
# Large Models for Single-Cell Omics and Drug Discovery

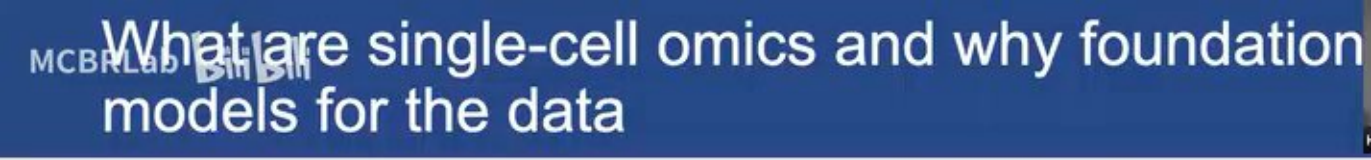
## Data, Pretraining, and Closed-Loop Environment

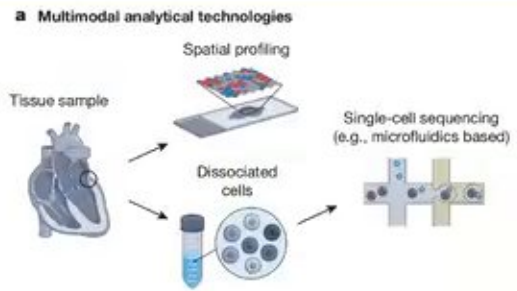
Haotian Cui

Sep 17, 2025

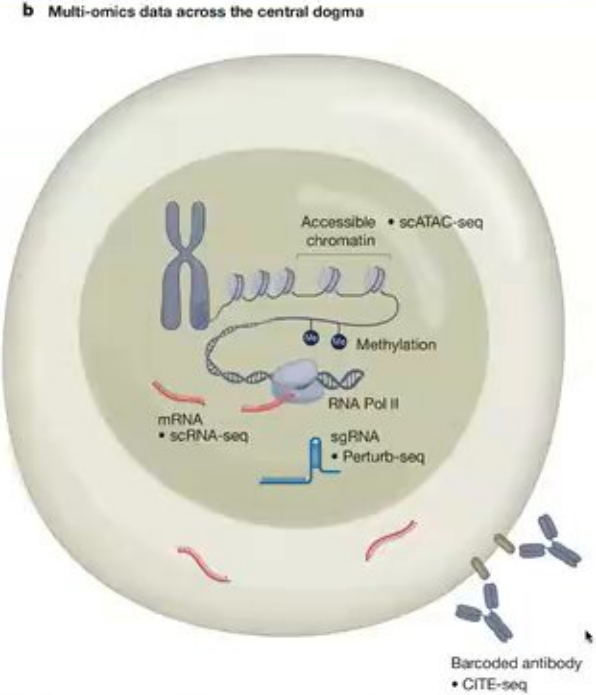

# Lessons from scGPT

# What are single-cell omics and why foundation models for the data



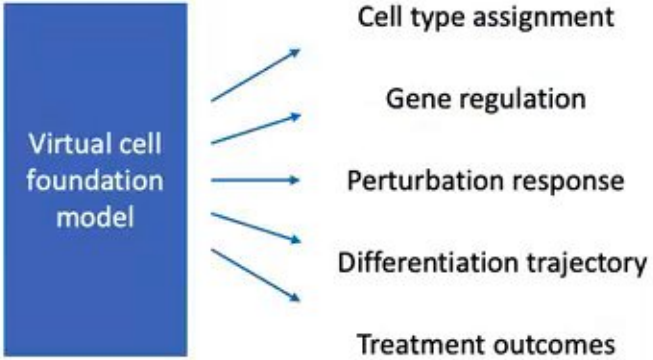

- **Cellular resolution**: revealing cell state heterogeneity
- **Muti-omics** enabled






Cui, H., Tejada-Lapuerta, A., Brbić, M., Saez-Rodriguez, J., Cristea, S., Goodarzi, H., Lotfollahi, M., Theis, F.J., and Wang, B. 2025. Towards Multimodal Foundation Models in Molecular Cell Biology. *Nature*, 640(8059), pp.623-633.

# What are single-cell omics and why foundation models for the data

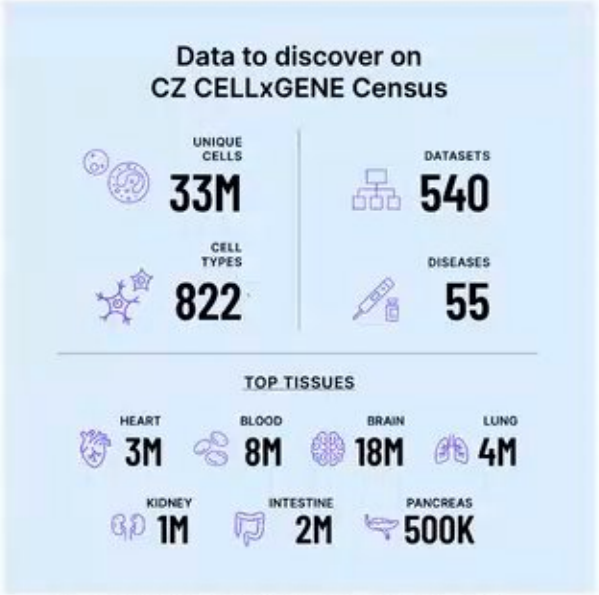



One FM supports varying applications

# Data and model arch consideration


- While texts are made up of words, cells can be characterized by genes.

- This is feasible based on two considerations:
  - **Transformers would be suitable** to learn interactions between gene tokens.
  - **Vast volume** of sequencing data, diverse cell types and conditions, and growing exponentially.

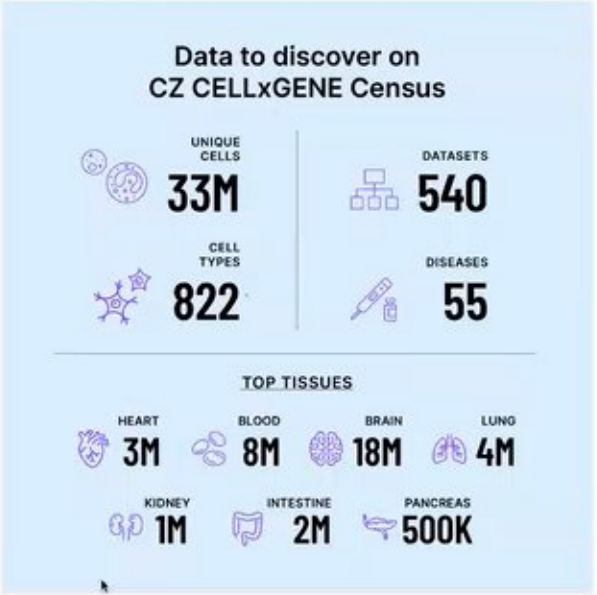



We are particularly excited about the recent launch of CellxGene census. Figure credit to @JCoolScience

# Data and model arch consideration


- While texts are made up of words, cells can be characterized by genes.

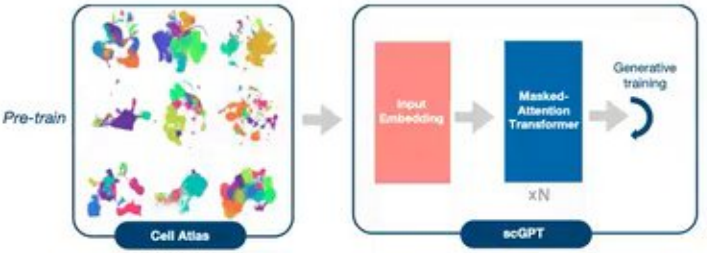
- This is feasible based on two considerations:
  - **Transformers would be suitable** to learn interactions between gene tokens.
  - **Vast volume** of sequencing data, diverse cell types and conditions, and growing exponentially.

