Attention on Key Genes

Unveiling Key Genes For Cancer Cell-state Predictions of the Geneformer Model by Inspecting the Attention Weights

Marian A. Trützschler von Falkenstein

Supervisor(s): Marcel J. T. Reinders, Niek Brouwer

[1]**EEMCS, Intelligent Systems, 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: Marian A. Trützschler von Falkenstein
Final project course: CSE3000 Research Project
Thesis committee: Marcel J. T. Reinders, Niek Brouwer, Merve Gürel

Abstract

Geneformer is a transformer which is pretrained on Geneformer-30M, a dataset consisting of 29.9 million healthy cells. This paper focuses on how Geneformer shifts its attention, when fine-tuned on a dataset of cancer cells, whose gene expression is expected to be distinct, and which genes are key when making cell-state predictions within such an environment. In this paper we compare the shift in attention, and which genes receive the most attention, in a weight-based analysis.

The observed shift in attention was significant, however the accuracy of the prediction increased minimally. The attention weights were mapped back to individual genes, which showed that while Geneformer shifts its attention towards key genes, largely it is still subject of a batch-effect, namely the amount of expressed genes. Further research into designing a data representation of consistent size might be beneficial.

1 Introduction

Cancer was cause of nearly one out of six deaths globally in the year 2020 [1]. According to the World Health Organisation, effective treatment of cancer starts with early detection. Machine learning models could help to identify potential therapeutic targets with increased throughput [2]. However, machine learning models are faced with the need for diverse datasets for training. In this context, transfer learning emerges as a powerful approach in computational intelligence [3], an example is the so-called *Geneformer* model [4].

Geneformer is pretrained on a large dataset of general data, Genecorpus-30M, which is comprised of single-cell transcriptomic data of 29.9 million cells across diverse cell lines. During pretraining, the model acquires foundational knowledge of gene network biology. While the knowledge gathered during pretraining is retained, the model can then be fine tuned for downstream tasks, using a narrow amount of task-specific data.

The combined knowledge gained during pretraining and fine-tuning allows the model to make cell-state predictions. Since the model has learned which genes are correlated, and possibly causally related during pretraining, the model can potentially be used to study the effect on cell-state of a perturbation, the alteration of a gene's activity.

Within the model, the transciptome of a single cell is encoded as a rank-value encoding, wherein genes are ranked by their expressing within that cell, normalised by their expression over the entire corpus. This encoding therefore gives priority to genes that distinguish cell state, as omnipresent housekeeping genes are deprioritised.

The model is trained using a self-supervised learning method wherein 15% of genes within a transcriptome are masked and the model is tuned towards predicting a masked gene in the specific cell state by using the other unmasked genes as the context. The aim of this approach is to make the model context-aware. In other words, the model learns the interaction between genes. Genes of a single-cell transcriptome are encoded as rank values. These encodings then pass through six layers of transformer encoder units, which then outputs the contextual gene embeddings, or a prediction. The attention weights that the model uses to make a prediction can also be inspected. The transformer encoder unit consists of a self-attention layer, normalization layers and a feed-forward neural network. See figure 2 and section 3.1 for an overview of the architecture of the model. A more in depth description of the architecture can be found in section 3.1.

Since Geneformer is pretrained only with transcriptomic data of healthy cells, it is reasonable to question whether it is possible for the model to learn how to interpret the embeddings of cancer cells, since the transcription factors of such a cell may be vastly different. To address this, the model was trained on data of the Sciplex-2 dataset. which contains transciptomic data of cancer cells which have been subjected to various drugs. The question has been formalised by the aim of this research: '*How does Geneformer shift its attention when fine-tuned on a dataset of cancer cells and which genes are key when making cell-state predictions?*'

Interpretation of the attention weights of the model remains largely unexplored in the Geneformer paper [4]. Here, we analysed the attention weights of the Geneformer model to uncover which part of the input data is considered the most important by the model when predicting cell states of drug-treated cancer cells. Analysis of these attention weights gives an insight on how the model adapts dynamically to distinct inputs [5] [6] [7]. Furthermore an analysis of the parts of the input data that is thoroughly attended by the model could reveal key genes, which could help identify target genes for perturbations. Additionally the analysis could provide further evidence for the context-awareness of the model, for example if the most attended genes correspond to biological pathways of genes affected by the drug or the disease. This evidence could be used as a foundation for further research into the application of transformer-based models in the field of cancer research.

