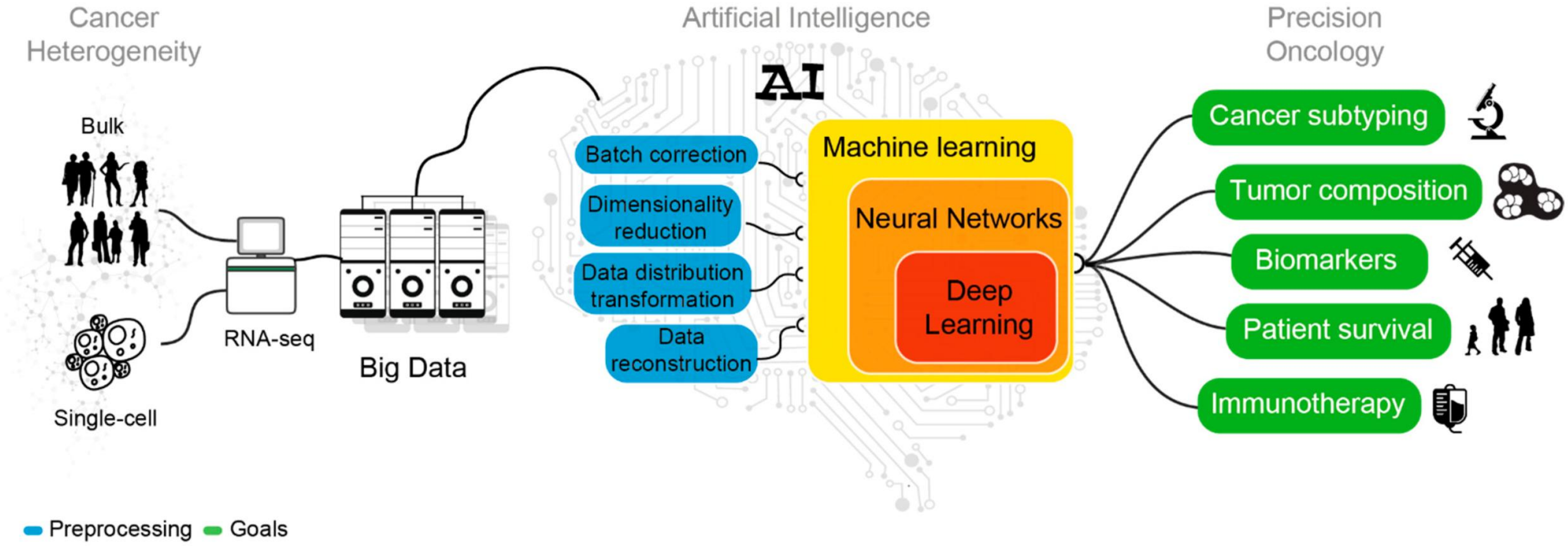
# Perturb-sequencing

- AI in scRNA-seq



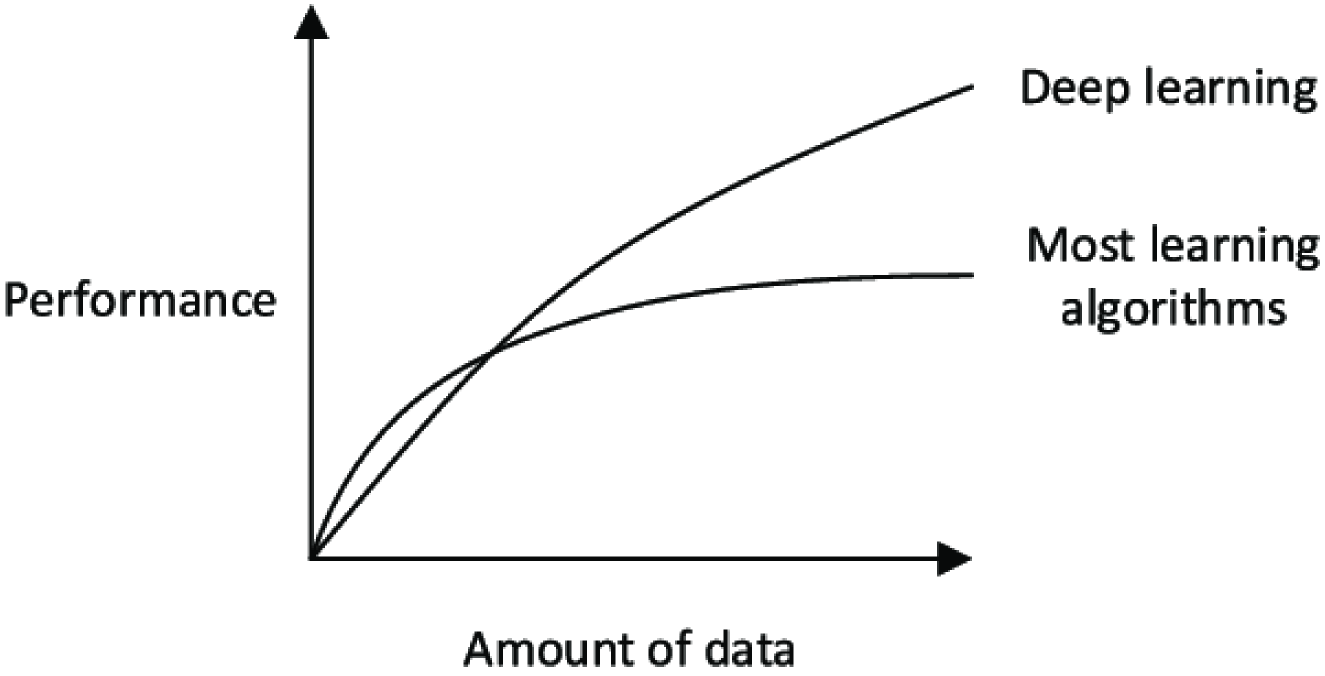

Artificial Intelligence in Bulk and Single-Cell RNA-Sequencing Data to Foster Precision Oncology

- AI in scRNA-seq

OpenAI
ChatGPT 4.0

Large language model


Computer vision

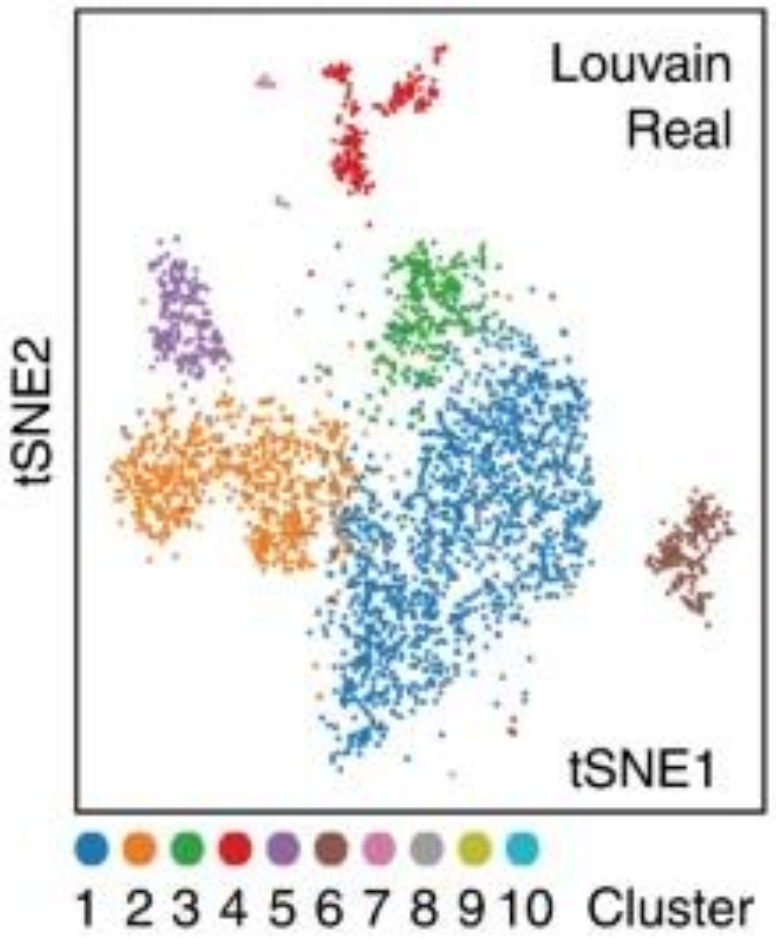


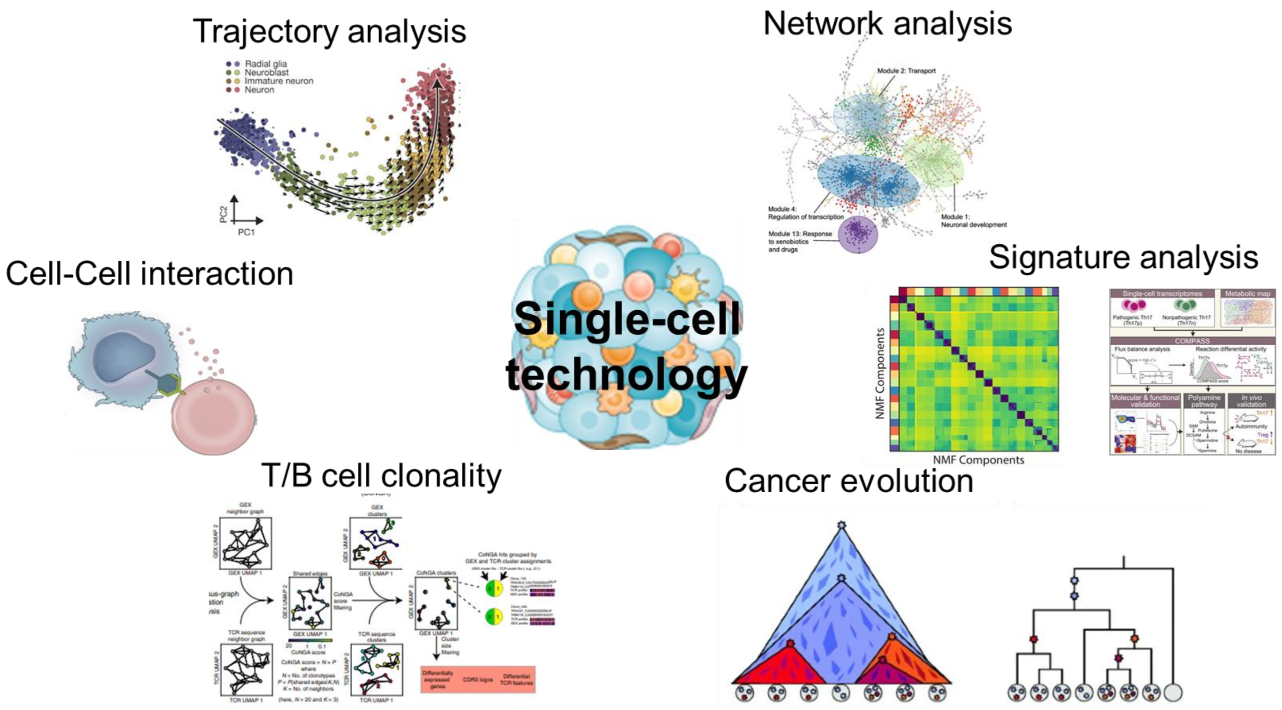
- AI in scRNA-seq


?

- Simulation

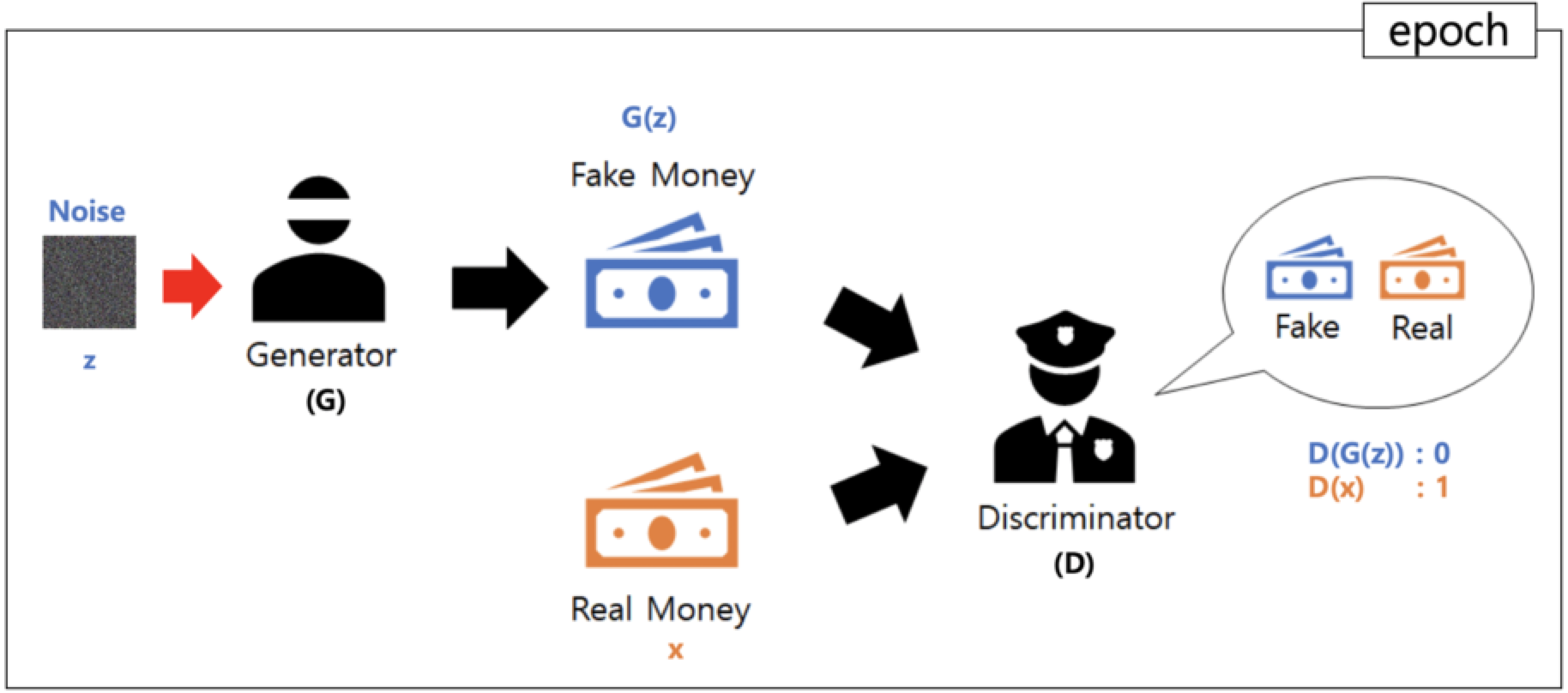
-Simulation: provides ground-truth for method development



