ThinkBio.Ai

# Geneformer: Advancing Predictions for Rare Diseases and Inaccessible Tissues

**Published:** dd – mm - yyyy

**Contributors:** Vasudev R, Dr Amitabha Chaudhuri, John Rekesh, Vidhi Malik

Geneformer is a transformer-based biomedical foundational model pretrained on large scale single cell transcriptomic data to understand gene-gene interactions in different healthy cells and use this foundational knowledge as the part transfer learning to study these interactions in unhealthy scenarios where data is limited.

## Overview

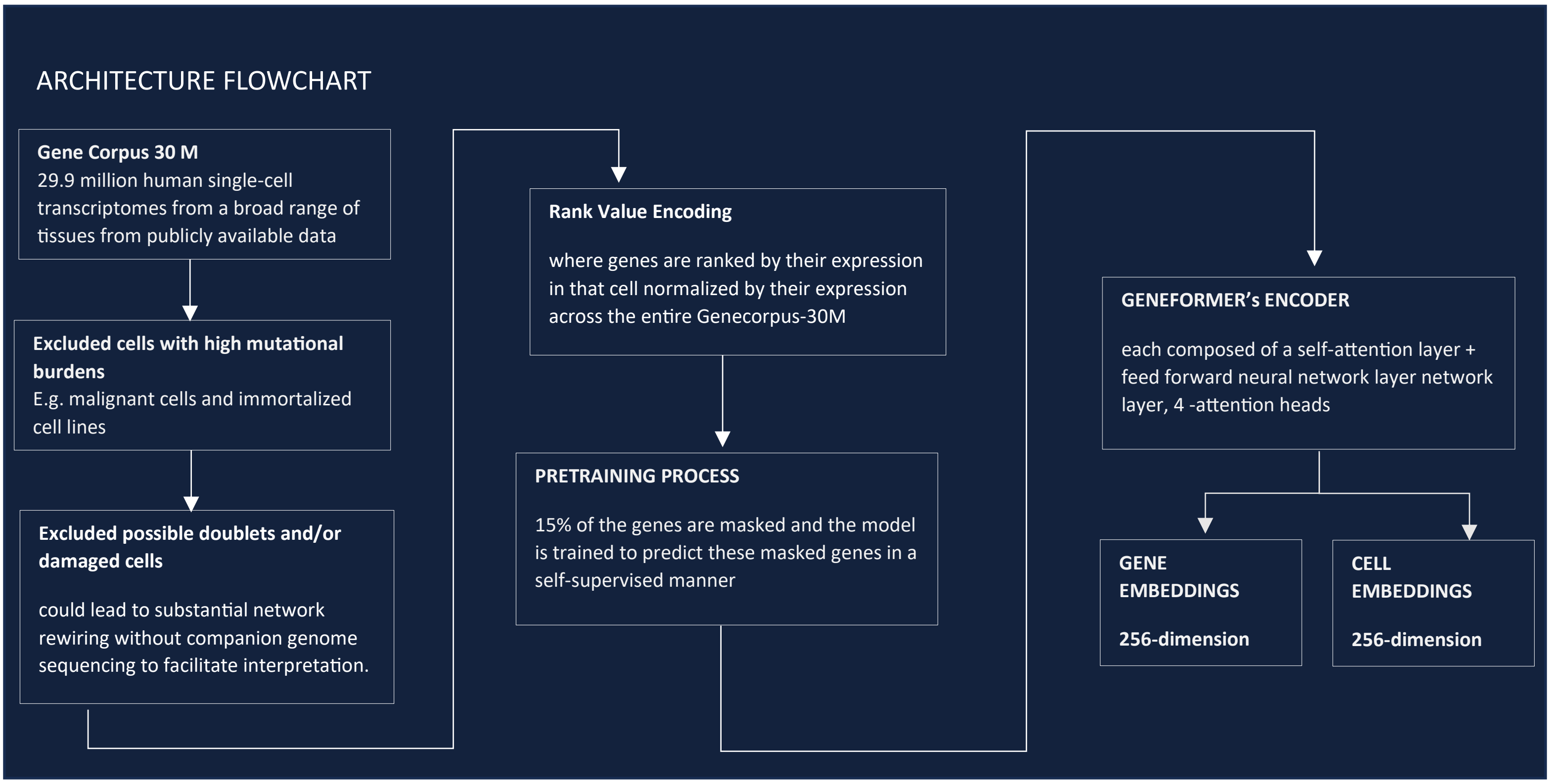
Pinpointing the core regulatory elements by understanding hierarchy of disease specific- GRNs (Gene Regulatory Networks) can help researchers to uncover the underlying mechanism that drive disease progression and develop targeted therapies aimed at modulating gene expression at these points.

Creating such networks will require large amount of single cell transcriptomic data to learn connections between the genes. But in the case of rare diseases and clinically inaccessible tissues such data is very limited, thus creating GRN in such cases is a challenging problem. This challenge is addressed by the transfer learning capability of Geneformer.

Geneformer is a Foundation model, pretrained on large scale single transcriptomic data in healthy tissues (as recent advancements in sequencing technologies provided rapid expansion in the availability of such transcriptomic data from wide range of human tissues). These learnings of geneformer can be foundation upon which various limited data disease cases can be studied at gene-to-gene interaction level.

This article provides a high-level architectural overview of Geneformer and explores its transfer learning capabilities in various downstream tasks within network biology.

# 1. Architecture



ARCHITECTURE FLOWCHART

**Gene Corpus 30 M**
29.9 million human single-cell transcriptomes from a broad range of tissues from publicly available data

**Excluded cells with high mutational burdens**
E.g. malignant cells and immortalized cell lines

**Excluded possible doublets and/or damaged cells**

could lead to substantial network rewiring without companion genome sequencing to facilitate interpretation.

**Rank Value Encoding**

where genes are ranked by their expression in that cell normalized by their expression across the entire Genecorpus-30M

**PRETRAINING PROCESS**

15% of the genes are masked and the model is trained to predict these masked genes in a self-supervised manner

**GENEFORMER's ENCODER**

each composed of a self-attention layer + feed forward neural network layer network layer, 4 -attention heads

**GENE EMBEDDINGS**

**256-dimension**

**CELL EMBEDDINGS**

**256-dimension**

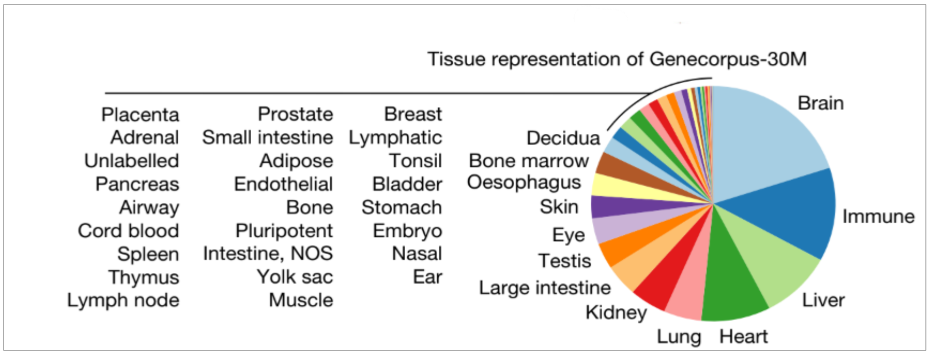


**Fig 1:** Various tissues considered for Genecorpus -30 M

### 1.1 Rank Value Encodings

How the rank is calculated

1. The non-zero median value of expression of each detected gene across all cells is taken
2. Aggregated the transcript count distribution for each gene in a memory-efficient manner
3. Normalizing the gene transcript counts in each cell by the total transcript count of that cell
4. Then normalized the genes in each single-cell transcriptome by the non-zero median value of expression of that gene across Genecorpus-30M.
5. ordered the genes by the rank of their normalized expression in that specific cell.

