Strategies for Fine-Tuning Geneformer to Predict the Exposure Level of Cancer Cells to Treatments

A Comparison of Different Fine-Tuning Strategies for Foundation Models

Octavian-Teodor Dragon[1]

Supervisor(s): Prof.dr.ir. Marcel J.T. Reinders[1]**, Niek Brouwer**[1]

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2024

Name of the student: Octavian-Teodor Dragon
Final project course: CSE3000 Research Project
Thesis committee: Prof.dr.ir. Marcel J.T. Reinders, Niek Brouwer, Dr. Nezihe Merve Gürel

Abstract

Studying the interactions of genes within a cell is an area of significant interest in the field of medicine as it can provide answers to what exactly drives the behaviour of a cell under specific circumstances, such as diseases. Once understood, gene interactions can enable the synthesis of efficient, possibly personalized treatments for these diseases [1] and other disorders. However, studying gene interactions requires a large amount of samples which might be costly and laborious to obtain in the case of rare disorders for which there is not much recorded data. Geneformer, a ”context-aware, attention-based deep-learning model” [2], was created specifically for solving this problem. The model makes use of transfer learning to apply any relevant knowledge gained from a larger, similar domain onto a downstream domain with limited data which can be used to further train the model. In this paper, we assessed four fine-tuning strategies, including the one used throughout the in silico experiments presented in the original Geneformer paper. We did this to assess whether the accuracy of Geneformer on the downstream task of predicting the sensitivity of cancer cells to different treatments can be improved versus the default implementation as found within the model's paper. The model was firstly fine-tuned using a training dataset compiled from the sciplex2 dataset [3], followed by the prediction of the dosage levels to which samples from a test set were exposed. Upon performing the experiment, we concluded that, depending on the way in which knowledge from the source domain is stored inside the pre-trained model and the similarity between the source and the target domains, different fine-tuning strategies were suitable for a given task. Hence, there is no single optimal fine-tuning method which can be used to predict the level to which cancer cells were exposed to treatments such as nutlin-3A.

1 Introduction

Exploring how the interaction of genes inside a cell is influenced by the state of the cell is a valuable topic for medical research. It can provide a clear explanation to the behaviour of a cell under specific circumstances, such as diseases [4]. Understanding these interactions is important as it can enable (1) creating efficient, possibly personalized treatments for these diseases [1] and other disorders, and (2) assessing the effects of gene-drug-drug interactions over the human body. These applications are important from both a medical and social standpoint, especially when considering that adverse drug reactions have previously been reported to represent the fourth leading cause of death in the US [5]. However, studying these interactions requires a large amount of samples (i.e. both normal and diseased cells & tissues) which might not be possible to obtain in the case of rare disorders [2]. Existing works [2; 6; 7; 8; 9; 10] attempted to solve this problem through transfer learning, a method for learning from a larger corpus of general data and using any relevant information to solve a downstream task which has limited data of its own [2; 7].

Transfer learning is a versatile approach which can be used to apply knowledge from a domain with large amounts of data to a similar domain with limited amounts of data [6; 7]. Thanks to its ability to reduce the reliance on extensive labelled medical datasets that are costly and time-consuming to gather, it has seen a successful use in domains such as medical imaging, where it was able to provide decision support [6; 8; 9]. However, transfer learning should not be regarded as a universal solution, since significant differences between the target and the source domains can lead to a decrease in the overall task performance because of negative transfer [6; 11; 10; 12]. There are other challenges to consider too: pairing the model with an architecture which is unsuitable for the task at hand (i.e. too simple or too complex), the presence of distribution shifts [11] (i.e. data distributions in the source and the target domain being different [13], even though the domains themselves might be related), and improperly tuned hyper-parameters [14]. A part of the aforementioned issues (i.e. the task-model mismatch) can be avoided by ensuring that the environment used for an experiment complements the type of the task and domain. This is done through trying multiple environment configurations [14].

There are different techniques for transfer learning which can be grouped into multiple categories, two of which being: (1) iterative approaches, in which the parameter space is explored using a heuristic function and the model layers are be selectively fine-tuned; these approaches are effective but require a precise definition of the search space and are computationally expensive [6]; and (2) non-iterative approaches, in which the parameters are adjusted on-the-go, without exhaustive exploration [6]; they are less computationally demanding.

Geneformer, a ”context-aware, attention-based deep-learning model” [2], was created specifically for solving the problem of limited data in the context of mapping gene networks. The model has already shown promising results by accurately predicting how a human cardiomyocyte's gene netowrk reacts to the deletion or activation of specific genes [2].

A prerequisite to seeing the adoption of Geneformer into the medical field is optimizing it such that it can be used with great confidence in real-life scenarios [15; 16]. One way of doing so is coming up with a fine-tuning strategy such that the predictive accuracy is increased for a desired task. Better performance would promote human-AI collaboration through an increased level of trust.

The existing works so far have focused on the ways in which different strategies of implementing transfer learning and fine-tuning can increase or decrease the accuracy of the model for tasks such as medical image classification [6; 8; 9], surgical workflow analysis [17; 18], object identification [19; 20], and segmentation [21]. Therefore, in this study we focused our attention on whether the fine-tuning strategies mentioned in other works also yield notable performance improvements in the context of predicting the sensitivity of human lung adenocarcinoma cells to a treatment by looking at non-iterative fine-tuning approaches.

In order to answer this question we will look at the following key aspects:

- Given the same amount of data, which fine-tuning strategy performs the best? Can we do better than the strategy presented in Geneformer's manuscript [2]?
- Are there any differences among the fine-tuning strategies which seem to be particularly contributing towards a higher accuracy for the task of predicting the sensitivity of cancer cells to treatments? Can we gain any insights from them?

In our study, we compared four different fine-tuning strategies by starting from the same pre-trained Geneformer model [22] and the sciplex2 dataset [3] (data accessible at NCBI GEO database [23], accession GSM4150377). We also reproduced one of the classification experiments presented in the Geneformer's manuscript [2] in order to establish a baseline metric. The fine-tuning strategies evaluated in this paper are: (1) freezing the first two layers of the model (the implementation showcased in Geneformer's manuscript [2]), (2) Linear Probing [6; 24], (3) Gradual Unfreeze (last/all)(LP-FT) (gradually unfreezing the pre-trained layers of Geneformer during the fine-tuning process) [6; 24], and (4) Full Fine-tuning [6]. The methods have shown potential in the context of image classification [6].