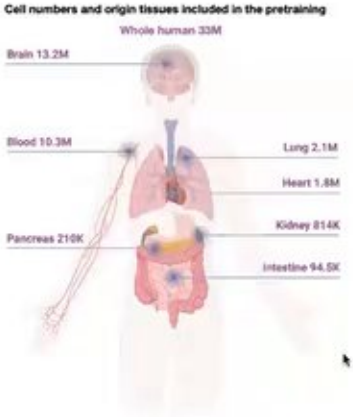



33M (2023)
to
108M (2025)

# Generative pretraining in scGPT


- In the pre-training:
  - Large-scale expression matrices were collected and preprocessed
  - We selected around 33 million normal human cells of 51 organs/tissues from the CZ CellxGene collection
  - Numbers of cells from major tissues shown on the right

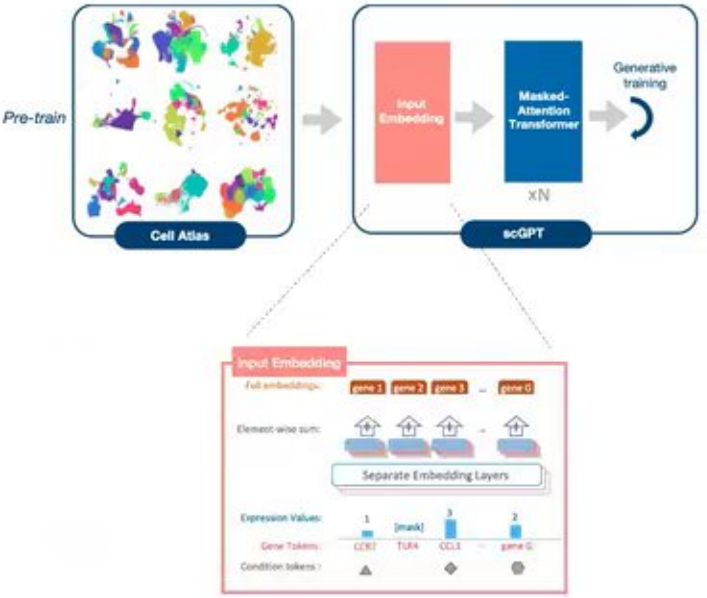


# Generative pretraining in scGPT


- In the pre-training:
  - Large-scale expression matrix were collected and preprocessed
  - We selected around 33 million normal human cells of 51 organs/tissues from the CZ CellxGene collection
  - The input contains three sets of tokens for: (1) gene identifiers (2) expression values (3) condition flags

$$\boldsymbol{h}^{(i)} = \mathbf{emb}_{\mathrm{g}}(\boldsymbol{t}_g^{(i)}) + \mathbf{emb}_{\mathrm{x}}(\boldsymbol{x}^{(i)}) + \mathbf{emb}_{\mathrm{c}}(\boldsymbol{t}_c^{(i)}).$$



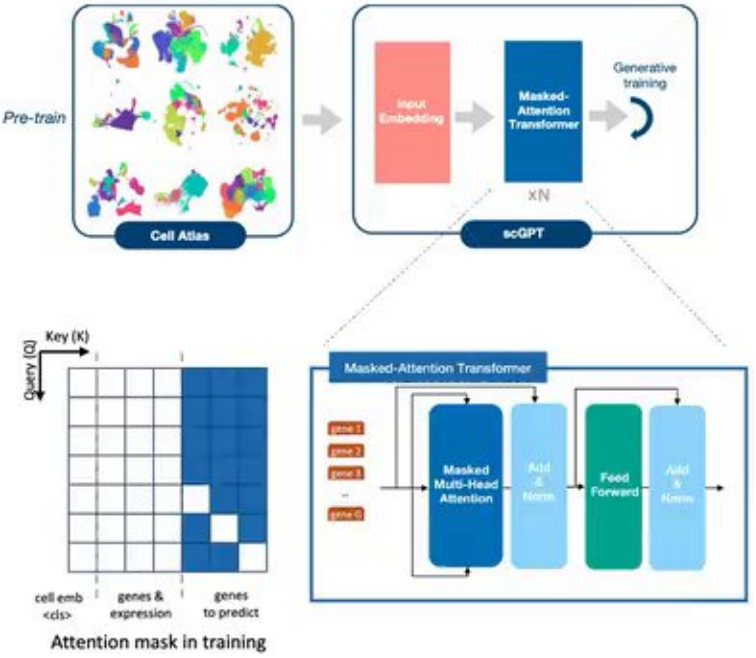

**Cui, H.**, Wang, C., Maan, H., Pang, K., Luo, F., Duan, N. and Wang, B., 2024. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. ***Nature Methods***, pp.1-11.

# Generative pretraining in scGPT


- In the pre-training:
  - Large-scale expression matrix were collected and preprocessed
  - We selected around 33 million normal human cells of 51 organs/tissues from the CZ CellxGene collection
  - The input contains three sets of tokens for: (1) gene identifiers (2) expression values (3) condition flags
  - The core model are 12 stacked transformer blocks, with **customized masked attention for generative training**

$$\boldsymbol{h}_0^{(i)} = \boldsymbol{h}^{(i)}$$

$$\boldsymbol{h}_l^{(i)} = \texttt{transformer\_block}(\boldsymbol{h}_{l-1}^{(i)}) \quad \forall l \in [1, n]$$

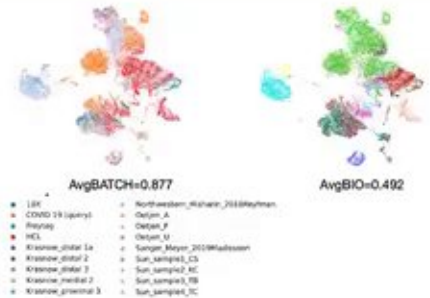



Attention mask in training

**Cui, H.**, Wang, C., Maan, H., Pang, K., Luo, F., Duan, N. and Wang, B., 2024. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. ***Nature Methods***, pp.1-11.

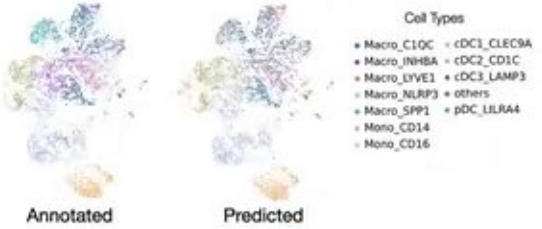
# scGPT supports varying applications


Data integration




Cell atlas query

Cell type assignment




Model based gene set enrichment

# Zero-shot ref map to cell atlas


- Zero-shot embeddings can further be used to map cells and propagate metainfo from large cell atlas
- With engineering efforts like fast ANN search (enabled by faiss), this can be used to query new cells super fast, https://github.com/bowang-lab/scGPT/blob/main/tutorials/Tutorial_Reference_Mapping.ipynb

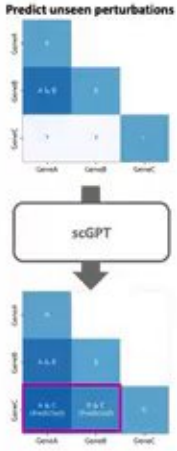
The search runs super fast, especially on GPU. Here the similarity search for 4,000 query cells within the whole reference of millions should take around 7 second on CPU and 0.1 second on GPU.

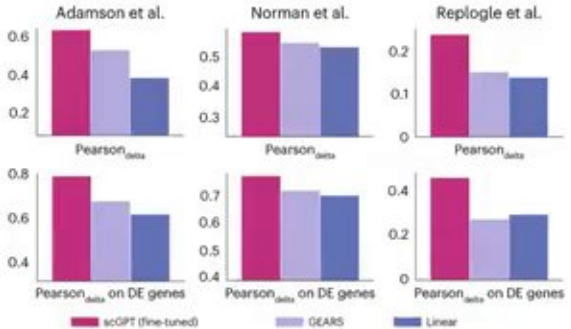
In [10]:
```
%%time
k = 50
# test with the first 100 cells
distances, idx = index.search(test_emebd, k)
```

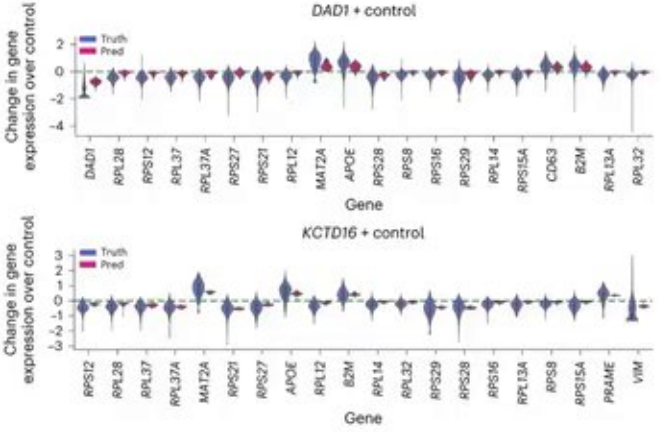
```
CPU times: user 252 ms, sys: 493 ms, total: 745 ms
Wall time: 133 ms
```

# Predicting genetic perturbation responses


- Finetune on some gene (combination) perturbations, predict on unseen ones



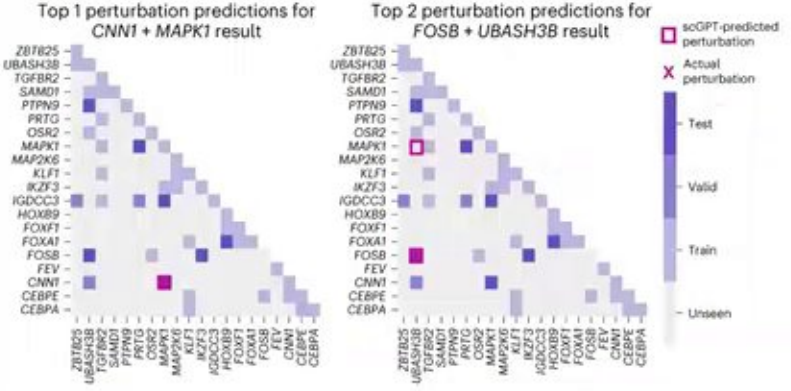



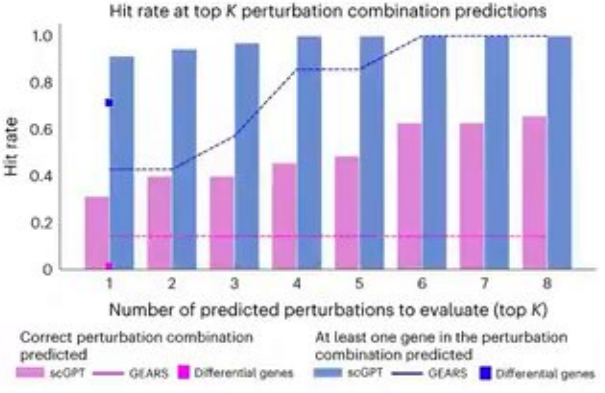
# Initial success of "reverse perturbation"


Reverse perturbation:

- Given the result cell state, predict the source perturbation (mimicking target id).




