Final Project For ECE228: Track Number 1

Group Number 32:
Geetika Agrawal
Pragnya Pathak
Vaishnavi Jahagirdar
Vivekanand Sahu

Abstract

In this project, we employed GeneFormer for cell and gene classification to enhance our understanding of diseases and accelerate therapeutic discovery. This project addresses the challenge of limited data in rare diseases and inaccessible tissues, facilitating the discovery of therapeutic targets. By applying transfer learning to the GeneFormer model, transcriptomic and genomic data, particularly single-cell RNA sequencing (scRNA-seq) datasets from the pancreas, were leveraged to enhance the accuracy and reliability of cell type identification. This method enabled precise classification of pancreatic cells and identification of dosage sensitivity associated with the pancreas, thereby improving the understanding of pancreatic diseases and development of targeted therapies.

0.1 Evaluation

I certify that I have filled the evaluation.

1 Introduction

In recent years, comprehending the complex networks of gene regulation that drives disease progression has become crucial for developing effective treatments. Traditional approaches often have shortcomings as they target peripheral downstream effects rather than the core elements critical for disease modification. Our project aims to address this problem by leveraging the Geneformer [5] model for gene and cell classification. This approach harnesses the power of transfer learning, allowing deep learning models pretrained on extensive datasets like Genecorpus-30M [4] (used in our project, Figure 1) to be fine-tuned for specific tasks having limited data. By using transcriptomic and genomic data, particularly from single-cell RNA sequencing (scRNA-seq) of pancreas cells [3], we aim to enhance the accuracy and reliability of cell type identification and gene dosage sensitivity.

The prime motivation behind this project is twofold: to improve our understanding of cellular specialization and to develop targeted therapies that can significantly impact disease treatment. Traditional techniques are generally hampered by the lack of data, especially for rare diseases or tissues that are very difficult to access. Our approach helps in mitigating these limitations by using a robust model pretrained on a large corpus of single-cell transcriptomes, which can then be fine-tuned to perform well even with limited task-specific data.

Our methodology involves preprocessing single-cell transcriptomes into rank-value encodings and using masked learning techniques to predict gene expression patterns. Enhancements such as layer freezing and cross-validation are utilized to refine the model's performance. Our results are promising, with cell classification accuracy reaching approximately **97%** and gene dosage sensitivity at around **94%**. These findings not only provided us with valuable insights into the roles of specific cell types in diseases but also helped in highlighting any potential biomarkers and therapeutic targets, paving the way for more precise and effective treatments.

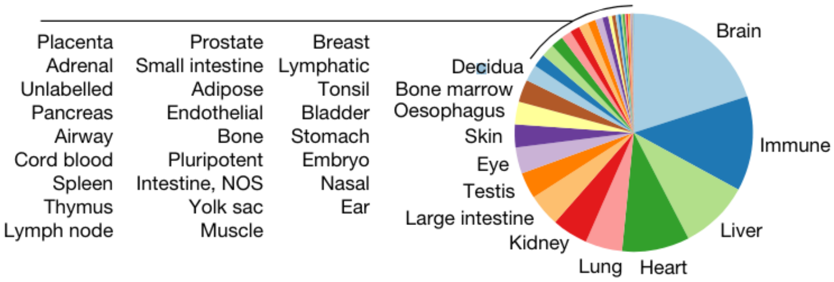

Figure 1: Tissue representation of Genecorpus-30M

2 Related Work

Comprehensive review of several key models like ScGPT [1], an autoregressive transformer model with 51 million parameters. Trained on 33 million normal human cells from 51 tissues and 441 studies, scGPT is proficient in cell-type annotation, multi-batch integration, multi-omic integration, perturbation prediction, and gene regulatory network (GRN) inference. This model's broad applicability highlights the effectiveness of transformer architectures in single-cell analysis.

ScBERT [6], a model based on the Performer architecture, which allows for longer input sequences. With 5 million parameters, scBERT was trained on 209 human single-cell datasets comprising 74 tissues and over 1 million cells. scBERT's primary application is cell type annotation, demonstrating the model's ability to handle diverse single-cell data effectively but with a narrower focus compared to GeneFormer.

ScFoundation [2], a transformer model with 100 million parameters, trained on 50 million human cells from over 100 tissue types, both normal and diseased. scFoundation addresses clustering, perturbation prediction, and drug response, showcasing the versatility and scalability of transformer models in handling extensive and varied datasets.

