Evaluating Machine Learning Approaches to Drug Response Prediction in Cancer Cells

Author Samuel Banas (s.banas@student.tudelft.nl), **Supervisors** Marcel Reinders, Niek Brouwer

TU Delft

1 Scientific Background

Accurately predicting cellular responses to drug treatments is vital for personalized medicine and drug discovery. **Gene expression** in cells is closely linked to diseases, and drugs that target proteins encoded by genes.

Transcriptomic data, which holds information about gene expression in cells, can be used by machine learning models to understand and model the effects of drug treatment.

Geneformer is a transformer-based deep learning model, pre-trained on a dataset of 30M cells. It can be fine-tuned on a smaller dataset to understand and predict the relationships between genes. Its self-attention mechanism has the potential to enhance performance in predicting the treatment label of cells.

This paper evaluates **Geneformer**'s performance against **SVM** by performing **single-gene perturbation experiments**.

2 Research Questions

1. What are the **features** and **characteristics** of the Sciplex2 **dataset**?

2. What are the **accuracies** of Geneformer and SVM in predicting the **treatment label of cells**?

3. How can we **simulate single-gene perturbations** to replicate the effect of drugs?

4. What are the **differences** in the ability of **Geneformer** and **SVM** to **identify significant genes** based on perturbations?

3 Methods

Sciplex2 Dataset: Human lung cancer cells treated with BMS-345541, Nutlin-3a, and SAHA drugs

Geneformer: Fine-tuned on the data to predict treatment label

SVM: Trained on the data to predict treatment label

Both models: Perform overexpression and deletion perturbations and identify significant genes

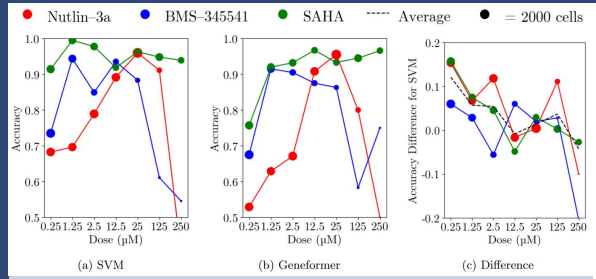
5 Conclusions

While the **SVM** model showed **higher accuracy** in predicting the treatment label of cells, it was unable to match **Geneformer** in **identifying significant genes** that exhibit similar effects to the BMS-345541, Nutlin-3a, and SAHA drugs.

After **deletion** and **overexpression** perturbations of untreated cells, the 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, 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.

4 Results

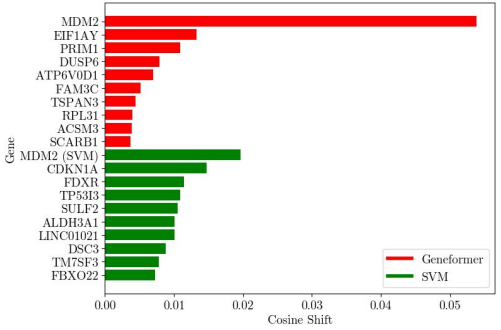


(a) SVM (b) Geneformer (c) Difference

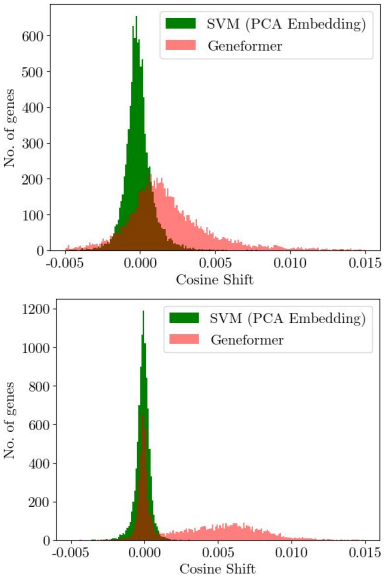
Top: SVM came out on top with weighted accuracy higher by 4.75%.

Bottom: Significant genes identified by the models differ in their similarity to highly differentially expressed genes. Out of the top 50 genes based on cosine shifts, SVM had a mean 26.5-gene overlap with the highest/lowest DGE genes, while Geneformer had a mean 2-gene overlap in the top 50 genes.

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



Top: MDM2 is the main target gene of Nutlin-3a, causing an indirect upregulation of MDM2 expression. With overexpression perturbations on the Nutlin-3a model, Geneformer cosine shifts identified MDM2 as substantially more significant than any other gene.



Left: For most drugs and perturbation types, Geneformer outperforms the SVM in cosine shift. Here, you can see a visible number of genes having a substantially positive cosine shift. Showing overexpression for SAHA (top) and deletion for BMS-345541 (bottom).