# マウス版 Geneformer の構築と転移学習による細胞解析への応用

伊藤　啓太†　　平川　　翼†　　山下　隆義†　　藤吉　弘亘†

† 中部大学　〒 487–0027 愛知県春日井市松本町 1200

E-mail: †{itokeita,hirakawa}@mprg.cs.chubu.ac.jp, ††{takayoshi,fujiyoshi}@isc.chubu.ac.jp

**あらまし**　Single-cell RNA-sequence 解析は，細胞 1 つ 1 つの遺伝子発現量を次元圧縮やクラスタリングなどの数理的な手法を利用して解析する手法である．この解析により，新規の細胞種や異常な遺伝子の特定が可能となった．Single-cell RNA-sequence 解析手法には，機械学習や深層学習の技術を用いた手法も存在し，深層学習の技術を利用した解析手法として Attention ベースの事前学習モデル Geneformer がある．しかし，Geneformer は，人の Single-cell RNA-sequence データを用いて事前学習を行っているため，ファインチューニング時に，マウスの Single-cell RNA-sequence データを用いることはできない．そこで，本研究では，マウスの Single-cell RNA-sequence データを用いて事前学習したモデルとしてマウス版 Geneformer を構築する．これにより，従来解析手法よりも高精度に細胞型の分類を行うことが可能であることを示す．また，in silico 摂動実験により病気の原因遺伝子を同定することが可能であることを示す．

**キーワード**　自然言語処理応用，医療・ヘルスケア，Single-cell RNA-sequence 解析，Geneformer

## 1　はじめに

病気や障害の原因は，異常な細胞の遺伝子情報をタンパク質として発現していることである．病気や障害の原因究明には，異常なタンパク質を発現している異常な細胞を特定する必要がある．その代表的な解析手法として，Single-cell RNA-sequence 解析 [1] がある．Single-cell RNA-sequence 解析を用いることで，遺伝子の発現状態に基づいた細胞型の分類や，新規の細胞種の同定，異常な細胞の推定を行うことが可能である．遺伝子発現量の解析には，Cell Ranger [2] や Seurat [3] のような信頼性の高い解析ツールを使用することが主流となっている．インフルエンザや日本脳炎，風邪といった治すことが可能な病気は，現在の解析手法で原因遺伝子を発見することが可能である．しかし，アルツハイマー型認知症やパーキンソン病といった，疾患の原因が完全に判明していない病気は難病とされており，現在の解析手法では原因遺伝子を発見することができていない．特に，アルツハイマー型認知症は高齢者に多く，認知症全体のうち 60% を占めており，重要な疾患の 1 つであると言える．このことから，現在の Single-cell RNA-sequence 解析手法に代わる新しい解析手法の需要が高まりを見せている．

新しい Single-cell RNA-sequence 解析手法として，機械学習や深層学習などの技術を取り入れた手法が注目されている [4]．その中の 1 つに，Geneformer [5] がある．Geneformer は，自然言語処理の分野で広く用いられている Attention ベースの深層学習モデル BERT [6] のように Transformer Encoder を利用した事前学習モデルである．Geneformer は，Attention 機構により遺伝子間の関係を学習する．そのため遺伝子ネットワークの構造を理解することが可能となる．転移学習により，限られた環境下でのデータを用いて，事前学習で作成した遺伝子ネットワークをファインチューニングすることで，生物学実験と同様のシミュレーション実験が可能となり，病気の原因遺伝子を特定するまでのプロセスを効率化することが可能となる．さらに，治療法の発見や新薬の開発などにもつなげることも可能となる．

Geneformer [5] は，人の単一細胞データを事前学習に使用している．そのため，人以外の単一細胞データを転移学習に使用しては，Geneformer の性能を最大限に発揮することができない．また，遺伝子操作などの生物実験は，人よりもマウスを使用するのが一般的である．そのため，単一細胞データはマウスの方が豊富であり，疾患マウスの単一細胞データの取得も容易である．このことから，マウスの単一細胞データを事前学習に使用した Genformer であるマウス版 Geneformer は，Single-cell RNA-sequence 解析を用いた疾患研究の発展には，非常に有用であると言える．

そこで，本稿では，マウス版 Geneformer の構築を目的とする．マウス版 Geneformer の構築により，マウスの単一細胞データを正確に細胞型に分類することが可能となる．また，疾患マウスモデルを使用した，生物学実験をシミュレーション実験する in silico 摂動実験が可能になる．そして，疾患マウスモデルの in silico 摂動実験により，疾患モデルがマウスしか存在しない疾患の原因遺伝子を同定することが可能となる．

本稿では次の構成をとる．2 節で 従来の Single-cell RNA-sequence 解析と新しい解析手法の研究を説明し，3 節でマウス版 Geneformer の構築について述べる．続いて，4 節で評価実験の概要および実験結果について説明し，5 節で本稿をまとめる．

## 2　関連研究

本章では，Single-cell RNA-sequence 解析，従来の解析手法

について，機械学習と深層学習を取り入れた解析手法について，Attention ベースの解析手法についてまとめる．

### 2.1 Single-cell RNA-sequence 解析

Single-cell RNA-sequence 解析 [1] とは，個々の細胞が保持している messenger RNA (mRNA) 全体を質的，量的に網羅的に調べる手法である．これは，遺伝子の発現状態に基づいた細胞型の分類を，次元削減やクラスタリングとなどの数理学的解析を用いることにより可能としている．遺伝子の発現とは，細胞核内の DNA の塩基配列に基づき，細胞ごとにさまざまな生体機能を持つタンパク質を合成し，遺伝子情報が具体的に現れることである．Single-cell RNA-sequence 解析により，今まで知り得なかった希少な細胞の検出や，細胞の遺伝子発現量の変化，病気の原因となる異常な細胞，遺伝子の抽出が可能となった．