After evaluating these models, GeneFormer[5] which is attention-based deep learning model pretrained on Genecorpus-30M, which includes 29.9 million human single-cell transcriptomes emerged as the most balanced and specialized for our needs, combining a manageable parameter size with targeted capabilities in gene dosage sensitivity and network dynamics, making it the optimal choice for our project's objectives.

3 Methodology

3.1 Preprocessing

The input consists of single-cell transcriptomes from the pancreas dataset, which includes 16,382 cells and 19,093 genes measured across all cells. Each cell's transcriptome is represented as a **rank value encoding** i.e, ranking genes within each cell based on their expression levels and normalizing by their expression across the Pancreas dataset. This method highlights the important genes for identifying different cell types while downplaying common genes. The rank-based representation is non-parametric and robust against technical artifacts, ensuring consistent analysis across different datasets.

To achieve this, the non-zero median expression value of each detected gene across all cells is calculated. The gene expression in each cell is normalized by the total transcript count of that cell to account for varying sequencing depth, and the normalized gene counts are then ranked. This approach minimizes errors from technical issues and ensures efficient storage and processing.

3.2 Model Architecture

The Geneformer architecture is based on the BERT model fine-tuned for sequence classification. It consists of **6 transformer layers** with a hidden size of 256 and 4 attention heads per layer as shown in Figure 2. The model supports sequences up to 2048 tokens and operates with a vocabulary size of

25426 tokens. Dropout probabilities for attention and hidden layers are 0.02 to prevent overfitting, and ReLU is used as the activation function in the hidden layers. Absolute position embeddings and layer normalization (epsilon = 1e-12) are used. Model weights are initialized between -0.02 and 0.02. It's designed for single-label classification tasks where label mappings and caching mechanisms are implemented for efficient processing.

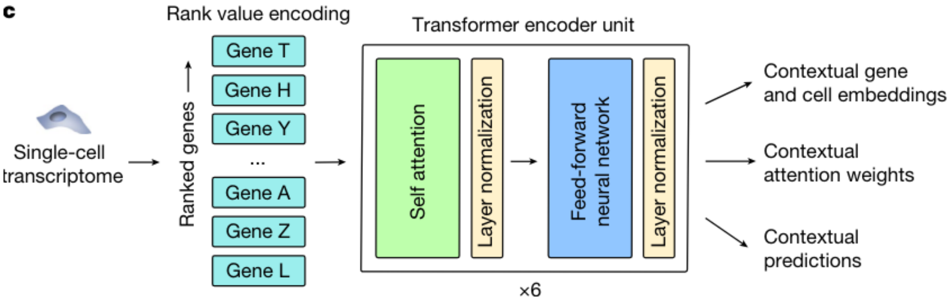

Figure 2: Geneformer [5]

The key innovation of our architecture include:

- **Masked Learning:** During pretraining, 15% of the genes within each transcriptome are masked. The model is trained to predict the masked genes using the context of the remaining unmasked genes. This self-supervised approach leverages large-scale unlabelled data to gain a fundamental understanding of entire network dynamics. The loss function for masked learning can be represented as:

$$L = -\sum_{i \in M} \log P(g_i | G \setminus g_i),$$

  where M is the set of masked genes, g_i is the masked gene, and $G \setminus g_i$ represents the transcriptome with the masked gene removed.

- **BertForSequenceClassification:** This component classifies cells based on gene expression data from scRNA-seq into specific types or states, producing a single classification for each cell, such as "neuron" or "immune cell." The classification loss function used is the Cross-Entropy Loss:

$$L_{CE} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i),$$

  where y_i is the true label, and $\hat{y}_i$ is the predicted probability for label i

3.3 Training Enhancements

Cross Validation: We employed fivefold cross-validation to cyclically split the data into training and validation sets, ensuring robust evaluation and mitigating overfitting.

Layer Freezing: To retain previously learned features of the Geneformer model as shown in Figure 3, the last few layers of the model are kept unchanged during training. We experimented with the number of frozen and unfrozen layers based on the optimal computational resources and optimized the model's performance.

