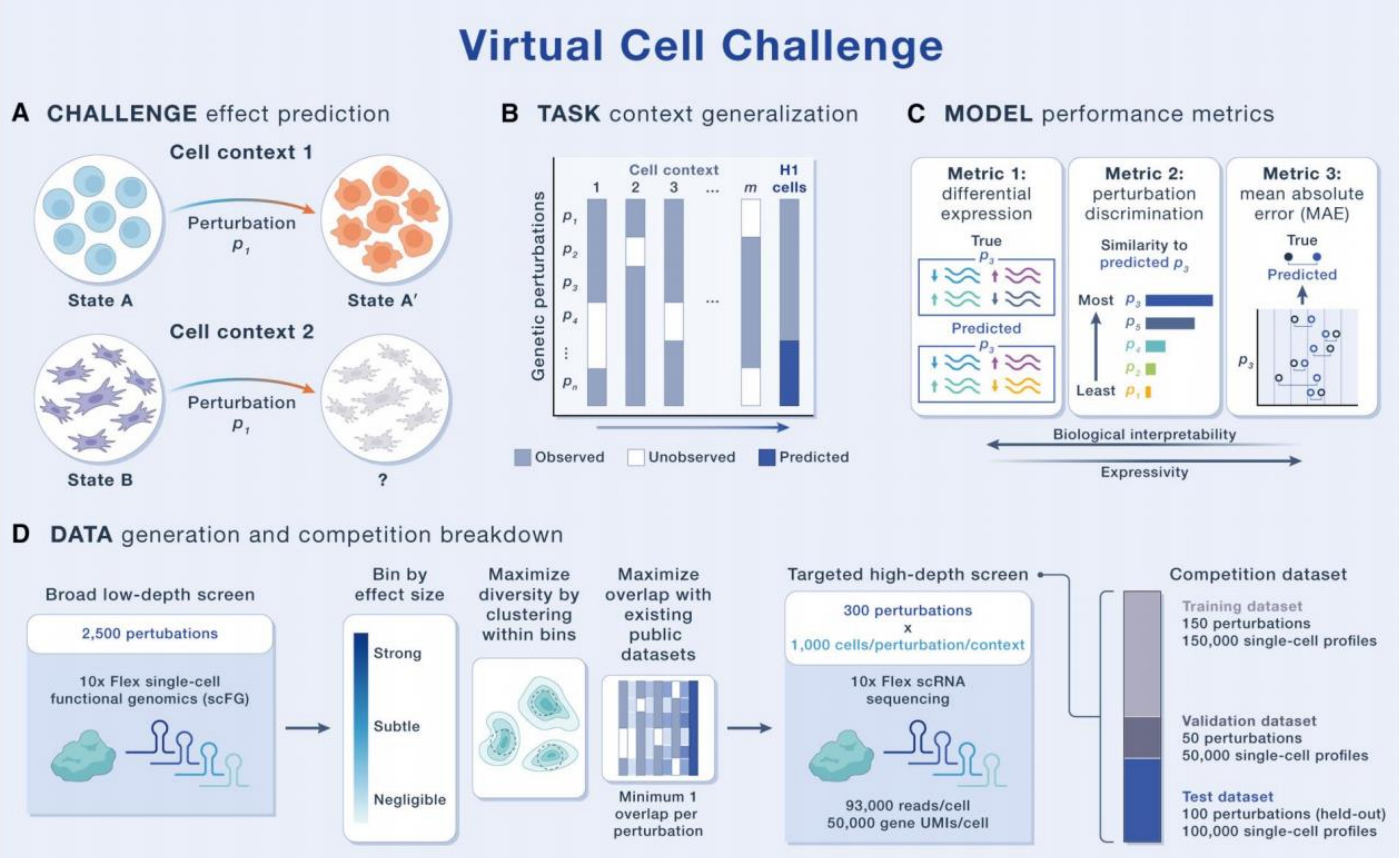
Predicting the Response of Cellular Transcriptome to Gene Perturbations

Anrui Wang, Jiawen Dai, Yetong Teng, Yanzhi Yin

Background

- Predicting cellular responses to genetic or chemical perturbations is challenging
- Advances in single-cell technologies and machine learning have enabled progress
- Generalization to unseen perturbations remains an open problem
- Virtual Cell Challenge (VCC)
  - Launched by the Arc Institute
  - ~300,000 human embryonic stem cells
  - Goal: develop models that generalize beyond observed perturbations

Based on single-cell RNA seq data, predict gene expression changes under unknown perturbations.



