

Predicting Transcriptional Responses to Single-Gene Perturbations via Set-Based Transformers

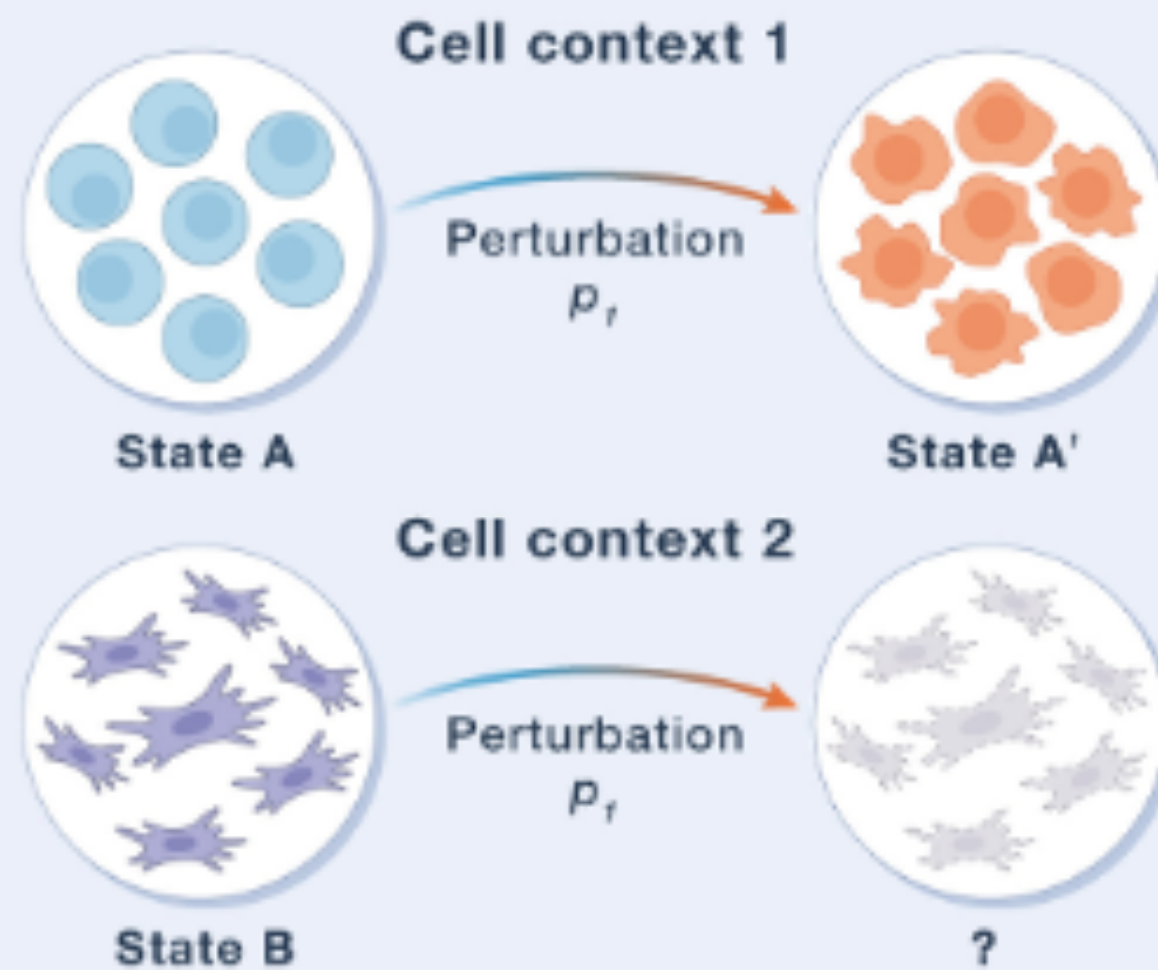
A Population-Centric Approach to the Virtual Cell Challenge

