CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

# Gene Regulatory Networks Inference using Bidirectional Encoder Representations from Transformers

## Understanding Alzheimer's Disease Pathways through Foundation Models-Geneformer

Master's thesis in Computer science and engineering

DIMITRIOS STEFANOU

Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2025

Master's thesis 2025

# Gene Regulatory Networks Inference using Bidirectional Encoder Representations from Transformers

Understanding Alzheimer's Disease Pathways through Foundation Models-Geneformer

DIMITRIOS STEFANOU

UNIVERSITY OF
GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Computer Science and Engineering
Chalmers University of Technology
University of Gothenburg
Gothenburg, Sweden 2025

Gene Regulatory Networks Inference using Bidirectional Encoder Representations from Transformers
Understanding Alzheimer's Disease Pathways through Foundation Models-Geneformer
DIMITRIOS STEFANOU



Supervisors: Annikka Polster, Department of Life Sciences
Danish Anwer, Department of Life Sciences
Examiner: Rocío Mercado Oropeza, Department of Computer Science and Engineering

Master's Thesis 2025
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Typeset in LaTeX
Gothenburg, Sweden 2025

Gene Regulatory Networks Inference using Bidirectional Encoder Representations from Transformers
Understanding Alzheimer's Disease Pathways through
Foundation Models-Geneformer
DIMITRIOS STEFANOU
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg

# Abstract

Alzheimers disease is characterised by complex molecular mechanisms that are only partially understood. This thesis leverages single-cell RNA sequencing data from the newly released ROSMAP dataset, using an equally new BERT framework, to uncover potential drivers of disease onset, progression or therapeutic targets. A pre-trained transformer model (Geneformer) is finetuned to classify major cell types and attempt patient classification. Geneformer achieves robust classification of major cell types, identifying molecular signals within a restricted gene set. but does not generalise effectively for patient classification. Performance is compared on reduced and full datasets to examine resource trade-offs. Molecular markers and candidate genes through perturbation analysis are presented through in silico perturbation. Future work may integrate updated versions of Geneformer with expanded gene inclusion and deeper architecture. This approach contributes insights into determinants of Alzheimers disease.

Keywords: Data science, machine learning, bioinformatics, transformers, genomics, project, thesis.

# Acknowledgements

I would like to express my heartfelt gratitude to the following individuals for their invaluable contributions to this work:

- Annikka Polster, for her supervision, critical review and insightful guidance, which helped shape the direction and rigour of this project.
- Danish Anwer, for his patience and support in clarifying complex programming concepts and serving as an ever-reliable sounding board during the development process.
- Iván Domenzain Del Castillo Cerecer, for his suggestions and constructive input on the assessment of in silico perturbations, which greatly enriched this study.

Dimitrios Stefanou, Gothenburg, 2025-01-21

# Contents

# List of Figures

# List of Tables

# 1

# Introduction

Alzheimer's disease (AD) is a progressive neurodegenerative condition that accounts for the majority of dementia cases worldwide. Despite its prevalence, there are currently no definitive treatments or universally accepted preventative measures. The underlying causes of Alzheimer's disease remain elusive, though ageing, characterised by time-dependent damage accumulation, is recognised as a primary risk factor, alongside potential genetic influences. This lack of clarity highlights the need for innovative approaches to unravel the complex molecular mechanisms involved in the disease.

In light of these knowledge gaps, advances in computational biology and machine learning offer promising avenues for exploration. We are aiming to better understand it through molecular determinants of cell function. A similar approach has already seen use with artificial neural networks[1], where structured gene expression data is argued to help in classification tasks, while also providing visualisation techniques to understand gene importance. Transformers can also incorporate such complex biological data, as seen in the case of Geneformer, the model we will be implementing. The Geneformer model is a cutting-edge machine learning framework designed for transfer learning tasks in network biology. Trained on extensive transcriptomic datasets, Geneformer excels at identifying intricate gene expression patterns, making it a valuable tool for probing the molecular intricacies of neurodegenerative processes like Alzheimer's disease[2].

Gene expression reflects the functional state of cells and can be quantified through sequencing technologies, enabling the analysis of molecular mechanisms underlying diseases[3]. Bulk RNA sequencing, the traditional transcriptomic approach, has inherent limitations when applied to complex, heterogeneous tissues like the brain. Bulk RNA sequencing aggregates gene expression signals across a large population of cells, often masking the diversity and biological differences present at the cellular level. This can be particularly relevant in diseases like AD, where the brain's heterogeneous tissues, the interaction of different cell types and their distinct gene expression profiles under pathological conditions might contribute to the onset and progression of the disease.

Single-cell RNA sequencing (scRNA-seq) addresses this limitation by enabling gene expression analysis at the level of individual cells, providing a more nuanced view of cellular heterogeneity. This method has revealed biologically relevant differences among cells, even within seemingly uniform populations, and has become increas-

ingly accessible with advancements in microfluidics and sequencing protocols[4]. By utilising scRNA-seq data, we can identify rare cell populations, uncover gene regulatory mechanisms, and explore cellular states with greater precision, making it a powerful tool for studying complex diseases like AD[4], [5].

This thesis leverages scRNA-seq data from the Religious Orders Study and Rush Memory and Aging Project (ROSMAP) dataset[5], integrating methodologies from data science and biological research, to fine-tune the Geneformer model for classification tasks. These tasks include identifying individuals with Alzheimer's disease and stratifying patient groups based on molecular features. Such stratification could uncover novel gene targets, paving the way for the development of therapeutic interventions.

However, analysing large-scale transcriptomic data is still challenging. The raw data often contains noise and artifacts that must be addressed during preprocessing to ensure robust downstream analysis. Accordingly, this thesis documents the ROSMAP papers's steps of data cleaning and preparation, culminating in a comprehensive processed dataset ready for analysis with the Geneformer model. Through this interdisciplinary approach, combining computational and biological insights, this study aspires to uncover new molecular insights into Alzheimer's disease, ultimately contributing to the broader understanding and potential biomarkers for treatment of AD.

# 2

# Methods

## 2.1 Background

Most applications of transformers in the context of AD or generally dementia tackle the problem through vision, comparing the results with relevant CNNs, or through speech which take advantage of inherent NLP (natural language processing) capabilities of transformers[6][7]. Our approach uniquely focuses on gene expressions for AD classification.

This project is one of two conducted in parallel to address the same objective. In addition to Geneformer, a variational autoencoder (VAE) was implemented from scratch for AD classification, achieving more than 70% accuracy when trained on a single major cell type. This accuracy served as the benchmark for Geneformer to surpass.

The project is based on gene expression data given by the ROSMAP dataset, to be processed by the Geneformer. For the analysis, a multitude of Python packages is utilised, some specialised for scRNA-seq (anndata, ScanPy, scrublet), as well as packages required by Geneformer itself (transformers, hyperopt, ray etc.).

## 2.2 Dataset

### 2.2.1 Study overview

The ROSMAP dataset includes data from 427 people and has over 2.3 million nuclei from the prefrontal cortex, providing a detailed view of the brain as it ages. It is comprised by individuals without any diagnosed dementia at the time of enrolment. Participants are subjected to an annual, comprehensive medical and psychological evaluations and have consented to the posthumous donation of their brains. Furthermore, the protocol includes the annual collection of blood samples, facilitating the preservation of serum, plasma, and cellular components. This dataset categorises people based on their Alzheimer's disease (AD) progression: 115 are in early stages (non-AD and early-AD), 198 are in intermediate stages, and 114 are in late stages. Additionally, these participants are classified based on their cognitive status at death into three groups: 146 with no cognitive impairment (NCI), 102 with mild cognitive impairment (MCI), and 144 with AD dementia.[8]

The ROSMAP dataset is comprised by specific major cell types. It is shown how neurons, glial cells, and vascular cells each respond differently to AD pathology, which includes amyloid-beta plaque formations and abnormal accumulations of tau protein. These findings not only present the complexity of AD, but also point to potential pathways for therapeutic intervention by targeting specific cell types.

Post-mortem analysis of prefrontal cortex tissue identified cellular and molecular correlates of AD pathology, cognitive decline, and resilience, using the detailed clinical and pathological data from the ROSMAP cohort. The study reported widespread gene expression changes across 54 cell types, grouped into 7 major cell classes, associated with various measures of AD pathology. These included alterations shared between excitatory neuron subtypes, a coordinated increase of the cohesin complex and DNA damage response factors in excitatory neurons and oligodendrocytes, and pathways linked to high cognitive function, dementia, and resilience to AD pathology. These findings highlight the mechanisms underlying cognitive impairment and preservation during AD progression.[8]

### 2.2.2 Ethics considerations

Informed consent was obtained from all participants, and ROSMAP received approval from the Institutional Review Board (IRB) of Rush University Medical Center. Participants also provided consent through an Anatomic Gift Act and a repository agreement, permitting their data to be used for further research[8]. The dataset used in this study is publicly available, and all metadata are anonymised, containing no names or other identifiers directly referencing the individuals involved.

### 2.2.3 Major cell types

**Astrocytes** add structural support in the central nervous system (CNS), maintain the blood-brain barrier, regulate neurotransmitter levels and influence synaptic activity. They are present in homeostasis, supporting nerve cell metabolism and response to injury[9]. Gene expression in astrocytes is associated with lipid metabolism, e.g. in cholesterol biosynthesis genes. These changes are correlated with cognitive function, suggesting astrocytes may contribute to resilience mechanisms in AD brains. Altered lipid metabolism in astrocytes supports the maintenance and integrity of nerve cell after the onset of AD.

**Microglia**, central immune cells of the CNS, respond to injuries and infections by clearing cellular debris, releasing cytokines, and modulating neuroinflammation, thus maintaining the health of nerve cells[10]. Increased microglial cell levels are observed in individuals with AD, within the ROSMAP sample. Microglia display gene expression patterns consistent with an immune response aimed at clearing amyloid-beta plaques, although this response may also contribute to inflammatory processes that accelerate the disease's progression.

**Vascular cells** appear in the blood-brain barrier and regulate the molecular flow of the brain. They maintain CNS homeostasis and support neural function through blood flow adjustments[11]. Vascular cell subtypes—endothelial cells, pericytes,

smooth muscle cells—show AD-related gene expression changes that may influence the integrity of the blood-brain barrier. Dysfunctions in vascular cells may contribute to AD pathology, generally impacting nerve cell health.

**Inhibitory neurons** mainly release gamma-aminobutyric acid (GABA) to dampen nerve cell activity. This prevents over-excitation and contributes to neural circuit balance, which is essential for cognition and motor control[12]. Two inhibitory neuron subtypes, particularly somatostatin (SST) and LAMP5 RELN groups, are linked with high cognitive resilience. The study shows that these neurons are more prevalent in the prefrontal cortex of individuals with preserved cognitive function late in life, which suggests deceleration of AD progression. In contrast, vulnerable inhibitory neuron subtypes were diminished in AD brains, aligning with increased cognitive impairment.

**Excitatory neurons** primarily release glutamate and facilitate synaptic transmission, learning, memory and sensory processing through synaptic plasticity[13]. The study links gene expression changes in excitatory neurons with AD pathology. These neurons display increased gene expression associated with DNA damage response mechanisms and structural genome alterations, such as variations in chromosomal structures and disruptions in 3D genome organisation. Furthermore, some excitatory neuron subtypes may be involved in preserving cognitive abilities in spite of AD, as evidenced by the favourable correlations found between specific synaptic genes and overall cognitive function.

**Oligodendrocytes** create myelin layers (also called sheaths) around neuron axons, which accelerates signal transmission and enhances neural connectivity[14]. These cells exhibit notable expression changes related to lipid metabolism, such as cholesterol biosynthesis, with some of these alterations also associated with AD pathology. Oligodendrocytes show an increase in cohesin complex expression as AD pathology advances, possibly as a response to DNA damage or structural alterations.

**Oligodendrocyte progenitor cell (OPCs)** are progenitor cells that generate oligodendrocytes and participate in remyelination and CNS repair, particularly after injuries or in demyelinating diseases[15]. OPCs demonstrate changes in gene expression related to lipid metabolism. These alterations suggest that OPCs may attempt to support myelin maintenance in the ageing brain, though further research is required to understand their specific role in AD resilience or susceptibility.

### 2.2.4 Count matrix and metadata

The ROSMAP study dataset is comprised of the scRNA-seq data, which includes the count matrix, along with several subject and batch-level metadata files. This sparse matrix contains columns which correspond to genes and rows to individual cells. Each entry records the raw count of transcripts for a gene within a specific cell), cell observations and gene names.

Notable features in observations, the **bc** column provides a unique barcode for each cell, distinguishing cells after sequencing. Each cell is assigned a **batch** identifier to trace its processing origin. The **n_counts** field records total gene expression

per cell, serving as a measure of transcriptional activity. Quality control is further provided by **pct_ribo** and **pct_mito**, indicating the percentage of reads derived from ribosomal RNA and mitochondrial genes, respectively. Elevated ribosomal count may signal cellular stress, while high mitochondrial levels are often a marker of stress or apoptosis.