- GAN

-Generative Adversarial Networks




-Noise: Gaussian distribution

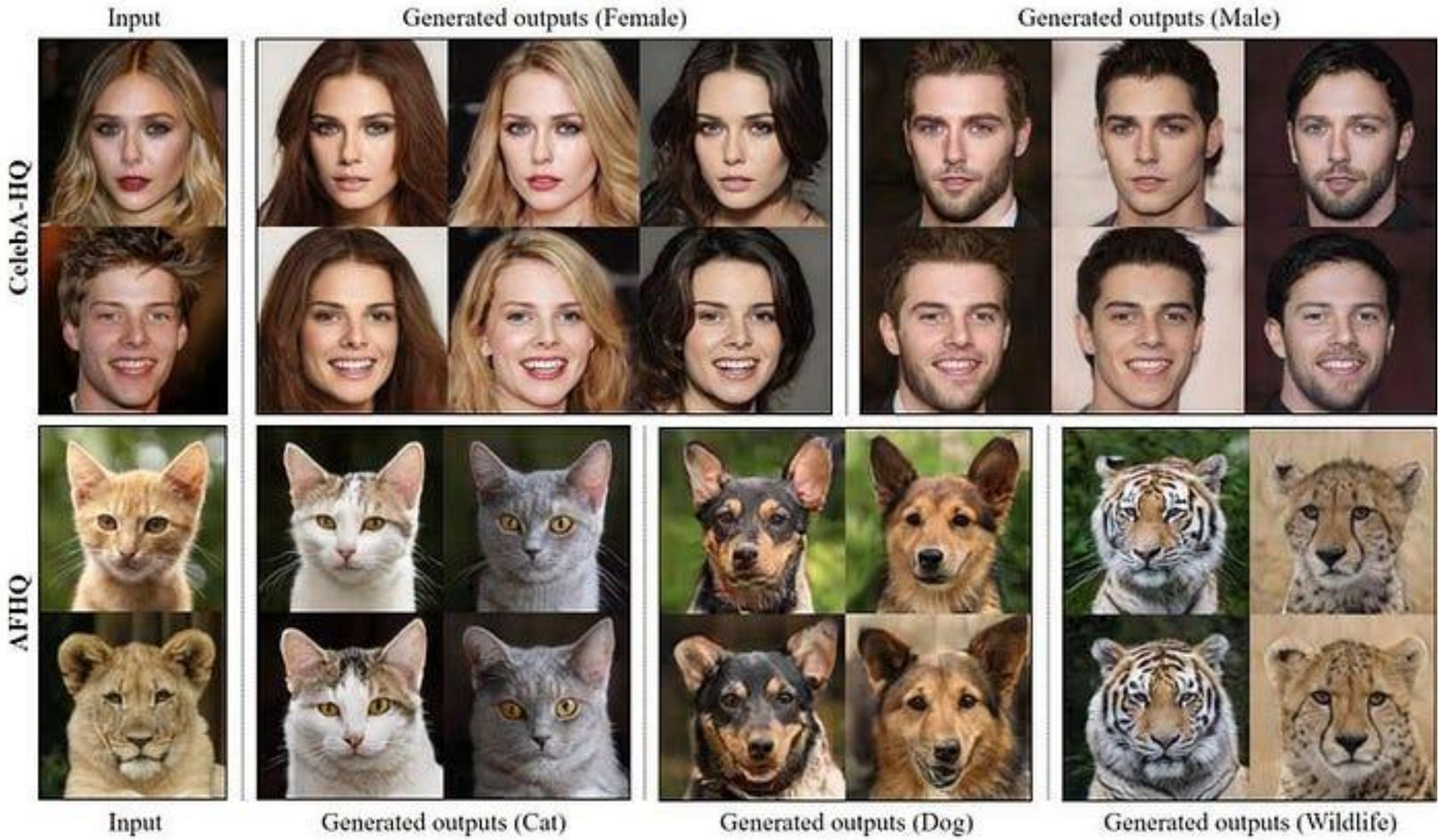
-Generator: make fake Money → similar to real money

-Discriminator: distinguish between fake and real (train first)

→ finally 0.5 vs 0.5 (cannot distinguish)

→ After training, generator can simulate the data

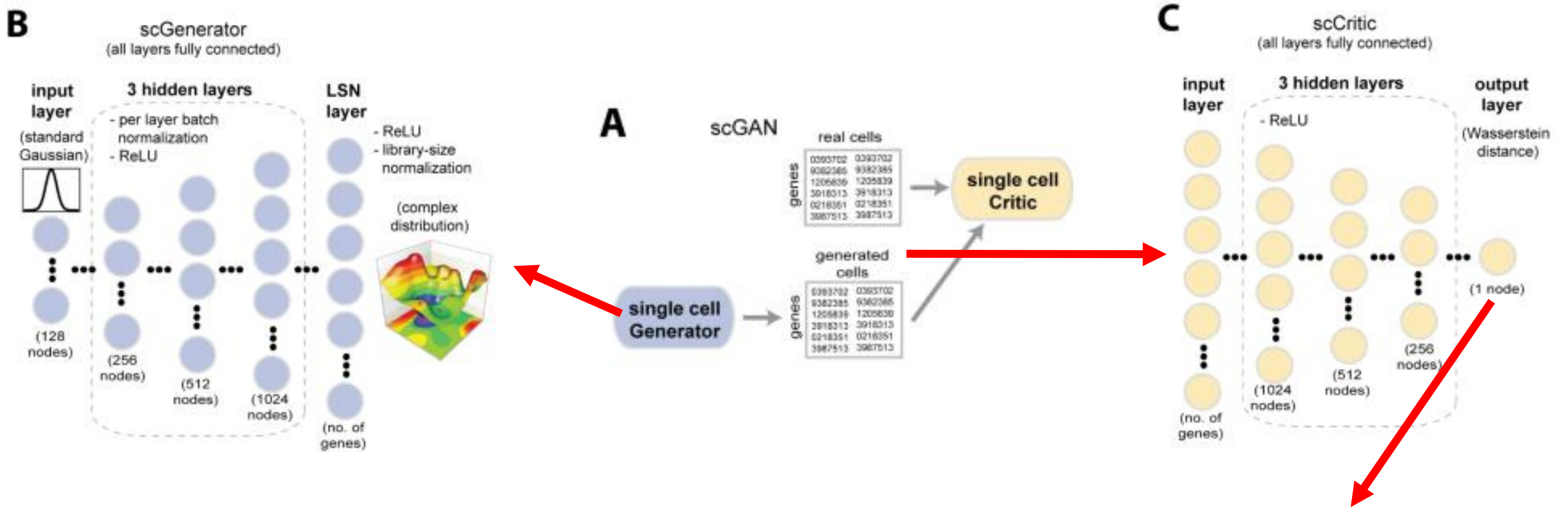
- GAN


Input
Generated outputs (Female)
Generated outputs (Male)
CelebA-HQ
AFHQ
Input
Generated outputs (Cat)
Generated outputs (Dog)
Generated outputs (Wildlife)

• Simulation

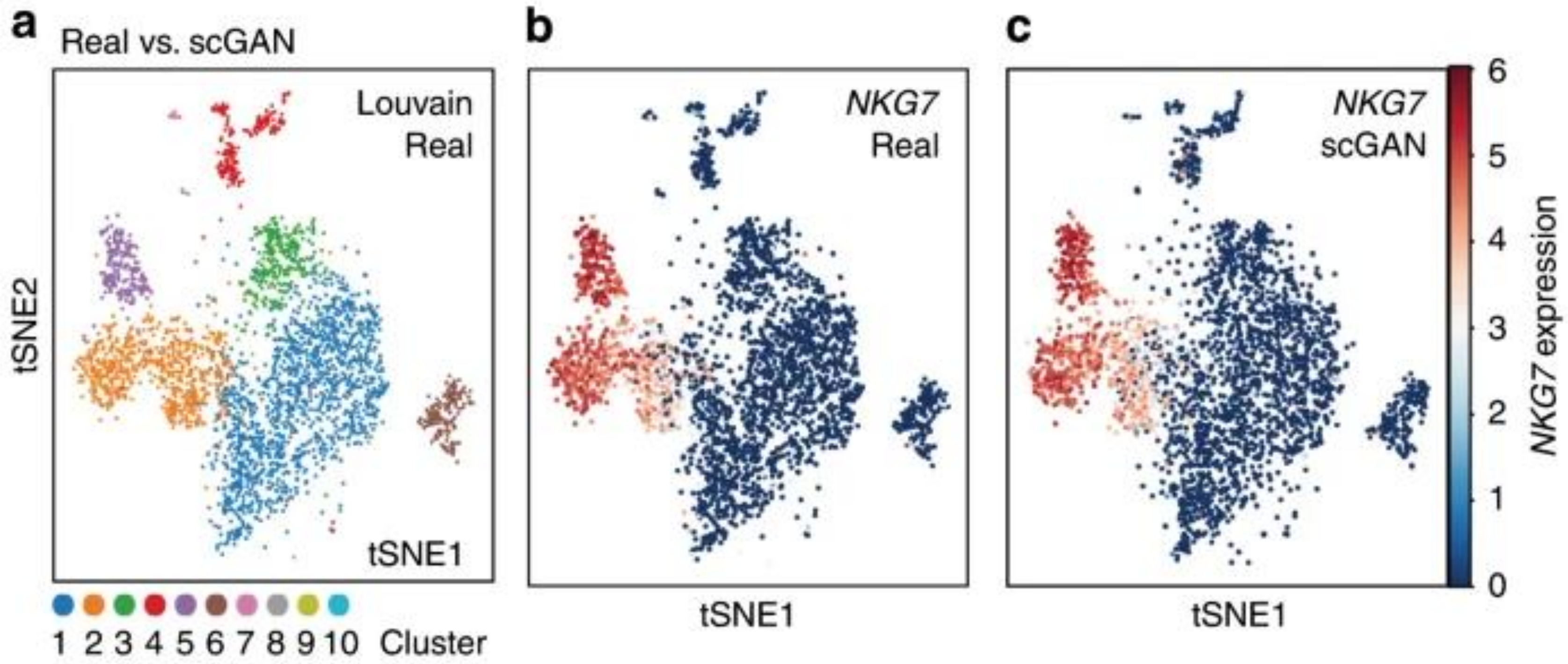
Article | Open access | Published: 09 January 2020

# Realistic in silico generation and augmentation of single-cell RNA-seq data using generative adversarial networks




-Discriminator: distinguish between fake and real

- Simulation



a
Real vs. scGAN
Louvain
Real
tSNE2
tSNE1
1 2 3 4 5 6 7 8 9 10 Cluster
b
NKG7
Real
tSNE1
c
NKG7
scGAN
tSNE1
6
5
4
3
2
1
0
NKG7 expression