2 Methodology

2.1 The sciplex-2 dataset

The sciplex-2 dataset [8] was used to finetune the model. Sciplex-2 contains single-cell transcriptomic data of cancer cells that have been subjected to the drugs Nutlin-3a, Dexamethasone, SAHA, and BMS. The drugs have been applied in the following dosages in µM [0; 0.25, 1.25, 2.5, 12.5, 25, 125, 250]. Since the dataset contains only diseased cells, this data is dissimilar to Genecorpus-30M, which contains only healthy cells, and can therefore be considered suitable to asses how Geneformer adapts to thus far unseen data.

2.2 Preprocessing

Only cells subjected to the Nutlin-3a drug were used for training the model. The training objective was to identify the dosage each individual cell was subjected to. Cells with a dosage larger than, or equal to 50 µM were excluded from the training and test set, since their presence within the dataset was disproportionate to the rest of the classes. Additionally dosage 0.25 µM was excluded from the training and test data, since such a small dosage has very little effect. A UMAP of the data into two-dimensional space is depicted in figure 1

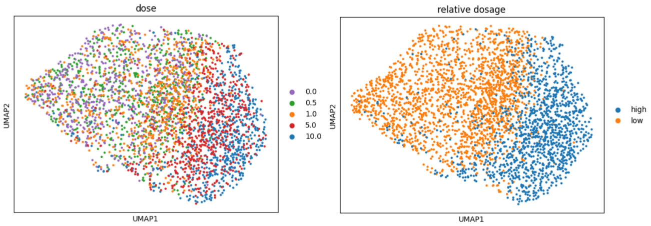


Figure 1: UMAP into two-dimensional space of the data after prepossessing. On the left, the dosages have been labeled, while on the right the dosages are clustered as follows: low dosage, [0 - 2,5] µM, and high dosage [12,5 - 25] µM

.

2.3 Fine tuning geneformer for cell classification

The Geneformer model was fine-tuned to predict the dosage of the Nutlin-3a drug to which the cell was subjected. Application of the Wilcoxon test for differential gene expression [9] showed which genes were most deferentially expressed within cells of each distinct dosage. The fine-tuning strategy allows us to assess whether Geneformer pays attention to the same genes. Additionally the test showed that the data mainly clustered in two groups: low dosage, [0 - 2,5] µM, and high dosage [12,5 - 25] µM. The model was also fine-tuned to distinguish between the two groups. For each of the scenarios, the model was trained once with the pretrained layers frozen, which means that the weights within these layers do not change. This setup allows us to asses which genes are regarded as important based on Genecorpus-30M. Then, the model was trained once with all layers unfrozen to observe how Geneformer shifts its attention to adapt to thus far unseen input types in order to boost accuracy.

2.4 Interpreting the attention weights

The pretrained Geneformer model consists of 6 transformer encoder layers, each with four attention heads. The attention weights of each individual head were extracted and analysed for both of the training settings. The shift in attention was analysed by weight-based analysis, wherein was inspected which part of the encoding received the most attention. The significance of these changes were quantified by norm-based analysis. These two methods were described in detail in sections 3.3 and 3.4, respectively. This method revealed which parts of the input the model considers of importance across different training schemes.

3 Background

3.1 The underlying architecture of the Geneformer model

Single-cell transciptomic is transformed into a rank-value encoding of size 2048 wherein genes are ranked based on their expression within a cell, divided by the median expression within Genecorpus-30M. Each gene is embedded as vector of size 256, which captures the context of the surrounding genes. Initially, in the input data, the gene expression is a sparce matrix containing 10,000 values, however the largest count of genes expressed within the dataset was only 1430. This means that a large portion of the encoding is made up of padding.

