Team 32

Transformer based cellular classification using RNA sequencing data

Geetika Agrawal, Pragnya Pathak, Vaishnavi Jahagirdar, Vivekanand Sahu

Introduction

Background & Motivation

- **Enhancing Disease Understanding and Treatment**
  - Understanding cellular specialization
  - Developing targeted therapies
- **Overcoming Data Limitations :**
  - Rare diseases and inaccessible tissues
  - Accelerating therapeutic target discovery

Our Objective

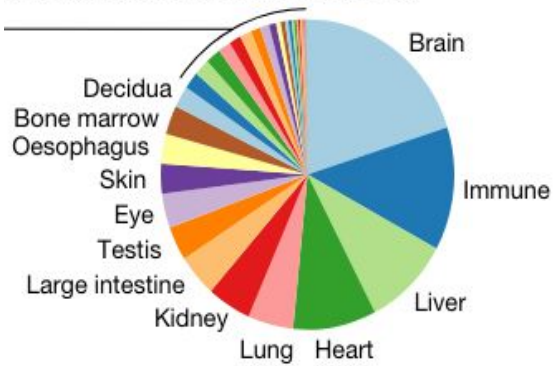
- Transfer learning on the **Geneformer model** for classifying cells and genes
- Uses **transcriptomic** and **genomic** data
- Improve accuracy and reliability of cell type identification

Input & Output

- **Input:** Transcriptomic and genomic data, particularly single-cell RNA sequencing (scRNA-seq) pancreas dataset
- **Output:** Accurate classification of cells and dosage-sensitivity associated with pancreas

Literature review

Model	Parameters	Training Data	Architecture	Downstream Tasks
scGPT[4]	51 million	33 million normal human cells (51 tissues, 441 studies)	Autoregressive transformer	Cell-type annotation, multi-batch integration, multi-omic integration, perturbation prediction, and GRN inference
scBERT[6]	5 million	209 human single-cell datasets comprising 74 tissues with over 1 million cells	Performer (allowing for longer inputs)	Cell type annotation
GeneFormer[1]	**40 million**	**Genecorpus-30M,** 29.9 million human single-cell transcriptomes	**Transformer**	**Gene dosage sensitivity, chromatin dynamics, network dynamics**
scFoundation[5]	100 million	50 million human cells (100+ tissue types, normal and disease)	Transformer	Clustering, perturbation prediction, drug response