Table 2.1: **Mitochondrial gene count distribution for AD and NCI: Training Set.**

| **Mito. Percentage Range** | **AD Counts (%)** | **NCI Counts (%)** |
|---|---|---|
| (−0.0012, 0.0333] | 793,558 (76.04%) | 729,918 (78.33%) |
| (0.0333, 0.0667] | 158,440 (15.18%) | 133,906 (14.37%) |
| (0.0667, 0.1] | 48,049 (4.60%) | 36,372 (3.90%) |
| (0.1, 0.133] | 22,882 (2.19%) | 16,254 (1.74%) |
| (0.133, 0.167] | 12,781 (1.22%) | 9,416 (1.01%) |
| (0.167, 0.2] | 7,856 (0.75%) | 5,944 (0.64%) |
| **Statistic** | **AD** | **NCI** |
| Count | 1,043,566 | 931,810 |
| Mean | 0.025674 | 0.023448 |
| Std. Dev. | 0.031109 | 0.029260 |
| Min | 0.000000 | 0.000000 |
| 25% | 0.006084 | 0.005197 |
| 50% | 0.015077 | 0.013487 |
| 75% | 0.032186 | 0.029720 |
| Max | 0.199951 | 0.199934 |
| High Mito.% (>10%) | 4.17% | 3.39% |

Table 2.2: **Mitochondrial gene count distribution for AD and NCI: Testing Set.**

| **Mito. Percentage Range** | **AD Counts (%)** | **NCI Counts (%)** |
|---|---|---|
| (−0.0012, 0.0333] | 234,493 (73.23%) | 143,368 (78.89%) |
| (0.0333, 0.0667] | 53,621 (16.75%) | 26,409 (14.53%) |
| (0.0667, 0.1] | 16,779 (5.24%) | 6,987 (3.84%) |
| (0.1, 0.133] | 8,049 (2.51%) | 2,681 (1.48%) |
| (0.133, 0.167] | 4,446 (1.39%) | 1,406 (0.77%) |
| (0.167, 0.2] | 2,806 (0.88%) | 887 (0.49%) |
| **Statistic** | **AD** | **NCI** |
| Count | 320,194 | 181,738 |
| Mean | 0.027719 | 0.022863 |
| Std. Dev. | 0.032556 | 0.027388 |
| Min | 0.000000 | 0.000000 |
| 25% | 0.006650 | 0.005657 |
| 50% | 0.016553 | 0.013719 |
| 75% | 0.035266 | 0.029227 |
| Max | 0.199957 | 0.199847 |
| High Mito.% (>10%) | 4.77% | 2.73% |

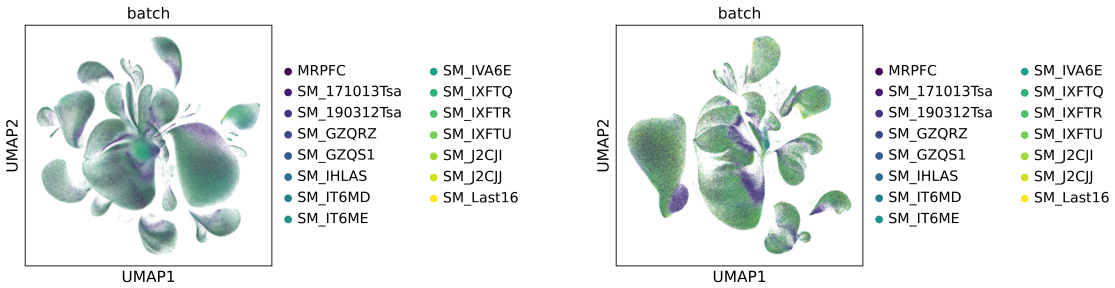
Subject-level metadata provides demographic and clinical context for the cellular observations. Each individual is assigned a unique **subject** identifier (e.g., ROSMAP-10132), which enables alignment with single-cell data. The **msex**, **age_death**, **pmi** (post-mortem interval) and **race** columns denote several personal characteristics of each individual, useful when controlling for specific subgroups that display significant genetic diversity. Last but not least, **Pathologic_diagnosis_of_AD** indicates a neuropathological diagnosis of Alzheimer's Disease (AD), serving as a label for binary classification between AD and non-AD subjects.

Batch-level metadata links batches with the individual subjects. Each batch is identified by a **batch** label to trace each cell's processing batch. The **subject** column provides a unique identifier for each individual in the study. The dataset column serves as an additional identifier to track which subjects were processed in each batch, as some subjects appear in multiple batches. This structure ensures traceability and allows for joining and appending the metadata of each subject to their respective cells.

Finally, the study includes an additional metadata file that assigns a cell type label to each cell observation. These labels were derived from the study's analytical results, after preprocessing, and are not available for every observation in the raw dataset. As we do not replicate the study's cell-type identification process, we incorporate these labels following our own preprocessing steps.

## 2.3 Data exploration and preprocessing

The raw dataset contains $\sim 84GB$ worth of trascriptomic data, with a gene expression count of **18934 genes** over **2580107 cells**. Since our aim was to fine-tune a machine learning model, we first split the dataset into training and testing parts (roughly **80-20 split**). Each split was preprocessed independently, as a number of bioinformatics-related preprocessing algorithms depend on the interconnections and relative distance between data points. To ensure an unbiased and stratified split, we sampled from every batch and kept patients that might be present in multiple batches in one set.



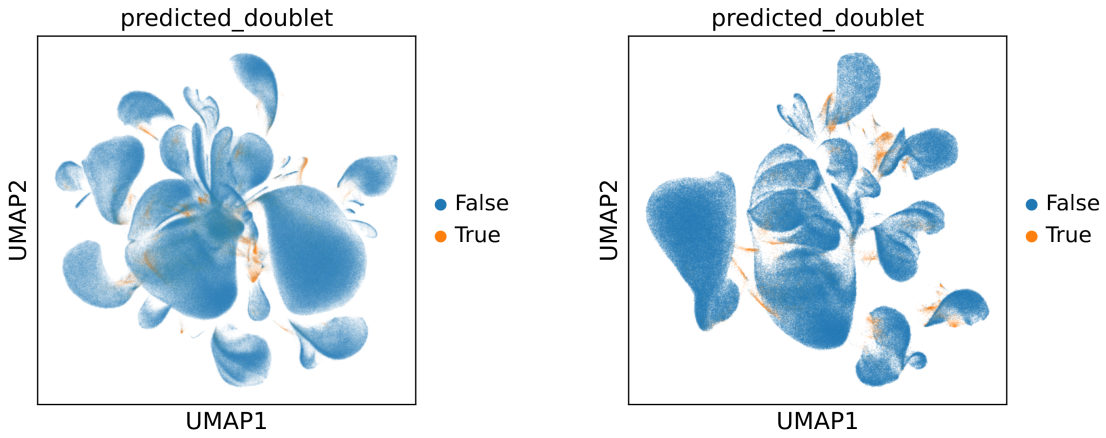

(a) Training Dataset (b) Testing Dataset

Figure 2.1: **UMAP of batches.** Although batch correction is usually employed when handling genomics data, it is not applied in our case study. The large number of batches in a low-dimensional space makes detection of any batch effect more difficult, and it can effectively be bypassed by Geneformer's rank value encoding.

Before feeding data to Geneformer's tokenisation function, we preprocessed the dataset in accordance to the original paper's pipeline. We began the preprocessing of the two sets by mapping gene names in .var to their respective **Ensembl IDs**, necessary for the Geneformer model. A total of 751 genes without a matching Ensembl ID were dropped. We additionally added metadata in .obs, found in separate .csv files.

UMAP embeddings were computed throughout the cleaning process to reduce dimensions and to visualise the data in 2-dimensional space, providing a clean graphical representation that highlights the relationships between indicated clusters.

Doublet detection in the dataset was performed using Scrublet. Unlike the procedure described in the original study which processed the entire dataset collectively, our approach involved scrubbing doublets by batch to bypass memory constraints. This could potentially be caused by package version updates rather than computational limitations. Batch scrubbing is an option available through a parameter of the function, thus being a sound alternative and not affecting the end result significantly, if at all. After doublet removal, we verified that cluster composition remained consistent by comparing cluster assignments before and after the filtering step.




(a) Training Dataset

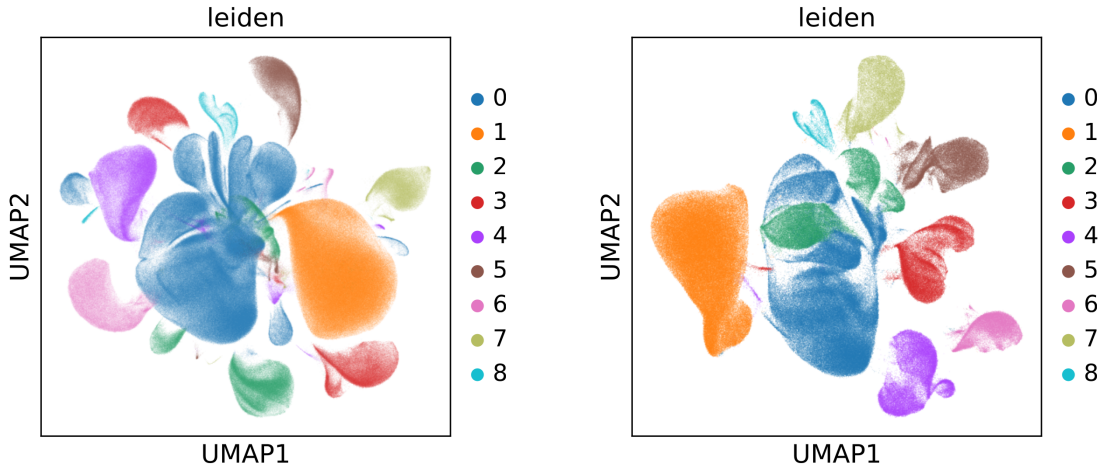
(b) Testing Dataset

Figure 2.2: **UMAP of predicted doublets.** Most clusters of scrubbed cells appear to be distinct separations from their closest large neighbors in both training and testing datasets.

Subsequently, we applied standard quality control measures to remove low-quality cells, often marked by either an unusually low or high number of detected genes indicative of dying cells or technical artifacts. Normalisation was performed to ensure uniform gene expression data across all cells, followed by a log transformation to stabilise variance. The next step involved identifying and removing highly variable genes to maintain focus on stable expression patterns. Afterwards, unwanted variations were regressed out, using simple linear regression. We concluded with the standardisation, adjusting the data for dimensionality reduction and plotting.

Principal component analysis (PCA) was used to reduce the dimensionality of the dataset, extracting key features that capture the most variance among the gene expression data. Following PCA, k-nearest neighbours were calculated for k=100, to measure the proximity between cells based on their principal component scores, which facilitated the detection of local cell groups or similar gene expressions.

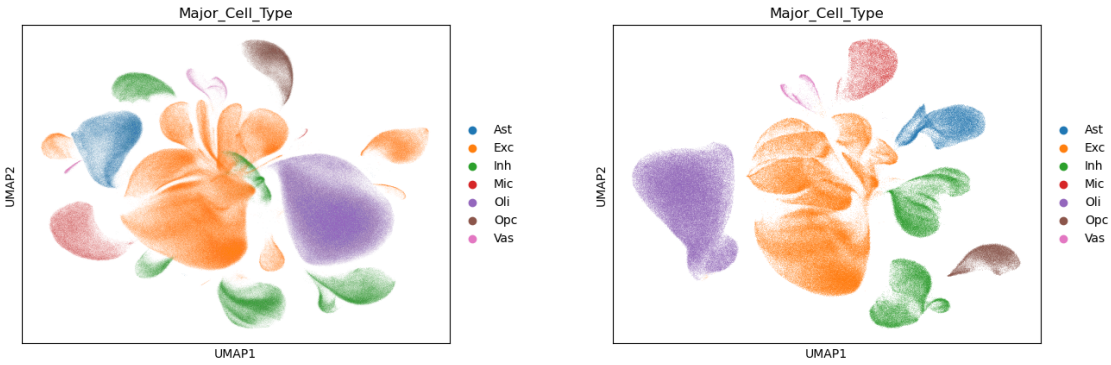
We then applied an implementation of the Leiden algorithm—an upgraded form of the Louvain algorithm, used for community detection—to identify clusters within the data. The Leiden algorithm partitions the data into highly connected networks of cells, identifying groups that share similar expression patterns. In the original paper, the first 30 possible Leiden groups are retained, although that upper limit is not reached our processes. The most prominent clusters that likely represent distinct cellular states or types.