We ran two classification experiments involving human adenocarcinoma cells [3]. In the first experiment, the samples have been exposed to Nutlin-3A at two different dosage levels, those being 0µM and 25µM. In the second experiment, four dosage levels were used: 0.25µM, 2.5µM, 25µM, and 125µM respectively. The pre-trained Geneformer model [2] was fine-tuned on samples from the task-specific dataset using each of the strategies discussed. Subsequently, it was tasked with predicting the dosage levels to which the samples in the test set were exposed.

In Section 2 we present the problem in more detail, as well as discussing the methodology used for conducting the experiments. We provide a summary of the dataset used, the selected fine-tuning strategies, and the implications of the experiment. Section 3 elaborates on the evaluated fine-tuning strategies in more detail: their unique features, as well as advantages and weaknesses that inherently come with using each method. The results are presented in Section 4, followed by Section 5 which discusses the implications of the experiment results. The conclusions and further work are specified in Section 6, while Section 7 shows the ways in which we ensured that our conclusions are creditable and representative of a responsible research process.

2 Problem Description

On an abstract level, the main problem lies within finding an efficient way to transfer knowledge from a source to a target domain. Thus, the focus of this paper is assessing how using different fine-tuning strategies impacts the prediction of the dosage levels at which human lung adenocarcinoma cells which have been exposed to the compounds nutlin-3A and the DMSO vehicle control [23; 3]. The dosage levels range from 0µM to 125µM. The first experiment includes two dosage levels, making it a two-class problem, whereas the second experiment resembles a multi-classification problem with four levels. Based on these tasks, the aforementioned fine-tuning strategies (selective fine-tuning [6], Linear Probing [6; 24], Gradual Unfreeze (last/all)(LP-FT) [6; 24], and Full Fine-tuning [6]) were compared to one another.

2.1 Classifying cardiomyocytes: performance baseline for Geneformer

One of the practical examples discussed in the Geneformer paper is the analysis of in silico treatments. As a prerequisite of this experiment, the model was trained to distinguish between cardiomyocytes [25] from non-failing heart cells (n=9) and cells affected by hypertrophic (n=11) or dilated (n=9) cardiomyopathy with an overall out-of-sample accuracy of 90% [2]. The experiment was performed in order to see whether Geneformer's in silico perturbation strategy could be applied to model human disease and reveal candidate therapeutic targets [2]. We began our studies by reproducing the first part of the experiment to set a performance baseline for the selective fine-tuning configuration used in the paper and observe how the predictive performance changes with a different dataset, more specifically sciplex2 [3].

3 Materials and methods

This section elaborates on the model used for the prediction task and its architecture, as well as the fine-tuning strategies implemented and the dataset chosen for the experiment.

3.1 The Geneformer model

Geneformer consists of six transformer encoder units, each containing a self-attention layer and a feed forward neural network layer with 4 attention heads per layer. Implementation was done using *pytorch* [26; 27] and the *Huggingface Transformers* library [28] was used for model configuration, data loading, and training. The model works by taking in single-cell transcriptomes, encoded in the form of rank values which represent the expression of each gene within the cell, and turning the encoded input into a 256-dimensional embedding in which the relationships between the cell's genes are highlighted.

On top of the original Geneformer model, an additional task-specific final transformer layer is then added [2]. In the context of our experiment, this seventh layer acts as the classifier which predicts whether a cancer cell was exposed to nutlin-3A and has shown sensitivity to it or not. This final task-specific transformer layer, together with a chosen amount of pre-trained layers from the original model, is to be fine-tuned using the human lung adenocarcinoma samples.

A representation of Geneformer's architecure can be seen in Figure 1.

3.2 The sciplex2 dataset

Sciplex2 is a publicly available repository of A549 human lung adenocarcinoma cells which have been exposed to BMS345541, dexamethasone, nutlin-3A, SAHA, or DMSO

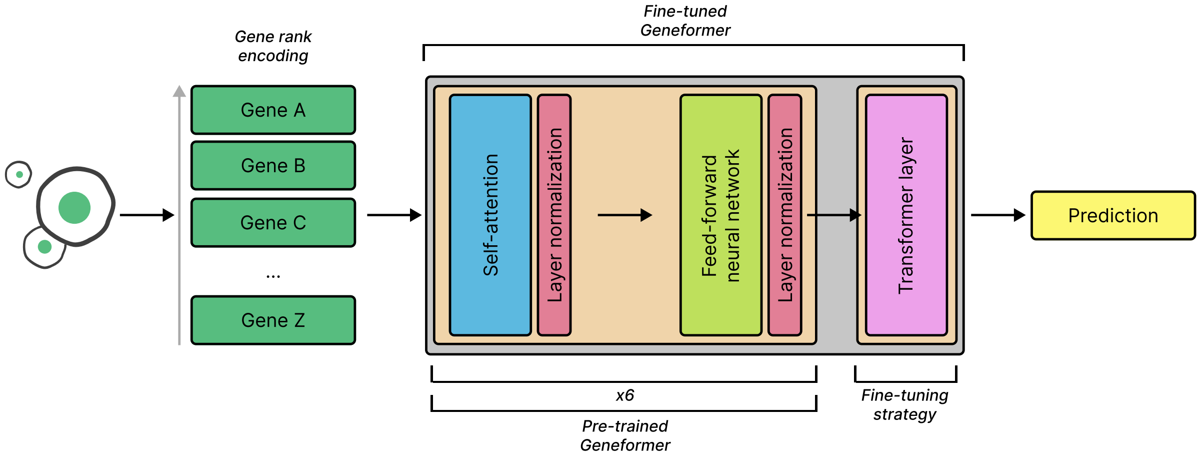

Figure 1: Geneformer's architecture. The model takes in a single-cell transcriptome, in the form of rank values which represent the gene expressions, and turns it into a 256-dimensional embedding. The embedding is then used as input for the final transformer layer as part of the classification task.

vehicle control with varying dosages, ranging from 0µM to 250µM [3]. Sciplex2 is the result of ”sci-Plex”, a method of increasing the cost-effectiveness of high-throughput screens (HTSs) with scRNA-seq-based phenotyping by quantifying the transcriptional responses to thousands of independent perturbations at single-cell resoltuion using ”nuclear hashing”, as described by Cao et al. [3] The cell sample entries were generated according to the protocol described in the manuscript [3], as well as in another work authored by Cao et al. [29]

The dataset is accessible at the NCBI GEO database [23] (accession GSM4150377).