Notable properties of Rank Value Encoding

- Non Parametric representation for each single cell
- High Rank - Genes with high cell state distinguishing power but are lowly expressed
- Low Rank - Genes with low cell state distinguishing power but are highly expressed

### 1.2. The Pretraining Process

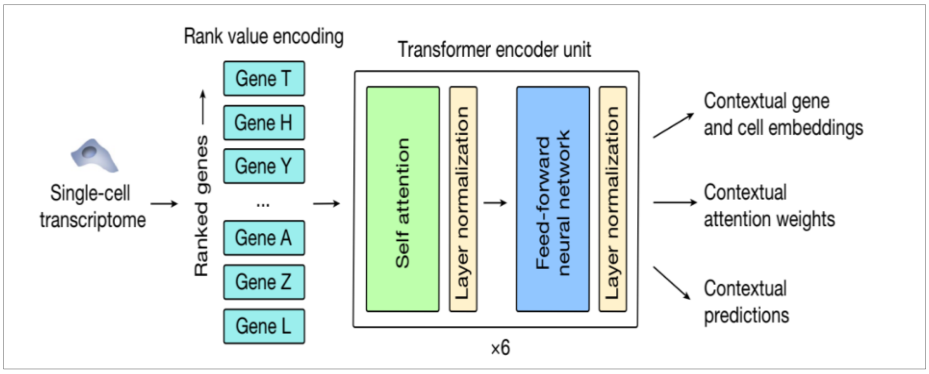


**Fig 2:** Rank value encodings presented to the GeneFormer

The model uses a self-supervised approach called masked learning objective where 15% of the genes in each single cell transcriptome is masked and the model is trained to predict which gene was in that masked position.

By using this approach the model learns

- **Gene-to-Gene interactions** : learning the underlying connections and regulatory network between the genes.
- **Contextual Relationships** : learns how genes collectively influence cellular state by using the surrounding gene expressions to infer masked genes.

These are generalized learnings which may not be available in limited data cases. During transfer learning these learnings make sure that model is not learning anything that it hasn't seen before (by acting as a foundation), thus it reduces the chances of overfitting.

---

# 2. In Silico Perturbation

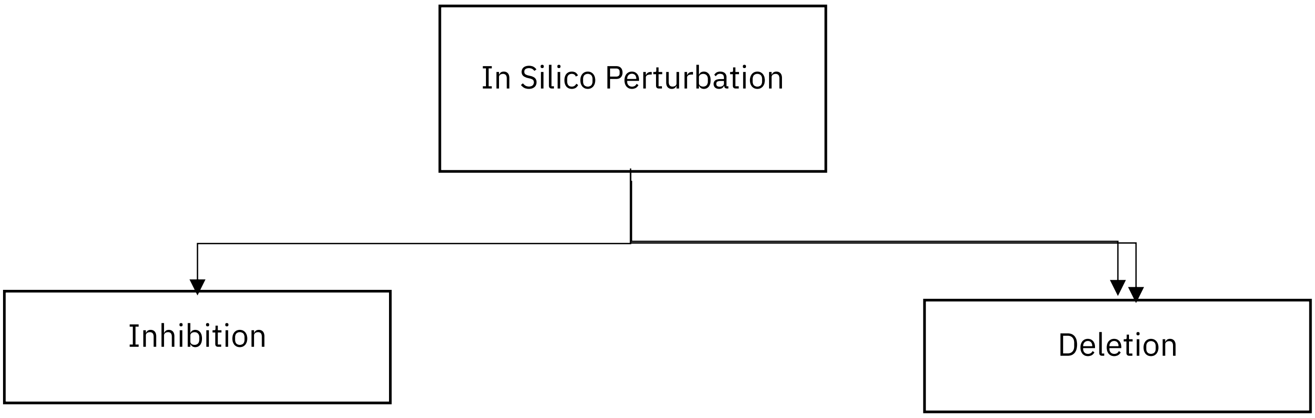
In silico perturbation is an additional step incorporated into certain downstream tasks to delete or inhibit specific genes. This is achieved by modifying the generated rank value encodings of genes for each cell under investigation. Perturbation effects are studied at both cell embeddings and gene embeddings to model how the perturbation influences the cell state and the regulation of downstream genes within the gene network.

### In Silico perturbation strategies

In silico perturbation strategies can be categorized based on how they manipulate gene rank value encodings to model different biological scenarios.

- **In silico Deletion:** Simulated by removing one or more genes from the rank value encodings.

- **In silico Inhibition or activation:** Simulated by moving one or more genes to the top of their rank value encodings.



### Quantifying the Perturbation Effects

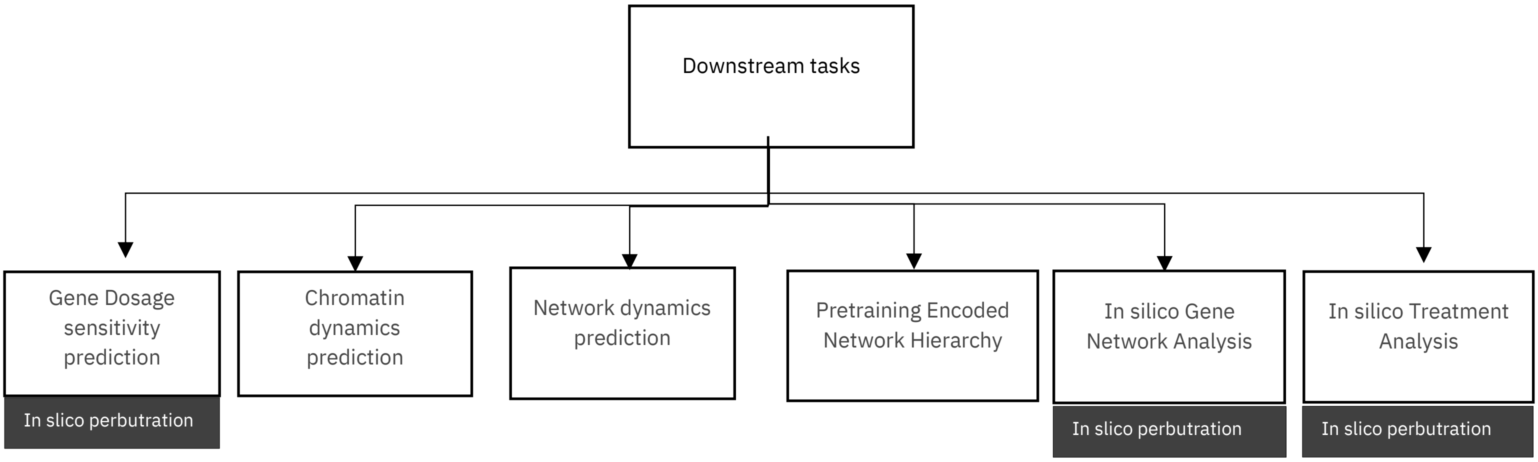
Effects of Perturbation can be quantified by calculating the cosine similarity between the original and perturbed values in both cell embeddings and gene embeddings.

- **Cell embeddings:** For predicting the deleterious effect of gene in the cell contexts.
- **Gene Embeddings:** For predicting which genes are most sensitive to in silico deletion of a given gene.

---

# 3. Geneformer Downstream tasks

Geneformer is a pre-trained Foundation model designed for fine-tuning across a broad range of applications, facilitating the discovery of key network regulators and candidate therapeutic targets. Below, we explain some downstream tasks related to chromatin and network dynamics, highlighting its adaptability with task-specific data.