(a) Training Dataset (b) Testing Dataset

Figure 2.3: **UMAP of leiden clusters.**

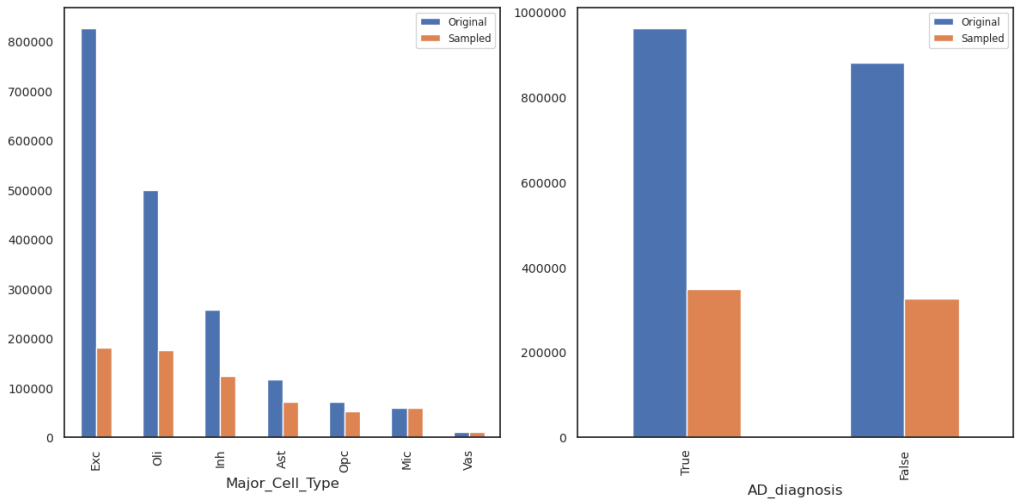
We retained only the data present in ROSMAP study's processed result. This allowed for vital metadata to be carried over, including the major cell type class of each cell data point, which would act as our labels for classification. Attempting cell type identification from scratch would be exceptionally difficult and outside the scope of the project, as we were not aiming to replicate the entire ROSMAP paper. The final preprocessed training and testing count matrices have sizes 1841718 Œ 18183 and 475758 Œ 18183 respectively.



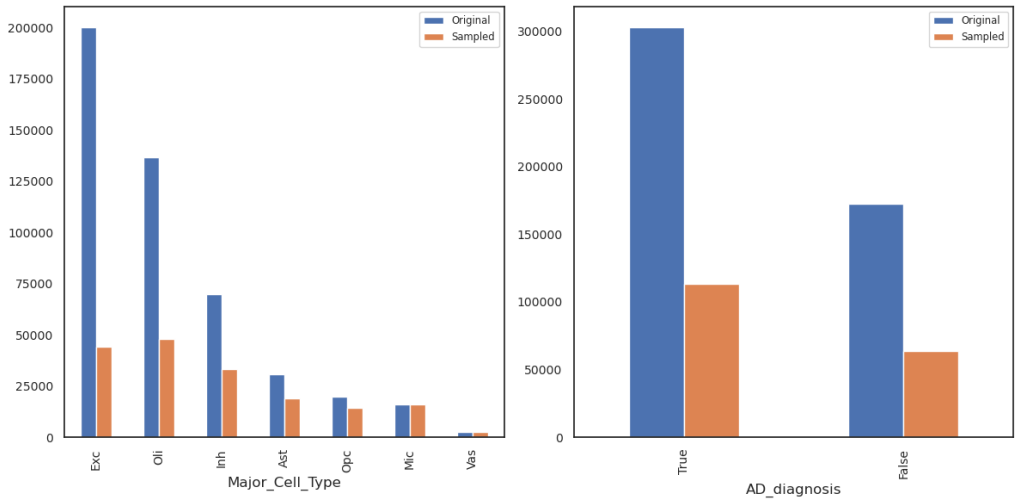

(a) Training Dataset (b) Testing Dataset

Figure 2.4: **UMAP of the final clean clusters, coloured by each major cell type.**

The last UMAP is arguably the most important one; we can verify that replicating the cleaning pipeline has succeeded, as the main clusters are homogeneous. Future plotting will incorporate a broader variety in plot types, but UMAPs will continue to be the most frequently utilised.




(a) Training Dataset




(b) Testing Dataset

Figure 2.5: **Undersampling performed.**

Finally, to address computational limitations, we performed undersampling on both datasets. Although Geneformer has integrated undersampling functions, these are superseded by the tokenisation method, which does not incorporate undersampling and utilises CPU resources. Consequently, the tokenisation process becomes excessively slow due to its computational demands. Moreover, we seek to determine whether fine-tuning on small amounts of data would still yield positive results.

## 2.4 Geneformer

With the dataset established and a processed sample obtained, we shift our attention to the theoretical background underlying of Geneformer.

### 2.4.1 Transformers

The Transformer model, as introduced in *Attention is All You Need*[16], utilises a self-attention mechanism that computes dependencies between input elements in parallel. The core of this model is the scaled dot-product attention, defined mathematically as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where $Q$, $K$, and $V$ represent query, key, and value matrices, respectively, and $d_k$ is the dimension of the keys. The scaling factor mitigates issues of vanishing gradients and ensures the values input to the softmax function are in a manageable range. This mechanism allows each position in the sequence to attend to every other position, which that eliminates the need for recurrent computations traditionally used in sequence models.

Transformers further take advantage of their multi-head attention to extend the model's ability to focus on information from different representation subspaces. The multi-head attention mechanism splits $Q$, $K$, and $V$ into multiple heads, computes the attention function for each, and concatenates the results:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O$$

where each head is calculated as $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$, with learned projection matrices $W_i^Q$, $W_i^K$, $W_i^V$, and $W^O$. This multi-head approach enables the model to capture complex patterns by attending to information from multiple perspectives.

Position encoding is added to the input embeddings to preserve sequence order, as the attention mechanism itself does not inherently encode positional information. This encoding is applied by adding a sinus function to each input embedding position, allowing the model to distinguish the sequence order in the input.

Transformers achieve high parallelization and reduce training times compared to previous methods, which has proven effective across various tasks in natural language processing, image classification and generation, time series analyses, as well as biomedical tasks, such as our current endeavour.

### 2.4.2 BERT

Building upon the Transformer's architecture, BERT (Bidirectional Encoder Representations from Transformers)[17] extends the attention mechanism through bidirectional training. BERT is trained for context cue recognition of a word relative to both its preceding and succeeding words by masking certain tokens in the input sequence. The model is thus guided to predict these masked tokens based on their surrounding context. The masked language modelling is central to BERT's

pre-training, as it enables the model to capture bidirectional dependencies for downstream tasks.

The training objective for masked language modelling (MLM) is for BERT to predict the original values of certain tokens that are randomly masked in the input. Specifically, a portion (typically 15%) of the tokens in each input sequence is selected. Among them, 80% are replaced with a special [MASK] token, 10% are replaced with a random token, with the remaining 10% left intact. The loss is computed only over masked tokens to prevent the model from overfitting on unmasked tokens.[17].

BERT also introduces a special [CLS] token at the beginning of every input sequence. This [CLS] token acts as a sequence-level representation; its embedding from the final layer serves as a summary of the entire input, making it suitable for tasks that require comprehensive understanding. Formally, given an input sequence $\{x_{[\text{CLS}]}, x_1, x_2, \ldots, x_n\}$, BERT computes embeddings for all tokens through successive transformer layers. At each layer $l$, the embeddings are updated collectively:

$$h^{(l)} = \text{TransformerLayer}\left(h^{(l-1)}\right)$$

where $h^{(l-1)}$ represents the embeddings from the previous layer. The TransformerLayer has self-attention mechanisms, allowing each token's embedding to integrate information from the entire sequence. After processing through all layers, the embedding of the [CLS] token, $h^{(L)}_{[\text{CLS}]}$ where $L$ is the final layer, is used as the sequence-level representation. The [CLS] embedding can then be input of a softmax layer for classification or other output functions, leveraging the bidirectional capabilities of the model.

### 2.4.3 Geneformer

Geneformer is a pretrained BERT-like model designed for network biology tasks. It leverages self-supervised learning and transfer learning techniques, pre-trained on an extensive dataset of about 30 million single-cell transcriptomes. The model aims to enhance the predictive accuracy for a range of downstream tasks, especially in the context of limited data scenarios.

Geneformer's architecture and training procedures as shown in the paper are present in Figure 2.6. The authors detail how the model encodes biological network dynamics and hierarchical structures through its training process. The approach combines the power of large-scale transcriptomic data with advanced machine learning techniques. After fine-tuning, the model can distinguish dosage-sensitive transcription factors, while being context-aware.

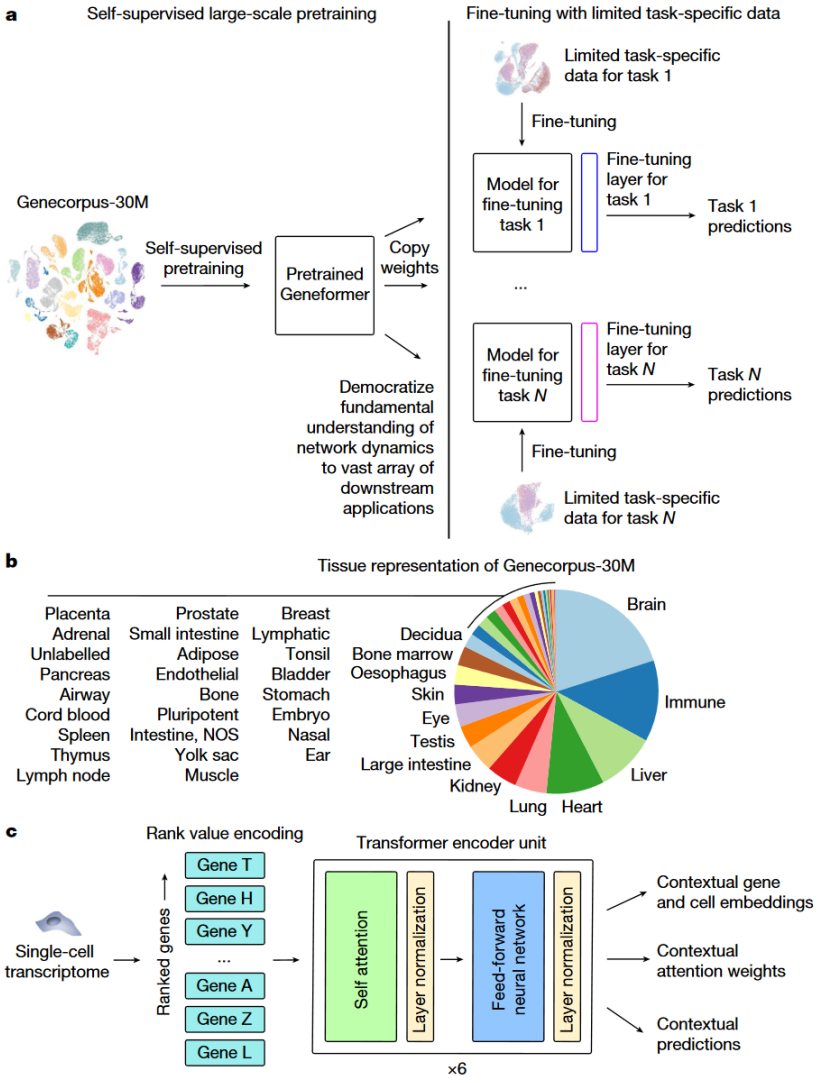



Figure 2.6: **Geneformer architecture and transfer learning strategy, taken from article.[2] a.** Transfer learning strategy involves: initial self-supervised pretraining, transferring weights to task-specific models, adding fine-tuning layers, and fine-tuning with task-specific data, enabling the model to apply fundamental domain knowledge to diverse downstream tasks. **b.** Tissue representation of Genecorpus-30M. **c.** Geneformer architecture: encodes single-cell transcriptomes into rank values, processes through six transformer layers (input size 2048, 256 embedding dimensions, six attention heads, feed-forward size 512), uses dense self-attention, and produces contextual embeddings, attention weights, and predictions.

Geneformer employs a transformer-based architecture tailored for transcriptomic data. Each input is a rank value encoding of 2048 genes. The model comprises six transformer encoder layers, each containing a self-attention mechanism with four attention heads, layer normalization, and a feed-forward network of size 512. The embeddings have 256 dimensions, and the model applies dense self-attention across all input features. Pretraining is performed using a masked language modelling objective, with 15% of genes masked—similar to the original BERT—and the model

trained to predict the masked genes. Pretraining uses categorical cross-entropy as loss function, applied to the predicted probabilities of the masked genes. Fine-tuning involves transferring pretrained weights, adding a task-specific output layer, and using cross-entropy loss for measuring performance(Figure 2.6). Outputs can be contextual embeddings, attention weights, and predictions.

