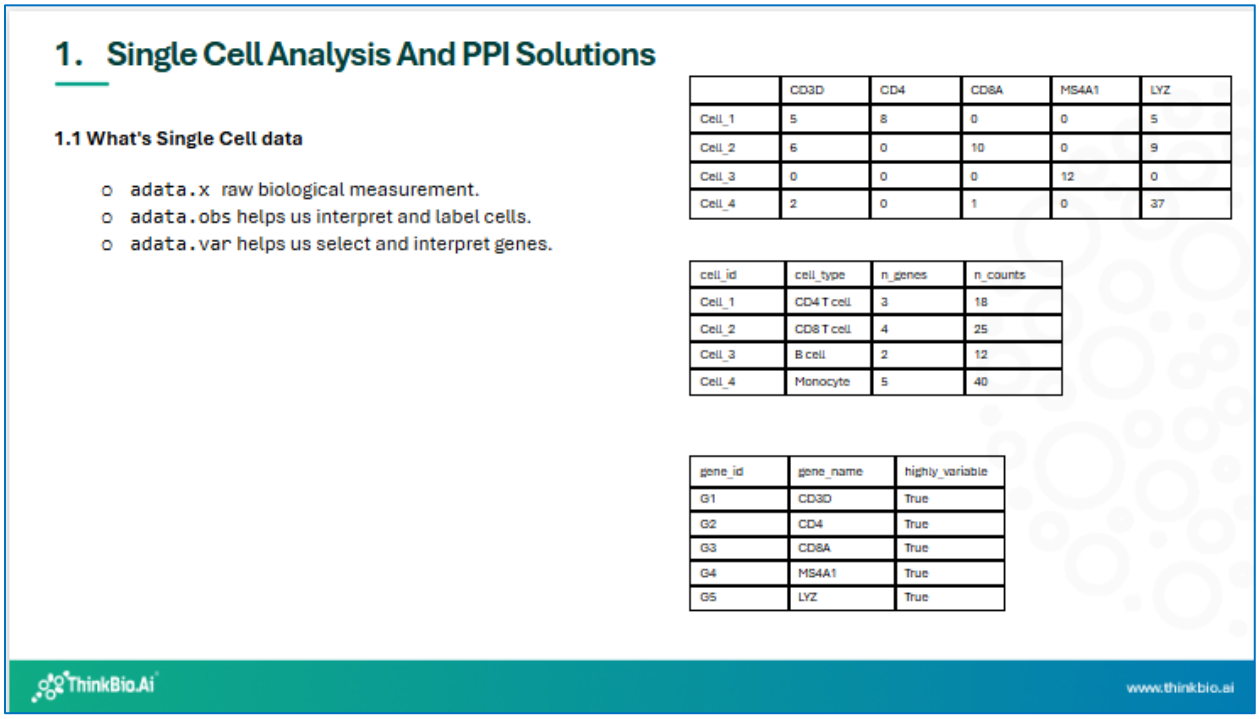
First three slides will be small intro like whats single cell transcriptomics data is? What are the challenges? Excisiting solutions for these challenges and ScNET model comes in which category



- So this slide is more for people who are completely new to Single Cell Transcriptomics data
- In single-cell transcriptomics, gene expression for thousands of individual cells is measured. To organize this high-dimensional data, we commonly use a data structure called AnnData, which stores everything as tables.
- coming to adata.x

  “The most important table is adata.X.

  Here, rows represent cells, columns represent genes, and each value indicates how strongly a gene is expressed in a given cell.”

  “For example, if a cell shows high expression of the gene CD4, it is likely a CD4 T cell.”

  “Next is adata.obs,

  which contains metadata about each cell.

  Each row corresponds to the same cell in adata.X.”