The encoding passes through six layers of transformer encoder units of the underlying model BERT [10]. The encoder is depicted in detail in figure 2. Such an encoder starts with a self-attention layer consisting of four attention-heads which apply self-attention to the data in parallel. The input embedding is then added to the output of the self-attention layer via residual connection, and normalised. The normalised output passes through a fully-connected feed forward neural network (FFN), which connects the output of the separate attention heads and adds non-linearity by applying the GELU [11] activation function. Finally, the input of the FFN is added to the output of the FFN, via a second residual connection, and normalised. The output of this last step is the hidden state, the input for the transformer encoder unit of the next layer.

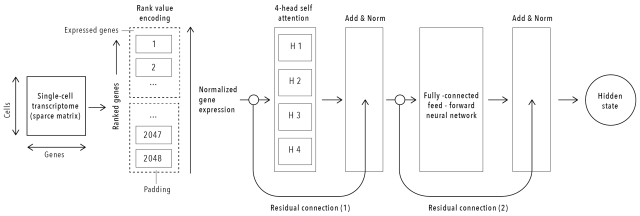


Figure 2: A detailed overview of the transformer encoder unit. A single-cell transciptome is encoded as a rank-value. The transformer encoder applies the attention mechanism and normalises the output before passing it through a fully-connected feed-forward neural network and normalising again. The encoder outputs a hidden state.

.

3.2 The attention mechanism

The self-attention mechanism is core to the transformer architecture and thus Geneformer [4]. Through this mechanism the model learns to attend to relevant information from the input [5], in this case the rank value encoding. With the attention mechanism, the input vector is transformed according to the following formula, as proposed by [5]:

$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (1)$$

,
wherein Q, K, and V represent the Query, Key and Value matrix, and d_k the dimension of Q and K.

Since Geneformer makes use of self-attention, Q and K represent a mapping of the input vector into query and key space. In particular Q and K are a result of

$$Q := xW_Q + b_Q \tag{2}$$

$$K := xW_K + b_K \tag{3}$$

wherein x represents the input vector, W and b the weight and bias term of the key, and query matrix. Through the attention mechanism, Geneformer analyses the correlation between the Q and K matrices, and thus aims to capture the correlation between the genes in the input.

The V matrix represent the re-projection into the original space, and is a result of

$$V := xW_V + b_V \tag{4}$$

The attention weights are updated using gradient descent. During fine-tuning the model uses cross-entropy loss.

3.3 Weight-based analysis

We extracted the attention weights directly from the underlying model BERT. Specifically this means A (from equation 1) is extracted and analysed. This analysis shows how the attention shifts in the trained genes and which genes receive the most attention.

3.4 Analysis using vector norms

While a weight-based analysis of the attention weights can reveal which parts of the input the model are most heavily attended by the model when making a prediction [7]. However, Kobayashi et al. [6] showed that a weight-based approach does not account for the scale difference between individual vectors. A larger attention-output will have a larger effect on the prediction outcome, therefore the observed changes were quantified by analysis of $||xA||$.

4 Experimental Setup and Analysis

4.1 Training and validation

The model was trained twice on the sciplex-2 dataset which was preprocessed as described in section 2.2. Both times the model was trained for 5 epochs, validated with one fold, and evaluated with an independent test set. First, the model was trained once using linear probing, meaning that all 6 layers of the pretrained Geneformer model were frozen and their weights were not updated. Then the model was trained once without freezing any layers. This approach aimed to compare attention weights of an untrained layer to a trained model, while also a measure of the accuracy of each model could be provided.

The training objective was to classify the dosage of the drug. The confusion matrices of the respective trained and untrained model are shown in figure 3. Comparing the predictions of the model trained by linear probing against the model in which the layers were also trained. The linearly-probed model was already able to distinguish between high [12.5, 25] and low [0, 1.25, 2.5] dosage of the drug, as most of the confusion is within these groups. After the layers were trained the model does not improve upon this. The accuracy of both models is 0.5..