3.3 Steps to classify transcriptomic data from cancer cells

As mentioned in Section 2, Geneformer was evaluated on the 2-class problem of differentiating between human lung adenocarcinoma cells exposed to either nutlin-3A or the DMSO vehicle control with dosages of 0µM and 25µM, and on the multi-classification task which implied exposure to nutlin-3A at dosage levels of 0.25µM, 2.5µM, 25µM, and 125µM respectively [3]. For the 2-class problem, we began with preprocessing the sciplex2 dataset by filtering its samples into a subset containing only the relevant transcriptomic data (cancer cells exposed to DSMO vehicle control and nutlin-3a with the specified dosages) and then tokenizing those into a format which could be used by the deep-learning model using the Geneformer's tokenizer implementation [2]. The subset was split into an evaluation set and a test set with a ratio of 8:2. We then fine-tuned the model, which was pre-trained using the Genecorpus-30M dataset comprising of 29.9 million human single-cell transcriptomes from a broad range of tissues from publicly available data [2], on the training set using the fine-tuning methods mentioned. The fine-tuned model's task performance was then evaluated on the test set, followed by compiling the confusion matrices and gathering metrics for the accuracy and F1-score.

The same steps were followed in the case of the 4-class problem. The only difference lied in the set of samples that were used for training and evaluating the model.

3.4 Fine-tuning strategies

Given their efficiency in avoiding multiple training iterations which are resource-intensive [6], these non-iterative fine-tuning strategies are appealing in mapping gene networks. Some of the methods, like Full Fine-tuning, are already a standard in transfer learning [6], and thus it is worth evaluating them for the task at hand.

Selective Fine-tuning [6]

Selective fine-tuning is a strategy where parameters such as the learning rate and the number of frozen layers are fixed beforehand, following that they stay the same throughout the retraining [6]. This strategy was adopted in the Geneformer's manuscript; it was used in different applications, among which was also the cardiomyocytes classification task which was reproduced in our study to establish a baseline.

The effectiveness of this method is highly dependent on the type of downstream task, the similarity of the domains, and the hyper-parameters chosen. Acceptable accuracy rates can be achieved, provided that suitable parameters are chosen. However, this strategy implies the time-consuming task of manually selecting and testing hyper-parameter configurations for the model.

Geneformer's manuscript reported that an out-of-sample accuracy of 90% was achieved by freezing the first two layers of the model and fine-tuning the rest [2].

Linear Probing [24]

Linear probing is similar to selective fine-tuning: the strategy consists of freezing all pre-trained layers and only training the additional classifier layer on the target dataset. An advantage of this strategy is that it is less computationally-demanding since only the last layer is trained. The original features of the pre-trained model are fully preserved. However, the model might not fully capture the nuances of the downstream task [6], and thus the predictive performance can be lower.

Gradual Unfreeze (last/all) (LP-FT) [24]

This strategy is the result of combining Linear Probing and Full Fine-tuning. Initially, all but the last classifier layer are frozen. The last layer is trained for an arbitrary amount of epochs after which all the layers are unfrozen. The fine-tuning then continues over the whole model for the remaining number of epochs. This approach enables tuning the pre-trained model for the downstream task while ensuring that the model weights still hold a specific degree of knowledge from the source domain.

Full Fine-tuning [6]

Full Fine-tuning is a standard, widely-used method [6] which implies training all the layers of the pre-trained model, as well as the additional classifier layer, on the task-specific dataset. Depending on the type of downstream task, it is hypothesized that this method could yield the highest accuracy rate, given that all layers are tuned specifically to solve the downstream task. However, when compared to the other strategies, an increased risk of overfitting and losing the original features [30; 6] should be taken into account.

Hyper-parameters for Fine-tuning

Throughout our experiments, all hyper-parameters besides the number of frozen layers were fixed as follows: max learning rate, 5×10^{-5}; learning scheduler, linear with warmup; optimizer, Adam with weight decay fix160; warmup steps, 500; weight decay, 0.001; batch size, 12. These parameters are the same as the ones chosen in Geneformer's manuscript [2].

4 Results

4.1 Classification of cardiomyocytes from non-failing hearts and hearts affected by hypertrophic or dilated cardiomyopathy & Selective Fine-tuning

The reproduction experiment yielded comparable results as was reported by the developers of Geneformer. We obtained an out-of-sample accuracy of 87.3%, with an F1-score of 0.85. The confusion matrix can be seen in Figure 2. The configuration was the same as the one presented in the original manuscript: the first two layers of the model were frozen, and the original hyperparameters were kept in place.

The model obtained an overall good accuracy rate, highlighting the power of transfer learning when applied in a situation with similar domains and a potential distribution shift. The domains for this task were the Genecorpus-30M dataset and the single-nucleus cell profiling dataset representing both healthy human hearts and hearts affected by dilated and hypertrophic cardiomyopathy.

4.2 Prediction of human lung adenocarcinoma cell exposure to nutlin-3A/DSMO vehicle control with two dosage levels

We firstly fine-tuned the pre-trained model with the original configuration as presented in the Geneformer manuscript: this implied fine-tuning the model throughout only one epoch in order to avoid overfitting [2]. We then tried the same configuration while using 10 epochs. The performance of the model was boosted by a considerable amount, as can be seen in Table 1. This was not an unexpected result, as it was mentioned in Geneformer's manuscript that ”hyperparameter tuning for deep learning applications generally significantly enhances learning and so it is likely that the maximum predictive potential of Geneformer in these downstream applications is significantly underestimated” [2].