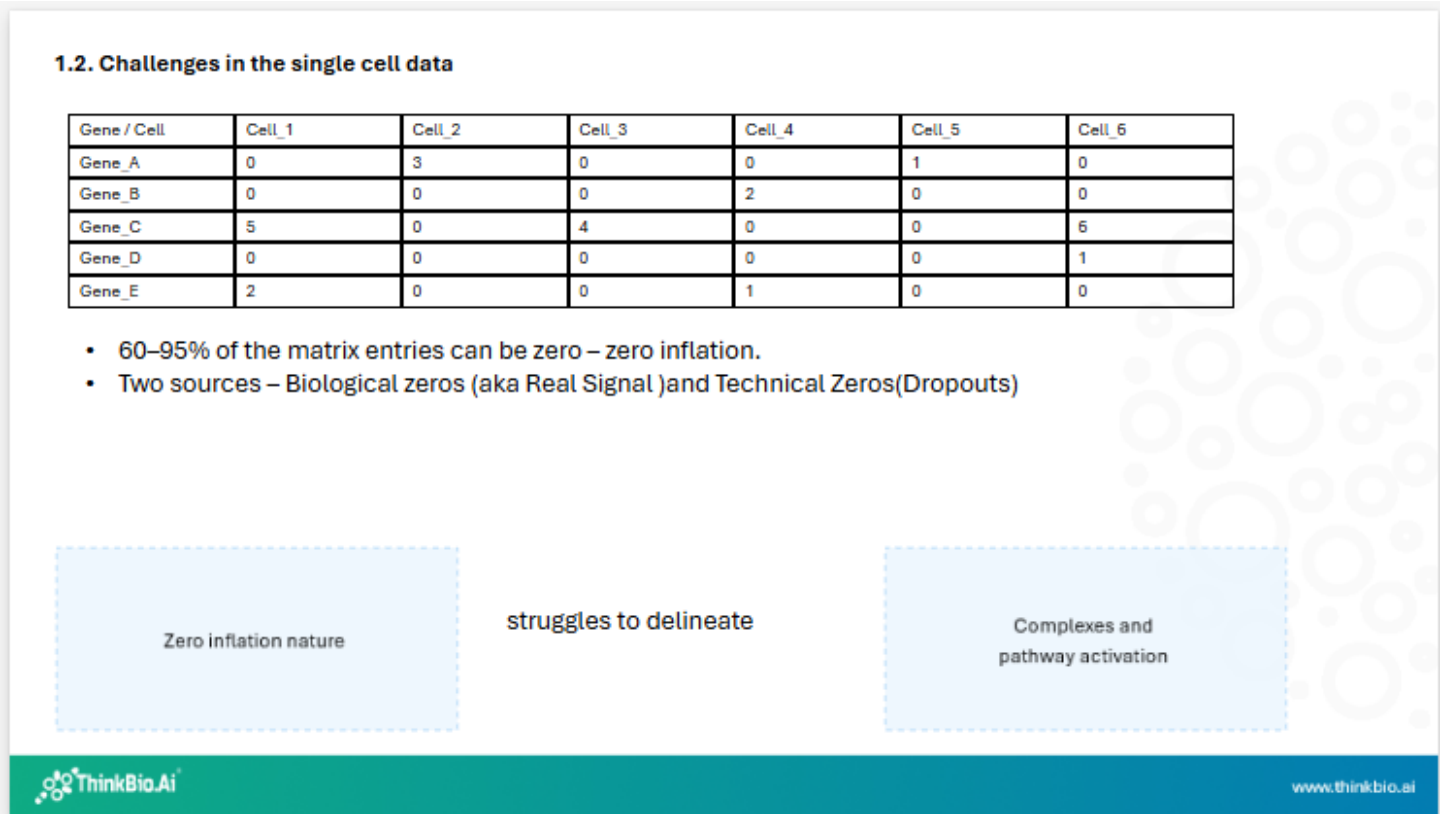
  “This includes information such as cell type annotations, total number of genes detected, or sequencing depth.”

- “Finally, adata.var contains metadata about genes.

  Each row corresponds to a gene in adata.X.”

  “In this example, it includes gene names or whether a gene is highly variable and therefore useful for clustering.”
- “All three tables are linked by their rows and columns, allowing us to connect gene expression values with both cell-level and gene-level information.”

- Single-cell RNA-seq (scRNA-seq) has become extremely widespread over the last ~8–10 years.
- Today, thousands of single-cell datasets are publicly available across many tissues, species, and diseases. Most of the foundation models here we use in thinkBio uses single cell transcriptomics data.



- But in reality those people who have worked with the count matrix knows 60-95% of values are zero -this is zero inflation
- By definition zero inflation refers to the fact that the data contain far more zero values than expected, even for genes that are truly expressed in a cell.
- There are two zeros for zero inflation - Biological zeros (aka Real Signal)and Technical Zeros(Dropouts)

1. Biological zeros

-also called as real signal

- These zeros are genuine

because

gene that is truly not expressed in that cell

Different cell types express different genes

Cell state (cycle, activation, differentiation) matters

2. Technical zeros (dropouts)

- also called as dropouts

-These are false zeros caused by limitations of the technology in terms of

-Low mRNA capture efficiency

-Inefficient reverse transcription

- Limited sequencing depth

simply putting A gene was expressed, but its transcripts were not captured or sequenced → recorded as zero also called a dropout event