However, when the models were fine-tuned to predict the relative dosage, high or low, both models performed with increased accuracy. Without training the layers, Geneformer reached an accuracy of 0.85, while training the layers resulted in an accuracy of 0.91.

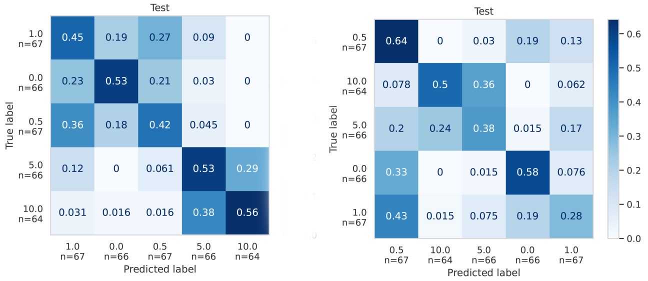


Figure 3: A comparison of the confusion matrix of the fully trained model (left) against the model trained with linear probing (right). Both models were trained and tested on the same respective train and test sets. Both models reached an accuracy of 0.5 only.

4.2 Attention weight comparison

The attention weights were extracted directly from the underlying BERT model [12] for comparison. We computed cosine similarity between the attention weights of each head of each layer, a visualisation is available in figure 4. This comparison showed that the largest differences between the attention weights are within the middle, and the last layer, presumably because the changes of the preceding layers propagate towards the subsequent layers. To verify this last assumption the model was fine-tuned once more with the first five layers frozen, meaning only the last layer was trained. When comparing the last layer of this model to the frozen layers, only a very small change was observed.

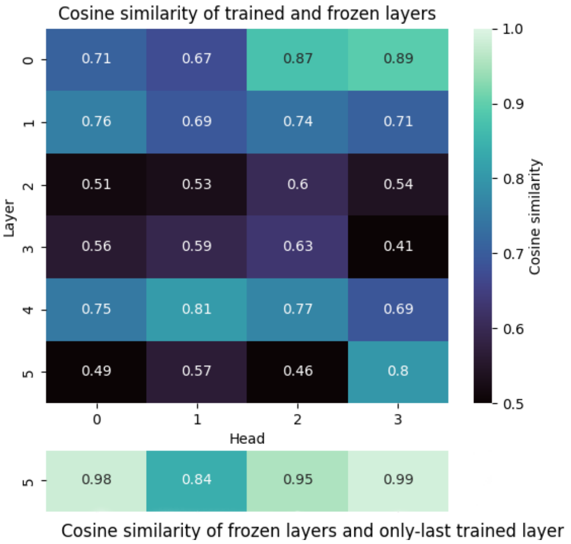


Figure 4: Heatmap showing the cosine similarity between the frozen layers against the fine-tuned layers, and below between the last layer of the model with the first five layers frozen between the corresponding frozen layer.

.

4.3 Norm-based analysis

To quantify the significance of the changes, $||\sum xA||$ of each layer was compared. This analysis showed that the changes in the first layer had a smaller impact to the prediction outcome compared to the subsequent layers. As the embeddings move through the layers of the model, they take in more meaning from the context and have become more distinct. In particular, $||A||$ within layer 1 is the largest, which causes the changes in the subsequent layers to be more significant.

.

4.4 Overall weight-change comparison

To uncover in which area of the encoding the largest changes occur, for each model, the attentions matrices of all heads were summed together, then the column-wise sum of the resulting matrix was calculated, which produced a one dimensional array, containing the total attention each rank of the encoding received.

Figure 6 shows the percentage difference between the frozen layers and the trained layers. The largest changes in attention occur within the rank 210. Genes ranked below the median lowest ranked gene of a single cell in the dataset, received less attention on average. However noticeable is a small peak in increased attention just below the median. Possibly the model is observing the length of the genome.

.