While Geneformer adapts BERT's architecture for modelling single-cell transcriptomic data, the [CLS] token is replaced by mean pooling of gene embeddings. This change addresses the limitation of using a single token to encapsulate complex, context-dependent information present in an entire transcriptome. Within this context, determining how a cell's embedding shifts following a perturbation requires examining the embeddings of all genes except the perturbed one, ensuring that the change in context is accurately captured. Consequently, relying on a single CLS token would not suffice, as it would merge all genes into an indistinguishable representation. Geneformer computes the cell embedding by averaging all gene token embeddings:

$$h_{\text{cell}} = \frac{1}{n_{\text{genes}}} \sum_{i=1}^{n_{\text{genes}}} h_i$$

This mean-pooling approach maintains the contribution of each gene in representing the overall cell state. Instead of condensing information into a single token, Geneformer enables modifications to the cell state by adjusting the embeddings of individual genes. By utilising distributed representations across all input genes, the model effectively captures complex gene expression patterns essential for transcriptomic data analysis and in silico perturbations[18].

The model demonstrates exceptional predictive capabilities in various biomedical applications. The focus of the original project is on cardiomyopathy, where Geneformer's predictions have potentially uncovered new insight into disease mechanisms. Multiple comparative analyses are presented, highlighting Geneformer's superior performance against existing methods (XGBoost, SVM, RF, LR, non-pretrained deep learning models with similar architecture) in tasks such as gene dosage sensitivity prediction, chromatin dynamics, and network dynamics prediction. These comparisons showcase the advantages of Geneformer's methodology in capturing complex biological phenomena. Geneformer can predict bivalent chromatin genes solely by fine-tuning on a small amount of labelled transcriptional data. It also understands network dynamics and hierarchy without perturbation data, with detailed case studies.

The authors finally analyse in silico gene networks and treatment through the Geneformer findings. They argue that in silico perturbation can be applied to model gene network connections, and the possibility of disease modelling and identification of candidate therapeutic targets.

### 2.4.4 Tokenisation and rank value encoding

The rank value encoding algorithm transforms single cell transcriptome data into a tokenised sequence suitable for a transformer. For each cell, raw gene expression counts are normalised by the total read counts in that cell and scaled to a target sum (e.g. 10,000) to account for varying sequencing depths.

Each gene's normalised expression is further adjusted by dividing by the gene's non-zero median expression across a large corpus (Genecorpus-30M), representing typical expression levels across diverse cell types and conditions. Only protein-coding and miRNA genes present in a predefined gene dictionary are retained. After ranking, only the **top 2048 genes** are kept for each cell to match the model's input size constraints.

Within each cell, genes are ranked in descending order based on their normalised expression values. This ranking prioritises genes with higher expression relative to their typical levels across the corpus. The ranked genes are mapped to unique token identifiers using the gene dictionary, resulting in a sequence that represents the cell's transcriptome in a format suitable for transformer models.

This algorithm emphasises genes that have elevated expression in a specific cell compared to their median expression across Genecorpus-30M, thereby highlighting genes which define a cell's identity or state. Ubiquitously expressed housekeeping genes are deprioritised because their relative expression is not elevated compared to the corpus median. In contrast, genes that are lowly expressed on average, but show higher expression in specific cell types, receive higher priority in the ranking.

An important aspect of this approach is its robustness against batch effects; non-biological variations arising from differences in experimental conditions or sample handling. Addressing batch effects is usually addressed during data preprocessing, because they introduce noise that can lead to misleading interpretations if not properly accounted for. Traditional preprocessing steps involve batch effect correction, but Geneformer's rank value encoding algorithm inherently mitigates these effects.

By assigning ranks to genes within each cell based on relative expression, the algorithm focuses on the order of gene expression rather than absolute abundance, which is susceptible to technical variability across sequencing platforms and preservation methods. This encoding strategy emphasises genes critical for defining cell states, such as transcription factors that may have low absolute expression but significant influence on cell phenotype.

## In silico perturbation

In silico perturbation is a computational method used with a pretrained or finetuned model to simulate gene perturbations, through the deletion or overexpression of genes. Geneformer includes this technique to study the effects of perturbations on gene networks and cell states by observing changes in gene and cell embeddingsvector representations. This approach can identify potential biomarkers which are critical in discerning cell states or types.[2]

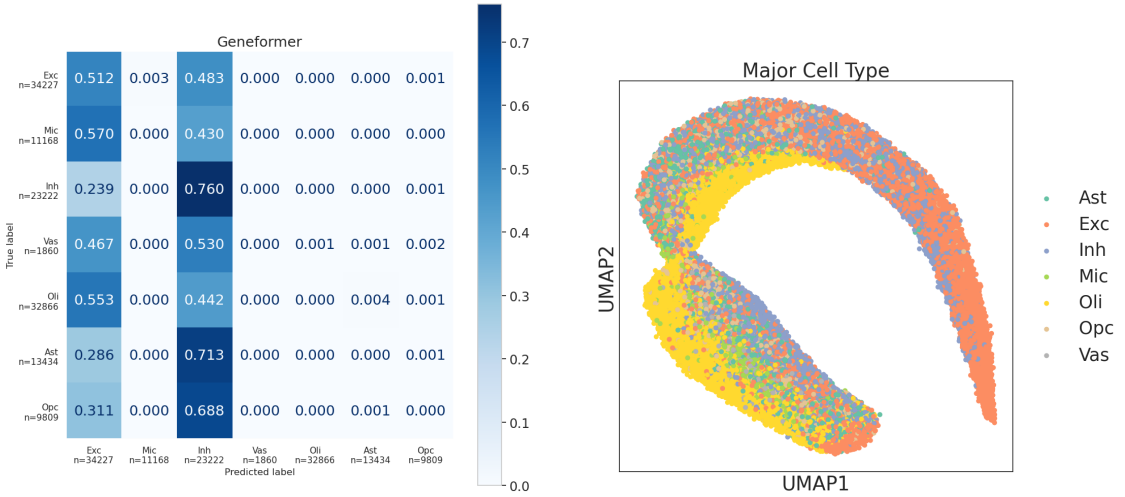
Geneformer's algorithm operates in two main phases; in silico perturbation and in silico analysis. During in silico perturbation, each gene is represented by a unique token attributed by rank value encoding, indicating its relative expression level. To delete genes, the algorithm removes gene tokens from the encoding, imitating a loss of cell function scenario. Overexpression is simulated by elevating tokens up in rank value encoding, increasing their influence in the model. The algorithm can perturb individual or batches of genes, supporting multiple specific scenarios of gene interactions.

During in silico analysis, the perturbed gene expression data is given to Geneformer, which processes it to generate updated gene or cell embeddings. The algorithm compares the new embeddings with the original, using cosine similarity to quantify the differences. Changes in gene embeddings show how the perturbation affects interactions of specific genes in embedded space, while shifts can display the perturbation influences on the overall cell states. The algorithm aggregates the cosine similarity scores across all perturbations for statistical analysis. The significance of the observed changes are determined through hypothesis testing (Wilcoxon test), and adjustments are applied to control the false discovery rate.

Simulating disease-associated perturbations allowed the authors to study pathogenic mechanisms and identify genes whose perturbation shifts cells from a diseased state towards a healthy state, making them therapeutic target candidates[2]. In our case study, we will attempt to differentiate between major cell types and potentially pinpoint AD biomarkers.

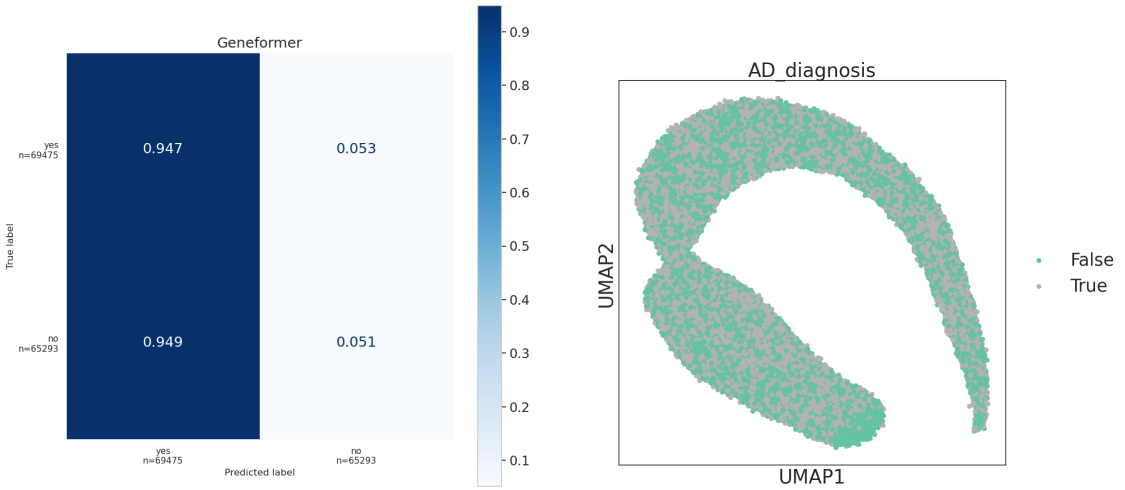
## Baselines

The pretrained Geneformer cannot work without fine-tuning. The confusion matrices show that the model does not differentiate between cell types (Figure 2.7) or disease states (Figure 2.8). Instead, it appears to randomly pick a label, usually favouring those with higher counts, and assign it to all predicted labels. The distribution across predicted classes lacks any meaningful pattern, indicating the model is effectively guessing rather than recognising distinct cell or disease characteristics.




(a) Confusion Matrix (b) UMAP of embeddings

Figure 2.7: **Pretrained Geneformer cell prediction plots.** No discernable cell types can be predicted, as they are too close in embedding space.




(a) Confusion Matrix (b) UMAP of embeddings

Figure 2.8: **Pretrained Geneformer AD prediction plots.** Similarly to cell classification, disease classification is not possible.

# 3

# Results

## Hyperparameter tuning

Geneformer uses ray and hyperopt for hyperparameter tuning. The important hyperparameters are the following:

- **num_train_epochs**: The total number of times the model will iterate over the entire training dataset. Original value by the authors was 1 epoch, but was changed to 50.
- **learning_rate**: The initial step size used by the optimiser to update the model weights during training. Sampled from a log-uniform distribution between $1 \times 10^{-6}$ and $1 \times 10^{-3}$.
- **weight_decay**: Also known as L2 regularisation, it adds a penalty proportional to the squared magnitude of the weights to the loss function, helping to prevent overfitting. It is sampled uniformly between 0.0 and 0.3.
- **lr_scheduler_type**: Specifies the strategy for adjusting the learning rate during training to improve convergence. The possible choices given are `"polynomial"`, `"cosine"`, or `"linear"`. The learning rate is adjusted after every optimisation step, which typically corresponds to every batch.
  - **Linear Scheduler**: Decreases learning rate linearly from its initial value to zero during the training process.
  - **Cosine Scheduler:** Adjusts the learning rate following a cosine function between the initial value and zero. This approach provides gradual decreases with restarts, helping the optimiser to escape local minima.
  - **Polynomial Scheduler:** Decreases the learning rate with a polynomial decay function. Provides flexibility in how quickly the learning rate decreases based on the polynomial degree.
- **warmup_steps**: The number of training steps during which the learning rate increases linearly from zero to the initial learning rate, which can help stabilise training early on. It is sampled uniformly between 100 and 2000 steps.
- **seed**: A value used to initialise the random number generator for reproducibility of the training process. It is sampled uniformly between 0 and 100.

Early stopping has been appended in the code with lenient parameters to encourage training, but avoid excessive overfit. It checks the evaluation loss for standard deviation of 0.005 and a grace period of 40 epochs.