Cases showing success of identifications




Higher hit rate among seven test cases

# Reflections on scGPT capabilities


- A general pre-training workflow for single-cell data supports varying tasks

pypi package 0.2.4 | downloads 56k | WebApp up | Starred 1.2k

Growing popularity gives me the chance to communicate and learn lessons from users

# Reflections on scGPT capabilities


- A general pre-training workflow for single-cell data supports varying tasks

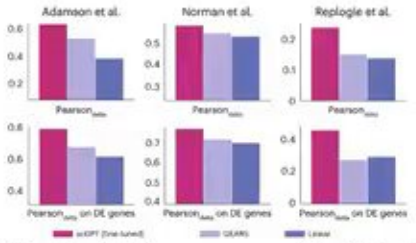

Growing popularity gives me the chance to communicate and learn lessons from users

What we have so far:

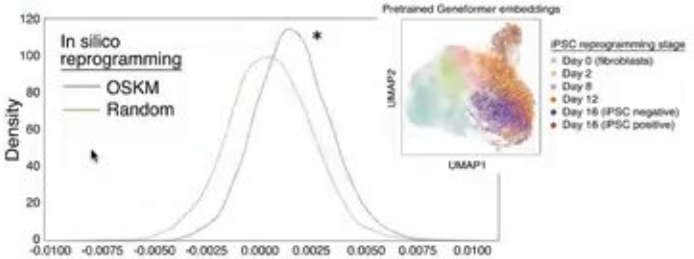
- "Biologically meaningful" cell and gene embeddings
- Several initial successes about perturbational prediction and driving gene id

What is missing:

- More subtle cell state capturing
- Reliable perturbational and dynamics modeling right after pretraining



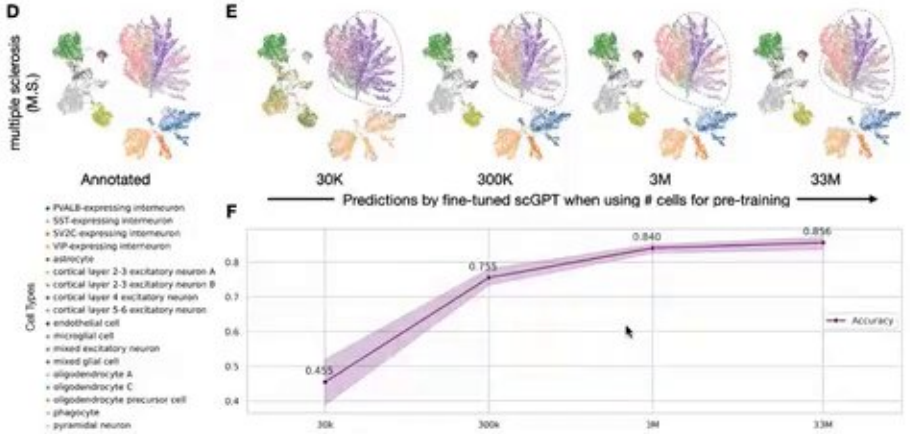

Evaluation of perturbation response prediction finetuning




Credit to GeneFromer, *in silico* perturbation towards iPSCs

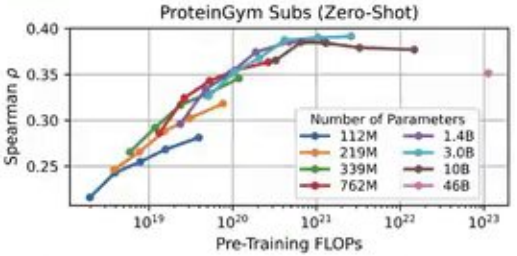
# Reflections on scaling law


- **Scaling pre-training data size**: larger pre-training data sizes yield superior pre-trained embeddings and further lead to improved performance on downstream tasks.



# Reflections on scaling law


- Scaling model size gives mixed results.
  - Based on new papers of varying size models.
  - Reports from ProGen3 in the related field of PLMs.
  - Similar phenomena in single-cell FMs and PLMs



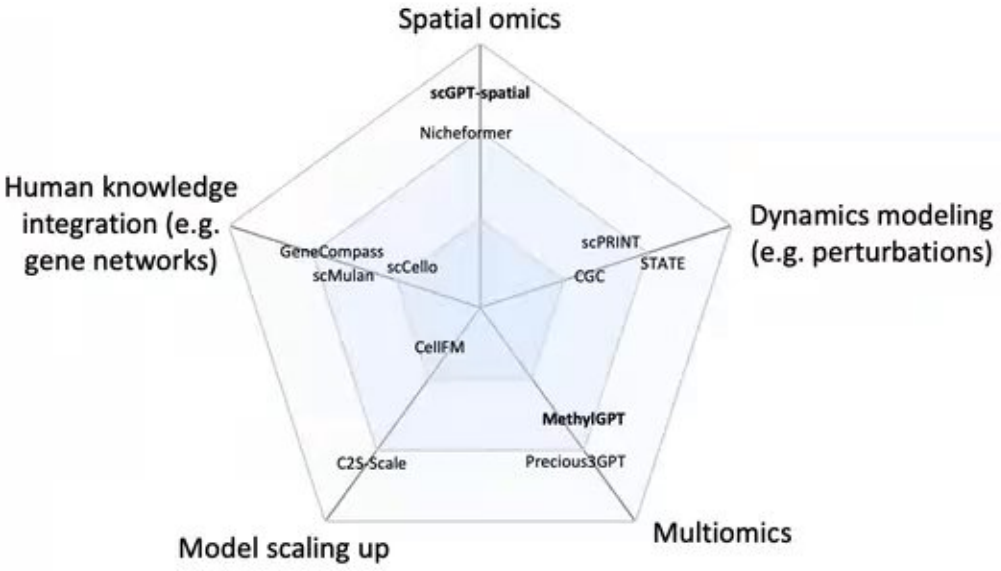

Credit to ProGen3, @_judewells. No benefit scaling **beyond 300m params.**


# Towards multimodal multitask pretraining

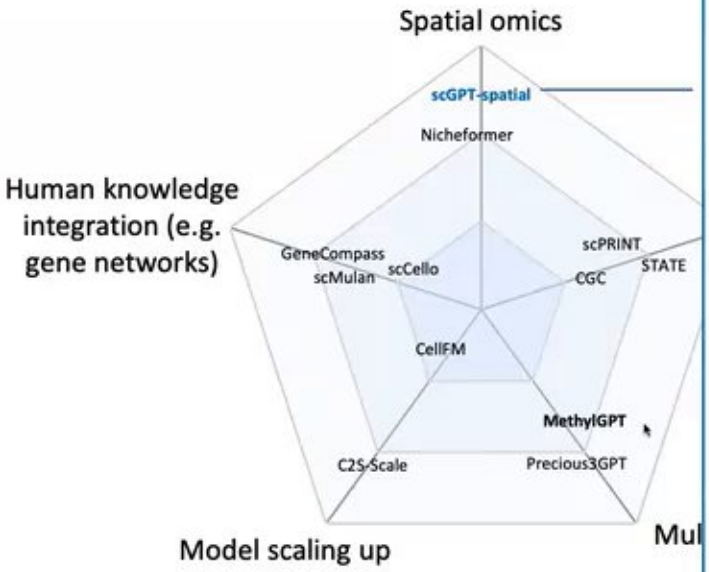
Towards building virtual cells

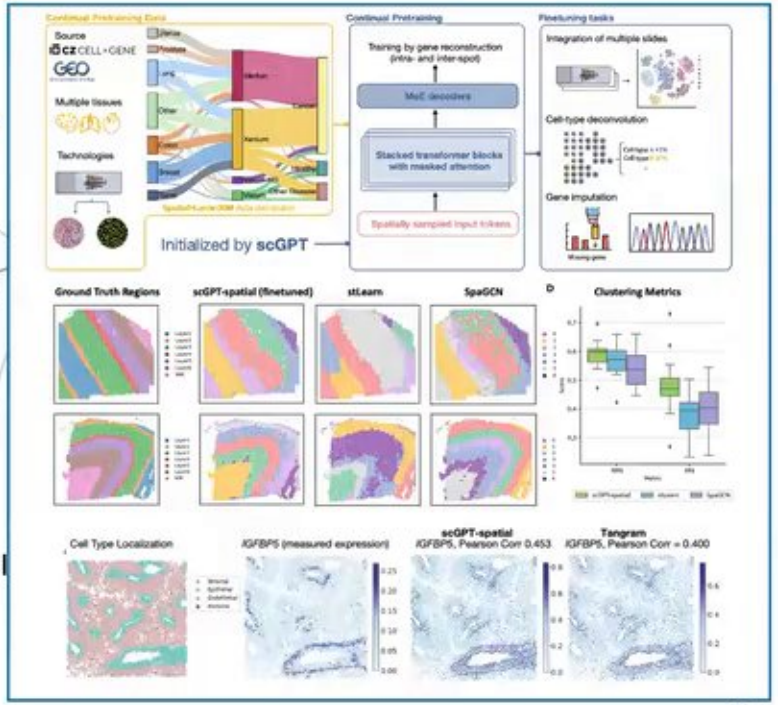
# Recent progress from us and others in the



