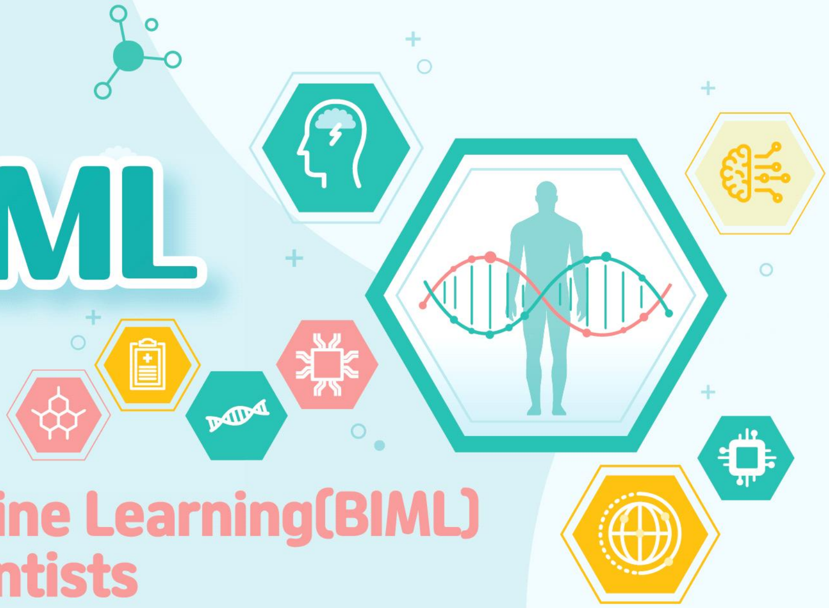
KSBi-BIML 2026



Bioinformatics & Machine Learning(BIML)
Workshop for Life Scientists

생명정보학 & 머신러닝 워크샵 (온라인)

Single Cell, Pre-training, and Foundation Model

이현주 _ GIST

KSBI
KOREAN SOCIETY FOR BIOINFORMATICS
한국생명정보학회

본 강의 자료는 한국생명정보학회가 주관하는 BIML 2026 워크샵을 목적으로 제작된 것으로 해당 목적 이외의 다른 용도로 사용할 수 없음을 분명하게 알립니다.

이를 다른 사람과 공유하거나 복제, 배포, 전송할 수 없으며 만약 이러한 사항을 위반할 경우 발생하는 **모든 법적 책임은 행위자 본인에게 있음**을 알립니다.

KSBi-BIML 2026

Bioinformatics & Machine Learning (BIML) Workshop for Life Scientists

한국생명정보학회가 주최하는 BIML-2026 동계 Bioinformatics &Machine Learning 교육 워크숍에 여러분을 초대합니다.

BIML 워크숍은 생명정보학 연구자들이 최신 AI바이오 분야의 인공지능 기반 분석 기술과 바이오 데이터 분석 기법을 이론과 실습을 통해 체계적으로 배울 수 있는 전문 교육 프로그램입니다. 2015년에 시작된 BIML 워크숍은 올해로 12년 차를 맞이하며, 국내 생명정보학 분야의 최초이자 최고 수준의 교육 프로그램으로 자리 잡았습니다. 이번 워크숍은 크게 인공지능바이오(AI바이오) 분야와 디지털바이오 분야, 두 분야로 구성됩니다.

AI바이오 분야에서는 생명정보 분석에 폭넓게 응용되고 있는 다양한 인공지능 기반 자료 모델링 기법을 다룰 예정입니다. 특히, 인공지능 심층학습을 활용한 단백질 구조 예측, 유전체 분석, 신약 개발에 대한 이론 및 실습 강의가 진행됩니다.

또한 디지털바이오 분야에서는 단일세포오믹스, 공간오믹스, 멀티오믹스, 메타오믹스에 대한 강의도 마련되어 있어, 연구자들의 분석 역량 강화에 실질적인 도움을 줄 것으로 기대됩니다.

또한 2024년부터 추가된 의료정보 자료 분석을 다루는 강의를 올해도 지속해서 운영하고자 합니다. 이는 최근 의료정보 자료 분석에 관한 연구 수요 증가를 반영한 것으로, 관련 연구를 수행하는 의과학자 및 의료정보 연구자들에게 유용한 지침을 제공할 것입니다.

또한, 올해도 생명정보학 기술의 다양화에 발맞춰 온라인 강좌를 대폭 확대했습니다. 올해는 무료 강좌 10개를 포함한 총 40개 이상의 강좌가 개설되며, 연구 주제에 맞는 강좌 추천과 강연료 할인 혜택도 제공합니다.

BIML-2026는 국내 주요 연구 중심 대학의 전임 교수 및 각 분야 최고 전문가들의 강의로 구성되어 있으며, 기초 이론부터 최신 연구 동향까지 아우르는 심도 있는 교육의 장이 될 것으로 확신합니다.

여러분의 많은 관심과 참여를 기대합니다!

2026년 2월

한국생명정보학회장 류 성 호

강의개요

Single Cell, Pre-training, and Foundation Model

Single-cell RNA sequencing(scRNA-seq)은 다양한 생물학적 환경에서 세포 이질성과 조직 기능을 규명하기에 유용하지만, scRNA-seq 데이터의 희소성(sparsity) 및 라벨을 가진 주석 데이터(annotation data)의 부족에 의해서, 전사체 분석의 정확도에 한계를 가진다. 이러한 문제를 해결하기 위해 단일세포를 위한 사전 학습 (pre-training) 방법론 및 파운데이션 모델 (foundation model)이 활발히 연구되고 있다. 사전 학습은 대규모 데이터를 활용해 자가 지도 학습(self-supervised) 방식으로 일반적인 표현(representation)을 학습함으로써 라벨이 없는 scRNA-seq 데이터에도 효과적으로 적용될 수 있다. 최근 발표된 단일 세포 기반 파운데이션 모델은 대형 scRNA-seq 데이터셋을 바탕으로 gene embedding과 cell embedding을 학습해 다양한 downstream task의 성능을 크게 향상시켰다. 이러한 모델들은 dropout으로 인해 결측된 발현값을 더 의미 있는 표현 공간에서 보정하고, 세포 타입 주석, 배치 보정, 희귀 세포 탐지 등에서 기존 방법보다 높은 정확도를 보여준다.

본 강의에서는 단일 세포 파운데이션 모델에서 사용되는 자가 지도 학습 방법론과 주요 모델 및 다양한 downstream task을 소개한다. 본 강의를 통해서 파운데이션 모델을 구축하기 위한 데이터셋과 학습 방법론들과 이를 생물학 지식으로 변환하는 연구들을 이해하는 것을 목표로 한다.

- Single-cell RNA sequencing 소개
- 사전 학습 방법론 및 파운데이션 모델 구축 방법론
- scRNA-seq 기반 파운데이션 모델 사례
- Downstream task 및 성능
- 질병 관련 연구에의 응용

* 강의 난이도: 중급

* 강의: 이현주 교수 (광주과학기술원 AI융합학과)

Curriculum Vitae

Speaker Name: Hyunju Lee, Ph.D.

▶ Personal Info

Name	Hyunju Lee
Title	Professor
Affiliation	Gwangju Institute of Science and Technology

▶ Contact Information

Address	123 Cheomdangwagi-ro, Buk-gu, Gwangju, 61005
Email	hyunjulee@gist.ac.kr

Research Interest

Bioinformatics, Machine learning, and Text Mining

Educational Experience

1997	B.S. in Computer Science, KAIST, South Korea
1999	M.A. in Computer Engineering, Seoul National University, South Korea
2006	Ph.D. in Computer Science, University of Southern California, USA

Professional Experience

2006-2007	Post-doc Researcher, Brigham and Women's Hospital and Harvard Medical School, USA
2007-	Professor, Dept of AI Convergence, Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology

Selected Publications (3 maximum)

1. Yeonghun Lee, Sung-Hye Park, Hyunju Lee. Prediction of the 3D cancer genome from whole-genome sequencing using InfoHiC. Molecular Systems Biology, 20(11):1156-1172, 2024
2. Sejin Park, Hyunju Lee. Robust self-supervised learning strategy to tackle the inherent sparsity in single-cell RNA-seq data. Briefings in Bioinformatics, 25(6): bbae586, 2024.
3. Yeonghun Lee and Hyunju Lee. Integrative reconstruction of cancer genome karyotypes using InfoGenomeR. Nature Communications, 12:2467, 2021.

KSBI | 한국생명정보학회

KSBi-BIML 2026

Single-Cell, Pre-training, and Foundation Models

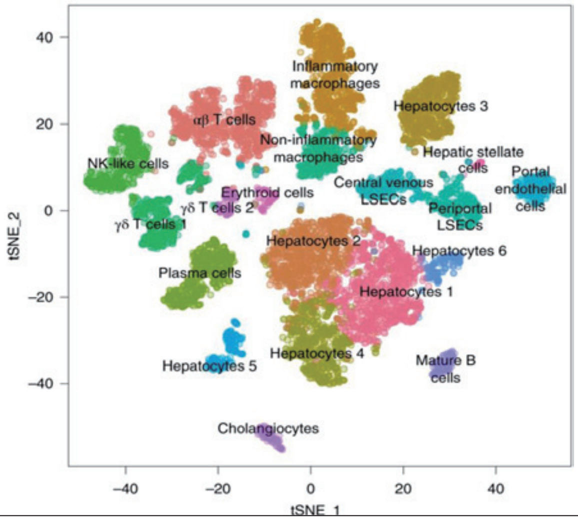
이현주 (광주과학기술원 AI 융합학과, AI 대학원)

Contents

- scRNA-seq basics
- Representative single cell foundation models
    - Pretraining foundation models (geneformer, scGPT, scFoundation)
    - Downstream tasks
    - Disease applications

Single-Cell RNA Sequencing (scRNA-seq)

- Key technology to dissect cellular heterogeneity and tissue function
- Single-cell resolution enables discovery of rare and transient cell states
- Sparsity and Dropout
  - Many zeros due to both technical limitations and true biological absence
  - Zero inflation: observed expression underestimates true expression
  - Impacts clustering, dimensionality reduction, and differential expression

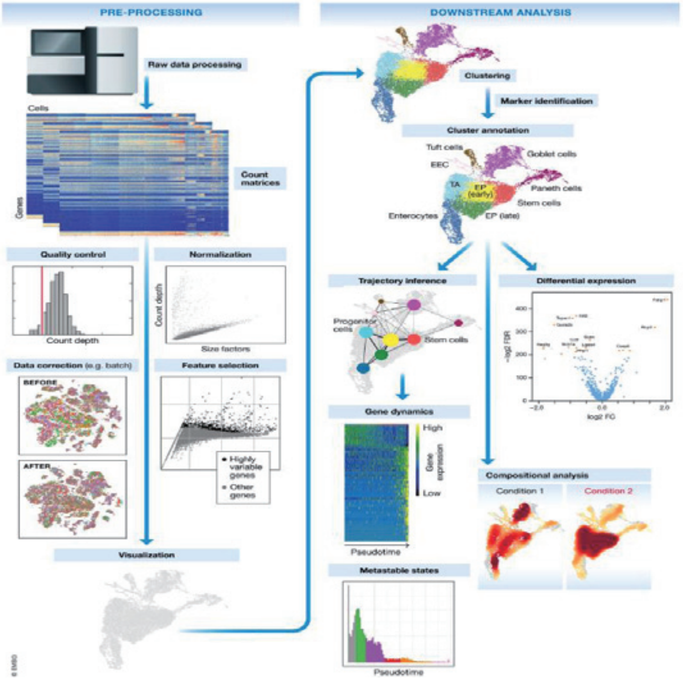


single-cell transcriptomic map of the human liver from data generated using scRNA-seq applied to five human liver samples (8,444 cells) reported in MacParland et al.

Each cluster corresponds to a biologically distinct cell type, such as hepatocytes, macrophages, endothelial cells, and immune cells

Standard scRNA-seq Analysis Pipeline

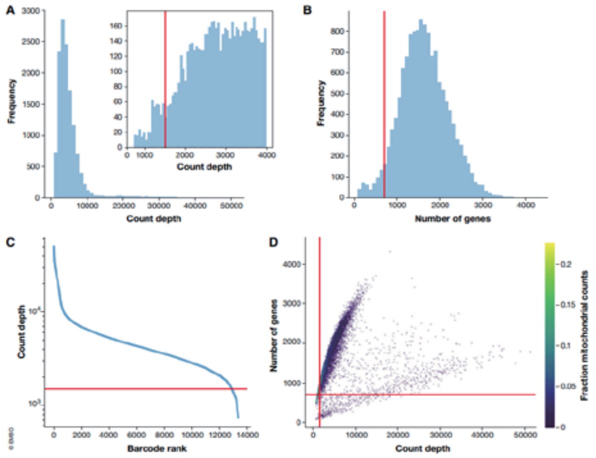
- Quality control (cell and gene filtering)
- Normalization and selection of highly variable genes
- Dimensionality reduction (PCA, UMAP, t-SNE)
- Batch correction with integration tools (e.g., Seurat, Harmony)



Luecken, M.D., Theis, F.J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol Syst Biol* **15**, MSB188746 (2019). https://doi.org/10.15252/msb.20188746

Standard scRNA-seq Analysis Pipeline
: Quality control (cell filtering)

- Cell QC based on three QC covariates
  - number of read counts per barcode (count depth)
  - number of genes per barcode
  - fraction of counts from mitochondrial genes per barcode
- Examined for outlier peaks that are filtered out by thresholding
  - correspond to dying cells, cells whose membranes are broken, or doublets
- Consider these covariates jointly instead of separately.