Loss Functions: We experimented with Dice, Binary Cross-Entropy (BCE), and Focal loss functions to address class imbalance as can be seen from Figure 4, we found Focal Loss particularly effective in addressing our class imbalance as it down-weighted the easy examples and focusing more on hard, misclassified examples. It is defined as:

$$L_{Focal}(p_t) = -\alpha_t \, (1 - p_t)^{\gamma} \log(p_t)$$

where:

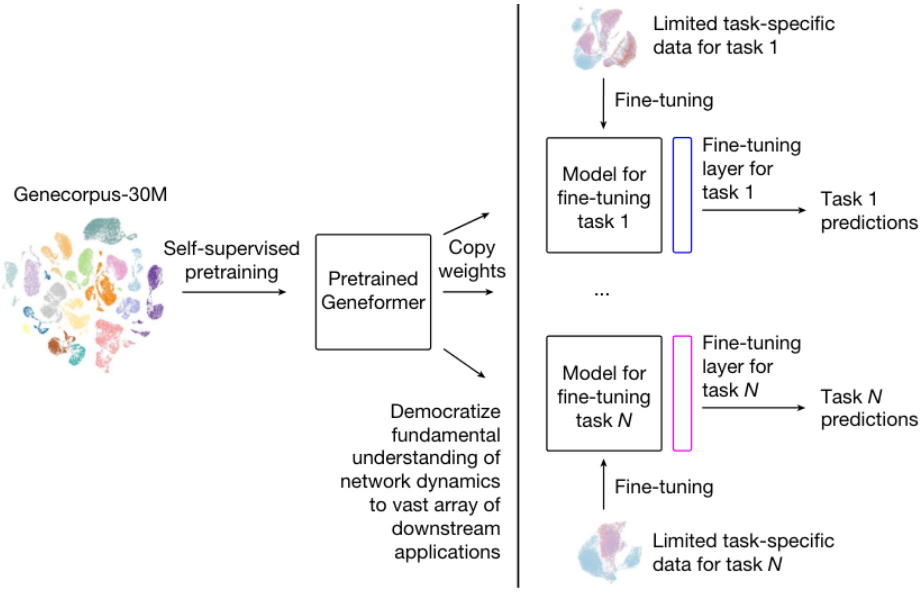
Genecorpus-30M
Self-supervised pretraining
Pretrained Geneformer
Copy weights
Democratize fundamental understanding of network dynamics to vast array of downstream applications
Limited task-specific data for task 1
Fine-tuning
Model for fine-tuning task 1
Fine-tuning layer for task 1
Task 1 predictions
...
Model for fine-tuning task *N*
Fine-tuning layer for task *N*
Task *N* predictions
Fine-tuning
Limited task-specific data for task *N*

Figure 3: Fine Tuning of Geneformer Model [5]

- p_t represents the predicted probability for the true class,
- α_t is a balancing factor to adjust the importance of different classes,
- γ is a focusing parameter that adjusts the rate at which easy examples are down-weighted compared to hard examples.

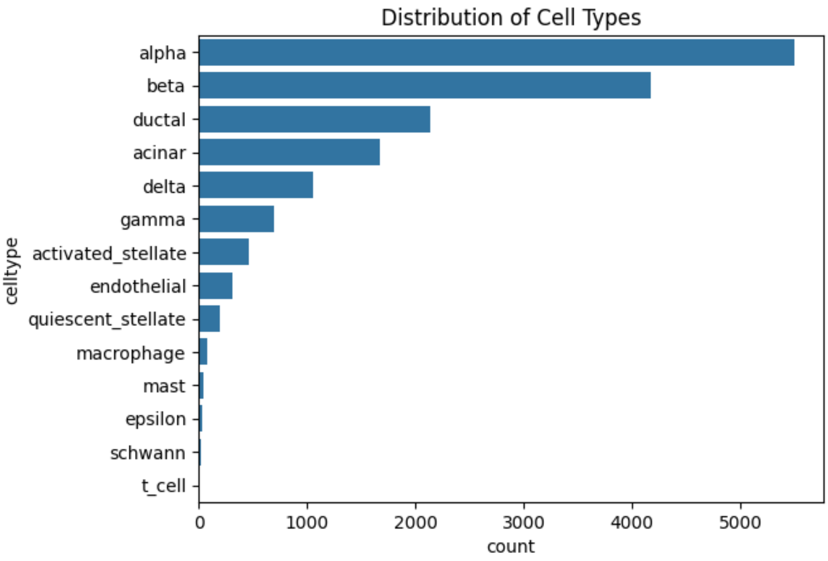

Figure 4: Distribution of cell types depicting Class Imbalance

3.4 Proposed Method

Our proposed method uses GeneFormer pretrained on the extensive Genecorpus-30M dataset. By fine-tuning GeneFormer on single-cell RNA sequencing (scRNA-seq) data from the pancreas, we aimed to classify cell types accurately and identify dosage-sensitive genes. The model's architecture, which includes masked learning and BertForSequenceClassification, enhances its ability to classify cells accurately. Layer freezing helps retain valuable pre-learned features, while cross-validation and upsampling techniques ensure robust evaluation and improved model performance. This approach effectively addressed class imbalance, making it a strong choice for cell and gene classification tasks.