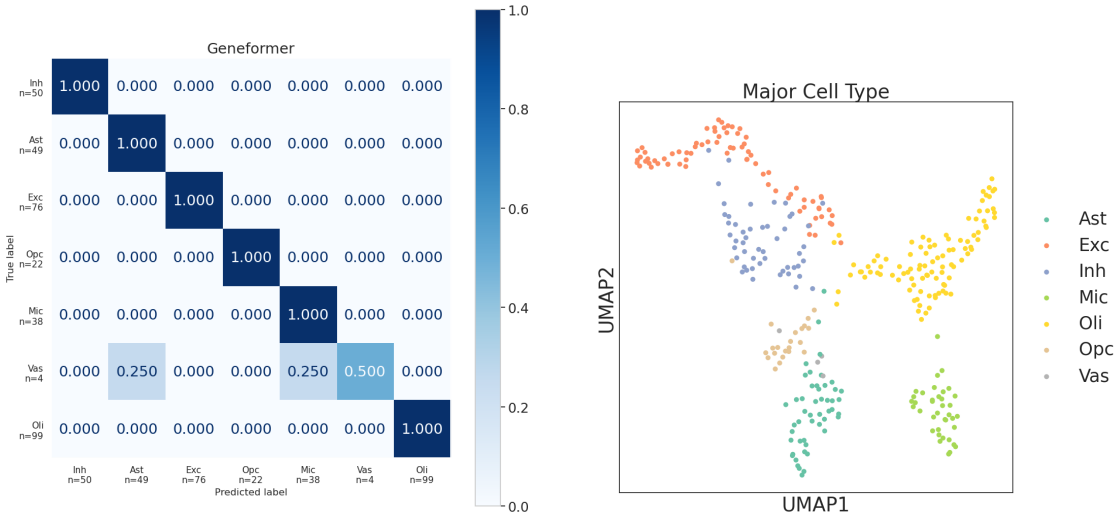
## Major cell type classification

Geneformer demonstrates robust performance in major cell type classification, achieving consistently high scores in metrics during finetuning. The model generalises effectively across multiple hyperparameter configurations (Table 3.1), even with a relatively small dataset. The majority of metric loss can be attributed to mislabelled vascular cells, which represent a particularly small set of the dataset, as shown previously in Figure 2.5.

| Learning Rate | LR Scheduler Type | Seed | Warmup Steps | Weight Decay | Loss | Accuracy | Macro F1 |
|---|---|---|---|---|---|---|---|
| 0.000115449 | linear | 94.4258 | 1196.01 | 0.0409301 | 0.010861 | 0.994083 | 0.994577 |
| 0.0000691624 | linear | 93.0338 | 542.778 | 0.0268241 | 0.0137652 | 0.997041 | 0.997623 |
| 0.0000782912 | polynomial | 7.7118 | 1796.35 | 0.138426 | 0.021211 | 0.991124 | 0.994991 |
| 0.000709494 | polynomial | 70.0329 | 1144.02 | 0.188615 | 0.0354514 | 0.991124 | 0.973412 |
| 0.000151046 | polynomial | 39.2859 | 1790.94 | 0.0661244 | 0.040937 | 0.991124 | 0.956753 |
| 0.000781482 | cosine | 33.2222 | 544.988 | 0.29709 | 0.0422633 | 0.991124 | 0.947419 |
| 0.000174829 | polynomial | 85.0009 | 552.423 | 0.103084 | 0.0430641 | 0.988166 | 0.988732 |
| 0.000193809 | polynomial | 51.0178 | 1770.87 | 0.0496014 | 0.0458824 | 0.988166 | 0.945487 |
| 0.0000836017 | cosine | 55.7845 | 773.235 | 0.240202 | 0.0498496 | 0.982249 | 0.927865 |
| 0.0000418065 | cosine | 87.4294 | 390.52 | 0.293528 | 0.0738606 | 0.967456 | 0.83142 |

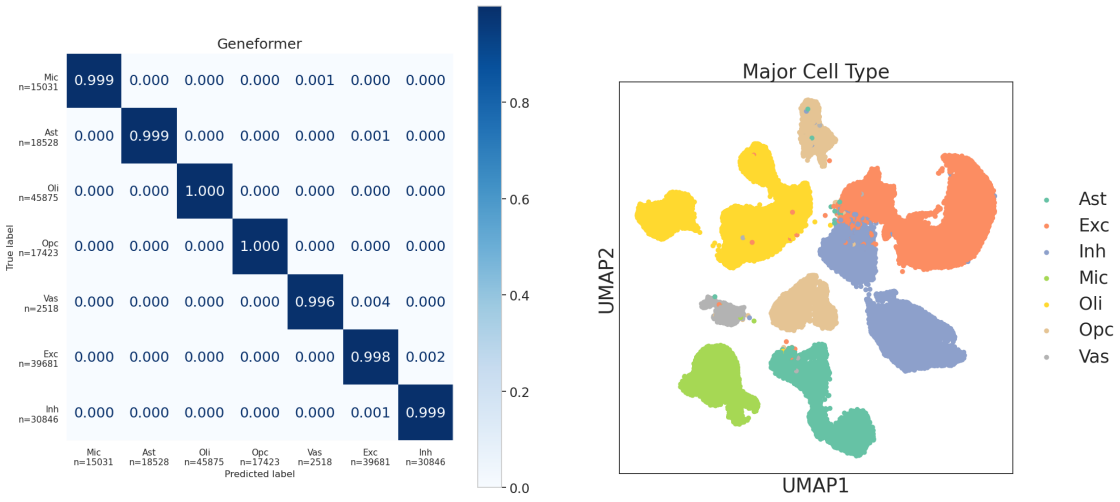
Table 3.1: **Trial results when finetuning with 2000 cells:** Consistently high accuracy and macro F1 scores across diverse hyperparameters (learning rates, scheduler types, seeds). Results indicate robust performance and effective generalization without overfitting.

This issue is mitigated when training on the whole training dataset, as the increased representation of vascular cells separate them from other cell types. Improved data balance ensures more robust classification across all cell types, and an equally robust model for in silico analysis or downstream tasks.




(a) Confusion Matrix (b) UMAP of embeddings

Figure 3.1: **Finetuned Geneformer prediction plots, 2000 cells.** Discernable cell types, barring vascular cells.




(a) Confusion Matrix (b) UMAP of embeddings

Figure 3.2: **Finetuned Geneformer prediction plots, 800k cells.** Fully discernable cell types, with few outliers.

After successfully finetuning Geneformer on major cell type classification, we use in silico perturbation to identify significant genes. Due to the computational cost of the algorithm, a limited subset of 1500 genes was randomly selected. The process was repeated multiple times, with each major cell type serving as a starting state and excitatory neurons as the target goal. Cell types that do not serve as starting states are designated as alternate states instead. A distinct separation is observed in the expression levels of significant genes across each major cell type (Figure 3.3).

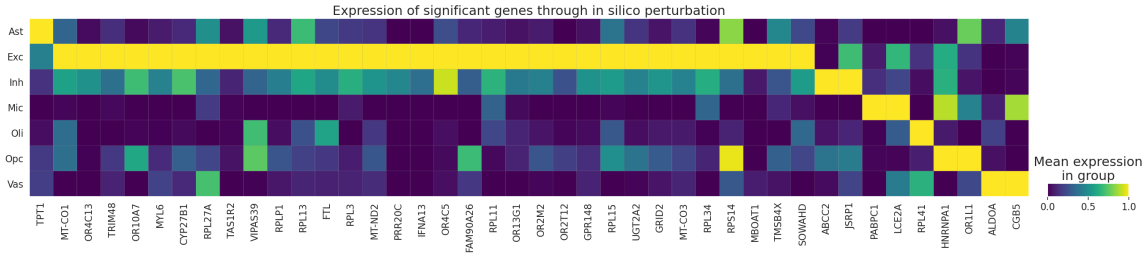



Figure 3.3: **Gene expression matrix of in silico perturbation results, with alt states:** Aggregated top 7 significant cell types for each perturbation. Duplicates found between runs were not allowed for variety.

| Major Cell Type | Gene | Gene name | Ensembl ID | Shift to goal end | Goal end vs random pval | Goal end FDR | N Detections | Cell Types with Top Gene |
|---|---|---|---|---|---|---|---|---|
| Ast | 17905 | TMSB4X | ENSG00000205542 | -0.015724 | 1.055288e-95 | 1.446166e-92 | 158 | [Ast, Mic, OPC, Oli, Vas] |
| Ast | 21114 | OR4C13 | ENSG00000258817 | -0.011607 | 5.880756e-125 | 1.343165e-121 | 264 | [Ast, OPC] |
| Ast | 9454 | GRID2 | ENSG00000152208 | -0.011116 | 1.454634e-32 | 6.040697e-30 | 59 | [Ast, Mic, OPC, Oli] |
| Ast | 3664 | RPL34 | ENSG00000109475 | 0.010325 | 1.851745e-07 | 1.256253e-05 | 12 | [Ast, Inh] |
| Ast | 9220 | ALDOA | ENSG00000149925 | 0.010242 | 3.827119e-44 | 2.280297e-41 | 96 | [Ast, Vas] |
| Ast | 2257 | RPL3 | ENSG00000100316 | -0.009952 | 1.676366e-04 | 5.860439e-03 | 6 | [Ast, Vas] |
| Ast | 14487 | OR10A7 | ENSG00000179919 | -0.008947 | 1.899938e-06 | 1.017060e-04 | 21 | [Ast] |
| Inh | 12019 | JSRP1 | ENSG00000167476 | 0.013816 | 0.000000e+00 | 0.000000e+00 | 1454 | [Inh] |
| Inh | 1720 | FTL | ENSG00000087086 | 0.010523 | 1.084390e-22 | 3.417846e-20 | 47 | [Inh] |
| Inh | 13620 | RPL15 | ENSG00000174748 | 0.010290 | 9.916486e-26 | 3.446106e-23 | 49 | [Inh] |
| Inh | 1984 | MYL6 | ENSG00000092841 | -0.010115 | 0.000000e+00 | 0.000000e+00 | 1127 | [Inh, OPC, Oli, Vas] |
| Inh | 436 | ABCC2 | ENSG00000023839 | 0.009840 | 3.118477e-188 | 1.056618e-184 | 785 | [Inh] |
| Inh | 11319 | RPS14 | ENSG00000164587 | -0.009617 | 1.731331e-03 | 3.872067e-02 | 5 | [Inh, OPC, Oli] |
| Inh | 8231 | RPL11 | ENSG00000142676 | -0.008760 | 3.675325e-20 | 8.588220e-18 | 63 | [Mic] |
| Mic | 16342 | CGB5 | ENSG00000189052 | 0.016401 | 1.024397e-03 | 1.794284e-02 | 4 | [Mic] |
| Mic | 7502 | RPLP1 | ENSG00000137818 | -0.012448 | 1.103683e-03 | 1.896404e-02 | 4 | [Mic] |
| Mic | 3831 | CYP27B1 | ENSG00000111012 | 0.011365 | 3.540695e-03 | 4.795325e-02 | 4 | [Mic] |
| Mic | 17228 | MT-ND2 | ENSG00000198763 | -0.011000 | 4.027665e-51 | 2.374046e-48 | 101 | [Mic] |
| Mic | 19843 | PRR20C | ENSG00000229665 | -0.009984 | 7.121710e-157 | 1.379272e-153 | 543 | NaN |
| Mic | 13445 | OR1L1 | ENSG00000173679 | -0.009296 | 3.343215e-04 | 7.286812e-03 | 7 | NaN |
| Mic | 17326 | MT-CO3 | ENSG00000198938 | -0.009086 | 2.392164e-04 | 5.543688e-03 | 8 | [Oli, Vas] |
| OPC | 9355 | VIPAS39 | ENSG00000151445 | -0.011143 | 0.000000e+00 | 0.000000e+00 | 1497 | [OPC] |
| OPC | 7055 | HNRNPA1 | ENSG00000135486 | 0.007140 | 0.000000e+00 | 0.000000e+00 | 912 | [OPC] |
| OPC | 23863 | UGT2A2 | ENSG00000271271 | -0.007094 | 9.180796e-04 | 2.153081e-02 | 13 | NaN |
| OPC | 19992 | IFNA13 | ENSG00000233816 | -0.007052 | 2.888205e-11 | 3.104936e-09 | 37 | NaN |
| OPC | 9238 | TRIM48 | ENSG00000150244 | -0.006779 | 2.121278e-15 | 3.393539e-13 | 58 | NaN |
| OPC | 19823 | RPL41 | ENSG00000229117 | 0.006696 | 1.418512e-39 | 7.942486e-37 | 92 | NaN |
| OPC | 13133 | MBOAT1 | ENSG00000172197 | -0.006417 | 8.504789e-08 | 6.144481e-06 | 36 | NaN |
| Oli | 12030 | RPL13 | ENSG00000167526 | -0.013283 | 1.183167e-04 | 3.271526e-03 | 6 | [Oli] |
| Oli | 17247 | MT-CO1 | ENSG00000198804 | -0.011452 | 2.692206e-06 | 1.106034e-04 | 11 | [Oli] |
| Oli | 14346 | TAS1R2 | ENSG00000179002 | -0.011033 | 2.514180e-03 | 4.248795e-02 | 7 | NaN |
| Oli | 15952 | LCE2A | ENSG00000187173 | -0.010245 | 5.702476e-04 | 1.264458e-02 | 6 | [Vas] |
| Oli | 1129 | PABPC1 | ENSG00000070756 | -0.009648 | 6.703169e-07 | 3.074505e-05 | 12 | [Vas] |
| Oli | 17167 | OR2M2 | ENSG00000198601 | -0.009376 | 2.918413e-04 | 7.125222e-03 | 8 | NaN |
| Oli | 13934 | OR4C5 | ENSG00000176540 | -0.008712 | 2.579611e-45 | 1.439530e-42 | 185 | NaN |
| Vas | 14057 | OR2T12 | ENSG00000177201 | -0.011536 | 1.722700e-13 | 2.185596e-11 | 30 | NaN |
| Vas | 13372 | GPR148 | ENSG00000173302 | -0.010499 | 2.810499e-07 | 1.631757e-05 | 16 | NaN |
| Vas | 16824 | OR13G1 | ENSG00000197437 | -0.010437 | 4.088132e-11 | 4.180267e-09 | 38 | NaN |
| Vas | 19850 | FAM90A26 | ENSG00000229924 | -0.009362 | 6.857068e-04 | 1.622721e-02 | 6 | NaN |
| Vas | 16070 | SOWAHD | ENSG00000187808 | -0.009224 | 7.835720e-04 | 1.801427e-02 | 7 | NaN |
| Vas | 11785 | RPL27A | ENSG00000166441 | -0.009058 | 1.013644e-83 | 1.068380e-80 | 225 | NaN |
| Vas | 6681 | TPT1 | ENSG00000133112 | -0.009047 | 7.431883e-05 | 2.489772e-03 | 8 | NaN |

Table 3.2: **Detailed results of in silico perturbation, with alt states:** Displaying gene information, cosine shifts, p-values, and the other cell types where the gene was deemed significant. Positive shifts denote the gene is significant for excitatory cell classification, negative shifts for differentiating the cell type from excitatory neurons.

