As a cell, is it better to be single?

Exploring the feasibility of fine-tuning Geneformer on bulk RNA sequencing data

Alan Kuźnicki[1]
Supervisors: Prof.dr.ir. Marcel 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: Alan Kuźnicki
Final project course: CSE3000 Research Project
Thesis committee: Prof.dr.ir. Marcel Reinders, Niek Brouwer, Dr. Merve Gürel

Abstract

Powerful new machine learning models in biomedicine are being developed constantly, further hastened by the advent of transformer-based architectures. These advanced systems can be used for various applications, from diagnostics to assessing drug effectiveness. Many of these are fundamentally cell classification problems. Models like Geneformer [1] use gene expression data to learn how to distinguish between these cell classes. This information is usually obtained through single-cell RNA sequencing. However, the alternative source, bulk RNA sequencing, offers some advantages that make exploring the feasibility of using it to train Geneformer enticing, such as its greater availability and lower cost.

In this paper, pseudo-bulk datasets are created from single-cell data by aggregation of gene expressions. A method to generate synthetic single-cell-like data from a bulk dataset is used to create new datasets. Some remain purely synthetic, while others are mixed with real single-cell data. Geneformer is fine-tuned on all generated datasets separately, and its performance in a cell classification problem is measured. It is shown that the more a dataset resembles real single-cell data, the better the model's performance. Using bulk data to fine-tune Geneformer is proven to be infeasible. The synthetic data fails to effectively fine-tune the model and is proven to not have a meaningful impact when added to a single-cell dataset. It is concluded that the generated synthetic data is of too low quality and that alternative generation methods should be explored.

1 Introduction

Using machine learning models in biomedicine is often a major challenge to researchers, but has attracted significant attention due to its potential to aid in a large variety of tasks, from patient diagnosis to drug discovery [2, 3]. In recent years, models utilizing the transformer architecture [4] have been developed and used successfully in various biomedical applications, such as diagnostics [5]. Many of the tasks these models can be used for include the fundamental problem of cell classification. For example, the effectiveness of a novel drug can be assessed by verifying whether the cells to which it was applied have been classified as "sensitive" or "resistant" [6].

Geneformer [1] is a recently published model based on the transformer architecture that can be used effectively in predicting cell labels. It comprises 6 transformer layers for pre-training on a large, general dataset, and one additional layer for prediction. Problem-specific data is used only to fine-tune the pre-trained model, meaning Geneformer does not need to be fully re-trained for each task, unlike earlier models [7, 8]. The pre-trained dataset and data used to fine-tune the model consist of gene expression values provided per cell. In cell classification problems, Geneformer learns to distinguish between classes based on the differences in their gene expressions. The data used to fine-tune Geneformer thus greatly influences its effectiveness. Given the model's wide potential applicability, it is important to explore its limitations, particularly concerning how the characteristics of the fine-tuning dataset influence its prediction performance.

Gene expression data is obtained through RNA sequencing (RNA-seq,) during which the abundance of respective RNA molecules is measured in a cell or tissue. Based on this information, relative expression values can be assigned to the corresponding genes [9]. Two major methods of gathering this data are single-cell RNA-seq (scRNA-seq) and bulk RNA-seq. In the former, gene expression information is obtained for each cell individually, while the latter provides it for groups of cells. Figure 1 illustrates the difference between these approaches. The method used to obtain gene expression data thus significantly impacts its characteristics and applicability. For example, as data from bulk RNA-seq is available only for groups and not individual cells, it is unlikely that it could be used to effectively fine-tune Geneformer for cell classification, a task requiring the separation of cells.

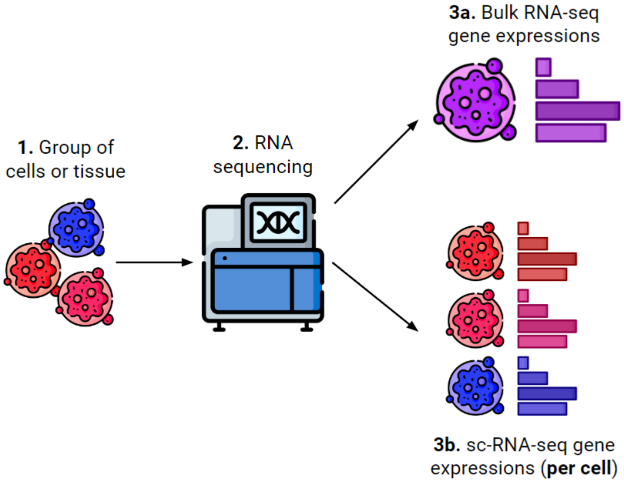
3a. Bulk RNA-seq gene expressions
1. Group of cells or tissue
2. RNA sequencing
3b. sc-RNA-seq gene expressions (**per cell**)