Virtual Cell Challenge
A CHALLENGE effect prediction
Cell context 1
Perturbation p_1
State A
State A′
Cell context 2
Perturbation p_1
State B
?
B TASK context generalization
Cell context
1
2
3
...
m
H1 cells
Genetic perturbations
p_1
p_2
p_3
p_4
p_n
Observed
Unobserved
Predicted
C MODEL performance metrics
Metric 1: differential expression
True
p_3
Predicted
p_3
Metric 2: perturbation discrimination
Similarity to predicted p_3
Most
p_3
p_5
p_4
p_2
Least
p_1
Metric 3: mean absolute error (MAE)
True
Predicted
p_3
Biological interpretability
Expressivity
D DATA generation and competition breakdown
Broad low-depth screen
2,500 pertubations
10x Flex single-cell functional genomics (scFG)
Bin by effect size
Strong
Subtle
Negligible
Maximize diversity by clustering within bins
Maximize overlap with existing public datasets
Minimum 1 overlap per perturbation
Targeted high-depth screen
300 perturbations
x
1,000 cells/perturbation/context
10x Flex scRNA sequencing
93,000 reads/cell
50,000 gene UMIs/cell
Competition dataset
Training dataset
150 perturbations
150,000 single-cell profiles
Validation dataset
50 perturbations
50,000 single-cell profiles
Test dataset
100 perturbations (held-out)
100,000 single-cell profiles