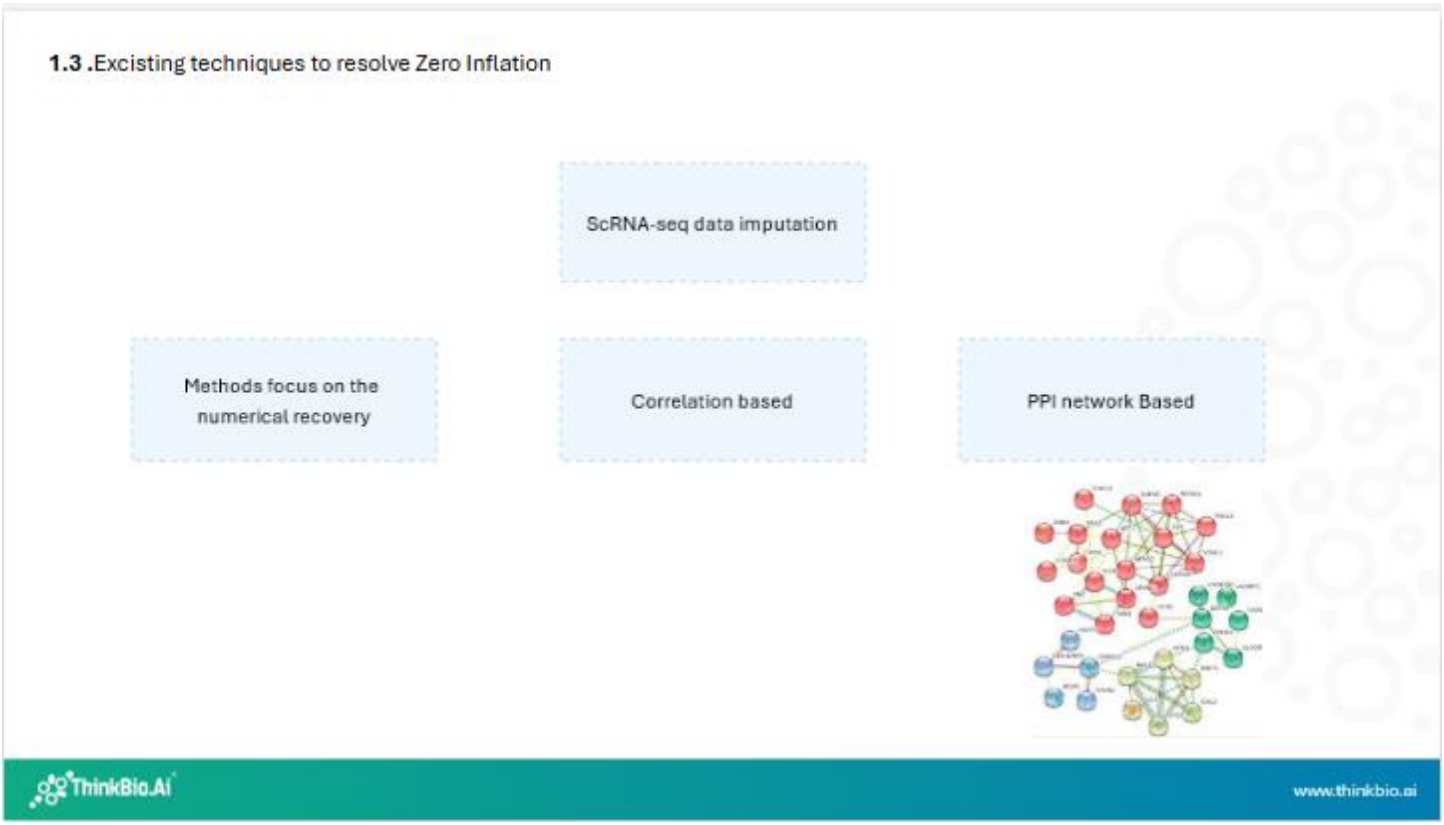
- This High sparsity and dropout in scRNA-seq data makes it difficult to delineate protein complexes and pathway activation states. Pathways and complexes rely on coordinated expression of multiple genes and Dropouts break that coordination

This weakens:

Co-expression signals

Gene set enrichment

Network inference



And to resolve zero inflation - Imputation methods are used

1. methods focus on the numerical recovery

1.1 like zero inflated probabilistic models

1.2. nearest neighbour-based smoothening of expression profiles

- Impute gene expression by borrowing information from transcriptionally similar cells.

1.3. dense latent representations

- Learn low-dimensional latent spaces that capture biological structure and reconstruct denoised expression profiles. It more like an auto encoder model

2. Correlation based

considers sequence depth and read errors to infer the coexpression which is more useful when separating true biological signal from noise

3. PPI Networks

So what are PPI networks in short nodes are proteins and edges are the interactions between the protiens like Physical Binding and Functional pathways

and ScNET is PPI network based

And what separate PPI network based imputation from rest?

FIRST POINT

PPI networks can capture functional context of genes in terms of

* Signal Transduction

PPI networks represent signal transduction as chains of interacting proteins that transmit information from receptors to effectors.

Edges show how signals propagate through physical protein interactions.

* Pathway Activation

Pathway activation appears in PPI networks as coordinated activation of a connected subnetwork corresponding to a known signaling pathway.

Changes in node activity (expression, phosphorylation) highlight which pathway modules are “on.”

* Protein Complex Activation

Protein complex activation is captured as densely connected clusters of proteins within the PPI network.

Activation reflects the assembly or functional engagement of these interacting proteins.

SECOND POINT

Talking about PPI most of these PPI networks are constructed at global scale- means these are constructed at organism level not wrt to the cell type or it does not reflect on the dynamic changes happening across different cell types and biological conditions

THIRD POINT

So by integrating PPIs with ScRNA has its own potential - combines two things contextual information from PPI + Strong functional annotation from ScRNA-seq

FOURTH POINT

Excisting studies on PPI integration with Rna -seq include

* idea of propogating gene expression across PPI edges.