Figure 1: *An illustrative explanation of the difference between the gene expression data obtained through scRNA-seq and bulk RNA-seq. In the former, cells are treated individually, and expressions are provided for each one separately. In the latter, the expressions are provided for a group of cells together, obscuring their heterogeneity. This figure has been designed using images from Flaticon.com.*

While it is probable that Geneformer fine-tuned on bulk data would not perform acceptably in cell classification problems, there are reasons why the feasibility of using it for this task instead of or together with single-cell data should be explored. Bulk data is generally available for a larger variety of problems and acquiring it is less labor-intensive and cheaper than single-cell data [10]. Measuring the impact of using it to fine-tune Geneformer on the model's prediction performance could shed light on the characteristics that fine-tuning datasets need for Geneformer to perform well. This information would be valuable for the model's future applications, especially if fine-tuning on some bulk datasets was shown not to impact the prediction performance negatively to a significant degree. Furthermore, methods to extract useful, not immediately available information from bulk data exist, for example, bulk deconvolution, which can reveal the abundance of different cell types within a bulk sample [11]. While the outcomes of bulk deconvolution are not directly useful for this research, their existence indicates that bulk data has the potential to be processed to yield a dataset more suitable for fine-tuning Geneformer.

This research aims to first, explore the usability of bulk datasets for fine-tuning Geneformer for cell classification and second, to verify the feasibility of processing such datasets to yield data that the model can more effectively be fine-tuned on. Of additional interest to the former goal is whether the "bulkiness" of the data, defined as the average number of cells per data point in the dataset, influences its usability. As the bulkiness decreases, the dataset resembles a single-cell one more. This leads to the hypothesis that the bulkier the data, the worse Geneformer will perform in cell classification, and that it will be possible to determine approximately at what point said data becomes prohibitively ineffective. To test this hypothesis, "pseudo-bulk" datasets are generated from a single-cell dataset by aggregating cells into groups of varying sizes. This yields several datasets of varying bulkiness. Geneformer is then fine-tuned on each individually and its prediction performance in cell classification is measured. The results of these experiments should yield enough evidence to either confirm or refute the hypothesis. In addition, this information ought to provide substantial insight into Geneformer's limitations, particularly regarding the quantity and quality of fine-tuning data. To address the second aim of this research, a simple probabilistic method to generate synthetic single-cell data from a bulk dataset is implemented. It is used in two separate experiments: firstly, Geneformer is fine-tuned on it alone, and secondly, the synthetic dataset is added to a pure single-cell one and Geneformer is fine-tuned on both. Analyzing the model's performance in predicting cell labels in a pure single-cell dataset should verify whether synthesizing data from bulk datasets can feasibly improve Geneformer's performance.

2 Methodology

2.1 Data and pre-processing

All data used in this research is taken from the Sciplex2 dataset [12]. Sciplex2 contains gene expression information from over 20,000 cells exposed to four different drugs: BMS-345541 (label: BMS), Nutlin-3a (label: Nutlin), suberoylanilide hydroxamic acid (label: SAHA), and Dexamethasone (label: Dex) at eight doses. One of these doses is 0.0; these cells are labeled "Untreated" and constitute a fifth class for the label prediction problem. Thus, the name of the drug used to treat a given cell or a lack thereof, is used as the class label that Geneformer is tasked with predicting.

The Sciplex2 dataset is pre-processed to remove certain cells and limit the number of genes in consideration. Firstly, all cells that do not have a provided drug label or dose are removed. Secondly, cells with fewer than 500 or more than 12,000 total gene expression counts are filtered out. Lastly, cells where more than half of all recorded expressions were of mitochondrial genes are also removed. The resulting dataset is saved.

From the pre-processed dataset, a representative 10% of cells are set aside as a benchmark/test dataset. The rest becomes the training dataset and is used for aggregation and synthetic generation. The test dataset is constant and used for all evaluations.

2.2 Aggregation of single-cell into pseudo-bulk data

Pseudo-bulk data is generated from the training single-cell dataset by grouping cells and aggregating each group into one data point. The average number of cells per group is determined by the aggregation factor **k**, while the actual sizes are randomized. The aggregation is performed for **k** = {2, 5, 10, 25, 100}, yielding datasets of increasing bulkiness. Three aggregation strategies are explored separately: summing gene expressions, averaging them with full knowledge of how many cells are in a group, and averaging approximately through using **k** instead of the real group size. Figure 2 illustrates the difference between the sum and average approaches. In all three cases, the grouping is done over cells within a class and dose. The member cells are chosen randomly within that subset of the data. Importantly, this approach does not provide any benefits in real applications and is taken to give more insights about Geneformer and supply more information relevant to the research.

Real group sizes: The number of cells taken into a group is generated by taking a random value from a Gaussian distribution centered at 0.0 with a standard deviation of 0.25. The resulting value is interpreted as a percentage change in group size from the base **k**. The new group size is capped from the bottom at a single cell, and from the top at 2**k** - 1 cells. The randomization of group sizes is performed to simulate real bulk data, where the exact number of cells within a sample is known only approximately.