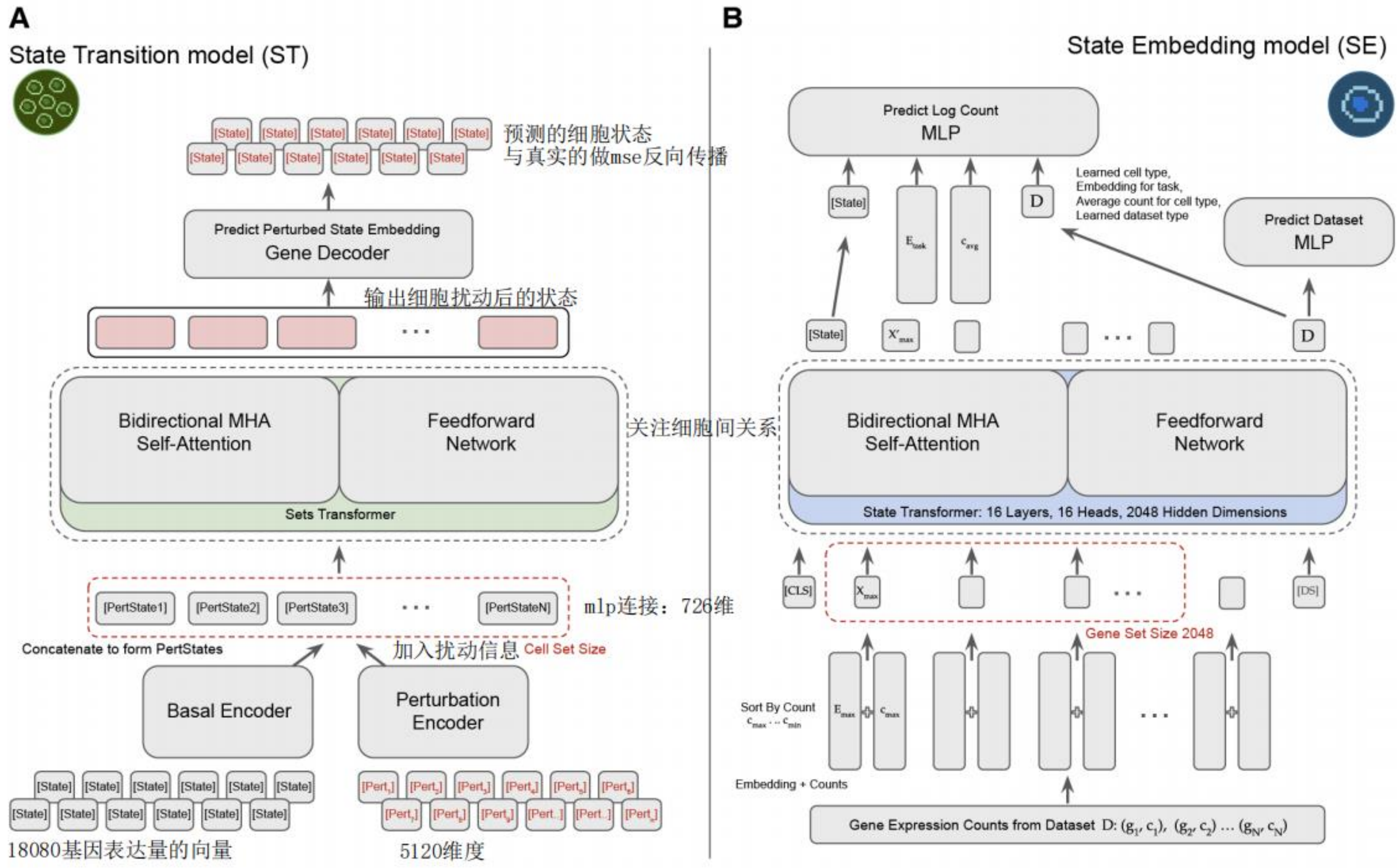
Dataset Overview

- Source: Virtual Cell Challenge 2025
- Cell Type: H1 human embryonic stem cells (hESCs)
- Dimensions: 221,273 cells × 18,080 genes
- Perturbations: 150 target genes (120 training, 30 testing)
- Controls: 38,176 non-targeting/control cells

Data Distribution Challenges

- Severe imbalance in perturbation samples:
  - Some perturbations have >1,000 cells.
  - Many perturbations have <500 cells.
- Large variation in DEG (Differentially Expressed Genes) counts:
  - Different perturbations yield drastically different numbers of DEGs.
  - Low signal examples: MED13 (204 DEGs: 70 up / 134 down).
  - High signal examples: KDM1A (3,053 DEGs: 2,194 up / 859 down).
  - This heterogeneity complicates model training and evaluation.

STATE Model Architecture



A

State Transition model (ST)

[State] [State] [State] [State] [State] [State]
[State] [State] [State] [State] [State] [State]

预测的细胞状态
与真实的做mse反向传播

Predict Perturbed State Embedding
Gene Decoder

输出细胞扰动后的状态

Bidirectional MHA Self-Attention

Feedforward Network

Sets Transformer

关注细胞间关系

[PertState1] [PertState2] [PertState3] ··· [PertStateN]

mlp连接：726维

Concatenate to form PertStates

加入扰动信息 Cell Set Size

Basal Encoder

Perturbation Encoder

[State] [State] [State] [State] [State] [State]
[State] [State] [State] [State] [State] [State]

[$Pert_1$] [$Pert_2$] [$Pert_3$] [$Pert_4$] [$Pert_5$] [$Pert_6$]
[$Pert_7$] [$Pert_8$] [$Pert_9$] [$Pert_{10}$] [$Pert_{11}$] [$Pert_{12}$]

18080基因表达量的向量

5120维度

B

State Embedding model (SE)

Predict Log Count
MLP

[State] E_{task} c_{avg} D

Learned cell type,
Embedding for task,
Average count for cell type,
Learned dataset type

Predict Dataset
MLP

[State] X'_{max} ··· D

Bidirectional MHA Self-Attention

Feedforward Network

State Transformer: 16 Layers, 16 Heads, 2048 Hidden Dimensions

[CLS] X_{max} ··· [DS]

