



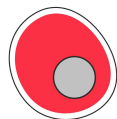
2024 Datasets and Benchmarks Track

# A benchmark for prediction of transcriptomic responses to chemical perturbations across cell types

Artur Szalata\* · Andrew Benz\* · Robrecht Cannoodt · Mauricio Cortes · Jason Fong · Sunil Kuppasani · Richard Lieberman · Tianyu Liu · Javier Mas-Rosario · Rico Meinl · Jalil Nourisa · Jared Tumiel · Tin M. Tunjic · Mengbo Wang · Noah Weber · Hongyu Zhao · Benedict Anchang · Fabian J. Theis+ · Malte D. Luecken+ · Daniel Burkhardt+

[openproblems.bio/results/perturbation\\_prediction](https://openproblems.bio/results/perturbation_prediction)

**HELMHOLTZ  
MUNICH**



**Open Problems** in  
Single-Cell Analysis

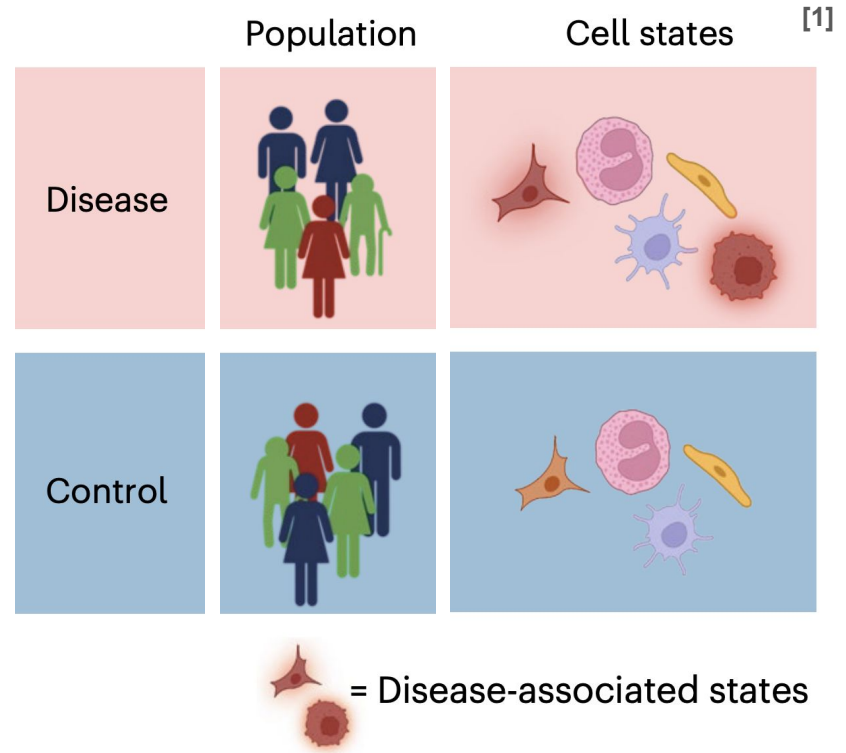
**Cellarity**  
A Flagship Pioneering Company



