Transfer learning enables predictions in network biology

Anshul Saini and Krrish Agarwalla

Seminar on (Large) Language Model Applications for Text and Biological Data

Supervisor: Roberto Olayo Alarcon

Ludwig-Maximilians-Universität München (LMU)

January 14, 2026

Abstract

The application of large language models (LLMs) to biological data has revolutionized the analysis of single-cell transcriptomics. This report presents a reproduction and extension of *Geneformer*, a context-aware, attention-based deep learning model pretrained on 30 million single-cell transcriptomes [1]. Geneformer leverages transfer learning to make predictions in network biology settings with limited data, such as gene dosage sensitivity and chromatin dynamics [1].

While the original authors demonstrated state-of-the-art performance, the computational cost of training and deploying such models remains a barrier. In this project, we reproduce the core findings of Geneformer, validating its rank-value encoding, pretraining strategies, and diverse fine-tuning methodologies, including dosage sensitivity predictions, chromatin dynamics modeling, and *in silico* perturbation analysis for therapeutic discovery [1]. Furthermore, we apply **Knowledge Distillation (KD)** to compress the model for resource-constrained environments. We successfully distilled the 10-million parameter Teacher model into multiple lightweight Student models with 4.3M, 3M, and 2M parameters. Our results demonstrate that the 4.3M distilled model retains 98% of the teacher's performance (84.53% vs 86.14% accuracy) on the downstream task of classifying Cardiomyopathy phenotypes, while requiring significantly fewer computational resources. This work paves the way for the democratization of genomic foundation models on consumer-grade hardware.

1 Geneformer: Explanation and Context

1.1 The Biological Context: The Need for Transfer Learning

Mapping and identifying key gene regulators within regulatory networks is fundamental to understanding disease mechanisms. However, mapping these architectures often requires massive amounts of transcriptomic data to learn the network dynamics between genes, which impedes discoveries in rare diseases or clinically inaccessible tissues.

Standard supervised learning trains a new model from scratch for each specific task. In contrast, **Transfer Learning** allows a model to learn fundamental knowledge from a large, general dataset (pretraining) and apply it to a specific task with limited data (fine-tuning). Geneformer applies this concept to transcriptomics, treating a single cell as a "sentence" and genes as "words" to model context-specific network dynamics.

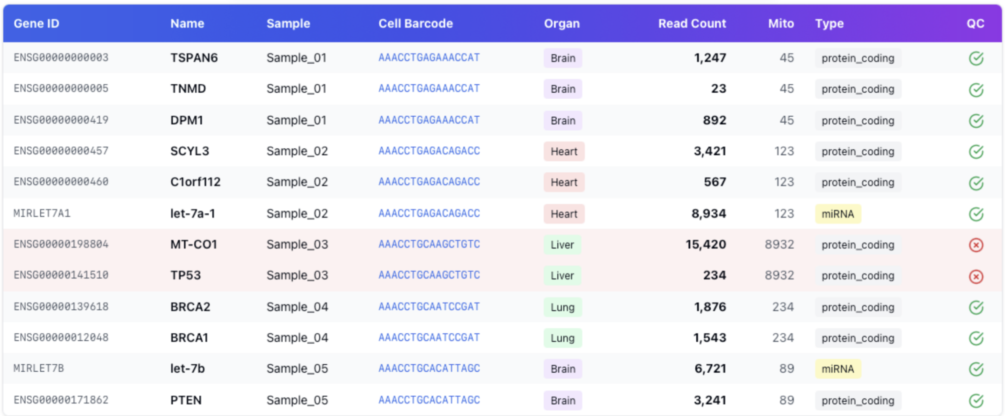
1.2 Data Curation: Genecorpus-30M

The model is built upon *Genecorpus-30M*, a large-scale pretraining corpus comprising 29.9 million human single-cell transcriptomes aggregated from 561 public datasets (including CellxGene, PanglaoDB, and others) [1]. The curation process involves a rigorous, standardized pipeline to ensure data quality and comparability across diverse sequencing platforms:

- **Filtering Criteria:** To remove low-quality data, cells were excluded based on three specific metrics:
  - *Mitochondrial Content:* Cells with mitochondrial read percentages greater than 3 standard deviations above the mean within a dataset were removed, as high mitochondrial content typically indicates cell death (apoptosis) or membrane rupture [1].
  - *Read Depth:* Cells with total read counts outside of 3 standard deviations from the dataset mean were excluded to filter out empty droplets or potential doublets [1].
  - *Gene Count:* Cells with fewer than seven detected protein-coding or miRNA genes were removed, as the model's 15% masking objective would be ineffective on such sparse data [1].

- **Exclusion of High Mutational Burden:** Cells with high mutational burdens, such as malignant cancer cells or immortalized cell lines, were deliberately excluded. This ensures the model learns ”normal” network topology without the confounding effects of substantial network rewiring found in cancer [1].
- **Normalization Strategy:** To handle technical variation across droplet-based sequencing platforms, raw transcript counts were normalized by the total read count per cell. This accounts for varying sequencing depths between experiments while preserving the relative abundance of genes [1].

Ultimately, 27.4 million cells passed these quality filters to form the final training corpus. The data structure includes gene IDs, read counts, and extensive metadata (organ, sequencing platform, etc.), as visualized below.



Gene ID	Name	Sample	Cell Barcode	Organ	Read Count	Mito	Type	QC
ENSG00000000003	TSPAN6	Sample_01	AAACCTGAGAAACCAT	Brain	1,247	45	protein_coding	✓
ENSG00000000005	TNMD	Sample_01	AAACCTGAGAAACCAT	Brain	23	45	protein_coding	✓
ENSG00000000419	DPM1	Sample_01	AAACCTGAGAAACCAT	Brain	892	45	protein_coding	✓
ENSG00000000457	SCYL3	Sample_02	AAACCTGAGACAGACC	Heart	3,421	123	protein_coding	✓
ENSG00000000460	C1orf112	Sample_02	AAACCTGAGACAGACC	Heart	567	123	protein_coding	✓
MIRLET7A1	let-7a-1	Sample_02	AAACCTGAGACAGACC	Heart	8,934	123	miRNA	✓
ENSG00000198804	MT-CO1	Sample_03	AAACCTGCAAGCTGTC	Liver	15,420	8932	protein_coding	⊗
ENSG00000141510	TP53	Sample_03	AAACCTGCAAGCTGTC	Liver	234	8932	protein_coding	⊗
ENSG00000139618	BRCA2	Sample_04	AAACCTGCAATCCGAT	Lung	1,876	234	protein_coding	✓
ENSG00000012048	BRCA1	Sample_04	AAACCTGCAATCCGAT	Lung	1,543	234	protein_coding	✓
MIRLET7B	let-7b	Sample_05	AAACCTGCACATTAGC	Brain	6,721	89	miRNA	✓
ENSG00000171862	PTEN	Sample_05	AAACCTGCACATTAGC	Brain	3,241	89	protein_coding	✓

Figure 1: Example of the transcriptomic dataset structure. Each row represents a gene’s expression within a specific cell, annotated with organ and quality control metrics [2].