Gene Set Size 2048

Sort By Count
$c_{max} \cdots c_{min}$

E_{max} + c_{max} ···

Embedding + Counts

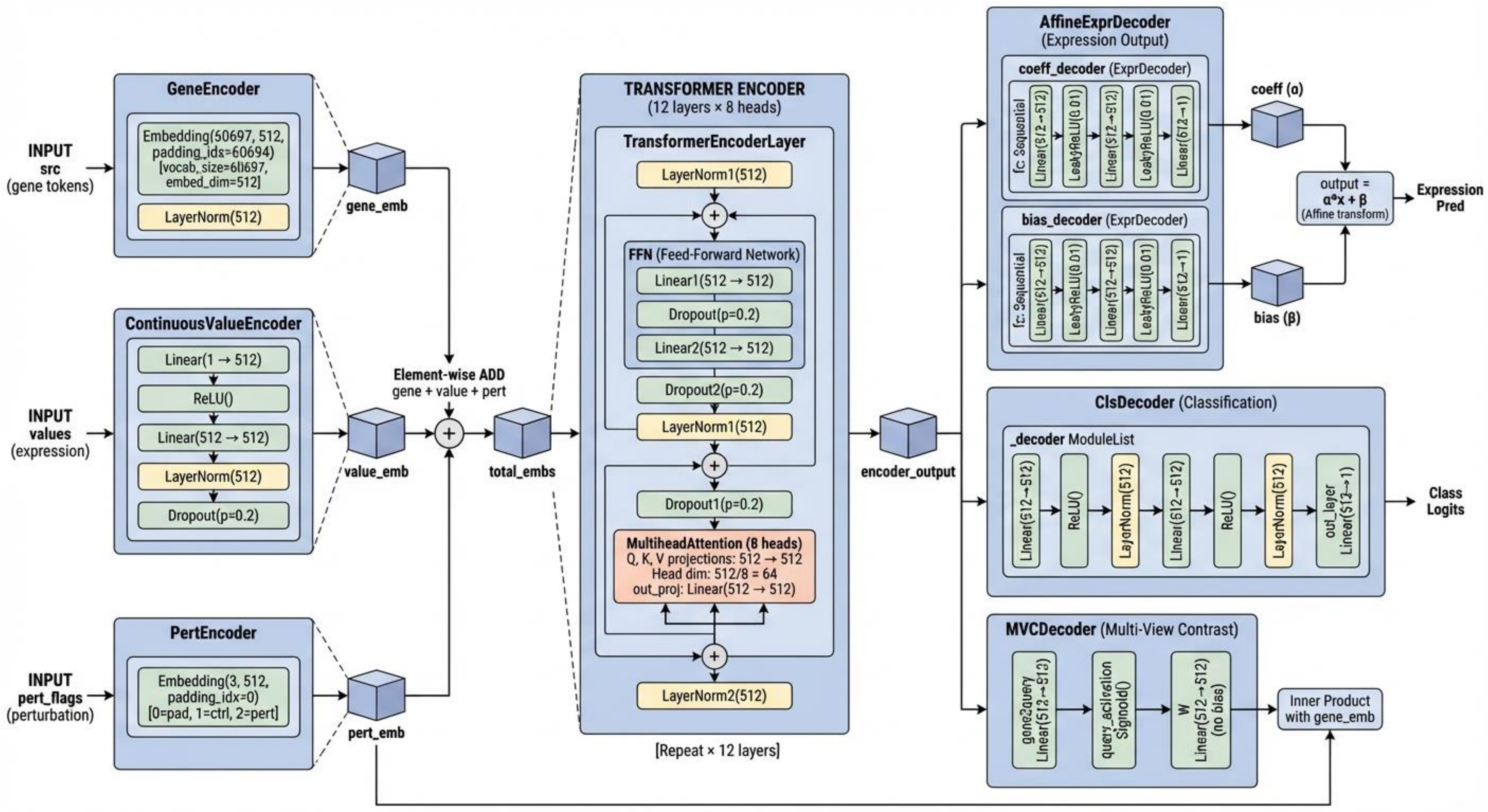
Gene Expression Counts from Dataset D: $(g_1, c_1), (g_2, c_2) \dots (g_N, c_N)$

STATE Model Details

Goal: Model how genetic perturbations reshape single-cell transcriptomes.

- SE (State Embedding):
  - Learns biologically informed cell embeddings using gene-level ESM2 features + Transformer.
- ST (State Transition):
  - Learns how a perturbation transforms control cells to perturbed cells.
  - Architecture: Transformer + MMD loss.
  - Inputs: Control cell embedding and perturbation embedding.
  - Process: Projects to shared hidden dim -> Transformer models "state shift".
  - Loss: MMD (Maximum Mean Discrepancy) between predicted vs real distributions.