**Chan  
Zuckerberg  
Initiative**

# Single-cell RNA sequencing has the potential to unlock novel treatments for complex diseases

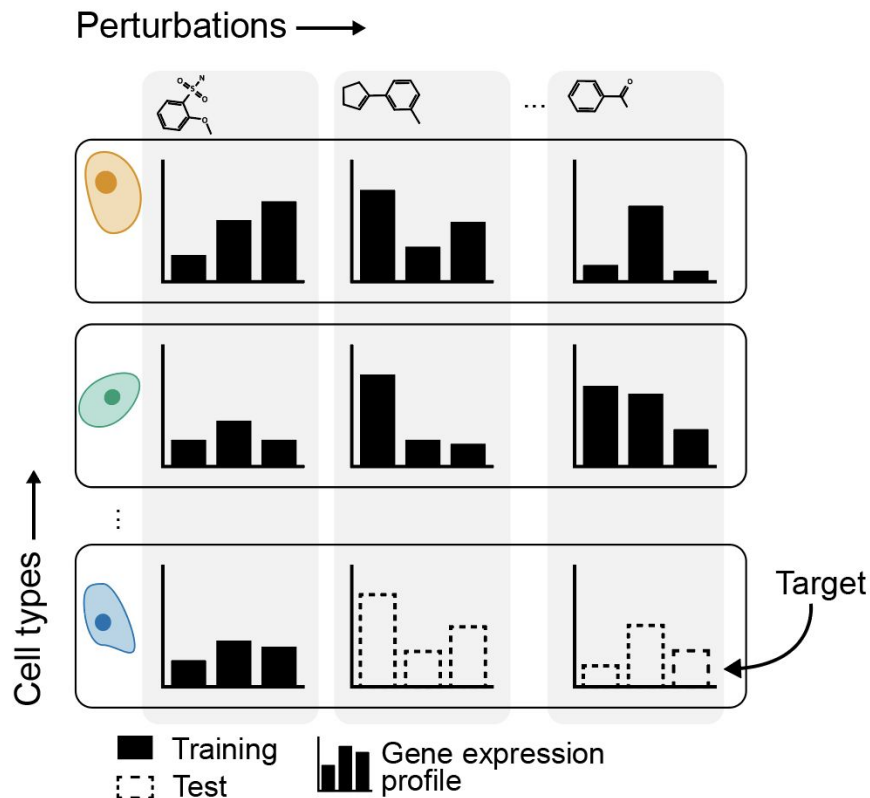
- Identification of transitions from disease to healthy cell states
- This information can be used for drug screening (inducing disease-to-healthy transitions)
- This could treat complex diseases that involve multiple biological pathways



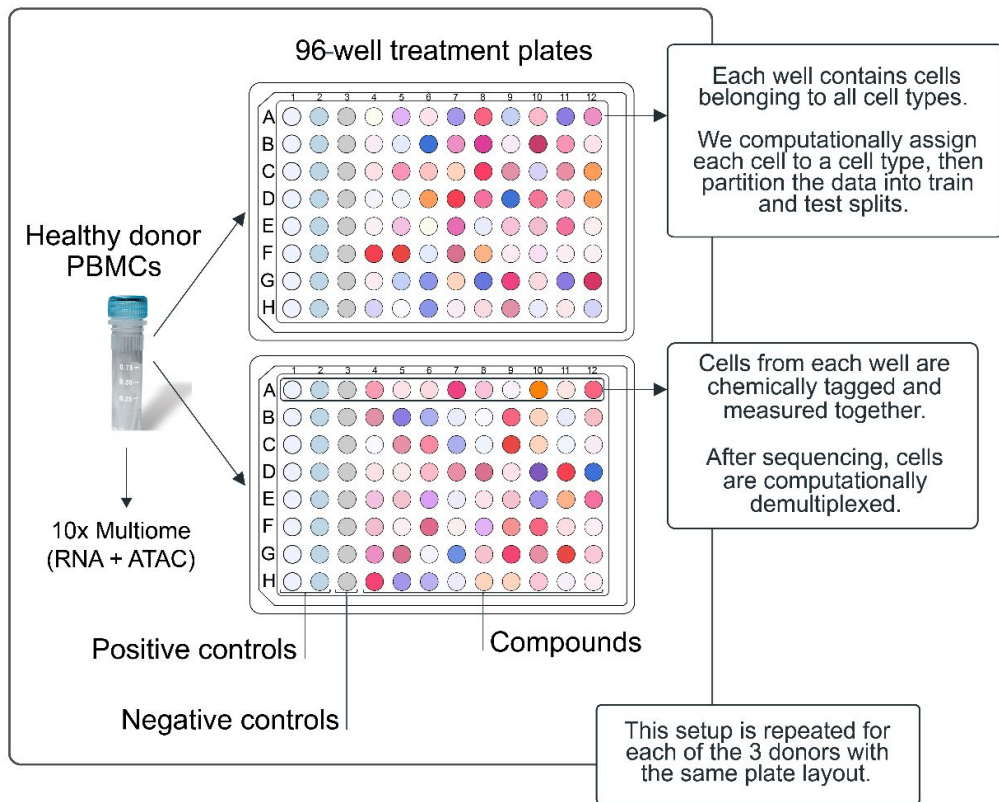
[1] Adapted from: Dann, Emma, et al. "Precise identification of cell states altered in disease using healthy single-cell references." *Nature Genetics* (2023)