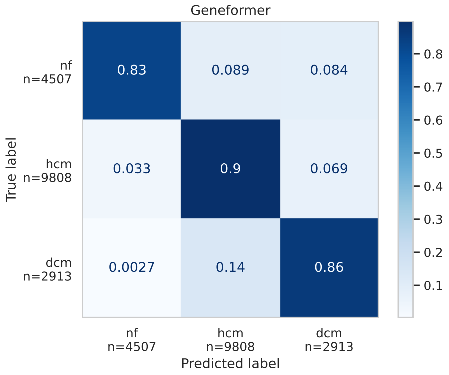

Figure 2: Confusion matrix for the cardiomyocytes classification task.

It can be argued that the increase in accuracy could have been the result of overfitting, but that would have lead to a test performance substantially worse than the levels of accuracy reached during the training stage. No significant difference between the two scores has been noticed.

Accuracies and F1-scores		
Strategy	**Accuracy**	**F1-score**
Two layers frozen (1 epoch)	93.3%	0.89
Two layers frozen (10 epochs) (Default)	96.0%	0.94
Linear Probing (10 epochs)	96.6%	0.95
Gradual Unfreeze (last/all) (LP-FT) (10 epochs)	97.3%	0.96
Full Fine-tuning (10 epochs)	95.1%	0.93

Table 1: Comparison of the different fine-tuning methods for the two-class human lung adenocarcinoma cell classification task.

Linear Probing obtained a very good accuracy, proving its suitability for the task at hand wherever a lower resource consumption is of utmost importance. However, despite the remarkable scores, it should be noted that the method had a tendency to return more false negatives compared to other strategies, as can be seen in Figure 5A. One must ensure that the model does not become too biased towards the ”Untreated” category, as this could lead to serious implications in a real-life setting.

The Gradual Unfreeze strategy obtained the highest accu-

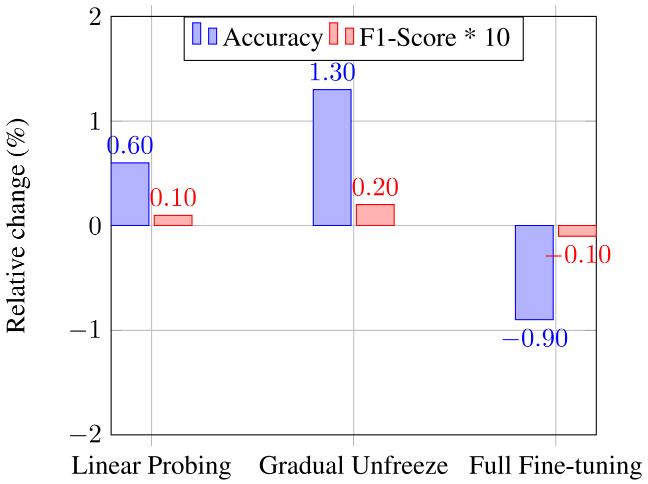

Figure 3: Comparison of the fine-tuning strategies relative to the default configuration of freezing the first two layers of Geneformer and training the rest over 10 epochs, % score change, two dosage level prediction task.

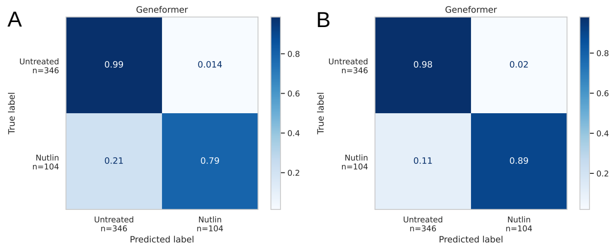

Figure 4: Confusion matrices for the 2-class cancer cell classification task (Left: 2 frozen layers, one epoch; Right: two frozen layers, 10 epochs).

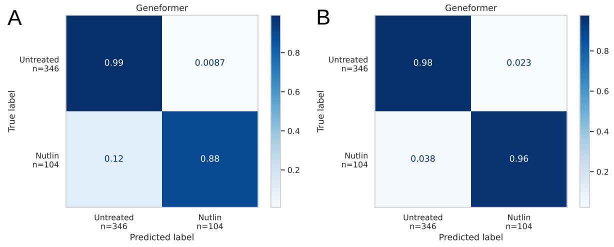

Figure 5: Confusion matrices for the 2-class cancer cell classification task (Left: Linear Probing, 10 epochs; Right: Gradual Unfreeze (last/all)(LP-FT), 10 epochs).

racy rate. This means that preserving the pre-trained weights to a higher degree was better for our task than fine-tuning all the layers from the beginning on the task-specific training set. In our experiment setup for the Gradual Unfreeze method, we decided to freeze the pre-trained layers for 5 epochs before unfreezing them and proceeding with a full model retraining.

Given that Full Fine-tuning performed better only when compared to the original Geneformer configuration which was trained using one epoch, it would be advisable to avoid employing the strategy in this particular context. The greater effort of training the whole model on the task-specific dataset did not lead to the performance improvement that would justify using it. Quite the contrary, the fully fine-tuned Geneformer model performed worse than its original counterpart with two layers frozen when trained over 10 epochs. This is indicative of a loss of relevant knowledge from the first two layers, an aspect which we elaborated on in Section 5.

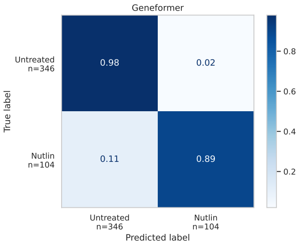

Figure 6: Confusion matrices for the 2-class cancer cell classification task (Full Fine-tuning, 10 epochs).

Overall, the results from all fine-tuning methods proved to be promising, indicating the ability to use less computationally-demanding strategies such as linear probing with little to no compromise with regard to the predictive accuracy. Surprisingly, using Linear Probing in our situation led to a better model performance compared to all but one other fine-tuning method.