Aggregation with summing: Gene expression data collected through bulk sequencing is comparable to a sum of the gene expressions of the individual cells. As such, aggregating single-cell data by summing the gene expressions yields pseudo-bulk data that resembles real bulk data.

Aggregation with approximate averaging: While

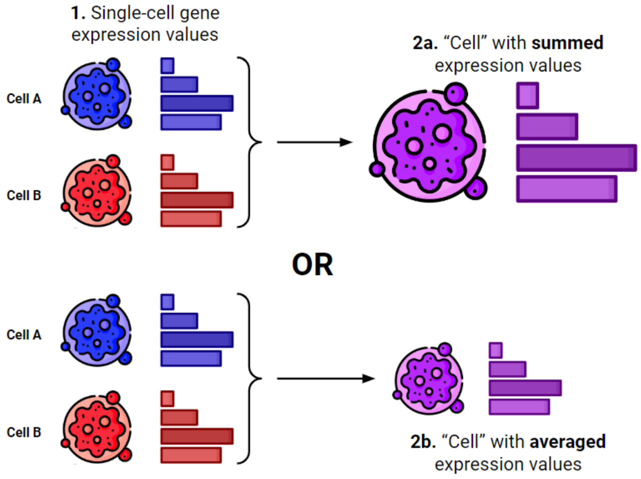
1. Single-cell gene expression values
Cell A
Cell B
2a. "Cell" with summed expression values
OR
Cell A
Cell B
2b. "Cell" with averaged expression values

Figure 2: *The main difference between the aggregation approaches. In the top half, aggregation is performed by summing the expressions, yielding a large "cell" with expressions from all component cells. The bottom half shows averaging aggregation, giving a normal-sized "cell". This figure has been designed using images from Flaticon.com.*

summing gene expressions creates realistic pseudo-bulk data, the aggregated dataset will have gene expression values significantly higher than the single-cell dataset, which may negatively influence the model's prediction performance. An alternative approach is to average the gene expressions within a group. As noted before, the precise number of cells comprising a bulk data point is unknown. To simulate that, the aggregation factor is used instead of the group size when averaging, resulting in only approximate averages.

Aggregation with exact averaging: The final explored aggregation method also utilizes averaging of the gene expressions but assumes that the number of cells per group is known precisely. Thus, the resulting averages are exact. Given that this information is not realistically available, this approach is intended to be a hypothetical best-case scenario for averaging.

2.3 Generating synthetic single-cell data from a bulk dataset

The bulk dataset is used as a basis for the synthetic data generator. It is first created by aggregating (with a known average) the entire single-cell training set into a single data point per label per dose, meaning seven points for each of the four drugs, and one for the untreated cells. Each of these is used as the means for the generator. From pure single-cell data, the covariance matrices are calculated per label and dose, and the max expression values for each gene are saved. The resulting covariance matrices are extremely large (10001x10001), so only the variances of the gene expressions are kept. This certainly impaired the quality of generated synthetic data but made generating many more data points within a reasonable time possible. A synthetic cell is generated by sampling each gene expression from a Gaussian centered on the bulk point's value, and with the corresponding variance. If the sampled value is smaller than zero or larger than the saved max value, it is resampled. This process is repeated several times for each bulk data point. Figure 3 provides an overview of this procedure.

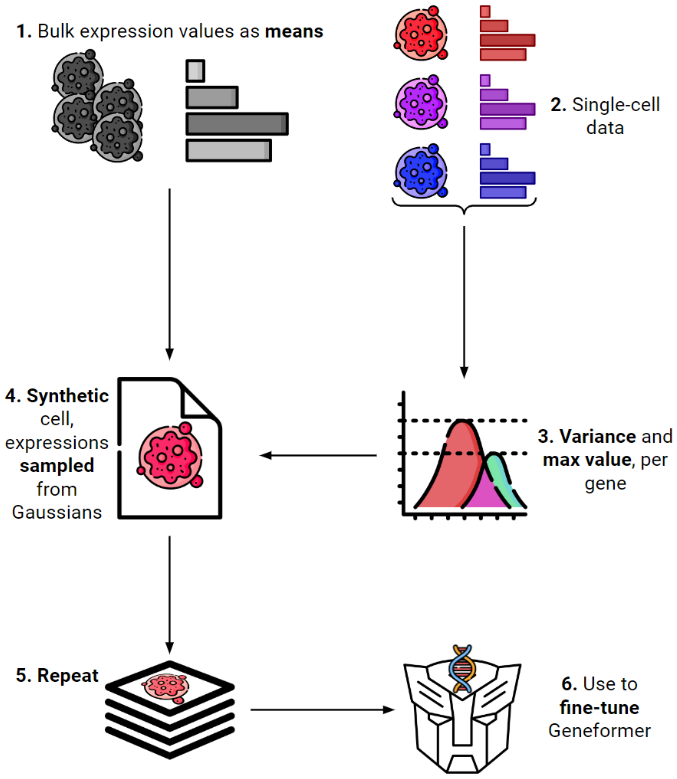
1. Bulk expression values as means
2. Single-cell data
3. Variance and max value, per gene
4. Synthetic cell, expressions sampled from Gaussians
5. Repeat
6. Use to fine-tune Geneformer

