# Evaluating Machine Learning Approaches for Predicting Drug Response in Cancer Cells

## A Comparative Analysis of Geneformer and Support Vector Machine

**Samuel Banas**[1]
**Supervisor(s): 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: Samuel Banas
Final project course: CSE3000 Research Project
Thesis committee: Marcel Reinders, Niek Brouwer, Merve Gürel

**Abstract**

Accurately predicting how cancer cells respond to drug treatment is important to advance drug development. This paper presents a comparative analysis of Geneformer, a deep-learning transformer pre-trained on transcriptomic data, and Support Vector Machine. Using the Sciplex2 dataset, which includes transcriptomic data from lung cancer cells treated with three drugs, both models were trained to predict the response of cancer cells to drug treatments.

This paper investigates how Geneformer and SVM perform in predicting the treatment label of cells across different drugs and doses, which drug doses are suitable for conducting single-gene perturbation experiments, how accurately can these experiments replicate drug effects, and what are the differences in results between Geneformer and SVM regarding their ability to identify significant genes affecting drug response.

Results indicate that while SVM generally achieves higher accuracy in predicting treatment labels of cells, Geneformer demonstrates better capability in identifying genes whose perturbations mimic drug effects. Geneformer's embeddings show significant shifts towards treated cell states after single-gene perturbations, indicating a deeper understanding of gene interactions in drug response. On the other hand, SVM's predictions rely more on differential gene expression. This comparative analysis underscores the strengths and limitations of each approach in modelling complex biological systems and predicting the drug response of cancer cells.

# 1 Introduction

Cancer is the leading cause of death worldwide, accounting for nearly one in six deaths in 2020, totalling 10 million [1]. As the field of medicine evolves, the use of machine learning methods in this field has become more predominant [2]. Predicting the response of cancer cells to drug treatment is a critical challenge in biomedical research, with significant implications for personalized medicine and drug discovery. The ability to accurately predict how cells respond to drugs has the potential to streamline the drug development process, reduce costs, and ultimately lead to more effective treatments for various diseases, including cancer. With the ability to model complex biological systems (such as gene networks) and predict drug responses based on transcriptomic data, machine learning methods have the potential to notably advance cancer drug research.

One promising approach involves using deep learning models like transformers, which have shown remarkable success in natural language processing (such as GPT-4 [3]) and are now being adapted for biological applications. This paper focuses on the Geneformer model, a 6-layer transformer neural network [4]. Pre-trained on transcriptomic data of 30 million cells, Geneformer is designed to learn the underlying relationships between genes, potentially allowing it to better predict the effects of drug treatment by understanding the downstream effects of single-gene perturbations (changes in individual gene expressions). However, while deep learning models like Geneformer show promise, their performance and interpretability compared to more traditional machine learning methods, such as Support Vector Machine, remain underexplored. Support Vector Machine has long been a staple in machine learning due to its robustness, low complexity, and effectiveness in high-dimensional spaces [5]. It has already been successfully applied to gene expression analysis [6]. Despite its success, there is a need for a direct comparative analysis with newer models like Geneformer to understand their relative strengths and limitations in the context of drug response prediction.

Using transcriptomic data of lung cancer cells treated by three different drugs from the Sciplex2 dataset, both SVM and Geneformer can be trained or fine-tuned to predict the treatment label of cells. These trained models can then be evaluated on their ability to understand and predict the effects of these drugs, by performing single-gene perturbation experiments. By changing the expression values of specific genes, and using these experiments to identify significant genes, we can compare them to

target genes of drugs from the Sciplex2 dataset, and comparatively evaluate the extent of drug effect understanding by the models.

This paper aims to answer the following research questions:

- What are the features and characteristics of the data included in the Sciplex2 dataset, and how can it be used to train the machine learning models? What preprocessing is required before the training?
- What are the accuracies of Geneformer and SVM in predicting the treatment label of cells? How do the accuracies vary based on the drug and dose?
- Which of the trained models (based on drug and dose) can be used for perturbation experiments? How can we simulate perturbations to replicate the effect of drugs? What data can be used for the experiment?
- What are the differences in the results of the experiments between Geneformer/SVM? Which aspects of the methods/data contribute to the results?

# 2 Scientific Background

## 2.1 Transcriptomic Data

Transcriptomics is the study of the transcriptome, which is the complete set of RNA transcripts transcribed based on the genome [7]. RNA transcripts are copies of gene sequences, made when genes are expressed. This field focuses on understanding how genes are regulated in different cells and conditions, and how these patterns of gene expression change in response to various factors, including drugs. In simpler terms, if we think of genes as instructions in a recipe book, transcriptomics is about studying which recipes (genes) are being read (expressed) and cooked (transcribed into RNA) at a particular moment. The patterns of these read instructions can tell us a lot about what is happening inside a cell at that time.

## 2.2 Genes, Diseases, and Drugs

Genes play a crucial role in disease treatment. Cancer, diabetes, heart disease, and many more are connected to gene expression, where some genes are overexpressed (producing more RNA) or underexpressed (producing less RNA) in diseased cells compared to healthy cells [8]. Many drugs are designed to interact with specific molecules in the body, especially proteins encoded by genes. By binding to the proteins, drugs can modify their activity, which can in turn change the expression of genes that interact with the proteins [9]. Studying the transcriptome of cells allows us to observe how gene expressions change when cells are treated with different drugs. By performing single-gene perturbation experiments, we can mimic the effects of drugs that affect specific genes, and use machine learning to understand their effects on a deeper level.