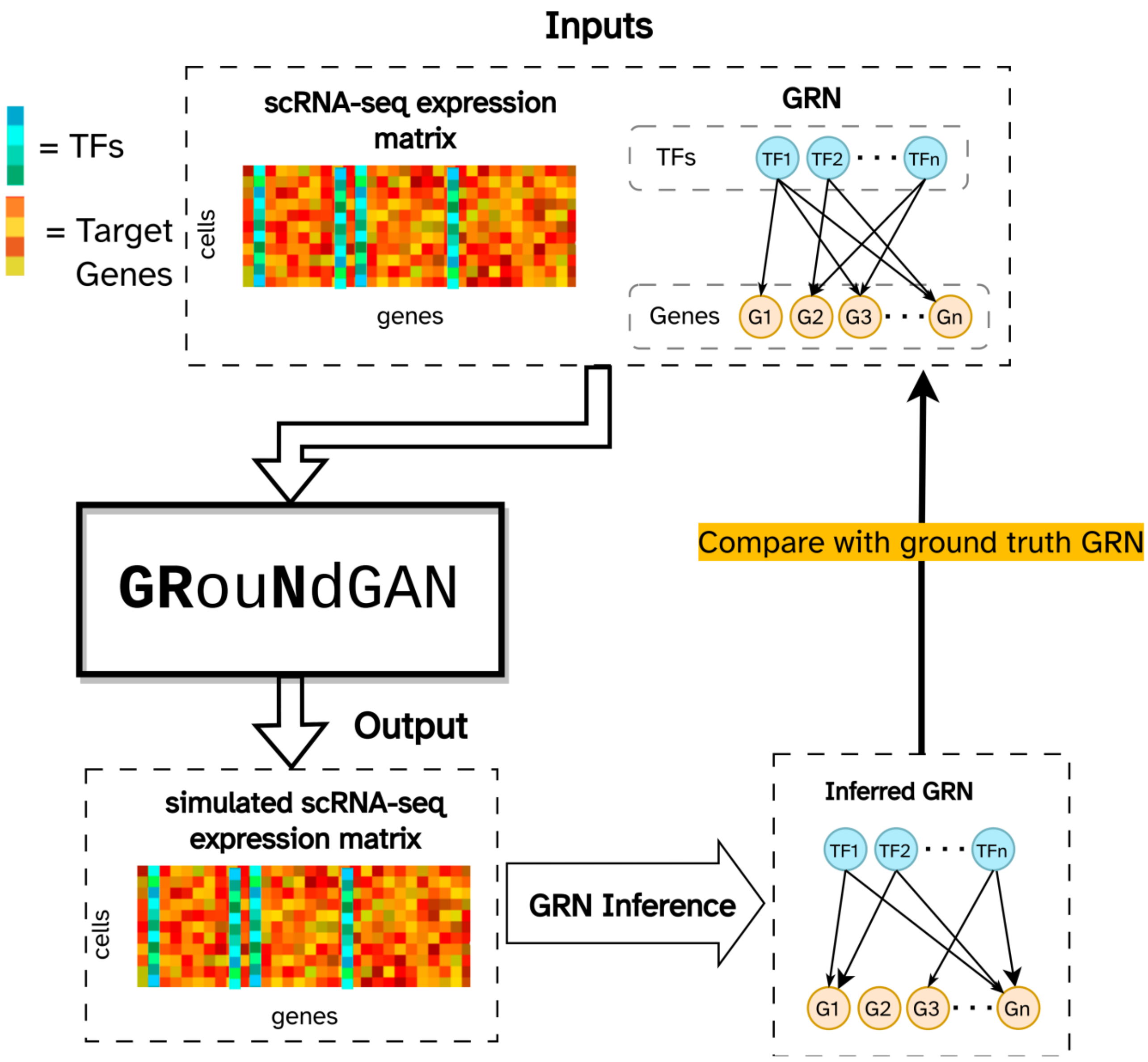
- Simulation

nature communications

Article https://doi.org/10.1038/s41467-0

# GRouNdGAN: GRN-guided simulation of single-cell RNA-seq data using causal generative adversarial networks

-Add GRN information together (TF → gene)
→ Gene expression + GRN



- Simulation

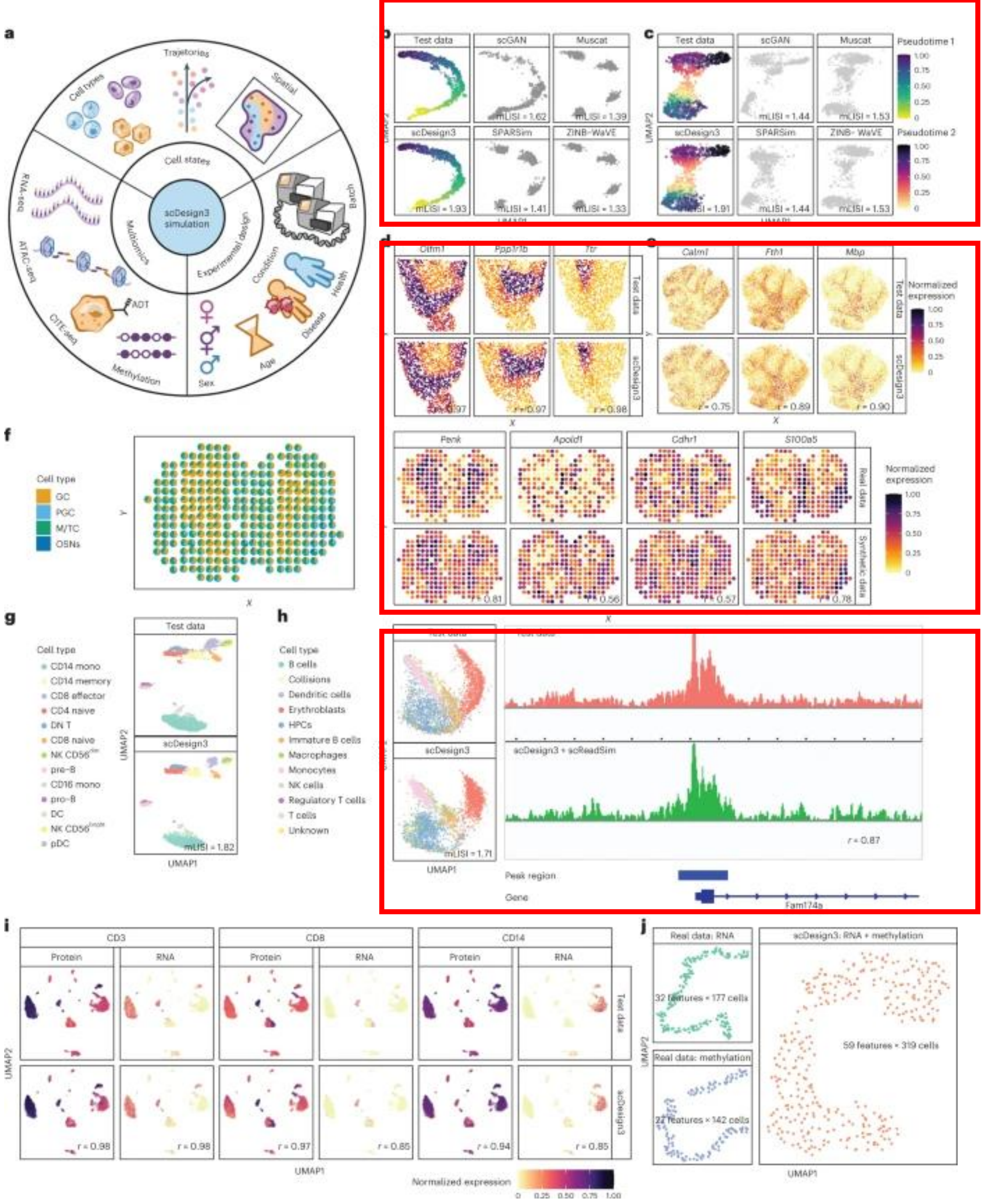
nature biotechnology

Brief Communication https://doi.org/10.1038/s41587-023-01772-1

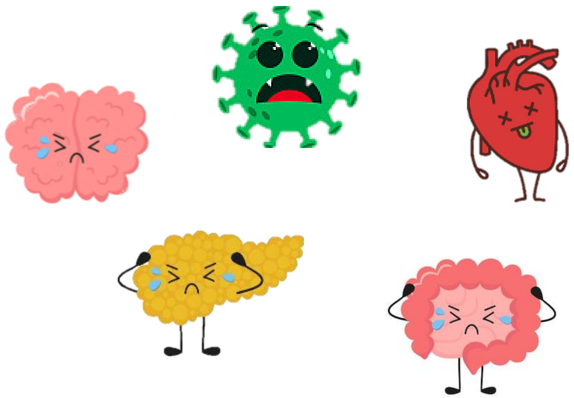
# scDesign3 generates realistic in silico data for multimodal single-cell and spatial omics

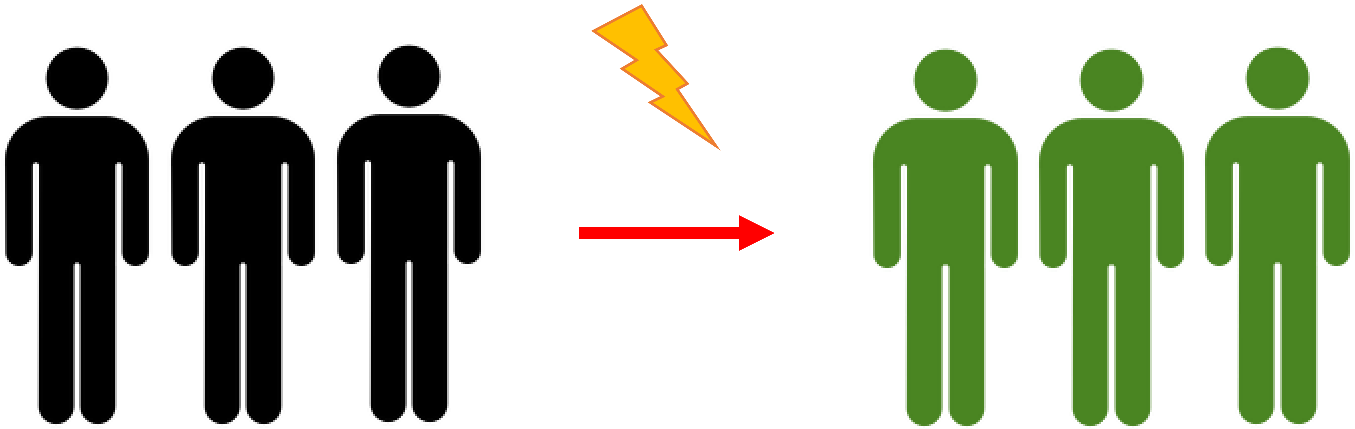
| | | | |
|---|---|---|---|
| Feature distribution | Gamma-Normal mixture | Poisson, ZIP NB, ZINB | Poisson, ZIP NB, ZINB **Bernoulli, Normal** |

-Statistic model can be also used for simulation
Multi-modality: rna, atac, methyl, spatial (spot, cell)
→ Require specific statistic model