# Perturbation prediction makes screening compounds against a change in gene expression tractable





- Goal: understand how chemical perturbations impact gene expression
- Challenge: biological and chemical space are both very large
- Solution: measure a fraction of possible perturbations and infer the rest
- However, existing perturbation datasets are limited by size and data quality issues



# A new single-cell dataset: 146 drug perturbations in human peripheral blood cells from 3 donors with 4 distinct cell types



## Cell types

-  B cells
-  Myeloid cells
-  NK cells
-  T cells

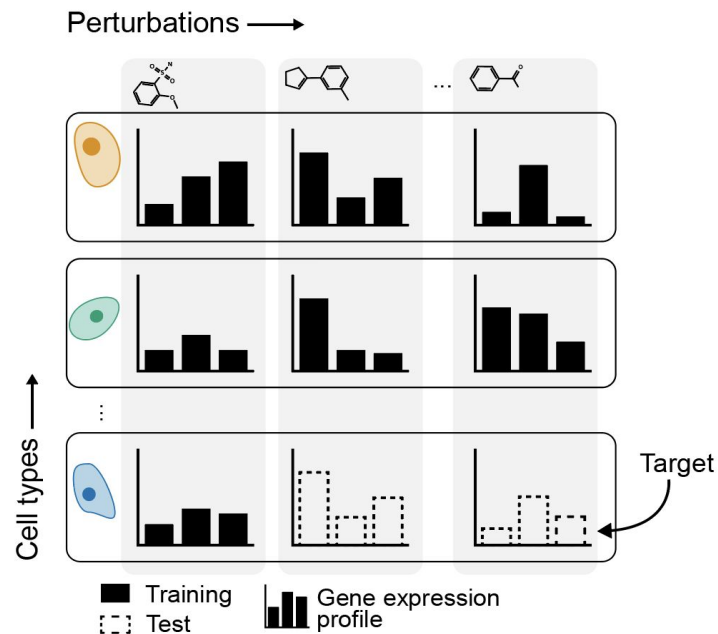
## Donors

-  Donor 1 ♀
-  Donor 2 ♂
-  Donor 3 ♂



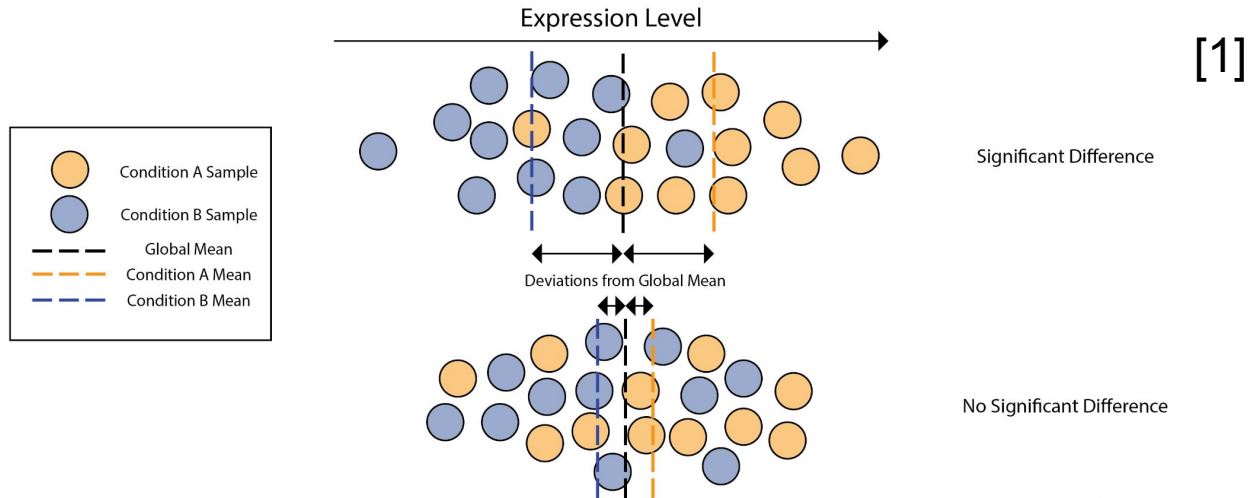
# We developed a robust benchmark around this new dataset