## 2.3 Differential Gene Expression

This paper uses differential gene expression as a benchmark to evaluate different models and metrics of identifying important genes through single-gene perturbation experiments. Differential gene expression (DGE) refers to the difference between the mean expression of a certain gene in treated cells and untreated cells. Therefore, a positive DGE of a certain gene means it is generally more expressed

in treated cells, while the opposite stands for negative DGE. In terms of the importance of genes in the treatment of cells, DGE can provide basic first-look importance based on expression values. It is, however, unable to pick up any relationships between genes, that may influence each other's expression when perturbed. Overlap of DGE-identified important genes with genes identified by other metrics and models therefore serves as a guideline for whether the evaluated technique managed to understand more about the gene interactions than simply gene expression. Genes with the highest (positive) DGE are used for the evaluation of overexpressions, and genes with the lowest (negative) DGE for deletions.

### 2.4 The Geneformer Model

Geneformer is an advanced deep learning model that uses transformer architecture to predict gene network dynamics, particularly in data-limited scenarios. Pre-trained on Genecorpus-30M, a dataset of approximately 30 million single-cell transcriptomes, Geneformer encodes gene expression as rank values to normalize data across the entire corpus, prioritizing contextually significant genes. The model comprises six transformer encoder units, each with a self-attention layer and a feedforward neural network. The self-attention mechanism enables Geneformer to weigh the importance of each gene within the context of the entire transcriptome, learning to focus on genes crucial for specific cellular contexts. During pre-training, Geneformer utilizes a masked learning objective, where 15% of genes in each transcriptome are masked and the model predicts these genes based on the remaining context, allowing it to learn the relationships between genes in a self-supervised manner [4].

Geneformer's context awareness is a key feature, with embeddings reflecting gene characteristics specific to cellular contexts, mitigating the effect of technical artefacts and patient variability. This context-aware embedding allows for integrating smaller datasets of transcriptomic data, maintaining generalizability even after fine-tuning. Fine-tuning on limited data adapts the pre-trained knowledge for specific tasks, such as identifying genes significant to drug treatment. Compared to models trained from scratch, Geneformer has already shown superior performance in cell-type annotation and other complex classification tasks due to its extensive pre-training [4].

## 3 Methods

### 3.1 The Sciplex2 Dataset

The Sciplex2 dataset used for model training consists of human lung adenocarcinoma (a type of lung cancer) cells from the A549 cell line. The cells were treated with increasing doses of 3 drugs - BMS-345541, Nutlin-3a, and suberoylanilide hydroxamic acid (SAHA). The dataset also includes cells treated with dexamethasone, but since its effects are more complex, and would be better simulated with multi-gene perturbations, the data was excluded from the research. In total, the dataset contains 2608 cells treated with BMS-345541, 3878 cells treated with SAHA, 4354 cells treated with Nutlin-3a, and 3581 untreated cells [10].

### 3.2 SVM and Geneformer Setup

Before the SVM training process, the Sciplex2 dataset was prepared for the training. Since the drugs have different effects on the cell's genome depending on the dose, the models were trained separately for each drug and dose. After the data was normalised, a specific drug and dose were selected, after which the untreated cells were downsampled to keep an even count of treated and untreated cells. The resulting dataset, usually consisting of 1000-2000 cells, was then used to train the SVM, with 85% of cells used for training and 15% for testing. Due to the simplicity of the model and its ability to reach

high accuracy (>95%) without hyperparameter optimization, the validation set was omitted in favour of training and testing. The training was performed on a 256-dimensional embedding of the original data, which was created using PCA from the *Sklearn* library. This decision was made to enable a fair comparison with Geneformer, which is also fine-tuned on a 256-dimensional embedding of the data, and to make the perturbation experiment feasible by reducing the runtime ∼100-fold.

For the comparison of Geneformer and SVM to be insightful, the models are trained/fine-tuned on the same dataset. Similarly to SVM, the Geneformer model was fine-tuned separately for different drugs and doses. The preprocessing of the data is identical to the SVM, including drug and dose separation, normalization, and downsampling. Afterwards, the pre-trained Geneformer model was fine-tuned separately on each drug and dose, with a dataset split of 80% training, 10% validation (for hyperparameter optimisation), and 10% testing. The number of epochs was increased from 0.9 to 10 to produce higher model accuracy.

### 3.3 Single-Gene Perturbation Experiment Setup

To simulate the effects of drugs with both inhibition and activation properties, this paper focuses on two kinds of single-gene perturbations - deletion and overexpression. Deletion perturbations mimic the inhibitory effects of drugs, similar to gene knockdown or knockout experiments, by eliminating the expression of a specific gene. Conversely, overexpression perturbations emulate the activation effects of drugs by boosting the expression of a typically inactive gene to its maximum observed level.

The drug doses to perform single-gene perturbation experiments were chosen based on the accuracies of the models. For each drug, a dose with reasonable accuracy (>90%) for both SVM and Geneformer models was selected. Perturbations were performed on the Nutlin-3a model at dose 25.0 µM, the BMS-345541 model at dose 1.25 µM, and the SAHA model at dose 1.25 µM. Deletions were performed by taking untreated cells with a chosen gene having a higher-than-median expression value and changing it to 0.0. Similarly, overexpressions were performed by taking untreated cells with a chosen gene having a lower-than-median expression value and changing the expression value to the maximum expression value of the perturbed gene in the dataset. The median values were calculated based on cells which had a gene expression value higher than zero, and only these cells were used for perturbations to retain equal cell counts for deletion and overexpression. After the cells were perturbed in this way, the chosen models (based on drug and dose) were used to predict their treated/untreated label. For some genes, the amount of cells that could be perturbed was too low to be included in the results. Samples, where the number of perturbed cells was <100, were discarded to keep the results significant.

