Greene <casey.s.greene@cuanschutz.edu>.

# Abstract

# Introduction

The most recent iteration of AI models ('foundation' and 'world' models) is exciting, and the field is constantly putting out newer, larger models; however, the models do not generalize to out-of-distribution tests and do not outperform simpler models across tasks. For example, Steiner et al. [1] and Ahlmann-Eltze et al. [2] both reported that linear baselines outperformed single-cell foundation models [3,4,5,6,7] on downstream tasks using data the models had not yet seen. The Virtual Cell Challenge in 2025 found similar results on a crowd sourced model evaluation for prediction in an unseen stem cell context [8]. Translation to personalized medicine is an even more difficult goal. The prediction sets are out-of-distribution of the training data (i.e., transcriptional profiles from observational samples or interventions on immortalized cell lines). Therefore, generalization is the bar these models must meet to impact personalized medicine.

This proposal creates the evaluation framework to prospectively evaluate personalized medicine applications for foundation models, with a particular focus in precision oncology. We are recruiting collaborators to contribute across the evaluation framework: new tasks and held-out data tranches, foundation or world models to benchmark, and evaluation harnesses that adapt scoring to new phenotypes and modalities, all to test out-of-distribution generalization. Our goal is a precision medicine "acid test" [9] for foundation models, encouraging model builders and users to demonstrate their performance on the prospective releases of test data tranches. We will pair the continual evaluation results from the companion repository with the collaborative creation of a benchmarking manuscript.

# Data tranches for model evaluation

Each tranche predicts response from pre-treatment expression.

**Table 1:** Data tranches for evaluating foundation model generalization to personalized medicine. Each tranche predicts drug response from pre-treatment expression, with a defined held-out axis and a leakage assessment.

| Tranche | Data source | Input | Output | Prediction expectation | Held-out axis | Leakage |
|---|---|---|---|---|---|---|
| **Pre-Tranche** Retrospective community data | Retrospective PDTO and cell line drug response datasets [10]. | Pre-treatment expression and the compound identity. | Drug response as one fixed sensitivity metric (dose-response AUC). | Predict across cohorts, drugs, organoids/subtype, and protocols. | Held-out cohort, drug, organoid/subtype, model system. | Need to assess. Transcriptomic models may train on perturbation data from cell lines. |

| Tranche | Data source | Input | Output | Prediction expectation | Held-out axis | Leakage |
|---|---|---|---|---|---|---|
| **Tranche 1** CRC drug PDO lines | CRC drug screen: 2 cell lines and 9 patient organoids, 100 compounds, pre-treatment expression per line. | Pre-treatment expression of the line, plus compound identity. | Drug response as one fixed sensitivity metric (dose-response AUC). | Given a line's baseline expression and a compound, predict that line's response. | Held-out compound, organoid, drug. | Retrospective but unpublished, so unseen. Lock splits and predictions before unblinding. |
| **Tranche 2** ER+ breast, mechanism | Oliphant ER+ patient-derived xenograft organoids in BTOM-ER medium [11], with pre-treatment expression. | Pre-treatment expression, plus the ER-pathway perturbation (fulvestrant dose, or estrogen withdrawal). | Response as CellTiter-Glo viability dose-response, same metric as Tranche 1. | Predict response to the ER-pathway perturbation, and test whether the prediction tracks the ER-dependence mechanism. | Held-out mechanism: an ER-pathway or resistance state not in training. Assess explainability of predictions. | Need to assess. Existing and curated mechanism data. May leak into training. |
| **Tranche 3** Prospective experiment | A prospective organoid experiment designed and run after predictions are locked. | Pre-treatment expression and compound identity. | Drug response as one fixed sensitivity metric (dose-response AUC). | Predict response on a not-yet-run, sealed experiment. | New patients and new contexts, tested prospectively. | Prospective and sealed. No leakage possible. Predictions registered before the assay. |
| **Future Periodic Tranches** | Prospective experiments and multi-lab datasets across multiple personalized medicine modalities. | Pre-treatment expression and the compound identity. | Perturbation response. The particular phenotype is a moving target requiring a proper evaluation adapter. | Predict response on a not-yet-run, sealed experiment or across cohorts and centers contributed by the community. | New patients and new contexts, tested prospectively, or new centers and new cohorts. Held-out mechanisms. | Prospective and sealed or unpublished. |

**Data.** Pre-perturbation state (e.g., transcriptome). Add other modalities or clinical features in the future.

**Benchmarks.** Baseline linear and nonlinear statistical models on expression. Foundation and world models (STACK, STATE, X-Cell). Reported prediction performance motivates future model inclusion.