- Task: predict perturbation effects for held out (cell type, compound) pairs
- Perturbation effects are derived from a generalized linear model contrasting treatment and control conditions
- p-value: probability that observed effects in treatment vs. control are random
- fold-change: magnitude of effect difference between treatment and control



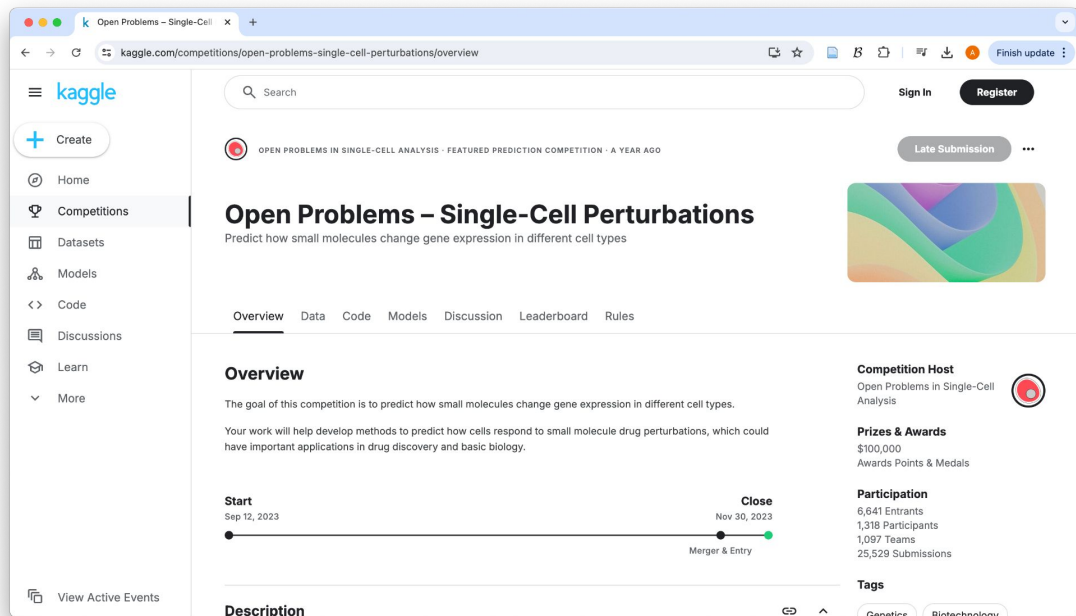
# Perturbation representation

- Perturbation effect is measured with a generalized linear model
- We propose cross-donor retrieval to evaluate the representation of perturbation effects
- Best representation:  $-\log_{10}(p\text{-value}) \times \text{sign}(\log\text{-fold change})$



# Competition on Kaggle based on this benchmark

- Sourcing effective models for perturbation prediction from Kaggle competitors
- Improving dataset and benchmark based on competitor feedback