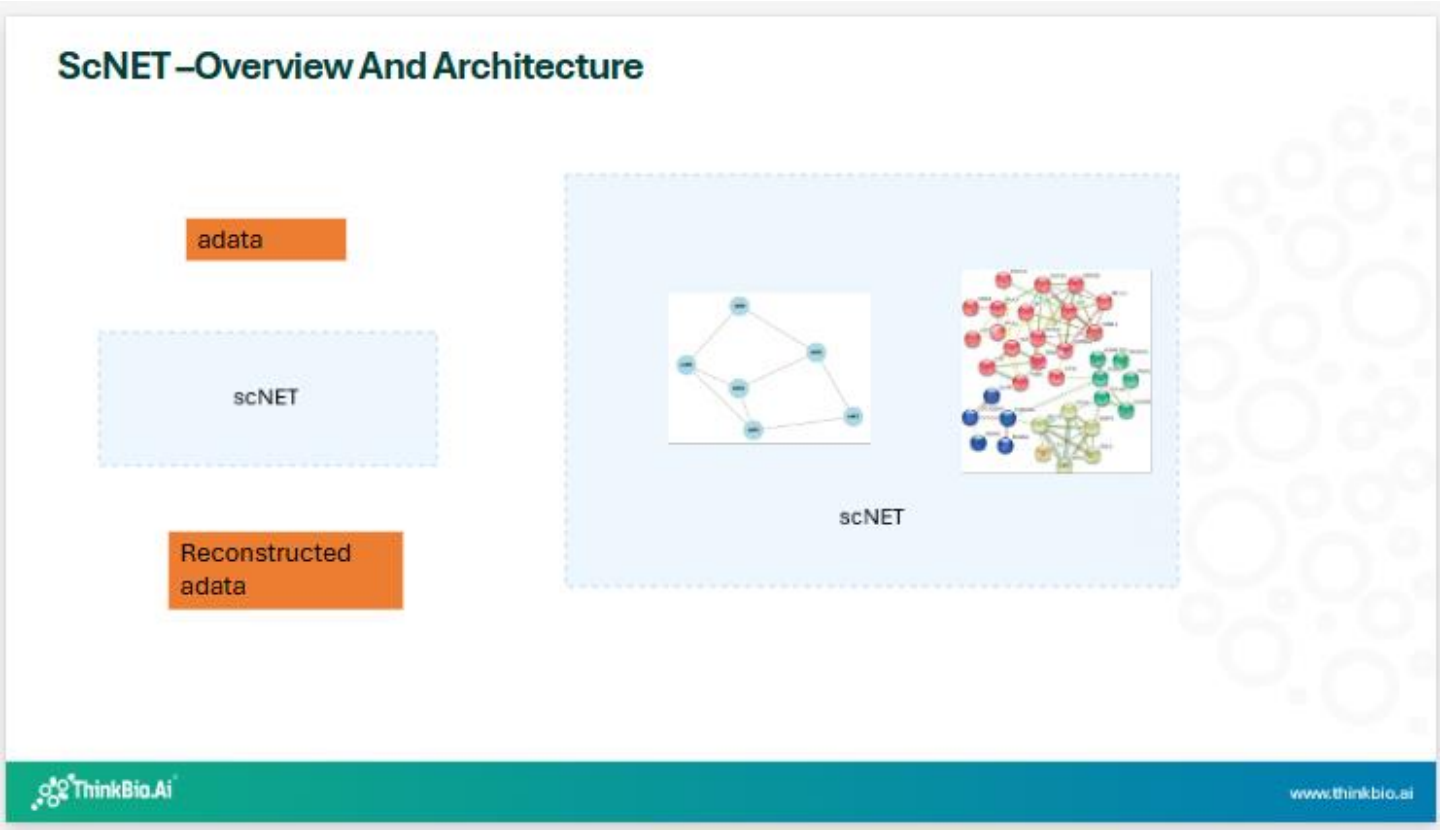
* How PPi networks with differ expressed genes to predict patient's survival rate in NscLC

* How it helps in dimensionality reduction, dropout imputation , cell-cell similarity analysis.

* ScLine recent model that integrates RNA-seq with different biological networks cell-cell and gene-gene relationship.

* Pinnacle - paper describes Advantages Integrating RNA-seq with PPI information at atlas level

* ScNET is about PPI integration of PPI at Dataset level .



Coming to ScNET – it's a PPI network based scRNA-seq data imputation deep learning framework primarily to handly the zero-inflation problem.

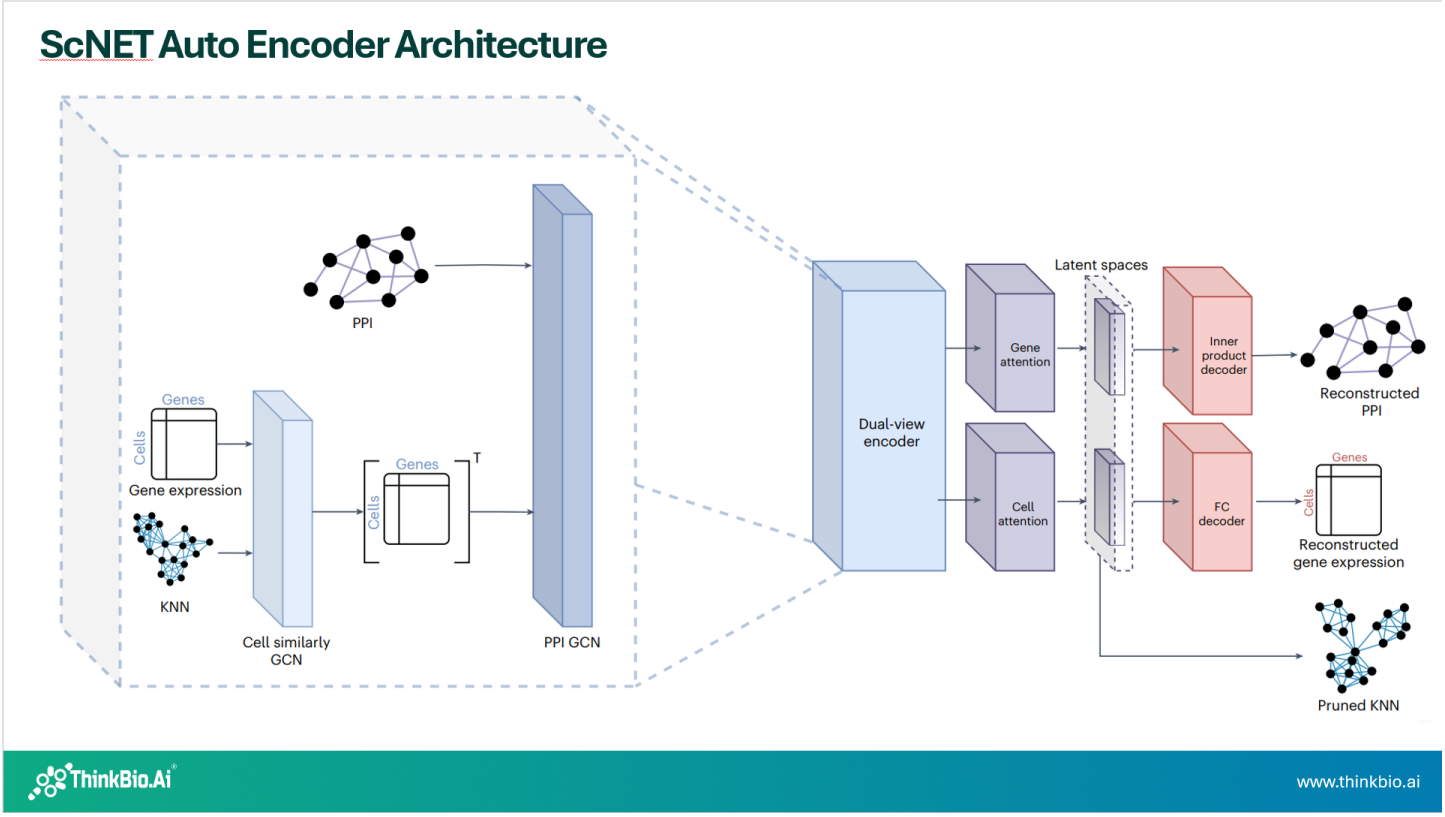
It employs a dual view encoder to integrate RNA-seq with PPI information.