scGPT Architecture

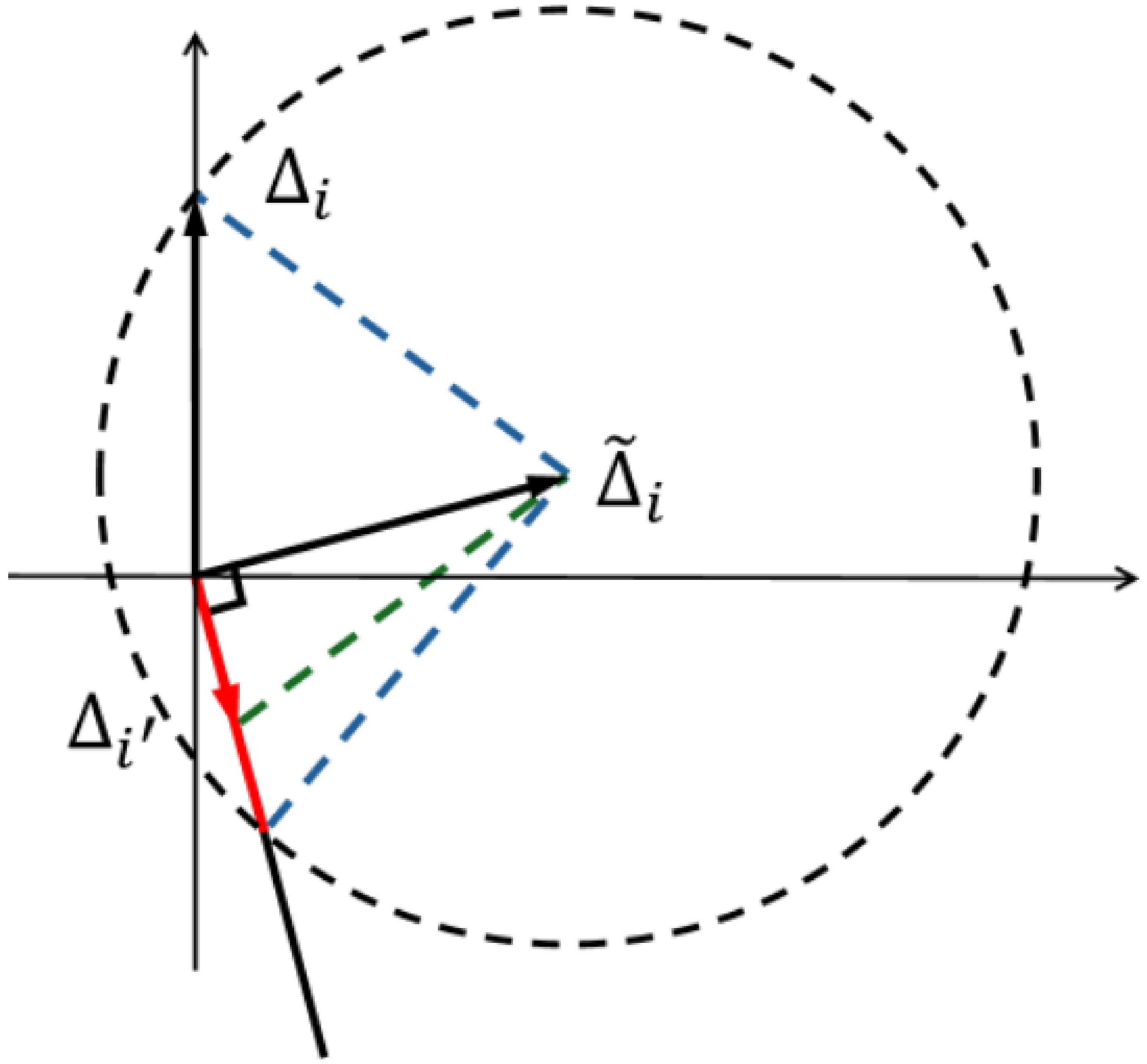


scGPT Configuration

- Frozen Components (Encoder):
  - Parameters with prefixes: encoder, value_encoder, transformer_encoder.
- Trainable Components:
  - pert_encoder: Embedding for perturbation flags.
  - decoder: Heads for expression prediction.
- Loss Components:
  - sw1: Sliced Wasserstein-1 (distribution alignment).
  - proto: ProtoInfoNCE on pseudobulk deltas.
  - de_rank: DE rank loss (used when DE gene map is available).
  - dir: DE direction loss (used when DE gene map is available).
- Total Loss Formula: 0.60 sw1 + 0.25 proto + 0.10 de_rank + 0.05 dir