Downstream tasks
Gene Dosage sensitivity prediction
In slico perbutration
Chromatin dynamics prediction
Network dynamics prediction
Pretraining Encoded Network Hierarchy
In silico Gene Network Analysis
In slico perbutration
In silico Treatment Analysis
In slico perbutration

## 3.1 Gene Dosage Sensitivity Prediction

This section covers the key biological concepts required, the fine-tuning process, as well as the observations and applications of the fine-tuned model.

### 3.1.1. Key Biological Concepts

- **Genome**: It includes the complete set of genetic material (DNA) in an organism, containing all the instructions needed for growth, development, and functioning. It includes all the genes and non-coding sequences.
- **CNVs (Copy Number Variants)** are genetic variations where parts of the genome are duplicated or deleted, affecting the number of gene copies, potentially influencing health and disease.
- **Genetic diagnosis:** process involves testing for testing for mutations, changes or variations in genes that may cause diseases or affect health.
- **Dosage sensitive genes:** are the genes where any changes in their copies can significantly affect the normal biological functions and in **dosage insensitive genes** can tolerate these changes without causing major issue.

### 3.1.2. Fine tuning Process

**Dataset:** used from studies [references 1-3]
**No of samples:** 10,000 single cell transcriptomes
**Classification labels:** Dosage Sensitive vs Dosage insensitive genes
Interpreting CNVs is a challenging problem in Genetic diagnosis, as CNVs are important for classifying whether a gene is dosage sensitive or dosage insensitive. Geneformer was fine-tuned to classify dosage-insensitive vs. dosage-sensitive genes in a manner that captured Dosage sensitive vs insensitive factors:

- Vary across different cell states,
- Provide insights into which specific tissues would be impacted by changes in gene dosage.

### 3.1.3. Results and Observations

ROC_ AUC is the measure to evaluate performance of classification algorithms. More the area under the curve better the classification algorithm is. Below Fig shows the Geneformer boosted the ability to predict dosage sensitive compared to alternative methods with AUC 0.91. Another important observation was this predictive accuracy can be improved by pre-training the Geneformer with larger and more diverse datasets.

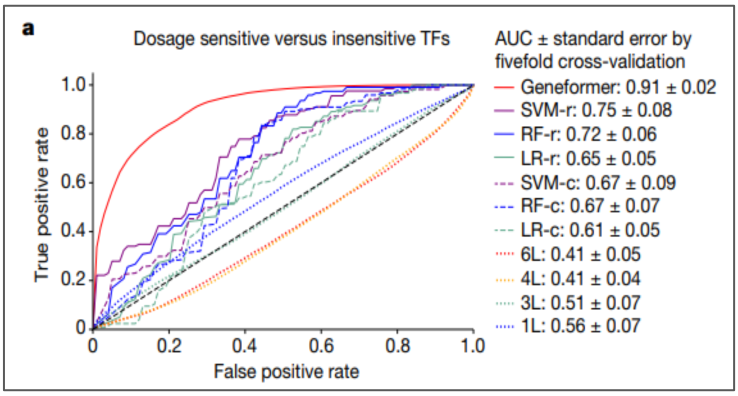


**Fig 3:** Geneformer's classification performance using AUC-ROC is compared with other algorithms**.**

### 3.1.4. Fine-tuned model applications

Fine tuned Geneformer can be used for two applications : ( I ) Neurodevelopmental disease and ( II ) In Silico deletion study in Cardiomyocytes.

#### ( I ) Neurodevelopmental disease

Gene dosage sensitivity was evaluated by deleting specific genes, with these deletions being primarily linked to neurodevelopmental disorders with high or moderate confidence. Collins Et al study of CNVs from 753,994 indivduals were used as reference. [ reference 4]

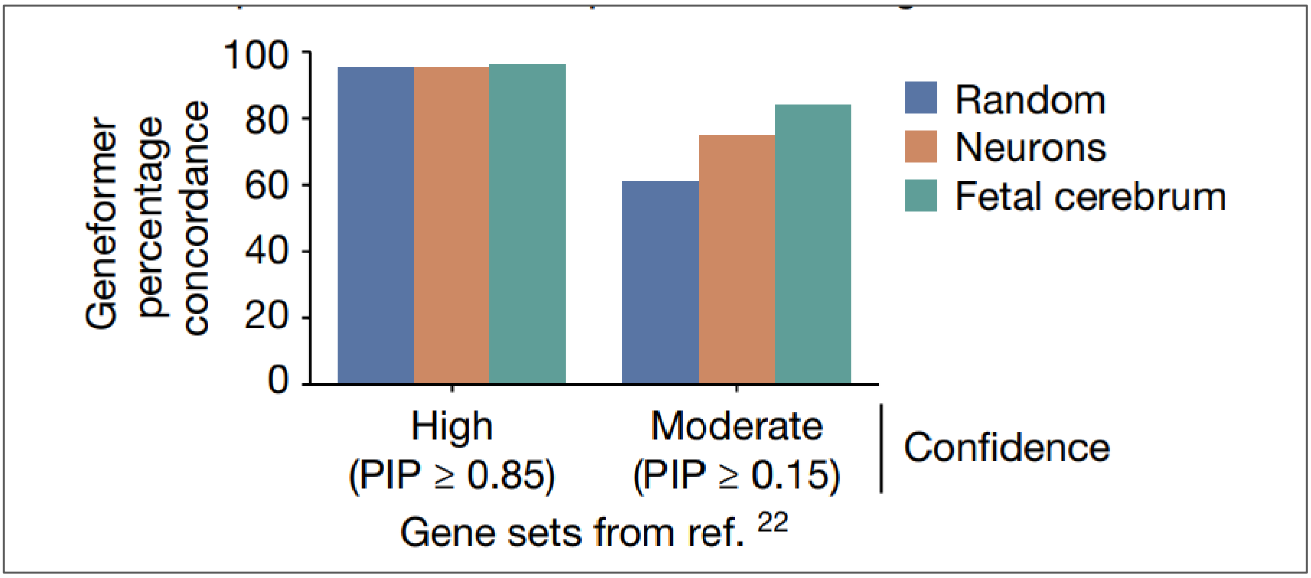


**Fig 4:** Geneformer was fine-tuned for dosage sensitivity predictions in (1) random cell types, (2) neurons (including adult), and (3) fetal cerebrum for neurodevelopmental genes from ref. 22, with PIP scores.

#### Observations

#### Concordance

- Fetal cerebral cells (High confidence) - 96% concordance with the reference study
- Other cells (High Confidence) – 95% confidence with reference study
- Fetal cerebral cells (Mod confidence) - 84% concordance with reference study

#### Context specificity highlighted

Although high confidence genes may have more stronger dosage sensitive effect, moderate confidence genes showed more context specificity as dosage sensitivity of Fetal cerebral cells greater than Neurons. This is consistent with the association of these genes with primarily neurodevelopmental phenotypes, where adult neurons may be less relevant.

#### ( II ) In silico deletion in Cardiomyocytes

Cardiomyocytes are known as cardiac muscle cells, these are cells that make up the heart muscle cells, they are essential for heart’s ability to pump the blood effectively. In silico deletion approach was done to cardiomyocytes to identify if the gene has deleterious effect on the context of cell. Genes known to cardiomyopathy (disease of heart muscle that affects its ability to pump blood efficiently) and structural heart diseases had more deleterious effect on the normal functioning of cardiomyocytes cells than control set of know hyperlipidaemia genes(a medical condition characterized by abnormal levels of lipids in the blood), This was measured using the cosine similarity of deleted genes w.r.t to original values.