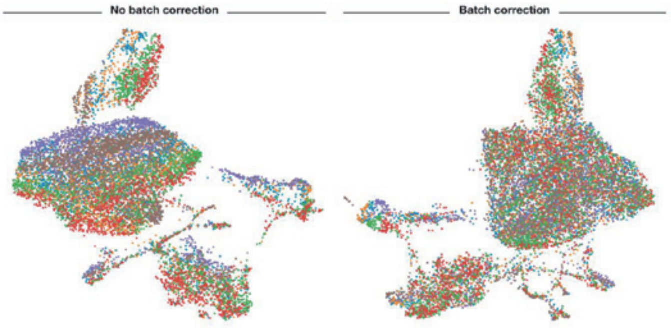


Plots of quality control metrics with filtering decisions for a mouse intestinal epithelium dataset from Haber *et al* (2017)

Luecken, M.D., Theis, F.J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol Syst Biol* **15**, MSB188746 (2019). https://doi.org/10.15252/msb.20188746

Standard scRNA-seq Analysis Pipeline
: Batch effects correction

- Batch effects arise when cells are processed in separate groups
  - Different chips, sequencing lanes, harvesting time points, etc.
- Batch correction: between samples/cells within the same experiment (bulk RNA-seq style)
  - A benchmark study shows **ComBat** performs well for low-to-medium complexity scRNA-seq datasets
  - recommend ComBat when cell type and state compositions between batches are consistent
  - ComBat model:
    - linear model with batch effects captured in both mean and variance



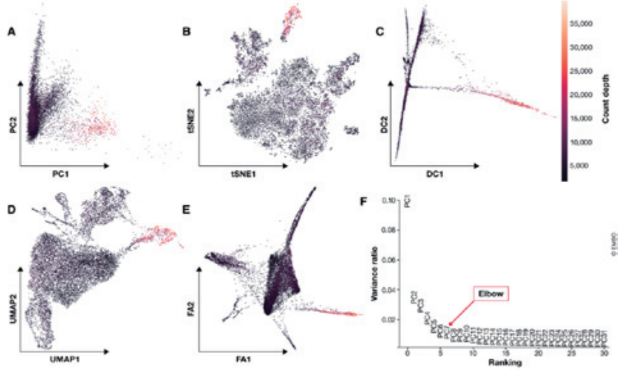
Cells are coloured by sample of origin. Separation of batches is clearly visible before batch correction and less visible afterwards. Batch correction was performed using ComBat on mouse intestinal epithelium data from Haber *et al* (2017).

Luecken, M.D., Theis, F.J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol Syst Biol* **15**, MSB188746 (2019). https://doi.org/10.15252/msb.20188746

Standard scRNA-seq Analysis Pipeline
- Feature selection, dimensionality reduction

- Feature Selection
  - Reduce dimensionality by selecting informative genes
  - Use Highly Variable Genes (HVGs) (typically 1,000–5,000)
  - Scanpy/Seurat HVG: bin by mean expression → select high variance-to-mean genes
  - Select HVGs after technical correction to avoid batch-driven HVGs
- Dimensionality Reduction
  - Embed cells into low-dimensional space capturing underlying structure
  - scRNA-seq is inherently low-dimensional (biological manifold)
  - Summarization (PCA): supports downstream analysis
  - Visualization (UMAP/t-SNE): 2D/3D plotting
  - 2D visualization should not be used as a dataset summary

Common visualization methods for scRNA-seq data

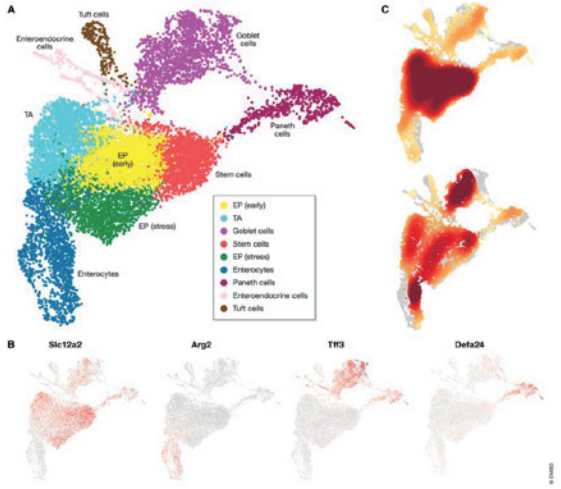


Luecken, M.D., Theis, F.J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol Syst Biol* **15**, MSB188746 (2019). https://doi.org/10.15252/msb.20188746

Standard scRNA-seq Analysis Pipeline
- Cluster analysis

Clustering of mouse intestinal epithelium dataset from Haber *et al* (2017)

- Clustering is the first key intermediate result: group cells by expression similarity
- Similarity is computed in reduced space (often PCA) using distance metrics
- Two approaches:
  - **Classical clustering** (e.g., k-means) on distance matrix
  - **Graph-based community detection** on KNN graph (default in Scanpy/Seurat)
- Leiden (or Louvain) modularity optimization is widely used, with **resolution** controlling cluster granularity
- Subclustering can reveal finer states but may introduce noise artifacts



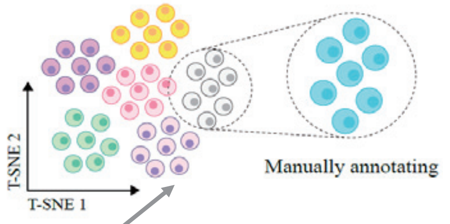
(A) Annotated cell-identity clusters found by Louvain clustering visualized in a UMAP representation.
(B) Cell-identity marker expression to identify stem cells (Slc12a2), enterocytes (Arg2), goblet cells (Tff3) and Paneth cells (Defa24).

Luecken, M.D., Theis, F.J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol Syst Biol* **15**, MSB188746 (2019). https://doi.org/10.15252/msb.20188746

scRNA-seq and Cell Type Annotation

- Manual cell type annotation requires expert knowledge and marker genes

- Normalizing the data
- Identification of highly variable features (feature selection)
- Perform linear dimensional reduction using PCA
- Determine the 'dimensionality' of the dataset
- Cluster the cells
- Finding differentially expressed features (cluster biomarkers)
- Assigning cell type identity to clusters



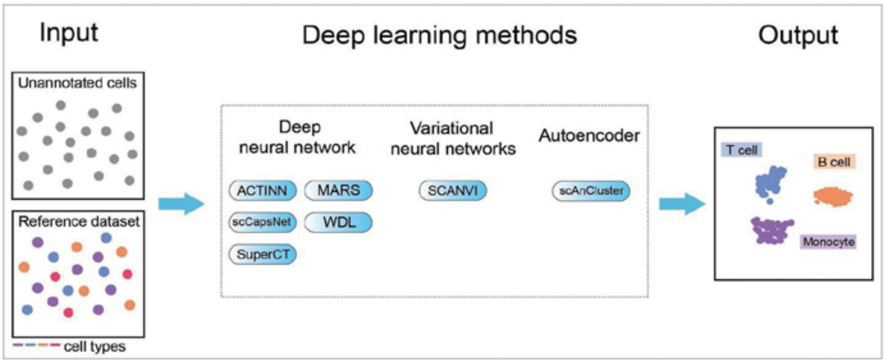
Marker genes database

Cluster ID	Markers	Cell Type
0	IL7R, CCR7	Naive CD4+ T
1	CD14, LYZ	CD14+ Mono
2	IL7R, S100A4	Memory CD4+
3	MS4A1	B
4	CD8A	CD8+ T
5	FCGR3A, MS4A7	FCGR3A+ Mono
6	GNLY, NKG7	NK
7	FCER1A, CST3	DC
8	PPBP	Platelet

Briefings in Bioinformatics, 2022, 23,1–13

Reference Data-Driven Learning-Based Methods

- Machine Learning Algorithms
  - Data-driven learning methods leverage machine learning algorithms to automate the identification of cell types in biological datasets.
  - These methods are particularly effective at managing the complexities and high dimensionality associated with single-cell datasets.



Input

Deep learning methods

Output

Unannotated cells

Reference dataset

cell types

Deep neural network

Variational neural networks

Autoencoder

ACTINN

MARS

scCapsNet

WDL

SuperCT

SCANVI

scAnCluster

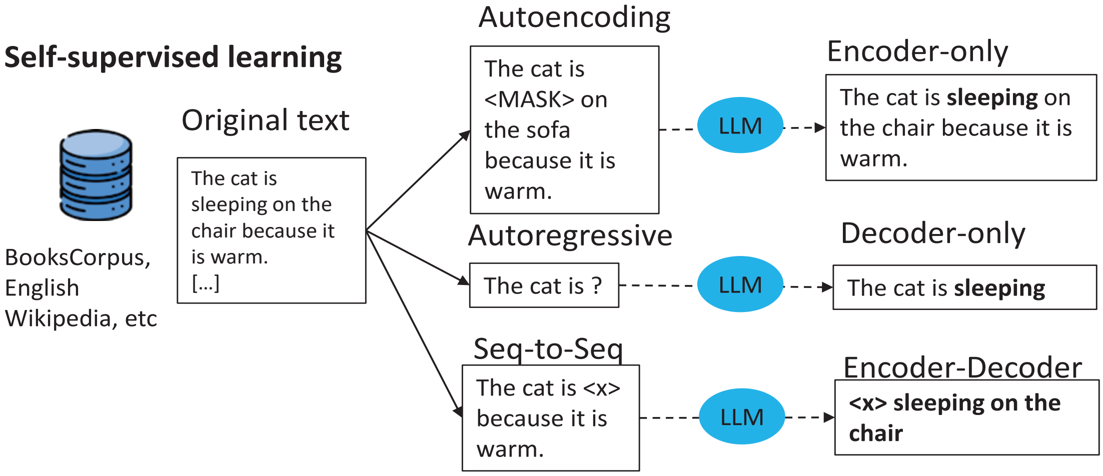
Briefings in Bioinformatics, 2022, 23,1–13

Motivation for Single-Cell Foundation Models

- Explosion of scRNA-seq datasets across tissues, species, and diseases
- Severe lack of reliable labels for cell types and states
- Sparsity and dropout limit classical transcriptome analysis
- Foundation models: "Train with large-scale unlabeled scRNA-seq, transfer to many downstream tasks"

Large-Scale Pre-training Methods

- **Advancements in AI Models**
  - Recent advancements in AI have enabled the creation of large-scale pre-trained language models.
  - These models utilize vast amounts of data, allowing them to generalize effectively across various downstream tasks for improved performance.



Self-supervised learning

BooksCorpus, English Wikipedia, etc

Original text

The cat is sleeping on the chair because it is warm. [...]

Autoencoding

The cat is <MASK> on the sofa because it is warm.

LLM

Encoder-only

The cat is **sleeping** on the chair because it is warm.

Autoregressive

The cat is ?

LLM

Decoder-only

The cat is **sleeping**

Seq-to-Seq

The cat is <x> because it is warm.

LLM

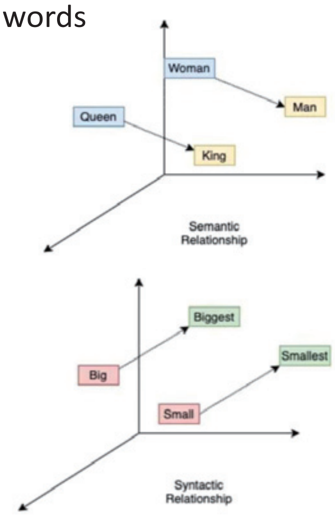
Encoder-Decoder

<x> sleeping on the chair

Large-Scale Pre-training Methods

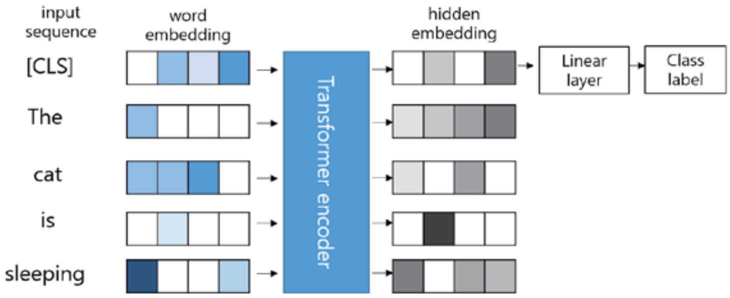
Word Embeddings

Captures both syntactic and semantic similarities between words



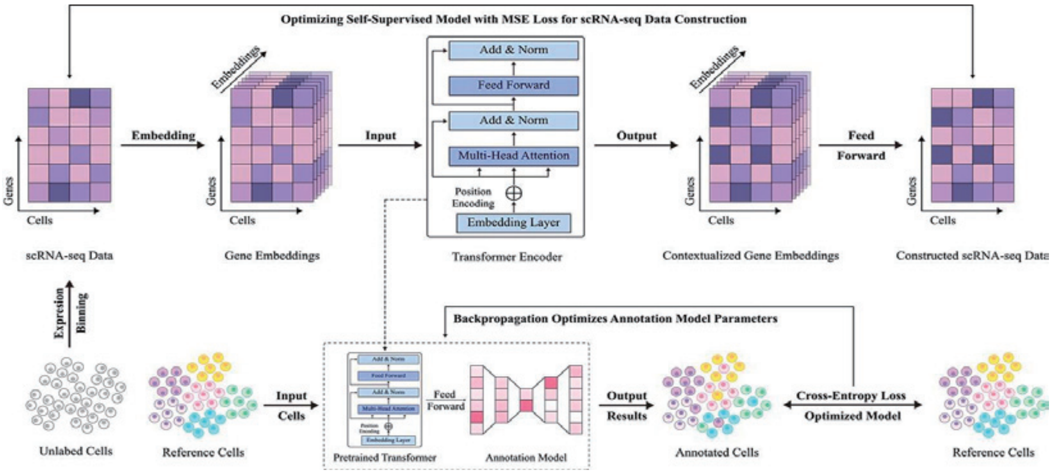
Language Transformer Model

Hidden embedding corresponding to the CLS token can be used to fine-tune a classification model.