### 3.4 Libraries and Computation Methods

The SVM was implemented using the Python library *Sklearn*. The *Sklearn* library offers an API that simplifies the process of training and testing models. It provides extensive documentation and built-in tools for model evaluation and hyperparameter tuning. Additionally, *Sklearn* is highly optimized for performance, making it efficient for handling large datasets and complex computations.

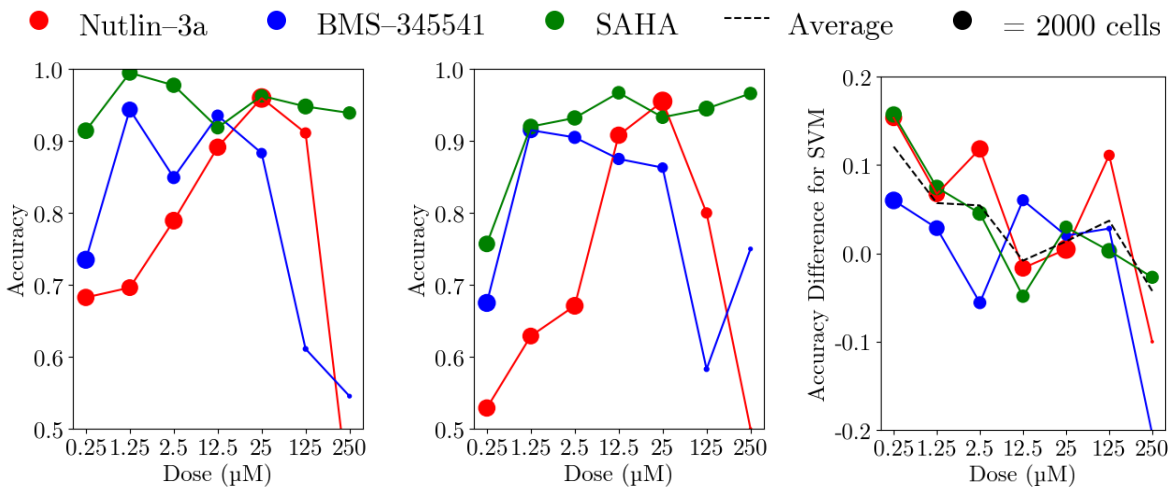
The *Scanpy* library was used to read the data from Sciplex2, which is stored in the .h5ad format, and normalize the data. The *NumPy* library was used for smaller mathematical tasks. The *Matplotlib* library was used to plot the figures.

Since the Geneformer model is fairly computationally intensive, the DAIC computation cluster provided by TU Delft was used for fine-tuning of the model. Before the fine-tuning phase, the Sciplex2 dataset needed to be tokenized using the TranscriptomeTokenizer from the Geneformer library.

# 4 Results

## 4.1 Cell Classification Results

The resulting accuracies of all Geneformer and SVM models for specific drugs and doses can be seen in Figure 1.




(a) SVM (b) Geneformer (c) Difference

Figure 1: Accuracy comparison of SVM and Geneformer models. The point size represents the cell count. **(a)** - Accuracy of the SVM Model for different doses of each drug. **(b)** - Accuracy of the Geneformer Model for different doses of each drug. **(c)** - Accuracy difference between the SVM and Geneformer models for different doses of each drug (in favour of SVM). The black line denotes the weighted average difference of all 3 drugs for each dose.

The accuracy progression based on the drug dose is similar for both methods. For all 3 drugs, the accuracy is relatively low for the smallest 0.25 µM dose. While the BMS-345541 and SAHA drugs already reach high accuracy around the 1.25 µM dose, Nutlin-3a starts producing higher accuracies only above the 12.5 µM level. On average, the SVM model outperforms Geneformer by 4.75%. Since both models become volatile for doses with low sample sizes, the accuracies were weighted by cell count to produce a weighted average.

## 4.2 Single-Gene Perturbation Experiment Results

### 4.2.1 SVM Re-Classification

After the single-gene perturbations were performed on the SVM model, the model was used to classify the perturbed untreated cells. The re-classification percentage refers to the percentage of cells, which were re-classified from untreated to treated after perturbing a certain gene. This value is used to identify genes, whose perturbation has the most similar effect to the drugs underpinning the model, according to the SVM model. The resulting re-classification percentages can be seen in Table 1. Out of the top 10 genes identified by the perturbation experiments based on re-classification percentages

(Figure 2), 6 were directly mentioned in cancer research using Nutlin-3a or SAHA (DHRS2 [11], AC138819.1 [12], CDKN1A [13], NEU1 [14], CYB5R1 [15], CTGF [16]).

| Drug @ Dose | Nutlin-3a @ 25.0 µM | | BMS-345541 @ 1.25 µM | | SAHA @ 1.25 µM | |
|---|---|---|---|---|---|---|
| Perturbation | Overexp. | Deletion | Overexp. | Deletion | Overexp. | Deletion |
| Mean Re-classification | 5.6% | 4.3% | 4.8% | 3.5% | 5.2% | 3.1% |
| Max Re-classification | 15.8% | 17.9% | 14.7% | 11.5% | 16.3% | 33.0% |

Table 1: Re-classification percentages of the SVM models based on different drugs and doses. Row 1 indicates the drug and dose used to train the model. Row 2 denotes the overexpression/deletion type of single-gene perturbation. Finally, Rows 3 and 4 contain the mean and maximum re-classification percentage for the indicated perturbation type, drug, and dose.