### 2.2 従来の Single-cell RNA-sequence 解析手法

1 つの細胞が持つ整体物質を解明し，定量しようとする試みは古くからあり，1960 年代に蛍光活性化セルソーティングが開発された．蛍光活性化セルソーティングは，識別抗体などのプローブとの組み合わせにより，多数の細胞集団の中で 1 つの細胞が保持している生体分子の種類や量について断片的な研究が可能となった．そして，2006 年に次世代シーケンサーが誕生した．次世代シーケンサーとは，数千から数百万もの DNA の塩基配列を同時に解析することが可能な技術である．この技術が開発されたことにより，塩基配列単位，細胞単位での遺伝子発現量の解析が容易になった．次世代シーケンサーの誕生により，希少な mRNA やノンコーディング RNA を含めた未知の転写産物の高感度検出が可能となった．また，mRNA 前駆体から不必要な塩基配列が切り離され，mRNA が熟成していく過程など，転写産物の種類だけでなく，転写産物の構造的差異の解析も可能となった．

次世代シーケンサーにより，個々の細胞が持つ遺伝子発現パターンを獲得することが可能となった．この技術を利用した Single-cell RNA-sequence 解析として遺伝子発現パターン解析技術 Cell Ranger [2] や，Seurat [3] が誕生した．これらは，個々の細胞の遺伝子発現パターンを次元削減やクラスタリングと言った数理学的解析を用いて，細胞の分類や異常な遺伝子の抽出を行うソフトウェアである．医学界では，信頼性の高い解析ツールとして様々な解析に広く用いられている．

### 2.3 機械学習や深層学習を取り入れた解析手法

機械学習や深層学習などの情報分野の技術も急速に発展している．機械学習は，教師あり学習と教師なし学習に分類され，教師あり学習として，サポートベクタマシン（SVM）や勾配ブースティング決定木が研究された．この技術を Single-cell RNA-sequence 解析に取り入れた手法が，CaSTLe [7] である．CaSTLe が誕生した時の単一細胞データのアノテーション手法は，手動でクラスタリングを行いアノテーションをするか，蛍光活性化セルソーティングを用いてアノテーションを行っていた．しかし，これらはアノテータの知識に依存したり，別で実験が必要なため，Single-cell RNA-sequence 解析に組み込むことができなかった．この課題を解決する手法として CaSTLe が誕生した．CaSTLe は，一般でも入手可能な 9 つの単一細胞データセットを用いて事前学習したモデルである．転移学習を用いて異なる単一細胞データセットのモデルを構築し，細胞型の分類を実現した．具体的には，一変量特徴選択手法を用いて特徴量を抽出し，弱学習器として決定木を用いた XGBoost 分類機 [8] を構築する．この手法は，単一細胞データのラベリングをより効率的かつ正確に行うことが可能となり，転移学習により事前学習時には存在しなかった細胞型の分類も行うことが可能である．

深層学習では，Convolution Neural Network (CNN) [9] が研究された．CNN は，深層学習モデルの 1 つで，畳み込み層，プーリング層，全結合層で構成され，画像認識やパターン認識などに利用されている．CNN を Single-cell RNA-sequence 解析に取り入れた手法として scDeepInsight [10] が存在する．この手法は，単一細胞データのような非画像データを画像に変換し，深層学習モデルに適用する手法が注目されるようになり，単一細胞データを視覚的に理解しやすくする手法が求められたことで誕生した手法である．scDeepInsight では，はじめに単一細胞データに対してフィルタリングと正規化を実施し，主成分分析や t-SNE [11] や UMAP [12] などで 2 次元に圧縮し，2 次元に埋め込んだ単一細胞データを画像に変換する．画像に変換した単一細胞データを EfficientNet-b3 [13] などの CNN モデルを用いて，特徴を抽出し，細胞型の識別を行う．また，scDeepInsight は，DeepInsight [14] という，遺伝子発現データを 2 次元の空間に埋め込むための手法に基づいて設計されている．scDeepInsight は，画像処理の分野で活躍している CNN ベースのモデルを活用しているため，他の従来手法に比べて細胞型の分類精度が高く，希少な細胞型や未知の細胞型を検出することも可能である．

また， グラフニューラルネットワーク (GNN) [15] の研究は 2010 年代から急速に発展していった．GNN の中でも重み付きグラフニューラルネットワークを用いた手法として，scDeepSort [16] が存在する．この手法は，従来の細胞型のアノテーションで必要であったマーカー遺伝子や RNA-sequence プロファイルなどを必要とせず，容易に細胞型のアノテーションを可能にした手法である．GNN を使用する理由は，単一細胞データは細胞と遺伝子発現量の行列として表現され，各行が細胞を表し各列が遺伝子を表す．このような行列データは，潜在的にグラフ構造を持っており，GNN を使用することで細胞間の関係性を考慮した細胞型のアノテーションが可能になるためである．scDeepSort は，単一細胞データのための事前学習済みの細胞型の分類手法であり，重み付きグラフニューラルネットワークを用いた深層学習モデルである．このモデルは，埋め込み層と重みグラフ集約層と線形分類層で構築されいる．埋め込み層では細胞と遺伝子を同じ次元に圧縮を行い，重みグラフ集約層では GraphSAGE [17] をバックボーンの GNN フレームワークとして使用し，新しい細胞表現を生成する．線形分類層では，新しい細胞表現を事前定義された細胞型に分類する．この手法

は，高い細胞型の分類性能と堅牢性を示し，細胞型のアノテーション手法として新しいアプローチを示した．

深層生成モデルでは，AutoEncoder (AE) [18]，Variational AutoEncoder (VAE) [19] や Gaussian-mixture VAE (GMVAE) [20] が研究された．これらの手法は，異常検知の分野で成功を収めている．深層生成モデルの VAE や GMVAE を Single-cell RNA-sequence 解析に利用した手法として，Single-cell VAE (scVAE) [21] が存在する．従来の解析手法では単一細胞データに対して次元削減を行い，情報を抽出する作法が一般的であった．しかし，それでは単一細胞データの希少で重要な情報を失う可能性があり，解析結果に影響を与える．そこで，次元削減前の生の単一細胞データを直接使用する手法として scVAE が誕生した．scVAE は，Single-cell RNA-sequence 解析用にカスタマイズされた VAE や GMVAE を構築した．通常の VAE や GMVAE との違いとして，使用する尤度関数をポアソン分布や負の二項分布などのカウント尤度関数を使用した点である．これは，生の遺伝子発現量データが離散的でかつスパースなデータであり，連続的なデータでないためである．負の二項分布を使用した GMVAE は，潜在空間でのクラスタリングが生物学的に適した細胞型の分布となっていることを示した．また，scVAE は，単一細胞データ内に存在する Human Error により生成されたデータをネットワーク内で除去するため，ノイズとなるデータを予め除去する必要がない．そのため，結果を得るためのプロセスが簡略化され，細胞の専門家でなくても解析が可能となった．