Basic workflow of large-scale pretraining methods

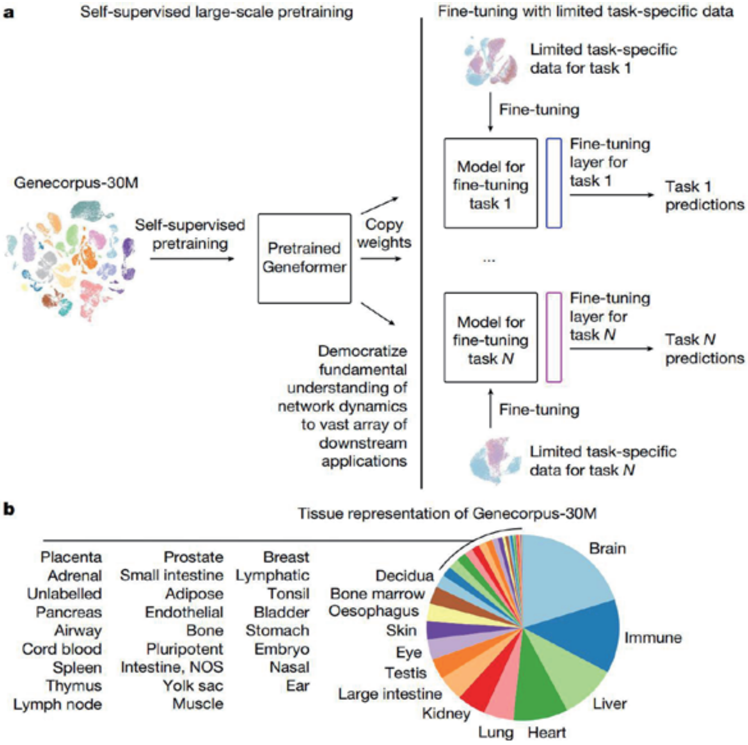
- Large-scale unlabeled scRNA-seq data are used as the initial feature source
- Gene embeddings with a transformer encoder framework reconstruct data via self-supervised learning
- A Transformer encoder is pretrained to learn deep cellular representations
- The pretrained model is fine-tuned for cell type annotation using supervised learning



Brief Bioinform, Volume 26, Issue 3, May 2025, bbaf207, https://doi.org/10.1093/bib/bbaf207

Geneformer

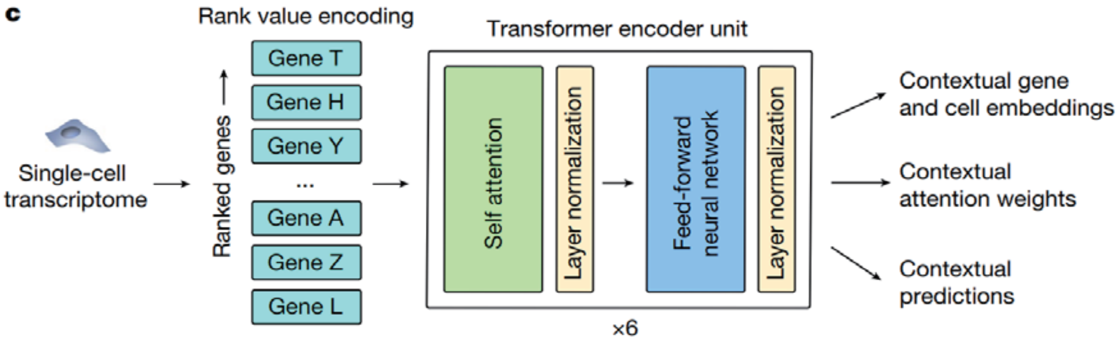
- Single-cell transcriptome-based foundation transformer
- Pre-trained on 30 millions of single-cell expression profiles
  - excluded cells with high mutational burdens (malignant cells and immortalized cell lines)
- Can be fine-tuned towards a vast array of downstream tasks with limited task-specific data


a
Self-supervised large-scale pretraining
Fine-tuning with limited task-specific data
Genecorpus-30M
Self-supervised pretraining
Pretrained Geneformer
Copy weights
Democratize fundamental understanding of network dynamics to vast array of downstream applications
Limited task-specific data for task 1
Fine-tuning
Model for fine-tuning task 1
Fine-tuning layer for task 1
Task 1 predictions
...
Model for fine-tuning task N
Fine-tuning layer for task N
Task N predictions
Fine-tuning
Limited task-specific data for task N
b
Tissue representation of Genecorpus-30M
Placenta
Adrenal
Unlabelled
Pancreas
Airway
Cord blood
Spleen
Thymus
Lymph node
Prostate
Small intestine
Adipose
Endothelial
Bone
Pluripotent
Intestine, NOS
Yolk sac
Muscle
Breast
Lymphatic
Tonsil
Bladder
Stomach
Embryo
Nasal
Ear
Decidua
Bone marrow
Oesophagus
Skin
Eye
Testis
Large intestine
Kidney
Lung
Heart
Liver
Immune
Brain

Theodoris, C.V., Xiao, L., Chopra, A. *et al.* Transfer learning enables predictions in network biology. *Nature* **618**, 616–624 (2023).

Geneformer Architecture and Pre-training

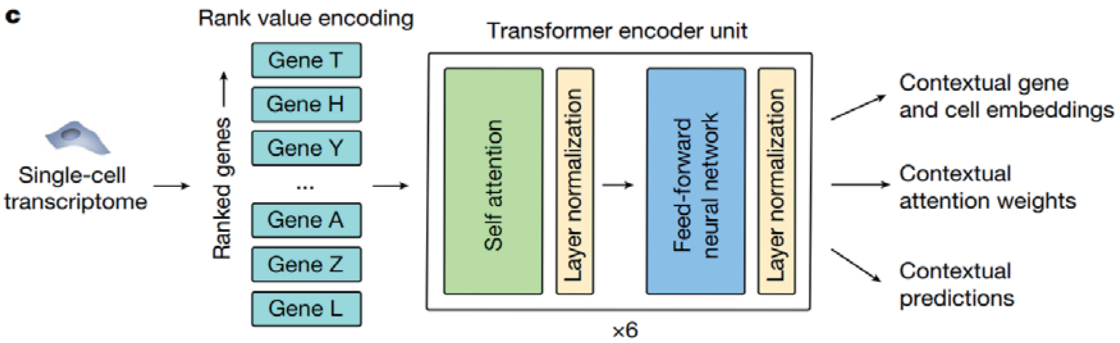
- Input: gene rankings by expression encoded as token sequences
- Rank value encoding of single-cell transcriptomes
  - Ranking genes by their expression within that cell normalized by non-zero median value of expression of each gene across all cells from Genecorpus-30M
  - Deprioritize ubiquitously highly expressed housekeeping genes
  - Highly rank transcription factors that may be expressed at low levels but have a high power to distinguish cell state



Theodoris, C.V., Xiao, L., Chopra, A. *et al.* Transfer learning enables predictions in network biology. *Nature* **618**, 616–624 (2023).

Geneformer Architecture and Pre-training

- BERT-style transformer encoder
  - Transformer encoder: 6 layers
  - Input size of 2,048 (fully represents 93% of rank value encodings in Geneformer-30M cells)
  - Embedding dimension: 256, attention heads: 4 per layer
  - Feed-forward size: 512
  - Full dense self-attention across the input size of 2,048
- Self-supervised masked gene prediction objective
  - Masking 15% of the genes within each transcriptome
  - Model was trained to predict which gene should be within each masked position in that specific cell state.



c

Rank value encoding

Transformer encoder unit

Single-cell transcriptome

Ranked genes

Gene T

Gene H

Gene Y

...

Gene A

Gene Z

Gene L

Self attention

Layer normalization

Feed-forward neural network

Layer normalization

×6

Contextual gene and cell embeddings

Contextual attention weights

Contextual predictions

Theodoris, C.V., Xiao, L., Chopra, A. *et al.* Transfer learning enables predictions in network biology. *Nature* **618**, 616–624 (2023).

Geneformer Gene and Cell Embeddings

- Contextual Geneformer gene embeddings
  - 256-dim vector that encodes the characteristics of the gene specific to the context of the cell.
  - Hidden state weights for each gene within the given single-cell evaluated by forward pass through the model.
  - Extracted from the second to last layer of the models
- Cell embeddings
  - Encode characteristics of the state of the given single cell
  - 256-dim vector: by averaging the embeddings of each gene detected in that cell
- Attention weights
  - Contextual Geneformer attention weights are extracted for each attention head within each self-attention layer for each gene within the given single-cell transcriptome evaluated by forward pass through the Geneformer model.

Theodoris, C.V., Xiao, L., Chopra, A. *et al.* Transfer learning enables predictions in network biology. *Nature* **618**, 616–624 (2023).

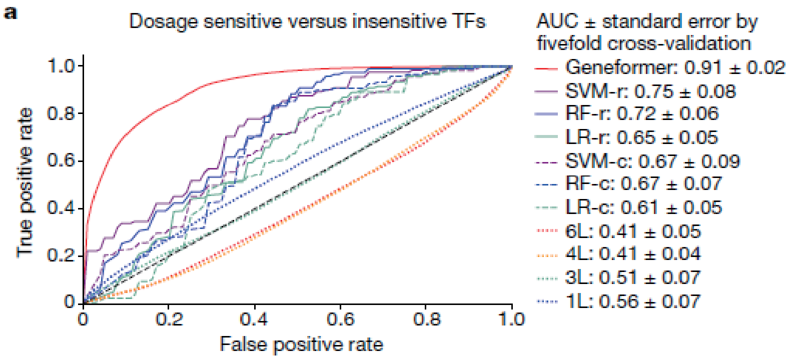
Geneformer Fine-tuning for Downstream Tasks

- Initialization
  - Use pretrained Geneformer weights
  - Add task-specific final transformer layer
- Tasks
  - Gene-level or cell-level classification
  - for example: dosage sensitivity, cell type annotation
- Single epoch fine-tuning (to avoid overfitting)
- Layer freezing strategy
  - Downstream tasks that are more relevant to pretraining task → freeze more layers
  - task that are more distance to pretraining→ fine-tune more layers
- Evaluation (gene-level tasks)
  - 5-fold cross-validation
  - Train 80% / Test 20% (gene labels)
  - Metrics: AUC ± SD, F1 score

Theodoris, C.V., Xiao, L., Chopra, A. *et al.* Transfer learning enables predictions in network biology. *Nature* **618**, 616–624 (2023).

Downstream Tasks: Dosage Sensitivity Prediction

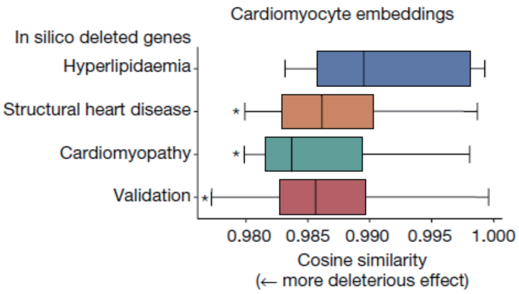
- Interpreting copy number variations (CNVs) requires identifying dosage-sensitive genes
- Traditional features (conservation, allele frequency) are cell-state invariant
- Geneformer captures contextual, transcriptional dynamics
- Fine-tuning
  - Trained on previously reported dosage-sensitive vs insensitive TFs
  - Used only 10,000 random single-cell transcriptomes
- Goal: classify dosage-sensitive vs dosage-insensitive genes
  - Geneformer significantly outperformed alternative methods

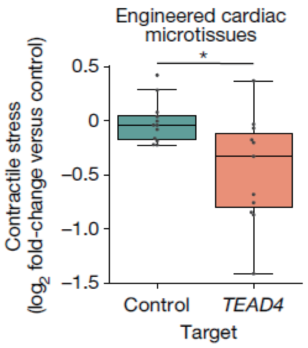


Theodoris, C.V., Xiao, L., Chopra, A. *et al.* Transfer learning enables predictions in network biology. *Nature* **618**, 616–624 (2023).