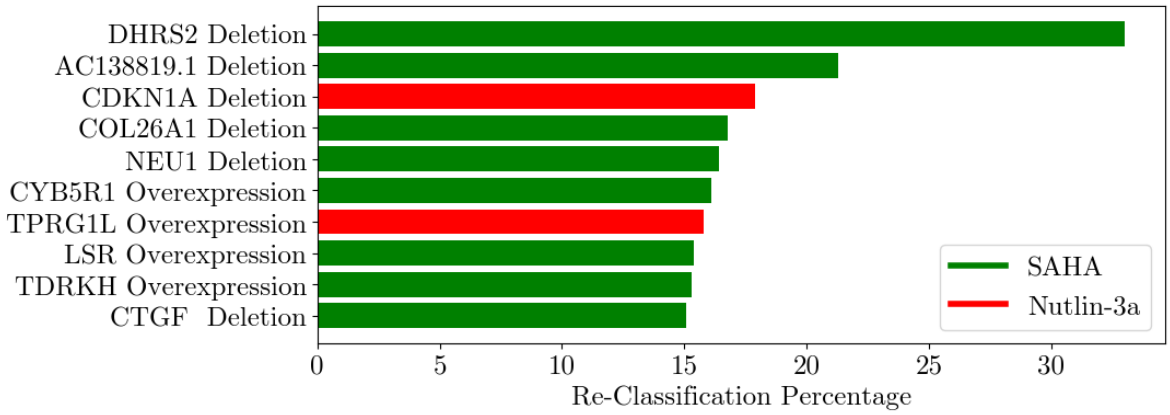
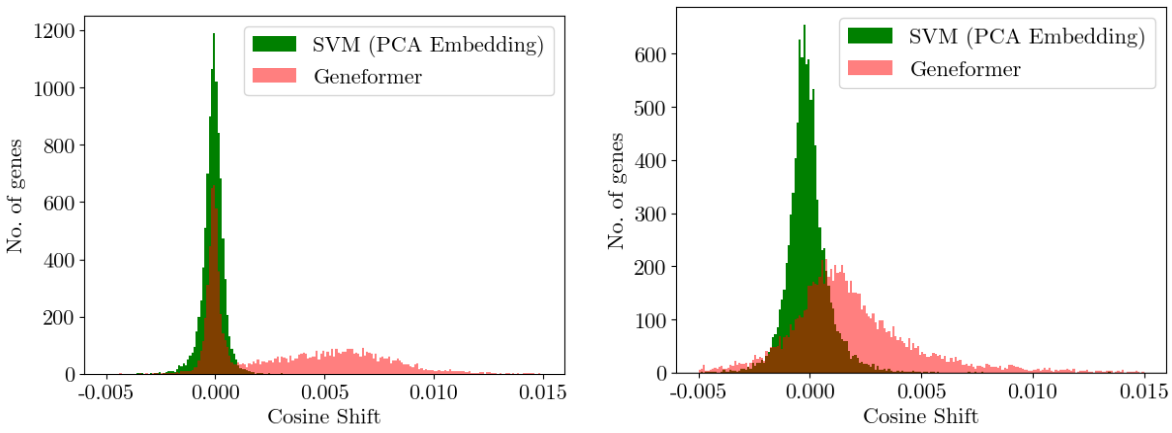



Figure 2: Top 10 perturbed genes based on re-classification percentage of the SVM model.

#### 4.2.2 PCA and Geneformer Cosine Shifts

The Geneformer and SVM models can be directly compared using cosine shifts in their embeddings. Cosine similarity is a metric which describes how close two embedding vectors are to each other. For each gene, we can calculate the mean cosine similarity with the treated cells before the perturbation, and after the perturbation. The difference between these two values results in the cosine shift, which describes how much the cells have moved towards the treated state. This method has its limitations since the mean embedding of the treated cells can be in an area which is not occupied by any treated cells (e.g. if the cluster of treated cells is moon-shaped). However, it will generally point towards genes which have a significant effect on the embedding.

For most drug and perturbation type combinations, the Geneformer model had significantly higher maximum cosine shifts. The cosine shifts of the PCA embeddings always follow a natural-like distribution around 0, while some of the Geneformer embedding cosine shifts are visibly shifted to the right. Figure 3 shows the cosine shift distributions of overexpression and deletion perturbations for the BMS-345541 and SAHA drugs respectively. The remaining distributions can be found in Appendix A.




(a) Deletion with BMS-345541

(b) Overexpression with SAHA

Figure 3: Post-perturbation cosine shift of cell embeddings towards a mean embedding of treated cells. The graph shows the distribution of genes with different cosine shifts for both the SVM and Geneformer models. **(a)** shows the cosine shift of gene deletion using BMS-345541-treated cells, while **(b)** shows the cosine shift of gene overexpression using SAHA-treated cells.

A sizeable difference between the Geneformer and SVM models can be found in their similarity to highly differentially expressed genes. Out of the top 50 genes based on cosine shifts in all perturbation types and drugs, SVM had a mean 26.5-gene overlap with the highest/lowest (depending on perturbation type) differentially expressed genes. In comparison, Geneformer had a mean 2-gene overlap in the top 50 genes. All DGE overlap values are included in Table 2.

| Drug @ Dose | Nutlin-3a @ 25.0 µM | | | | BMS-345541 @ 1.25 µM | | | | SAHA @ 1.25 µM | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Perturbation | Overexp. | | Delet. | | Overexp. | | Delet. | | Overexp. | | Delet. | |
| Model | S | G | S | G | S | G | S | G | S | G | S | G |
| DGE Overlap | 32 | 1 | 25 | 1 | 15 | 0 | 23 | 4 | 28 | 2 | 36 | 4 |

Table 2: Overlap of 50 genes with the highest cosine shift with 50 genes with the highest differential gene expression for overexpression, and lowest differential gene expression for deletion. Information varies based on SVM/Geneformer model and different drugs and doses. Row 1 indicates the drug and dose used to train the model. Row 2 denotes the overexpression/deletion type of single-gene perturbation. Row 3 differentiates between the SVM (S) model and the Geneformer (G) model. Finally, row 4 contains the overlap with DGE genes.