3.5 Training and Testing Algorithm

Training Algorithm:

- Initialize GeneFormer with pretrained weights.
- Add a final task-specific transformer layer for fine-tuning.
- Employ hyperparameters: max learning rate (5×10^{-5}), linear learning scheduler with warmup, Adam optimizer with weight decay fix, warmup steps (500), weight decay (0.001), and batch size (12).
- Use a single training epoch to avoid overfitting.
- Apply cross-validation to cyclically split the data into training and validation sets.
- Implement upsampling to balance underrepresented classes.
- Use Dice, Binary Cross-Entropy (BCE), and Focal loss functions to address class imbalance. Exact Parameters are shown in Table 1.

Parameter	**Cell Classification**	**Gene Classification**
Cross Val Split	`5`	`5`
Batch Size	`batch_size = 64`	`batch_size = 64`
Optimizer	`optimizer = Adam`	`optimizer = Adam`
Frozen Layers	`2`	`4`

Table 1: Cell Classification and Gene Classification Parameters

Testing Algorithm:

- Evaluated the fine-tuned model using fivefold cross-validation.
- Reported results as AUCs ± standard deviation and F1 score.
- Tested the model for generalization to out-of-sample data.
- Predict dosage sensitivity of disease genes in various cell contexts.

3.6 Experiments

Different experiments we conducted to optimize our model and adapt it to resource constraints and performance requirements.

3.7 Model Optimization Experimentation

1. Freeze Layers of Pretrained Model

Initially, we froze 5 layers of the Geneformer model, which was pre-trained on a large Genecorpus (30M parameters). However, due to resource constraints and the model's rich pre-training on a vast dataset, we found that freezing only **2 layers and 4 layers** yielded satisfactory results in our classification task. This adjustment not only optimized training under limited resources but also leveraged the model's extensive pre-training effectively.

2. Loss Function Selection

We experimented with different loss functions including Binary Cross-Entropy (BCEWithLogitsLoss), Dice Loss, and Focal Loss. BCEWithLogitsLoss struggled with our imbalanced dataset, treating all errors equally and causing bias towards the majority class. Dice Loss, although useful for segmentation, did not sufficiently penalize easy-to-classify examples in our imbalanced data. Focal Loss, designed to handle class imbalance, down-weighted well-classified examples and focused on hard-to-classify samples, ensuring greater emphasis on misclassified instances. This approach improved our model's precision, recall, and F1-score.

3. Cross-Validation

We started with 1-fold cross-validation as per original paper. But to obtain a more robust evaluation, we increased this to **5-fold cross-validation**, which provided us a well-rounded assessment of our model's generalization ability.

4. Batch Size Optimization

Batch size optimization was crucial factor under our limited resources. We experimented with batch sizes ranging from 128 (initially) to 4 (which slowed down training excessively) and finally settled on **64**, which balanced training efficiency and resource constraints effectively.

3.8 Hardware-Aware Optimization

1. BERT Quantization

We explored BERT quantization to reduce model size and inference time. Quantization reduced the **model size from 438 to 181**, leading to a **43% decrease** in batch inference time. However, this came with a slight drop in accuracy, prompting us to prioritize balance between model size and accuracy for future.

2. Hardware Resource Adjustment

Initially deployed on 2 NVIDIA GeForce RTX 2080 Ti GPUs with approximately 15-20 GB memory each, we optimized to run on a reduced hardware setup with 8 GB GPU memory, meeting our computational constraints.

4 Results

4.1 Dosage Sensitivity Results

- The enhanced model achieved a accuracy of **94%** over the baseline model.
- The baseline model achieved an AUC of **0.79 ± 0.10**, while the enhanced model improved this to an AUC of **0.83 ± 0.08** as in Figure 5.