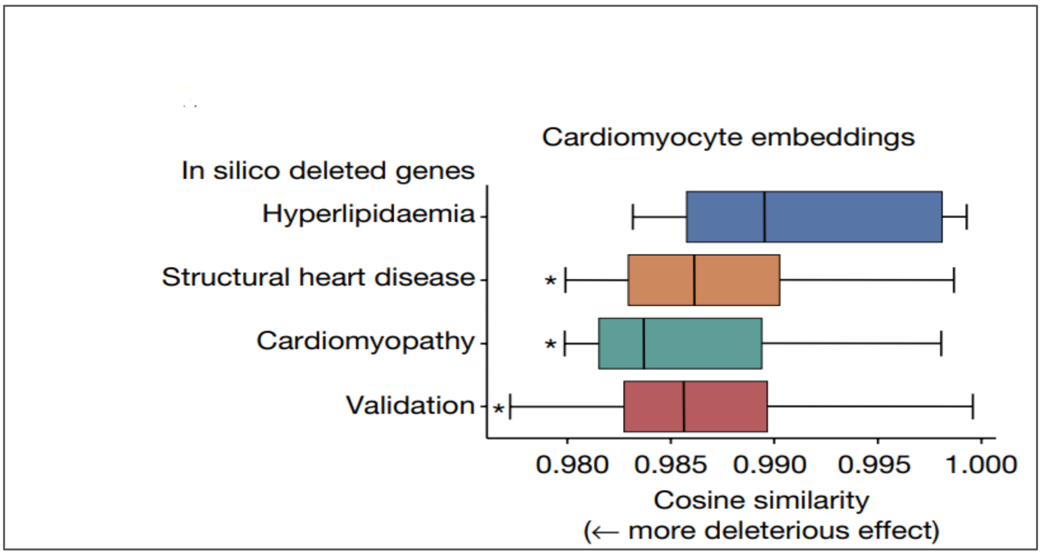


**Fig 5.** Cosine similarity measured between original values and in silico deleted values in cardiomyocytes

This study was further enriched for human phenotype. And top 25 deleted genes, most significant were transcription factors known to regulate **myocardial development** (heart muscles that forms and matures during embryonic development) e.g. FOXM1 and TEAD4

# 3.2. Chromatin Dynamics Prediction

### 3.2.1. Key Biological Concepts

- **Chromatin structure:** Double Helix structure of DNA wrapped around proteins called Histones(also called structural protein that give DNA its shape). These histone proteins are wrapped more and more depending on what stage we are in the cell's life.

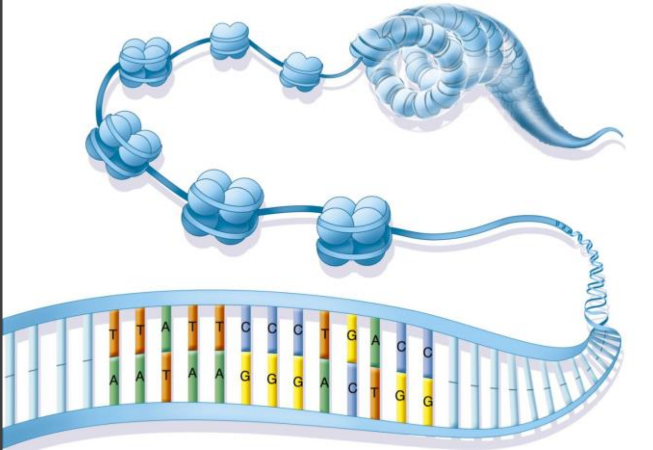


**Fig 5.** Chromatin Structure Diagram

### 3.2.2. Fine tuning Process

**Dataset:** Embryonic stem cells [reference 6] included genes found in conserved regions [reference 5]

**No of samples:** 15,000 single cell transcriptomes

**Classification labels:** bivalently marked genes vs (promoters were unmethylated or marked solely by H3K4me3 )

### 2.3. Results and Observations

ROC-AUC was used to measure the performance and comparison was done with other classification algorithms.

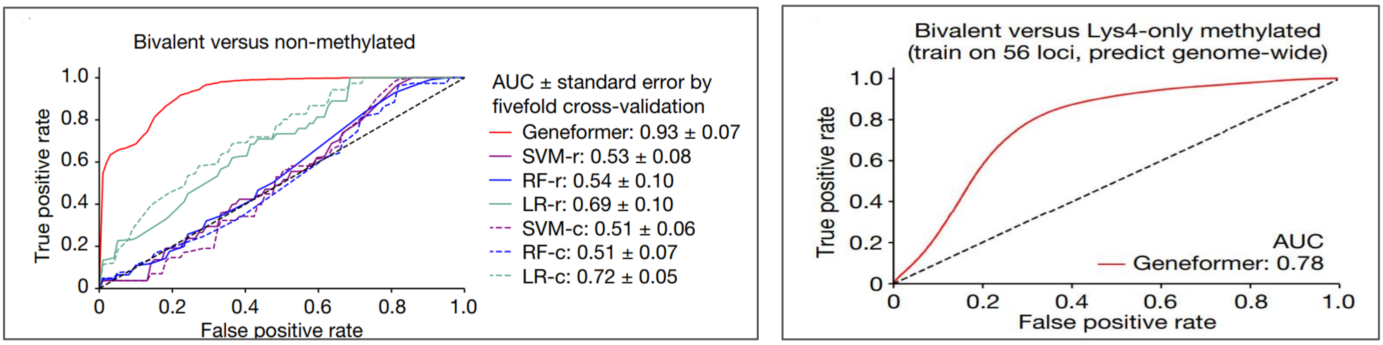


**Fig 6.** ROC – AUC Graph for evaluating Bivalent vs non-methylated and Bivalent vs Lys4

### ( I ) Another Fine-tuning process related to Chromatin

**Dataset:** cells undergoing iPSC to cardiomyocyte differentiation [7] with no associated Chip–seq or genomic distance data

**No of samples:** 34,000 single cell transcriptomes

**Classification labels:** long range vs short range transcription factors (based on genomic distance)

Determining the genomic distance over which transcription factor binding influences the downstream expression which valuable for interpreting regulatory variants and inferring target genes from genome occupancy data

### Results and Observations

ROC-AUC was used to measure the performance and comparison was done with other classification algorithms.

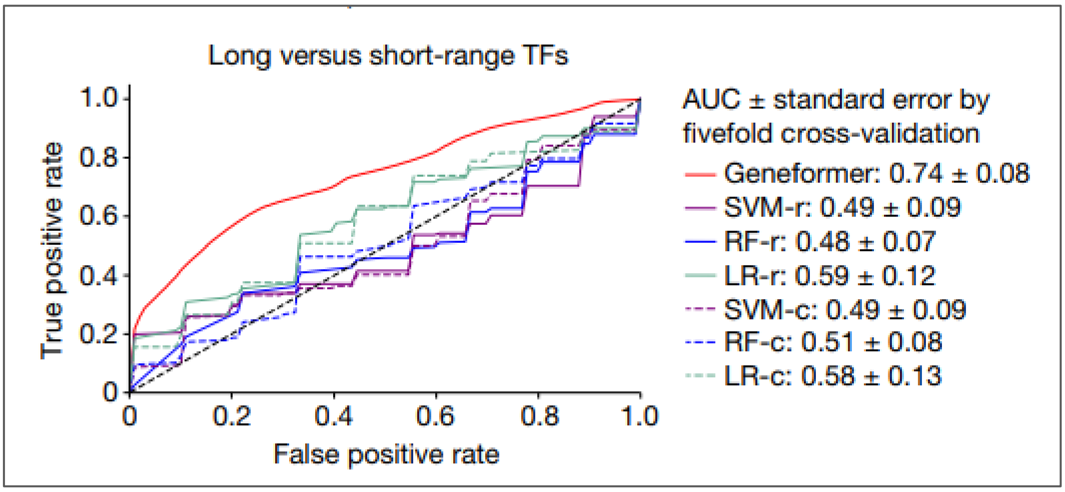


**Fig 7.** ROC – AUC Graph for classification long range vs short range TFs

This showcases geneformer's ability to improve predictions even for higher order transcription factor property of regulatory range. This is a challenging problem to infer from transcriptional data alone.