# Recent progress from us and others in the

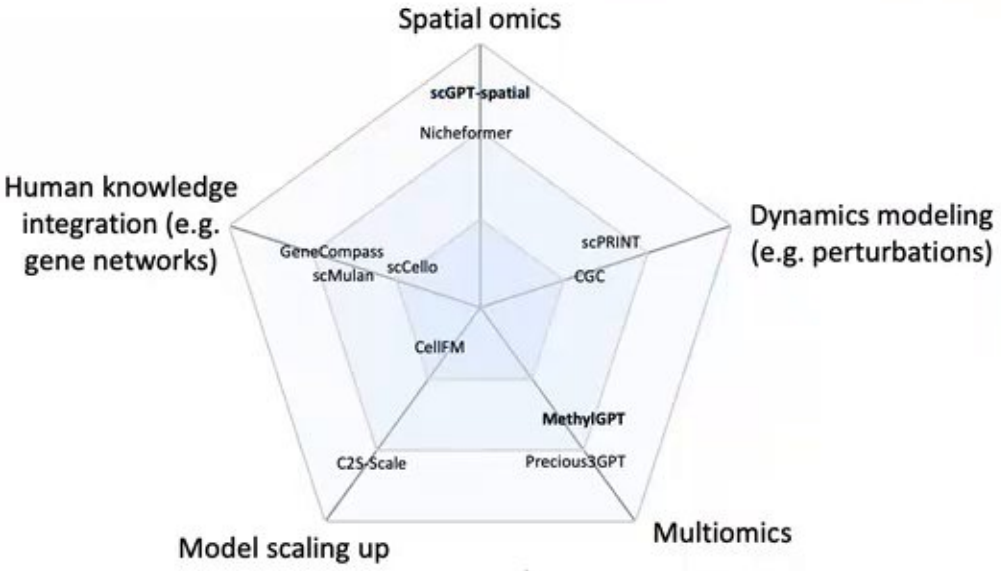





Wang, C.X.*, Cui, H.*, Zhang, A.H., Xie, R., Goodarzi, H. and Wang, B., 2025. scGPT-spatial: Continual Pretraining of Single-Cell Foundation Model for Spatial Transcriptomics. *bioRxiv*, pp.2025-02.

# Recent progress from us and others in the

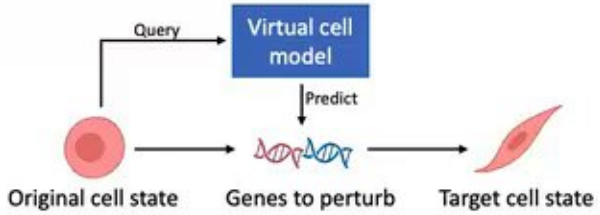




The definition of virtual cells can be diverse, what is the most relevant next steps?

# Why perturbational prediction, or dynamics mo in general

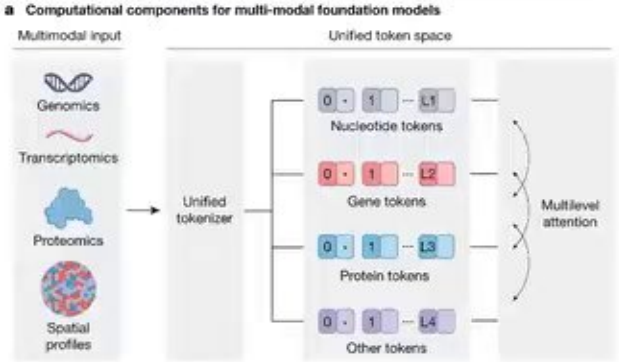
MCBRLab


- **Important new dimension**




Recall "one formula for many therapeutic applications"

# Model arch for unified multimodal pretraini




**Cui, H.**, Tejada-Lapuerta, A., Brbić, M., Saez-Rodriguez, J., Cristea, S., Goodarzi, H., Lotfollahi, M., Theis, F.J., and Wang, B. 2025. Towards Multimodal Foundation Models in Molecular Cell Biology. ***Nature***, 640(8059), pp.623-633.

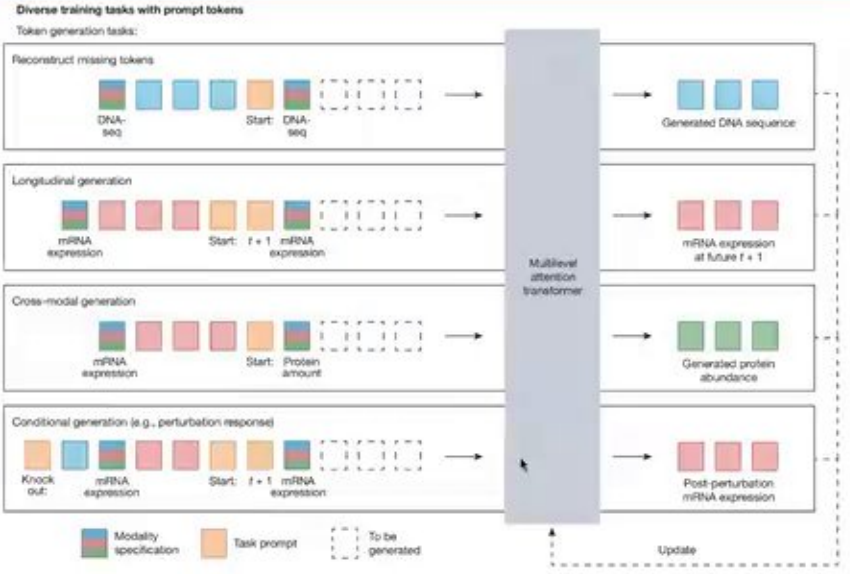
# Unified mutitask training


Gene expression reconstruction:

Temporal prediction:

Cross-modal generation:

**Perturbational training:**






**Cui, H.**, Tejada-Lapuerta, A., Brbić, M., Saez-Rodriguez, J., Cristea, S., Goodarzi, H., Lotfollahi, M., Theis, F.J., and Wang, B. 2025. Towards Multimodal Foundation Models in Molecular Cell Biology. ***Nature***, 640(8059), pp.623-633.


# Model guided data collection

An outer loop of active learning

# Improving efficiency with an outer loop of active learning

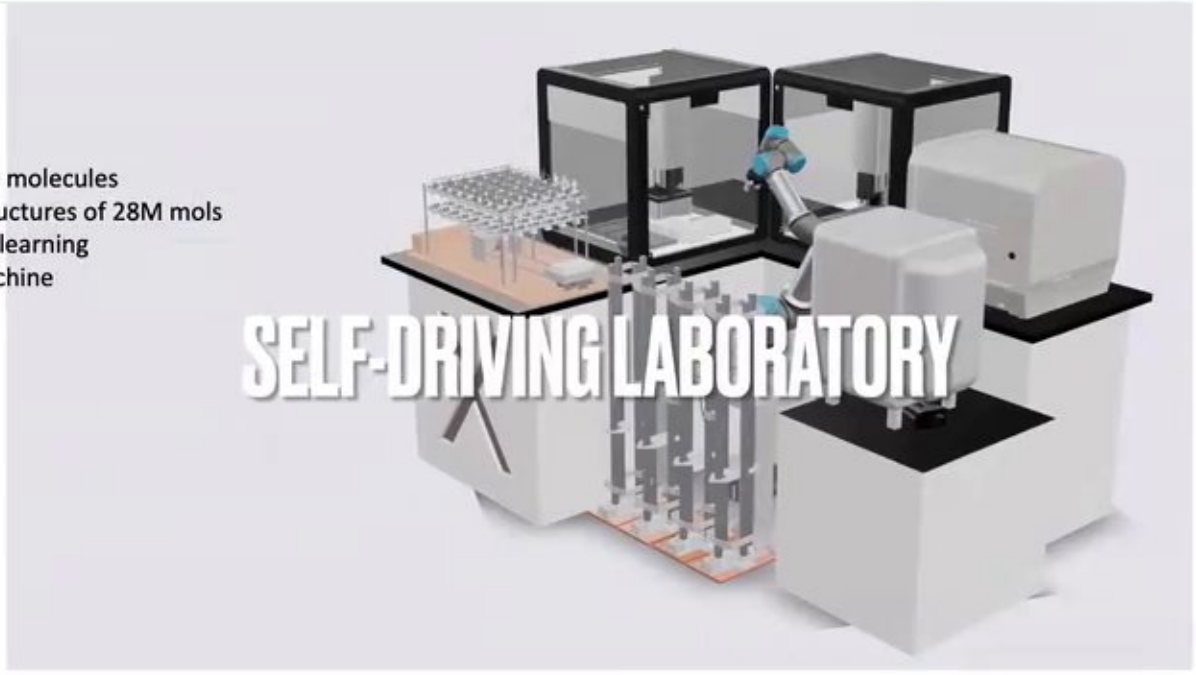
MCBRLab BiliBili


**Cui, H.***, Xu, Y.*, Pang, K., Li, G., Gong, F., Wang, B., Li, B., 2025. LUMI-lab: a Foundation Model-Driven Autonomous Platform Enabling Discovery of New Ionizable Lipid Designs for mRNA Delivery. ***bioRxiv***, pp.2025-02. (Under review, ***Cell***)