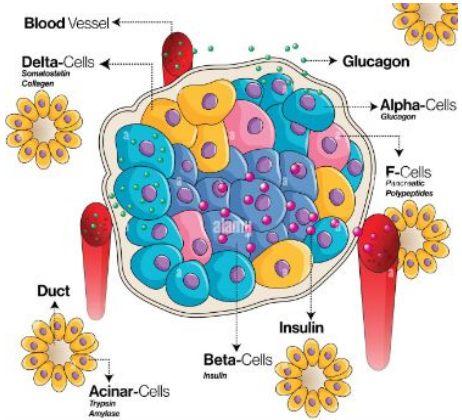


Genecorpus-30M

Our Dataset

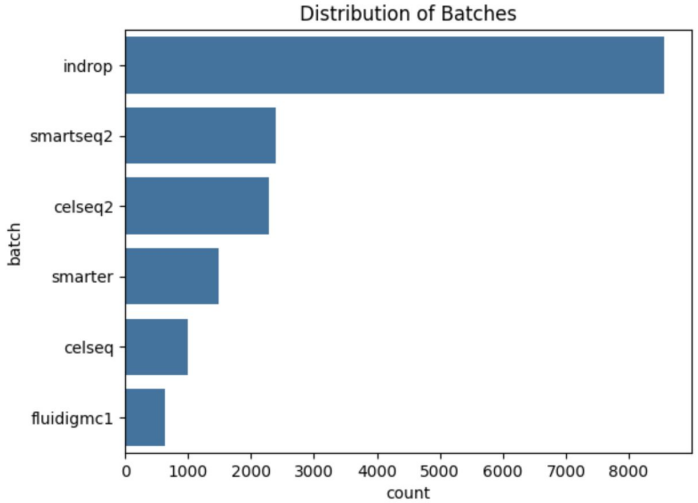
Pancreas Dataset[3] Description

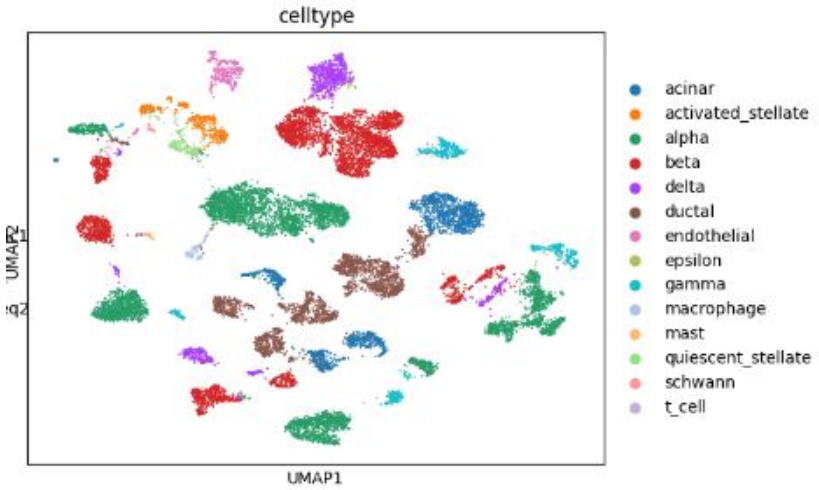
- **n_obs (observations)**: Number of cells in our dataset, which is 16,382 cells.
- **n_vars (variables)**: Number of genes measured across all cells, which is 19,093 genes.



Understanding importance of each cell type:

- **Diabetes:** Alpha, Beta, Delta, Gamma, Epsilon Cells
- **Cancer:** Macrophages, T cells
- **Cardiovascular Diseases:** Endothelial Cells
- **Neurodegenerative Diseases:** Schwann Cells





Methodology

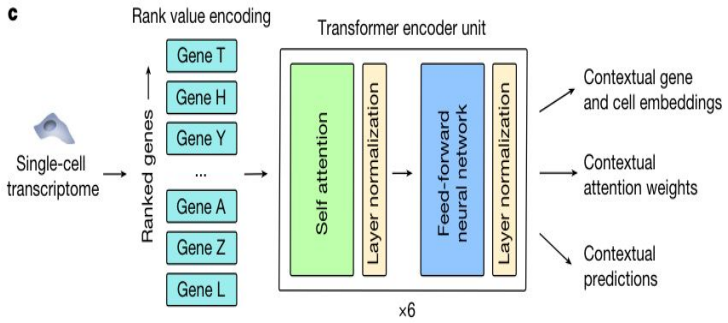
Pre Preprocessing:

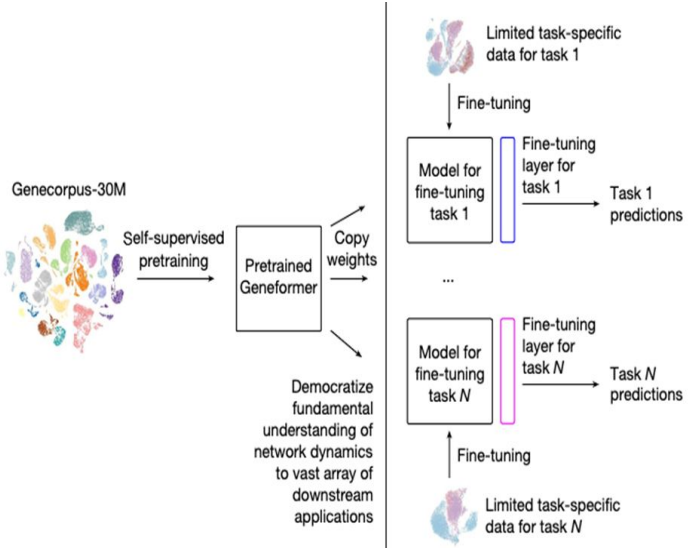
- Input: single-cell transcriptomes
- Each cell's transcriptome is represented as a rank value encoding.