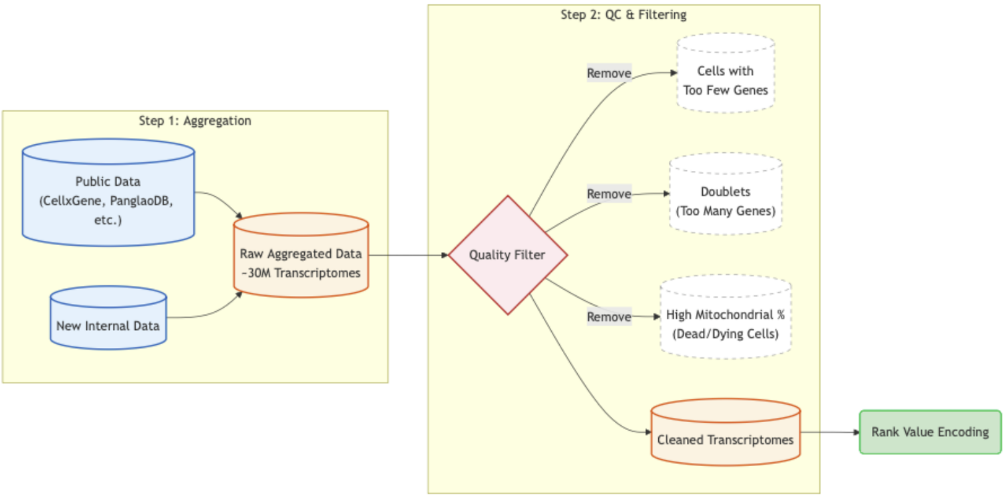


Figure 2: The data curation pipeline used to generate Genecorpus-30M [2]. Raw data is aggregated, filtered for quality control, and passed to Rank Value Encoding.

1.3 Rank Value Encoding

A critical innovation of Geneformer is the **Rank Value Encoding** [1]. Traditional transcriptomic analyses often rely on raw expression counts, which can be heavily biased by ubiquitously expressed "housekeeping" genes. These genes consume a large proportion of sequencing reads but often provide little information about the specific cellular state or disease phenotype. To mitigate this bias and focus the model's attention on biologically informative regulatory genes (transcription factors), Geneformer employs a rank-based input representation.

The encoding process transforms raw data into a sequence of tokens through the following mathematical steps [2]:

1. **Normalization:** First, the raw read count of a gene is normalized by the sequencing depth of that cell (total reads) and scaled by a gene-specific median factor. For a gene g in cell c, the normalized value V_{gc} is calculated as:

   $$V_{gc} = \frac{R_{gc}}{D_c} \times \frac{10,000}{M_g} \tag{1}$$

   Where:

   - R_{gc} is the raw read count of gene g in cell c.
   - D_c is the total read depth of cell c.
   - M_g is the non-zero median expression of gene g across the entire Genecorpus-30M.

   This step effectively deprioritizes housekeeping genes, which typically have high median expression (M_g), thereby lowering their normalized value V_{gc}.

2. **Ranking:** The genes within each cell are then sorted in descending order based on their normalized values V_{gc}.

   $$\text{Rank}(c) = \text{argsort}_{\text{desc}}(V_{1c}, V_{2c}, ..., V_{nc}) \tag{2}$$

3. **Tokenization:** The resulting ranked list of genes is converted into a sequence of tokens. Highly expressed, cell-state-defining genes (like transcription factors) appear earlier in the sequence, while housekeeping genes are pushed towards the end [1].

This rank-based representation is robust to variations in sequencing depth and ensures that the model learns the hierarchical structure of gene regulation rather than absolute expression levels.

1.4 Model Architecture and Pretraining

Geneformer is built upon a standard Transformer encoder architecture, specifically adapted for the constraints of single-cell data. The model consists of a stack of *six transformer encoder units*, a depth chosen based on the maximum data scale available for effective pretraining.

The specific architectural hyperparameters are as follows:

- **Input Size:** 2,048 tokens. This length was selected to fully represent 93% of the rank value encodings in the Genecorpus-30M dataset.
- **Embedding Dimension (d_{model}):** 256. Each gene token is embedded into a 256-dimensional vector space.
- **Attention Heads:** 4 heads per layer.
- **Feed-Forward Dimension (d_{ff}):** 512.
- **Regularization:** A dropout probability of 0.02 is applied to fully connected layers and attention probabilities to prevent overfitting.

The model utilizes full dense self-attention across the entire input sequence. Training was optimized using the AdamW optimizer with a linear learning rate scheduler (max learning rate 1×10^{-3}) and a warmup period of 10,000 steps in the paper.