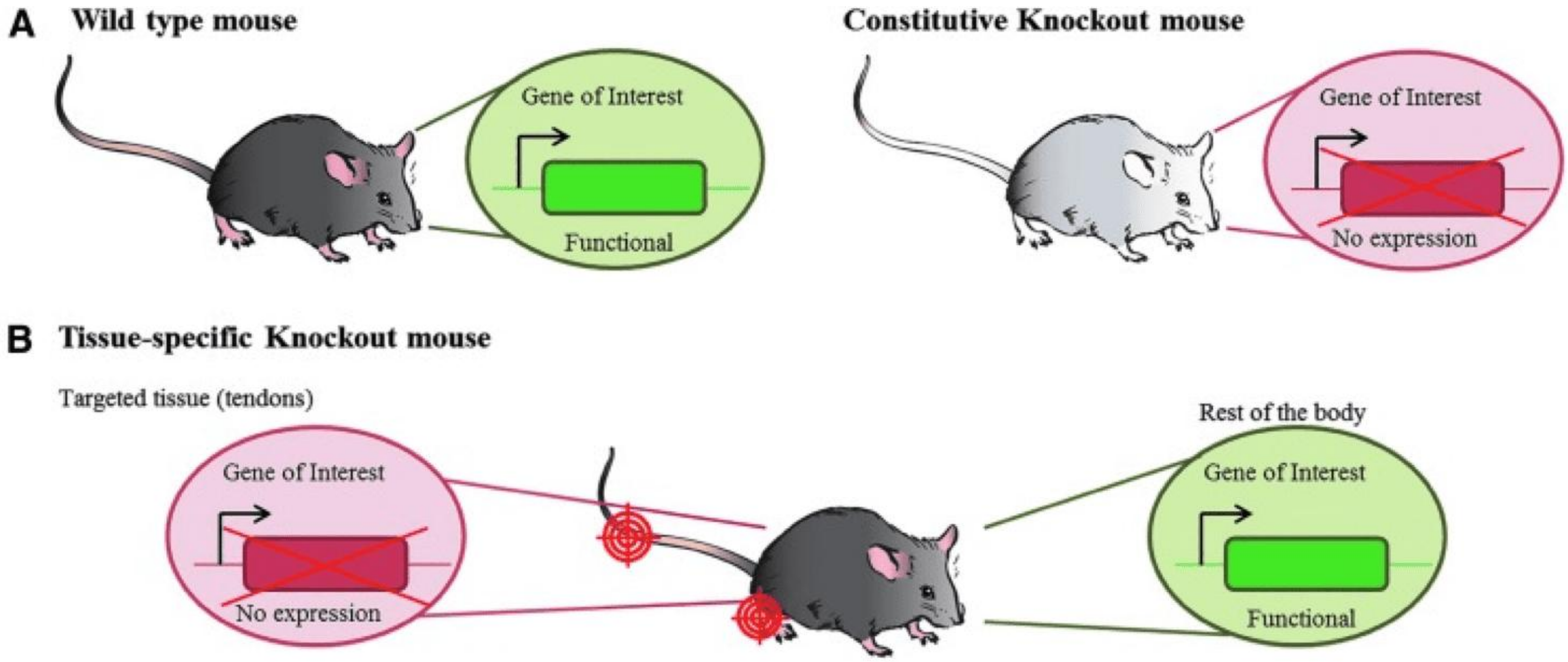


- Perturbation

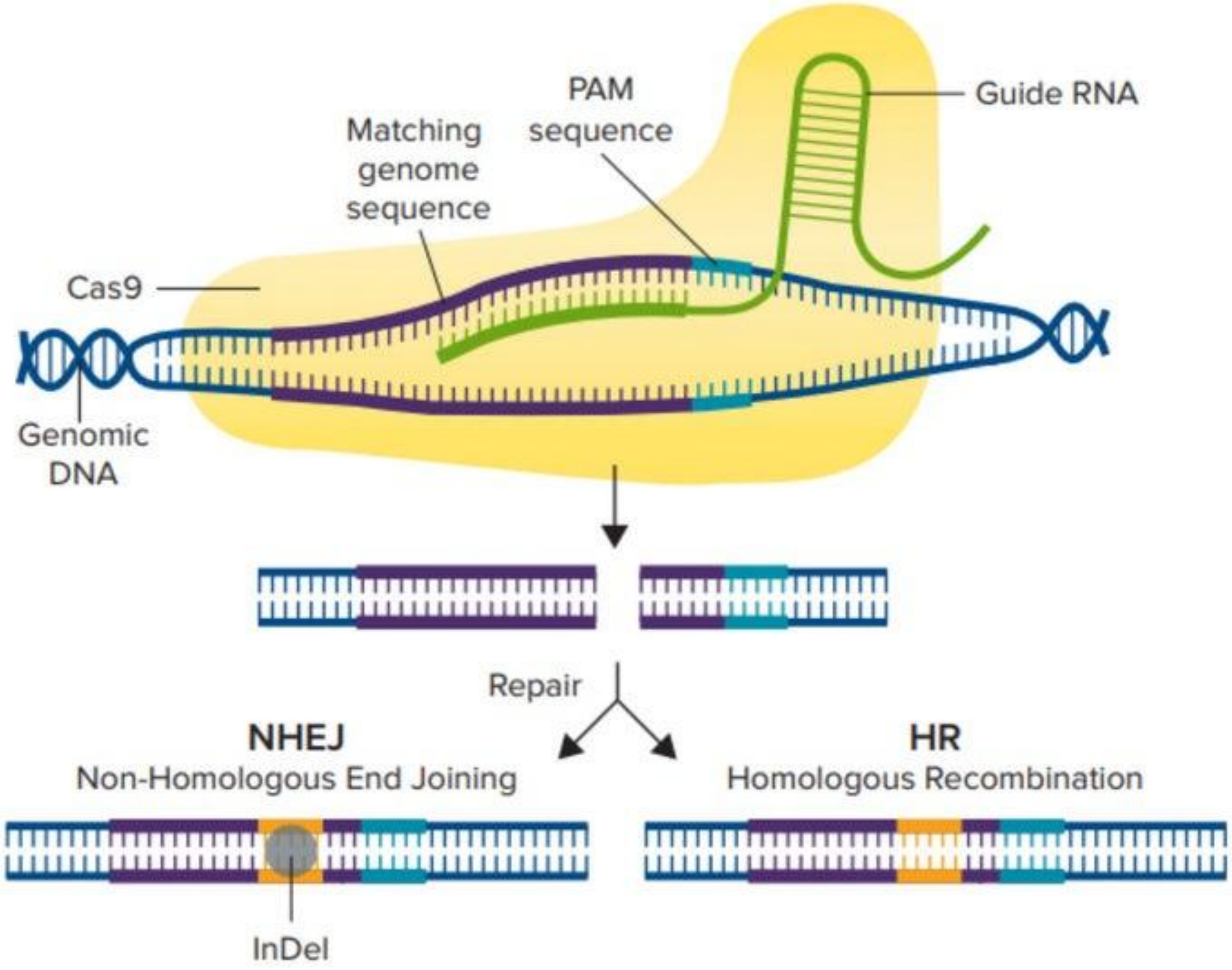




- Gene KO experiment



- CRISPR-Cas9




Guide RNA → detect the target region
→ Cas9 cut the DNA
→ Repair → Gene KO

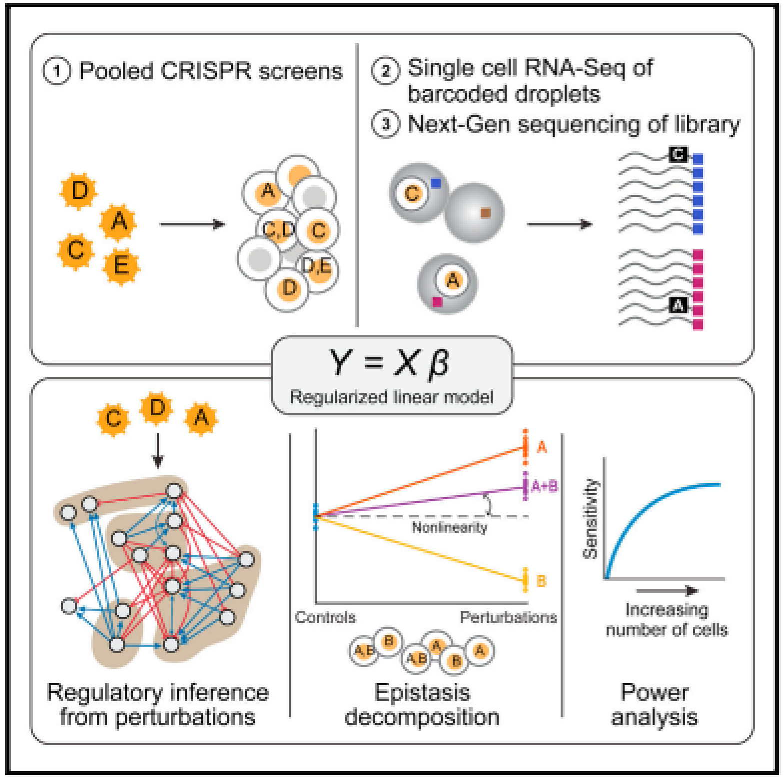
- Perturb-seq

Cell

Resource

# Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens

## Graphical Abstract




## Authors

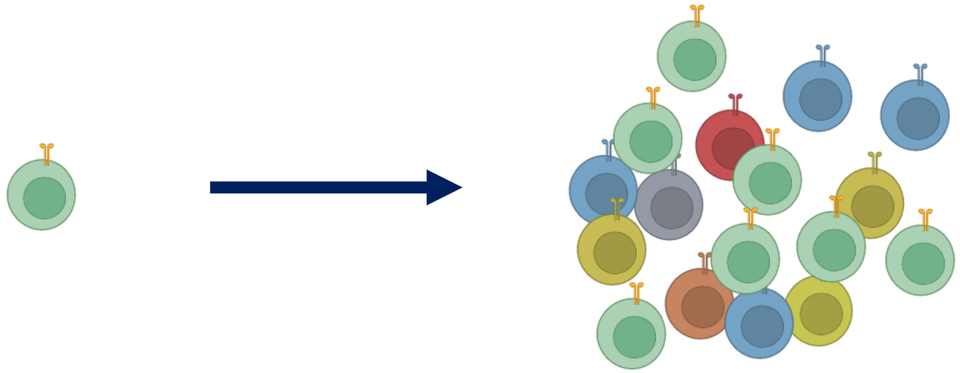
Atray Dixit, Oren Parnas, Biyu Li, ..., Jonathan S. Weissman, Nir Friedman, Aviv Regev

## Correspondence

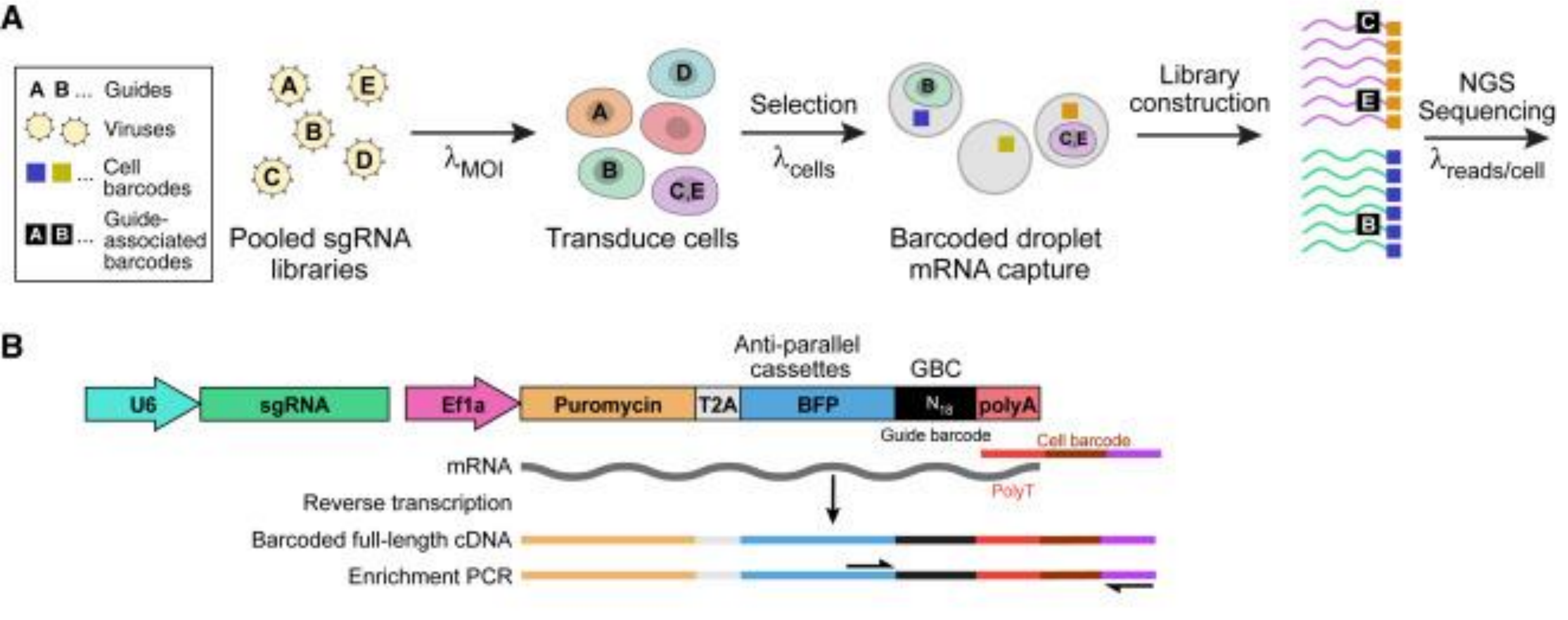
aregev@broadinstitute.org

## In Brief

A technology combining single-cell RNA sequencing with CRISPR-based perturbations termed Perturb-seq makes analyzing complex phenotypes at a large scale possible



- Perturb-seq

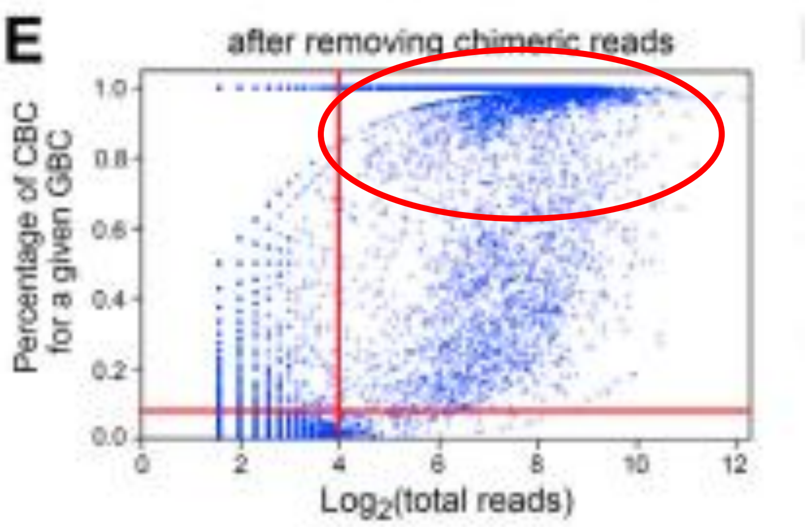


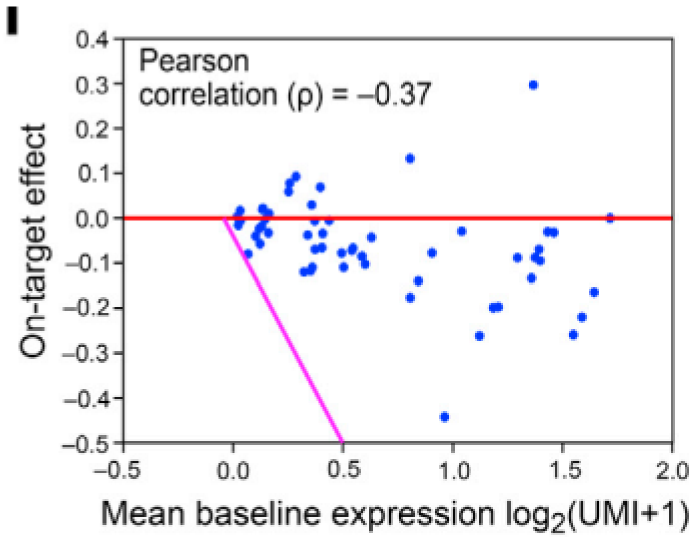

-Each sgRNA: Guide barcode (GBC) → which will be expressed → detect during the alignment
-Each sgRNA → each cell (cell barcode)
-Each cell: different genetic perturbation
-Obtain various perturbation of cells at the same time

# • Perturb-seq

*Technical comments
- 3 guides / gene (different part of the gene)
- Negative ctrl: non-target gRNA: do not target the genome, targeting intergenic region
- Pre-sorting: sgRNA+, Cas9+, CD8+, viable cells → FACS sorting

But! gRNA drop out + multiple guides




X: gRNA exp

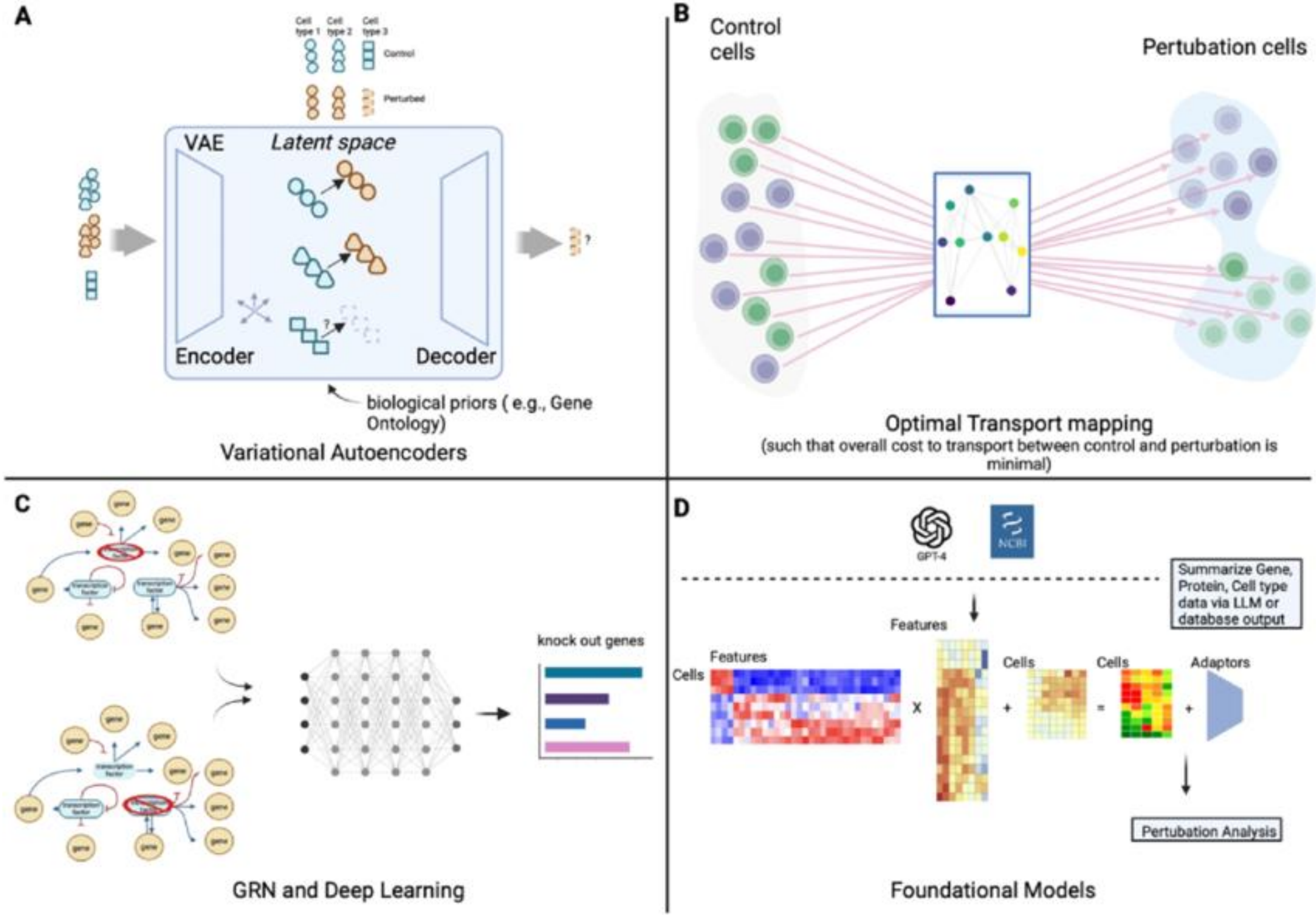
Y: target gRNA / total gRNA (proportion)

→ Both high expression → good cell!