### 2.4 Geneformer

自己教師あり学習は，近年注目を集めている研究分野である．自己教師あり学習では，データから正解ラベルを自動で生成できるプレテキストタスクを用いて事前学習を行い，様々な下流タスクにファインチューニングすることで，より良い特徴量を抽出することが可能となる．そして，2018 年に自然言語処理の分野で Attention ベースの事前学習モデル BERT [6] が発表された．BERT は，Transformer Encoder [22] を使用した構造となっている．この Transformer Encoder を Single-cell RNA-sequence 解析に利用した手法が Geneformer [5] である．この手法は，遺伝子ネットワークのマッピングに焦点を当て，遺伝子間の関係を学習する手法である．Geneformer は，大規模な人の単一細胞データをテキストとして表現し，そのテキストデータを用いて事前学習した Attention ベースの深層学習モデルである．Geneformer は，Transformer Encoder [22] を使用した構造となっている．そのため，Attention 機構により遺伝子間の関係性の学習が可能となり，人の細胞の遺伝子ネットワークを正確に予測することが可能となる．正確に遺伝子間の関係を学習するために，遺伝子のトークン化には，細胞の状態を示す遺伝子を積極的にトークンとして使用する Rank Value Encoding を使用している．Geneformer は，転移学習することで限られた単一細胞データの遺伝子ネットワークを予測することが可能となり，従来の解析手法よりも細胞型の分類精度が向上することを示した．また，実際の生物学的実験をシミュレートすることが可能となる in silico 摂動実験が可能となり，実際の人の心筋症モデル [23] を用いて，心筋症の原因遺伝子を同定するなどして，有効性を示した．

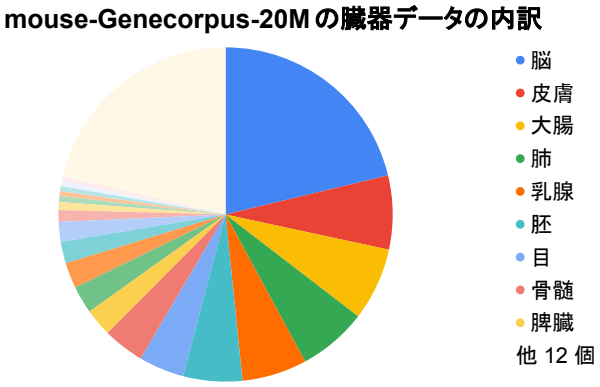


図 1　mouse-Genecorpus-20M の臓器データの内訳

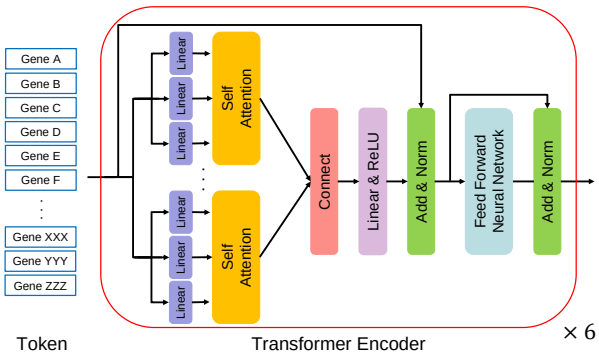


図 2　マウス版 Geneformer のネットワーク図

## 3　マウス版 Geneformer

本章では，マウス版 Geneformer の事前学習に使用したデータセットである mouse-Genecorpus-20M やアーキテクチャ，マウス版 Geneformer の事前学習について述べる．

### 3.1 mouse-Genecorpus-20M

mouse-Genecorpus-20M は，マウス版 Geneformer を事前学習するときに使用するデータセットである．このデータセットは，一般に公開されているマウスの Single-cell RNA-sequence データベースから，正常なマウスの全身の生の単一細胞データを約 2,300 万個取得し，構築している．mouse-Genecorpus-20M の構築に使用したマウスの臓器の内訳を図 1 に示す．また，主に使用した Single-cell RNA-sequence データベースを以下に示す．

- PanglaoDB
- Single Cell Expression Atlas
- Single Cell Portal
- ENCODE project
- 10x Genomics

生の単一細胞データは遺伝子発現量マトリックスとして表現されている．遺伝子発現量マトリックスは，生物学や遺伝学

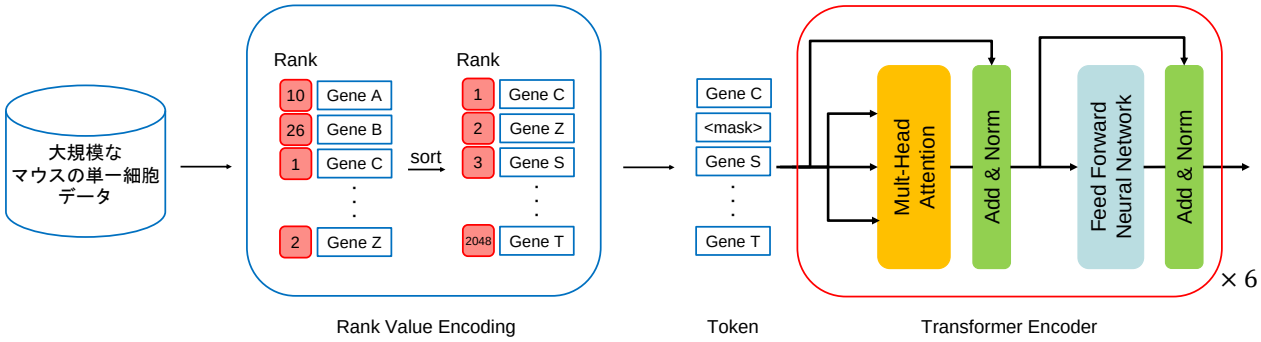


図 3　マウス版 Geneformer のアーキテクチャ

の研究において重要なデータ構造の一つである．行方向に異なる遺伝子を，列方向に異なる細胞を表し，対応する細胞における遺伝子の発現量が数値として記載されている．mouse-Genecorpus-20M は，正常なマウスの遺伝子ネットワークを学習させるための単一細胞データのコーパスなため，癌細胞などの遺伝子ネットワークを再編する原因となる単一細胞データは除外する必要がある．