Short summaries of important information regarding some of the significant genes:

- **TPT1:** This gene encodes a protein that regulates cellular growth, proliferation, and microtubule stability. It is involved in apoptosis, protein synthesis, and cell division, with a role in mitotic and meiotic progression. Associated with carcinogenesis and is upregulated in certain cancers.
- **CYP27B1:** This gene encodes a cytochrome P450 enzyme that synthesises the active form of vitamin D3, regulating calcium metabolism. Mutations can

cause vitamin D-dependent rickets type I. Prognostic for carcinoma.

- **VIPAS39:** Encodes a protein involved in the sorting of lysosomal proteins. Associated mutations with ARCS2 (arthrogryposis, renal dysfunction, cholestasis-2). Expressed in oligodendrocytes.
- **PRR20C:** This gene is part of a cluster of five identical loci on chromosome 13q21.1. The predicted proline-rich protein contains dopamine D4 receptor signatures and PRINTS domains. Not detected in cell type RNA expression.
- **GRID2:** This gene encodes an ionotropic glutamate receptor primarily expressed in cerebellar Purkinje cells, playing a role in synapse organisation. Mutations cause cerebellar ataxia in humans and neuronal apoptosis in mice, with severe phenotypes leading to embryonic or postnatal death. Cen be expressed in most major cell types of the dataset (Oligodendrocyte precursor cells, Inhibitory neurons, Excitatory neurons, Oligodendrocytes, Microglial cells)
- **ABCC2:** This gene encodes a protein in the ATP-binding cassette (ABC) transporter superfamily, specifically the MRP subfamily, which contributes to multidrug resistance. Expressed in hepatocyte canaliculi, it functions in biliary transport and impacts drug resistance. Mutations in this gene are associated with Dubin-Johnson syndrome, characterised by conjugated hyperbilirubinemia.
- **JSRP1:** This gene encodes a protein involved in excitation-contraction coupling in skeletal muscle, regulating calcium influx and efflux at the sarcoplasmic reticulum through interactions with CACNA1S, CACNB1, and calsequestrin.
- **PABPC1:** This gene encodes a poly(A) binding protein that facilitates translation initiation by binding to the 3' poly(A) tail of mRNAs and promotes mRNA decay through poly(A) shortening. It shuttles between the nucleus and cytoplasm and belongs to a small gene family with related pseudogenes.
- **LCE2A:** Predicted to be involved in keratinization.
- **RPL41:** This gene encodes a cytoplasmic ribosomal protein of the 60S subunit, belonging to the L41E family and sharing similarity with yeast YL41. It interacts with protein kinase CKII to stimulate DNA topoisomerase II-alpha phosphorylation.
- **HNRNPA1:** This gene encodes a core heterogeneous nuclear ribonucleoprotein (hnRNP) involved in pre-mRNA processing, mRNA metabolism, and alternative splicing regulation. Mutations are linked to amyotrophic lateral sclerosis 20, and multiple transcript variants exist due to alternative splicing.
- **OR1L1:** Olfactory receptors are G-protein-coupled receptors that detect odorant molecules and mediate smell perception. They belong to the largest gene family in the genome. Not detected in cell type RNA expression.

- **ALDOA:** This gene encodes a glycolytic enzyme that converts fructose-1,6-bisphosphate into glyceraldehyde 3-phosphate and dihydroxyacetone phosphate. Mutations are linked to Glycogen Storage Disease XII and various cancers. Related pseudogenes are located on chromosomes 3 and 10.

- **CGB5:** This gene encodes the beta 5 subunit of chorionic gonadotropin (CG), a glycoprotein hormone produced by placental trophoblastic cells to support pregnancy. It is part of a gene cluster on chromosome 19q13.3, arranged alongside luteinizing hormone beta subunit genes.

Gene function summaries are provided by the Human Protein Atlas and its sources[19]. The expression levels may be context-specific and relevant only to the analysed sample. Most genes found do not relate to our major cell types directly, or to AD pathology, according to the existing body of literature in the atlas.

We repeat the process once more, but do so strictly for each start-end goal pair without alternate states: no other cell types included. Between the two trials, only one gene (RPS15A) among the top differs (Figure 3.4). The cosine shifts to goals ends are also similarly positive or negative in both tables. These results indicate that the in silico perturbation is robust, consistently identifying similar patterns in significance despite the absence of alternate states.

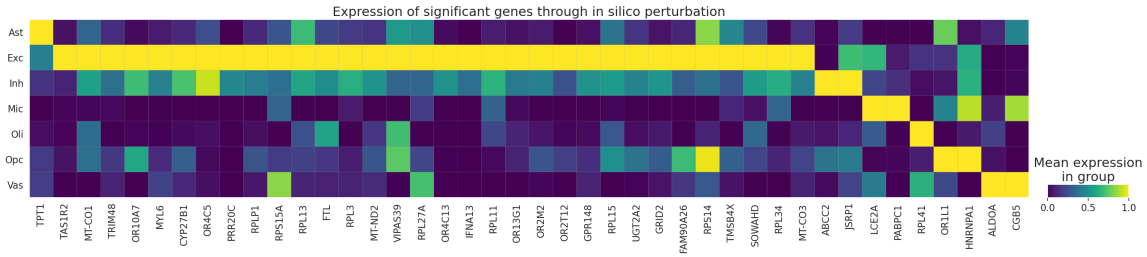



Figure 3.4: **Gene expression matrix of in silico perturbation results, no alt states:** Aggregated top 7 significant cell types for each perturbation. Duplicates found between runs were not allowed for variety.

| Major Cell Type | Gene | Gene name | Ensembl ID | Shift to goal end | Goal end vs random pval | Goal end FDR | N Detections | Cell Types with Top Gene |
|---|---|---|---|---|---|---|---|---|
| Ast | 17905 | TMSB4X | ENSG00000205542 | -0.015838 | 6.744859e-96 | 9.243154e-93 | 158 | [Ast, Mic, OPC, Oli, Vas] |
| Ast | 21114 | OR4C13 | ENSG00000258817 | -0.011605 | 6.646020e-125 | 1.517951e-121 | 264 | [Ast, OPC] |
| Ast | 9454 | GRID2 | ENSG00000152208 | -0.011135 | 1.449894e-32 | 6.021013e-30 | 59 | [Ast, Mic, OPC, Oli] |
| Ast | 9220 | ALDOA | ENSG00000149925 | 0.010367 | 2.405999e-44 | 1.498719e-41 | 96 | [Ast, Vas] |
| Ast | 3664 | RPL34 | ENSG00000109475 | 0.010322 | 1.926570e-07 | 1.269313e-05 | 12 | [Ast, Inh] |
| Ast | 2257 | RPL3 | ENSG00000100316 | -0.010009 | 1.652766e-04 | 5.807567e-03 | 6 | [Ast, Vas] |
| Ast | 14487 | OR10A7 | ENSG00000179919 | -0.008950 | 1.861370e-06 | 1.022378e-04 | 21 | [Ast] |
| Inh | 12019 | JSRP1 | ENSG00000167476 | 0.013817 | 0.000000e+00 | 0.000000e+00 | 1454 | [Inh] |
| Inh | 1720 | FTL | ENSG00000087086 | 0.010412 | 1.730868e-22 | 5.099664e-20 | 47 | [Inh] |
| Inh | 13620 | RPL15 | ENSG00000174748 | 0.010212 | 1.337534e-25 | 4.648102e-23 | 49 | [Inh] |
| Inh | 1984 | MYL6 | ENSG00000092841 | -0.010170 | 0.000000e+00 | 0.000000e+00 | 1127 | [Inh, OPC, Oli, Vas] |
| Inh | 436 | ABCC2 | ENSG00000023839 | 0.009855 | 2.646684e-188 | 8.967626e-185 | 785 | [Inh] |
| Inh | 11319 | RPS14 | ENSG00000164587 | -0.009596 | 1.710477e-03 | 3.732164e-02 | 5 | [Inh, OPC, Oli] |
| Inh | 8231 | RPL11 | ENSG00000142676 | -0.008701 | 4.006244e-20 | 9.525723e-18 | 63 | [Mic] |
| Mic | 16342 | CGB5 | ENSG00000189052 | 0.016410 | 1.027644e-03 | 1.781556e-02 | 4 | [Mic] |
| Mic | 7502 | RPLP1 | ENSG00000137818 | -0.012462 | 1.097695e-03 | 1.871881e-02 | 4 | [Mic] |
| Mic | 3831 | CYP27B1 | ENSG00000111012 | 0.011356 | 3.481489e-03 | 4.734056e-02 | 4 | [Mic] |
| Mic | 17228 | MT-ND2 | ENSG00000198763 | -0.010915 | 5.969246e-51 | 3.518481e-48 | 101 | [Mic] |
| Mic | 19843 | PRR20C | ENSG00000229665 | -0.009980 | 3.544627e-157 | 6.864930e-154 | 543 | NaN |
| Mic | 13445 | OR1L1 | ENSG00000173679 | -0.009287 | 3.213566e-04 | 7.049567e-03 | 7 | NaN |
| Mic | 17326 | MT-CO3 | ENSG00000198938 | -0.009071 | 2.321304e-04 | 5.397928e-03 | 8 | [Oli, Vas] |
| OPC | 9355 | VIPAS39 | ENSG00000151445 | -0.011071 | 0.000000e+00 | 0.000000e+00 | 1497 | [OPC] |
| OPC | 23863 | UGT2A2 | ENSG00000271271 | -0.007079 | 9.180953e-04 | 2.157470e-02 | 13 | [OPC] |
| OPC | 7055 | HNRNPA1 | ENSG00000135486 | 0.007078 | 0.000000e+00 | 0.000000e+00 | 912 | NaN |
| OPC | 19992 | IFNA13 | ENSG00000233816 | -0.007048 | 2.711695e-11 | 2.986866e-09 | 37 | NaN |
| OPC | 19823 | RPL41 | ENSG00000229117 | 0.006777 | 6.381975e-40 | 3.573374e-37 | 92 | NaN |
| OPC | 9238 | TRIM48 | ENSG00000150244 | -0.006757 | 2.277354e-15 | 3.643224e-13 | 58 | NaN |
| OPC | 6861 | RPS15A | ENSG00000134419 | 0.006394 | 1.202431e-25 | 3.512666e-23 | 54 | NaN |
| Oli | 12030 | RPL13 | ENSG00000167526 | -0.013229 | 1.206016e-04 | 3.316668e-03 | 6 | [Oli] |
| Oli | 17247 | MT-CO1 | ENSG00000198804 | -0.011369 | 2.945605e-06 | 1.199103e-04 | 11 | [Oli] |
| Oli | 14346 | TAS1R2 | ENSG00000179002 | -0.011015 | 2.532891e-03 | 4.294052e-02 | 7 | NaN |
| Oli | 15952 | LCE2A | ENSG00000187173 | -0.010253 | 5.709044e-04 | 1.263822e-02 | 6 | [Vas] |
| Oli | 1129 | PABPC1 | ENSG00000070756 | -0.009648 | 6.643280e-07 | 3.016049e-05 | 12 | [Vas] |
| Oli | 17167 | OR2M2 | ENSG00000198601 | -0.009370 | 2.900222e-04 | 6.986092e-03 | 8 | NaN |
| Oli | 13934 | OR4C5 | ENSG00000176540 | -0.008688 | 4.260665e-45 | 2.377628e-42 | 185 | NaN |
| Vas | 14057 | OR2T12 | ENSG00000177201 | -0.011569 | 1.286081e-13 | 1.646905e-11 | 30 | NaN |
| Vas | 13372 | GPR148 | ENSG00000173302 | -0.010497 | 2.534519e-07 | 1.485365e-05 | 16 | NaN |
| Vas | 16824 | OR13G1 | ENSG00000197437 | -0.010432 | 3.434354e-11 | 3.511755e-09 | 38 | NaN |
| Vas | 19850 | FAM90A26 | ENSG00000229924 | -0.009393 | 6.367585e-04 | 1.506885e-02 | 6 | NaN |
| Vas | 16070 | SOWAHD | ENSG00000187808 | -0.009243 | 7.472048e-04 | 1.717819e-02 | 7 | NaN |
| Vas | 11785 | RPL27A | ENSG00000166441 | -0.009102 | 6.811284e-85 | 7.179094e-82 | 225 | NaN |
| Vas | 6681 | TPT1 | ENSG00000133112 | -0.009042 | 6.827465e-05 | 2.292890e-03 | 8 | NaN |

Table 3.3: **Detailed results of in silico perturbation, no alt states:** Displaying gene information, cosine shifts, p-values, and the other cell types where the gene was deemed significant. Positive shifts denote the gene is significant for excitatory cell classification, negative shifts for differentiating the cell type from excitatory neurons.