This approach models gene-gene relationships under specific biological contexts and refines cell-cell relationships using an attention mechanism

By using such a reconstructed expression it captures better gene annotation, pathway characterization and gene–gene relationship identification, while improving cell clustering and pathway analysis across diverse cell types and biological conditions.

- So the diagram on the left is a kind of high level understanding or blackbox for whoever wanted to use ScNET
    - Input is your RNA seq data and output is the reconstructed RNA-seq expression data
- On the right opening the scNET black box – you can see two networks
    - One is cell to cell interaction based on K-Nearest neighbour and gene-gene interaction network, these two networks are for learning cell-cell interactions on one hand and gene-gene interactions on the other hand.
    - So as you know the heart of GNN is message passing, like each node receive information from neighbours , Aggregates it and updates its own representation like that these two network work as GNN.

- And also one more thing the gene expression information is propgated to both the network alternatively
- The cell -cell network also bring edge attention mechasim to refine the cell-cell relationships
    - And from my undrsatanding these attention mechanism actually gives the idea to the cells, not all interactions in cells are equally important, and it also relaxes the condition of fixed number of connections per cell.



Now coming to the diagram in the paper- labelled as Autoencoder Architecture, so Scnet also follows an Autoencoder model architecture, as you might know autoencoders are models that learn compressed represenatations of data by encoding the input into a lower dimensional bottleneck and then decodeing it back to rconstruct the original , aiming to minimize the reconstruction error

And this is the most confusing partabout ScNET if its an autoencoder then why its not an USL- and the answer is – ScNET is unsupervised model and its also an autoencoder model.

Means it does not use labels (no cell types , conditions or phenotypes during the training).

Training is purely based on the structure of the two networks.

The objective is to **learn embeddings** that preserve network topology and biological relationships, not to predict a supervised target.

So more specifically → its an **unsupervised representation learning**.

Coming to the workflow

- This flow diagram looks little confusing just look at the inputs
    - KNN graph , PPI and the adata.
- This is input to the dual view encoder
- Then you can see two attention layers these two are employed to extract the latent representations of both cells and genes
- Then from the latent spaces the inner product decoder is used to reconstruct the network connections

- Whereas the fully connected layer is responsible for reconstructing the genes expression
- Like I told about the need of attention at cell-cell essential connections are only kept thus the KNN graph will cut so much connections and finally apruned KNN graph is the result.

The Dual view encoder capture both the network structure and expression information while reducing the noise level of data.

Before that you can see two Convolution layers one is specific to cell KNN and the other is specific to PPI or gene

The gene one is Graph convolution layer and the cell one is graph attention layer.

You might be thinking why different convolution layers

scNET uses

GCN aggrates averaged info from all the nodes around it while GAT uses weight and selects only certain nodes. And why it is here is based on the intuition that

GCN for PPI (gene–gene graph) → because PPIs are global, static, and trusted

GAT for cell–cell KNN graph → because cell similarity is **local, noisy, and heterogeneous**

GCN = “average reliable neighbors”
GAT = “pay attention to the most relevant neighbors**”**

KNN graph pruning using attention coefficients.

Now talking about the pruned KNN graph output , that it selects only K other cells in the dataset. This assumption may be wrong as the as cells from different populations and states may be represented in varying numbers within the data,thus this value might also vary so its not biologically accurate but at the same time we need datset specific

So the learned attention coefficients prune only low quality edges.

Network Evaluation

So to asses the predictive power of networks generated, they considered KEGG pathways.