### 3.2　アーキテクチャ

マウス版 Geneformer は，大規模なマウスの単一細胞データを用いて事前学習した，Attention ベースの深層学習モデルである．マウス版 Geneformer のネットワーク図を図 2 に示す．転移学習をすることで，希少な細胞や遺伝子が存在するデータの遺伝子ネットワークを正確に予測することが可能である．マウス版 Geneformer は，Attention 層とフィードフォワードニューラルネットワーク層を持つ Transformer Encoder を 6 つ使用した構成となっている．Attention Head の数は 4 つである．トークンには，細胞状態を正確に表現する 2,048 個の遺伝子を使用している．よって，1 層目の Transformer Encoder の入力次元数は 2,048 次元である．出力次元数で 256 次元であり，2 層目以降の Transformer Encoder の入力と出力の次元数は 256 次元である．

### 3.3　事 前 学 習

マウス版 Genformer の概要図を図 3 に示す．マウス版 Geneformer の事前学習には，大規模なマウスの単一細胞データを用いる必要がある．そのため，はじめに，mouse-Genecorpus-20M を作成した．mouse-Genecorpus-20M は，癌細胞などの遺伝子ネットワークを再編する原因となる単一細胞データは除外する必要がある．

次に，mouse-Genecorpus-20M 内のデータをフィルタリングする．フィルタリングは，細胞に由来しない RNA や死細胞，複数細胞が混じった細胞（Doublet），空の液滴等で取得したデータを除外するために行う．これらの単一細胞データは必ず存在する．これらの細胞データは，Single-cell RNA-sequence 解析においてノイズとなるデータであり，細胞状態を正確に表現する特徴量の獲得を困難にする．これらから，フィルタリングを行う必要がある．具体的なフィルタリング条件を以下に示す．

- 各データセットにおいて，遺伝子の総発現量の平均から 3 標準偏差以外の細胞を除去
- 各データセットにおいて，ミトコンドリア RNA の遺伝子発現量の平均から 3 標準偏差以外の細胞を除去
- 各データセットにおいて，1 細胞で検出する遺伝子数が 7 未満の細胞を除去
- 各細胞の遺伝子発現総数が 20,000 超過の細胞を除去

フィルタリング後の単一細胞データ数は，22,446,161 細胞となり，この単一細胞データを事前学習に使用する．

次に，フィルタリング後の mouse-Genecorpus-20M からトークンを作成する．トークンには，細胞の状態を示す遺伝子を使用する．トークン化には，トークンとして使用する遺伝子を抽出する Rank Value Encoding を使用する．Rank Value Encoding は，mouse-Genecorpus-20M 内での遺伝子発現量に基づいて，各細胞内に存在する遺伝子に対してランク付けを行う．このトークン化手法により，細胞状態を区別するための遺伝子を優先的に選択することが可能となる．具体的な手順を以下に示す．

1. mouse-Genecorpus-20M の全細胞の非ゼロ遺伝子発現量から，各遺伝子の非ゼロ中央値を計算する
2. 各細胞の各遺伝子発現量を，その細胞の総遺伝子発現量で除算し，正規化をする
3. 正規化した遺伝子発現量を，その遺伝子の非ゼロ中央値で除算し，さらに正規化をする
4. 正規化した遺伝子発現量を 10,000 倍し，大きい順にソートし，ランク値を設定する

そして，ランク値 2,048 までの遺伝子が細胞の状態を示す遺伝子であり，トークンとして使用される．ランク値は 1 に近づくにつれて細胞状態をよく示す．

作成した mouse-Genecorpus-20M のトークンを用いて Geneformer を学習させる．学習に使用するプレテキストタスクには，マスクトークンの予測タスクを使用する．このタスクは，入力されたトークンの一部をランダムにマスクし，そのマスクトークンを他のトークンを用いて予測するタスクである．このタスクによりマウス版 Geneformer は，マウスの遺伝子の関連性や発現パターンを学習することが可能であると考える．事前学習では，各細胞のトークンをランダムに 15% の確率でマスクし，残りのマスクされていないトークンを用いてマスク

したトークンを予測させるようにネットワークの最適化を行う．

## 4 評価実験

本章では，マウス版 Geneformer の有効性と頑健性を示すための評価実験を行う．

### 4.1 実験概要

構築したマウス版 Geneformer の評価するための実験を以下に 3 つ示し，実施する．

- マウス版 Geneformer と従来手法による細胞型の分類実験
- 事前学習の有無による細胞型の分類実験
- マウス版 Geneformer による in silico 摂動実験

マウス版 Geneformer と従来手法による細胞型の分類実験では，マウスの多様な臓器の単一細胞データを使用し，マウス版 Geneformer が従来手法よりも分類精度が向上するかを評価する．事前学習の有無による細胞型の分類実験では，マウス版 Geneformer の事前学習が意味のあるものなのかをマウスの胚，腎臓，尿道と前立腺が混ざった単一細胞データの細胞型を分類することにより評価をする．これらの評価には，Accuracy を用いて評価を行う．また，細胞分布を可視化することで，マウス版 Geneformer の性能の分析や，使用するデータの分析を行う．マウス版 Geneformer による in silico 摂動実験では，生物学実験である in vivo 実験で同定した遺伝子が in silico 摂動実験で検出することが可能かを評価する．in vivo 実験とは，人やマウスなどの生物の体内で，医薬品の効果や変異させた遺伝子の影響，生物学的な遺伝子発現のプロセスの調査，評価を行うための実験のことである．