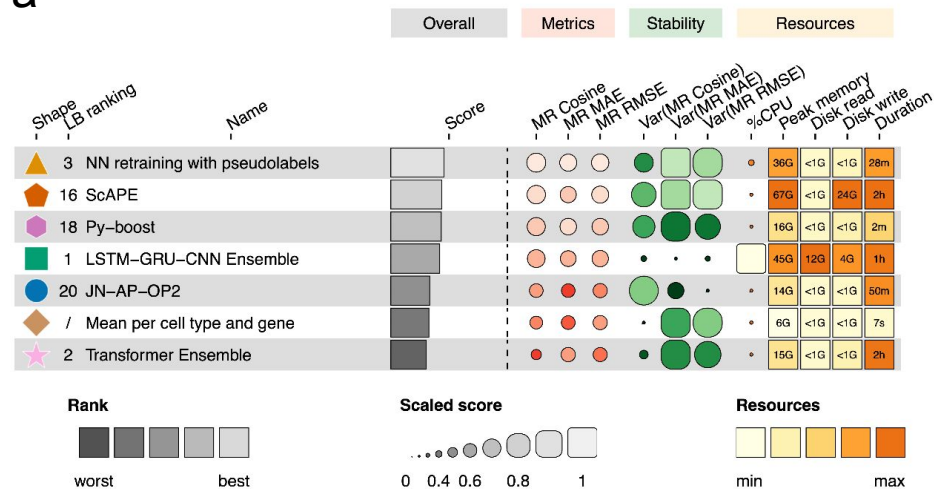


The screenshot shows the Kaggle website interface for a competition titled "Open Problems - Single-Cell Perturbations". The page is viewed in a browser window with the URL `kaggle.com/competitions/open-problems-single-cell-perturbations/overview`. The left sidebar contains navigation options: Home, Competitions, Datasets, Models, Code, Discussions, Learn, and More. The main content area features a search bar, a "Sign In" button, and a "Register" button. The competition title is prominently displayed, along with a subtitle: "Predict how small molecules change gene expression in different cell types". Below the title, there are tabs for "Overview", "Data", "Code", "Models", "Discussion", "Leaderboard", and "Rules". The "Overview" tab is selected, showing the competition's goal: "Your work will help develop methods to predict how cells respond to small molecule drug perturbations, which could have important applications in drug discovery and basic biology." A timeline indicates the competition started on "Sep 12, 2023" and ends on "Nov 30, 2023", with a "Merger & Entry" point. On the right side, the "Competition Host" is identified as "Open Problems in Single-Cell Analysis". The "Prizes & Awards" section lists a prize of "\$100,000" and "Awards Points & Medals". The "Participation" section shows "6,641 Entrants", "1,318 Participants", "1,097 Teams", and "25,529 Submissions". At the bottom, there are "Tags" for "Genetics" and "Biotechnology".

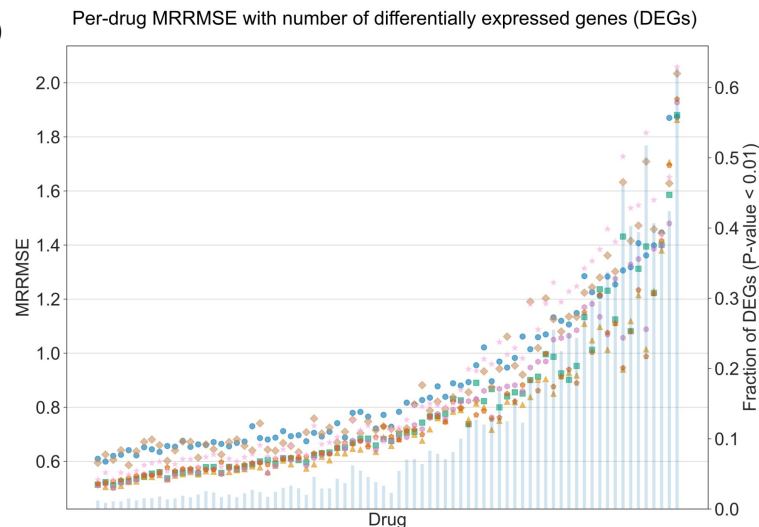
# A new benchmarking platform: Open Problems Perturbation Prediction (OP3)