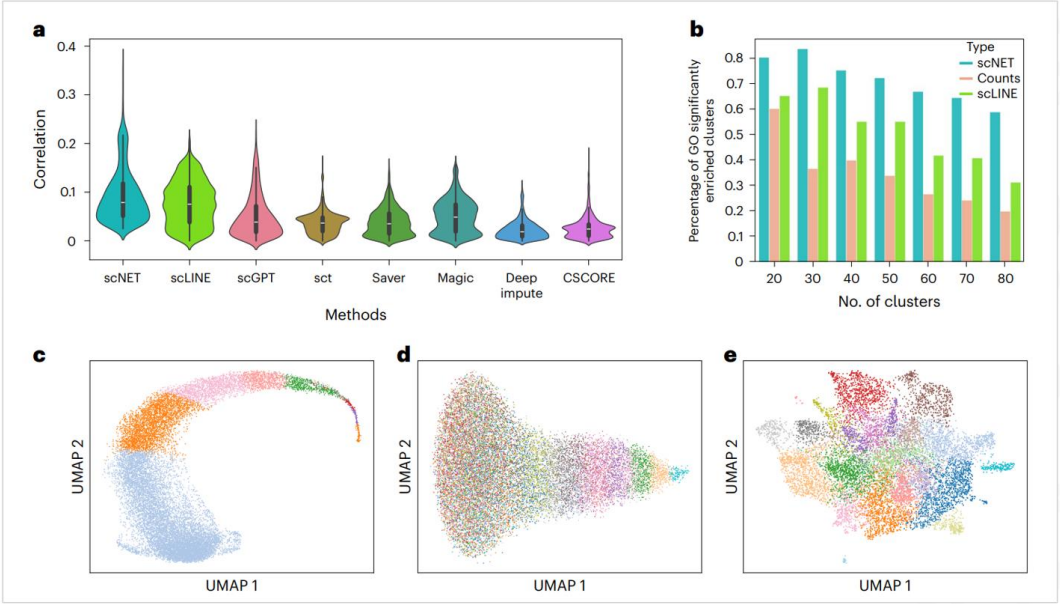
Each pathway was divided into training and test set in the ratio 2:3 and 1:3

Then Random walk with restart approach – and it’s the network based algorithm used to mesure node relevance or similarity in the graph by simulating a walker that repeatedly explores the network but occasionally jumps back to the starting node

why the restart matters -it makes the walk biased towards the sources and thus captures the local and global connectivity and noise removal in terms of connectivity and thus the membership value is passed to all other nodes, and this is the idea for using ROC-AUC for the predicted connection value

Gene Embeddings Improve Functional Annotations

- Gene Ontology (GO) semantic similarity value
- Co-embedded coefficient



This slide and all these results are about the utility of the gene embeddings. As you might know gene embeddings the compact, numerical vector representations of genes, which is specific to the biological contexts

FIRST VIOLIN PLOT

- The first violin plot is about correlations in the embedding space accurately reflected the known biological annotations and functions for this two metrics were used – GO semantic similarity value and coembedded coefficient for every gene pair
- Talking about GO semantic similarity value- it's the numerical score that quantifying how functionally alike two genes or GO terms are based on their shared position and information in hierarchical GO graph
    - Ranges from 0 -> 1
    - 0 -> no shared memory
    - 1 -> identical
    - Excisting methods use least common ancestor to mesure the relatedness and this is crucial step for building biological networks
- Coembedded Coefficients are of Gene-coexpression network and Relatedness and Linkage
    - In Gene coexpression networks – coeffient in this context is similarity score or correlation coefficient is calculated for evry pair of genes based on their expression data and other properties

These two values were calculated for every gene pair in the embeddingsand then analyzed the distribution by comparing it with the other zero-inflation imputation tools

And as you can see the ScNET has substantially higher mean correlation.

- Graph 2 is about how the embedding space captures the functional annotations by clustering genes
- So for this experimentation they used k-means algorithm with cluster numbers ranging from 20-80.

- measured the percentage of clusters significantly enriched for one or more GO terms.and

  GO (Gene Ontology) terms by the way are **labels for gene functions**, such as:

  - “DNA repair”
  - “Cell cycle”
  - “Immune response”

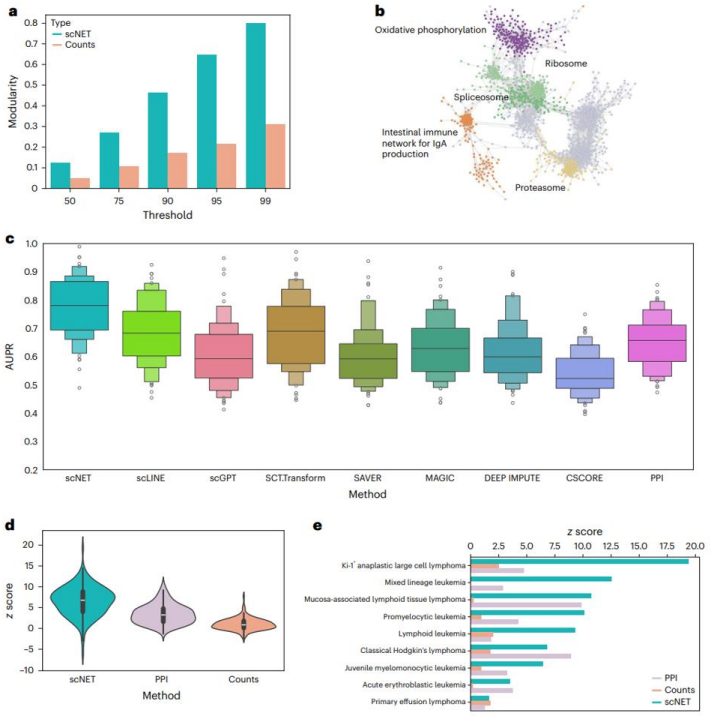
This enrichment was done using GSEA tool – and GSEA is the computuational method that identifies whether a predefined group of genes represents some pathways or functions

And if you look at the diagram here you can see its showing enhanced clustering efficacy of ScNETs Gene embedding compared to other imputation techniques

Also a visual representation also showing smaller more well defined clusters following a 30-means clustering

- ***Gene Embeddings Improve Functional Annotations***

- Co-embedded Network



Coming to coembedded network

Its the final **gene network** is learned in a space.