Figure 3: *An overview of the synthetic data generation process. Bulk RNA-seq data is used directly as a base for the synthetic cells, while single-cell data provides statistical metrics for the generator. The process is repeated a desired number of times. This figure has been designed using images from Flaticon.com.*

2.4 Experimental setup and procedure

All experiments follow a similar structure. Firstly, the pure single-cell dataset is prepared by either aggregating into pseudo-bulk or generating synthetic data. One in nine (11.1%) data points are marked as the validation set, while the rest are marked as the training set. The test set put aside during pre-processing is used to test the model's performance. Geneformer is fine-tuned on the training data, with the validation set being used for the model's built-in k-fold cross-validation. The fine-tuning shifts Geneformer's attention toward genes that more effectively differentiate the cell classes, which increases the prediction performance. A confusion matrix is generated, and the overall accuracy and F1 score are computed. Figure 4 outlines this approach.

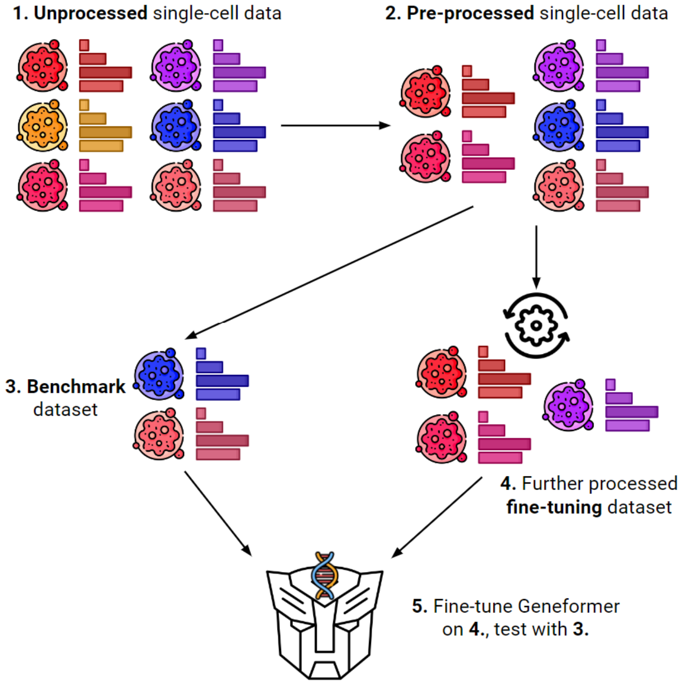
1. Unprocessed single-cell data
2. Pre-processed single-cell data
3. Benchmark dataset
4. Further processed **fine-tuning** dataset
5. Fine-tune Geneformer on **4.**, test with **3.**

Figure 4: *A simple scheme of the experimental setup. Original single-cell data is pre-processed and split into a test/benchmark set and training set, which can be further processed. The latter is used to fine-tune Geneformer, while the former evaluates its performance. This figure has been designed using images from Flaticon.com.*

3 Results

Geneformer is fine-tuned on a number of generated datasets and its prediction performance is measured on a single-cell test set. The fine-tuning datasets are created through either the process of aggregating single-cell data into pseudo-bulk data points, or generating synthetic single-cell-like data. Some generated synthetic datasets are also combined with the single-cell training set and used for fine-tuning together.

3.1 Pseudo-bulk data generated by aggregating single-cell information

Figures 5 and 6 show the accuracies and F1-scores, respectively, of the evaluation of Geneformer for the full five-label cell classification problem. The model was fine-tuned on datasets generated through the three aggregation methods: summing, exact averaging, and approximate averaging for different aggregation factor values.

A clear relationship between the performance metrics and the bulkiness of the fine-tuning dataset can be observed; the bulkier the data, the worse Geneformer performs. This holds for all three aggregation methods. This provides strong evidence in support of the hypothesis. As both the accuracy and F1-score drop below 0.7 already for a **k**=5, it appears likely that bulk data in general is not suitable for fine-tuning Geneformer for cell classification. Similar results for all aggregation factors have been obtained in different runs with varied hyperparameter configurations.

Another piece of evidence against the feasibility of effectively using bulk data to fine-tune Geneformer is the validation metrics recorded during the model's fine-tuning. Especially for the two most bulky datasets, the validation accuracies and F1-scores often exceeded 0.9. A follow-up experiment was conducted in which the validation set was boosted with some single-cell data. This was intended to verify whether the model could be made to perform better, even if the training set remained fully pseudo-bulk. The resulting performance metrics did not diverge significantly from the baseline, supporting the hypothesis.