4.3 Prediction of human lung adenocarcinoma cell exposure to nutlin-3A/DSMO vehicle control with four dosage levels

The multi-dosage prediction task proved to be more difficult for Geneformer which obtained an average accuracy of 68%. The average does not include the one epoch fine-tuning configuration. The scores for each method can be seen in Table 2.

Accuracies and F1-scores

Strategy	**Accuracy**	**F1-score**
Two layers frozen (1 epoch)	52.6%	0.40
Two layers frozen (10 epochs) (Default)	71.0%	0.70
Linear Probing (10 epochs)	65.8%	0.62
Gradual Unfreeze (last/all) (LP-FT) (10 epochs)	66.5%	0.63
Full Fine-tuning (10 epochs)	68.9%	0.65

Table 2: Comparison of the different fine-tuning methods for the multi-class human lung adenocarcinoma cell classification task.

Here, freezing the first two layers and training the rest of the model seemed to fare best both in terms of accuracy and the F1-score. The strategy led to an accuracy of 71%. As can be seen in Figure 8B, the model was able to differentiate

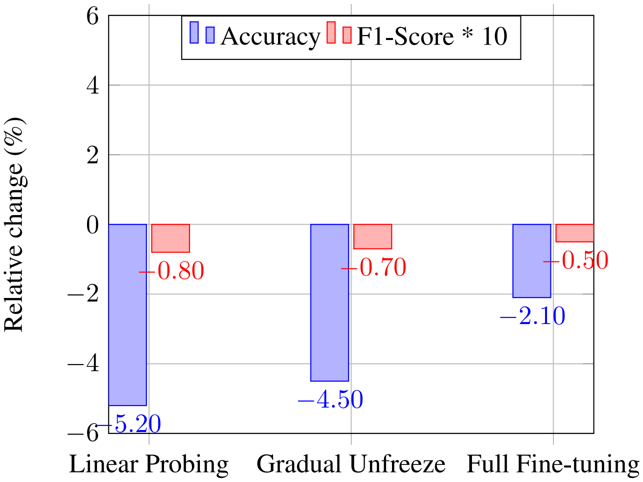

Figure 7: Comparison of the fine-tuning strategies relative to the default configuration of freezing the first two layers of Geneformer and training the rest over 10 epochs, % score change, four dosage level prediction task.

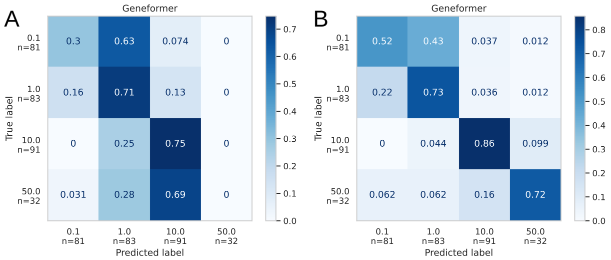

Figure 8: Confusion matrices for the multi-class cancer cell classification task (Left: 2 frozen layers, one epoch; Right: two frozen layers, 10 epochs).

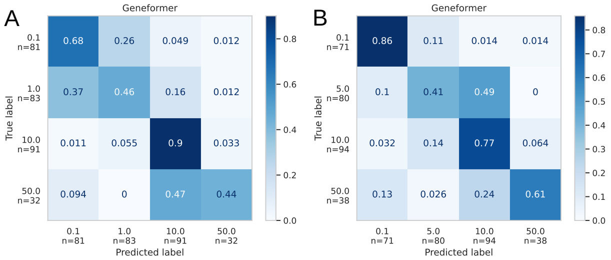

Figure 9: Confusion matrices for the multi-class cancer cell classification task (Left: Linear Probing, 10 epochs; Right: Gradual Unfreeze (last/all)(LP-FT), 10 epochs).

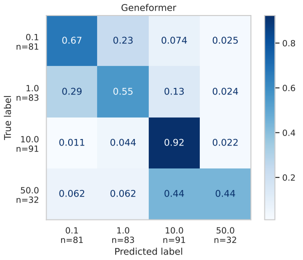

Figure 10: Confusion matrices for the multi-class cancer cell classification task (Full Fine-tuning, 10 epochs).

between cells exposed to nutlin-3A at dosage levels < 10.0 and cell exposed at levels ≥ 10.0 with relative ease. However, within each of those two ranges, the predictive accuracy worsened. This is indicative of a general grouping of the sample data into two main clusters, a trend highlighted when the same configuration was used to tune Geneformer over a single epoch (Figure 8A). This was discussed in more detail in Section 5.

The second best strategy was Full Fine-tuning with an accuracy of 68.9%, F1-score 0.65, followed by Gradual Unfreeze with an accuracy of 66.5%, F1-score 0.63 (Figure 9B, Figure 10). Corroborating those results with the scores obtained by the best configuration, we noticed an importance of choosing what layers are frozen during the fine-tuning and the number of epochs for which they are frozen: interestingly enough, fully training the weights on the target domain seems to be removing features from the pre-trained weights that are relevant to differentiating between dosage levels 10.0 and 50.0. On the other hand, gradually unfreezing the layers improved the prediction of samples exposed to dosage levels 0.1 and 50.0 at the cost of a higher rate of error for dosage levels 1.0 and 10.0.

Linear Probing behaved in a manner similar to training Geneformer with the first two layers frozen. The model could discriminate the cluster of samples exposed to dosages < 10.0 from the one of samples exposed to dosages ≥ 10.0 with higher accuracy than the predictions made within the pairs $\{0.1, 1.0\}$ and $\{10.0, 50.0\}$.

5 Discussion

Our findings show that no single fine-tuning method can outperform all other options across the different scenarios. Thus, there is no configuration which is optimal. The same conclusion was reached for tasks involving medical imaging processing and other real-world tasks involving distribution shifts [6; 30]. Depending on the application, various configurations should be considered and tested as they differently influence how information is being stored within the layers of Geneformer.

The suitability of Linear Probing as a fine-tuning method is mainly limited to tasks where the problem complexity is reduced or where the source and target domains share a high degree of similarity, given that all model layers besides the additional transformer layer at the end are frozen during the training process. For the two-fold classification task, the method managed to obtain the highest scores, however a noticeable decline in the predictive accuracy was observed after running the four class classification task.