## 3.1 Patient classification

Patient classification is a substantial challenge. To achieve more manageable classification conditions, the analysis focused on excitatory neurons, a homogeneous subpopulation expected to yield more successful model training metrics compared to diverse cell types. A notable trial achieved perfect accuracy on the training set (Table 3.4), confirming that overfitting is possible with the rank value encoding algorithm and model architecture. This implies that sufficient signal is retained to discern the training data effectively.

| Learning Rate | LR Scheduler Type | Seed | Warmup Steps | Weight Decay | Loss | Accuracy | Macro F1 |
|---|---|---|---|---|---|---|---|
| 0.00034152 | polynomial | 15.9191 | 1952.24 | 0.187833 | 0.000258712 | 1 | 1 |
| 0.000106839 | polynomial | 24.5637 | 1438.51 | 0.131524 | 0.00435945 | 0.999493 | 0.999473 |
| 0.000256832 | polynomial | 61.5161 | 1910.39 | 0.184694 | 0.0107432 | 0.995943 | 0.995794 |
| 0.000366788 | polynomial | 80.6359 | 1279.49 | 0.211489 | 0.0250268 | 0.994929 | 0.994723 |
| 0.000482551 | polynomial | 70.8157 | 1655.46 | 0.191817 | 0.0267574 | 0.990872 | 0.990539 |
| 0.00156075 | polynomial | 95.907 | 1943.18 | 0.278394 | 0.092122 | 0.963996 | 0.962675 |
| 0.00797425 | polynomial | 23.7498 | 1427.59 | 0.286615 | 0.674488 | 0.596349 | 0.373571 |
| 0.00893749 | polynomial | 25.7042 | 1821.62 | 0.25486 | 0.680027 | 0.596349 | 0.373571 |
| 0.00986826 | polynomial | 60.4795 | 798.837 | 0.144146 | 0.678313 | 0.596349 | 0.373571 |
| 0.00273875 | polynomial | 12.7131 | 1604.75 | 0.281221 | 0.674353 | 0.596349 | 0.373571 |

Table 3.4: **Trial results for overfitting on excitatory neurons:** While not definitive, overfitting on the training set is possible. This confirms that the output of the rank value encoding algorithm is distinct enough for AD diagnostic differences between patients to enable learning.

The validation of overfitting ensures the encoding's capacity to retain training data signals. However, Geneformer struggles to generalise on normal hyperparameter tuning trials. Possible explanations include the limited number (2048) of genes selected for rank value encoding, potentially omitting key biomarkers. The inability of the model to discriminate class differences suggests that either the selected features lack sufficient variation or that architectural changes in the model are needed to handle our scRNA-seq data. Literature corroborates the sensitivity of transformer-based models to input feature quality and context dependencies[2].

| Learning Rate | LR Scheduler Type | Seed | Warmup Steps | Weight Decay | Loss | Accuracy | Macro F1 |
|---|---|---|---|---|---|---|---|
| 5.76831e-06 | cosine | 16.0253 | 284.863 | 0.0770874 | 0.680301 | 0.571615 | 0.429808 |
| 3.0689e-06 | cosine | 27.2598 | 262.007 | 0.299489 | 0.682111 | 0.554688 | 0.356784 |
| 2.01827e-06 | cosine | 12.9055 | 160.731 | 0.0117983 | 0.682118 | 0.554688 | 0.356784 |
| 6.14853e-06 | cosine | 8.40538 | 372.935 | 0.234521 | 0.682888 | 0.5625 | 0.446154 |
| 1.40901e-06 | linear | 49.1782 | 406.849 | 0.275708 | 0.68332 | 0.554688 | 0.356784 |
| 1.38879e-06 | linear | 22.0441 | 1918.84 | 0.163265 | 0.684406 | 0.557292 | 0.363197 |
| 1.1236e-07 | linear | 30.2365 | 784.779 | 0.0795632 | 0.686822 | 0.553385 | 0.356245 |
| 8.18224e-07 | cosine | 15.5075 | 103.374 | 0.00086361 | 0.686918 | 0.554688 | 0.356784 |
| 2.13882e-07 | linear | 51.2208 | 1324.36 | 0.18674 | 0.687526 | 0.554688 | 0.356784 |
| 2.6152e-07 | cosine | 36.1625 | 1563.41 | 0.0679355 | 0.687827 | 0.554688 | 0.356784 |
| 0.00484532 | linear | 15.2331 | 1885.41 | 0.195204 | 0.689302 | 0.554688 | 0.356784 |
| 0.00804846 | cosine | 49.6939 | 1947.97 | 0.0925221 | 0.689406 | 0.554688 | 0.356784 |
| 8.26565e-06 | polynomial | 58.4741 | 1673.32 | 0.179499 | 0.690931 | 0.55599 | 0.492842 |
| 0.00364566 | cosine | 70.7435 | 871.099 | 0.163804 | 0.691542 | 0.554688 | 0.356784 |
| 1.09346e-05 | polynomial | 21.352 | 978.936 | 0.208995 | 0.696298 | 0.574219 | 0.539456 |
| 1.2245e-05 | polynomial | 69.1916 | 1809.04 | 0.00931146 | 0.721866 | 0.558594 | 0.533238 |
| 1.87671e-05 | polynomial | 5.04512 | 495.799 | 0.15985 | 0.921359 | 0.567708 | 0.550564 |
| 2.14146e-05 | polynomial | 41.3188 | 784.906 | 0.0742512 | 1.08502 | 0.541667 | 0.532766 |
| 3.48657e-05 | polynomial | 39.6723 | 458.296 | 0.183663 | 2.79169 | 0.559896 | 0.540445 |
| 4.01191e-05 | polynomial | 52.6649 | 965.926 | 0.118072 | 3.23118 | 0.545573 | 0.534776 |
| 0.00201032 | polynomial | 91.8507 | 1607.74 | 0.186914 | 3.62526 | 0.542969 | 0.539872 |
| 0.000132378 | cosine | 27.9603 | 605.555 | 0.0314321 | 3.76051 | 0.567708 | 0.555357 |
| 0.000122177 | cosine | 8.35798 | 107.824 | 0.246549 | 4.07972 | 0.557292 | 0.551127 |
| 0.000431666 | cosine | 46.6655 | 929.392 | 0.16115 | 4.15561 | 0.559896 | 0.549621 |
| 0.00437128 | linear | 1.18245 | 1452.29 | 0.0751053 | 4.20038 | 0.49349 | 0.465077 |

Table 3.5: **Trial results for 20K excitatory neurons with 5 layers frozen**

| Learning Rate | LR Scheduler Type | Seed | Warmup Steps | Weight Decay | Loss | Accuracy | Macro F1 |
|---|---|---|---|---|---|---|---|
| 1.70705e-06 | cosine | 44.2462 | 1087.71 | 0.0160589 | 0.678197 | 0.554688 | 0.356784 |
| 1.59611e-06 | linear | 36.0439 | 1336.38 | 0.168942 | 0.680583 | 0.55599 | 0.36264 |
| 1.28442e-06 | linear | 39.2947 | 986.617 | 0.0500438 | 0.682018 | 0.554688 | 0.356784 |
| 9.2264e-07 | linear | 31.6398 | 1350.27 | 0.149118 | 0.683429 | 0.554688 | 0.356784 |
| 4.69825e-07 | linear | 20.6763 | 1457.01 | 0.135294 | 0.685172 | 0.554688 | 0.356784 |
| 1.01512e-07 | linear | 10.0987 | 1098.31 | 0.0504333 | 0.68552 | 0.554688 | 0.356784 |
| 3.1524e-06 | linear | 63.4825 | 839.936 | 0.232146 | 0.686827 | 0.554688 | 0.463932 |
| 5.44495e-07 | linear | 98.0154 | 888.548 | 0.0914531 | 0.687326 | 0.554688 | 0.356784 |
| 1.97109e-07 | linear | 3.42791 | 1227.85 | 0.298867 | 0.687867 | 0.554688 | 0.356784 |
| 0.000876123 | cosine | 6.20099 | 1618.78 | 0.205211 | 0.688992 | 0.554688 | 0.356784 |
| 0.00434285 | cosine | 86.1734 | 850.593 | 0.216813 | 0.688994 | 0.554688 | 0.356784 |
| 0.00238929 | cosine | 78.6919 | 1450.63 | 0.0867017 | 0.689004 | 0.554688 | 0.356784 |
| 0.000539011 | cosine | 87.4956 | 774.334 | 0.191897 | 0.689124 | 0.554688 | 0.356784 |
| 0.00729533 | linear | 42.629 | 1727.19 | 0.0147546 | 0.68923 | 0.554688 | 0.356784 |
| 6.25083e-06 | cosine | 71.9813 | 225.015 | 0.105668 | 0.715395 | 0.552083 | 0.522222 |
| 0.000716384 | polynomial | 22.3296 | 601.731 | 0.0133209 | 0.726386 | 0.550781 | 0.539731 |
| 6.90017e-06 | cosine | 52.7201 | 243.755 | 0.0767803 | 0.741805 | 0.529948 | 0.50358 |
| 8.78795e-06 | linear | 86.0628 | 222.177 | 0.194997 | 0.766945 | 0.558594 | 0.535569 |
| 0.00028494 | polynomial | 93.084 | 378.183 | 0.124152 | 0.856423 | 0.572917 | 0.485877 |
| 0.000666523 | cosine | 45.466 | 343.335 | 0.183326 | 0.873425 | 0.566406 | 0.534523 |
| 1.82981e-05 | polynomial | 77.953 | 677.982 | 0.251454 | 1.53825 | 0.5625 | 0.547903 |
| 2.70995e-05 | polynomial | 22.0147 | 1832.19 | 0.0099962 | 2.77898 | 0.559896 | 0.54998 |
| 0.000569365 | cosine | 28.7685 | 493.465 | 0.249302 | 2.97225 | 0.550781 | 0.501714 |
| 6.71831e-05 | polynomial | 58.0523 | 1101.95 | 0.26848 | 3.97841 | 0.552083 | 0.540874 |
| 0.000161596 | cosine | 47.316 | 619.387 | 0.0741865 | 4.07604 | 0.529948 | 0.516301 |

Table 3.6: Trial results for 20K excitatory neurons with 2 layers frozen

| Learning Rate | LR Scheduler Type | Seed | Warmup Steps | Weight Decay | Loss | Accuracy | Macro F1 |
|---|---|---|---|---|---|---|---|
| 7.85501e-07 | polynomial | 95.5739 | 1676.19 | 0.24201 | 0.663871 | 0.58931 | 0.455207 |
| 1.12953e-07 | linear | 41.3737 | 718.734 | 0.0644596 | 0.674259 | 0.581846 | 0.372358 |
| 1.07865e-05 | linear | 11.8334 | 774.164 | 0.0660742 | 0.711499 | 0.624571 | 0.604043 |
| 0.000319417 | cosine | 61.6481 | 1541.78 | 0.153133 | 5.68559 | 0.604238 | 0.58781 |
| 1.24308e-06 | cosine | 99.2068 | 598.442 | 0.145516 | 0.660256 | 0.597203 | 0.515481 |
| 8.60674e-07 | polynomial | 18.5802 | 1372.3 | 0.215137 | 0.662311 | 0.598147 | 0.48229 |
| 1.35656e-07 | cosine | 47.2538 | 259.104 | 0.0945343 | 0.673138 | 0.581846 | 0.368969 |
| 0.000191843 | linear | 57.3682 | 337.832 | 0.226831 | 5.02011 | 0.61359 | 0.597304 |

Table 3.7: **Trial results for 160K excitatory neurons with 5 layers frozen**

| Learning Rate | LR Scheduler Type | Seed | Warmup Steps | Weight Decay | Loss | Accuracy | Macro F1 |
|---|---|---|---|---|---|---|---|
| 0.000133531 | cosine | 44.4689 | 1277.87 | 0.00613319 | 2.45353 | 0.623501 | 0.611182 |
| 2.85294e-05 | linear | 18.1739 | 500.922 | 0.156032 | 0.865784 | 0.609184 | 0.586734 |
| 0.000191777 | linear | 56.3456 | 1464.3 | 0.239821 | 3.74787 | 0.614601 | 0.585356 |
| 2.57024e-05 | polynomial | 61.9749 | 147.394 | 0.249124 | 0.654654 | 0.619115 | 0.56448 |
| 3.08694e-06 | linear | 98.638 | 1948.5 | 0.112038 | 0.667825 | 0.581452 | 0.377194 |

Table 3.8: **Trial results for 160K excitatory neurons with 2 layers frozen**

Multiple parameters, number of frozen layers and training cells were chosen for trials. In all configurations, the model failed to generalise or converge on meaningful diagnostic patterns. Slight fluctuations in accuracy or macro F1 scores across trials (e.g. Table 3.7) were attributable to class imbalances induced by the model labelling all instances as a single dominant class.