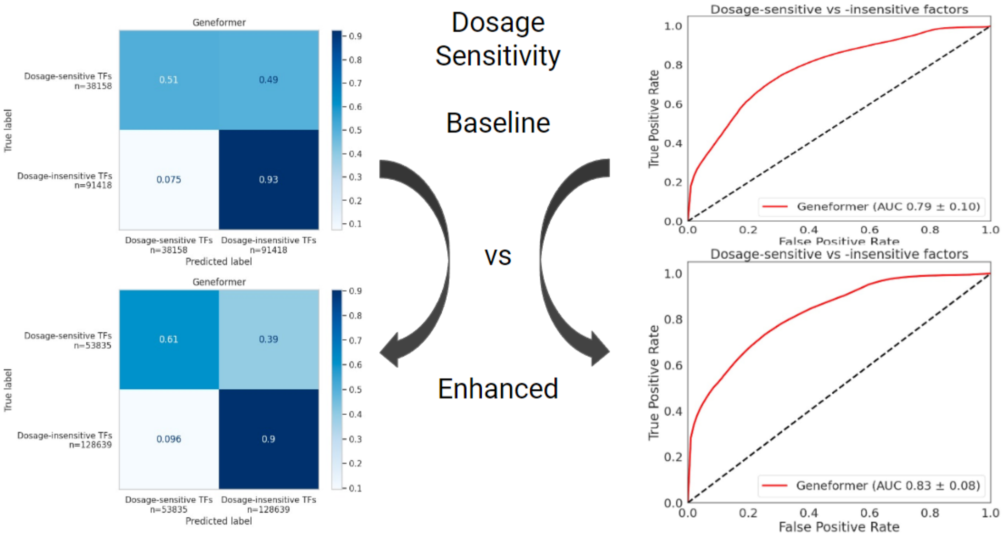

Figure 5: Dosage Sensitivity Baseline vs Enchanced - (a) Confusion Matrix (b) ROC Curve

4.2 Cell Classification Results

- The enhanced model achieved a higher accuracy of **97%** over the baseline model.
- The confusion matrices (Table 2) for both baseline and enhanced models showed improved precision and recall for various cell types as in Figure 6.

Table 2: Cell Classification Metrics

Cell Type	**Baseline Model**	**Enhanced Model**
Alpha	0.94	0.99
Beta	0.97	0.91
Delta	0.54	0.98
Gamma	0.35	0.97
Macrophage	0.70	1.00
Endothelial	0.95	0.94
Schwann	0.50	0.98
Mast	0.33	1.00
Acinar	0.89	1.00
Ductal	0.92	1.00
Activated Stellate	1.00	0.93

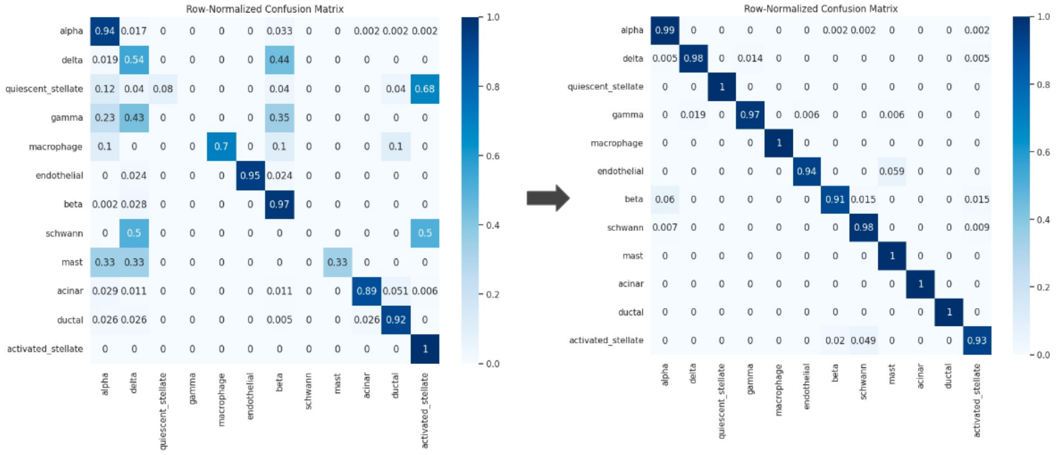

Figure 6: Cell Classification - (a) Baseline Confusion Matrix (b) Enhanced Confusion Matrix

5 Conclusion

We successfully utilized the Geneformer model for cell classification and disease-related gene expression profiling on our pancreas dataset using transfer learning. The final accuracy values for cell classification were approximately 97%, and for dosage sensitivity, it was around 94%. These findings provide valuable insights into the roles of specific cell types in various diseases and highlight potential biomarkers and therapeutic targets. For future work, we plan to use BERT quantization to better balance the trade-off between model size and accuracy. We can explore techniques like int-8 quantization and pruning to efficiently train the model with optimal resource utilization.

6 Code:

Github Link to our project: https://github.com/vaishjah3/Geneformer/tree/main

Please go through our README.md file before executing.

References

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

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

[3] Pancreas Dataset. Pancreas Dataset. figshare, 2022.

[4] Christina Theodoris and et al. Genecorpus-30M, 2022. Hugging Face.

[5] Christina V Theodoris, Ling Xiao, Abhishek Chopra, and et al. Transfer learning enables predictions in network biology. *Nature Communications*, 12(1):1–12, 2021.

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