Using Geneformer with Full Fine-tuning led to a performance decrease in both tasks. These results are contrary to the hypothesis that tuning all the layers on the downstream domain should improve the model's performance within the domain. Moreover, it implies that one should pay attention when updating the pre-trained weights: sometimes it is better to preserve those to a higher degree.

Keeping a balance between preserving the existing knowledge and tuning the model to the target domain can be best done using Gradual Unfreezing, as it offers a higher degree of freedom with regard to how much each layer is tuned by enabling the selection of the number of layers to freeze, as well as setting the amount of epochs for which they remain frozen. Despite this, an optimal configuration can be more difficult to find, since there are more parameters which have to be adjusted manually. This can be mitigated by automatically tuning the hyperparameters over multiple training iterations [6], a method out of the scope of this paper.

During both experiments, the number of epochs over which the first two layers of Geneformer were trained had a notable impact over the accuracy of the predictions. Even when no fine-tuning was performed on these layers, the model still attained a reasonable performance, a sign that the pre-trained weights already stored information which was closely related to knowledge specific of the target domain. Tuning those two layers over a limited amount of epochs could lead to a performance boost, however over-training them worsened the out-of-sample predictive accuracy.

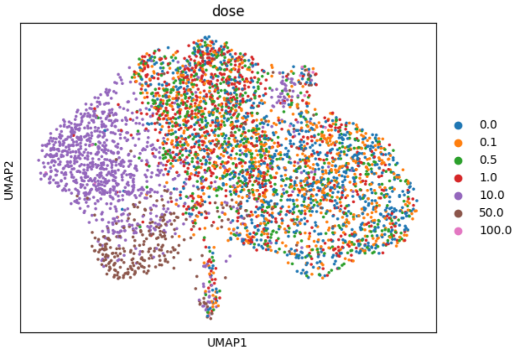

Figure 11: UMAP plot of the human adenocarcinoma cells exposed to nutlin-3A at dosage levels of 0μM, 0.25μM, 1.25μM, 2.5μM, 25μM, 125μM, and 250μM.

The confusion matrices generated by the fine-tuning strategies in the four-level dosage prediction task unveiled a tendency to group the test samples into two clusters: one of cells exposed to nutlin-3A with dosage levels < 10.0, and another one of cells exposed at levels ≥ 10.0. This could mean that rising the dosage level of the compound is not proportional to the changes in its impact over the cell, the biggest changes being observed within the range [1.0, 10.0]. This is supported by Figure 11, which shows a clear separation between the two clusters mentioned above.

It is possible that the performance obtained from the fine-tuning methods be further improved by also tuning the other training hyperparameters accordingly. They have not been specifically adapted for the tasks at hand, since the scope of our research was solely to conduct an objective comparison of the fine-tuning strategies themselves.

6 Conclusions & Future Work

Our study has evaluated the performance of Geneformer, a ”context-aware, attention-based deep-learning model” [2], over the task of predicting the effects of Nutlin-3A applied to human adenocarcinoma cells. We combined the model with widely-used fine-tuning methods and evaluated their impact over the predictive accuracy. We did this to assess Geneformer's ability to leverage the potential of transfer learning in the context of mapping gene networks for cancer cells. The model has already shown promising results when used for other tasks involving transcriptomic data, such as in silico cell perturbations and classifying healthy cardiomyocytes from diseased cardiomyocytes, and thus it may contribute towards developing personalized treatments in the future. Existing research has compared the fine-tuning methods shown here for medical imaging; our work expanded on it by exploring a different medical domain.

We sought to answer what different strategies can be employed to fine-tune Geneformer so that it is able to correctly identify a sample taken from a cancer cell line as either untreated/not sensitive to a treatment, or treated and sensitive to the treatment, as well as looking into features of the fine-tuning methods that seemed to particularly impact the overall accuracy.

We performed two in silico experiments by gathering relevant samples from the sciplex2 dataset, namely human adenocarcinoma cells exposed to the compound Nutlin-3A with varying dosage levels. We used those samples to create an evaluation and a test set. We then tuned the pre-trained Geneformer model using each of the fine-tuning methods discussed (freezing the first 2 layers, Linear Probing, Full Fine-tuning, Gradual Unfreeze (Last/All)(LP-FT)), following that we evaluated the configuration over the test set and plotted the results.

We concluded that, depending on the way in which knowledge from the source domain is stored inside the pre-trained model and the similarity between the source and the target domains, each fine-tuning strategy can either be suitable or not for a given task. There is no single optimal solution.

However, our study has limitations. Firstly, we focused our attention on a particular subset of fine-tuning strategies, those being non-iterative methods. We also trained and evaluated said methods on a relatively small set of samples, those being human adenocarcinoma cells exposed to two compounds: nutlin-3A and the DMSO vehicle control. We relied on just the Sciplex2 dataset in order to create the test set on which the performance of the model was measured. Further research could improve on this analysis by showcasing the effects of other types of fine-tuning strategies, i.e. iterative methods, over Geneformer's predictive accuracy. The model could also be tested in more diverse scenarios in order to assess its generalization capability. This could be done by training and testing Geneformer over different cancer cell lines exposed to other types of compounds. Finally, a more comprehensive exploration of the hyperparameter space should reveal whether an even better performance can be obtained out of Geneformer.

We believe that this paper will serve as a starting point for thoroughly assessing Geneformer's ability to be used for predicting the impact of new drugs over the phenotype of a cell in the emerging field of machine learning-aided pharmaceutical research and development.

7 Responsible Research

Performing research with ethics in mind from the start is crucial for ensuring that the results of the experiments are accurate, representative, and reproducible. Being critical with regard to the ethical implications of a study is something to be though about throughout the entire duration of the study and the researcher should always adhere to guidelines for responsible research [31]. Throughout the research presented in this paper, we ensured that the experiments we made are reproducible by the specialist reader. We also assessed the ethical implications of the study in a holistic manner: the main points are discussed in Section 7.2.