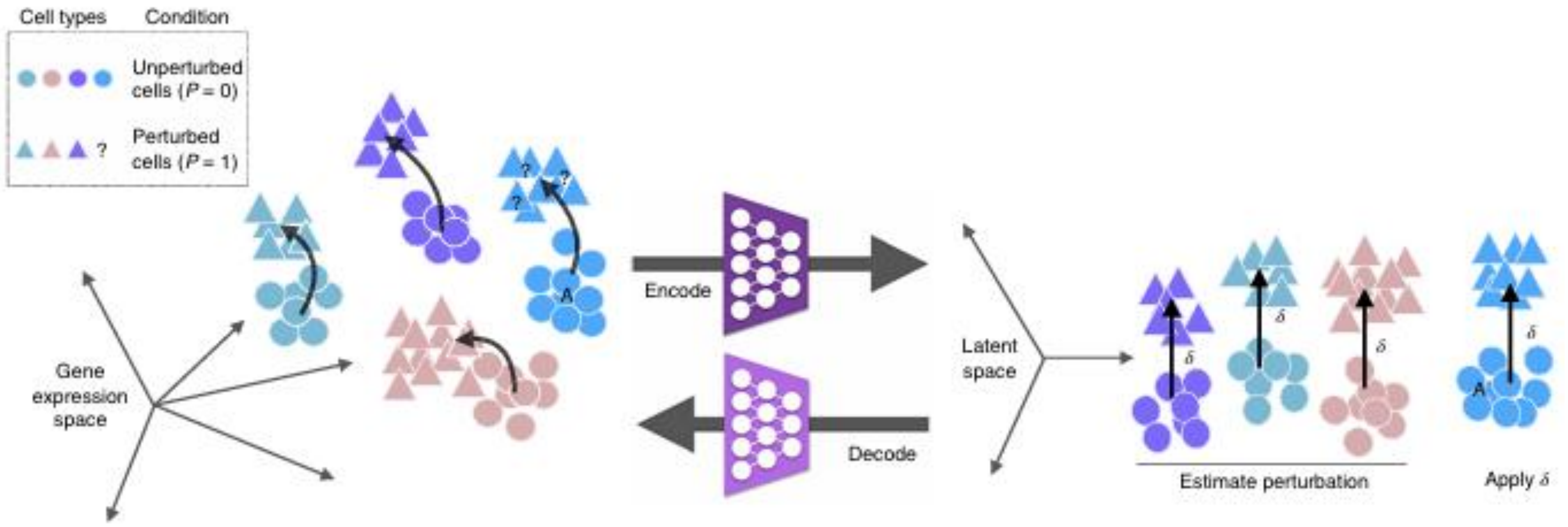
-target gene expression after on target gRNA

→ Negatively correlated (target → KO → no expression

- Perturbation modeling

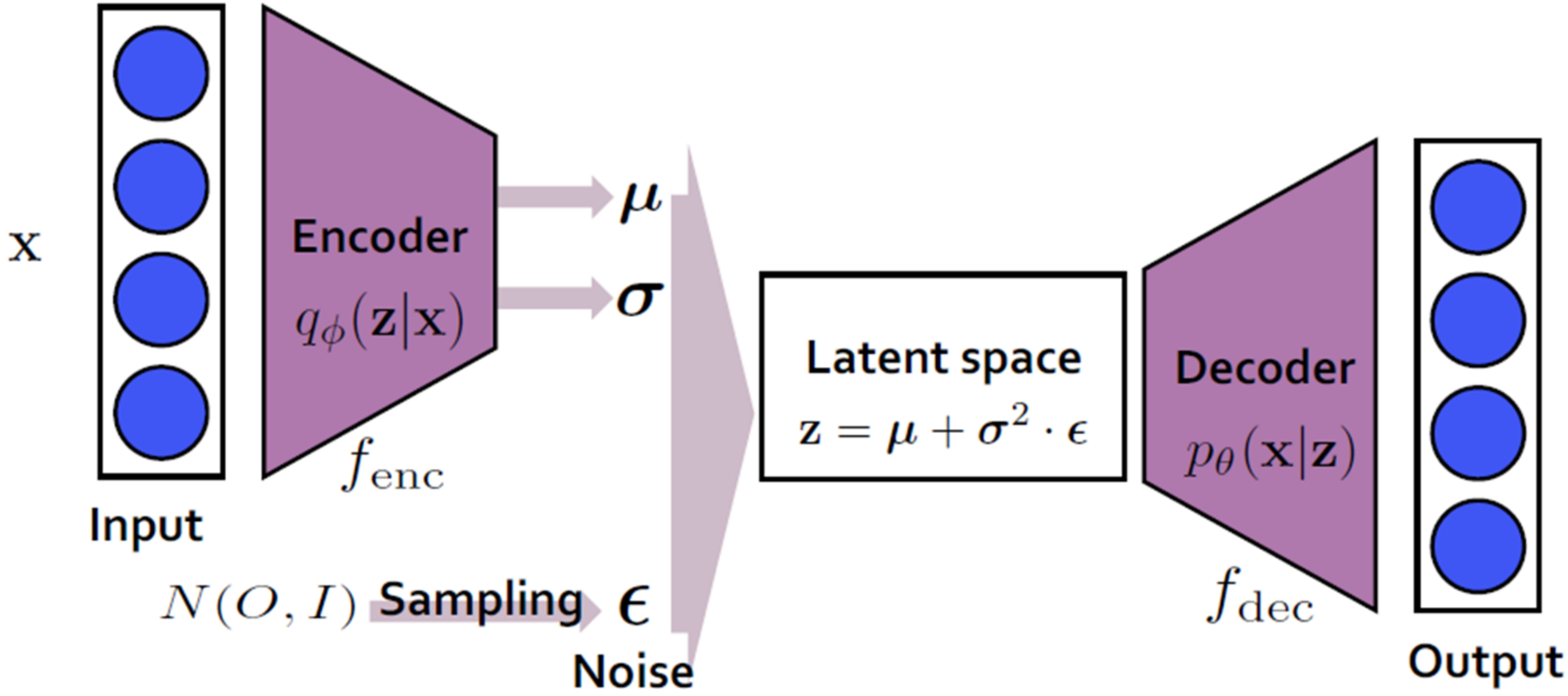


- scGen




-Variational AutoEncoder (VAE) based → simulation-based perturbation effect size

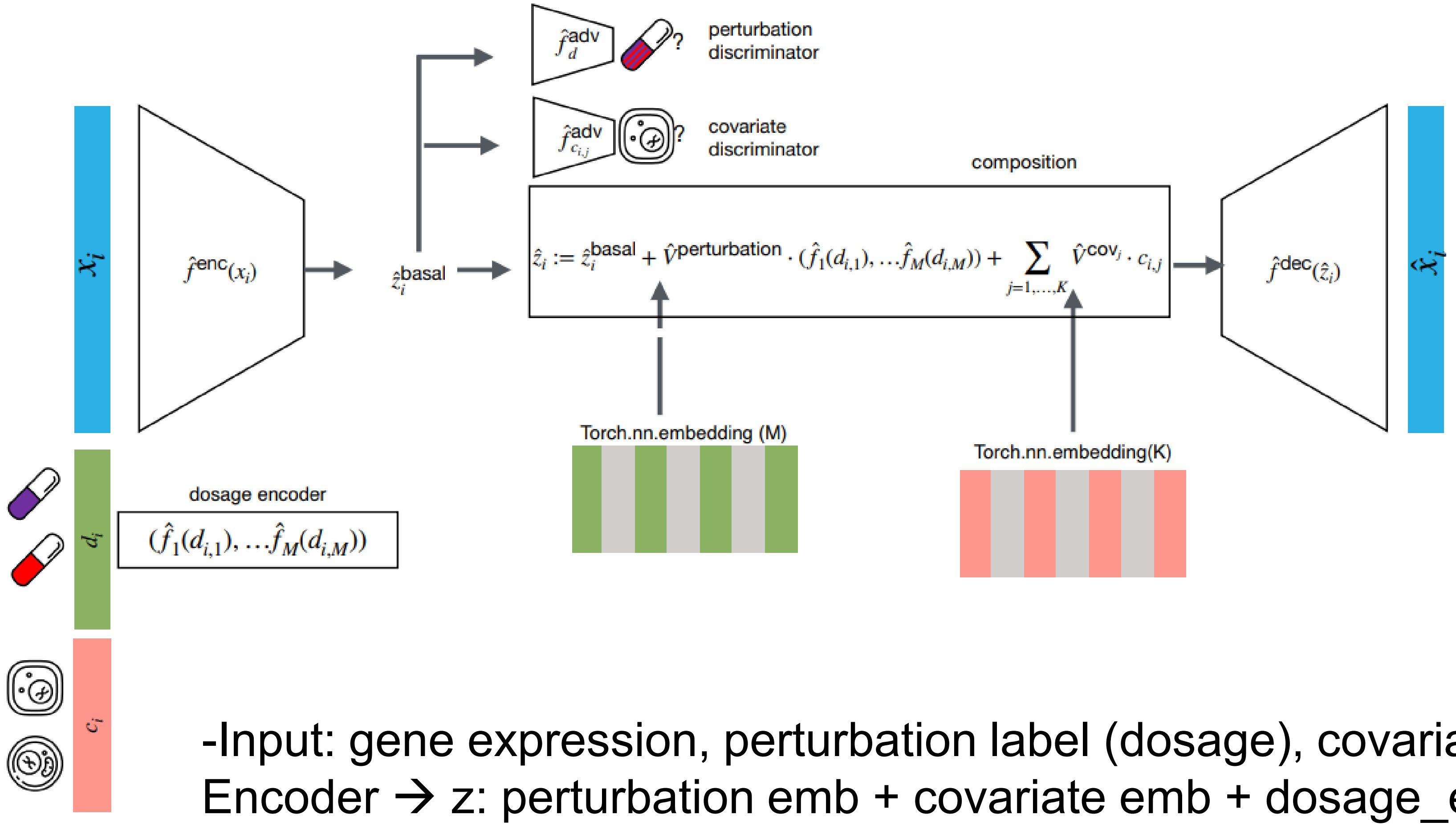
- scGen




-Input: gene expression → Encoder → Gaussian distribution (latent space)
→ Random noise sampling → Decoder → output (simulated gene expression)

*** Make “Input” & “Output” the same
→ Latent space: abstract of perturbation
→ Perturb – unperturb from latent space → perturbation effect size

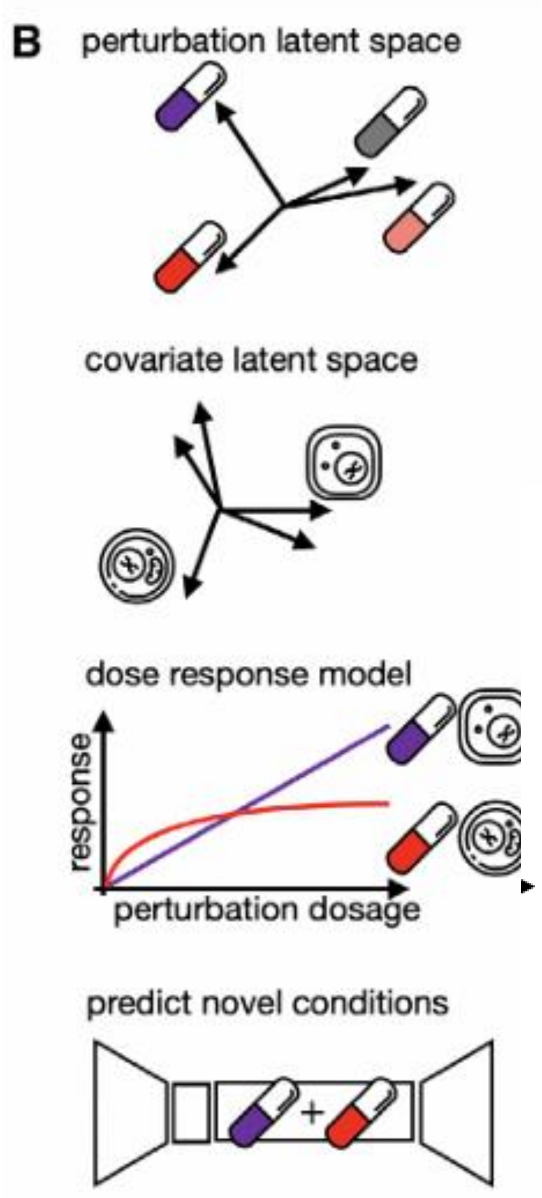
- CPA




-Input: gene expression, perturbation label (dosage), covariates
Encoder → z: perturbation emb + covariate emb + dosage_emb

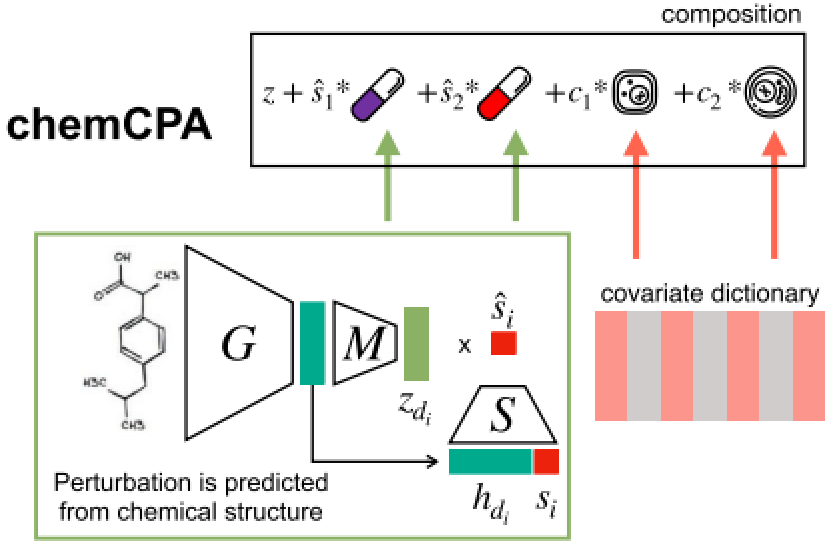
Loss fn: reconstruction error
Cross entropy: [f(z_latent) & perturb category] + [f(z_latent), cov]
→ Latent space: can distinguish perturbation & covariable