### 4.2 in silico 摂動実験の概要

in silico 摂動実験とは，コンピュータ上で行う遺伝子の機能を解析する手法の 1 つである．遺伝子の欠失や活性化などの変異をシミュレーションすることで，その遺伝子が遺伝子ネットワークにどのような影響を与えるかを予測する．in silico 摂動実験では，はじめに遺伝子発現量データを解析し，各遺伝子の発現量を Rank Value Encoding を用いてランク付けを行う．次に，遺伝子をランダムに何回も摂動させる．遺伝子を欠失させる場合は，その遺伝子のランク値を下げ，他の遺伝子のランク値を上げる操作を行う．一方，遺伝子を活性化させる場合は，その遺伝子のランク値を上げ，他の遺伝子のランク値を下げる操作を行う．遺伝子の摂動により，摂動後の細胞状態が特定の細胞状態に近づいた時，摂動した遺伝子は重要な遺伝子であることが分かる．このとき，コサイン類似度を用いて摂動後の細胞状態が特定の細胞状態に近づいたかを数値的に評価する．

### 4.3 実験条件

マウス版 Geneformer の事前学習を mouse-Genecorpus-20M を用いて行う．事前学習に使用した条件を表 1 に示す．mouse-Genecorpus-20M を Geneformer に学習させるために，我々が使用した GPU は，32 GB の NVIDIA V100 を 8 枚使用し，約 2 日間学習を行った．

表 1　事前学習に使用したパラメータ

| 実験条件 | パラメータ |
|---|---|
| 入力最大次元数 | 2,048 |
| Transformer Encoder の層数 | 6 |
| attention head 数 | 4 |
| 埋め込み次元数 | 256 |
| 活性化関数 | ReLU |
| Drop out 率 | 0.02 |
| Epoch 数 | 20 |
| 最大学習率 | 1e-3 |
| Warm up steps | 10,000 step |
| スケジューラ | cosine スケジューラ |
| バッチサイズ | 12 |
| 最適化手法 | AdamW |
| Weight decay | 1e-3 |

表 2　ファインチューニングに使用したパラメータ

| 実験条件 | パラメータ |
|---|---|
| 入力最大次元数 | 2,048 |
| 固定する層 | 0 |
| Epoch 数 | 10 |
| 最大学習率 | 5e-5 |
| Warm up steps | 500 step |
| スケジューラ | cosine スケジューラ |
| バッチサイズ | 12 |
| 最適化手法 | AdamW |
| Weight decay | 1e-3 |

マウス版 Geneformer と従来手法による細胞型の分類実験は，テキスト分類タスクを用いて細胞型の分類タスクへファインチューニングを行う．使用するデータは，マウスの舌，胸腺，乳腺，大腸，四肢の筋肉，脾臓，心臓，脳，腎臓の 9ヶ所の臓器の単一細胞データである．これらの臓器データを学習用とテスト用の 2 つに 80%，20% の確率でランダムに分割する．その他のファインチューニング条件を表 2 に示す．

事前学習の有無による細胞型の分類実験は，テキスト分類タスクを用いて細胞型の分類タスクへファインチューニングを行う．使用するデータは，マウスの尿道と前立腺が混ざった単一細胞データ，胚，腎臓の単一細胞データである．胚は，動物や人間の発生学において受精卵のことを指し，様々な臓器の細胞型に変化する細胞が存在する．

このことから，胚の細胞は，細胞型の特徴を完全に保持していないと考えられる．尿道と前立腺が混ざったデータは，2 種類の臓器が混ざったデータである．これらから，マウス版 Geneformer と従来手法による細胞型の分類実験で使用する臓器データよりも，細胞型を分類するのが困難な臓器データと考えられる．これらの臓器データを学習用とテスト用の 2 つに 80%，20% の確率でランダムに分割する．その他のファインチューニング条件は，表 2 の条件を使用する．

in silico 摂動実験は，テキスト分類タスクを用いて疾患型の分類タスクへファインチューニングを行う．使用するデータは，糖尿病性腎臓病と UMOD 腎臓病と正常な腎臓のデータ [24]，

表 3　マウス版 Geneformer と従来手法の分類精度の比較

| 臓器 | 細胞種数 | マウス版 Geneformer | scDeepSort | scVAE |
|---|---|---|---|---|
| 舌 | 3 | 93.08 | 76.69 | 80.44 |
| 胸腺 | 6 | 93.47 | 54.94 | 74.95 |
| 乳腺 | 7 | 98.16 | 47.76 | 74.25 |
| 大腸 | 7 | 90.91 | 49.78 | 59.00 |
| 四肢の筋肉 | 9 | 99.02 | 90.82 | 79.58 |
| 脾臓 | 10 | 97.83 | 81.01 | 76.47 |
| 心臓 | 11 | 95.55 | 79.55 | 79.42 |
| 脳 | 15 | 91.95 | 58.46 | 76.19 |
| 腎臓 | 18 | 92.82 | 58.01 | 56.25 |

タンパク質である COP1 をノックアウトしたデータと正常なデータ [25] である．前者のデータは，腎臓組織内の細胞型や疾患に特異的な遺伝子発現の変化を解析するためのデータである．後者のデータは，神経炎症を抑える役割のあるタンパク質 COP1 をノックアウトしたマイクログリアにおけるタンパク質 c/EBP$\beta$ を解析するためのデータである．これらのデータを学習用とテスト用の 2 つに 80%，20% の確率でランダムに分割する．その他のファインチューニング条件は，表 2 の条件を使用し，疾患型の分類精度が 90.00 ポイント以上となったモデルを in silico 摂動実験に使用する．糖尿病性腎臓病の in silico 摂動実験では，正常な腎臓細胞を糖尿病性腎臓病の細胞に近づける実験を行う．UMOD 腎臓病の実験では，UMOD 腎臓病の細胞を正常な腎臓細胞に近づける実験を行う．COP1 ノックアウトの実験では，COP1 をノックアウトした細胞を正常な細胞に近づける実験を行う．

### 4.4　細胞型の分類結果

構築したマウス版 Geneformer と，従来手法である scDeepSort と scVAE の細胞型の分類精度の比較結果を表 3 に示す．表 3 から，マウス版 Geneformer は，従来手法である scDeepSort や scVAE よりも細胞型の分類精度が全てのマウスの臓器において大幅に向上していることが分かる．このことから，マウス版 Geneformer は，細胞型の分類タスクにおいて有効であると考える．そして，マウス版 Geneformer は，細胞型の種類数に依存することなく分類精度が 90 ポイントを超えている．従来手法では，細胞型の種類数ごとに分類精度が変動している．これらから，マウス版 Geneformer は，分類するクラス数に頑健であると言える．