Evaluation Metric: PDS

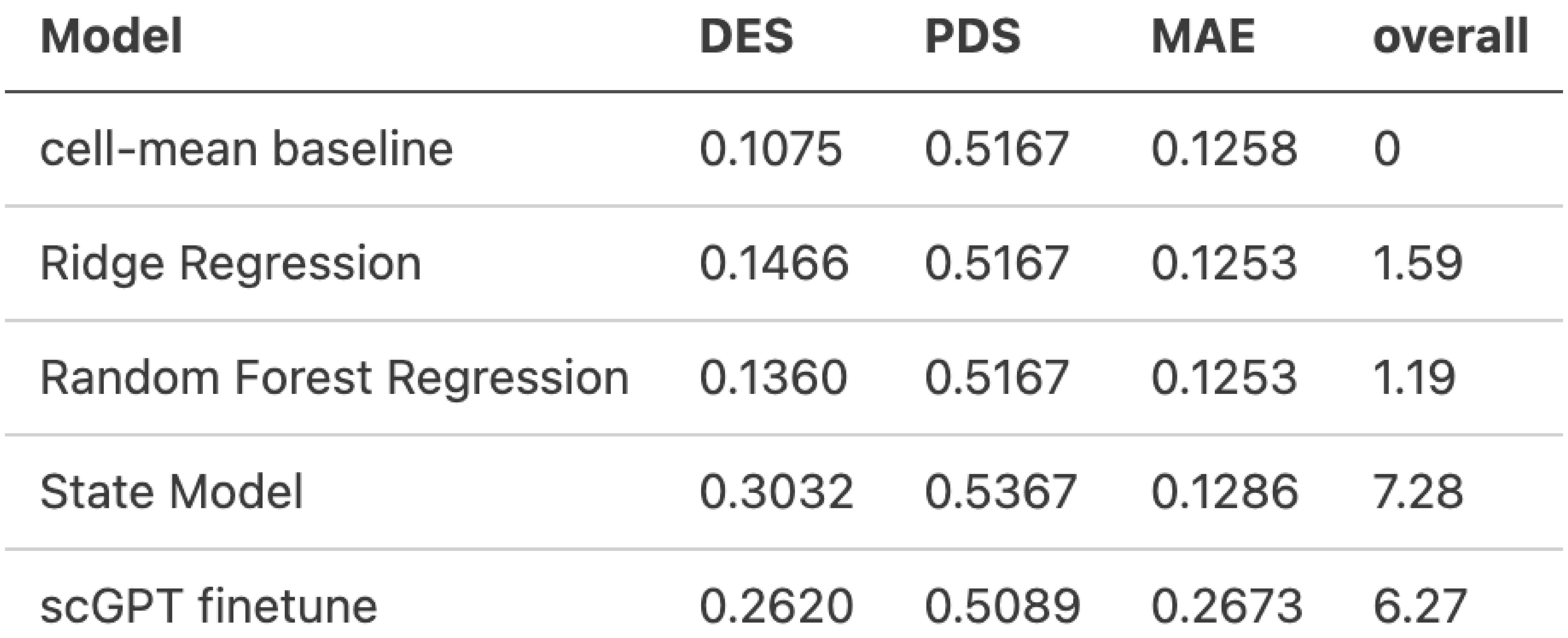
- Measures the degree of similarity in distribution patterns between the gene perturbation effects predicted by the model and the actual perturbation effects.



Other Evaluation Metrics

- DES (Differential Expression Score):
  - Measures whether the predicted differentially expressed genes (DEGs) match the real data DEGs.
- MAE (Mean Absolute Error of Top 2000 Genes):
  - Focuses on genes with the most drastic changes to observe error in predicting pseudo-batch expression.
- Overall Score Calculation:
  - S = (Scaled DES + Scaled PDS + Scaled MAE) / 3 * 100
  - Scaling is based on the cell-mean baseline model.

Evaluation Results



Model	DES	PDS	MAE	overall
cell-mean baseline	0.1075	0.5167	0.1258	0
Ridge Regression	0.1466	0.5167	0.1253	1.59
Random Forest Regression	0.1360	0.5167	0.1253	1.19
State Model	0.3032	0.5367	0.1286	7.28
scGPT finetune	0.2620	0.5089	0.2673	6.27