# Improving efficiency with an outer loop of active learning


LUMI-lab:

- **Active iterations** to make molecules
- **Model training** on 3D structures of 28M mols
- Propose, validate, online learning
- **100** new LNPs / day / machine

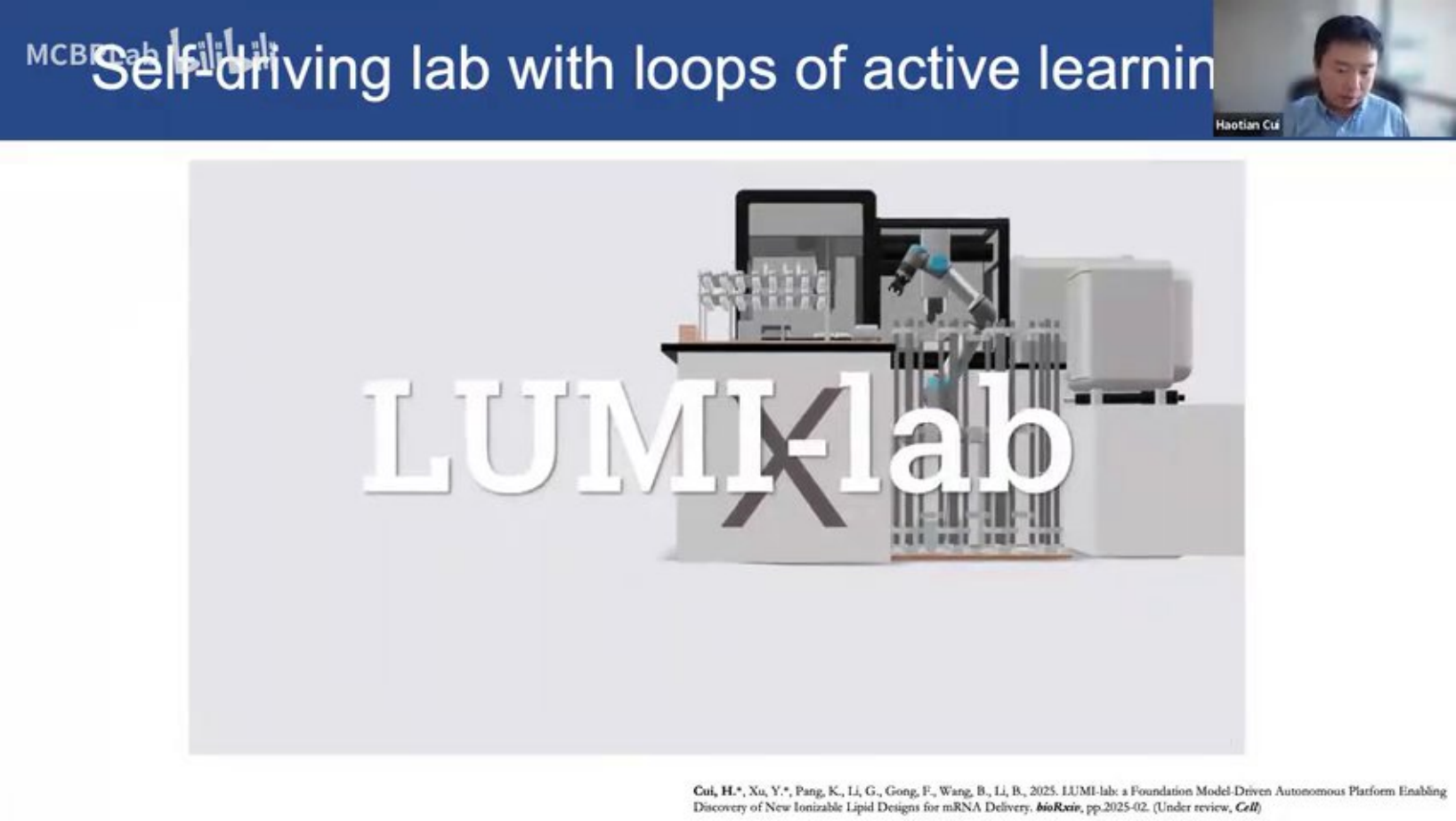



**Cui, H.***, Xu, Y.*, Pang, K., Li, G., Gong, F., Wang, B., Li, B., 2025. LUMI-lab: a Foundation Model-Driven Autonomous Platform Enabling Discovery of New Ionizable Lipid Designs for mRNA Delivery. ***bioRxiv***, pp.2025-02. (Under review, ***Cell***)

# Self-driving lab with loops of active learnin

**Cui, H.***, Xu, Y.*, Pang, K., Li, G., Gong, F., Wang, B., Li, B., 2025. LUMI-lab: a Foundation Model-Driven Autonomous Platform Enabling Discovery of New Ionizable Lipid Designs for mRNA Delivery. ***bioRxiv***, pp.2025-02. (Under review, ***Cell***)

# Self-driving lab with loops of active learnin







**Cui, H.***, Xu, Y.*, Pang, K., Li, G., Gong, F., Wang, B., Li, B., 2025. LUMI-lab: a Foundation Model-Driven Autonomous Platform Enabling Discovery of New Ionizable Lipid Designs for mRNA Delivery. ***bioRxiv***, pp.2025-02. (Under review, ***Cell***)

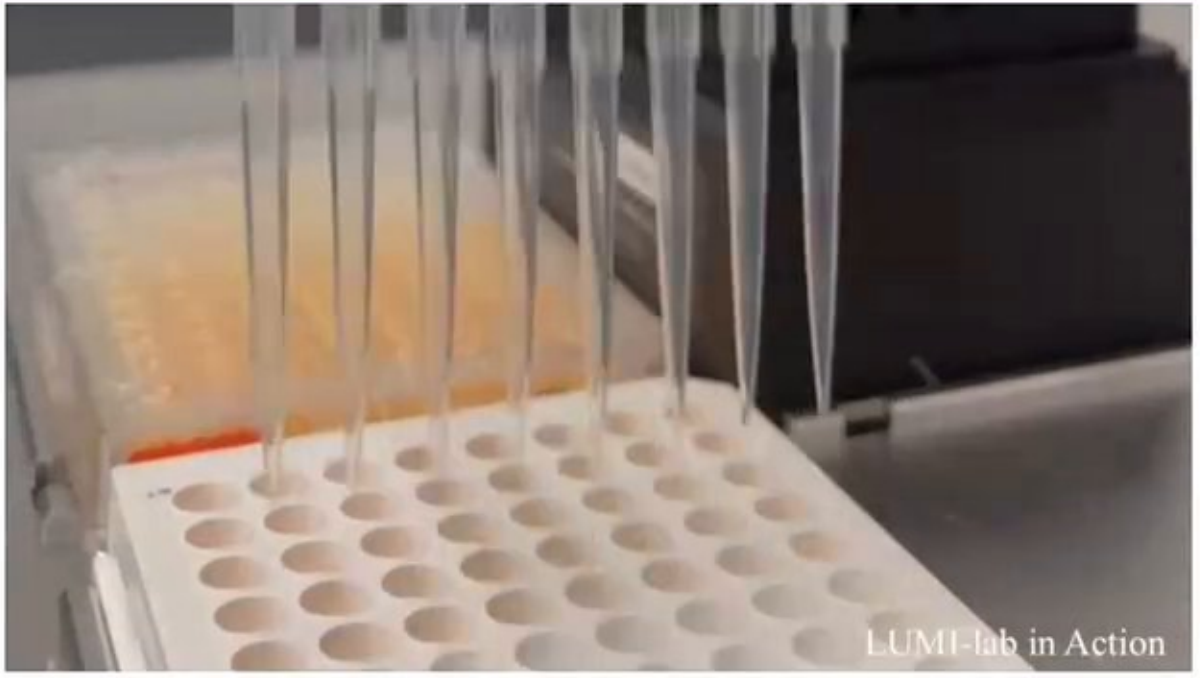
MCBR Lab

# Self-driving lab with loops of active learnin



**Cui, H.***, Xu, Y.*, Pang, K., Li, G., Gong, F., Wang, B., Li, B., 2025. LLMI-lab: a Foundation Model-Driven Autonomous Platform Enabling Discovery of New Ionizable Lipid Designs for mRNA Delivery. ***bioRxiv***, pp.2025-02. (Under review, ***Cell***)