また，scDeepSort や scVAE は，マウスの臓器ごとで細胞型の分類精度にばらつきがある．scDeepSort では最大 43.06 ポイントの分類精度差が存在し，scVAE では最大 24.19 ポイントの分類精度差が存在する．一方，マウス版 Geneformer は，マウスの臓器ごとで細胞型の分類精度のばらつきが抑えられ，最大分類精度差は 8.11 ポイントである．このことから，マウス版 Geneformer は，従来解析手法よりもマウスの臓器間の差に柔軟に対応することが可能であり，頑健であると言える．

さらに，舌と四肢の筋肉のデータにおいて細胞分布の可視化を行った．その結果を図 4 と図 5 に示す．舌の細胞分布の可視

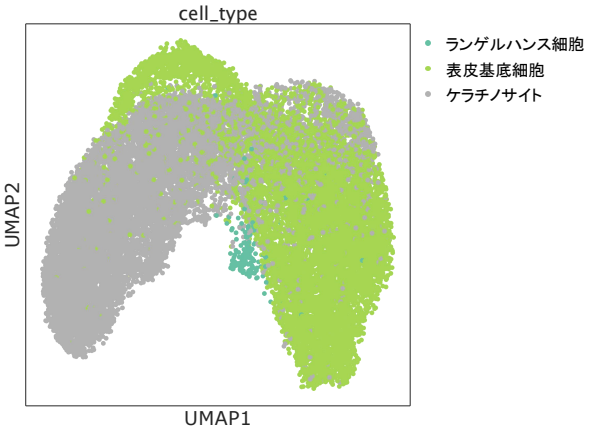


図 4　舌の細胞分布の可視化結果

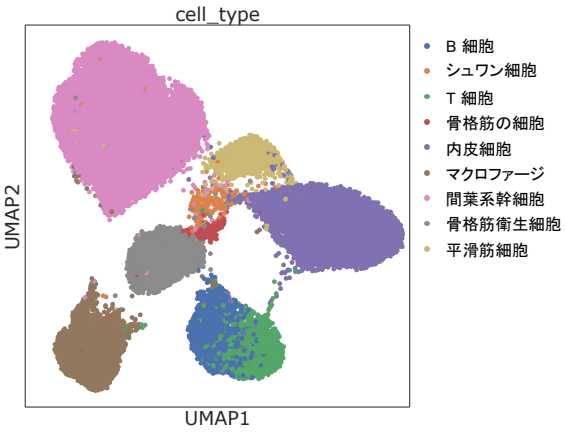


図 5　四肢の筋肉の細胞分布の可視化結果

化結果に注目すると，ケラチノサイトと表皮基地細胞が大まかに細胞型ごとに分かれていることが分かる．舌のデータの分類精度が 93.08 ポイントであることを考慮すると，細胞型を正確に予測している細胞はケラチノサイトと表皮基地細胞であると考える．一方，ランゲルハンス細胞が表皮基底細胞と被っている．これは，ランゲルハンス細胞と表皮基底細胞が非常に似た特徴量を持っていると考えられる．以上のことを踏まえるとマウス版 Geneformer は，細胞型の分布が重なるほど非常に似た特徴量を持つ細胞型の遺伝子ネットワークの予測は，困難であると考える．

次に，四肢の筋肉の細胞の可視化結果に注目する．こちらの細胞は，舌の細胞よりも細胞型に分けることができていることが分かる．また，シュワン細胞と骨格筋の細胞の分布が隣接し，B 細胞と T 細胞の分布が隣接し，細胞型の分類に困難と思われた．しかし，四肢の筋肉のデータのマウス版 Geneformer による分類精度が 99.02 ポイントであり，ほとんどの細胞を分類することが可能となっている．これらから，マウス版 Geneformer は，隣接している細胞型の特徴も詳細に捉える遺伝子ネットワークを予測することが可能であると考える．

### 4.5　事前学習の有無による細胞型の分類結果

事前学習ありのマウス版 Geneformer と事前学習なしのマウス版 Geneformer の細胞型の分類結果を表 4 に示す．表 4 から，事前学習したマウス版 Geneformer の分類精度は，事前学

表 4　事前学習の有無による分類精度の比較

| 臓器 | 細胞種数 | 事前学習あり | 事前学習なし |
|---|---|---|---|
| 尿道と前立腺 | 7 | 93.98 | 92.81 |
| 胚 | 9 | 77.22 | 69.78 |
| 腎臓 | 10 | 77.96 | 68.56 |

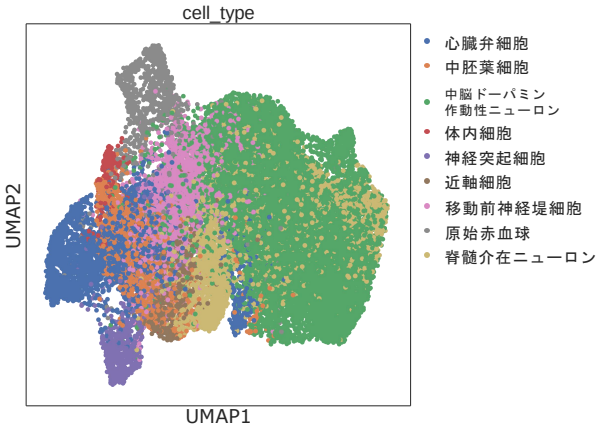


図 6　胚の細胞分布の可視化結果

習していないマウス版 Geneformer の分類精度よりもすべてのデータで高いことが分かる．最大の分類精度差は腎臓のデータで 9.40 ポイントで，最小の分類精度差は尿道と前立腺が混ざったデータで 1.17 ポイントである．このことから，事前学習モデルを使用することは，細胞型の分類を行うことが困難なデータになるにつれて効果的であると言える．よって，事前学習モデルの使用は，すべてのマウスの単一細胞データにおいて事前学習していないモデルよりも効率的に学習することが可能となり，分類精度の向上につながると考える．