## 3.3. Network Dynamics prediction

### 3.3.1. Key Biological Concepts

- **Gene regulatory networks (GRNs)** are complex network of molecular regulators that control the gene expression levels in a cell. The hierarchical structure can be used to organize the complex interactions within a cell.

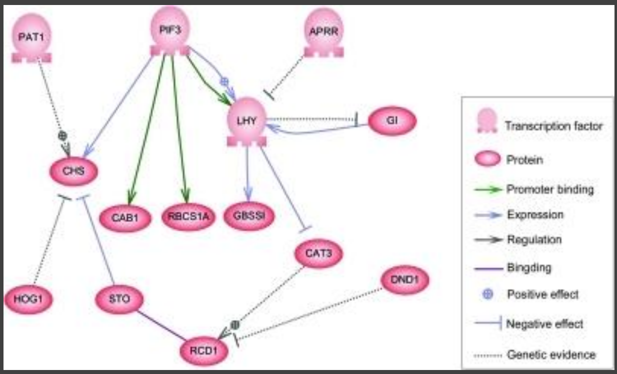


**Fig 8:** Example for GRN

- **Top level (or master regulators):** these are the genes that are at the top of the hierarchy that control expression of many other genes, they play a crucial role in determining the cell fate and functions.
- **Intermediate regulators:** These genes are regulated by master regulators; in turn they control expressions of other genes. They act as intermediators in the regulatory cascade.
- **Target genes:** They are at the bottom of the hierarchy; they often encode the proteins that perform specific cell functions

By identifying the hierarchy for disease, it can help to design therapies targeting the core regulatory elements which is driving the disease process rather than targeting the peripheral targets which may not be disease modifying.

### 3.3.2. Fine Tuning Process

**Dataset:** normal **endothelial cells (ECs)** from the **Heart Atlas** [reference 8]

**No of samples:** 30,000 single cell transcriptomes

**Classification labels:** central versus peripheral factors

Geneformer's pretraining process resolved the issue limited data. Fine tuning was done to classify central vs peripheral targets in endothelial cells, the comparison was done to

- classify N1 network central vs peripheral (ROC- AUC 0.81)
- N1 activated vs non target (ROC- AUC 0.81)
- additional study was to check the minimum amount of fine-tuning data required

### 3.3.3. Results and Observations

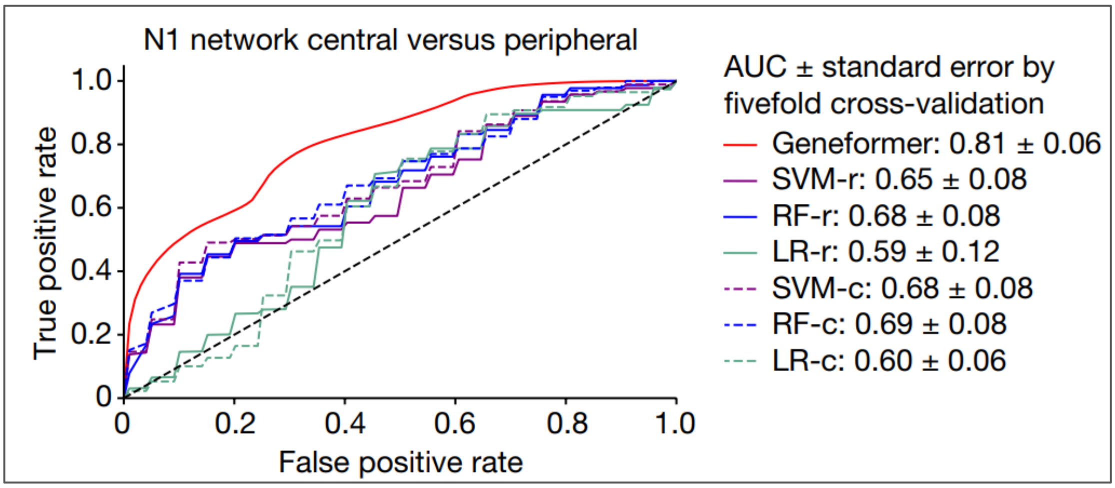


**Fig_9 :** Classification Performance N1 central vs peripheral

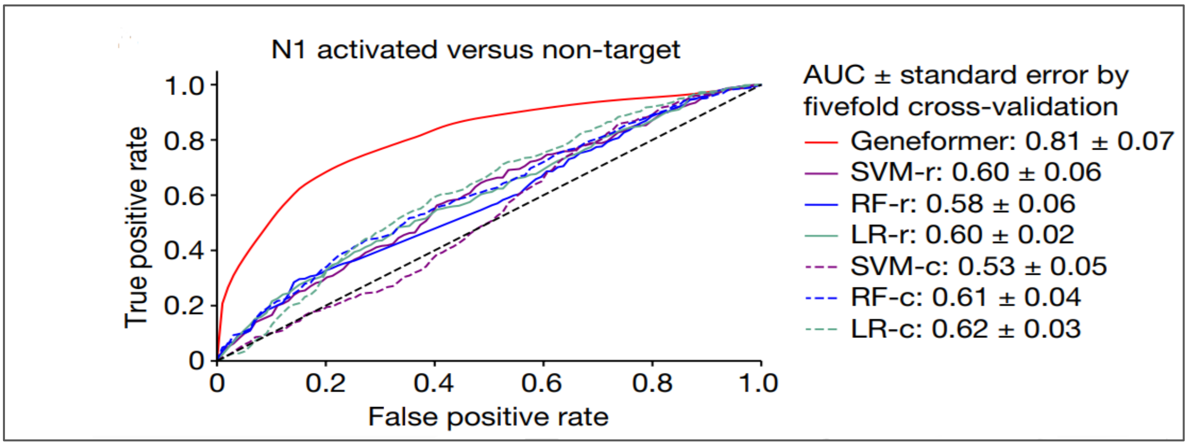


**Fig 10:** Classification Performance N1 activated vs non target

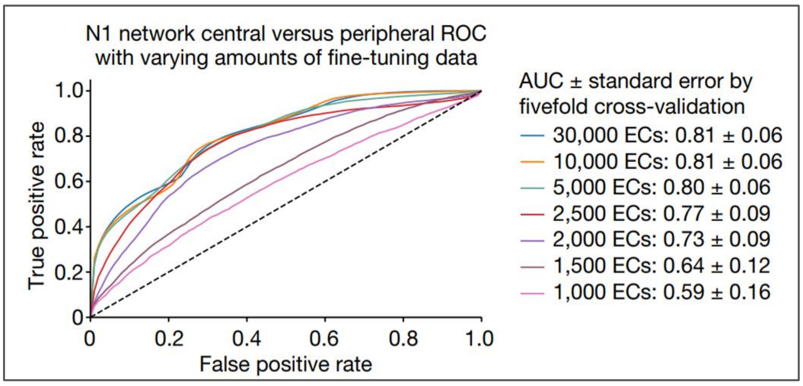


**Fig 11:** Checking minimum finetuning data threshold

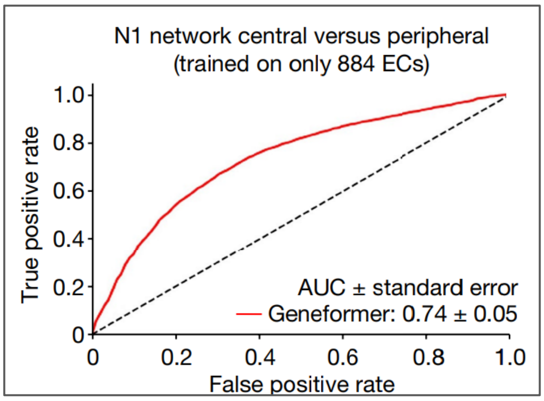


**Fig 12:** Performance for finetuning for 884 samples

- Fig 11 showcases iteratively decreasing the number of samples from 30,000 to 1,000 ECs. Predictive power remained the same while reducing till 5000 ECs.
- Fig 12. More minuscular number was also tested (884 ECs), Geneformer had better predictive power than other classification techniques which were using 30,000 ECs

**Conclusion:** Like any other Machine learning models Geneformer’s predictive power depends on the quality of the data not the quantity of the data.