Downstream Tasks: In Silico Deletion and Experimental Validation

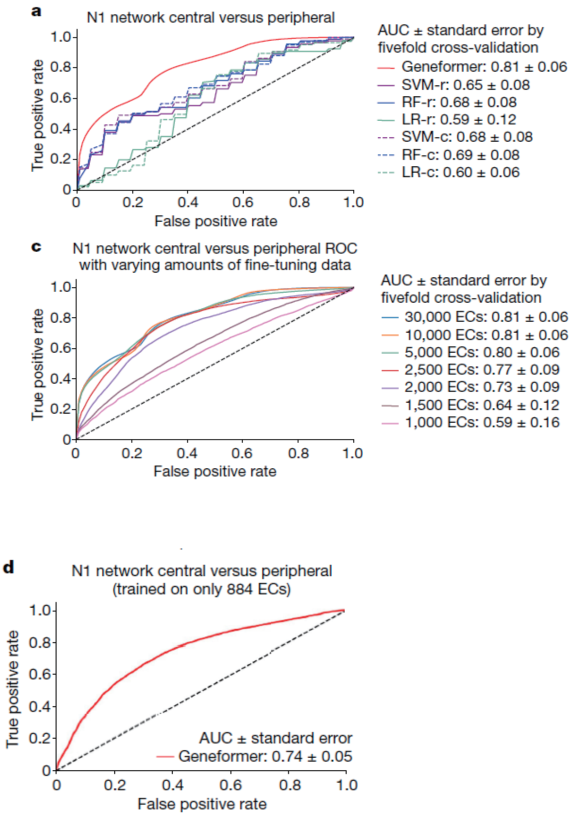
- In silico gene deletion
  - Simulated by removing a gene from rank-encoded transcripts
  - Measured impact on cell embeddings
- Performed in fetal cardiomyocyte (no fine-tuning)
  - Deletion of cardiomyopathy / structural heart disease genes
    → larger deleterious effects than control hyperlipidaemia genes
- Biological relevance
  - Top predicted genes enriched for cardiomyopathy phenotypes
  - Identified known regulators (e.g. FOXM1) and novel candidates (TEAD4)
- Experimental validation
  - CRISPR knockout of TEAD4 in iPSC-derived cardiac microtissues
  - Caused significant reduction in contractile stress





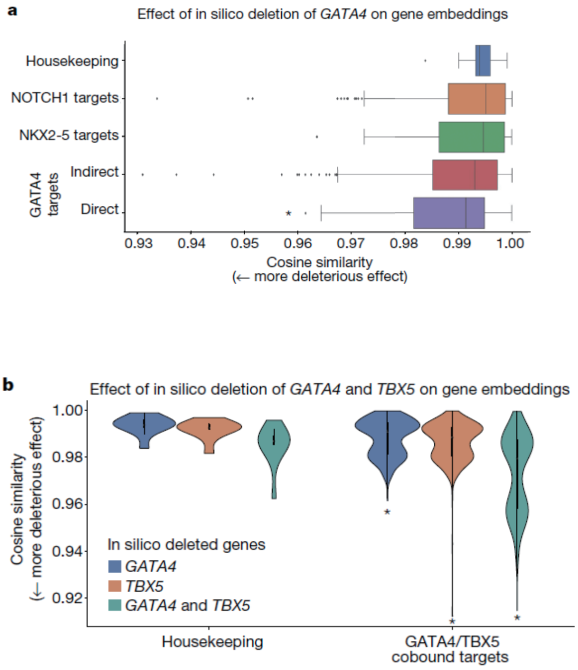
Downstream Tasks: Network dynamics predictions

- Identifying central vs peripheral genes in regulatory networks is critical for designing disease-modifying therapies
- NOTCH1 (N1) cardiac network
  - Previously mapped using large perturbation datasets
  - Central nodes were shown to have broad restorative effects
- Geneformer fine-tuning (no perturbation data)
  - Trained using ~30,000 normal endothelial cell (EC) single-cell transcriptomes
  - Successfully distinguished:
    Central vs peripheral genes
  - Comparable performance maintained with only 5,000 ECs
  - Using biologically more relevant data (884 ECs from healthy vs dilated aortas):
    - Geneformer outperformed other methods trained on ~30,000 ECs
  - Indicates that data relevance can outweigh data quantity



Downstream Tasks: In Silico Gene Network Analysis

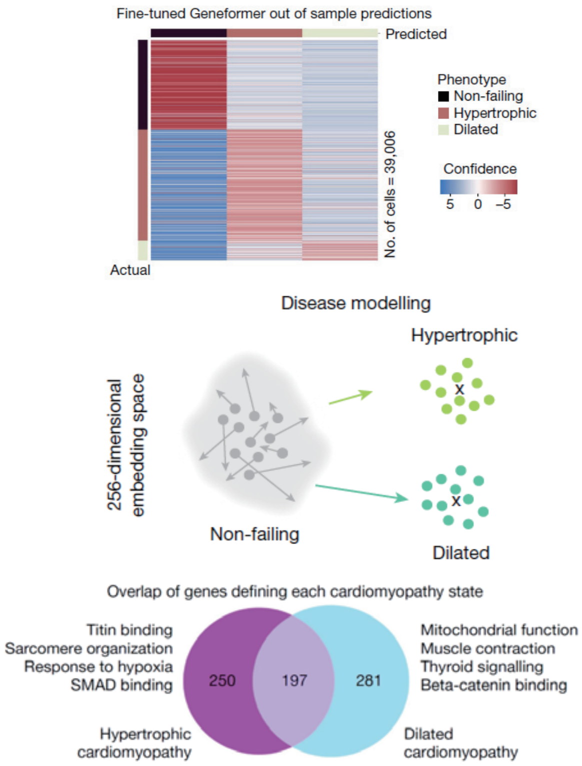
- Test whether pretrained Geneformer gene embeddings encode network connections between transcription factors (TFs) and their targets before fine-tuning
- No fine-tuning or perturbation data used
- In silico deletion
  - Computationally remove a TF from input
  - Measure impact on other genes via embedding changes
  - Larger embedding shift → stronger network dependency
- Single TF deletion (GATA4 or TBX5)
  - Deletion of GATA4 most strongly affected its direct targets (ChIP–seq validated)
  - Indirect targets showed weaker effects
  - Housekeeping genes were minimally affected
- Combined deletion of GATA4 + TBX5 caused:
  - Greater disruption of cobound target genes
  - Effects exceeding the sum of individual deletions
- Indicates Geneformer captures TF cooperativity (synergy)



Theodoris, C.V., Xiao, L., Chopra, A. *et al.* Transfer learning enables predictions in network biology. *Nature* **618**, 616–624 (2023).

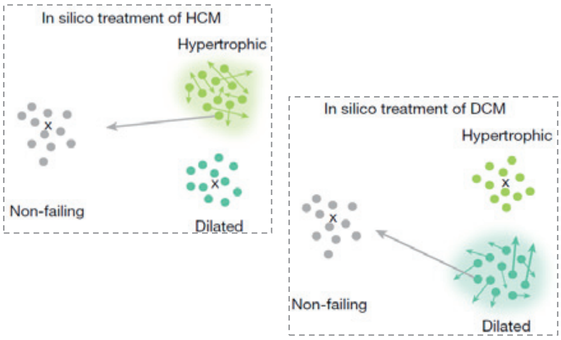
In silico treatment revealed candidate therapeutic targets

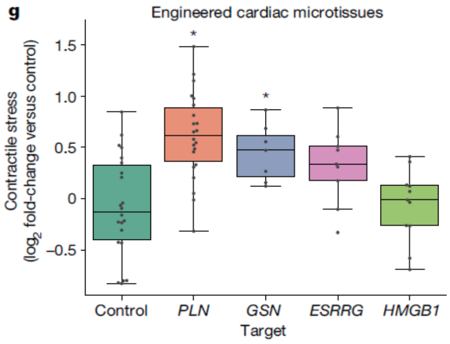
- Geneformer was fine-tuned to classify cardiomyocyte states:
  - Non-failing (n=9)
  - Hypertrophic cardiomyopathy (HCM) (n=11)
  - Dilated cardiomyopathy (DCM) (n=9)
  - Samples were randomly assigned as training or out-of-sample data **by patient** so that no single cells from the out-of-sample data were used for training.
  - Achieved ~90% out-of-sample accuracy using human heart single-cell data
- In silico perturbation to identify disease drivers
  - Gene deletion or activation was simulated in cardiomyocytes
- Measured whether cell embeddings shifted toward:
  - HCM state or DCM state
- Results:
  - 447 genes predicted to drive HCM when lost
    - Enriched for sarcomere organization and titin binding
  - 478 genes predicted to drive DCM when lost
    - Enriched for muscle contraction and mitochondrial function



In silico treatment revealed candidate therapeutic targets

- In silico treatment: reversing disease states
- Perturbations were simulated in HCM/DCM cardiomyocytes
  - Identify genes/pathways whose modulation shifts embeddings back toward non-failing state
- Experimental validation
  - Geneformer-predicted targets in DCM
    - GSN and PLN
  - Model system:
    - iPSC-derived cardiac microtissues with TTN truncation (TTN^{+}/−)
    - Known to cause reduced contractile stress (DCM model)
  - CRISPR knockout of GSN or PLN in TTN^{+}/− tissues:
    - Significantly improved contractile stress
    - Partial rescue toward non-failing phenotype





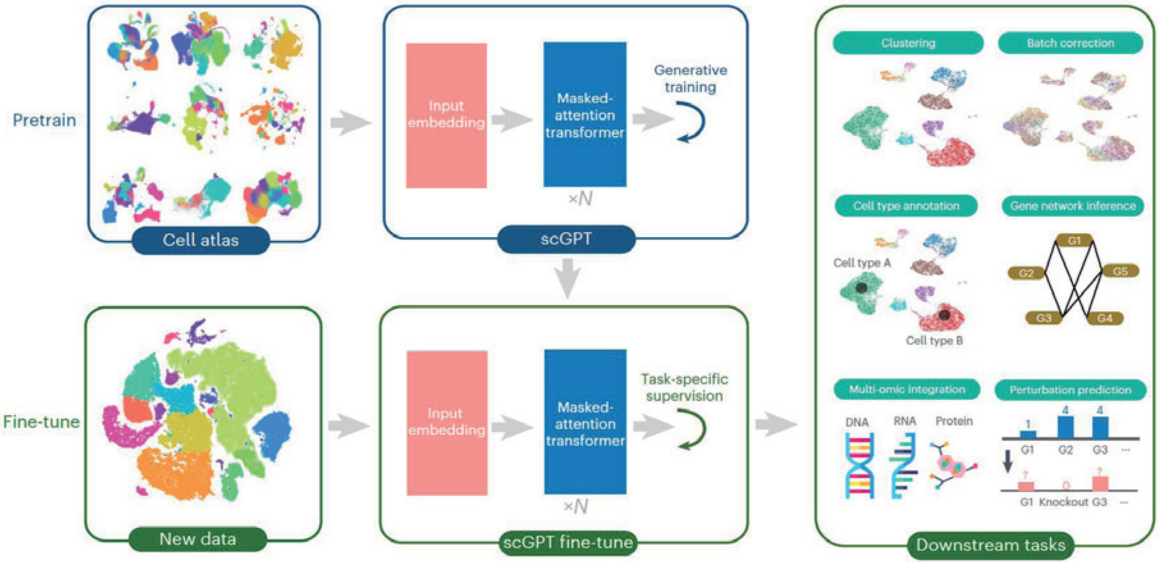
Theodoris, C.V., Xiao, L., Chopra, A. *et al.* Transfer learning enables predictions in network biology. *Nature* **618**, 616–624 (2023).

Geneformer repository

- https://huggingface.co/ctheodoris/Geneformer
- Geneformer-V1-10M: original model trained June 2021 on ~30M human single cell transcriptomes, 10M parameters, input size 2048, vocabulary ~25K protein-coding or non-coding RNA genes
- Geneformer-V2-104M and Geneformer-V2-316M: updated model trained Dec 2024 on ~104M human single cell transcriptomes, 104M or 316M parameters, input size 4096, vocabulary ~20K protein-coding genes

scGPT Introduction

- Single-cell foundation model by pretraining on over 33 million cells (CELLxGENE)
- Unified generative pretraining workflow specifically for non-sequential omics data
- Adapt the transformer architecture to simultaneously learn cell and gene representations

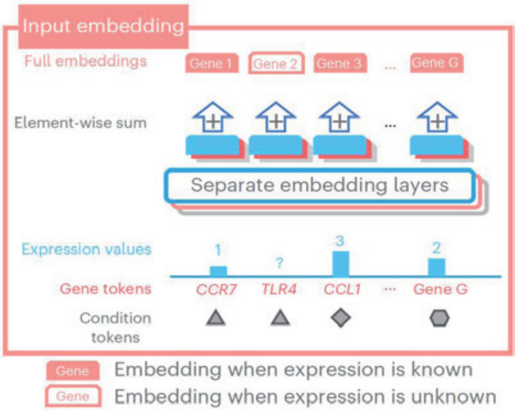


Cui, H., Wang, C., Maan, H. *et al.* scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat Methods* **21**, 1470–1480 (2024).

scGPT Input Embeddings and transformer

- scGPT models single-cell data using a transformer encoder
- Each cell is treated as a set of gene tokens
- Input embedding =

  gene identity (unique integer identifier) +
  binned expression value (cell-wise value binning)+
  condition embedding
- A special <cls> token appended to the beginning of the input tokens
  - cell embedding: final embedding at this position (aggregates gene information)



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

$$x_j^{(i)} = \begin{cases} k, \text{ if } X_{i,j} > 0 \text{ and } X_{i,j} \in [b_k, b_{k+1}]. \\ 0, \text{ if } X_{i,j} = 0. \end{cases}$$

$$\boldsymbol{t}_g^{(i)} = \left[\mathrm{id}(g_1^{(i)}), \mathrm{id}(g_2^{(i)}), \ldots, \mathrm{id}(g_M^{(i)})\right],$$

$$\boldsymbol{t}_c^{(i)} = \left[t_{c,1}^{(i)}, t_{c,2}^{(i)}, \ldots, t_{c,M}^{(i)}\right],$$

Cui, H., Wang, C., Maan, H. *et al.* scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat Methods* **21**, 1470–1480 (2024).