4.5 Evaluation of key genes

In order to evaluate which genes are key, the attentions were mapped back to the corresponding gene within one rank of a single transcriptome. This method shows the attention the

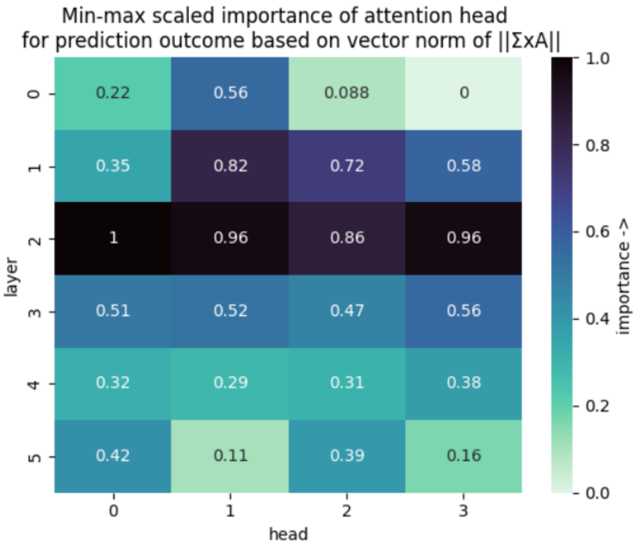


Figure 5: Overall percentage difference in attention over all heads and layers between the frozen and trained layers. The largest changes occur within rank 200. The median amount of expressed genes of a single cell within the dataset was 683, ranks after this on average receive less attention after fine-tuning the layer, except in the ranks just below the median.

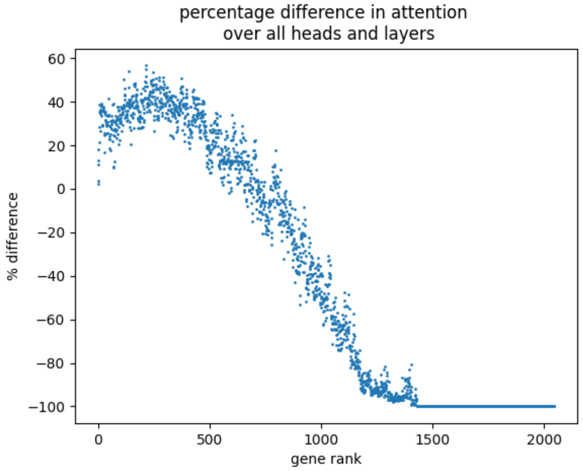


Figure 6: Overall percentage difference in attention over all heads and layers between the frozen and trained layers. The largest changes occur around within rank 210. The median amount of expressed genes of a single cell within the dataset was 683, ranks after this on average receive less attention after fine-tuning the layer, except in the ranks just below the median.

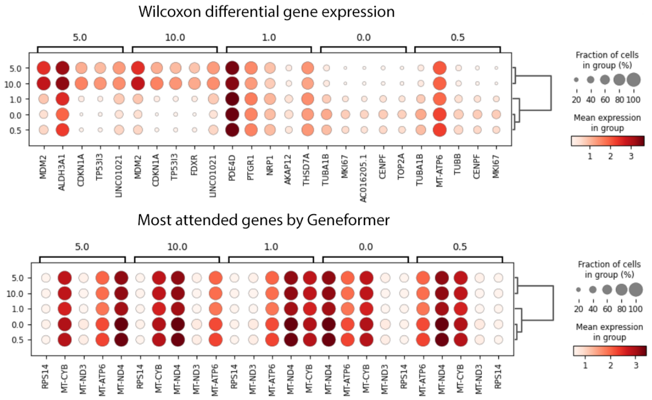


Figure 7: Most differentially expressed genes by result of the wilcoxon test compared to the most attended genes by Geneformer. MDM2 received the sixth most attention out of all genes.

model spends in total on a single gene. The result can be found in table 8 in the appendix A. Curiously, the model pays a large part of it's attention towards various mitochondrial genes, which are not directly affected by the p53 pathway which the Nutlin-3a drug is targeting, nor its downstream effect [13] [14].