Compared to a PPI-only network, you get:

- Cell-type–specific gene relationships
- Removal of irrelevant PPIs
- Strengthening of interactions active in your data

So why they are considering this network is it serves as a strong inference tool in Bulk RNA-seq data but has not been replicated in Single cell data.

For this Anlaysis they used Malaria associated Bcells and pairwise absolute value correlations were computed in original and embedding space and the thresholds were set at different percentiles. Then the modularity value was calculated at each percentiles modularity here means - A number that tells how clearly a network splits into well-defined communities (modules)

High modularity → clear communities

Low modularity → weak or no community structure

Typical values:

- ~0 → random-like network
- 0.3–0.6 → strong community structure

Here also ScNET compared to Counts give better results. And if you see at 99 the difference is too high and this percentile was employed as the threshold for subsequent analysis

The second image is again they used KEGG pathways and used in clusters atleast containing 30 genes expressed in the dataset

Coming to the third image

*Each pathway was separated into training and test set and each gene in the training set was assigned a membership value of 1 and the idea is to propogate these membership values to rest of the genes and evaluate the quality of reconstructing the test set and it is measured using AUPR

Here also compared to other imputation techniques scNET gives better results.

Finally the last two images its about the reconstructed gene lists associated with different diseases. This is not like KEGG , its rather the general associations of genes with the conditions and if you check for the second last graph Scnet achieved a mean score of 7

Last Figure is also about the reconstructed gene lists associated with different types of leukemias and lymphomas , these are c**ancers of the blood and immune system**, specifically arising from **white blood cells**.

Leukemia is “cancer in the blood”

Lymphoma — “cancer in lymph nodes”

These two are strongly associated with B-cells.

In this analysis 6 out of 9 gene lists scNET coembedded network performed better than both other networks.

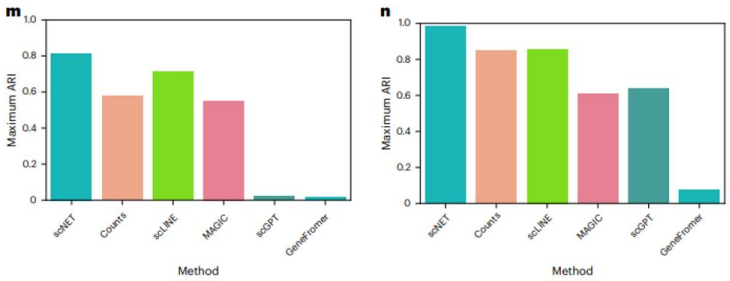
These results show the reconstructed expression demonstrates a synergistic effect in the integration.

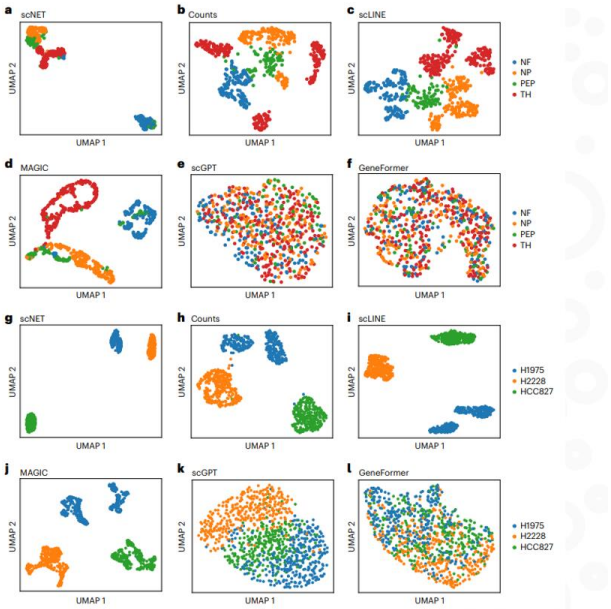
Evaluation of cell clustering

- *Cell clustering Evaluation*

- Dorsal root ganglia (DRG)

Type	Main Function	Sensation
NF	Touch & proprioception	Non-painful touch
NP	Mechanical pain & itch	Pain
PEP	Thermal & inflammatory pain	Pain
TH	Pleasant touch	Gentle touch





Cell clustering can be evaluated using the cell embedding where each cell in your dataset is projected into the high dimension vector. For evaluating the cell clustering they experimented on

Dorsal Root Ganglia (DRG) and these are are clusters of sensory neurons located near the spinal cord

They carry information such as:

- Touch
- Pain
- Temperature
- Itch
- Proprioception (body position)

Each DRG neuron type is specialized for a different kind of sensation. They used the well-established DRG sensory neuron classes for clustering

NF — Neurofilament neurons

NP — Non-peptidergic nociceptors

PEP — Peptidergic nociceptors

TH — Tyrosine hydroxylase neurons.