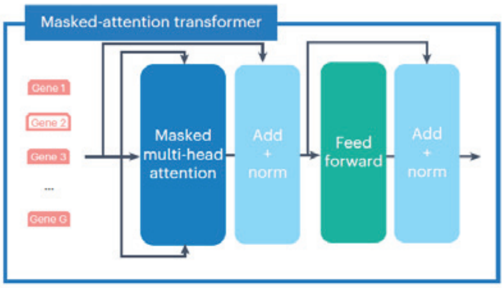
scGPT Pretraining

- Self-attention captures gene–gene interactions
- scGPT introduces a custom attention mask:
  - – Masked (unknown) genes attend only to known genes and themselves

$$Q = \boldsymbol{h}_l^{(i)} W_q,\ K = \boldsymbol{h}_l^{(i)} W_k,\ V = \boldsymbol{h}_l^{(i)} W_v,$$

$$\texttt{Attention}(Q,K,V) = \texttt{softmax}\left(\frac{QK^T}{\sqrt{d}} + \boldsymbol{A}_{\text{mask}}\right) V,$$

$$a_{i,j} = \begin{cases} 0, & \text{if } j \notin \text{unknown genes}, \\ 0, & \text{if } i = j \text{ and } j \in \text{unknown genes}, \\ -\inf, & \text{if } i \neq j \text{ and } j \in \text{unknown genes}. \end{cases}$$



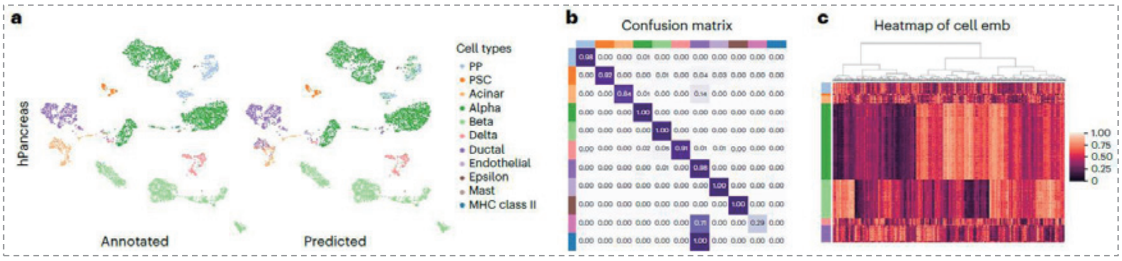
- Expression values of masked genes are predicted iteratively
  - – highest prediction confidence to be included as known genes
- Pretraining objective: mean squared error(MSE) loss on masked gene expression

$$\mathcal{L} = \frac{1}{|\mathcal{U}_{\text{unk}}|}\sum_{j\in\mathcal{U}_{\text{unk}}}\left(\text{MLP}\left(\boldsymbol{h}_n^{(i)}\right) - x_j^{(i)}\right)^2,$$

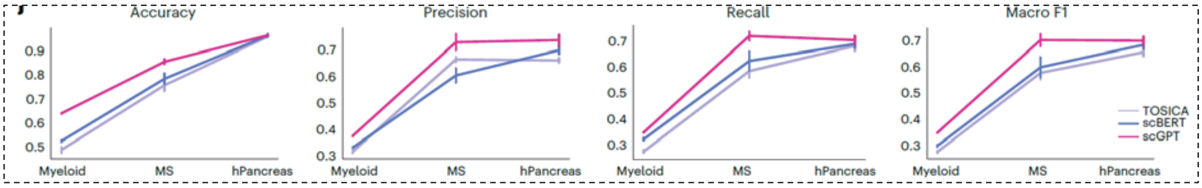
Cui, H., Wang, C., Maan, H. *et al.* scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat Methods* **21**, 1470–1480 (2024).

scGPT Cell type annotation

- Fine-tuning for cell type annotation
  - A neural network classifier is added on top of the scGPT transformer output cell embedding
  - **The whole model is fine-tuned using cross-entropy loss**
  - Training is performed on expert-annotated reference datasets
  - Cell types are predicted on held-out query datasets
- Performances
  - Human pancreas dataset: High precision (>0.8) for most cell types except for rare cell type

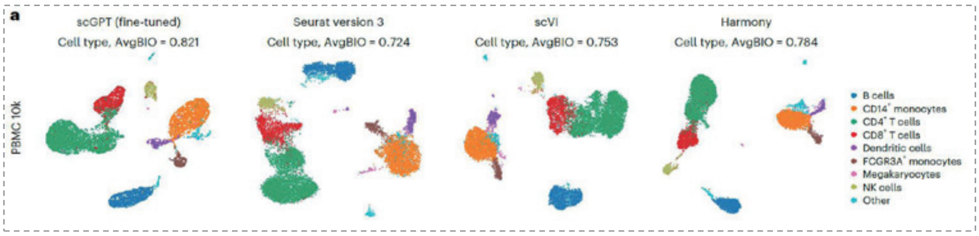


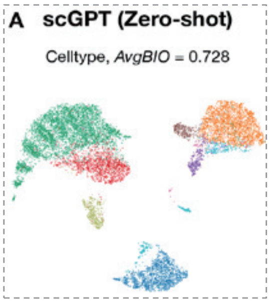
- Compared with **TOSICA** and **scBERT**
  - scGPT consistently outperformed both methods for hPancreas, myeloid, multiple sclerosis(MS)



Multi-batch scRNA-seq Integration with scGPT

- Integrate scRNA-seq datasets across batches while preserving biological structure and removing batch effects
- Self-supervised fine-tuning of scGPT using masked gene expression prediction (GEP), GEP for cell modeling (GEPC), domain adaptation via reverse back propagation (DAR), domain-specific normalization, etc
- Benchmark methods: scVI, Seurat, and Harmony
- PBMC 10k (2 batches): scGPT clearly separated all cell types with the highest biological conservation
- Performance Metric: Achieved an AvgBIO score of 0.821, outperforming other methods by 5–10%
- Generalization: Strong integration performance even without fine-tuning (zero-shot)
- scGPT effectively balances batch correction and biological signal preservation across diverse datasets





Cui, H., Wang, C., Maan, H. *et al.* scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat Methods* **21**, 1470–1480 (2024).

Single-cell Multi-modal Integration with scGPT

- Integrative Representation Learning for scMultiomic Data
  - scMultiomic datasets contain heterogeneous modalities across experimental batches
  - Joint integration of gene expression, chromatin accessibility (ATAC), and protein abundance
- Integration Settings
  - *Paired:* all cells share all modalities
  - *Mosaic:* batches share only partial modalities
- Tokenization: RNA genes, ATAC regions, and proteins are represented as distinct token types
- Embeddings: Pretrained RNA gene embeddings are reused; ATAC and protein embeddings are learned from scratch
- Modality awareness: Modality tokens indicate gene, region, or protein identity to guide masked prediction

$$\boldsymbol{t}_b^{(i)} = \left[t_{b,1}^{(i)}, t_{b,2}^{(i)}, \ldots, t_{b,M}^{(i)}\right] = \left[t_b^{(i)}, t_b^{(i)}, \ldots, t_b^{(i)}\right]$$

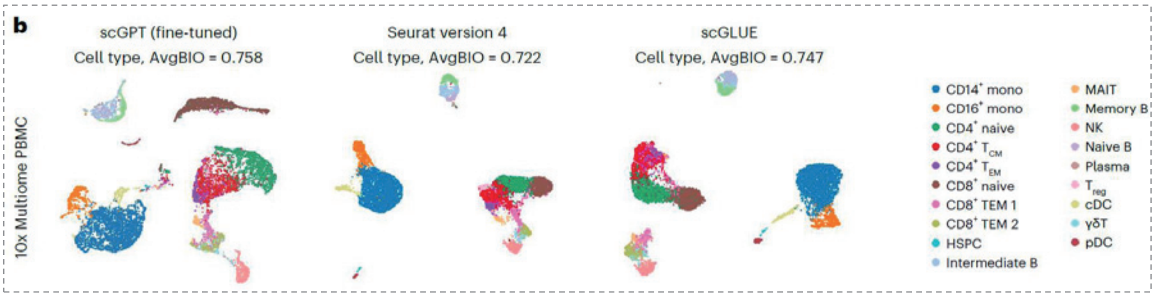
$$\boldsymbol{h}_n^{\prime(i)} = \mathrm{concat}\left(\boldsymbol{h}_n^{(i)}, \mathrm{emb_b}\left(\boldsymbol{t}_b^{(i)}\right) + \mathrm{emb_m}\left(\boldsymbol{t}_m^{(i)}\right)\right)$$

- Optimization: Model is fine-tuned using GEP and GEPC objectives. For multi-modal batch correction, domain adaptation via reverse back propagation (DAR) was applied.

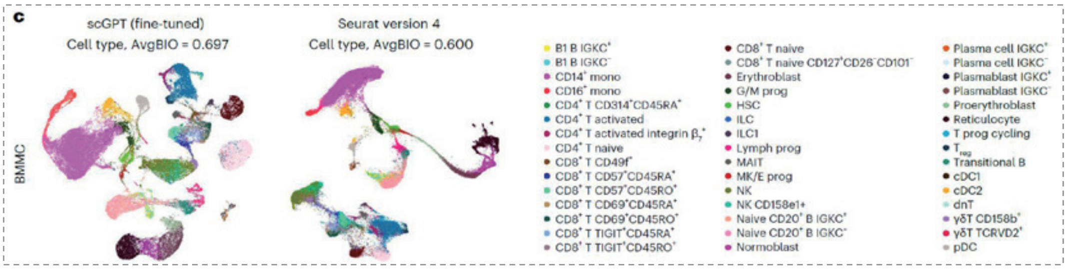
Cui, H., Wang, C., Maan, H. *et al.* scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat Methods* **21**, 1470–1480 (2024).

Single-cell Multi-modal Integration with scGPT

- Datasets: 10x Multiome PBMC (RNA + ATAC), BMMC (RNA + protein)
- Benchmark methods: scGLUE, Seurat v4
- 10x Multiome PBMC: scGPT uniquely resolved CD8$^+$ naïve T cells as a distinct cluster

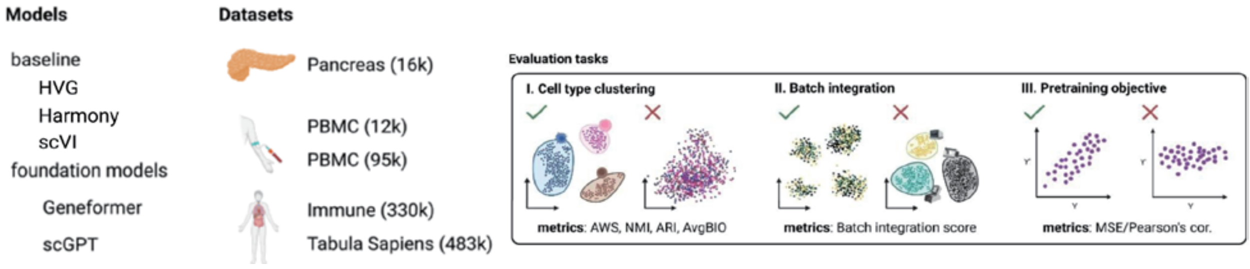


- BMMC (gene expression + protein abundance) : Improved biological conservation (≈9% higher AvgBIO) with clearer fine-grained immune 48 cell subtypes



Benchmarking of scFMs for zero-shot evaluation

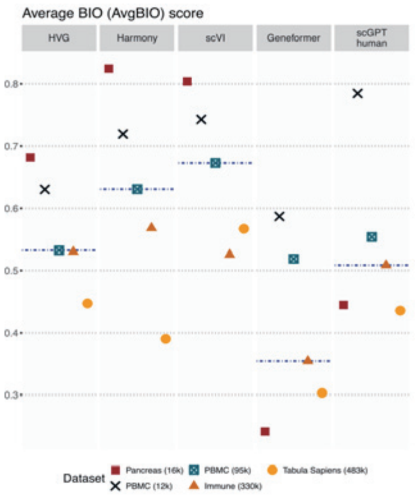
- Performance comparison of scGPT and Geneformer
  - Authors of two models claim
    - proposed models generate robust cell embeddings but also exhibit strong capabilities for generalizing to unseen datasets
  - Zero-shot evaluation by comparing with
    - highly variable genes (HVG)
    - established methods such as Harmony and scVI