For all used aggregation factors, all three differently-aggregated pseudo-bulk datasets tended to yield very similar performance metrics. Exceptions included the approximate average data set for **k**=10, which performed noticeably better than the others in terms of both the accuracy and F1-score. This is likely to have been caused by the inherent randomness of the fine-tuning process and does not provide evidence in support of this particular aggregation approach over the alternatives. The results do prove that there is no meaningful difference in terms of the type of pseudo-bulk data used, which indicates that Geneformer is resilient against variations in overall magnitude of the gene expressions between the training and test sets.

3.2 Synthetic single-cell data generated from a bulk dataset

The primary experiments conducted with synthetic single-cell data were first using a purely synthetic fine-tuning dataset, and second, adding some synthetic data points to real single-cell data to verify whether such padding could affect Geneformer's performance in cell classification.

Figure 7 shows the confusion matrix of Geneformer fine-tuned with a purely synthetic, two-label dataset (SAHA and Untreated.) Both of the accuracies exceed 0.8, showing that the synthetic data can feasibly be used to effectively fine-tune Geneformer. Further experiments with a full five-label dataset, however, resulted in an average accuracy and F1-score of no more than 0.45 each, providing strong evidence against the general usability of the generated synthetic data. Overall, it appears that synthesizing all of the fine-tuning data has some potential for simpler cell classification problems, but further work on the generation method is needed to expand its applicability to more complex problems.

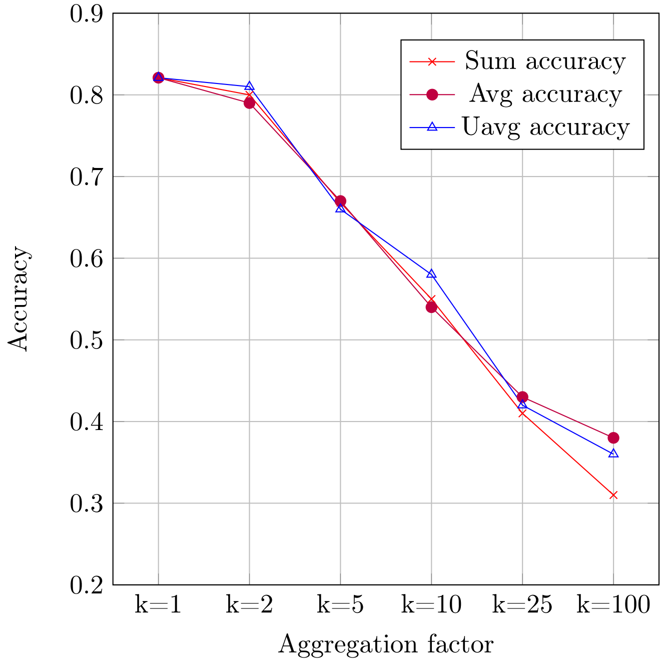

Figure 5: *The measured average accuracy of Geneformer over all five labels. "Sum accuracy" refers to the accuracy of the model fine-tuned on data aggregated by summing. "Avg accuracy" refers to the accuracy when fine-tuned on known-average aggregated data, while "Uavg accuracy" refers to the unknown-average aggregated case.*

The final experiment was conducted by adding 700 synthetic data points for the BMS (100 per dose) and Untreated classes. The model was fine-tuned on the combined dataset and each class's accuracy was compared to the baseline outcomes of fine-tuning Geneformer on pure single-cell data. Table 1 shows these accuracies for four synthetically-augmented datasets and the pure single-cell one (top row.) While Dex and SAHA appear to have not been meaningfully affected in any run of the experiment, more noticeable variance is observed in the accuracies of BMS, Untreated, and, interestingly, Nutlin. The addition of synthetic data points tended to slightly increase the accuracy of BMS, while Untreated's was worse than the baseline in all cases. Nutlin's accuracy varied the most, from dropping to 0.68 to growing to 0.81. These results show that using synthetic data together with single-cell does not reliably or meaningfully improve the model's performance for any class. This could be caused by the low quality of the synthetic data, or influenced by the significant overlap between much of the Nutlin, BMS, and Untreated classes.