Also they consider study on **H1975, H2228, and HCC827 which are human lung cancer cell lines**, commonly used in cancer and drug-response studies.

Cell line	Cancer type
H1975	Lung adenocarcinoma
H2228	Lung cancer
HCC827	Lung adenocarcinoma

The Foundation models Geneformer and ScGPT here they used finetuning. Leiden clustering was then employed with resolutions 0-1 and calculated maximum ARI with respect to original cell labels

Adjusted Rand Index (ARI) measures **how similar two clusterings are**.

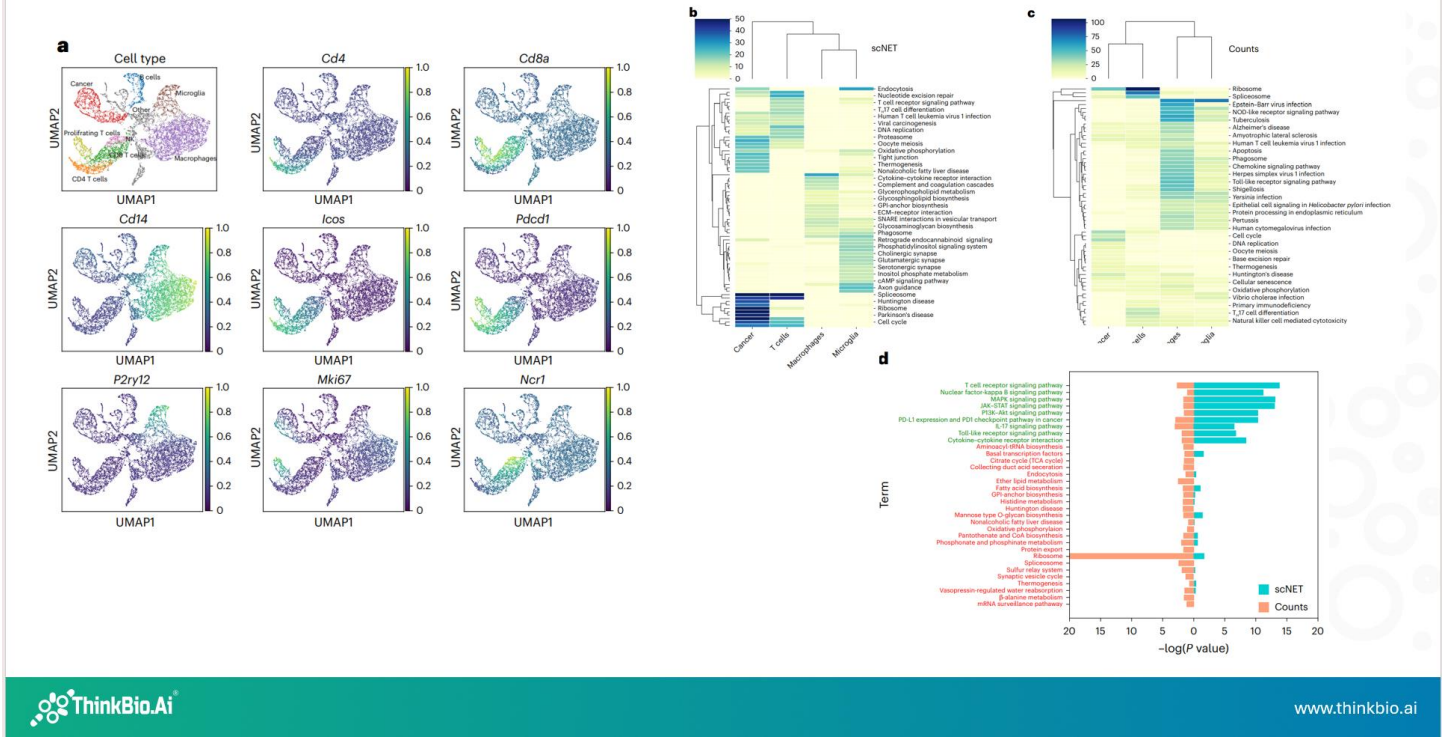
ARI value	Meaning
1.0	Perfect match
~0.5	Good agreement
0	Random clustering
< 0	Worse than random

So the conclusion here is even though these foundation models are trained on very large datsets zero shot prediction they give weak performance and there is the need for more unsupervised learning methods

And about the umap for Scnet the clusters are not further separating into separate communities

scNET reduces zero inflation and improve pathway analysis

- *scNET reduces zero inflation and improve pathway analysis*



This is the analysis we carried out using the same dataset. So its about whether the reconstruced gene expression accurately captures the unique expression dynamics of different cell population and the datset used was GL261a mouse brain tumor model dataset and visualize the reconstructed gene markers fr different cell populations

the expression of

- Cd4 and Cd8a corresponded accurately to their respective T cell subsets
- Cd14 effectively identified macrophages
- P2ry12 captured microglia cells with precision

Also AUPR for each cell type based on its respective marker gene was calculated

compared the AUPR scores across the original counts data – Here also ScNET gave consistently higher AUPR scores accross all cell types.

Coming to Figure 5b it shows the reconstructed expressions ability to better capture the differential pathways between clusters and cell populations, primarily the focus was on four main cell populations T cells, macrophages, microglia and cancer cells.

Standard differential gene expression analysis was done and the resulting differentially expressed genes for each cluster were used to calculate enriched KEGG pathways using GSEA and captured the top 20 most enriched pathways for each cell population.

5c is the the same analysis on the original gene expression data