さらに，胚のデータにおいて細胞分布の可視化を行った．その結果を図 6 に示す．胚の細胞分布の可視化結果から，四肢の筋肉の細胞分布のように細胞型がきれいに分布せず，全体的に 1ヶ所に集中したような分布となっていることが分かる．さらに，心臓弁細胞は 2ヶ所に分布し，中胚葉細胞や近軸細胞，脊髄介在ニューロンは，別の細胞型だが 1ヶ所に分布していることが分かる．これらから，胚のデータは細胞型の特徴を完全に保持していないと考えられる．また，胚のデータの事前学習ありの分類精度が 77.22 ポイントで事前学習なしの分類精度が 69.78 ポイントであることを考慮すると，マウス版 Geneformer は，事前学習を行うことで細胞型の特徴を完全に保持していないデータに対しても，事前学習で得た知識を使用して遺伝子ネットワークを予測することが可能であると考える．

これらから，マウス版 Geneformer は，事前学習を行うことで，多様なデータの遺伝子ネットワークを予測することが可能になると考える．

### 4.6　in silico 摂動実験の結果

#### 4.6.1　in silico 摂動実験による糖尿病性腎臓病の結果

糖尿病性腎臓病の in silico 摂動実験の結果を示す．正常な腎臓細胞に存在する遺伝子 Slc12a3 を欠失させた時，糖尿病性腎臓病の細胞に最も近づくことが分かった．生物実験である in vivo 実験では，糖尿病性腎臓病にかかると顆粒細胞と黄斑細胞で構成する糸球体装置に変化があり，その顆粒細胞には遺伝子 Ren1 が存在し遺伝子 Slc12a3 が存在しないと同定している [24]．このことから，糖尿病性腎臓病における in silico 摂動実験は成功し，糖尿病性腎臓病で変化した遺伝子を特定したと考える．

#### 4.6.2　in silico 摂動実験による UMOD 腎臓病の結果

UMOD 腎臓病の in silico 摂動実験の結果を示す．UMOD 腎臓病の細胞に存在する遺伝子 Slc35b1 を欠失させた時，正常な腎臓細胞に最も近づくことが分かった．in vivo 実験では，腎臓における変異型 UMOD タンパク質の蓄積により，Unfolded Protein Response (UPR) が変化し，UPR を活性化させる遺伝子 Slc35b1 と Slc3a2 の遺伝子の発現量が変化することが分かっている [24]．このことから，UMOD 腎臓病における in silico 摂動実験では，2 つの遺伝子のうち 1 つを特定したと考える．

#### 4.6.3　in silico 摂動実験による COP1 ノックアウトの結果

COP1 をノックアウトしたミクログリアの in silico 摂動実験の結果を示す．COP1 をノックアウトした細胞に存在する遺伝子 Apoe を欠失させた時，正常な細胞に最も近づくことが分かった．遺伝子 Apoe を欠失させたときのコサイン類似度は 0.36 であった．また，遺伝子 Cebpb，Clec7a，Cst7 の遺伝子の欠失も正常な細胞に近づくことが分かった．in vivo 実験では，COP1 をノックアウトすることにより c/EBP$\beta$ が脳内に溜まることにより，神経炎症が起こることが言われている．また，COP1 のノックアウトにより遺伝子 Apoe と Cebpb の発現量が増加し，神経変性関連遺伝子 Clec7a，Cst7 の発現量も変化することが分かっている [25]．遺伝子 Cebpb はタンパク質 c/EBP$\beta$ を発現する遺伝子であり，遺伝子 Apoe は神経変性疾患であるアルツハイマー型認知症の危険因子と言われている遺伝子である．これらから，in silico 摂動実験は成功し，in vivo 実験を再現したと考える．

#### 4.6.4　細胞分布の可視化結果

in silico 摂動実験で使用したデータの細胞分布の可視化結果を図 7，図 8 に示す．これらを見ると，正常な細胞と異常な細胞が別れて分類していることが分かる．特に，図 8 は正常な細胞である Cop1 WT と COP1 をノックアウトした細胞である Cop1 KO が完全に分離している．これらから，マウス版 Geneformer は，ファインチューニングしたことにより，疾患データ特有の特徴を詳細に捉えた遺伝子ネットワークを予測することが可能であると考える．

## 5　おわりに

本稿では，大規模なマウスの単一細胞データを用いて事前学習した，Attention ベースの深層学習モデルであるマウス版 Geneformer を構築した．マウス版 Geneformer と従来手法による細胞型の分類実験では，マウスの 9 つの臓器データを使用し，従来手法である scDeepSort，scVAE よりもマウス版 Geneformer の分類精度が向上していることを確認した．また，

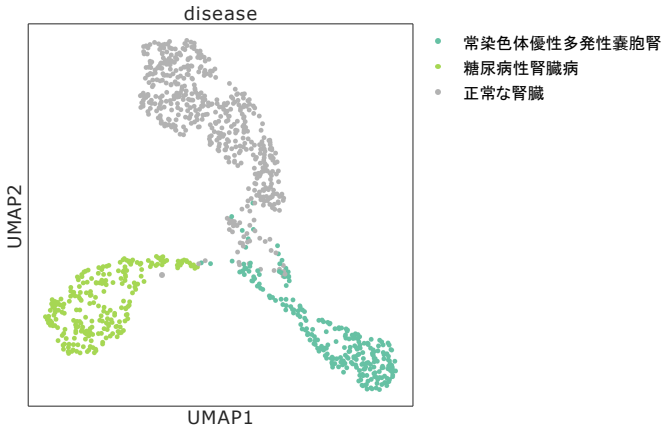


図 7　腎臓データの疾患分布の可視化結果

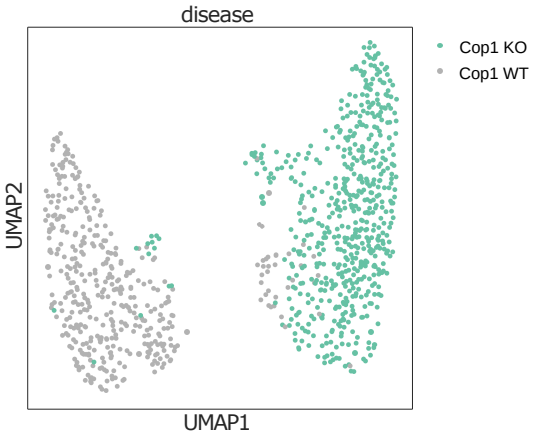


図 8　COP1 データの疾患分布の可視化結果

マウスの 9 つの臓器データで分類精度が一貫して 90 ポイントを超えていることから，マウス版 Geneformer は，マウスの臓器に依存しないことと，細胞型の種類数に依存しないことを確認した．これらから，マウス版 Geneformer の有効性と頑健性を示した．