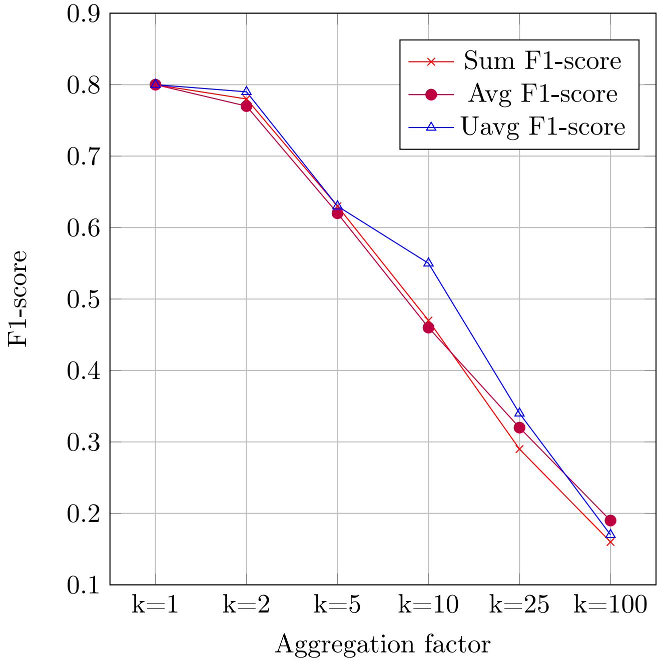

Figure 6: *The measured F1-score of Geneformer over all five labels. "Sum F1-score" refers to the summing aggregated fine-tuning data, "Avg F1-score" refers to the known-average aggregated data, while "Uavg F1-score" refers to the unknown-average aggregated data.*

Untreated	Nutlin	Dex	BMS	SAHA
0.68	0.73	0.94	0.70	0.92
0.62	0.68	0.93	0.66	0.88
0.64	0.72	0.94	0.73	0.90
0.57	0.81	0.92	0.73	0.93
0.61	0.79	0.94	0.76	0.91

Table 1: *Table showing the accuracies of Geneformer's prediction for each cell class for five datasets. The first row contains the metrics obtained by fine-tuning Geneformer on single-cell data, while the other four are separate synthetically-padded datasets. Notably, the model's overall performance is never meaningfully better than the baseline.*

4 Discussion

The experimental outcomes strongly support the hypothesis that using bulky data to fine-tune Geneformer is not feasible. Further, it is shown that while synthetic data generation has some potential in small, simple problems, like a two-label cell classification, it fails to generate data that can effectively fine-tune Geneformer for more complex tasks. Additional topics that require analysis are the limitations of the chosen approach to the topic, the used data, and the model's own restrictions.

Bulky dataset fine-tuning: Across all experiments where Geneformer was fine-tuned on aggregated pseudobulk data, a significant negative impact of the dataset's bulkiness on the model's prediction performance was observed. Given the extensive evidence, as well as the con-

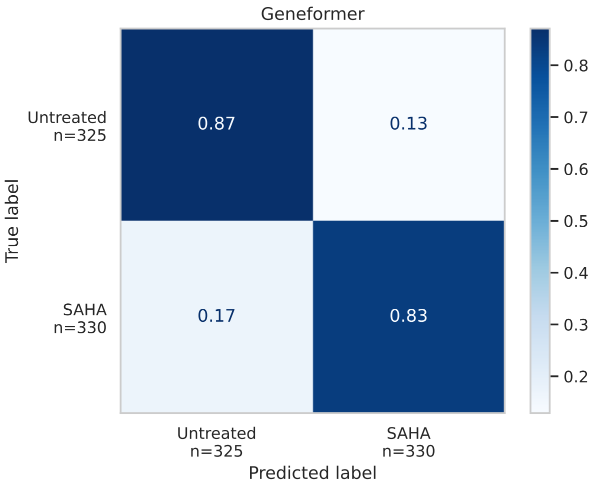

Figure 7: *The confusion matrix between SAHA and Untreated of Geneformer fine-tuned on synthetic single-cell data only.*

text of the chosen problem being cell label prediction, the hypothesis that bulky RNA-seq data is generally not a feasible choice of fine-tuning dataset for Geneformer can be conclusively confirmed.

Synthetic data generation: The generated synthetic data has shown minor promise in very simple applications, like a two-label classification problem, but did not deliver acceptable results in more complex tasks. Furthermore, adding such data to a single-cell dataset in an attempt to improve Geneformer's performance failed to affect it in a meaningful or consistent manner. One reason for these results could be the low quality of the synthetic data due to the chosen generation method. To conclusively determine whether the approach of synthesizing single-cell data from a bulk dataset is a dead-end, more advanced generation methods should be attempted. A good first future attempt could be to utilize the entire covariance matrices of the gene expression data, and not just the variances. The choice to only keep the latter made it feasible to generate a sufficient number of data points, but necessarily broke the inter-gene expression relationships. This has probably significantly impaired Geneformer's ability to successfully use this data, as it relies on these co-dependencies for making predictions.

This avenue of research does not appear to be meaningfully explored in literature. A paper introducing a method to generate synthetic single-cell data from bulk has been published very recently [13] and is currently pending review. Attempting to replicate the approach of the authors would be another interesting future research topic.