# Self-driving lab with loops of active learnin

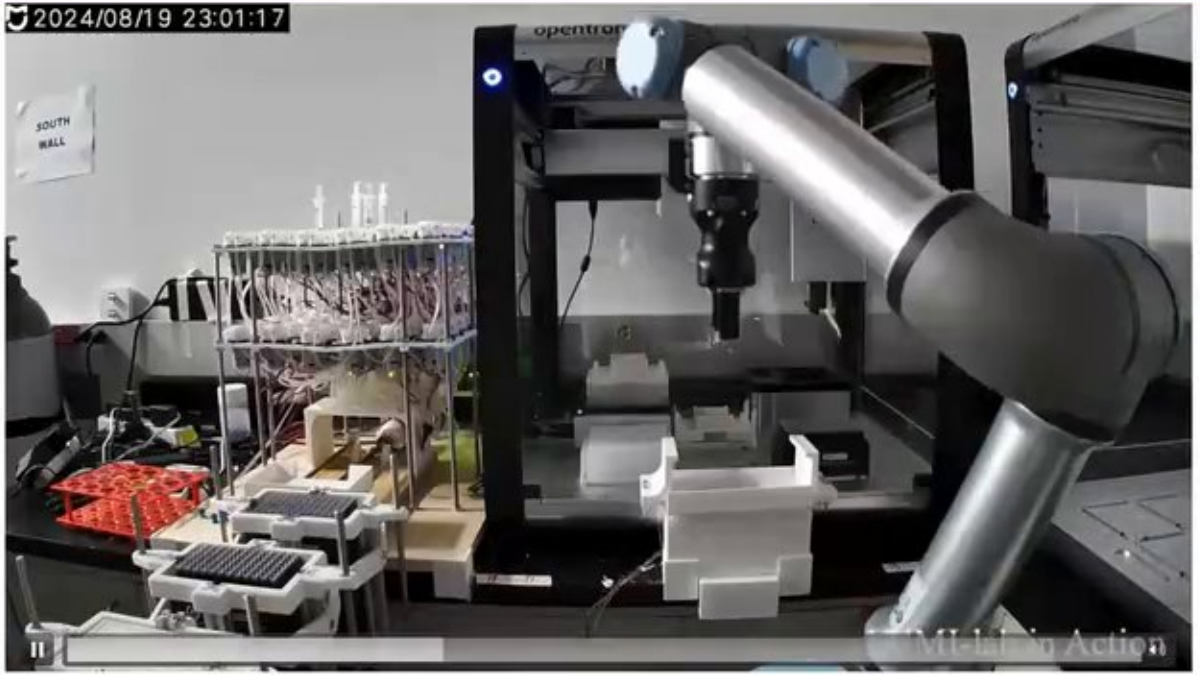

**Cui, H.***, Xu, Y.*, Pang, K., Li, G., Gong, F., Wang, B., Li, B., 2025. LUMI-lab: a Foundation Model-Driven Autonomous Platform Enabling Discovery of New Ionizable Lipid Designs for mRNA Delivery. ***bioRxiv***, pp.2025-02. (Under review, ***Cell***)

# Self-driving lab with loops of active learnin

**Cui, H.***, Xu, Y.*, Pang, K., Li, G., Gong, F., Wang, B., Li, B., 2025. LUMI-lab: a Foundation Model-Driven Autonomous Platform Enabling Discovery of New Ionizable Lipid Designs for mRNA Delivery. ***bioRxiv***, pp.2025-02. (Under review, ***Cell***)

MCBR Lab

# Self-driving lab with loops of active learnin


**Cui, H.***, Xu, Y.*, Pang, K., Li, G., Gong, F., Wang, B., Li, B., 2025. LUMI-lab: a Foundation Model-Driven Autonomous Platform Enabling Discovery of New Ionizable Lipid Designs for mRNA Delivery. ***bioRxiv***, pp.2025-02. (Under review, ***Cell***)

MCBR Lab

# Self-driving lab with loops of active learnin

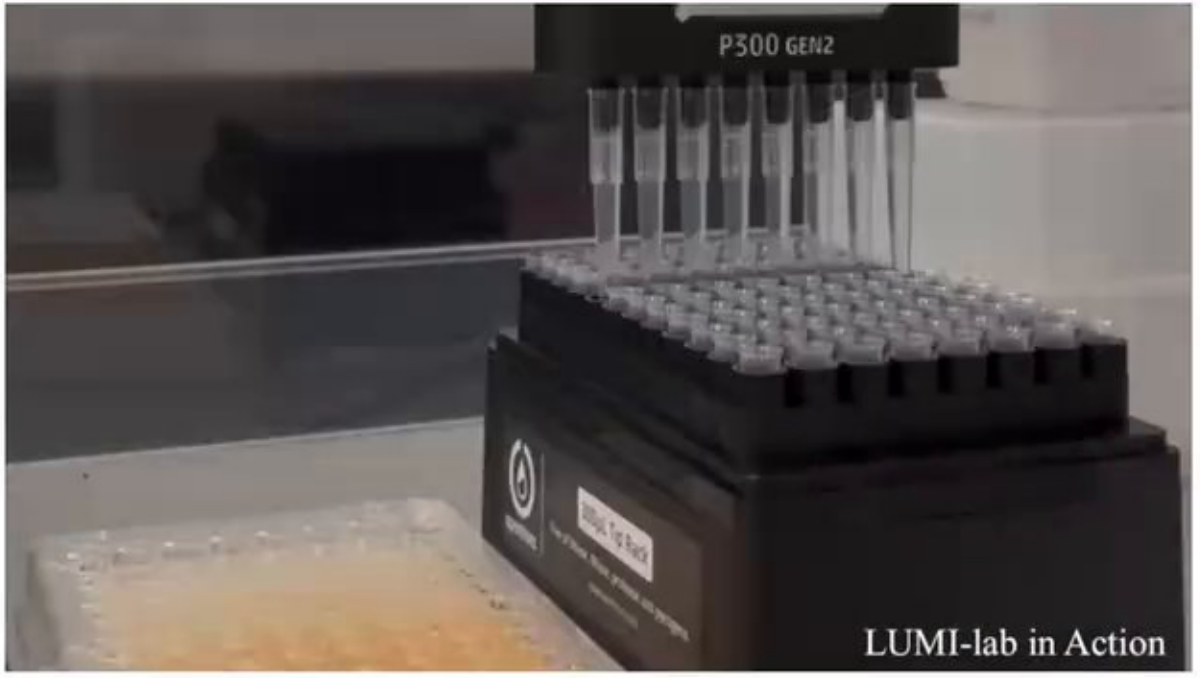

**Cui, H.***, Xu, Y.*, Pang, K., Li, G., Gong, F., Wang, B., Li, B., 2025. LUMI-lab: a Foundation Model-Driven Autonomous Platform Enabling Discovery of New Ionizable Lipid Designs for mRNA Delivery. ***bioRxiv***, pp.2025-02. (Under review, ***Cell***)

# Self-driving lab with loops of active learnin


**Cui, H.***, Xu, Y.*, Pang, K., Li, G., Gong, F., Wang, B., Li, B., 2025. LUMI-lab: a Foundation Model-Driven Autonomous Platform Enabling Discovery of New Ionizable Lipid Designs for mRNA Delivery. ***bioRxiv***, pp.2025-02. (Under review, ***Cell***)

MCBR Lab

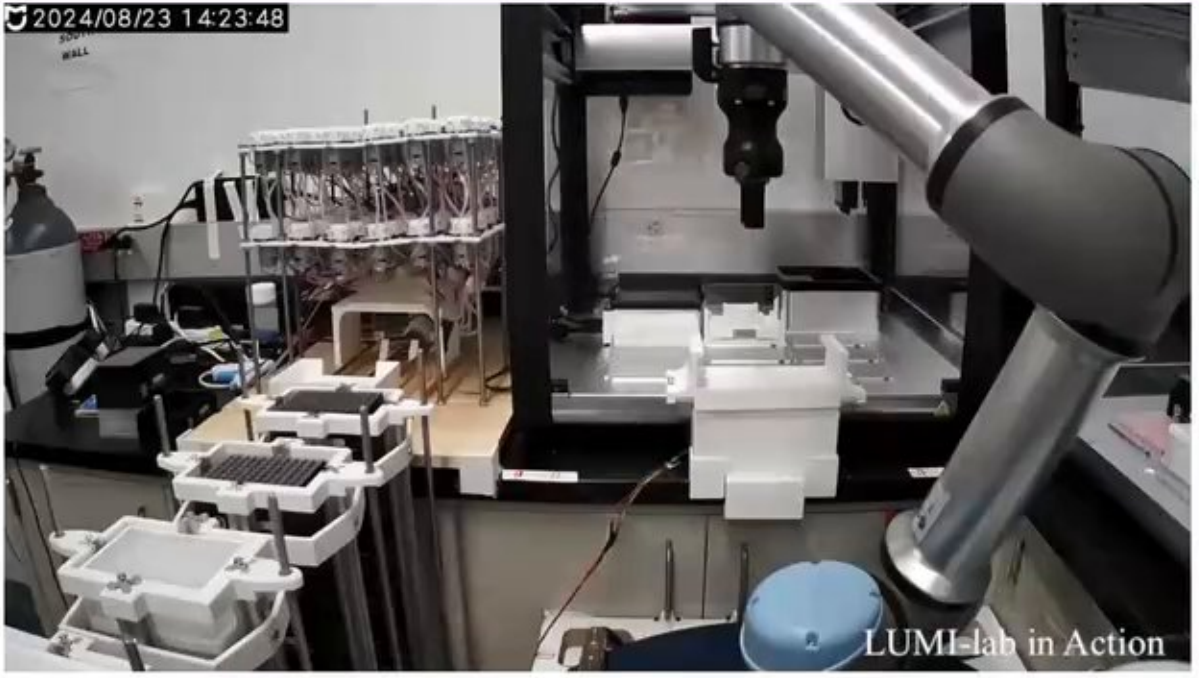
# Self-driving lab with loops of active learnin