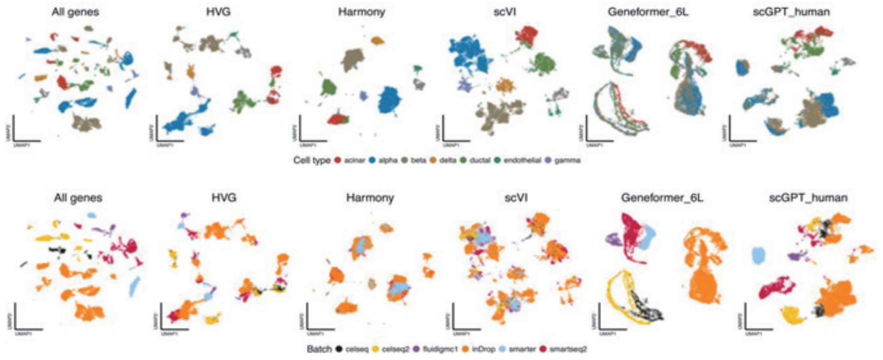
Kedzierska et al., Zero-shot evaluation reveals limitations of single-cell foundation models, Genome Biol, (2025) 26:101

Benchmarking of scFMs for zero-shot evaluation

- Performance comparison of scGPT and Geneformer: Cell type clustering
  - Zero-shot ability of cell embeddings to separate known cell types across multiple datasets



- AvgBIO score
  - scGPT performs better on PBMC (12k) compared to scVI, Harmony, and HVG.
  - However, across other datasets, scGPT is worse than scVI and Harmony.

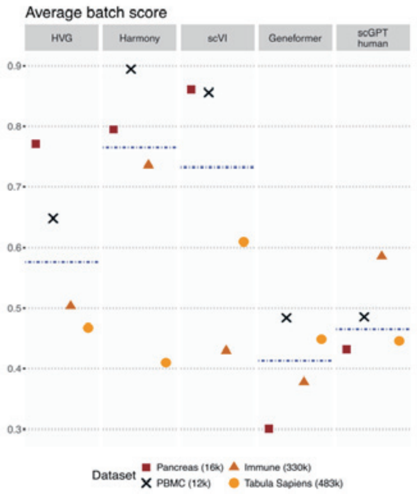


UMAP projections of the Pancreas (16k) using the cell embedding space

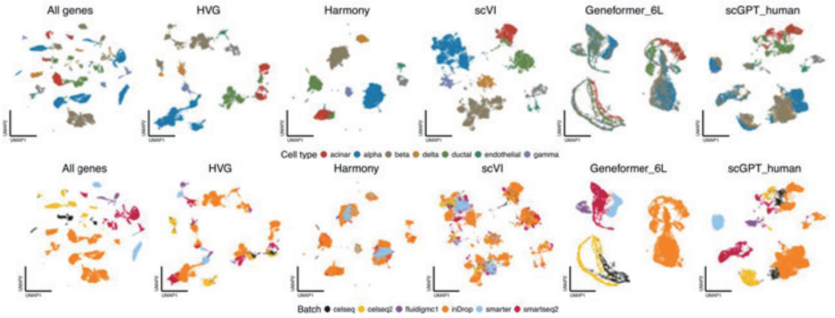
Kedzierska et al., Zero-shot evaluation reveals limitations of single-cell foundation models, Genome Biol, (2025) 26:101

Benchmarking of scFMs for zero-shot evaluation

- Performance comparison of scGPT and Geneformer: Batch integration
  - Eliminate batch effects from multiple data sources without removing meaningful biological differences



- Geneformer embeddings do not preserve cell-type structure; clustering is mainly driven by batch effects.
- scGPT embeddings show some cell-type separation, but the dominant structure is still batch-driven.
- In contrast, Harmony and scVI largely succeed in integrating the Pancreas dataset.

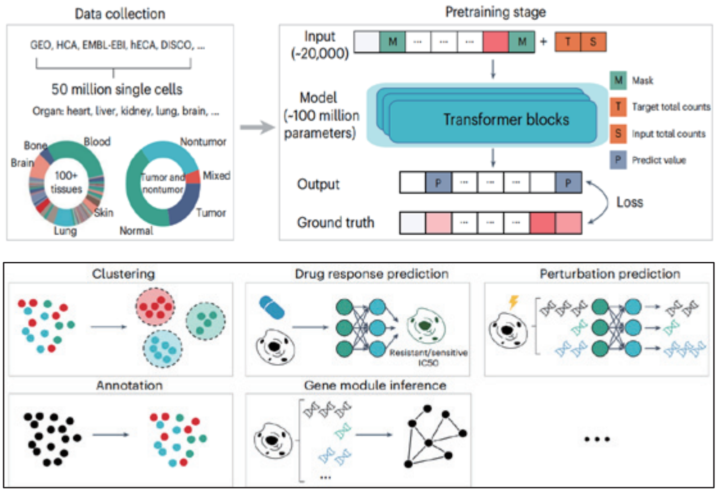


UMAP projections of the Pancreas (16k) using the cell embedding space

Kedzierska et al., Zero-shot evaluation reveals limitations of single-cell foundation models, Genome Biol, (2025) 26:101

scFoundation: Overview

- Large-scale foundation model on single-cell transcriptomics
- Pretrained on over 50 million human single-cell transcriptomic profiles
  - Downloaded from GEO, Single Cell Portal, HCA, and EMBL-DBI
- ~100 million parameters



Downstream tasks

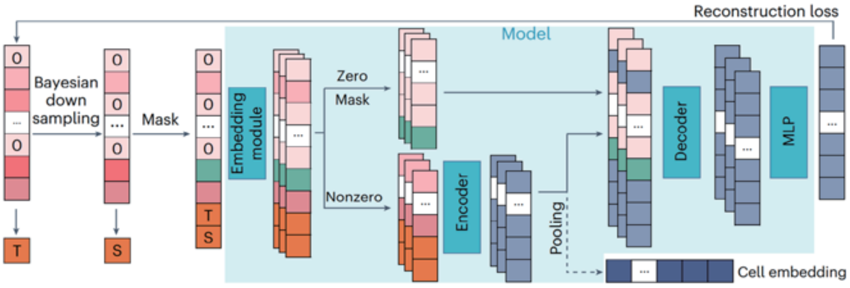
Hao, M., Gong, J., Zeng, X. *et al.* Large-scale foundation model on single-cell transcriptomics. *Nat Methods* **21**, 1481–1491 (2024).

scFoundation Pretraining

- Continuous raw gene expression vector serves as a training sample.
- Trained across diverse datasets and sequencing depths
  - Hierarchical Bayesian downsampling strategy generates the input sample.
  - First, Bernoulli distribution
  - Second, Binomial distribution

$$\boldsymbol{X}^{input} = \begin{cases} \boldsymbol{X} & if\ \gamma = 0 \\ [B(X_1, b), B(X_2, b), \ldots, B(X_N, b)] & if\ \gamma = 1 \end{cases} \qquad \gamma \sim Bernoulli(0.5) \ \ b \sim Beta(2,2)$$

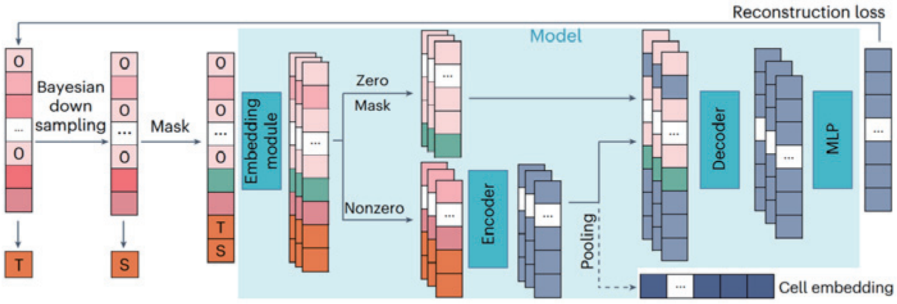
- Read depth-aware pre-training objective
  - Gene expression total counts (T and S) of the raw and input samples are computed.
  - Values in the input sample are randomly masked.



Hao, M., Gong, J., Zeng, X. *et al.* Large-scale foundation model on single-cell transcriptomics. *Nat Methods* **21**, 1481–1491 (2024).

scFoundation Pretraining

- Transformer-based encoder–decoder structure
  - Embedding module : value + gene identity
    - Converts each gene′s scalar expression into a *d*-dimensional embedding: E_i.
    - Zeros use a dedicated learnable embedding E^0; masked inputs use E^m.
    - Add gene identity via a learnable gene embedding T_i^G
    - E_i^{input} = E_i+ T_i^G; then stack across genes for the Transformer input.



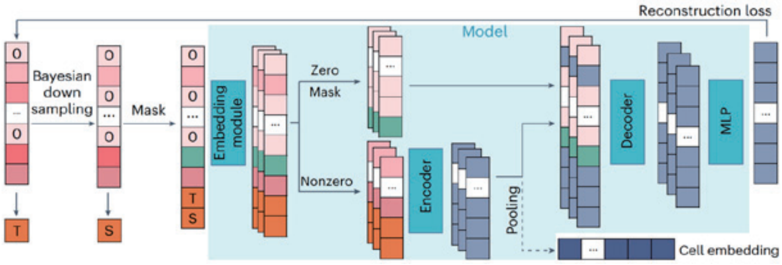
Hao, M., Gong, J., Zeng, X. *et al.* Large-scale foundation model on single-cell transcriptomics. *Nat Methods* **21**, 1481–1491 (2024).

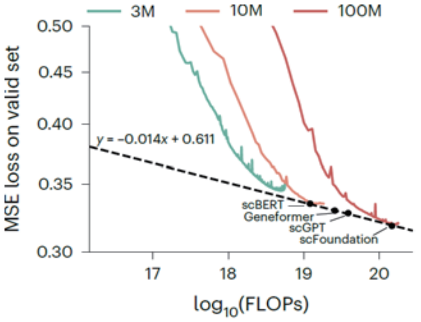
scFoundation Pretraining

- Transformer-based encoder–decoder structure
  - Embeddings of nonzero and nonmasked values (including T and S) are fed into the model encoder.
  - Output embeddings of the encoder are combined with mask and zero embeddings and fed into the decoder.
  - Encoder output can be pooled to generate a cell embedding for downstream usage.
  - Decoder output embeddings are projected to the gene expression value via a shared MLP layer.
  - The regression loss between the predicted and raw sample’s gene expression values is computed.

$$L = \frac{1}{|M|}\sum_{i=0}^{|M|}(X_i - P_i)^2$$

X_i : ground truth gene expression values
P_i : predicted gene expression values



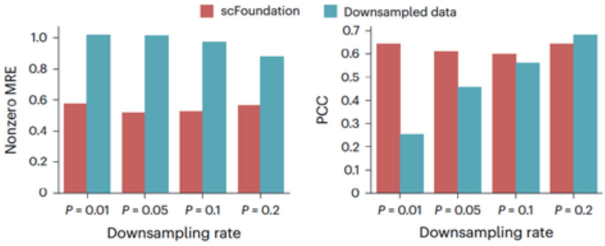


As the model parameters and the total number of floating-point operations (FLOPs) increased, the loss on the validation dataset exhibited a power-law decline.