Overall, the Geneformer model exhibited larger maximum cosine shifts for all drugs and perturbation types. On average, the maximum cosine shift of the Geneformer model was 5.7 times higher than the maximum cosine shift of the SVM model for a particular drug and perturbation type. All maximum cosine shift values are in Table 3.

| Drug @ Dose | Nutlin-3a @ 25.0 µM | | | | BMS-345541 @ 1.25 µM | | | | SAHA @ 1.25 µM | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Perturbation | Overexp. | | Delet. | | Overexp. | | Delet. | | Overexp. | | Delet. | |
| Model | S | G | S | G | S | G | S | G | S | G | S | G |
| Max. Cos. Shift | 0.020 | 0.054 | 0.003 | 0.029 | 0.009 | 0.112 | 0.008 | 0.043 | 0.013 | 0.037 | 0.008 | 0.009 |

Table 3: Cosine shifts of SVM (using PCA embedding) and Geneformer models based on different drugs and doses. Row 1 indicates the drug and dose used to train the model. Row 2 denotes the overexpression/deletion type of single-gene perturbation. Row 3 differentiates between the SVM (S) model and the Geneformer (G) model. Finally, Rows 4 and 5 contain the mean and maximum cosine shift for the indicated model, perturbation type, drug, and dose.

# 5 Discussion

In SVM training and Geneformer fine-tuning, we could observe varying model accuracy based on drug and dose. While the BMS-345541 and SAHA drugs already reach their highest accuracy around the 1.25 µM dose, Nutlin-3a starts producing higher accuracies only above the 12.5 µM level. This observation is supported by a study of chromatin changes [17], which found that while SAHA and BMS-345541 showed changes at relatively low doses (around 1 µM), Nutlin-3a induced very few detectable changes at these levels. Based on this, we can conclude that while BMS-345541 and SAHA are already potent at a relatively low dose, the Nutlin-3a drug has to be applied in a higher dose to have a noticeable effect on the genome.

The SVM has generally exhibited higher accuracy, which can be attributed to Geneformer being a pre-trained model - having transcriptome data from 30 million cells will inevitably result in some irrelevant data being used by the model, leading to some inaccuracy. Nevertheless, the higher accuracy achieved by SVM does not necessarily mean a better understanding of the gene network by the model.

After performing single-gene perturbation experiments on both Geneformer and SVM models, and evaluating them using different metrics, we can observe differences in their ability to identify genes that significantly impact the treatment label of cells.

Based on the re-classification percentages, we can conclude that the SVM model is able to identify some genes with significant impact for some drugs. The gene with the highest re-classification value of 33% was DHRS2 for gene deletion on the SAHA model. Interestingly, according to [11], HDAC inhibitor drugs, including SAHA, can increase the DHRS2 transcriptional expression level. Furthermore, [11] also reports that decreased DHRS2 expression is associated with HDAC inhibitor resistance (including SAHA) and poor prognosis in ovarian cancer. To summarize, the SAHA drug should have a lower effect on cells with low expression of DHRS2. Therefore, the genome of a cell treated by SAHA, which had a low expression of DHRS2, will change less significantly and will be more similar to an untreated cell. The machine learning model will then be more likely to classify an otherwise-unchanged untreated cell as treated, if the expression of DHRS2 is low, which explains the misinterpretation of DHRS2's effects by the model. We can see a similar phenomenon with gene CDKN1A, which had the third highest re-classification value of 17.9% for deletion perturbation on the Nutlin-3a model. According to [13], the gene expression of CDKN1A significantly increased in cells treated with Nutlin-3a. On the other hand, AC138819.1, which has the second highest re-classification value of 21.3% for gene deletion on the SAHA model, might be connected to tolerance of SAHA in cancer cells. LncRNA genes, which AC138819.1 belongs to, were found to be upregulated in SAHA-tolerant nasopharyngeal cancer cells [12]. The re-classification of cells to SAHA-treated by deletion of this gene therefore meets the expectations.

When we compare significant genes identified by the SVM's re-classification percentage and cosine shifts, we can see that using cosine shifts as a metric leads to more accurate gene identification. For SAHA, the DHRS2 gene gets correctly identified by gene overexpression using cosine shifts, and the same applies to Nutlin-3a and the CDKN1A gene. These findings suggest that to identify genes with similar effects to drug treatment, it is more beneficial to look at the size of the shift of the cells towards the treated state, rather than the proportion of cells which changed state entirely.

With the cosine shift metric showing more merit in gene identification, we can use it to compare SVM and Geneformer. The superior ability of the Geneformer model to understand relationships between genes can be observed by analysing gene overlap with differential gene expressions. Identifying significant genes solely by their differential expression can lead to some results, but this approach completely neglects the relationships between genes, which affect their gene expressions. While not assigning large cosine shifts to highly differentially expressed genes, Geneformer can still identify the significant genes with high confidence. MDM2 is the main target gene of Nutlin-3a - the drug inhibits the interaction of MDM2 with the p53 protein, leading to increased p53 levels, which results in an indirect upregulation of MDM2 expression [18]. With overexpression perturbations on the Nutlin-3a model, Geneformer cosine shifts identified MDM2 as substantially more significant than any other gene. With Geneformer, MDM2 had a cosine shift value of 0.054, which is 4 times higher than the next gene, and 2.7 times higher than the cosine shift of MDM2 with the PCA embedding used by the SVM. In comparison, the cosine shift of MDM2 was only 1.3 times higher than the next gene using SVM, showing that the model assigned a much lower relative significance to the gene. Cosine shift values of the top 10 genes for each model, drug, and perturbation type, can be found in Appendix B.