**Cui, H.***, Xu, Y.*, Pang, K., Li, G., Gong, F., Wang, B., Li, B., 2025. LUMI-lab: a Foundation Model-Driven Autonomous Platform Enabling Discovery of New Ionizable Lipid Designs for mRNA Delivery. ***bioRxiv***, pp.2025-02. (Under review, ***Cell***)

MCBE Lab

# Self-driving lab with loops of active learnin


**Cui, H.***, Xu, Y.*, Pang, K., Li, G., Gong, F., Wang, B., Li, B., 2025. LUMI-lab: a Foundation Model-Driven Autonomous Platform Enabling Discovery of New Ionizable Lipid Designs for mRNA Delivery. ***bioRxiv***, pp.2025-02. (Under review, ***Cell***)

MCBE Lab @ Pitt

# Self-driving lab with loops of active learnin



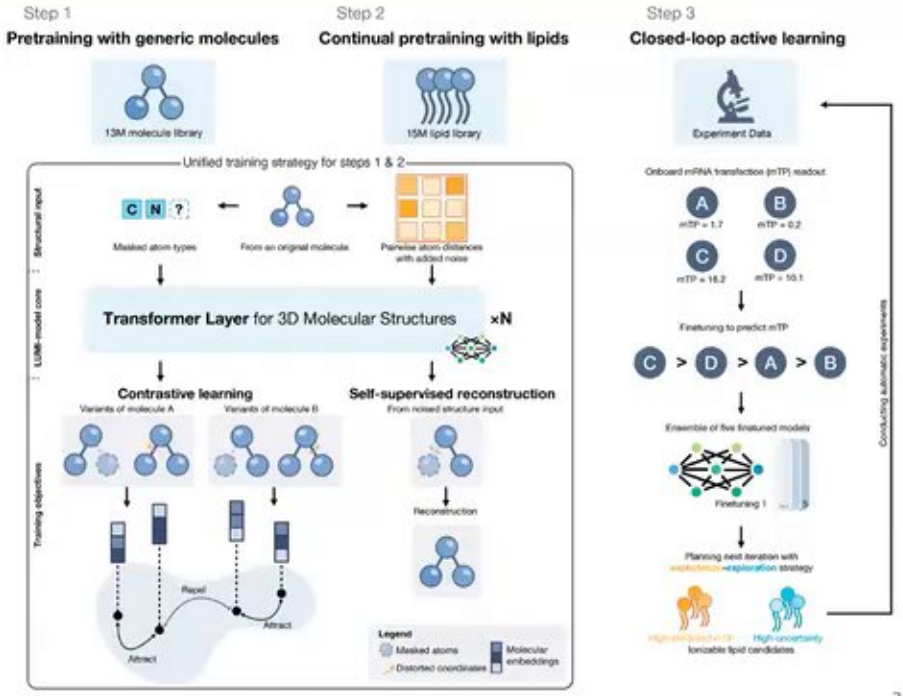
**Cui, H.***, Xu, Y.*, Pang, K., Li, G., Gong, F., Wang, B., Li, B., 2025. LUMI-lab: a Foundation Model-Driven Autonomous Platform Enabling Discovery of New Ionizable Lipid Designs for mRNA Delivery. ***bioRxiv***, pp.2025-02. (Under review, ***Cell***)

# Self-driving lab with loops of active learnin


**Cui, H.***, Xu, Y.*, Pang, K., Li, G., Gong, F., Wang, B., Li, B., 2025. LUMI-lab: a Foundation Model-Driven Autonomous Platform Enabling Discovery of New Ionizable Lipid Designs for mRNA Delivery. ***bioRxiv***, pp.2025-02. (Under review, ***Cell***)

# Three-stage training in LUMI-lab


Stage 1. 13M generic mols
Stage 2. 15M lipid-like mols
Stage 3. active learning with SDL

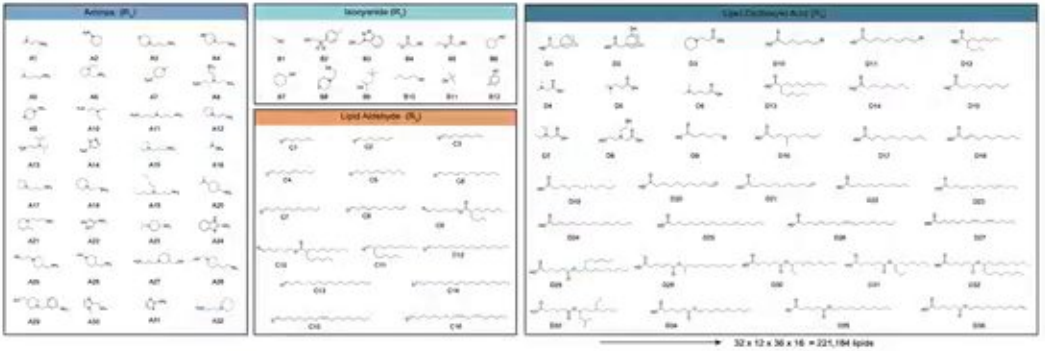




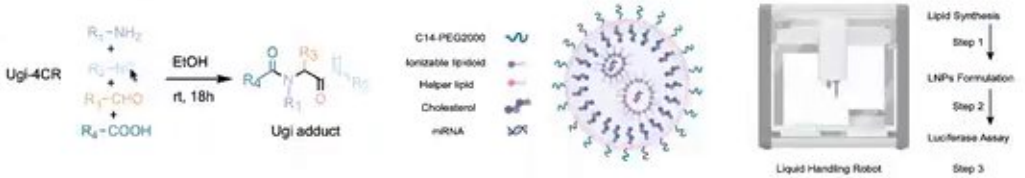

**Cui, H.***, Xu, Y.*, Pang, K., Li, G., Gong, F., Wang, B., Li, B., 2025. LUMI-lab: a Foundation Model-Driven Autonomous Platform Enabling Discovery of New Ionizable Lipid Designs for mRNA Delivery. ***bioRxiv***, pp.2025-02. (Under review, ***Cell***)

# Model proposing candidates with combinatorial molecular structures

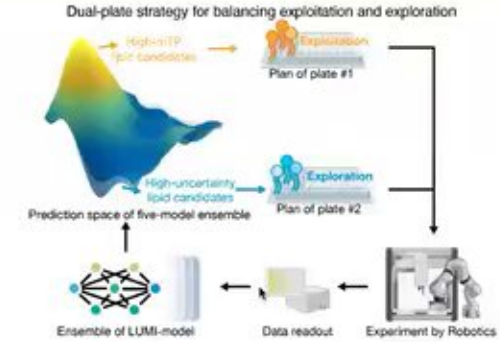

Possible experiment space of:
32 x 12 x 36 x 16 = 221,184 lipids




Unified synthesis approach for any candidate

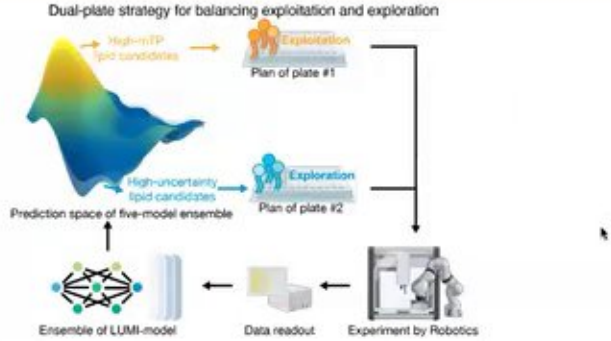


# Identifying new family of delivery vehicles

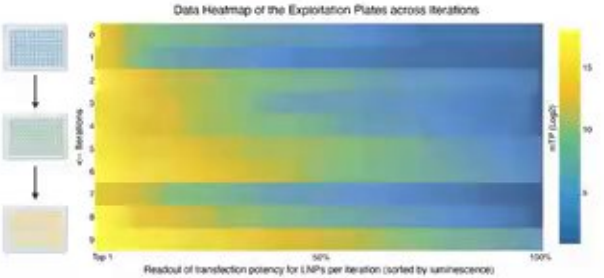




(1) Active learning of 10 iters, balancing exploration-exploitation

# Identifying new family of delivery vehicles


(1) Active learning of 10 iters, balancing exploration-exploitation




(2) Consistent increasing performance among iters

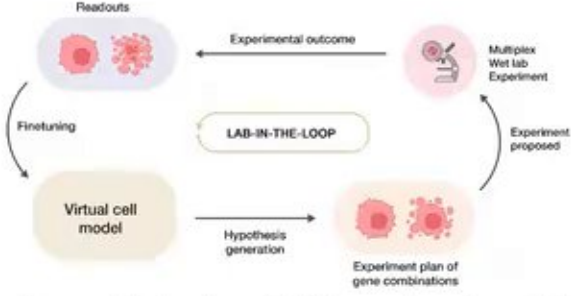
# Outer loop of active learning for perturbational training and exploration


- Empirical conclusion by LUMI-lab: guided exploration (**data collection**) and optimization (**model**) in a local minima (**question of interest**) of a large space can be efficient.
  - Another recent work with similar concept in protein design, EvolvePro (Jiang et al., Science 2024)
- The workflow for perturbational training:

**Initialization**: genome scale perturbation data (e.g. one-gene, cell line) that enabled a base modeling

**Given specific context:** costimulator (up to two gene combination) to enhance CAR-T. Select a pool of 1000 candidates.