## 3.4. Pretraining Encoded Network Hierarchy

### 3. 4.1. About Geneformer’s attention weights

GeneFormer’s attention weights are iteratively optimized according to the training objective. Geneformer’s 6 layers have 4 attention heads each and these pay attention to distinct classes of genes.

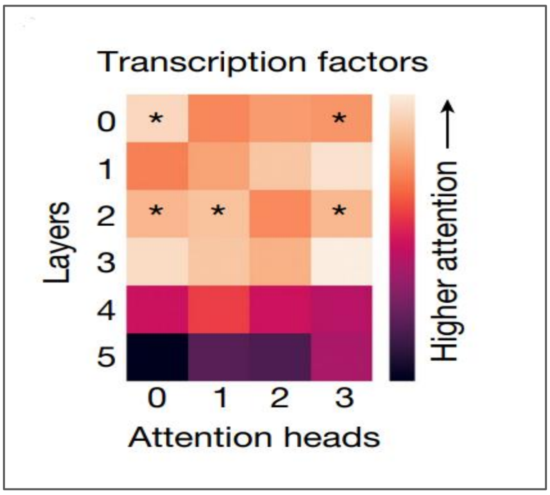
The trained attention weight for each reflects on

- which genes the gene pays attention to and
- which genes pay attention to that gene

### 3. 4.2. Fine tuning Process

**Dataset:** attention weights in aortic ECs [reference 9]

**Problem:** Examining the pretrained attention weights on whether its learning network dynamics



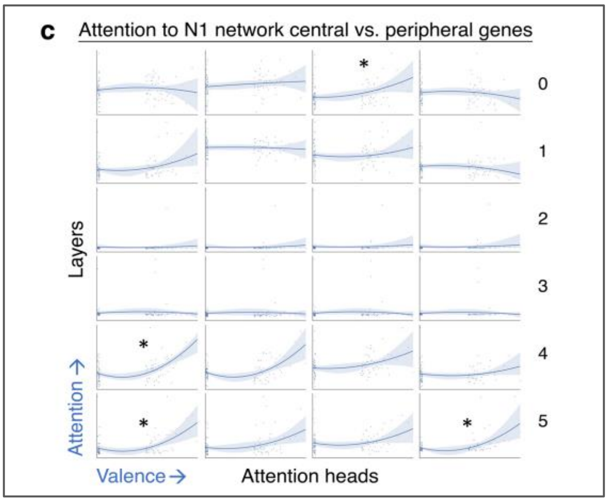


**Fig 13:** *ed are attention weights attending transcription factors

**Fig 14:** *ed are attention weights attending central nodes in N1 n/w

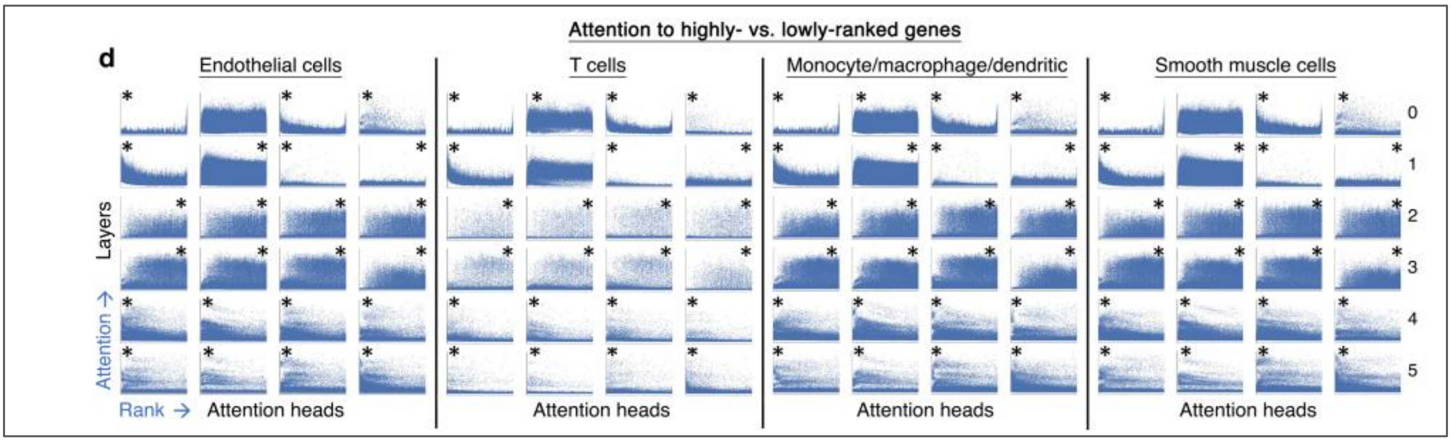


**Fig 15:** * are attention weights attending transcription factors in Endothelial, Tcells, Monocytes , smooth muscle cells

### 3.4.3. Results and Observations

- 20 % of attention heads attended transcription factors than genes (self-supervised manner giving relative importance to transcription factors in distinguishing cell states) in aortic ECs
- These centrality driven attention heads consistently attended highest ranked genes in each cells unique rank value encodings in smooth muscle cells, T cells, and macrophage, monocyte and dendritic cells
- In N1-dependent network specific attention heads attended peripheral central regulatory nodes than peripheral genes (reason for distinguishing central vs peripheral nodes in downstream task 3)

**Conclusion:** For each and every case majority of the attention heads in final layers were constantly attending the transcription factors and centrality driven factors, while attention heads in lower layers more diverse in gene ranks, they have attended suggesting its orienting different cell states

## 3.5. In Silico Gene network analysis

### 3.5.1. Key Biological Concepts

Gene embeddings reflect on the joint output of attention weights in the network.

**Key concepts**

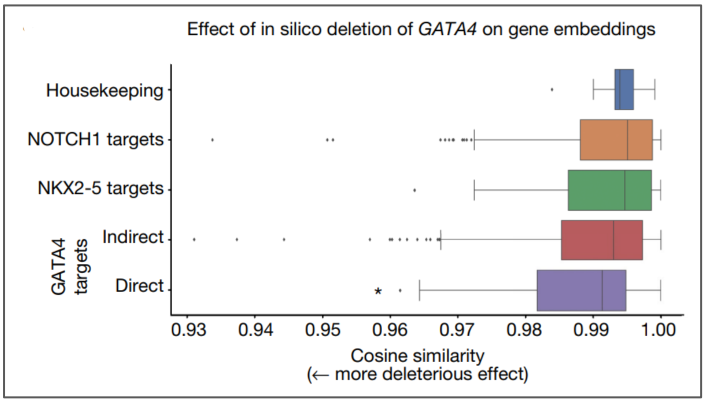
- **Transcription Factor:** is one of proteins that regulates gene expression by binding to specific DNA sequences near genes, mostly in the promoter or enhancer regions. They act as "gene switches," turning genes on or off by controlling the recruitment of RNA polymerase, the enzyme responsible for transcribing DNA into RNA.
- **Direct targets:** these are the genes whose promoters and enhancers are directly bound by the transcription factor
- **Indirect targets**: Genes whose expression is influenced indirectly through other intermediate proteins or pathway triggered by the transcription factors

### 3.5.2. Fine tuning Process

**Data used:** Embeddings in Fetal Cardiomyocytes [reference 10]

**Objective:** In silico deletion of the genes to determine whether the pretrained Geneformer has learned the connection between transcription factors and their targets.