# 6 Conclusions and Future Work

The comparative analysis of Geneformer and Support Vector Machine in predicting drug response through single-gene perturbations produced valuable insight into the advantages and disadvantages of both approaches. While the SVM model consistently showed higher accuracy in predicting the treatment label of cells, it was unable to match Geneformer in identifying gene perturbations that exhibit similar effects to the BMS-345541, Nutlin-3a, and SAHA drugs. The pre-training of the Geneformer model on large amounts of transcriptomic data seems to impair its classification ability, but results in a more complex embedding that better models the relationships between genes.

The Sciplex2 dataset turned out to be suitable for this comparison, providing a sufficient amount of samples, as well as drugs and doses that lead to models with reasonable efficiency (>90%).

In general, the SVM boasted higher model accuracies in predicting the treated/untreated label of cells than Geneformer. The accuracy of certain drug doses varied for different drugs, with Nutlin-3a only producing accurate models at relatively high doses.

Deletion and overexpression single-gene perturbations of untreated cells were performed, with three SVM models and three Geneformer drug-based models then used to classify the perturbed cells. The perturbed genes have shown a re-classification percentage of up to 33% using the SVM. Most of the genes with a high re-classification percentage could be linked to research regarding the SAHA and Nutlin-3a drugs, showing some level of understanding of their effects by the SVM model, but their effects were often misrepresented.

Cosine shifts were found to be a more suitable metric for significant gene identification, and they were used to directly compare Geneformer and SVM. Genes identified by Geneformer cosine shifts have shown a drastically lower overlap with highly differentially expressed genes, while still being able to identify significant genes correctly, which shows an understanding of gene relationships that goes beyond individual gene expression.

To gain a more complete understanding of Geneformer's and SVM's ability to identify significant genes, a more comprehensive analysis of the identified genes can be performed. Due to the time limitations of this research, it was not possible to exhaustively examine all genes affected by the studied drugs, and their reported significance by the models. The models' performance can also be evaluated on different datasets to validate the initial results from Sciplex2. To more closely simulate the effects of some drugs, deletion and overexpression perturbations can be extended with inhibition and activation. The embeddings of the models could be further analysed, as their shape influences the location of a mean treated embedding, which significantly affects the cosine shifts.

# 7 Responsible Research

Ensuring ethical integrity and reproducibility is especially important in research connected to medicine, particularly when predicting the drug response of cancer cells.

Some of the main ethical considerations include data privacy and result validation. The transcriptomic data used in this research was extracted from a publicly available Sciplex2 dataset. Regarding result validation, this paper serves solely as an empirical comparison tool of two machine learning approaches. Before implementing any results into the treatment of patients, the results should be more rigorously tested and confirmed by multiple studies.

Reproducibility is critical for the extension and confirmation of the initial results by other researchers. The methods of SVM implementation and single-gene perturbations were designed using publicly available Python libraries and carefully described for reproduction. The implementation of Geneformer can be found in the original paper [4].

# References


[1] World Health Organization. Cancer. `https://www.who.int/news-room/fact-sheets/detail/cancer`, February 2022. Accessed: 2024-06-23.

[2] Hafsa Habehh and Suril Gohel. Machine learning in healthcare. *Current Genomics*, 22(4):291–300, December 2021.

[3] OpenAI (2023). GPT-4 Technical Report. *arXiv (Cornell University)*, March 2023.

[4] Christina V. Theodoris, Ling Xiao, Anant Chopra, Mark D. Chaffin, Zeina R. Al Sayed, Matthew C. Hill, Helene Mantineo, Elizabeth M. Brydon, Zexian Zeng, X. Shirley Liu, and Patrick T. Ellinor. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, June 2023.

[5] Christo El Morr, Manar Jammal, Hossam Ali-Hassan, and Walid El-Hallak. *Support Vector Machine*, pages 385–411. Springer International Publishing, Cham, 2022.

[6] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1/3):389–422, January 2002.

[7] Transcriptome. `https://en.wikipedia.org/wiki/Transcriptome`, December 2023. Accessed: 2024-06-23.

[8] Maria Jackson, Leah Marks, Gerhard H. W. May, and Joanna B. Wilson. The genetic basis of disease. *Essays in Biochemistry*, 62(5):643–723, 12 2018.

[9] Gene therapy. https://en.wikipedia.org/wiki/Gene_therapy, June 2024. Accessed: 2024-06-23.

[10] José L. McFaline-Figueroa. Sample gsm4150377: Sciplex2 - a549 transcription modulators. National Center for Biotechnology Information, 2020. Accessed: 13 June 2024.

[11] Yingyan Han, Zhi Wang, Shujuan Sun, Zeyu Zhang, Jia Liu, Xin Jin, Peng Wu, Teng Ji, Wencheng Ding, Beibei Wang, and Qinglei Gao. Decreased dhrs2 expression is associated with hdaci resistance and poor prognosis in ovarian cancer. *Epigenetics*, 15(1-2):122–133, 2019.

[12] Fei Xue, You Cheng, Li Xu, Chuan Tian, Hongye Jiao, Rui Wang, and Xia Gao. Lncrna neat1/mir-129/bcl-2 signaling axis contributes to hdac inhibitor tolerance in nasopharyngeal cancer. *Aging*, 12(14):14174–14188, July 2020.

[13] Ada Lerma Clavero, Paula Lafqvist Boqvist, Katrine Ingelshed, Cecilia Bosdotter, Saikiran Sedimbi, Long Jiang, Fredrik Wermeling, Borivoj Vojtesek, David P. Lane, and Pavitra Kannan. Mdm2 inhibitors, nutlin-3a and navtemadelin, retain efficacy in human and mouse cancer cells cultured in hypoxia. *Scientific Reports*, 13, March 2023.