- CPA



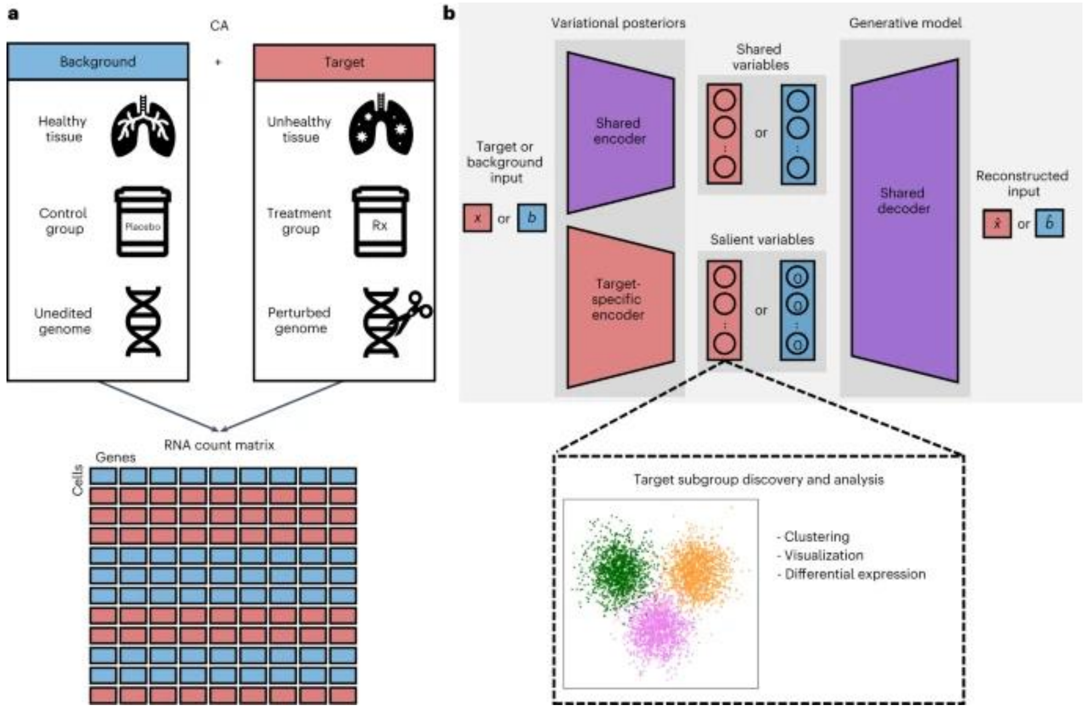

Decoder
1)Latent space (perturbed, non-perturbed: covariates)
→ Binary perturbation classification or multiple perturbation (multiple drug)

2)Dosage effect or time-dependent
3)Unseen drug prediction
4)drug-combination prediction




-Add Drug structure information

• ContrastiveVI



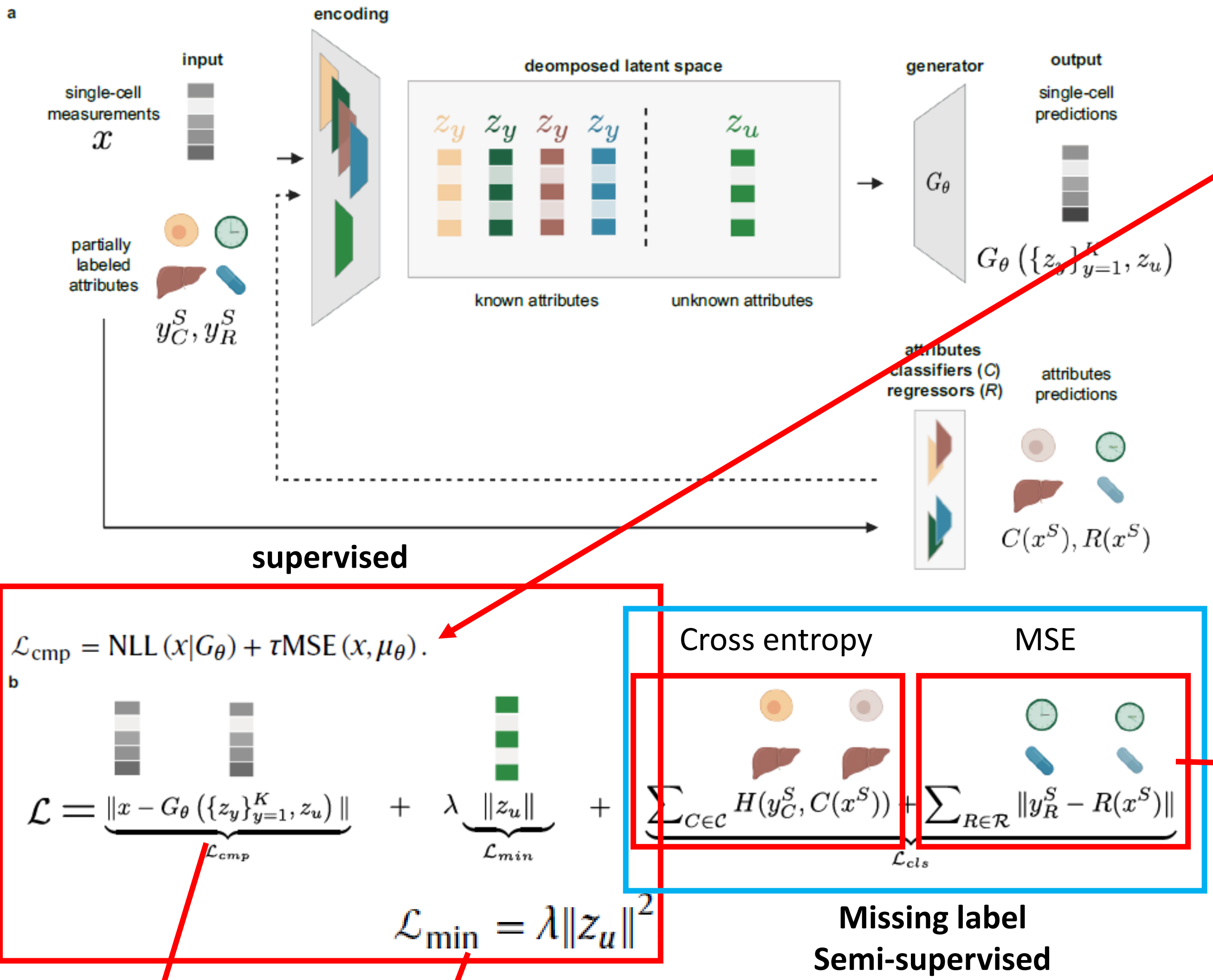

-VAE based

-Shared space: drug treatment (DMSO, drug) → well mixed

Perturbed space: WT, p53 Mut  classification

-cell type-specific response

- Biolord




$$\mathcal{L}_{\mathrm{cmp}} = \mathrm{NLL}\,(x|G_\theta) + \tau \mathrm{MSE}\,(x, \mu_\theta).$$

$$\mathcal{L} = \underbrace{\|x - G_\theta(\{z_y\}_{y=1}^{K}, z_u)\|}_{\mathcal{L}_{cmp}} + \lambda \underbrace{\|z_u\|}_{\mathcal{L}_{min}} + \underbrace{\sum_{C\in\mathcal{C}} H(y_C^S, C(x^S)) + \sum_{R\in\mathcal{R}} \|y_R^S - R(x^S)\|}_{\mathcal{L}_{cls}}$$

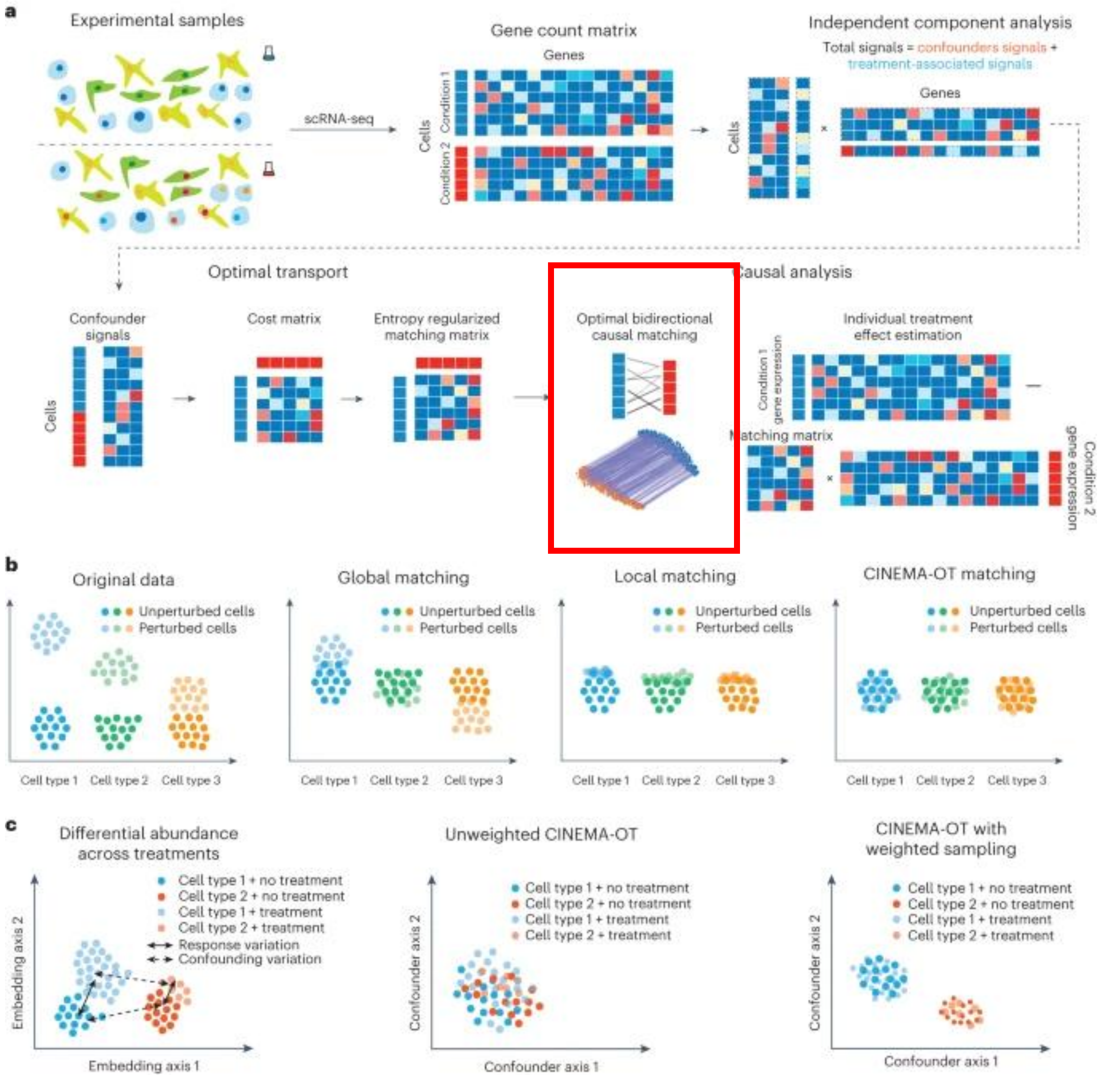
$$\mathcal{L}_{\mathrm{min}} = \lambda \|z_u\|^2$$

-Input distribution
Zu + n (gaussian noise)
Log-norm → gaussian
Raw count → ZINB
Peak → poisson

**-Missing label semisupervised learning**
C: classifier for categorical attribute (Cross Ent)
R: regressor for ordinal attribute (MSE)

-(known*unknown) Decomposed latent space: reconstruction error optimization
Completeness term: negative log-likelihood loss (NLL) per distribution
-Unknown attribute: L2 norm
Information sharing between known & unk