Hao, M., Gong, J., Zeng, X. *et al.* Large-scale foundation model on single-cell transcriptomics. *Nat Methods* **21**, 1481–1491 (2024).

scFoundation Pretraining

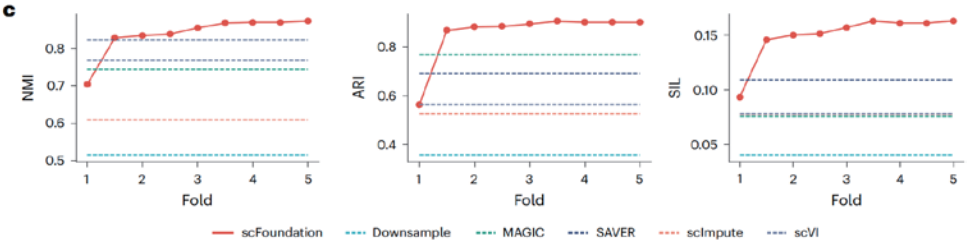
- Expression enhancement for low-depth datasets
  - Read-depth-aware (RDA) modeling: link the cells with different read depths
  - Enhance the read depth of the input cell by setting T (total) as a higher number than S (input)
  - S: downsampled the total counts to 1%, 5%, 10% and 20% of the original profiles
  - T: 1/p (p = sampling ratio)
  - MAE, MRE and PCC between predicted and actual nonzero gene expressions
  - Reduction of half the MAE and MRE from the downsampled data even when the downsampling rate was below 10%



Hao, M., Gong, J., Zeng, X. *et al.* Large-scale foundation model on single-cell transcriptomics. *Nat Methods* **21**, 1481–1491 (2024).

scFoundation : imputation

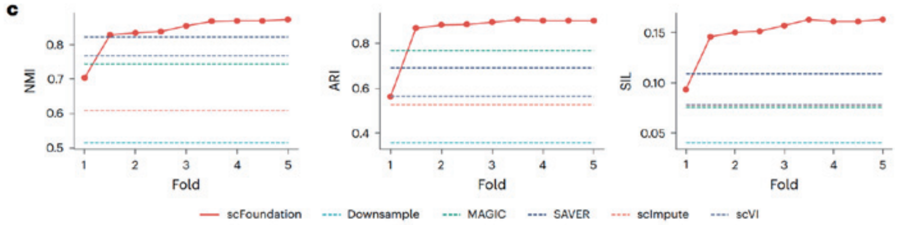
- Comparison with other imputation methods based on cell clustering metrics
  - Human pancreatic islet dataset
    - Manually generated downsampled gene expression profiles and their corresponding reference data (in the SAVER paper)
  - scFoundation
    - five sets of cell embeddings generated using non–fine-tuned pretrained encoder
    - varying the token parameter T relative to the downsampling factor S, with T/S folds ranging from 1 to 5.
  - Other methods
    - The downsampled data were used to train the method.
    - scVI produced imputed cell embeddings.
    - MAGIC, SAVER, and scImpute generated imputed gene expression matrices
  - Then, obtain clusters using scanpy.tl.leiden



Hao, M., Gong, J., Zeng, X. *et al.* Large-scale foundation model on single-cell transcriptomics. *Nat Methods* **21**, 1481–1491 (2024).

scFoundation : imputation

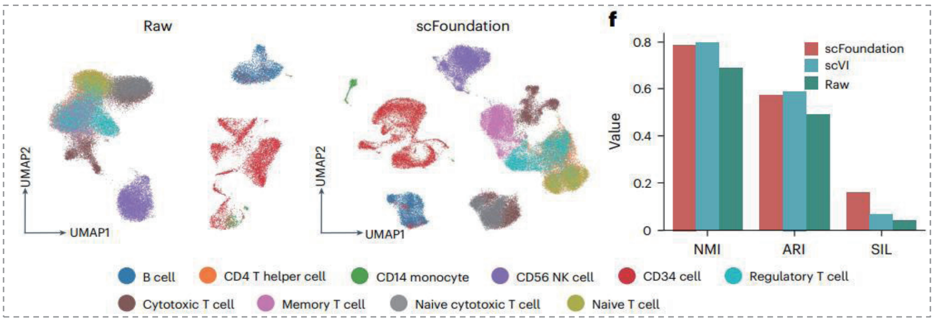
- Comparison with other imputation methods based on cell clustering metrics
  - The ground truth cluster labels were obtained from the reference data
  - Evaluation metrics
    - ARI and NMI (scikit-learn package): evaluate the degree of consistency between the clustering results and the actual cell type labels.
    - SIL: measures the aggregation degree of true cell type labels on the cell neighborhood maps
  - When T = S (fold = 1)
    - scFoundation outperformed both the baseline and scImpute across all metrics.
    - its performance was inferior to smaller models such as SAVER. This observation is consistent with previous reports showing that large models may not be advantageous when read depth is not increased.
  - As the T/S fold increased,
    - scFoundation's performance improved rapidly, surpassing all other methods.
    - Performance reached a plateau at higher folds, indicating that the learned cell embeddings became insensitive to T values beyond approximately 3.5×S.



Hao, M., Gong, J., Zeng, X. *et al.* Large-scale foundation model on single-cell transcriptomics. *Nat Methods* **21**, 1481–1491 (2024).

scFoundation : Clustering

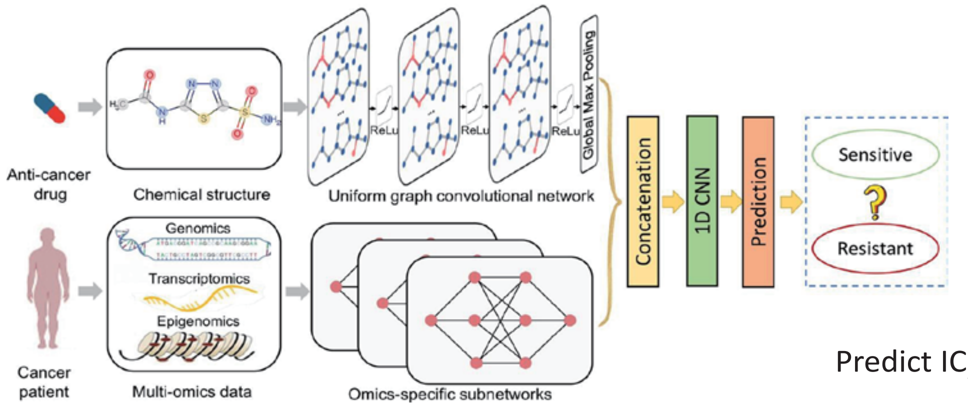
- Clustering in non-fine-tuning mode
  - Zheng68K dataset
    - comprising about 60,000 human peripheral blood mononuclear cells
    - early 10x Chromium platform.
    - Each cell: ~ 500 expressed genes, < 2,000 total reads, making cell type distinction challenging
  - scFoundation
    - without fine-tuning to enhance cell embeddings by setting the T value as 10,000.
    - scFoundation effectively separated memory T cells from other T cells and distinguished CD14 monocytes and CD34 cells better
  - scFoundation and scVI
    - Both methods outperformed the raw data in clustering.
    - While their NMI and ARI metrics were similar, scFoundation had a higher SIL score, showing its generalization ability in non-fine-tuning mode



Hao, M., Gong, J., Zeng, X. *et al.* Large-scale foundation model on single-cell transcriptomics. *Nat Methods* **21**, 1481–1491 (2024).

scFoundation: Drug response prediction

- Overview of drug response prediction task
  - DeepCDR: multi-omics (genomics, transcriptomics, epigenomics) and drug graph as input, uniform graph convolutional network (UGCN)

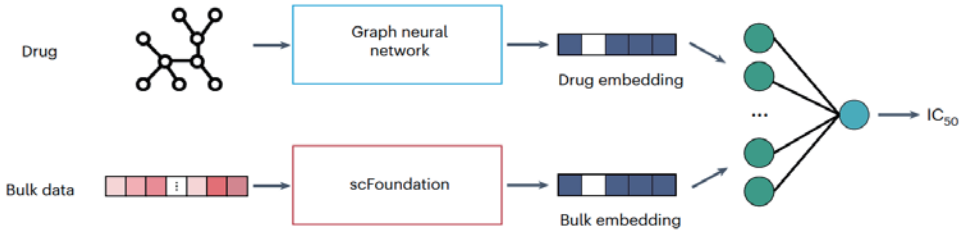


Predict IC_{50}

Qiao Liu, Zhiqiang Hu, Rui Jiang, and Mu Zhou. Deepcdr: a hybrid graph convolutional network for predicting cancer drug response. *Bioinformatics*, 36(Supplement 2):i911– i918, 2020.

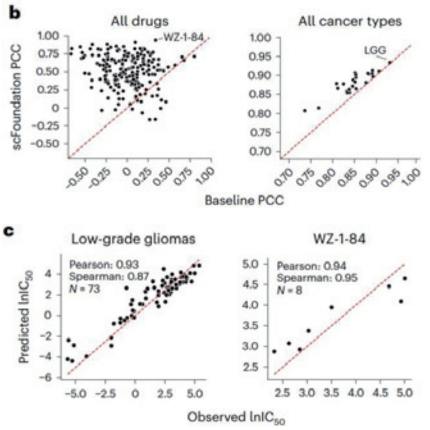
scFoundation: Drug response prediction

- Used the cell line and drug-paired data preprocessed by DeepCDR
  - Cancer Cell Line Encyclopedia (CCLE) and Genomics of Drug Sensitivity in Cancer (GDSC) datasets
  - Cell line data: 697 gene expression profiles
  - Drugs: represented as graphs with consistent feature matrices and adjacent matrix sizes.
  - 223 drugs and 561 cell lines data from 31 cancer types
  - Experimentally measured IC_{50} values :
    - randomly split 5% of data as the test set, resulting in 89,585 and 4,729 cell line-drug samples for training and testing, respectively.
- scFoundation-based drug response prediction model
  - Test whether single-cell–trained embeddings can transfer to bulk-level prediction tasks.
  - For each cell line, S and T equal to the sum of all gene expression values.
  - Fed the nonzero gene expression values and two indicators into the model encoder and got the context embedding for each gene. The bulk-level cell-line embedding was obtained by the max-pooling operation for each embedding dimension across all genes.
  - Replaced the gene expression with the cell-line embedding and trained the DeepCDR with the same setting.

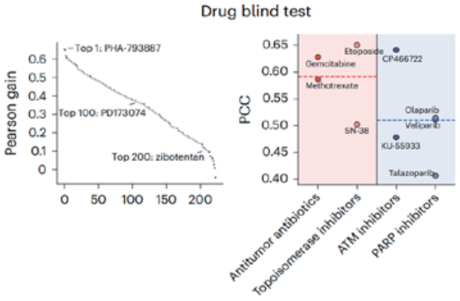


scFoundation: Drug response prediction

- Comparison with baseline DeepCDR model
  - DeepCDR used only gene expression data
  - **Most drugs** achieved higher PCC when using **scFoundation embeddings**
  - **All cancer types** benefited from scFoundation-based representations
- Best prediction case of drug and cancer types.
  - Regardless of high or low IC50, this model could predict accurate values and achieved a PCC above 0.93.

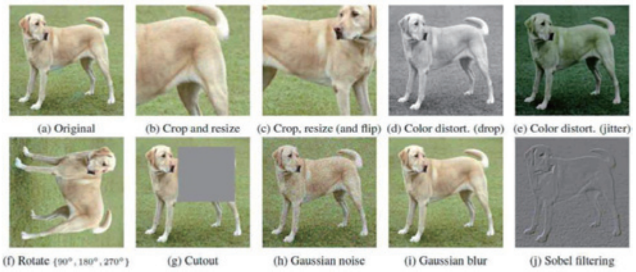


- Drug-blind test
  - scFoundation-based models consistently outperformed the original model.
  - Top 1 PCC-gaining drug PHA793887, a potent ATP-competitive CDK inhibitor
    - PCC improved from 0.07 to 0.73.
  - 200th-ranked drug zobotentan used for blocking endothelin A receptor activity
    - PCC improved from 0.49 to 0.64.
  - Drugs were grouped by therapeutic mechanism:
    - **Chemotherapy** (e.g., antitumor antibiotics, topoisomerase inhibitors)
    - **Targeted therapy** (e.g., ATM and PARP inhibitors)
    - Chemotherapy drugs showed **higher PCC values**
    - Targeted therapies performed worse when using gene expression alone



Cell Augmentation Techniques

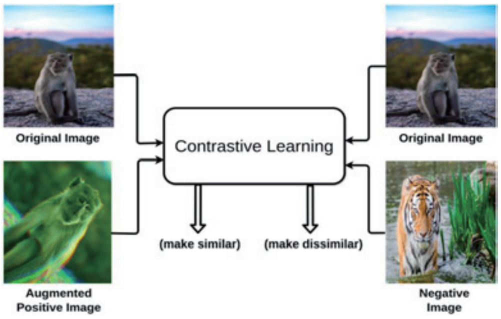
- Contrastive learning
  - Self-supervised learning
  - Generate augmented data
  - A discriminative approach that aims at grouping similar samples closer and diverse samples far from each other
- Contrastive learning for single cells
  - Construct different views of the same cell via augmentations
  - Contrastive loss: pull same-cell views together, push different cells apart



(a) Original (b) Crop and resize (c) Crop, resize (and flip) (d) Color distort. (drop) (e) Color distort. (jitter)

(f) Rotate {90°, 180°, 270°} (g) Cutout (h) Gaussian noise (i) Gaussian blur (j) Sobel filtering

A Simple Framework for Contrastive Learning of Visual Representations Chen, Ting and Kornblith, Simon and Norouzi, Mohammad and Hinton, Geoffrey (2020) Proceedings of the 37th International Conference on Machine Learning 119 1597--1607



Technologies **2021**, *9*(1), 2

ScRobust: Contrastive Learning for Cell Embedding

- **Cell augmentation**

– Makes various non-zero gene sets by random selection

– Adds the special token ‘CLS’ for cell embedding (summary vector)

- **Gene Embeddings**

Trainable unit embedding vectors (dimension *d* = 512)

- **Contrastive learning**

– scRobust learns a wide variety of local cell embeddings

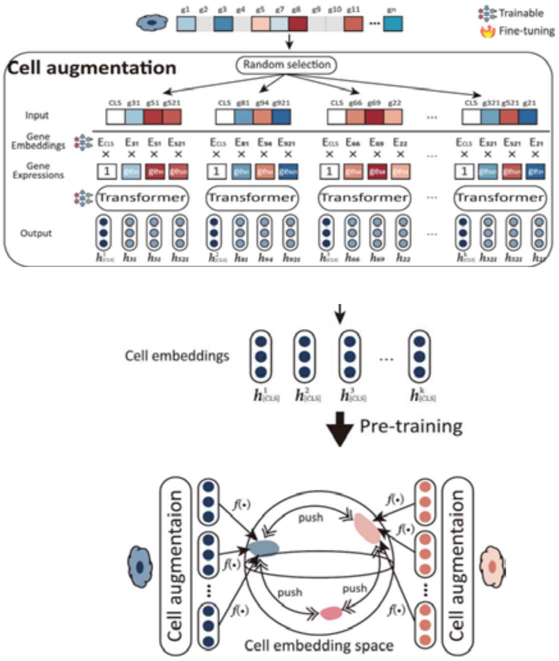
– Vector pair originating from the same cell is considered the “positive pair”

– Contrastive loss function

$$z_{i_1} = f(h^{i_1}_{[\mathrm{CLS}]})$$

$$L_{cl_{i_1,i_2}} = -\log \frac{\exp(\mathrm{sim}(z_{i_1}, z_{i_2})/\tau)}{\sum_{k=1}^{2N} \mathbf{1}_{[k \neq i_2]} \exp(\mathrm{sim}(z_{i_1}, z_k)/\tau)}$$

f : a fully connected network projects the cell embedding vector $h_{[\mathrm{CLS}]}$ into the cell embedding space.