At first glance, Geneformer pays significant attention to housekeeping genes. These genes often occupy the lowest ranks within the encoding. Possibly Geneformer is assessing the amount of genes expressed, just as shown in figure 6. It is also possible that Geneformer pays a lot of attention towards these genes as there is serious variance within the gene embedding in this area.

In figure 7 the genes which receive the most attention by Geneformer have been compared to the top 5 genes of the differential gene as found by the Wilcoxon signed-rank test. This analysis shows that the most attended genes do not show differential gene expression across the groups.

5 Discussion

5.1 Interpretation of the results

Analysis of the attention weights showed that when fine-tuning the layers caused a shift in attention, with the most significant changes being in layer 2. The largest shift attention was measured around rank 210, which corresponds to the average rank of the MDM2 gene, the target of the Nutlin-3a drug [13], within the encoding. This shows an aptitude of Geneformer to identify the most important gene within the encoding. However, a significant change was also observed towards the median end of the non-padding sequence of the encoding, which could indicate that Geneformer, is looking at the amount of expressed genes as well. This could also indicate a batch-effect, contrary to what is proposed by [4].

Possibly another way of encoding the expressed genes which ensures a consistent length of the non-padding sequence, would help Geneformer focus on more relevant genes, which could boost accuracy in perturbation experiments. For example, PCA could be used to map the expression into consistent feature space such that the input length of each encoding stays the same. Geneformer was able to identify genes which are key for predictions, however there may still be some improvement to be made: Geneformer was unable to outperform the statistical test in terms of finding genes which were differentially expressed.

5.2 Limitations

Inspection into the attention weights of the model revealed which part of the input vector the model attends to when making predictions. However Jain and Wallace [15] showed that the relationship between the attention weights and the model output is unclear for NLP models. In their study they also found that attention does not necessarily correlate to feature importance, another measure such as attribution score, as proposed by [16] is an alternative. This score indicates how much a given input feature contributes to the output.

6 Responsible Research

6.1 Data Usage

Such that no persons privacy may come to harm, only publicly accessible data has been used for the purpose of this research. The data was extracted from the Sciplex-2 dataset which is publicly available online [8].

6.2 Bias

In this research we aim to assess how the Geneformer model adapts dynamically to new inputs by shifting it's attention to the important aspects of the input. The model has been fed with real-world data and thus simulates a life setting, but a very small dataset, which contained only a single cell-line. However, when implementing the model in practice it is important to understand when and why the model is failing and adjust accordingly, especially within limited-data conditions, as algorithmic bias is a result of non-diversity within the training data [17]. In order to combat this bias, characteristics that are legally or ethically protected, such as ethnicity, sex and age, can be safeguarded using auditing algorithms or fairness constraints [18].

6.3 Reproducibility and Open Science Principles

A detailed description of the methodology is available in section 2, while the experimental setup is presented in section 4. Data input and code is available, see apppendix C. This paper, which contains detailed information on the research has been made publicly available.

7 Conclusions and Future Work

Weight-based analysis showed that Geneformer is able to adapt to thus far unseen input through fine-tuning and shift its attention to potentially relevant genes. This was done by training the model on the sciplex-2 dataset which contained cancer cells subjected to various drugs. Norm-based analysis showed that although the changes were significant, the accuracy of the fine-tuned model remained largely the same.

After the attentions were ampeed back to the input genes, observable was a noticeable portion of the attention which went to genes that might not be significant for the prediction outcome. This could be destructive for pertubation experiments, as the model does not focus largely on the differentially expressed genes across the classes. This neglect of differentially expressed genes could be attributed to the amount of genes expressed of the input single-cell transciptomic data, or the input size. Future work should focus on researching methods which combat variance in input size in a smart way. Possible option include subsampling the genes or applying PCA.

References

[1] W. H. Organisation, "Cancer." https://www.who.int/health-topics/cancer, Feb. 2022. [Accessed 21-05-2024].