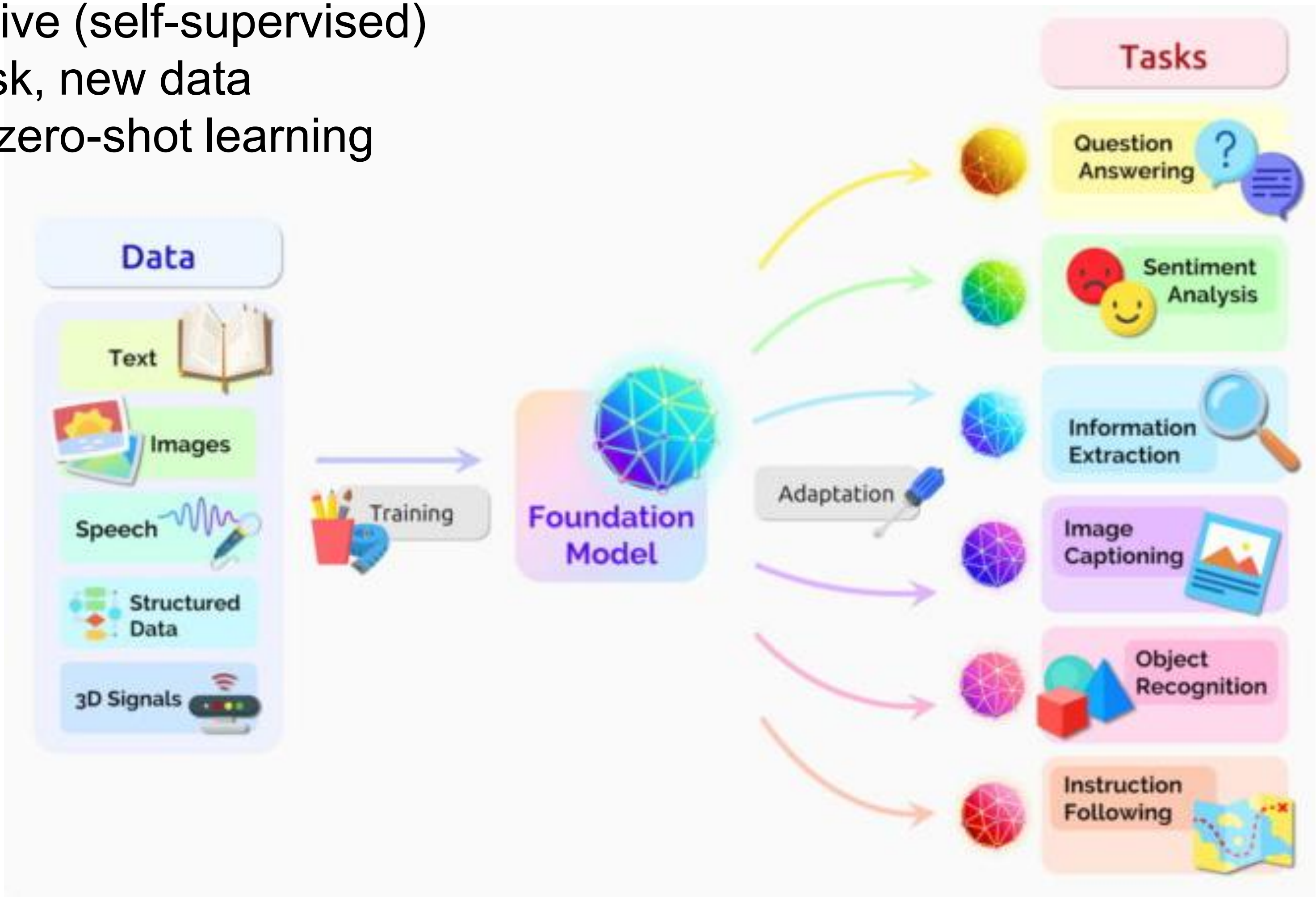
- CINEMA-OT



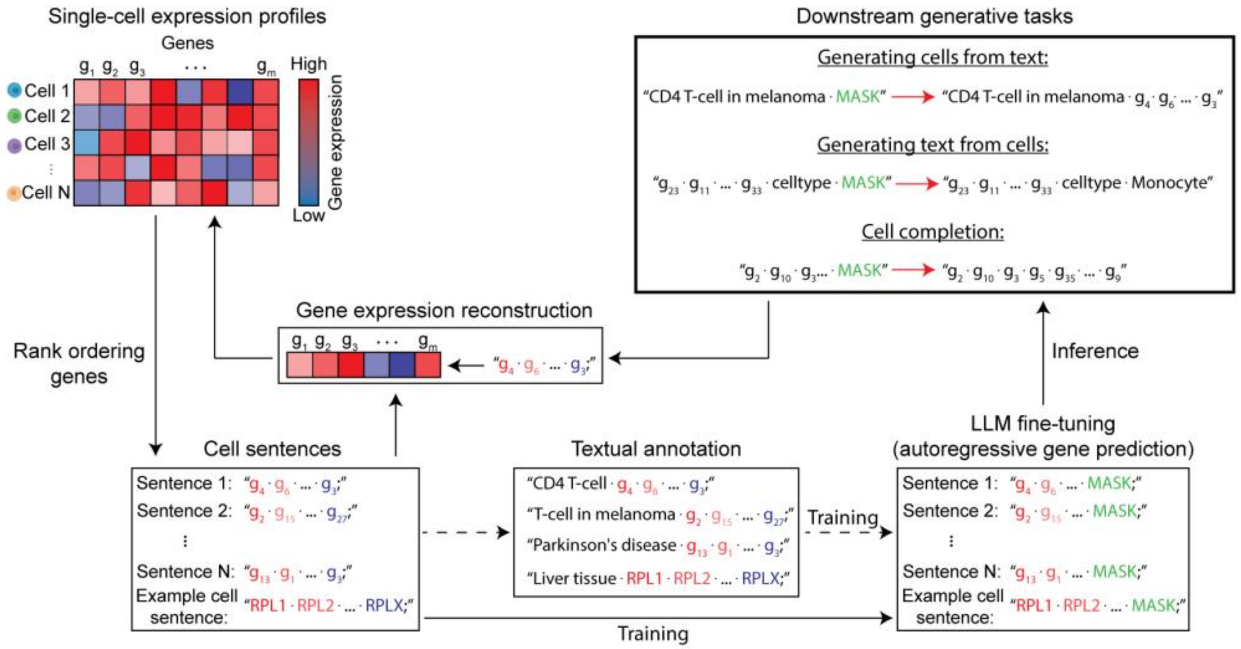

-Optimal transport algorithm based
-Move A → B
(cost function)

# • Foundation model in scRNA-seq

-Too many task
-Cannot train all kinds of task
→ Build versatile, general model for “every” task
→ Build with large enough data and parameters
-Training: autoregressive (self-supervised)
-Fine-tuning: same task, new data
-Prompt engineering: zero-shot learning

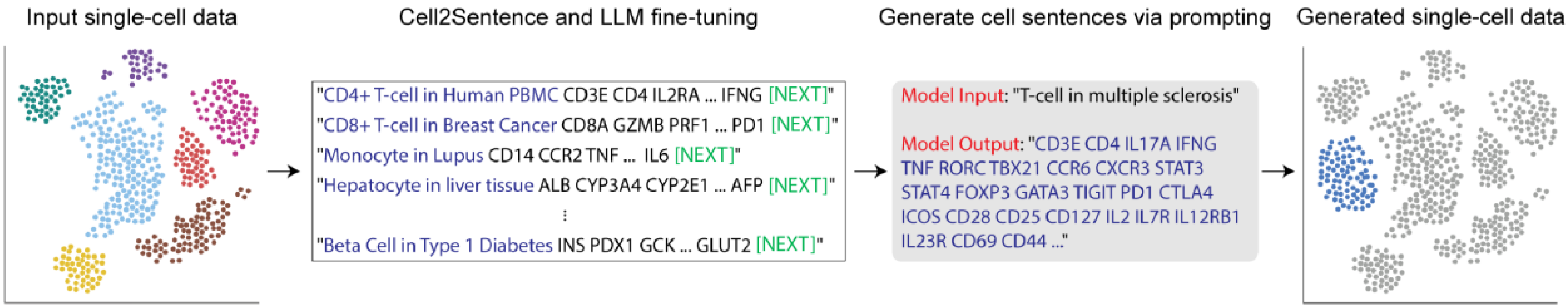


- Cell2Sentence




-Gene → log-norm → rank
-Celltype → gene sentence (convert embedding)
(Fine tuning by preexisting LLM: GPT-2)
-Usage: user cell type (text)
→ Cell type information (ex: marker genes)

- Cell2Sentence




**Cell Type Generation**

**Prompt:** Generate the 100 highest expressed genes listed in descending order for a long-lived plasma cell

**Response:** **MT-V1 RPS9 [...] RPS9 RPL8**

**Cell Label Prediction**

**Prompt:** Identify the cell type most likely associated with these 100 highly expressed genes listed in descending order: DIF3 RPS11 [. . . ] RPP4 RPS22

**Response:** **The cell type corresponding to these genes is a CD4-T cell.**

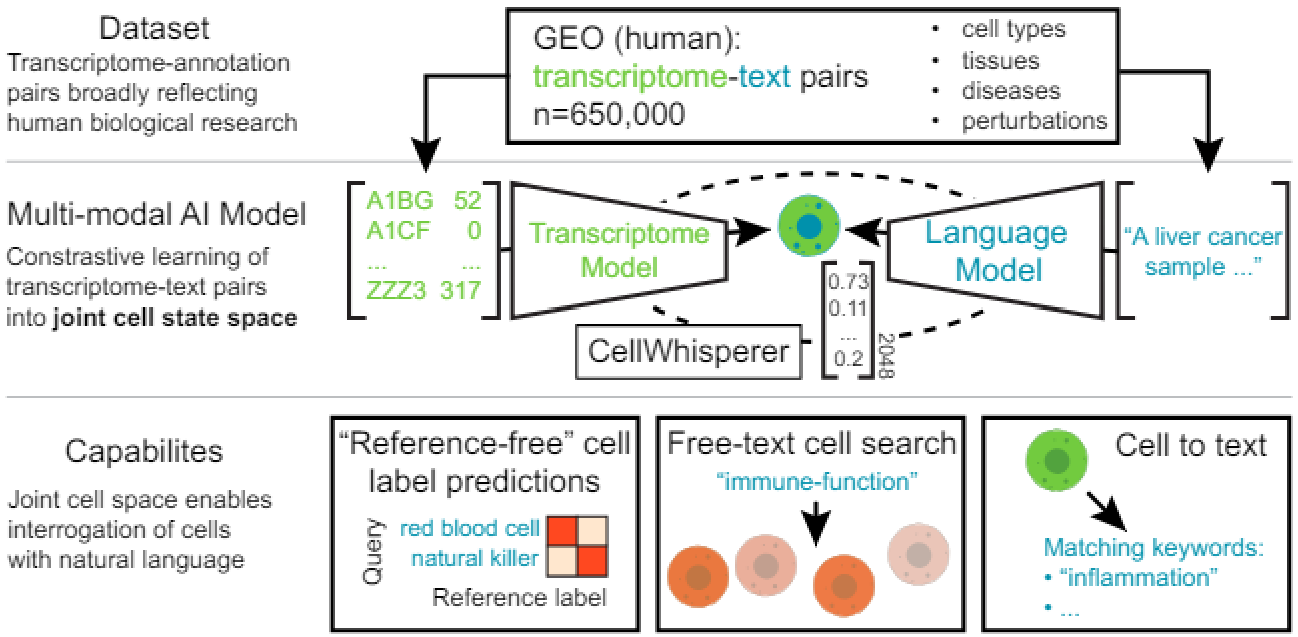
**Natural Language Insights from Data**

**Prompt:** Extract and summarize biological insights given this cell's 100 highest expressed genes: MALAT1 VIM [. . . ] RPL10A S100A6

**Response:** **This study used single-nucleus RNA sequencing to comprehensively map the cell types in the adult human brain.**

- CellWhisper

JOINT EMBEDDING OF TRANSCRIPTOMES AND TEXT ENABLES INTERACTIVE SINGLE-CELL RNA-SEQ DATA EXPLORATION VIA NATURAL LANGUAGE

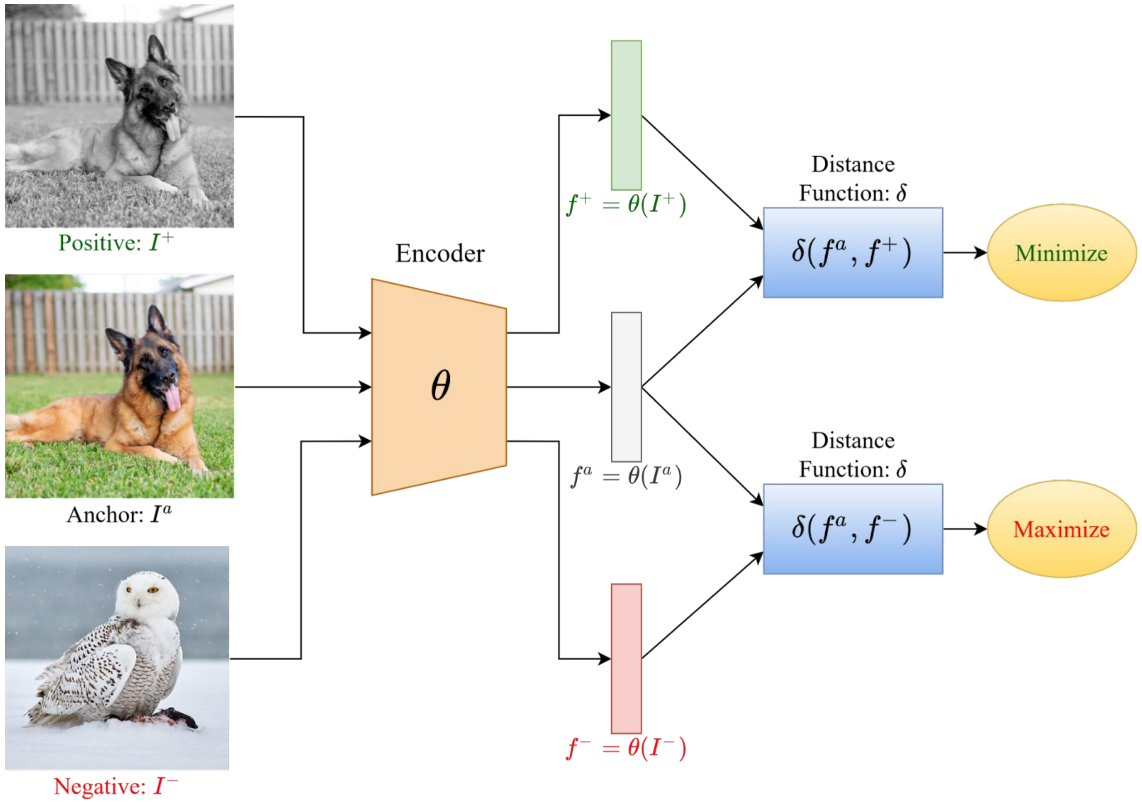


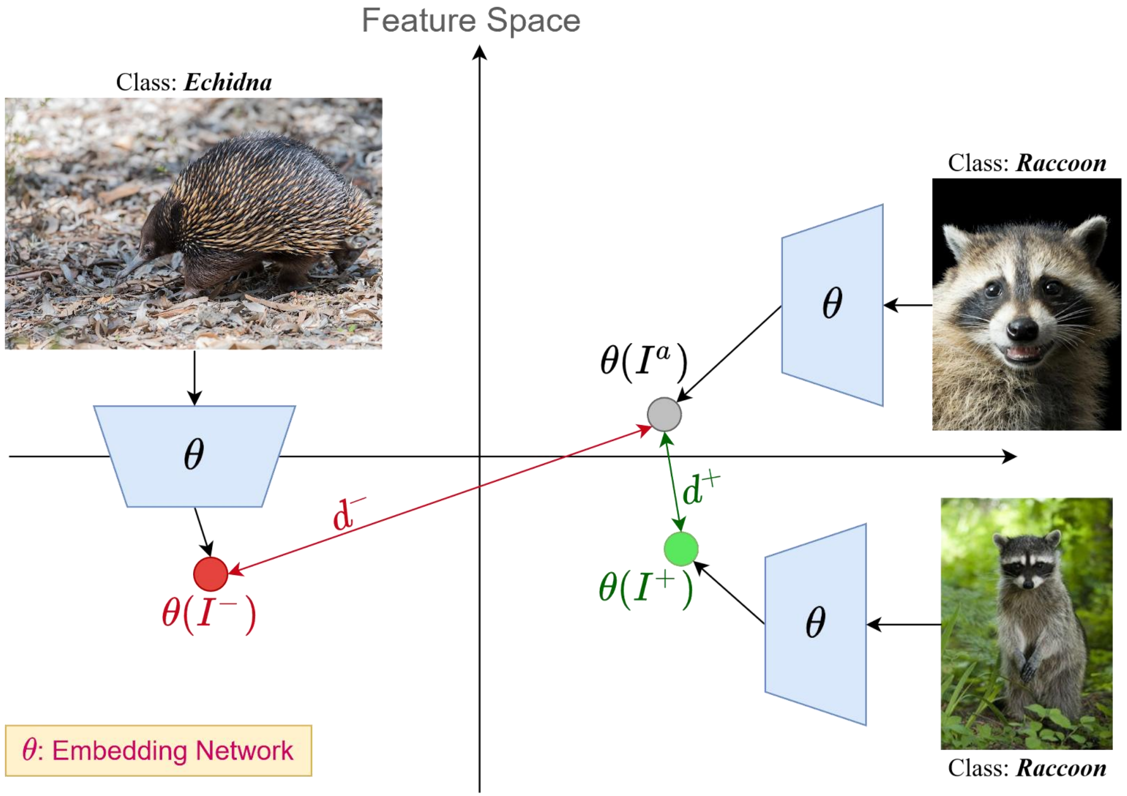

Geo (bulk data)
Transcriptome data (exp model: Geneformer)
Description (language model: BioBERT)
→ Joint space
Data is pairwise (description~transcriptome)