### 3.5.3. Results and Observations



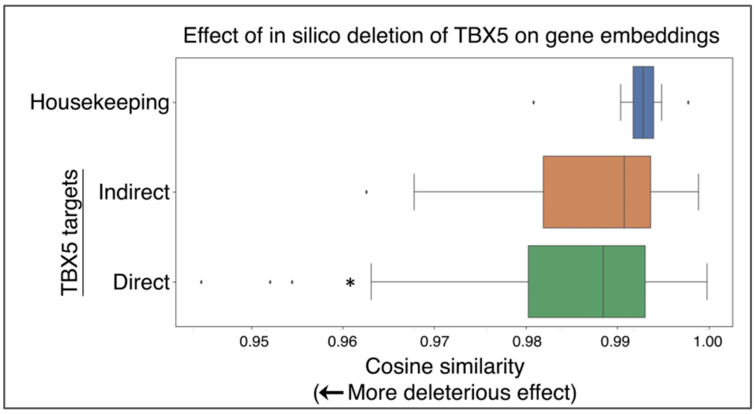


**Fig 16:** Direct targets of GATA4 had more deleterious effect than indirect targets

**Fig 17:** Deleterious effect of TBX5 Genes

- In Silico deletion of GATA4(a known conditional heart disease gene). Direct targets of GATA4 significantly have more effect than the indirect targets (**Fig 16**)
- In Silico deletion of TBX5 in cardiomyocytes, also showed similar results as direct targets were more significantly impacted than indirect targets (**Fig 17**)

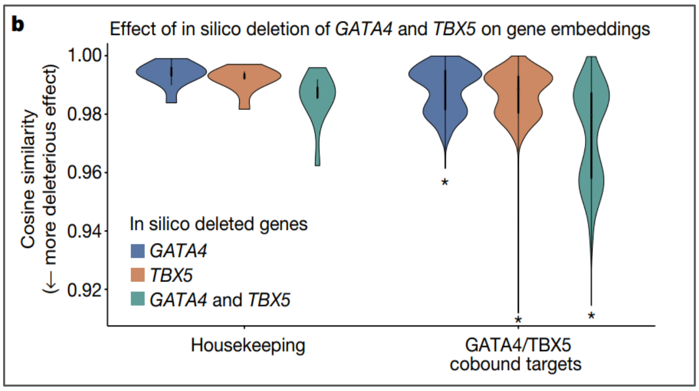


**Fig 18:** Combination deletion effect of TBX5 and GATA4

- Combined effect of In Silico Deletion of both TBX5 and GATA4 was also studied, indeed in silico deletion of TBX5 or GATA4 significantly had more deleterious effect on their known co-bound target compared to housekeeping genes (**Fig 18**)
- Also, in silico deletion of both GATA4 and TBX5 in combination even had greater impact on their known co-bound targets than the sum of their in-silico deletion, showcasing Genformer recognized their cooperative action at these co-bound targets

## 3.6. In Silico Treatment analysis

### 3.6.1. Key Biological Concepts

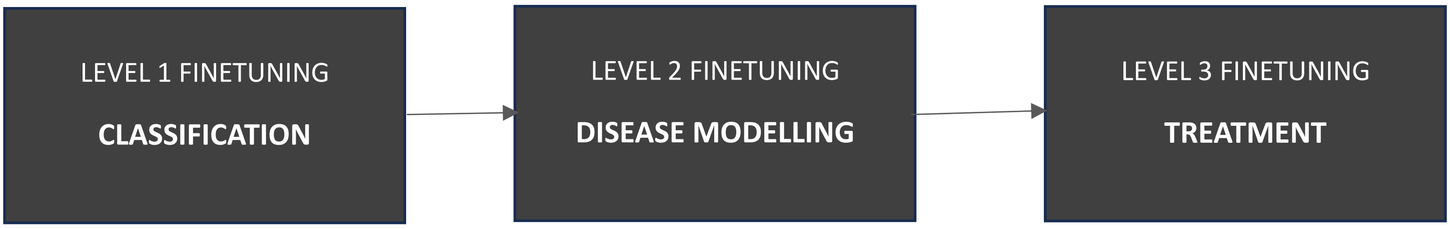
**Key concepts**

- **Cardiomyopathy:** A disease that makes it harder for the heart to pump blood to rest of the body efficiently.
- **Hypertrophic:** The heart muscle cells thicken abnormally which will obstruct the blood flow
- **Dilated:** The heart ventricle expands enlarge and weaken, reducing its effects to pump properly
- **Non- Failing Hearts**: Refers to the condition when heart functions normally and not experiencing any problem.

In silico treatment analysis is the downstream task to check if the **perturbation strategy** can be used to **model human diseases** and reveal the **candidate therapeutic targets**

### 3.6.2. Fine tuning Process

Three step Fine-tuning process:



LEVEL 1 FINETUNING

**CLASSIFICATION**

LEVEL 2 FINETUNING

**DISEASE MODELLING**

LEVEL 3 FINETUNING

**TREATMENT**

## LEVEL 1 FINETUNING: CLASSIFICATION

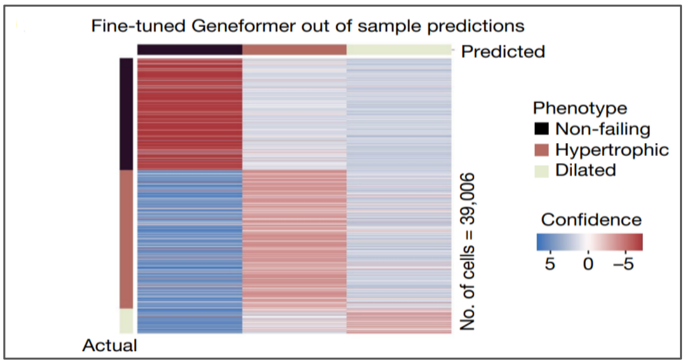


**Fig 19:** Out of sample classification performance

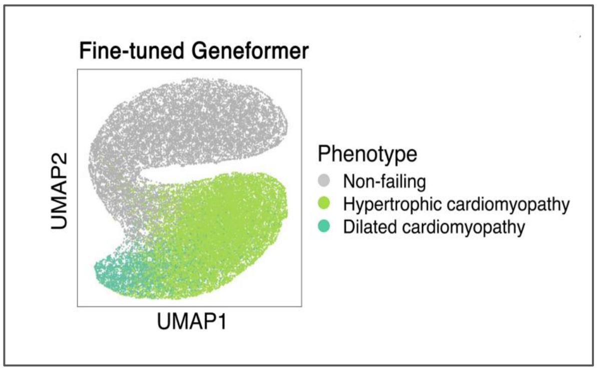


**Fig 20:** lower dimension representation

- Classification process to distinguish between **Cardiomyopathy cells** (dilated and hypertrophic) from **non-failing hearts**.
- out of sample accuracy =90%.

