Research Project 3000 G58

1 Background

Cancer poses the **highest clinical, social, and economic burden** among all human diseases in terms of cause-specific Disability-Adjusted Life Years (**DALYs**) [1].

Developing effective treatments is **crucial** to lower this burden. Understanding how **drugs interact with cancer cells** and their downstream effects is vital for **creating new treatments** and **overcoming resistance** to existing therapies.

Combining gene perturbations in cell transcriptomes is crucial for drug discovery. Unlike single-gene perturbations, combination analyses reveal **synergistic effects** and **resistance mechanisms**, aiding in the identification of **effective drug combinations**.

Geneformer, a model leveraging a large corpus of single-cell transcriptomes, excels in **context-specific predictions**, particularly in **data-limited scenarios** [2]. This study compares **Geneformer's predictive performance** against traditional **machine learning models** in predicting cancer cell responses to perturbation combinations.

This comparison aims to **enhance** the **overall understanding** of drugs and their effects on **cancer cells using the sciplex2 dataset**.

2 Research Question

How does the **predictive performance** of the **Geneformer model compare** to **traditional machine learning methods** in **predicting** the **response of cancer cells** to **perturbation combinations** using the **sciplex2** dataset?

Models

Random Forest
Support Vector Machine
Gradient Boosting Classifier
vs
Geneformer

Perturbation combinations

altering the expression levels of **2+ specific genes** within the transcriptome of a single cell

sciplex2 dataset [3]

single-cell transcriptomic data profiling the **response** of A549 human **lung cancer** cells to varied **drug perturbations** and dosages

3 Methodology

Dataset Analysis

- **Selected Drug**: "Dex" (*Dexamethasone*), dose of *125μM*

Dataset Preparation and Preprocessing

1. **Normalization**: Gene expression values normalized & log-transformed
2. **Labeling**: Binary labels for treated (1) and untreated (0) cells
3. **Class Balancing**: Downsampled untreated cells to match treated cells
4. **Dimensionality Reduction**: PCA reduced data to 256 components
5. **Data Splitting**: Training (70%), validation (15%), and test (15%) sets

Model Classification Performance Evaluation

1. **Metrics**: Accuracy, Precision, Recall, F1 Score, AUC-ROC
2. **Comparison**: Models classified *untreated* vs. *treated* cells

Differential Gene Expression (DGE)

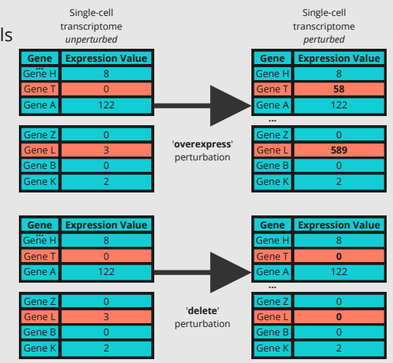
- Difference in the mean expression between treated and untreated cells

HDE Gene Pairs - top 500 gene **pairs** with the **highest** absolute **DGE**
HDE Single Genes - top 500 **single** genes with the **highest** absolute **DGE**

Model Comparison for Gene Perturbation Combinations

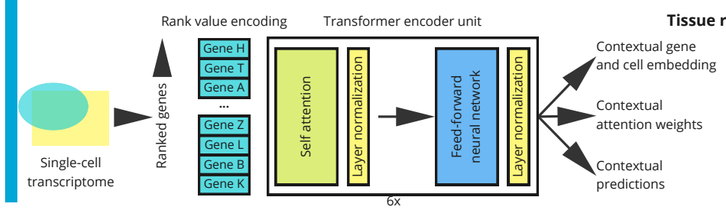
1 **Cosine Shift**: Identify changes in the embedding
2 **Shift Percentage**: Percentage of untreated cells reclassified to treated
3 **Importance of HDE genes**: Primitive models expected to put high importance on HDE genes

Perturbation Algorithm

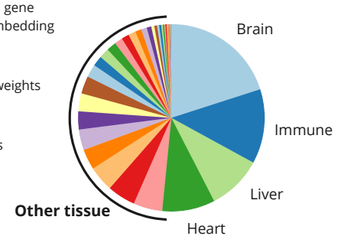


Geneformer [2]

context-aware, **attention-based** deep learning **model** pretrained on a large-scale corpus of approximately **30 million single-cell transcriptomes**