[2] A. M. Sebastian and D. Peter, "Artificial intelligence in cancer research: Trends, challenges and future directions," *Life*, vol. 12, p. 1991, Nov 2022. PubMed-not-MEDLINE.

[3] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[4] C. V. Theodoris, L. Xiao, A. Chopra, M. D. Chaffin, Z. R. Al Sayed, M. C. Hill, H. Mantineo, E. M. Brydon, Z. Zeng, X. S. Liu, and P. T. Ellinor, "Transfer learning enables predictions in network biology," *Nature*, vol. 618, pp. 616–624, Jun 2023.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[6] G. Kobayashi, T. Kuribayashi, S. Yokoi, and K. Inui, "Attention is not only a weight: Analyzing transformers with vector norms," 2020.

[7] O. Kovaleva, A. Romanov, A. Rogers, and A. Rumshisky, "Revealing the dark secrets of bert," 2019.

[8] V. Alexander, "scrnaseq exposed to multiple compounds." https://www.kaggle.com/datasets/alexandervc/scrnaseq-exposed-to-multiple-compounds, May 2021. Accessed: 08-05-2024.

[9] Y. Li, X. Ge, F. Peng, W. Li, and J. J. Li, "Exaggerated false positives by popular differential expression methods when analyzing human population samples," *Genome Biology*, vol. 23, p. 79, Mar. 2022.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

[11] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," 2023.

[12] H. Face, *BERT: Pre-trained Models for Natural Language Processing*, 2024. Accessed: 2024-06-03.

[13] T. Van Maerken, L. Ferdinande, J. Taildeman, I. Lambertz, N. Yigit, L. Vercruysse, A. Rihani, M. Michaelis, J. Cinatl, C. A. Cuvelier, J.-C. Marine, A. De Paepe, M. Bracke, F. Speleman, and J. Vandesompele, "Antitumor Activity of the Selective MDM2 Antagonist Nutlin-3 Against Chemoresistant Neuroblastoma With Wild-Type p53," *JNCI: Journal of the National Cancer Institute*, vol. 101, pp. 1562–1574, 11 2009.

[14] M. Fischer, "Census and evaluation of p53 target genes," *Oncogene*, vol. 36, pp. 3943–3956, Mar. 2017.

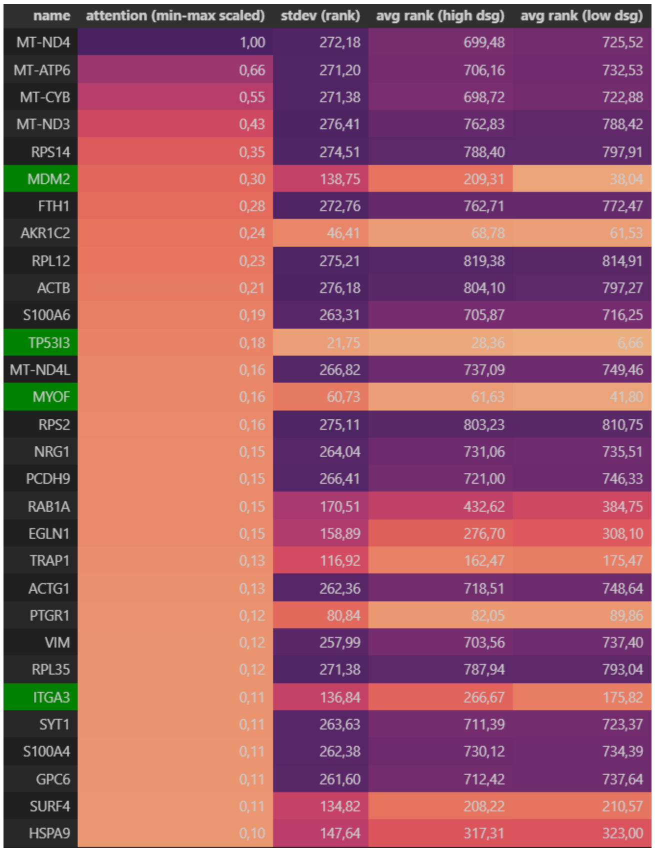
[15] S. Jain and B. C. Wallace, "Attention is not explanation," 2019.