## LEVEL 2 FINETUNING: DISEASE MODELLING

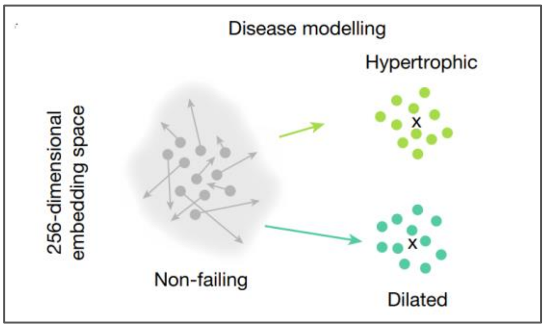


**Fig 21:** Embeddings shift from non-failing to both cardiomyopathy sates

In silico deletion was done in the embeddings of Non-failing Hearts to determine whether the deletion of certain genes can shift the embeddings from non-failing state to either dilated cardiomyopathy or hypertrophic cardiomyopathy.

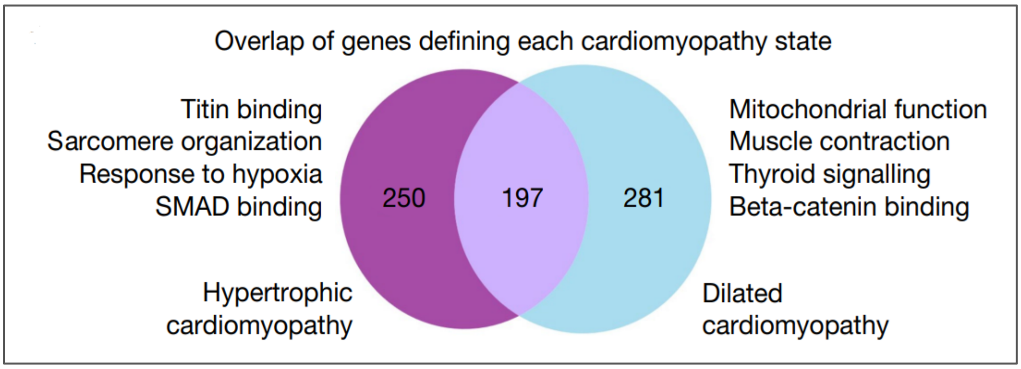


**Fig 22:** Pathways Enriched for deleted genes

- In Silico deletion of 447 genes shifted from non-failing hearts towards hypertrophic cardiomyopathy state. These were enriched for pathways including Titin Binding and sarcomere organization.
- In Silico deletion of 478 genes shifted from non-failing hearts towards dilated cardiomyopathy state. These were enriched for pathways involved in muscle contraction and mitochondrial function.
  8/*

**LEVEL 3 FINETUNING: TREATMENT**

Using the enriched pathways detected the next step was to determine whether the inhibition or activation of specific genes would shift the embeddings back towards the non-failing heart state. This makes way for generating candidate therapeutic targets.

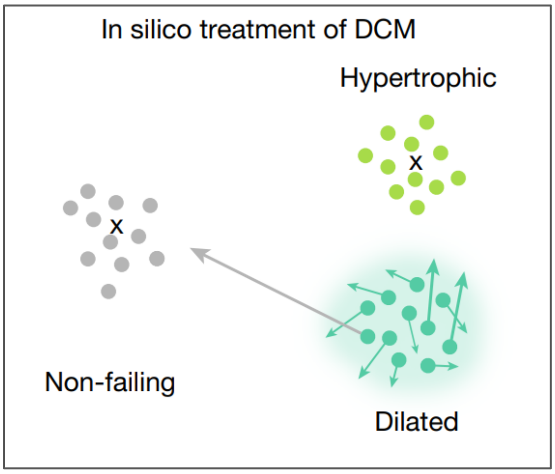


**Fig 23:** Embeddings shift from dilated to non-failing

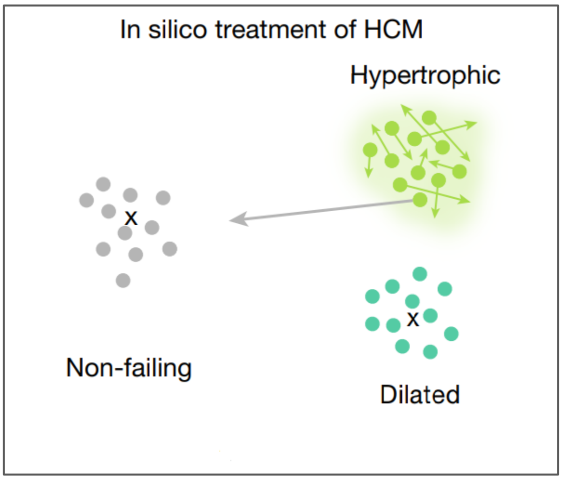


**Fig 24:** Embeddings shift from hypertrophic to non-failing

**Experimental validation of geneformer predicted therapeutic candidates for dilated cardiomyopathy**

- Titin(TTN) truncating mutations are leading cause of dilated cardiomyopathy in humans and are found in 20% of affected patients
- TTN+/- in the A-band are known to exhibit contractile stress
- CRISPR mediated knockout of GSN and PLN+/- significantly improved contractile stress of TTN=/- cardiac microtissues

**Conclusion:** These findings provide experimental validation in support of utility of Genformer as tool for discovery of candidate therapeutic targets in human disease.

## Conclusion

Geneformer's pretraining process served as the helping hand required in addressing the problem of limited data during various downstream tasks in network biology predictions. Analysis done on Downstream tasks can alone be further applied to other therapeutic specific research.

- Interpreting CNVs can help in identifying genes affected by the dosage changes, guiding targeted therapies , personalized treatments and drug development in CNV-related diseases.
- In silico perturbation strategy can be applied to a wide range of purposes, such as:
    - Assessing how specific genes affect normal cell function,
    - Investigating which disease states are caused by the deletion of certain genes,
    - Exploring whether adding specific genes can revert a disease state to a healthy one.

  The effects can be studied by measuring cosine similarity and embeddings shift.
- Identifying master regulators for any disease is useful for enabling targeted therapies, personalized medicine, improved drug development, broader disease modulation and identification of biomarkers for monitoring disease progression

At the end of the day for any machine learning model learns patterns according to the presented data it is trained on. Overall Geneformer covers the basic concepts of pretraining and transfer learning required for building a biomedical single cell Foundation models. The rank value encoding proposed in the model may not be optimal way to present single-cell transcriptomes, and domain experts could offer improved pretraining strategies for better performance. *scGPT*, *scBERT*, and *scBERT are* some of the latest foundation models based on Geneformer.

# REFERENECES

[ 1]. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature 536, 285–291 (2016).

[ 2 ] Shihab, H. A., Rogers, M. F., Campbell, C. & Gaunt, T. R. HIPred: an integrative approach to predicting haploinsufficient genes. Bioinformatics 33, 1751–1757 (2017).

[ 3 ] Ni, Z., Zhou, X. Y., Aslam, S. & Niu, D. K. Characterization of human dosage-sensitive transcription factor genes. Front. Genet. 10, 1208 (2019)

[ 4 ] Collins, R. L. et al. A cross-disorder dosage sensitivity map of the human genome. Cell 185, 3041–3055 (2022).

[ 5 ] Bernstein, B. E. et al. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell 125, 315–356 (2006).

[ 6 ] Franzén, O., Gan, L.-M. & Björkegren, J. L. M. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. Database 2019, baz406 (2019).

[ 7 ] Selewa, A. et al. Systematic comparison of high-throughput single-cell and singlenucleus transcriptomes during cardiomyocyte differentiation. Sci. Rep. 10, 1535 (2020)

[ 8 ] Litviňuková, M. et al. Cells of the adult human heart. Nature 588, 455–472 (2020).

[ 9 ] Li, Y. et al. Single-cell transcriptome analysis reveals dynamic cell populations and differential gene expression patterns in control and aneurysmal human aortic tissue. Circulation 142, 1374–1388 (2020).

[ 10 ] Cao, J. et al. A human cell atlas of fetal gene expression. Science 370, 808 (2020)