7.1 Reproducibility of experiments

The focus of this paper is the assessment of different fine-tuning strategies for Geneformer [2] such that an efficient, accurate classification of cancer cells exposed to different types of treatments as either sensitive or insensitive to the treatments can be performed, as discussed in Section 2. This implies several experiments where Geneformer was fine-tuned on part of the sciplex2 dataset [3] using different fine-tuning strategies and then evaluated on a test set. To ensure their reproducibility, the methods chosen and the steps taken were thoroughly documented in Section 3. The materials used by the methods were referenced such that they can be easily obtained by the reader.

7.2 Other ethical considerations

We firmly believe that the findings of our study can aid future technological breakthroughs in medicine. Accessible personalized treatments for disorders is something that we consider everyone should have access to at a reasonable cost. That being said, there are some important ethical considerations that come with using large medical databases for Machine Learning-related purposes:

- **Privacy:** the use of potentially sensitive medical data should be done in a way that protects the privacy of any patients whose genetic data could have been included in the datasets [32]. The patients should be aware that their genetic data, over which they have full authority, has been used.
- **Accurate representation of data:** the data used to train deep-learning models for the purpose of aiding with the treatment of disorders should be representative of a diverse population [33; 34], such that the model predictions remain trustworthy over a wide range of phenotypes and they are not biased towards a group of the population.
- **Clinical implications:** deep-learning models could be used to aid with decision making in clinical contexts in the future [33]. As such, it is crucial to think about their impact and the implications of Artificial Intelligence in this scenario. Misclassification can potentially lead to inappropriate treatment choices being made [35; 36; 37], and thus the accuracy and the reliability of AI models become crucial.

References

[1] Swen Jesse J, Huizinga Tom W, Gelderblom Hans, de Vries Elisabeth G. E, Assendelft Willem J. J, Kirchheiner Julia, and Guchelaar Henk-Jan. Translating pharmacogenomics: Challenges on the road to the clinic. *PLoS Medicine*, 4:1317–1324, 2007.

[2] Theodoris Christina V., Xiao Ling, Chopra Anant, Chaffin Mark D., Al Sayed Zeina R., Hill Matthew C., Mantineo Helene, Brydon Elizabeth M., Zeng Zexian, Liu X. Shirley, and Ellinor Patrick T. Transfer learning enables predictions in network biology. *Nature*, 618:616–624, 2023.

[3] Srivatsan Sanjay R., McFaline-Figueroa José L., Ramani Vijay, Saunders Lauren, Cao Junyue, Packer Jonathan, Pliner Hannah A., Jackson Dana L., Daza Riza M., Christiansen Lena, Zhang Fan, Steemers Frank, Shendure Jay, and Trapnell Cole. Massively multiplex chemical transcriptomics at single-cell resolution. *Science*, pages 45–51, 2019.

[4] Singh-Blom U. Martin, Natarajan Nagarajan, Tewari Ambuj, Woods John O., Dhillon Inderjit S., and Marcotte Edward M. Prediction and validation of gene-disease associations using methods inspired by social network analyses. *PLoS ONE*, 8, 2013.

[5] Malki Mustafa Adnan and Pearson Ewan Robert. Drug–drug–gene interactions and adverse drug reactions. *The Pharmacogenomics Journal*, 20:355–366, 2020.

[6] Davila Ana, Colan Jacinto, and Hasegawa Yasuhisa. Comparison of fine-tuning strategies for transfer learning in medical image classification. *Image and Vision Computing*, 146:105012, 2024.

[7] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[8] Padmavathi Kora, Chui Ping Ooi, Oliver Faust, U. Raghavendra, Anjan Gudigar, Wai Yee Chan, K. Meenakshi, K. Swaraja, Pawel Plawiak, and U. Rajendra Acharya. Transfer learning techniques for medical image analysis: A review. *Biocybernetics and Biomedical Engineering*, 42(1):79–107, 2022.

[9] Hee E. Kim, Alejandro Cosa-Linan, Nandhini Santhanam, Mahboubeh Jannesari, Mate E. Maros, and Thomas Ganslandt. Transfer learning for medical image classification: a literature review. *BMC Medical Imaging*, 22(69), 2022.

[10] Jindong Wang, Vincent W. Zheng, Yiqiang Chen, and Meiyu Huang. Deep transfer learning for cross-domain activity recognition. In *Proceedings of the 3rd International Conference on Crowd Science and Engineering*, ICCSE'18, New York, NY, USA, 2018. Association for Computing Machinery.

[11] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In Kamalika Chaudhuri and Ruslan

Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5389–5400. PMLR, 09–15 Jun 2019.

[12] David N Perkins, Gavriel Salomon, et al. Transfer of learning. *International encyclopedia of education*, 2:6452–6457, 1992.

[13] Quinonero-Candela Joaquin, Sugiyama Masashi, Schwaighofer Anton, and Lawrence Neil D. *Dataset Shift in Machine Learning*. MIT Press, 2022.

[14] John Hancock and Taghi M. Khoshgoftaar. Impact of hyperparameter tuning in classifying highly imbalanced big data. In *2021 IEEE 22nd International Conference on Information Reuse and Integration for Data Science (IRI)*, pages 348–354, 2021.

[15] Hafsa Habehh and Suril Gohel. Machine learning in healthcare. 2021.

[16] Abdullah Alanazi. Using machine learning for healthcare challenges and opportunities. *Informatics in Medicine Unlocked*, 30:100924, 2022.

[17] Yutaro Yamada, Jacinto Colan, Ana Davila, and Yasuhisa Hasegawa. Task segmentation based on transition state clustering for surgical robot assistance. In *2023 8th International Conference on Control and Robotics Engineering (ICCRE)*, pages 260–264, 2023.

[18] Dandan Zhang, Zicong Wu, Junhong Chen, Anzhu Gao, Xu Chen, Peichao Li, Zhaoyang Wang, Guitao Yang, Benny Lo, and Guang-Zhong Yang. Automatic microsurgical skill assessment based on cross-domain transfer learning. *IEEE Robotics and Automation Letters*, 5(3):4148–4155, 2020.