Model Architecture:

- **Masked Learning:** 15% of the genes within each transcriptome were masked, and the model was trained to predict which gene should be within each masked position in that specific cell state using the context of the remaining unmasked genes.

- **BertForSequenceClassification**: Classifying cells based on gene expression data from scRNA-seq into specific types or states. It produces a single classification for each cell, like "neuron" or "immune cell."





Training Enhancements

1) **Layer Freezing:**

   Specifies the number of layers in the model to be kept unchanged during training to retain previously learned features.

2) **Cross validation:**

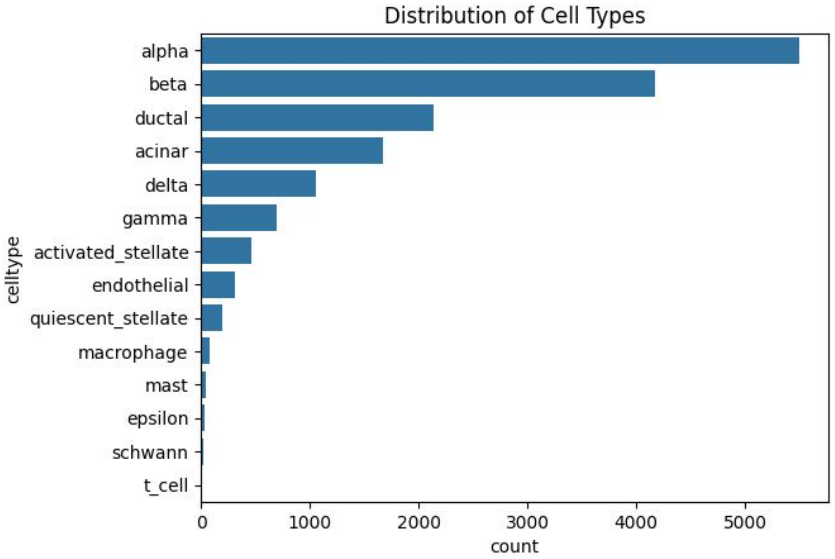
   A method that cyclically splits the data into training and validation sets to ensure robust evaluation and mitigate overfitting.

3) **Upsampling:**

   A technique to increase the number of samples in underrepresented classes to balance the dataset and improve model performance.