Limitations: Metrics

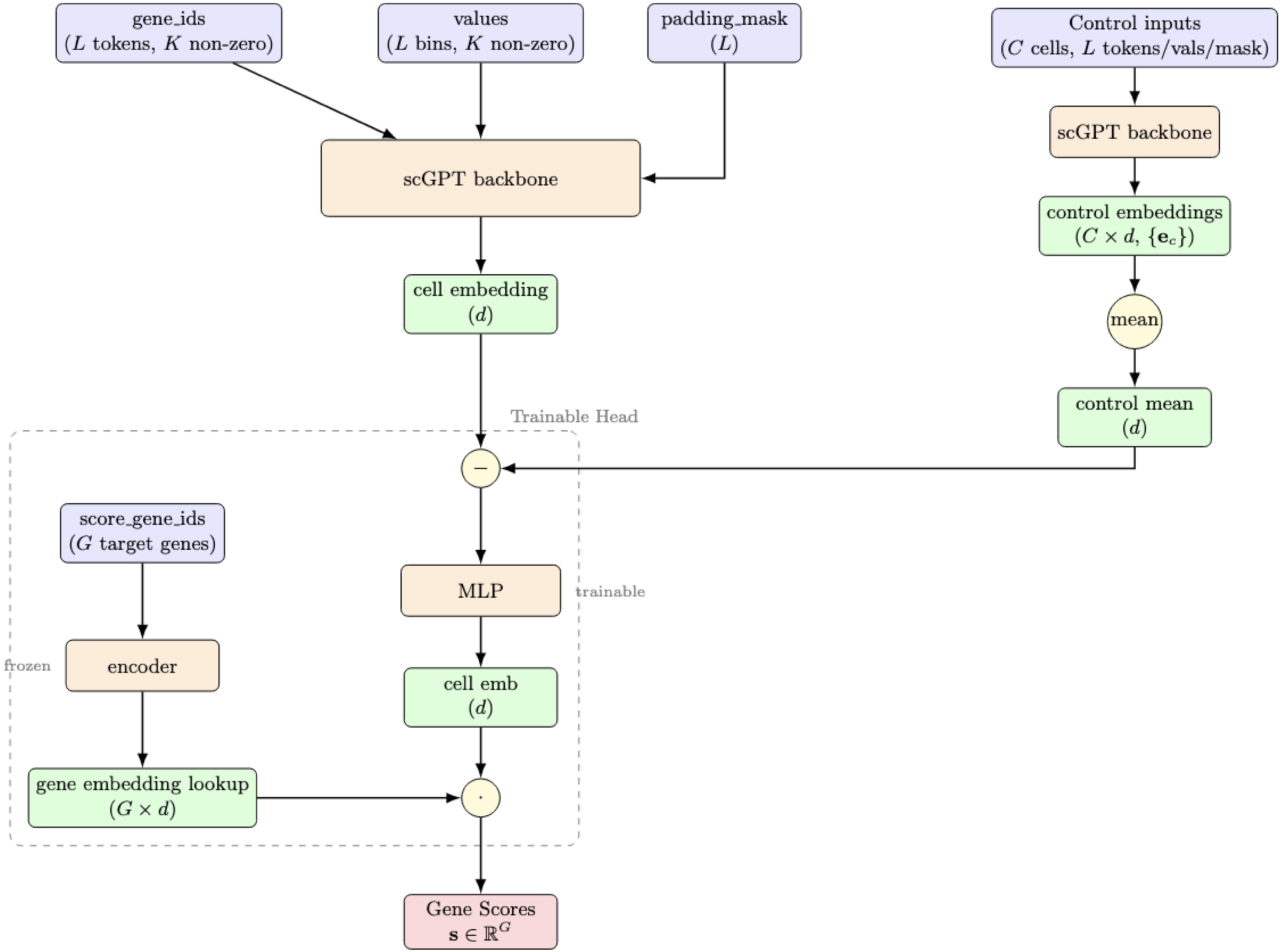
- PDS Dominance:
  - Metric scaling caused the leaderboard to be dominated by PDS.
  - PDS carried roughly twice the weight of DES.
  - Biased optimization toward matching pseudo-bulk magnitudes rather than accuracy.
- MAE Issues:
  - Bulk MAE is dominated by baseline expression levels.
  - Insensitive to biologically meaningful but subtle changes.
  - Fails to distinguish true replicates from unrelated perturbations.