[14] Rosario Mosca, Diantha van de Vlekkert, Yvan Campos, Leigh E. Fremuth, Jaclyn Cadaoas, Vish Koppaka, Emil Kakkis, Cynthia Tifft, Camilo Toro, Simona Allievi, Cinzia Gellera, Laura Canafoglia, Gepke Visser, Ida Annunziata, and Alessandra d'Azzo. Conventional and unconventional therapeutic strategies for sialidosis type i. *Journal of Clinical Medicine*, 9(3):695, March 2020.

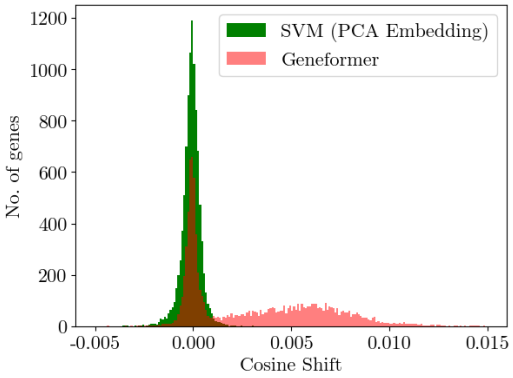
[15] Robert Jenke, Denys Oliinyk, Tamara Zenz, Justus Körfer, Linda Schäker-Hübner, Finn K. Hansen, Florian Lordick, Florian Meier-Rosar, Achim Aigner, and Thomas Büch. Hdac inhibitors activate lipid peroxidation and ferroptosis in gastric cancer. *Biochemical Pharmacology*, 225(Series):116257, July 2024.

[16] Claudiu Komorowsky, Matthias Ocker, and Margarete Goppelt-Struebe. Differential regulation of connective tissue growth factor in renal cells by histone deacetylase inhibitors. *Journal of Cellular and Molecular Medicine*, 13(8b):2353–2364, 2009.

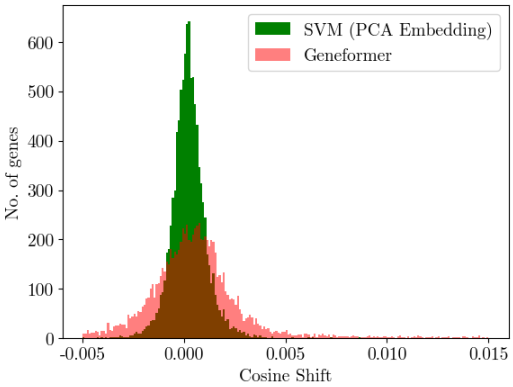
[17] Gregory T. Booth, Riza M. Daza, Sanjay R. Srivatsan, JosÃ© L. McFaline-Figueroa, Rula Green Gladden, Andrew C. Mullen, Scott N. Furlan, Jay Shendure, and Cole Trapnell. High-capacity sample multiplexing for single cell chromatin accessibility profiling. *BMC Genomics*, 24(1), December 2023.

[18] Artur Zajkowicz, Małgorzata Krześniak, Iwona Matuszczyk, Magdalena Głowala-Kosińska, Dorota Butkiewicz, and Marek Rusin. Nutlin-3a, an mdm2 antagonist and p53 activator, helps to preserve the replicative potential of cancer cells treated with a genotoxic dose of resveratrol. *Molecular Biology Reports*, 40(8):5013–5026, May 2013.

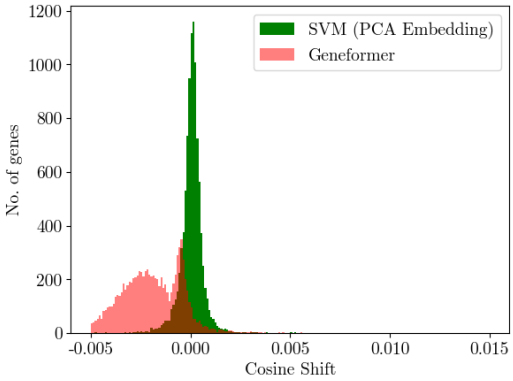
# Appendix A: Cosine Shift Distributions