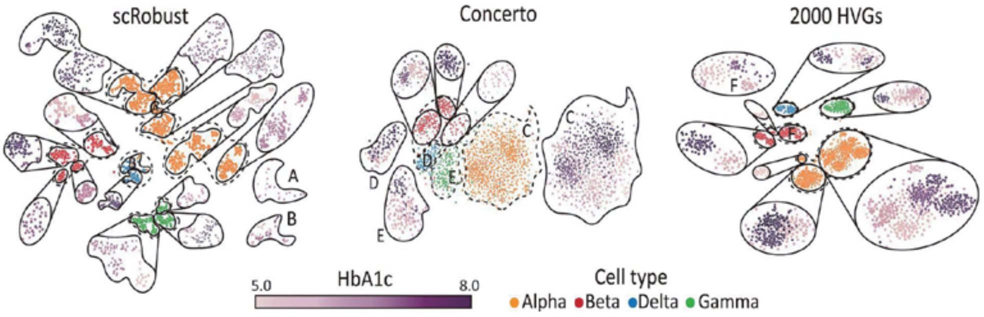


Sejin Park, Hyunju Lee, Robust self-supervised learning strategy to tackle the inherent sparsity in single-cell RNA-seq data, *Briefings in Bioinformatics*, 25 (6): bbae586, 2024

ScRobust: Impact of the pre-training tasks

- Dataset : Segerstolpe human islet cells
  - 2,089 single cells
  - Included cell types and HbA1c values (blood glucose level)

- scRobust clearly distinguished α, β, δ, γ cell types
- Better performance than Concerto and highly variable genes (HVGs)
- scRobust reflects both cell type and HbA1c
- Outperformed other methods in maintaining clarity



Sejin Park, Hyunju Lee, Robust self-supervised learning strategy to tackle the inherent sparsity in single-cell RNA-seq data, *Briefings in Bioinformatics*, 25 (6): bbae586, 2024

ScRobust: Highly unique genes for downstream task

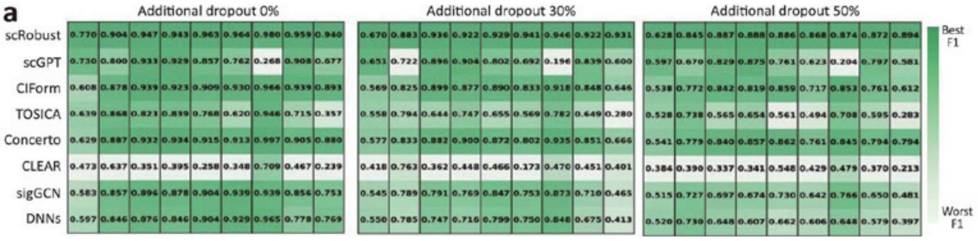
- Conventional approaches
  - Selects common input genes for all samples
  - *Highly variable genes* such as 2,000 genes are selected as input for all samples.
- ScRobust
  - Selects an individual gene set for each sample.
  - *Highly unique genes* are selected for each sample in a downstream task.

  - Pre-trains a model based on the contrastive learning, making it learn all genes**.**
  - In contrastive learning, it uses all genes through high-quality data augmentation.

Sejin Park, Hyunju Lee, Robust self-supervised learning strategy to tackle the inherent sparsity in single-cell RNA-seq data, *Briefings in Bioinformatics*, 25 (6): bbae586, 2024

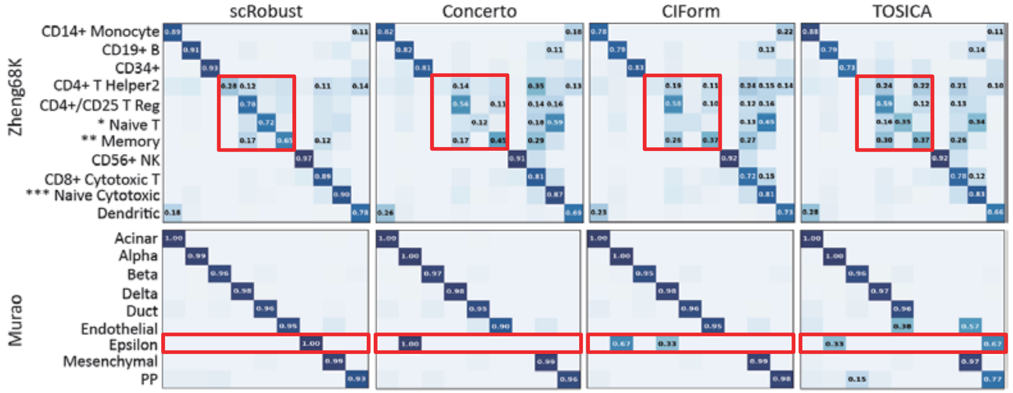
ScRobust performance for cell type annotation

- scRobust consistently achieves higher accuracy on nine benchmark datasets compared to previous methods

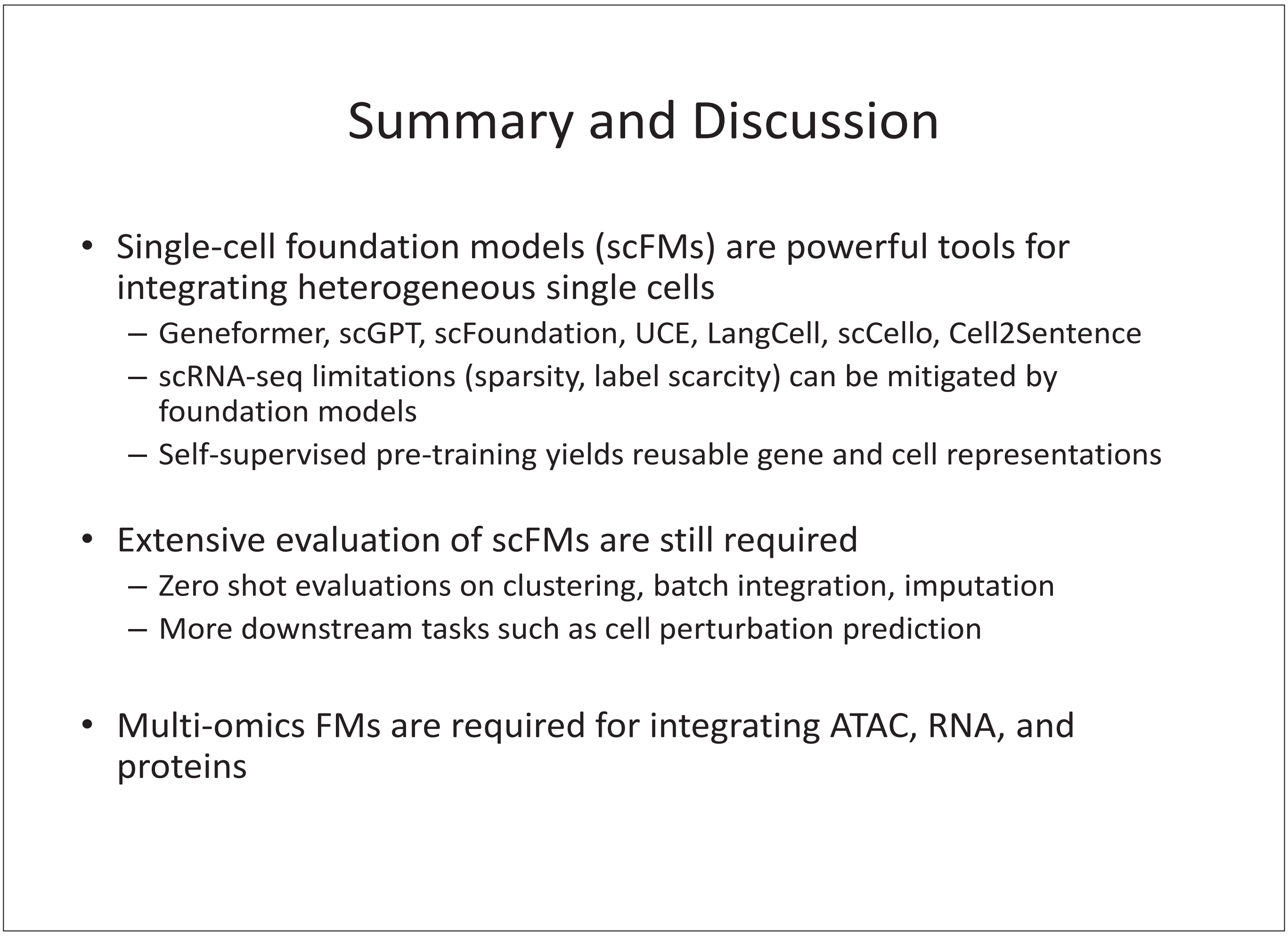

a
Best F1
Worst F1

	Additional dropout 0%									Additional dropout 30%									Additional dropout 50%								
scRobust	0.770	0.904	0.947	0.943	0.963	0.964	0.980	0.959	0.940	0.670	0.883	0.936	0.922	0.929	0.941	0.946	0.922	0.931	0.628	0.845	0.887	0.888	0.886	0.868	0.874	0.872	0.894
scGPT	0.730	0.800	0.933	0.929	0.857	0.762	0.268	0.908	0.677	0.651	0.722	0.896	0.904	0.802	0.692	0.196	0.839	0.600	0.597	0.670	0.829	0.875	0.761	0.623	0.204	0.797	0.581
CIForm	0.608	0.878	0.939	0.923	0.909	0.930	0.966	0.939	0.893	0.569	0.825	0.899	0.877	0.890	0.833	0.918	0.848	0.646	0.538	0.772	0.842	0.819	0.859	0.717	0.853	0.761	0.612
TOSICA	0.639	0.868	0.823	0.839	0.768	0.620	0.946	0.715	0.357	0.558	0.794	0.644	0.747	0.655	0.569	0.782	0.649	0.280	0.528	0.738	0.565	0.654	0.561	0.494	0.708	0.595	0.283
Concerto	0.629	0.887	0.932	0.934	0.915	0.913	0.997	0.905	0.880	0.577	0.833	0.882	0.900	0.872	0.802	0.935	0.851	0.666	0.541	0.779	0.840	0.857	0.862	0.761	0.845	0.794	0.794
CLEAR	0.473	0.637	0.351	0.395	0.258	0.348	0.709	0.467	0.239	0.418	0.763	0.362	0.448	0.466	0.173	0.470	0.451	0.401	0.384	0.390	0.337	0.341	0.548	0.429	0.479	0.370	0.213
sigGCN	0.583	0.857	0.896	0.878	0.904	0.939	0.939	0.856	0.753	0.545	0.789	0.791	0.769	0.847	0.753	0.873	0.710	0.465	0.515	0.727	0.697	0.674	0.730	0.642	0.766	0.650	0.481
DNNs	0.597	0.846	0.876	0.846	0.904	0.929	0.965	0.778	0.769	0.550	0.785	0.747	0.716	0.799	0.750	0.848	0.675	0.413	0.520	0.730	0.648	0.607	0.662	0.606	0.648	0.579	0.397

- scRobust is particularly effective for rare cell type annotation.



Sejin Park, Hyunju Lee, Robust self-supervised learning strategy to tackle the inherent sparsity in single-cell RNA-seq data, *Briefings in Bioinformatics*, 25 (6): bbae586, 2024



Summary and Discussion

- Single-cell foundation models (scFMs) are powerful tools for integrating heterogeneous single cells
  - Geneformer, scGPT, scFoundation, UCE, LangCell, scCello, Cell2Sentence
  - scRNA-seq limitations (sparsity, label scarcity) can be mitigated by foundation models
  - Self-supervised pre-training yields reusable gene and cell representations
- Extensive evaluation of scFMs are still required
  - Zero shot evaluations on clustering, batch integration, imputation
  - More downstream tasks such as cell perturbation prediction
- Multi-omics FMs are required for integrating ATAC, RNA, and proteins