また，マウス版 Geneformer の事前学習の有無による細胞型の分類実験では，2 つの臓器が混ざったデータと胚のデータ，腎臓のデータを使用し，事前学習していないマウス版 Geneformer よりも事前学習したマウス版 Geneformer の方が分類精度が向上することを確認した．また，胚の細胞分布の可視化結果から，事前学習を行うことで細胞型の特徴を完全に保持していないデータに対しても，事前学習で得た知識を使用して遺伝子ネットワークを予測することが可能であることを確認した．これらから，マウス版 Geneformer の事前学習の有効性を示した．

最後に，マウス版 Geneformer を使用した in silico 摂動実験では，糖尿病性腎臓病モデルと UMOD 腎臓病モデルと COP1 ノックアウトモデルを使用し，in vivo 実験を再現した．このことから，マウス版 Geneformer による in silico 摂動実験は，in vivo 実験と同等の結果を得られることを示した．また，疾患モデルの細胞分布の可視化結果から，正常な細胞と異常な細胞で別れて分布していることを確認した．これらから，マウス版 Geneformer による in silico 摂動実験の有効性を示した．

今後の展望として，Single-cell RNA-sequence データに特化したプレテキストタスクの開発や，本稿で提案したモデルを実際のマウスの疾患モデルに適用することが挙げられる．

## 文　献

[1] Saliba, Antoine-Emmanuel, et al. ”Single-cell RNA-seq: advances and future challenges.” Nucleic acids research 42.14 (2014): 8845-8860.

[2] Lepetit, Maxime et al. “scAN1.0: A reproducible and standardized pipeline for processing 10X single cell RNAseq data.” In silico biology vol. 15,1-2 (2023): 11-21. doi:10.3233/ISB-220252

[3] Gribov, Alexander, et al. ”SEURAT: visual analytics for the integrated analysis of microarray data.” BMC medical genomics 3 (2010): 1-6.

[4] Jiang, Peng, et al. ”MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features.” Nucleic acids research 35.suppl_2 (2007): W339-W344.

[5] Theodoris, Christina V., et al. ”Transfer learning enables predictions in network biology.” Nature (2023): 1-9.

[6] Devlin, Jacob, et al. ”Bert: Pre-training of deep bidirectional transformers for language understanding.” arXiv preprint arXiv:1810.04805 (2018).

[7] Lieberman, Yuval, Lior Rokach, and Tal Shay. ”CaSTLe–classification of single cells by transfer learning: harnessing the power of publicly available single cell RNA sequencing experiments to annotate new experiments.” PloS one 13.10 (2018): e0205499.

[8] Chen, Tianqi, and Carlos Guestrin. ”Xgboost: A scalable tree boosting system.” Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016.

[9] LeCun, Yann, et al. ”Gradient-based learning applied to document recognition.” Proceedings of the IEEE 86.11 (1998): 2278-2324.

[10] Jia, Shangru, et al. ”scDeepInsight: a supervised cell-type identification method for scRNA-seq data with deep learning.” Briefings in Bioinformatics 24.5 (2023): bbad266.

[11] Van der Maaten, Laurens, and Geoffrey Hinton. ”Visualizing data using t-SNE.” Journal of machine learning research 9.11 (2008).

[12] McInnes, Leland, John Healy, and James Melville. ”Umap: Uniform manifold approximation and projection for dimension reduction.” arXiv preprint arXiv:1802.03426 (2018).

[13] Tan, Mingxing, and Quoc Le. ”Efficientnet: Rethinking model scaling for convolutional neural networks.” International conference on machine learning. PMLR, 2019.

[14] Sharma, Alok, et al. ”DeepInsight: A methodology to transform a non-image data to an image for convolution neural network architecture.” Scientific reports 9.1 (2019): 11399.

[15] Kipf, Thomas N., and Max Welling. ”Semi-supervised classification with graph convolutional networks.” arXiv preprint arXiv:1609.02907 (2016).

[16] Shao, Xin, et al. ”scDeepSort: a pre-trained cell-type annotation method for single-cell transcriptomics using deep learning with a weighted graph neural network.” Nucleic acids research 49.21 (2021): e122-e122.

[17] Hamilton, Will, Zhitao Ying, and Jure Leskovec. ”Inductive representation learning on large graphs.” Advances in neural information processing systems 30 (2017).

[18] Hinton, Geoffrey E., and Ruslan R. Salakhutdinov. ”Reducing the dimensionality of data with neural networks.” science 313.5786 (2006): 504-507.

[19] Kingma, Diederik P., and Max Welling. ”Auto-encoding variational bayes.” arXiv preprint arXiv:1312.6114 (2013).

[20] Dilokthanakul, Nat, et al. ”Deep unsupervised clustering with gaussian mixture variational autoencoders.” arXiv preprint arXiv:1611.02648 (2016).

[21] Grønbech, Christopher Heje, et al. "scVAE: variational auto-encoders for single-cell gene expression data." Bioinformatics 36.16 (2020): 4415-4422.
[22] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).
[23] Chaffin, Mark, et al. "Single-nucleus profiling of human dilated and hypertrophic cardiomyopathy." Nature 608.7921 (2022): 174-180.
[24] Marshall, Jamie L., et al. "High-resolution Slide-seqV2 spatial transcriptomics enables discovery of disease-specific cell neighborhoods and pathways." Iscience 25.4 (2022).
[25] Ndoja, Ada, et al. "Ubiquitin ligase COP1 suppresses neuroinflammation by degrading c/EBP β in microglia." Cell 182.5 (2020): 1156-1169.