Chosen approach to the topic: The question of the feasibility of using bulk data to fine-tune Geneformer has been explored from the perspective of either using it directly as input, or as a source for synthetic data. However, a number of possible alternative approaches exist, which could have yielded different conclusions. For example, bulk data has been successfully used to improve the quality of single-cell data by filling in the gaps in gene expressions recorded in the latter [14]. This method of foregoing the direct use of bulk data to instead utilize it to augment the single-cell dataset would have significantly changed the outcomes of this research. Overall, it needs to be acknowledged that only the feasibility of using bulk data directly to fine-tune Geneformer has been explored and refuted.

Choice of problem: Choosing cell classification problems specifically naturally influenced the usability of bulk data for fine-tuning Geneformer. For this type of cell-level task, cell-level training data is logically needed; that is, scRNA-seq data, not bulk RNA-seq. However, there exist alternative problems for which the opposite is true. For example, if the tissue class were to be predicted, single-cell data would likely prove less useful for fine-tuning that bulk.

Limitations of the source data: Only one single-cell dataset, Sciplex2 [12], has been used over the course of this research. To provide more support to the presented conclusions, the experiments should be repeated on other single-cell datasets, for example [15, 16]. Furthermore, not real-life bulk data was utilized over the course of this research; all was generated by aggregating Sciplex2 single-cell data. A bulk dataset like the Cancer Cell Line Encyclopedia [17] could be a potential source for such data.

The model's problems: While Geneformer is an advanced model, it is affected by a few problems that influence how it can be used experimentally. For example, there appears to be a bug in the source code that results in a crash whenever fewer than 10 data points are provided for a particular class. The model has built-in k-fold cross-validation functionality, which is, however, restricted to either one or five folds. In addition to that another bug seems to be present, which causes an error whenever the five-fold setting is chosen. This means that the only possible choice is a single fold, significantly limiting cross-validation's usability. Lastly, the hyperparameters used during fine-tuning, such as the learning rate, can be automatically optimized by Geneformer. While conducting the experiments, however, a single optimization stage took several hours to complete, making it infeasible to do at the necessary scale. Altogether, the limitations of Geneformer did not make it impossible to answer the research questions but did negatively influence what evidence could be gathered from the experiments.

5 Conclusions and Future Work

In conclusion, the experimental results strongly support the hypothesis that fine-tuning Geneformer on bulk data negatively impacts its performance in cell label prediction problems. It is further shown that the bulkier the fine-tuning dataset, the worse the model's performance. With as few as five cells within a group, a significant drop in all performance metrics is identifiable. This shows that Geneformer is unlikely to be effective in cell classification if fine-tuned on a bulky dataset. The experiments further confirm that generating synthetic data is a feasible method of creating datasets that are effective in fine-tuning Geneformer for two-label problems. The model did not perform well in more complex tasks, such as including five labels. It is likely that the synthetic data generation method is too simplistic to create accurate approximations of real single-cell data. Furthermore, the addition of synthetic data to certain underrepresented classes did not result in a meaningful change in Geneformer's performance, suggesting that padding a pure single-cell dataset with synthetic data is not an effective way of boosting Geneformer's prediction capability.

There are several topics of particular interest for future research concerning using bulk data for fine-tuning Geneformer. Firstly, the experiments should be repeated on more datasets, especially more expansive ones. Secondly, a different approach to utilizing bulk for the fine-tuning process should be explored, for example by using it to improve the quality of existing single-cell data. Lastly, synthetic data generation should be explored further, in particular if a more sophisticated approach is implemented. A good starting point would be utilizing the full covariance matrices of the genes instead of only their variances.

6 Responsible Research

The three primary ethical considerations regarding this paper are the data sensitivity, the environmental and societal impact of the research and its outcomes, and the reproducibility of the study.

Data sensitivity: Geneformer is a tool intended for use in biomedicine, thus using real-world, human-sourced data to train on. When working with such models it is important to be aware of how sensitive this type of data is. Its source must be ethical and trustworthy, especially given that this data might inform the model's decision-making in situations such as patient treatment, or developing new drugs against cancer. In this paper, another potential issue is the generation of pseudo-bulk and synthetic single-cell data. Fortunately, the base data was sourced from the Sciplex2 database, ensuring its correctness and ethical procurement. The methods used in this paper also did not generate data to be used in reality, but rather to explore the model's limitations and the potential use of bulk data for fine-tuning. Because of this, as long as all the data has a proper source, and the employed data processing methods are well-understood, it is unlikely that any negative impacts may stem from how sensitive data was handled in this research.

Environmental and societal impact: With the recent proliferation of powerful AI systems, new light has been shed on how energy-expensive training these models is. This causes concerns of both environmental and societal nature. Firstly, most of the energy generated today still comes from fossil fuels, meaning that training powerful AI directly contributes to worsening the impact of climate change and pollution. Secondly, energy prices are increasing alongside the demand, which likely affects people living in poverty disproportionately. Geneformer shares its architecture with many of these models, and like them requires a lot of energy to be trained. During this research, dozens of hours of GPU time were used. While its impact is most likely minimal, it has to be acknowledged.