Contrast learning (cosine similarity)
→ Only pair → short distance
→ Wrong pair → long distance
→ Loss function

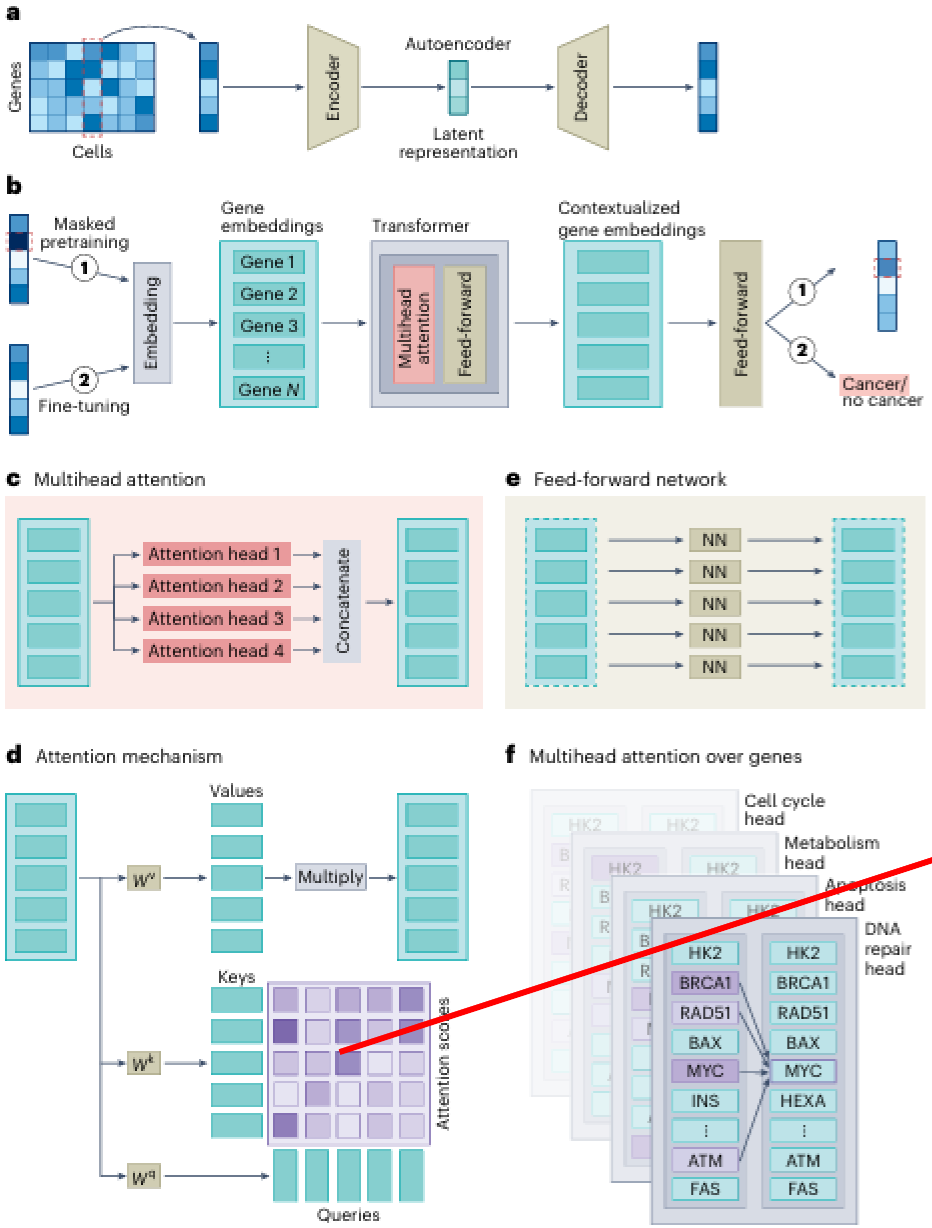
-Text → gene expression, celltype, tissue …

- Constrastive learning in scRNA-seq (text processing)



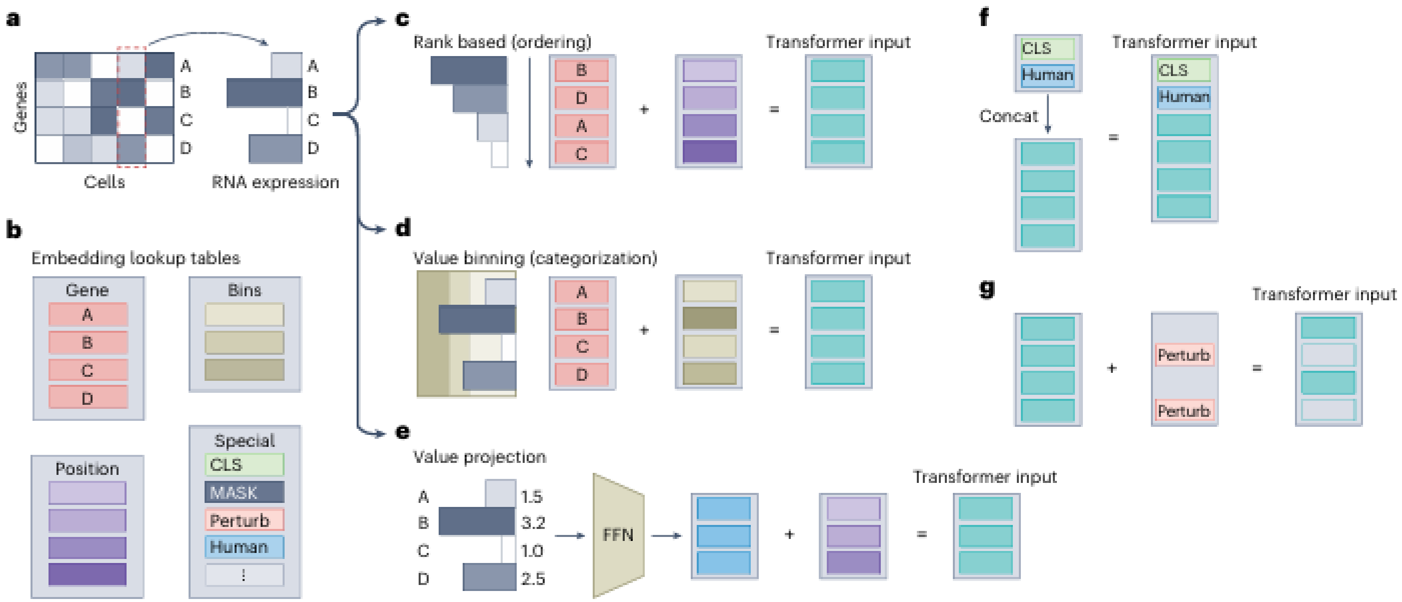
- Foundation model in scRNA-seq



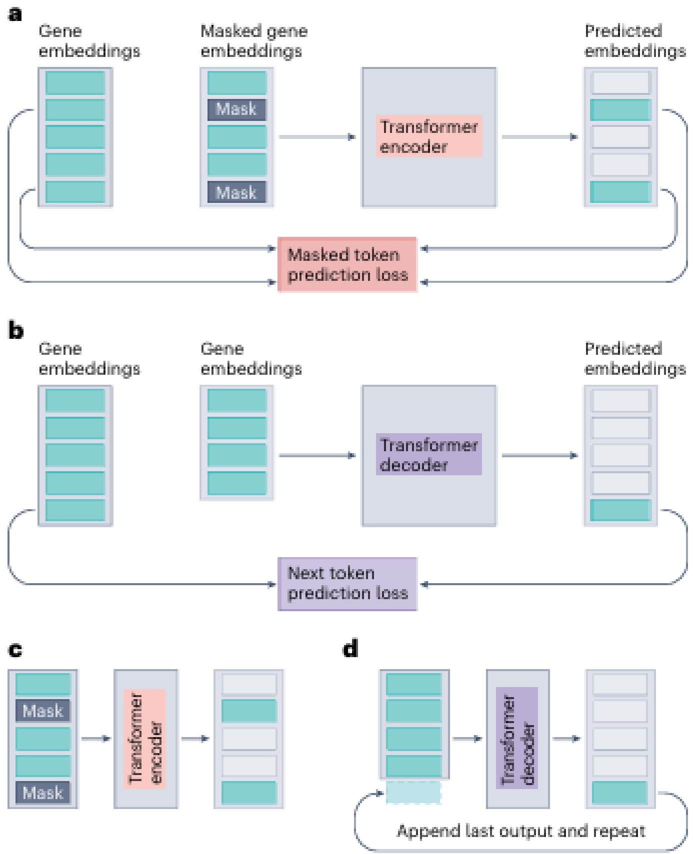

-Transformer based
→ Autoregressive learning: decoder → recapitulate the gene expression

- Attention score
→ gene-gene network

- Foundation model in scRNA-seq



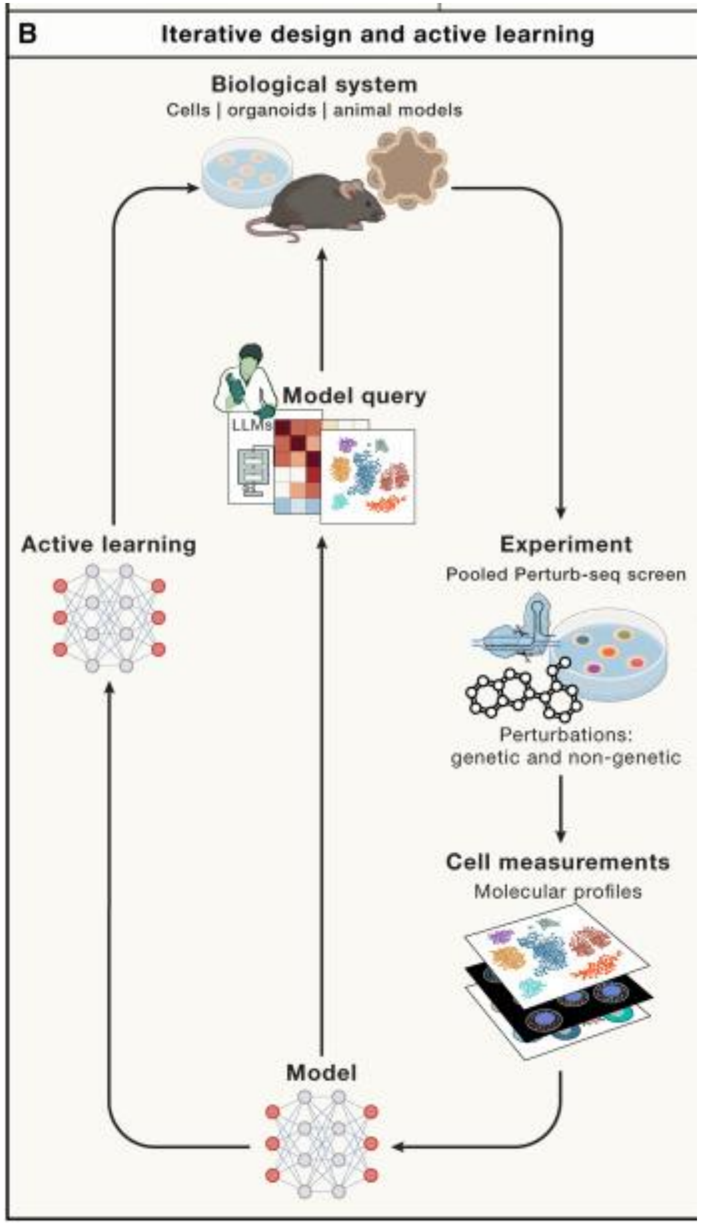

- How to give positional information




-masked attention → predict masked gene exp (endocder)

-Self-attention in decoder → predict next gene

# • Future direction of AI field in single-cell data




-In-silico experiment for unseen perturbation
→ Experimental validation
→ New hypothesis
→ In-silico experiment

Toward a foundation model of causal cell and tissue biology with a Perturbation Cell and Tissue Atlas

- Limitation of foundation model in scRNA-seq

nature methods

Brief Communication

https://doi.org/10.1038/s41592-025-02772-6

# Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines

Received: 11 October 2024

Accepted: 24 June 2025

Published online: 4 August 2025

Constantin Ahlmann-Eltze[1,2,3], Wolfgang Huber[2] & Simon Anders[1]

Recent research in deep-learning-based foundation models promises to learn representations of single-cell data that enable prediction of the effects