[16] Y. Hao, L. Dong, F. Wei, and K. Xu, "Self-attention attribution: Interpreting information interactions inside transformer," 2021.

[17] N. Norori, Q. Hu, F. M. Aellen, F. D. Faraci, and A. Tzovara, "Addressing bias in big data and AI for health care: A call for open science," *Patterns (N Y)*, vol. 2, p. 100347, Oct. 2021.

[18] A. Agarwal, M. Dudík, and Z. S. Wu, "Fair regression: Quantitative definitions and reduction-based algorithms," 2019.

A Tables



name	attention (min-max scaled)	stdev (rank)	avg rank (high dsg)	avg rank (low dsg)
MT-ND4	1,00	272,18	699,48	725,52
MT-ATP6	0,66	271,20	706,16	732,53
MT-CYB	0,55	271,38	698,72	722,88
MT-ND3	0,43	276,41	762,83	788,42
RPS14	0,35	274,51	788,40	797,91
MDM2	0,30	138,75	209,31	38,04
FTH1	0,28	272,76	762,71	772,47
AKR1C2	0,24	46,41	68,78	61,53
RPL12	0,23	275,21	819,38	814,91
ACTB	0,21	276,18	804,10	797,27
S100A6	0,19	263,31	705,87	716,25
TP53I3	0,18	21,75	28,36	6,66
MT-ND4L	0,16	266,82	737,09	749,46
MYOF	0,16	60,73	61,63	41,80
RPS2	0,16	275,11	803,23	810,75
NRG1	0,15	264,04	731,06	735,51
PCDH9	0,15	266,41	721,00	746,33
RAB1A	0,15	170,51	432,62	384,75
EGLN1	0,15	158,89	276,70	308,10
TRAP1	0,13	116,92	162,47	175,47
ACTG1	0,13	262,36	718,51	748,64
PTGR1	0,12	80,84	82,05	89,86
VIM	0,12	257,99	703,56	737,40
RPL35	0,12	271,38	787,94	793,04
ITGA3	0,11	136,84	266,67	175,82
SYT1	0,11	263,63	711,39	723,37
S100A4	0,11	262,38	730,12	734,39
GPC6	0,11	261,60	712,42	737,64
SURF4	0,11	134,82	208,22	210,57
HSPA9	0,10	147,64	317,31	323,00

Figure 8: Genes ranked by the amount of attention received, min-max scaled. The genes are sorted by the total amount of attention received. Genes that were included by Fischer in the census p53 target genes [14], were marked in green.

B Figures

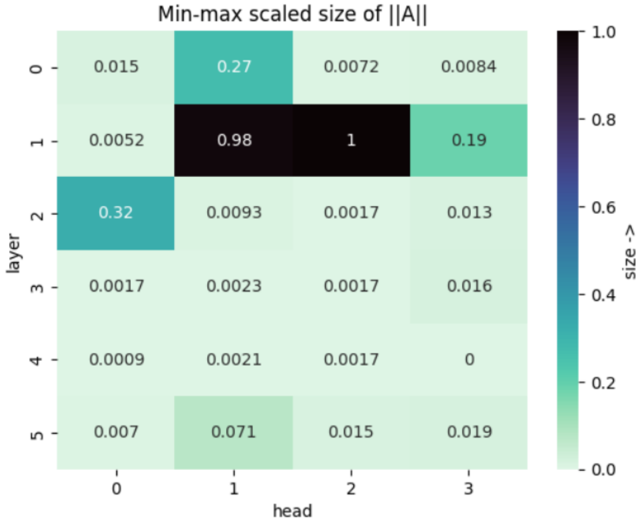


Figure 9: Heatmap plot of the min-max scaled attention norm, by far the second layer attention weights heave the largest norm.

C Code and Data Availability

Code and input data is availible. Please check out the Github repository.

The link to the repository will become available as the paper is uploaded to the Github repo.