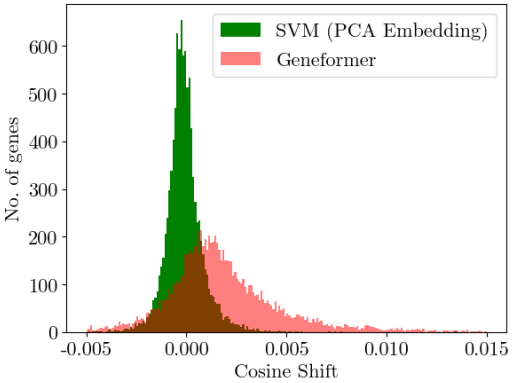

(a) Deletion with BMS-345541



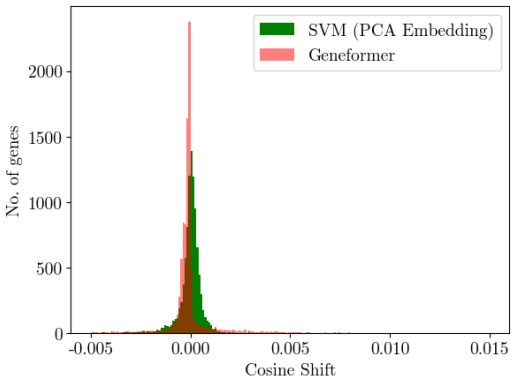

(b) Overexpression with BMS-345541



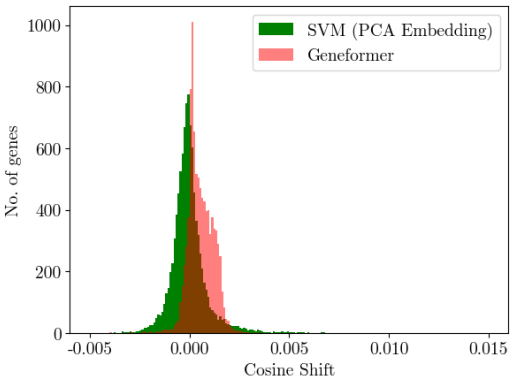

(c) Deletion with SAHA




(d) Overexpression with SAHA



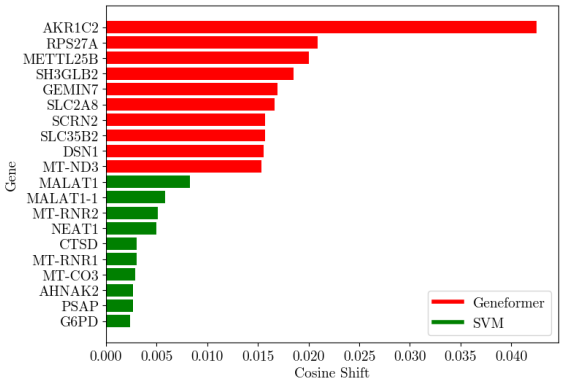

(e) Deletion with Nutlin-3a



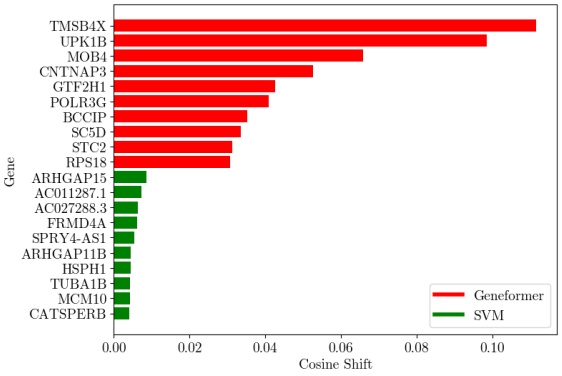

(f) Overexpression with Nutlin-3a

Figure 4: Post-perturbation cosine shift of cell embeddings towards a mean embedding of treated cells. The graphs show the distribution of genes with different cosine shifts for both the SVM and Geneformer models, based on varying drugs and perturbation types.

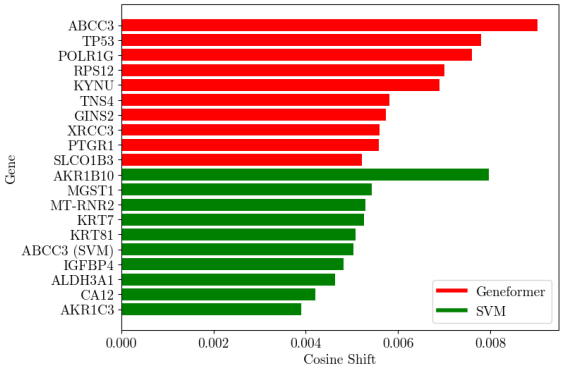
# Appendix B: Most Significant Genes Based on Cosine Shifts



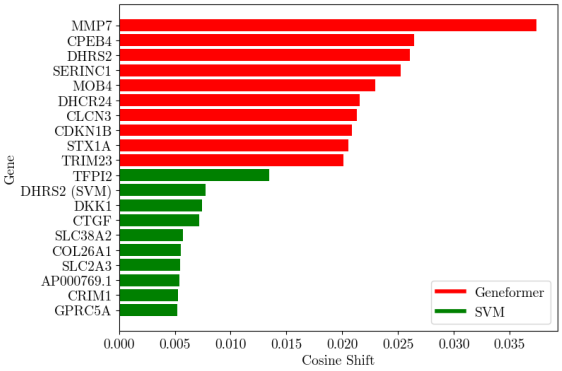

(a) Deletion with BMS-345541



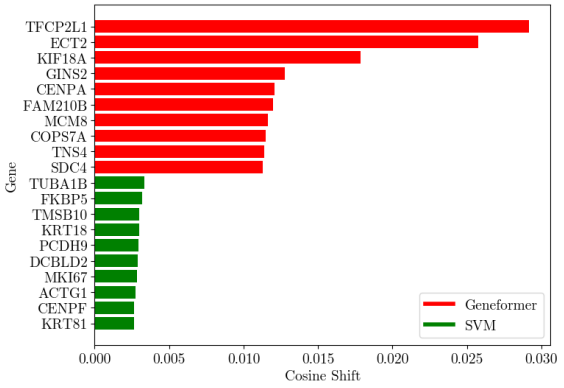

(b) Overexpression with BMS-345541



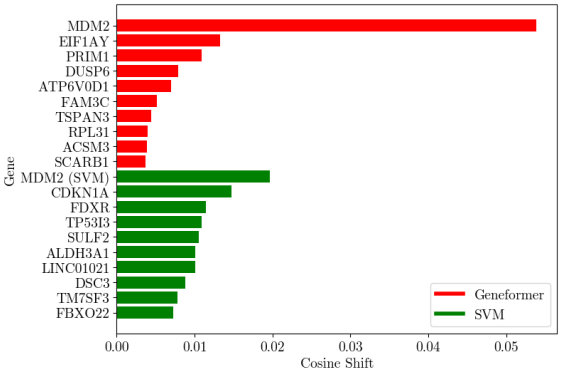

(c) Deletion with SAHA




(d) Overexpression with SAHA




(e) Deletion with Nutlin-3a




(f) Overexpression with Nutlin-3a

Figure 5: Post-perturbation cosine shift of cell embeddings towards a mean embedding of treated cells. The graphs show 10 genes with the highest cosine shifts for both the SVM and Geneformer models, based on varying drugs and perturbation types.