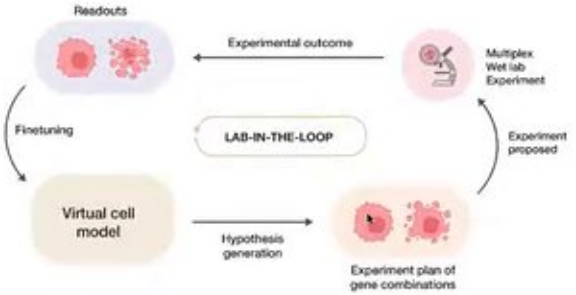
**Iterations:**




In each iter, only a multiplex (e.g. 96x2) experiments will be actively proposed

# Outer loop of active learning for perturbational training and exploration


**Iteration loops:**




**Features:**

- Since small scale, may go beyond cell lines and test on **primary cells or organoids**
- Accumulating most valuable data, can be used in future pretraining

# Summary


- Lessons learned from scGPT, importance of training and inference alignment
- Trends for building future virtual cell models, multimodal and multitask training
- An outer loop of active learning to enhance data and training efficiency


# Thank You!

**Acknowledgements:**

Dr. Bo Wang, Chloe Wang, Hassaan Maan, Ronald Xie, Kuan Pang, Reagan Li, Dr. York Xu, Dr. Bowen Li, Dr. Fabian Theis, Dr. Michael Taylor, Dr. Nan Duan, Dr. Gary Bader and other collaborators.

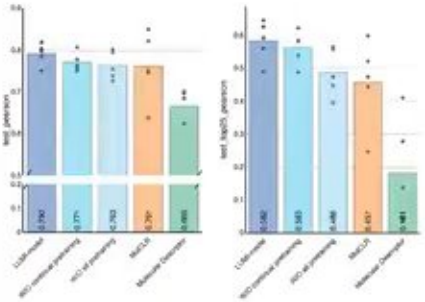
UNIVERSITY OF TORONTO

VECTOR INSTITUTE

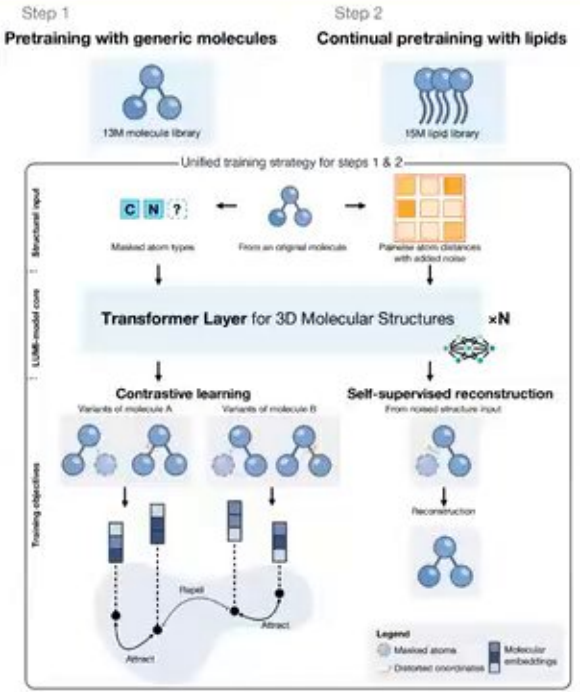
UHN

# Three-stage training in LUMI-lab


Stage 1. 13M generic mols
Stage 2. 15M lipid-like mols
Stage 3. active learning with SDL



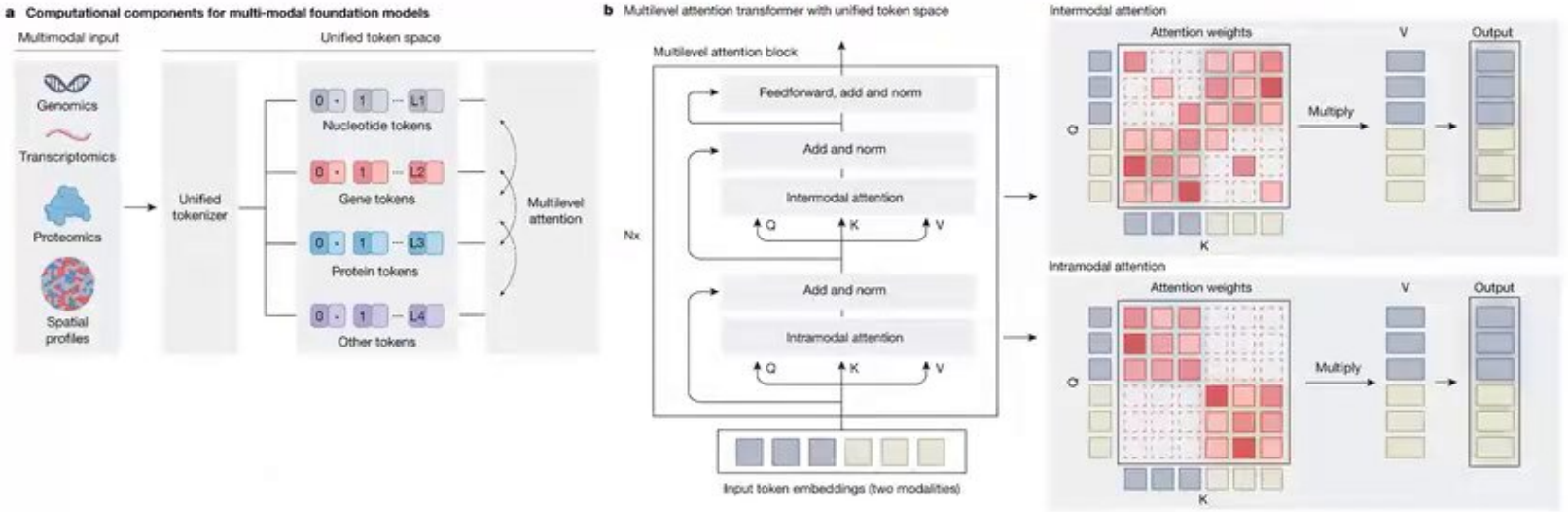
Effectiveness of stages 1 and 2.






**Cui, H.***, Xu, Y.*, Pang, K., Li, G., Gong, F., Wang, B., Li, B., 2025. LUMI-lab: a Foundation Model-Driven Autonomous Platform Enabling Discovery of New Ionizable Lipid Designs for mRNA Delivery. ***bioRxiv***, pp.2025-02. (Under review, ***Cell***)

# Model arch for unified multimodal pretraini

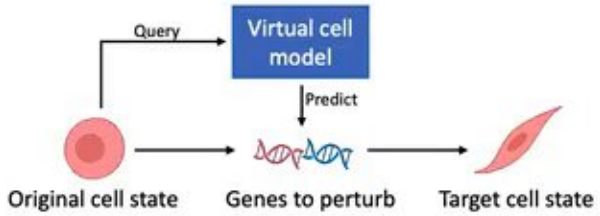



**Cui, H.**, Tejada-Lapuerta, A., Brbić, M., Saez-Rodriguez, J., Cristea, S., Goodarzi, H., Lotfollahi, M., Theis, F.J., and Wang, B. 2025. Towards Multimodal Foundation Models in Molecular Cell Biology. ***Nature***, 640(8059), pp.623-633.

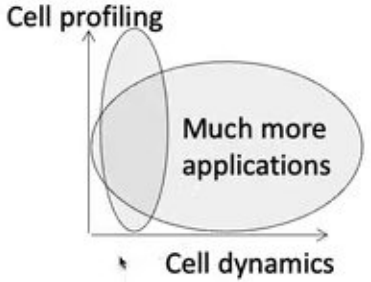
# Why perturbational prediction, or dynamics mo in general

MCBRLab BiliBili


- **Important new dimension**




Recall "one formula for many therapeutic applications"




Cell dynamics, either normal development or under external stimulus, opens a **new dimension of applications**

MCBRLab bilibili

# Join our mission to revolutionize Biomedicine

genbio.ai

Hugging Face

GitHub

@genbioai