# 4

# Conclusion

## 4.1 Discussion

We fine-tuned Geneformer on scRNA-seq data from the ROSMAP dataset to classify between 7 major cell types associated with AD. The model can achieve high accuracy and macro F1 scores when distinguishing between major cell types even without a large number of cells given as training input. This result underscores the effectiveness of the pretrained transformer-based architecture in capturing complex gene expression patterns in single-cell data.

The in silico perturbation analysis identified several genes with significant shifts in expression across major cell types. Some of these genes, such as TPT1 and CYP27B1, have known roles in cellular processes irrelevant to AD pathology. Most significant genes did not directly relate to the major cell types or AD according to existing literature. Furthermore, the representation of significant genes did not correlate with the cell types they were found to be significant for upon perturbation. However, it is important to note that only the top genes were presented in the plots. Investigation of the full list of significant genes is warranted for a complete view of the process. Genes not present in the tables, such as **KLK6** and **MOBP**, which are known to be related to dementia and a major cell type of the dataset, have been deemed significant through perturbation. These results suggest that the model may capture novel or less-characterised gene associations, which warrants further biological validation.

Despite the success in cell type classification, the model could not generalise well in patient classification tasks. While overfitting on the training set indicated that the model could capture patterns within the tokenised ranked genes, it failed to perform adequately on testing data. One possible reason is the rank value encoding algorithm, which limits input to the top 2048 genes. This constraint may exclude genes necessary for distinguishing AD patients from non-AD individuals. In addition, the imbalance in cell type representation, particularly the underrepresentation of vascular cells, affected the model's performance. Handling a large dataset posed challenges in terms of resource allocation for extensive hyperparameter tuning. Finally, all of the aforementioned data cleaning steps were replicated according to the ROSMAP paper to remain consistent with their analysis. Thus, some choices may be suboptimal for our objective of finetuning Geneformer.

These findings highlight the complexity of capturing disease-relevant signals from scRNA-seq data. The observed discrepancy between strong cell type classification performance and weaker patient-level classification emphasises the challenge of identifying underlying molecular signatures that differentiate AD patients from non-AD individuals, on a limited amount of available genes. Moreover, the lack of direct correspondence between many identified significant genes and known AD markers may indicate that current approaches miss potentially important but less-characterised molecular interactions. Integrating additional data sources or employing an updated Geneformer, which now incorporates double the number of genes in the rank value encoding, capable of handling a broader gene set may help uncover these hidden patterns.

In conclusion, this study demonstrates the potential of fine-tuning pretrained models like Geneformer for specific tasks in genomics. While challenges remain in applying the model to disease classification, the success in cell type classification highlights their applicability in understanding cellular heterogeneity in complex diseases like AD. Future work should address these limitations by leveraging the latest version of Geneformer. This update shows promise in capturing more informative biomarkers but became available too late during the study. Continued refinement and integration of diverse data sources will be essential for advancing the utility of transformer-based models in biomedical research.

## 4.2 Conclusions

A path for further research is Geneformer's newly introduced multitask classification capability. Since the model already excels at distinguishing major cell types, it may be possible to extend those same learned representations to handle patient classification if both tasks are trained together. By jointly modelling cell identity and disease status, the multitask approach could reveal transcriptomic signals that single-task training might overlook.

Another practical refinement involves revisiting the data preprocessing pipeline. The current approach follows replicated quality control procedures. While that ensures clean data according to the ROSMAP paper, it also risks discarding cells that would otherwise be informative to Geneformer. Relaxing some of these thresholds and retaining a broader cell population could expand Geneformer's view of AD-relevant heterogeneity. In parallel, incorporating more genes during rank value encoding and expanding the architecture's depth may capture a greater slice of the transcriptome and further improve disease-related predictions.

Finally, a gene regulatory network (GRN) can be constructed on the dataset using a tool such as SCENIC[20]. The embeddings and findings from Geneformer might highlight gene relationships that we can then compare to the GRN. This could reveal new regulatory interactions or confirm hypotheses about how genes control cell behaviour. Additionally, analysing the genes identified by Geneformer and their corresponding roles in the GRN can tell us which genes are key to a cell's identity or disease state. When certain genes and cells are flagged as important, they can

become targets for wet lab testing.

# Bibliography


[1] I. G.-V. Roni Wilentzik Müller, "Exploring neural networks and related visualization techniques in gene expression data," *Frontiers*, 2020.

[2] C. V. Theodoris, L. Xiao, A. Chopra, *et al.*, "Transfer learning enables predictions in network biology," *Nature*, 2023. [Online]. Available: https://www.nature.com/articles/s41586-023-06139-9.

[3] Z. Wang, M. Gerstein, and M. Snyder, "Rna-seq: A revolutionary tool for transcriptomics," *Nature Reviews Genetics*, 2009.

[4] N. Methods, "Method of the year 2013," *Nature Methods*, vol. 11, pp. 1–1, 2013. DOI: 10.1038/nmeth.2801.

[5] NIAGADS DSS, *Religious orders study/memory and aging project (rosmap)*, https://dss.niagads.org/cohorts/religious-orders-study-memory-and-aging-project-rosmap/.

[6] S. B. Roshanzamir Alireza Aghajan Hamid, "Transformer-based deep neural network language models for alzheimer's disease risk assessment from targeted speech," *BMC*, 2021.

[7] G.-R. K. Uttam Khatri, "Explainable vision transformer with self-supervised learning to predict alzheimer's disease progression using 18f-fdg pet," *MDPI*, 2023.

[8] H. Mathys, G. Abdelhady, X. Jiang, *et al.*, "Single-cell atlas reveals correlates of high cognitive function, dementia, and resilience to alzheimer's disease pathology," *Cell*, Sep. 2023. DOI: 10.1016/j.cell.2023.08.039. [Online]. Available: https://doi.org/10.1016/j.cell.2023.08.039.

[9] H. K. Kimelberg, "Functions of astrocytes in brain health and disease: A review," *Neurochemical Research*, vol. 35, no. 12, pp. 1946–1959, 2010. DOI: 10.1007/s11064-010-0377-5.

[10] A. Nimmerjahn, F. Kirchhoff, and F. Helmchen, "Resting microglial cells are highly dynamic surveillants of brain parenchyma in vivo," *Science*, vol. 308, no. 5726, pp. 1314–1318, 2005. DOI: 10.1126/science.1110647.

[11] N. J. Abbott, L. Rönnbäck, and E. Hansson, "Astrocyte-endothelial interactions at the blood-brain barrier," *Nature Reviews Neuroscience*, vol. 7, no. 1, pp. 41–53, 2010. DOI: 10.1038/nrn1824.

[12] A. Kepecs and G. Fishell, "Interneuron cell types in the neocortex and hippocampus," *Neuron*, vol. 82, no. 6, pp. 1283–1301, 2014. DOI: 10.1016/j.neuron.2014.05.027.

[13] T. V. Bliss and G. L. Collingridge, "A synaptic model of memory: Long-term potentiation in the hippocampus," *Nature*, vol. 361, no. 6407, pp. 31–39, 1993. DOI: 10.1038/361031a0.

[14] M. Simons and K.-A. Nave, "Oligodendrocytes: Myelination and axonal support," *Cold Spring Harbor Perspectives in Biology*, vol. 8, no. 1, a020479, 2016. DOI: 10.1101/cshperspect.a020479.

[15] R. J. Franklin and C. Ffrench-Constant, "Regenerating cns myelin: From mechanisms to experimental medicines," *Nature Reviews Neuroscience*, vol. 18, no. 12, pp. 753–769, 2017. DOI: 10.1038/nrn.2017.136.

[16] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017. DOI: 10.48550/arXiv.1706.03762. [Online]. Available: https://doi.org/10.48550/arXiv.1706.03762.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2019. [Online]. Available: https://arxiv.org/abs/1810.04805.

[18] C. V. Theodoris and contributors, *Geneformer: Discussion on CLS Token and Cell Embedding*, Accessed: 2024-08-21, 2023. [Online]. Available: https://huggingface.co/ctheodoris/Geneformer/discussions/227.

[19] H. P. Atlas, *The human protein atlas*, Accessed: 2024-12-02, 2024. [Online]. Available: https://www.proteinatlas.org/.

[20] S. Aibar, C. González-Blas, T. Moerman, *et al.*, "Scenic: Single-cell regulatory network inference and clustering," *Nature Methods*, 2017. DOI: 10.1038/nmeth.4463. [Online]. Available: https://doi.org/10.1038/nmeth.4463.

[21] I. Virshup, S. Rybakov, F. J. Theis, P. Angerer, and F. A. Wolf, "Anndata: Annotated data," *bioRxiv*, 2021. DOI: 10.1101/2021.12.16.473007. [Online]. Available: https://doi.org/10.1101/2021.12.16.473007.

[22] J. M. Leland McInnes John Healy, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv*, 2018.

[23] V. Bhardwaj *et al.*, "Gene regulatory networks and their applications: Understanding biological and medical problems in terms of networks," *Frontiers in Cell and Developmental Biology*, 2014. DOI: 10.3389/fcell.2014.00038. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fcell.2014.00038/full.

# A

# Appendix 1

## A.1 AnnData

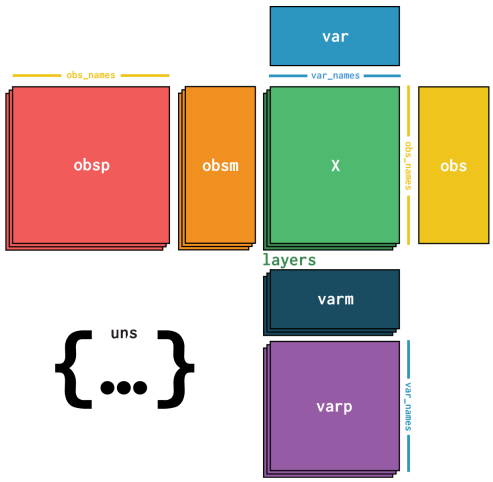



Figure A.1: AnnData object structure

Anndata objects are containers tailored for handling complex data typically found in single-cell genomics.[21] The primary data, usually gene expression counts, goes into the .X matrix. Each row corresponds to a single cell, and each column to a gene.

Metadata about cells, such as cell type, batch and patient ID, is stored in the .obs dataframe, while .var holds gene-related information, like gene names or their corresponding Ensembl ID. .Obsm and .varm are matrices for storing multidimensional annotations related to .obs and .var, respectively, which include principal components from PCA and UMAP coordinates. .Obsp and .varp are designated for storing pairwise relationships between observations or variables. Examples include distance matrices or adjacency matrices representing relationships like spatial proximity or co-expression networks, such as nearest neighbours relations.

The layers section is used for storing different transformations of your primary data. For example, if we normalised or performed feature scaling, we could keep the original counts in X and the processed data in layers. The uns slot is reserved for unstructured data, that does not fit into a tabular form. It can vary between various

metadata or parameters related to the analysis, such as Leiden clustering information or PCA parameters.

## A.2 UMAP

UMAP (Uniform Manifold Approximation and Projection) is a dimension reduction technique, scalable and suitable for real-world data.[22] The technique is constructed to preserve both local and global structure of high dimensional data, thus being advantageous for visualisation and as a general-purpose dimension reduction method for machine learning.

UMAP is based on manifold theory and topological data analysis. The process begins by approximating the manifold on which data presumably lies, through local manifold approximations. Then, it constructs a representation of the data by patching together local fuzzy simplicial sets. This representation is optimised in low dimensional space to mirror the high dimensional topological structure. Practically, UMAP constructs a weighted k-neighbour graph from the high-dimensional data, transforming it through non-linear dimensionality reduction computations to achieve a low dimensional layout. Cross-entropy between the two representations is used for optimisation.

The algorithm is efficient, and has been applied across various fields, and is a staple in genomics visual analysis, along with other methods, such as t-SNE. UMAP plots can be utilised to visually identify clusters, data distributions, insights into the the heterogeneity or uniformity of the data, relationships between data points or clusters and potential outliers.

## A.3 Gene regulatory network

A gene regulatory network (GRN) is a system of interconnected molecular regulators that dictate the expression of genes. GRNs consist of transcription factors, DNA regulatory elements, and signaling molecules. These components coordinate to regulate gene activation and repression. GRNs are crucial for cellular processes such as differentiation, development, and response to environmental stimuli. Their study helps explain biological mechanisms and their disruption in diseases[23].