[19] Jani Koskinen, Mastaneh Torkamani-Azar, Ahmed Hussein, Antti Huotarinen, and Roman Bednarik. Automated tool detection with deep learning for monitoring kinematics and eye-hand coordination in microsurgery. *Computers in Biology and Medicine*, 141:105121, 2022.

[20] Joël L. Lavanchy, Joël Zindel, Kadir Kirtaç, Isabell Twick, Enes Hosgor, Daniel Candinas, and Guido Beldi. Automation of surgical skill assessment using a three-stage machine learning algorithm. Mar 2021.

[21] Thomas Sanford, Ling Zhang, Stephanie A. Harmon, Jonathan Sackett, Dong Yang, Holger R. Roth, Ziyue Xu, Deepak Kesani, Sherif Mehralivand, Ronaldo Hueb Baroni, Tristan Barrett, Rossano Girometti, Aytekin Oto, Andrei S. Purysko, Sheng Xu, Peter A. Pinto, Daguang Xu, Bradford J. Wood, Peter L. Choyke, and Barış Türkbey. Data augmentation and transfer learning to improve generalizability of an automated prostate segmentation model. 2020.

[22] Theodoris Christina. Geneformer. https://huggingface.co/ctheodoris/Geneformer, 2013.

[23] Barrett Tanya, Wilhite Stephen E., Ledoux Pierre, Evangelista Carlos, Kim Irene F., Tomashevsky Maxim, Marshall Kimberly A., Phillippy Katherine H., Sherman Patti M., Holko Michelle, Yefanov Andrey, Lee Hyeseung, Zhang Naigong, Robertson Cynthia L., Serova Nadezhda, Davis Sean, and Soboleva Alexandra. Ncbi geo: archive for functional genomics data sets—update. *Nucleic Acids Research*, 41:D991–D995, 2013.

[24] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution, 2022.

[25] Chaffin Mark, Papangeli Irinna, Simonson Bridget, Akkad Amer-Denis, Hill Matthew C., Arduini Alessandro, Fleming Stephen J., Melanson Michelle, Hayat Sikander, Kost-Alimova Maria, Atwa Ondine, Ye Jiangchuan, and Virendar K. Kaushik Christian M. Stegmann Kenneth B. Margulies Nathan R. Tucker Patrick T. Ellinor Kenneth C. Bedi Jr, Matthias Nahrendorf. Single-nucleus profiling of human dilated and hypertrophic cardiomyopathy. *Nature*, 608:174–180, 2022.

[26] PyTorch Team. Pytorch.

[27] Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, Geeta Chauhan, Anjali Chourdia, Will Constable, Alban Desmaison, Zachary DeVito, Elias Ellison, Will Feng, Jiong Gong, Michael Gschwind, Brian Hirsh, Sherlock Huang, Kshiteej Kalambarkar, Laurent Kirsch, Michael Lazos, Mario Lezcano, Yanbo Liang, Jason Liang, Yinghai Lu, C. K. Luk, Bert Maher, Yunjie Pan, Christian Puhrsch, Matthias Reso, Mark Saroufim, Marcos Yukio Siraichi, Helen Suk, Shunting Zhang, Michael Suo, Phil Tillet, Xu Zhao, Eikan Wang, Keren Zhou, Richard Zou, Xiaodong Wang, Ajit Mathews, William Wen, Gregory Chanan, Peng Wu, and Soumith Chintala. Pytorch 2: Faster machine learning through dynamic python bytecode transformation and graph compilation. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, ASPLOS '24, page 929–947, New York, NY, USA, 2024. Association for Computing Machinery.

[28] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771, 2019.

[29] Junyue Cao, Jonathan S. Packer, Vijay Ramani, Darren A. Cusanovich, Chau Huynh, Riza Daza, Xiaojie Qiu, Choli Lee, Scott N. Furlan, Frank J. Steemers, Andrew Adey, Robert H. Waterston, Cole Trapnell, and Jay Shendure. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*, 357(6352):661–667, 2017.

[30] Yoonho Lee, Annie S. Chen, Fahim Tajwar, Ananya Kumar, Huaxiu Yao, Percy Liang, and Chelsea Finn.

Surgical fine-tuning improves adaptation to distribution shifts, 2023.

[31] Farrimond Hannah. *Doing Ethical Research*. Bloomsbury Publishing, 2012.

[32] *Policy for Sharing of Data Obtained in NIH Supported or Conducted Genome-Wide Association Studies (GWAS)*, number NOT-OD-07-088, 2007.

[33] Tran Khoa A., Kondrashova Olga, Bradley Andrew, Williams Elizabeth D., Pearson John V., and Waddell Nicola. Deep learning in cancer diagnosis, prognosis and treatment selection. *Genome Med*, 13, 2021.

[34] Gagnier Joel J, Moher David, Boon Heather, Beyene Joseph, and Bombardier Claire. Investigating clinical heterogeneity in systematic reviews: a methodologic review of guidance in the literature. *BMC Med Res Methodol*, 12, 2012.

[35] Emma Chen, Andy Kim, Rayan Krishnan, Jin Long, Andrew Y. Ng, and Pranav Rajpurkar. Chexbreak: Misclassification identification for deep learning models interpreting chest x-rays. In Ken Jung, Serena Yeung, Mark Sendak, Michael Sjoding, and Rajesh Ranganath, editors, *Proceedings of the 6th Machine Learning for Healthcare Conference*, volume 149 of *Proceedings of Machine Learning Research*, pages 103–125. PMLR, 06–07 Aug 2021.

[36] Yuhang Dong, W. David Pan, and Dongsheng Wu. Impact of misclassification rates on compression efficiency of red blood cell images of malaria infection using deep learning. *Entropy*, 21(11), 2019.

[37] Abhay Shah, Stephanie Lynch, Meindert Niemeijer, Ryan Amelon, Warren Clarida, James Folk, Stephen Russell, Xiaodong Wu, and Michael D. Abràmoff. Susceptibility to misdiagnosis of adversarial images by deep learning based retinal image analysis algorithms. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1454–1457, 2018.