Task2 Definition

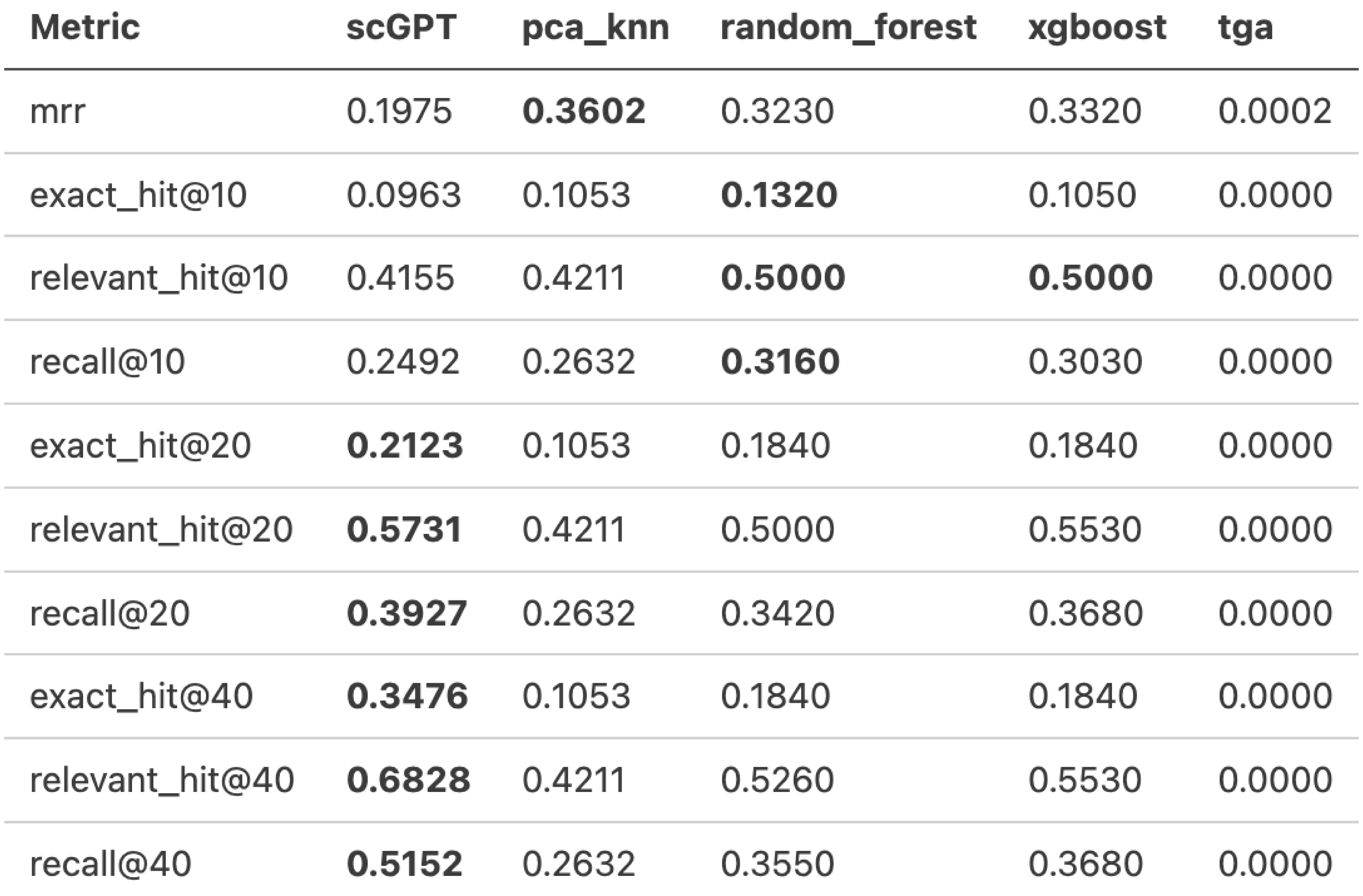
- Task: Given (x, y), predict p and rank genes to recover perturbation targets.
- Inputs (x, y): Expression tokens per perturbed cell plus matched controls.
- Output: Per-gene logits used for ranking against target genes.

$$\text{given } (x, y) \ \Rightarrow \ \hat{p} = g(x, y)$$

Model Architecture



Results Comparison



Metric	scGPT	pca_knn	random_forest	xgboost	tga
mrr	0.1975	**0.3602**	0.3230	0.3320	0.0002
exact_hit@10	0.0963	0.1053	**0.1320**	0.1050	0.0000
relevant_hit@10	0.4155	0.4211	**0.5000**	**0.5000**	0.0000
recall@10	0.2492	0.2632	**0.3160**	0.3030	0.0000
exact_hit@20	**0.2123**	0.1053	0.1840	0.1840	0.0000
relevant_hit@20	**0.5731**	0.4211	0.5000	0.5530	0.0000
recall@20	**0.3927**	0.2632	0.3420	0.3680	0.0000
exact_hit@40	**0.3476**	0.1053	0.1840	0.1840	0.0000
relevant_hit@40	**0.6828**	0.4211	0.5260	0.5530	0.0000
recall@40	**0.5152**	0.2632	0.3550	0.3680	0.0000

Summary

- Forward Prediction (Task 1):
  - Fine-tuned the scGPT foundation model for the Virtual Cell Challenge.
  - Achieved superior generalization on unseen perturbations, outperforming traditional baseline models.
- Critical Metric Analysis:
  - Identified limitations in official competition metrics (PDS & MAE).
  - Highlighted the trade-off between statistical distribution matching and true biological accuracy.
- Novel "Reverse" Prediction Task (Task 2):
  - Defined a new task to identify upstream genetic targets based on cellular expression changes.
  - Modified the foundation model architecture to support gene ranking, surpassing traditional retrieval methods.
- Demonstrated practical value for drug target identification and mechanism-of-action studies.