**Controls.** Negative: shuffled response among organoids within each drug. Positive: injected perturbation-specific effects to recapitulate.

**Scoring.** Perturbation by substrate interaction. For example, does the model predict that an organoid responds to a drug?

**Community call.** Find interested collaborators with unpublished data to assess prior to publication through social media platforms like LinkedIn. Include collaborators in Manubot-style evaluation for interpretation of results.

# References


1. **A Systematic Evaluation of Single-Cell Foundation Models on Cell-Type Classification Task**
Nicolas Steiner, Ziteng Li, Omid Vosoughi, Johanna Schrader, Soumyadeep Roy, Wolfgang Nejdl, Ming Tang
*Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining* (2025-03-10) https://doi.org/hb782r
DOI: 10.1145/3701551.3708811

2. **Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines**
Constantin Ahlmann-Eltze, Wolfgang Huber, Simon Anders
*Nature Methods* (2025-08) https://doi.org/g9w5p7
DOI: 10.1038/s41592-025-02772-6 · PMID: 40759747 · PMCID: PMC12328236

3. **scGPT: toward building a foundation model for single-cell multi-omics using generative AI**
Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, Bo Wang
*Nature Methods* (2024-02-26) https://doi.org/gtkxpk
DOI: 10.1038/s41592-024-02201-0 · PMID: 38409223

4. **Transfer learning enables predictions in network biology**
Christina V Theodoris, Ling Xiao, Anant Chopra, Mark D Chaffin, Zeina R Al Sayed, Matthew C Hill, Helene Mantineo, Elizabeth M Brydon, Zexian Zeng, XShirley Liu, Patrick T Ellinor
*Nature* (2023-05-31) https://doi.org/gr9x63
DOI: 10.1038/s41586-023-06139-9 · PMID: 37258680 · PMCID: PMC10949956

5. **Large-scale foundation model on single-cell transcriptomics**
Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, Le Song
*Nature Methods* (2024-06-06) https://doi.org/gtxwwj
DOI: 10.1038/s41592-024-02305-7 · PMID: 38844628

6. **Predicting transcriptional outcomes of novel multigene perturbations with GEARS**
Yusuf Roohani, Kexin Huang, Jure Leskovec
*Nature Biotechnology* (2023-08-17) https://doi.org/gtb96r
DOI: 10.1038/s41587-023-01905-6 · PMID: 37592036 · PMCID: PMC11180609

7. **scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data**
Fan Yang, Wenchuan Wang, Fang Wang, Yuan Fang, Duyu Tang, Junzhou Huang, Hui Lu, Jianhua Yao
*Nature Machine Intelligence* (2022-09-26) https://doi.org/grqzjp
DOI: 10.1038/s42256-022-00534-z

8. **Virtual Cell Challenge 2025 Wrap-Up: Winners and Reflections**
Arc Institute
(2025) https://arcinstitute.org/news/virtual-cell-challenge-2025-wrap-up

9. **Acid2**
Wikipedia
https://en.wikipedia.org/wiki/Acid2

10. **CoderData: Cancer Omics and Drug Experiment Response Data**
Sara Gosline, Yannick Mahlich, Jeremy Jacobson, Marissa Andrews

*PNNL-CompBio* (2025) https://pnnl-compbio.github.io/coderdata/index.html

11. **Establishing conditions for the generation and maintenance of estrogen receptor-positive organoid models of breast cancer**
Michael UJ Oliphant, Dipikaa Akshinthala, Senthil K Muthuswamy
*Breast Cancer Research* (2024-03-29) https://doi.org/hb782s
DOI: 10.1186/s13058-024-01798-6 · PMID: 38553763 · PMCID: PMC10979603