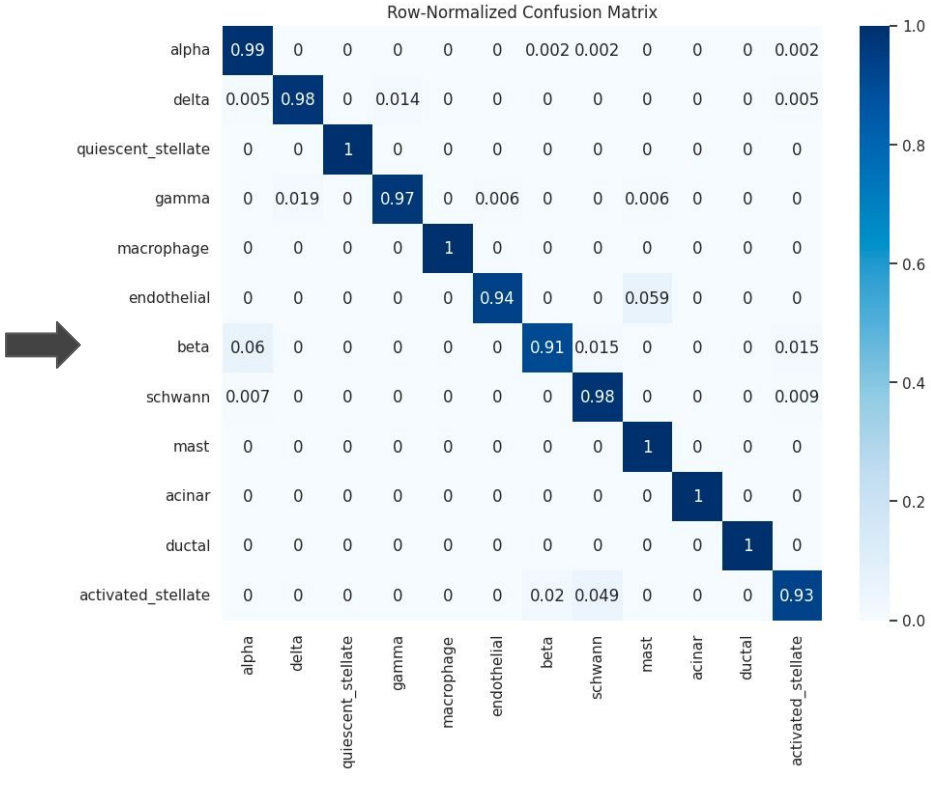
4) **Dice / CE / BCE losses :**

   These loss functions address class imbalance by giving more importance to minority classes, ensuring better model accuracy and fairness.

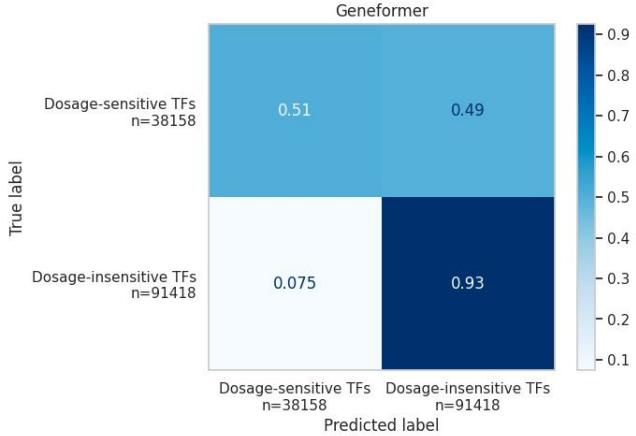


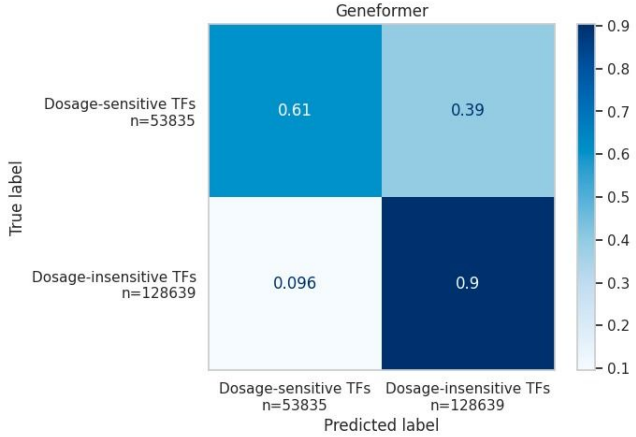
Results


Row-Normalized Confusion Matrix
alpha 0.94 0.017 0 0 0 0 0.033 0 0 0.002 0.002 0.002
delta 0.019 0.54 0 0 0 0 0.44 0 0 0 0 0
quiescent_stellate 0.12 0.04 0.08 0 0 0 0.04 0 0 0 0.04 0.68
gamma 0.23 0.43 0 0 0 0 0.35 0 0 0 0 0
macrophage 0.1 0 0 0 0.7 0 0.1 0 0 0 0.1 0
endothelial 0 0.024 0 0 0 0.95 0.024 0 0 0 0 0
beta 0.002 0.028 0 0 0 0 0.97 0 0 0 0 0
schwann 0 0.5 0 0 0 0 0 0 0 0 0 0.5
mast 0.33 0.33 0 0 0 0 0 0 0.33 0 0 0
acinar 0.029 0.011 0 0 0 0 0.011 0 0 0.89 0.051 0.006
ductal 0.026 0.026 0 0 0 0 0.005 0 0 0.026 0.92 0
activated_stellate 0 0 0 0 0 0 0 0 0 0 0 1
alpha delta quiescent_stellate gamma macrophage endothelial beta schwann mast acinar ductal activated_stellate
1.0 0.8 0.6 0.4 0.2 0.0