Wang Anrui

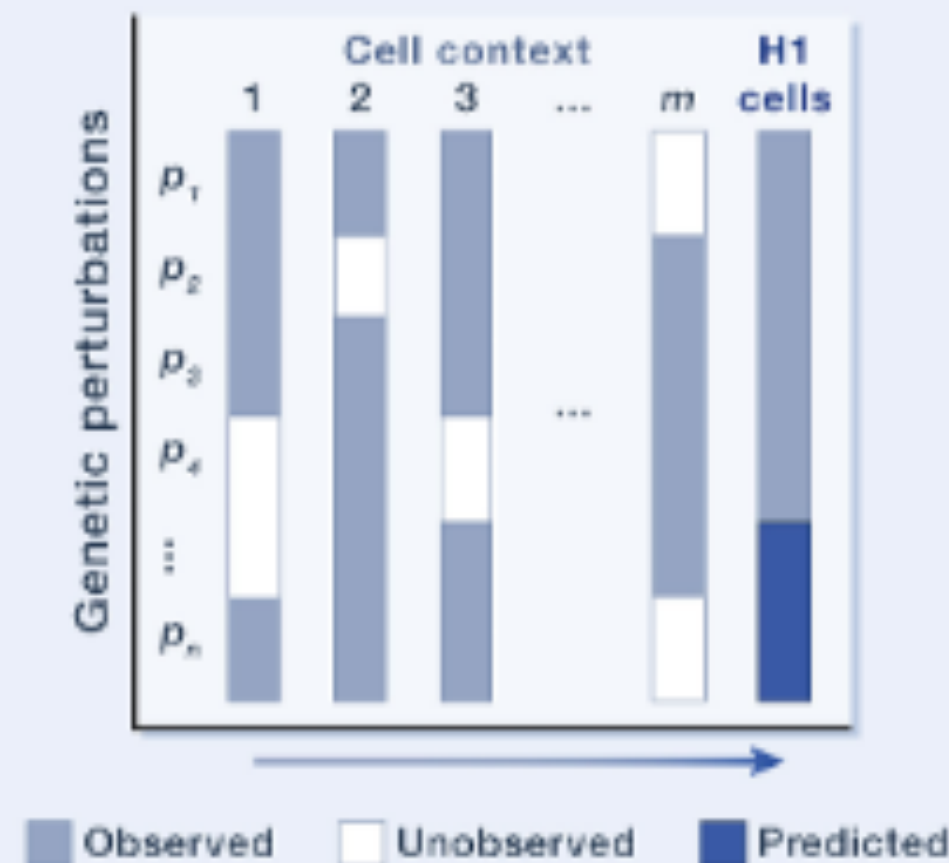
Problem Definition & Background

Virtual Cell Challenge

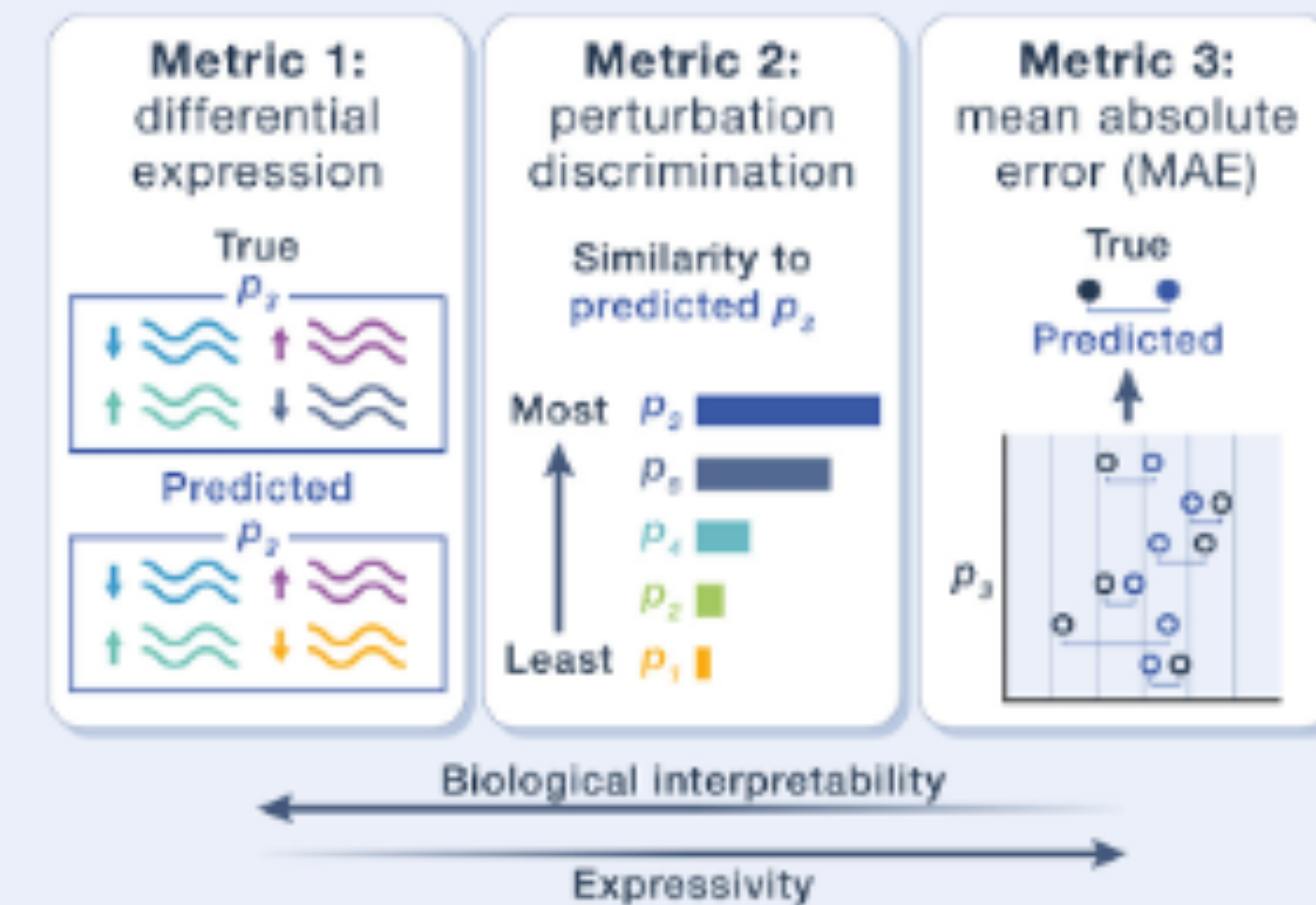
A CHALLENGE effect prediction



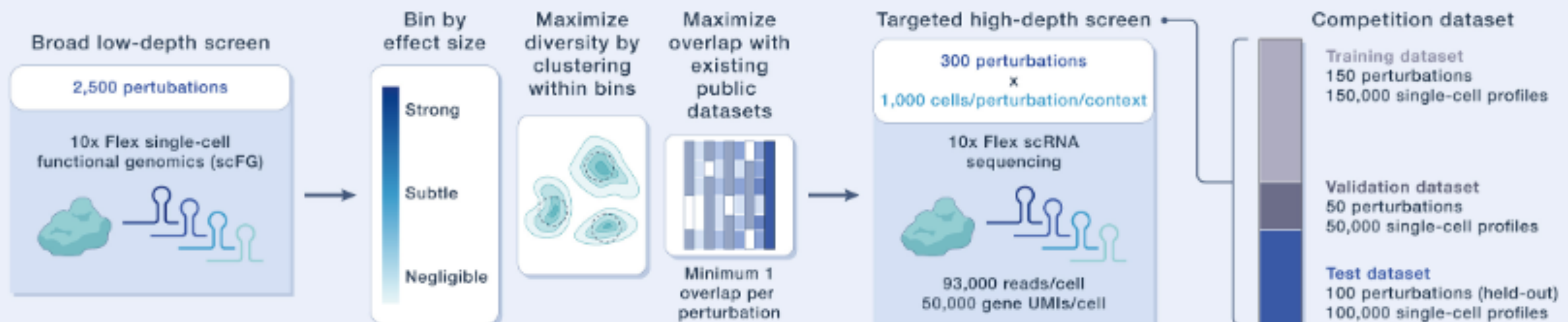
B TASK context generalization



C MODEL performance metrics



D DATA generation and competition breakdown



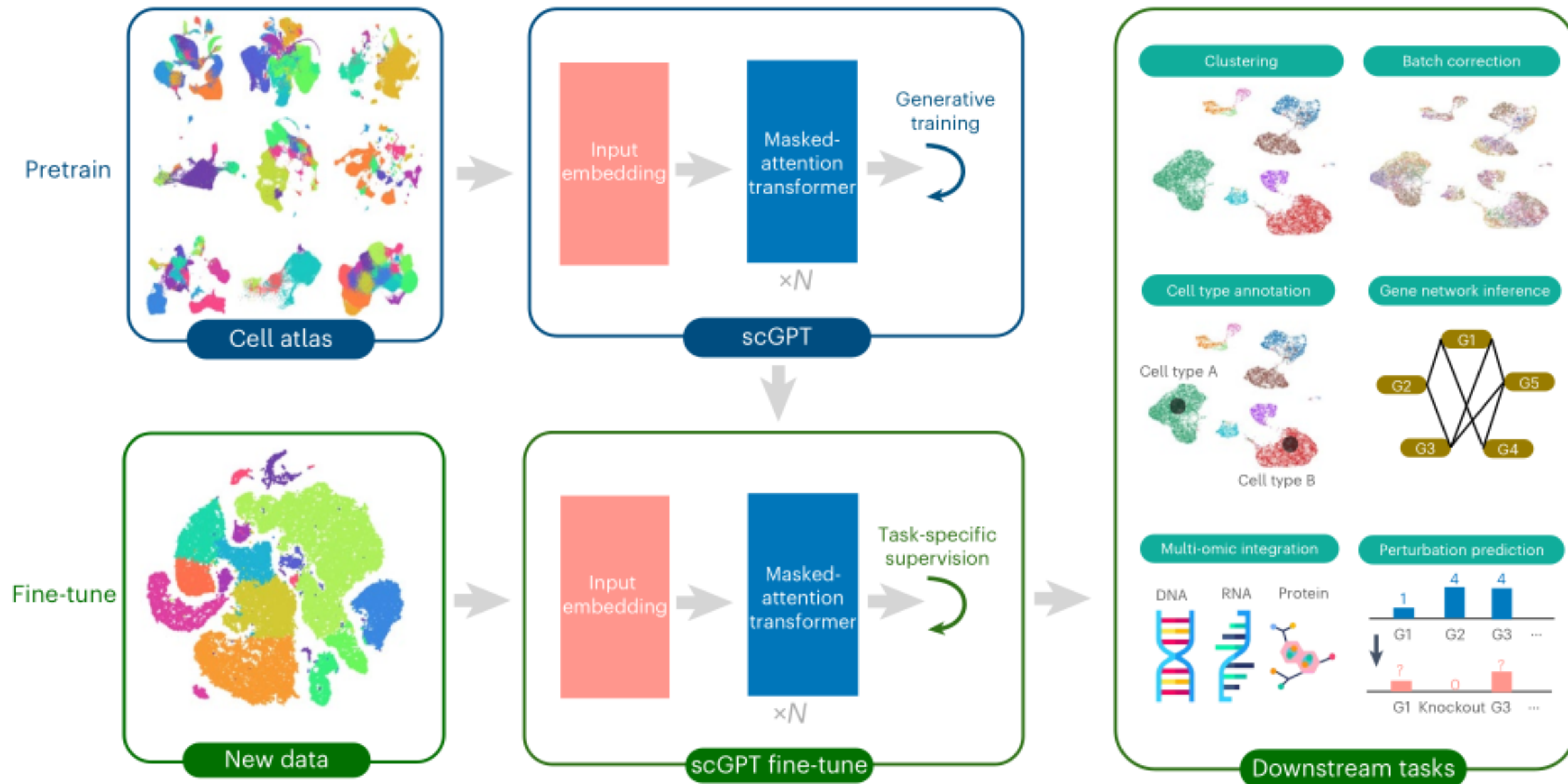
Dataset & Input Characteristics

G/C	C1	C2	C3	...	Cn
G1	2	5	0	...	8
G2	0	19	40	...	0
G3	230	0	398	...	0
....
Gm	0	0	0	3000

- In a single piece of data:
- m in the scale of 18K, sparse matrix with a lot of zeros
- n in the scale of 183K
- 150 gene perturbations in total

Methodology

a



Evaluation & Expected Outcomes

Baselines:

- Linear Models, GEARS (GNN-based), and standard scGPT(Foundation Model).

Biological Metrics (VCC Metrics):

- Differential Expression Score (DES): Accuracy of identifying up/down-regulated genes.
- Perturbation-Discrimination Score (PDS): Ability to distinguish specific perturbation signatures from background noise.
- Mean Absolute Error(MAE): Evaluates the overall accuracy of the predicted expression profile