Research reproducibility: The ability of peers to reproduce scientific research is crucial, particularly in fields as sensitive as biomedicine. To ensure that, the author has provided detailed descriptions of all steps taken during data processing, the generation of aggregated datasets, and synthetic information from bulk data. The dataset from which all original data was sourced, Sciplex2, is publicly available online, as is the paper that introduced it [12]. Likewise, Geneformer can be downloaded via Huggingface [1]. The model's creators provided several examples of how to use it, which were used as templates for the experiments in this paper. Taken together with the experimental setup presented in previous sections, this should provide sufficient information to ascertain that the results can be reproduced.

References

[1] C. Theodoris, L. Xiao, A. Chopra, *et al.*, "Transfer learning enables predictions in network biology", *Nature*, vol. 618, pp. 616–624, 2023. DOI: https://doi.org/10.1038/s41586-023-06139-9.

[2] J. Goecks, V. Jalili, L. M. Heiser, and J. Gray, "How machine learning will transform biomedicine", *Cell*, vol. 181(1), pp. 92–101, 2020. DOI: https://doi.org/10.1016/j.cell.2020.03.022.

[3] S. Ekins, A. C. Puhl, K. M. Zorn, *et al.*, "Exploiting machine learning for end-to-end drug discovery and development", *Nature Materials*, vol. 18, pp. 435–441, 2019. DOI: https://doi.org/10.1038/s41563-019-0338-z.

[4] A. Vaswani *et al.*, "Attention is all you need", in *Advances in Neural Information Processing Systems*, vol. 30, 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[1] https://huggingface.co/ctheodoris/Geneformer

[5] H. Zhou, Y. Yu, C. Wang, *et al.*, "A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics", *Nature Biomedical Engineering*, vol. 7, pp. 743–755, 2023. DOI: https://doi.org/10.1038/s41551-023-01045-x.

[6] Z. Huang, P. Zhang, and L. Deng, "Deepcovdr: Deep transfer learning with graph transformer and cross-attention for predicting covid-19 drug response", *Bioinformatics*, vol. 39(39 Suppl 1), pp. i475–i483, 2023. DOI: https://doi.org/10.1093/bioinformatics/btad244.

[7] Y. Lieberman, L. Rokach, and T. 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*, vol. 13, no. 10, pp. 1–16, 2018. DOI: 10.1371/journal.pone.0205499. [Online]. Available: https://doi.org/10.1371/journal.pone.0205499.

[8] X. Shao *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*, vol. 49, e122, 21 2021. DOI: https://doi.org/10.1093/nar/gkab775.

[9] K. R. Kukurba and S. B. Montgomery, "Rna sequencing and analysis", *Cold Spring Harbor protocols*, pp. 951–969, 2015. DOI: https://doi.org/10.1101/pdb.top084970.

[10] D. Lähnemann, J. Köster, E. Szczurek, *et al.*, "Eleven grand challenges in single-cell data science", *Genome biology*, vol. 21(1):31, 2020. DOI: https://doi.org/10.1186/s13059-020-1926-6.

[11] Y. Im and Y. Kim, "A comprehensive overview of rna deconvolution methods and their application", *Molecules and cells*, vol. 46(2), pp. 99–105, 2023. DOI: https://doi.org/10.14348/molcells.2023.2178.

[12] S. R. Srivatsan *et al.*, "Massively multiplex chemical transcriptomics at single-cell resolution", *Science*, vol. 367, pp. 45–51, 2020. DOI: https://doi.org/10.1126/science.aax6234.

[13] H. J. Cho, E. Xie, A. Zhang, and S. Bekiranov, "Generating synthetic single cell data from bulk rna-seq using a pretrained variational autoencoder", *bioRxiv*, 2024. DOI: 10.1101/2024.05.18.594837.

[14] T. Peng, Q. Zhu, P. Yin, *et al.*, "Scrabble: Single-cell rna-seq imputation constrained by bulk rna-seq data", *Genome biology*, vol. 20, p. 88, 2019. DOI: https://doi.org/10.1186/s13059-019-1681-8.

[15] K. Prazanowska and S. Lim, "An integrated single-cell transcriptomic dataset for non-small cell lung cancer", *Scientific Data*, vol. 10, no. 167, 2023. DOI: https://doi.org/10.1038/s41597-023-02074-6.

[16] J. McFarland, B. Paolella, A. Warren, *et al.*, "Multiplexed single-cell transcriptional response profiling to define cancer vulnerabilities and therapeutic mechanism of action", *Nature Communications*, vol. 11, no. 4296, 2020. DOI: https://doi.org/10.1038/s41467-020-17440-w.

[17] B. Institute, *Ccle 2019*, Accessed: April 25, 2024. [Online]. Available: https://portals.broadinstitute.org/ccle/data.