Row-Normalized Confusion Matrix
alpha 0.99 0 0 0 0 0 0.002 0.002 0 0 0 0.002
delta 0.005 0.98 0 0.014 0 0 0 0 0 0 0 0.005
quiescent_stellate 0 0 1 0 0 0 0 0 0 0 0 0
gamma 0 0.019 0 0.97 0 0.006 0 0 0.006 0 0 0
macrophage 0 0 0 0 1 0 0 0 0 0 0 0
endothelial 0 0 0 0 0 0.94 0 0 0.059 0 0 0
beta 0.06 0 0 0 0 0 0.91 0.015 0 0 0 0.015
schwann 0.007 0 0 0 0 0 0 0.98 0 0 0 0.009
mast 0 0 0 0 0 0 0 0 1 0 0 0
acinar 0 0 0 0 0 0 0 0 0 1 0 0
ductal 0 0 0 0 0 0 0 0 0 0 1 0
activated_stellate 0 0 0 0 0 0 0.02 0.049 0 0 0 0.93
alpha delta quiescent_stellate gamma macrophage endothelial beta schwann mast acinar ductal activated_stellate
1.0 0.8 0.6 0.4 0.2 0.0

Results



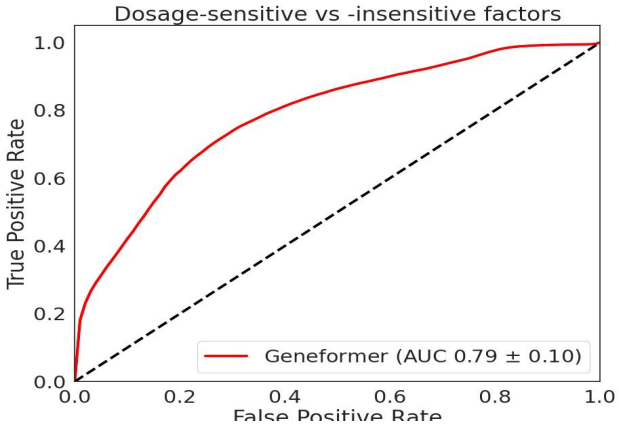


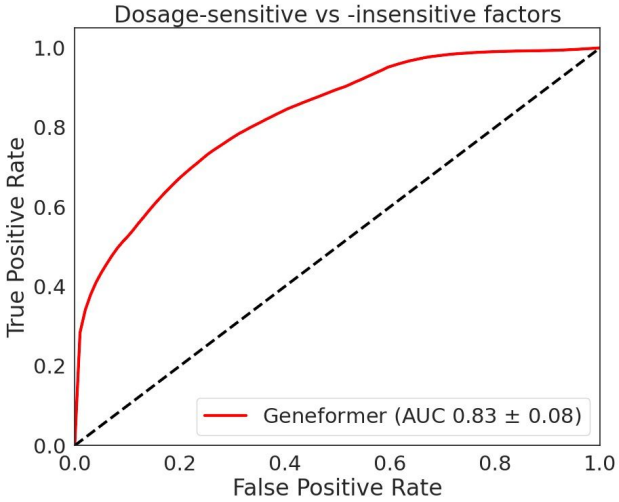
Dosage Sensitivity

Baseline

vs

Enhanced





Conclusion & Discussion

- We were able to utilize Geneformer capabilities cell classification and disease-related gene expression profiling on our pancreas dataset using Transfer Learning.
- Final accuracy values:
  - Cell classification ~ **97%**
  - Dosage-sensitivity ~ **94%**
- The findings provide valuable insights into the roles of specific cell types in various diseases and highlight potential biomarkers and therapeutic targets.

Future Scope:

- We can use techniques like int-8 quantization and pruning to efficiently train the model with optimal resource utilization.

References

[1] Theodoris, C.V., Xiao, L., Chopra, A. et al. Transfer learning enables predictions in network biology.

[2] Christina Theodoris, et al., "Genecorpus-30M", Hugging Face

[3] **Pancreas Dataset**: "Pancreas Dataset." figshare, 2022. https://doi.org/10.6084/m9.figshare.20184524.v1

[4] Cui, et al., 2023. scGPT: An Autoregressive Transformer Model for Single-Cell Analysis. *bioRxiv*.

[5] Hao, et al., 2023. scFoundation: A High-Parameter Transformer Model for Single-Cell Clustering and Drug Response Prediction. *bioRxiv*.

[6] Yang, et al., 2022. scBERT: A Performer-Based Model for Single-Cell Type Annotation. *Nature Machine Intelligence*.