- Top models from Kaggle competition were implemented for benchmark
- Dataset bootstrapping to assess robustness
- Key findings:
  - simple models tend to outperform more complex ones
  - drugs with larger effects are more difficult to predict correctly

a



b

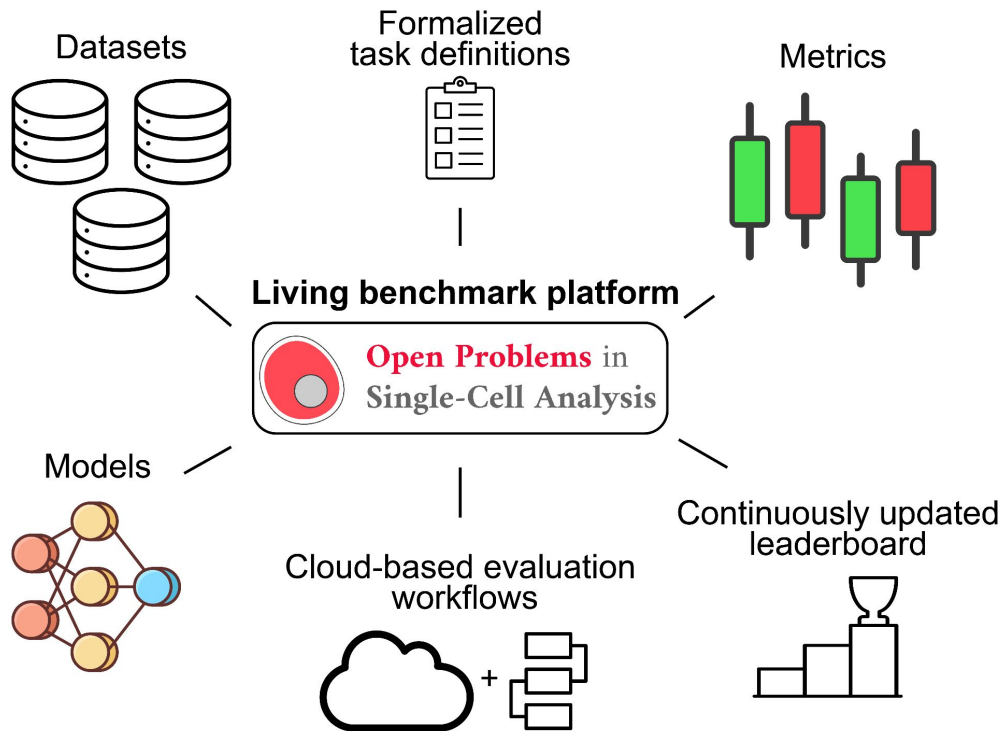


Glossary: Mean Rowwise (MR), Root Mean Squared Error (RMSE), Differentially Expressed Genes (DEGs)



# The OP3 benchmarking platform is a first step towards developing successful perturbation prediction models

- Best model predictions are still far from ground truth
- New methods can easily be added to the benchmarking platform via GitHub PRs
- More data is clearly needed, especially for the task of extrapolating across unseen chemical structures



## Acknowledgements

Chan  
Zuckerberg  
Initiative 

kaggle

 SaturnCloud

kaggle  
competitors

Cellarity   
A Flagship Pioneering Company

Yale



HELMHOLTZ  
MUNICH 

OLDEN LABS™

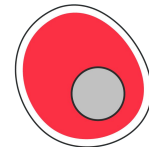
Retro  
BIOSCIENCES

 PURDUE  
UNIVERSITY®

TU  
WIEN

NIH 

  
Helmholtz-Zentrum  
hereon



Open Problems in  
Single-Cell Analysis