Tissue representation of Genecorpus-30M



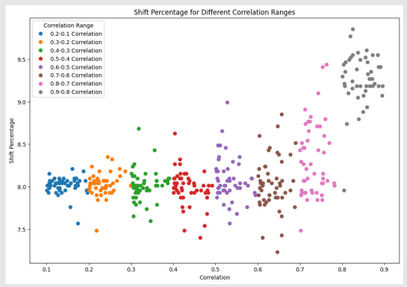
4 Results

1 **Classification Task**

1 SVM *Accuracy:* 0.9126 | *F1 Score:* 0.9143
2 GBM *Accuracy:* 0.9094 | *F1 Score:* 0.9091
3 RF *Accuracy:* 0.8706 | *F1 Score:* 0.8726
4 Geneformer *Accuracy* | 0.8544, *F1 Score:* 0.8530

2 **Perturbation Combinations**

Shift Percentage per Correlation Range



Experiment with **6000 perturbations**

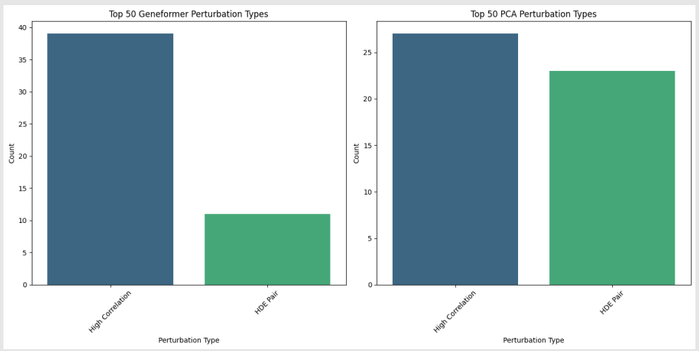
- **2500** most **expression-correlated** gene pairs
- **250 HDE Single** Genes
- **250 HDE** Gene **Pairs**

- 3000 gene/gene pairs

Top 100 most important perturbations per model

	Geneformer	RF	GBM	SVM
HDE Pairs	0.15	0.00	0.01	0.01
HDE Single	0.01	0.12	0.15	0.13
High Correlation	0.84	0.88	0.84	0.86

Cosine Shifts in PCA and Geneformer Embeddings



5 Conclusions

1 **Traditional ML surpassed Geneformer in Classification**
- Traditional ML is optimized for binary classification tasks
- Geneformer's pre-training data may introduce noise

2 **Highly correlated genes were the most impactful**
- Expression-correlated genes caused top perturbations
- Valuable for in-silico perturbation analysis of large datasets

3 **Geneformes demonstrated higher gene network understanding**
- Geneformer placed less emphasis on HDE genes compared to PCA
- Geneformer preferred HDE Gene Pairs over HDE Single Genes

6 Future work & Limitation

1 **More exhaustive search of the solution space**
- research covered less than 0.001% of 1.3B possible perturbations in **sciplex2**

2 **Evaluation of different perturbation combinations**
- research was limited to gene pair perturbations which are the least complex

3 **Utilizing larger and more diverse datasets**
- to validate findings and improve generalizability

References

[1] C. Mattiuzzi and G. Lippi, "Current Cancer Epidemiology," Journal of Epidemiology and Global Health, vol. 9, no. 4, pp. 217–222, Dec. 2019, doi: https://doi.org/10.2991/jegh.k.191008.001.
[2] C. V. Theodoris et al., "Transfer learning enables predictions in network biology," Nature, vol. 618, no. 7965, pp. 616–624, Jun. 2023, doi: https://doi.org/10.1038/s41586-023-06139-9.
[3] Jos'e L. McFaline-Figueroa. Sample gsm4150377: Sciplex2 - a549 transcription modulators. National Center for Biotechnology Information, 2020. Accessed: 13 June 2024.

TU Delft
Author: Michal Krkoska
m.krkoska@student.tudelft.nl