1.4.1 Attention Mechanism: Learning Network Dynamics

The core of Geneformer is the multi-head self-attention mechanism, which enables the model to be "context-aware." In natural language processing, attention allows a model to understand that the word "bank" has a different meaning in the context of "river" versus "money." Analogously, Geneformer uses attention to learn how the function of a specific gene changes based on the co-expression of other regulators in the cell.

The attention weights (A) for a given gene reflect which other genes it "attends" to (inputs) and which genes attend to it (regulatory influence). Analysis of the pretrained weights revealed that the model learns biological hierarchies in a completely self-supervised manner:

- **Transcription Factor Specialization:** Approximately 20% of the attention heads significantly prioritize transcription factors (TFs) over other genes, recognizing their role as global regulators.
- **Layer-wise Hierarchy:** The model exhibits a functional hierarchy across its six layers. Early layers tend to survey the broad input space (diverse attention), while deeper layers become “centrality-driven,” focusing heavily on the highest-ranked genes that uniquely define the cell state.

This demonstrates that the attention mechanism successfully encodes the gene regulatory network topology without any prior biological labeling.

1.4.2 Masked Language Modeling (MLM)

The model is pretrained using a **Masked Learning Objective**, similar to BERT in NLP [1].

- **Process:** 15% of the genes in a sequence are randomly masked [1].
- **Objective:** The model must predict the identity of the masked genes based solely on the context of the remaining unmasked genes.

This forces the model to learn the fundamental ”grammar” of gene networks in a self-supervised manner [1].

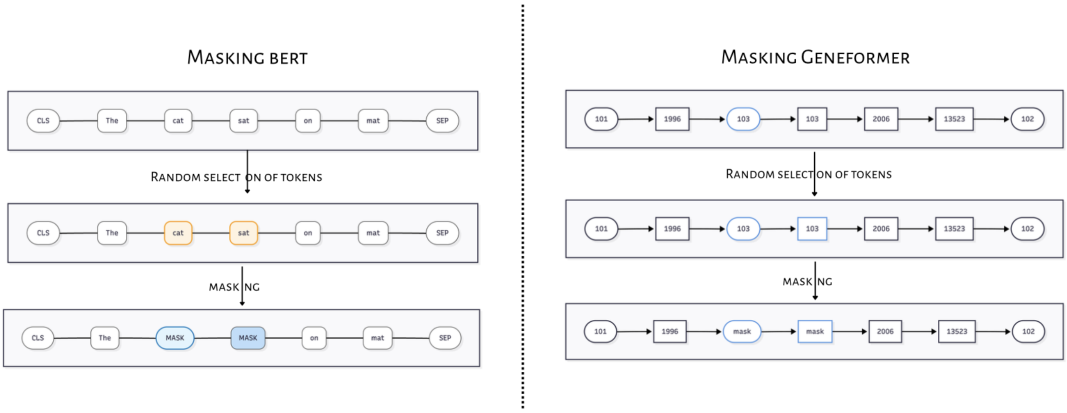


Figure 3: Geneformer utilizes a masking strategy where 15% of gene tokens are masked, forcing the model to learn network representations [2].

2 Our Contribution: Knowledge Distillation

2.1 Problem Statement

While Geneformer is the state of the art model, it is computationally expensive. The original model contains 10 million parameters and required training on 12 V100 GPUs for 3 days [2]. This creates a barrier to entry for researchers with limited resources. Additionally, the dataset is massive (30M rows), causing standard training loops to crash RAM on consumer hardware [2].

2.2 Proposed Solution: Teacher-Student Distillation

To address this, we implemented **Knowledge Distillation (KD)** [2]. This technique compresses the knowledge of a large "Teacher" model into a smaller, lighter "Student" model.

Knowledge Distillation can be understood through an intuitive analogy: consider an experienced teacher guiding a student preparing for an exam. The teacher doesn't simply tell the student "the answer is C" instead, they explain their reasoning: "C is most likely correct, but B is also plausible if you interpret the question differently, while A and D are clearly wrong for these reasons." This nuanced guidance helps the student understand not just what is correct, but why, and how to think about similar problems in the future.

Similarly, in KD, a large pre-trained Teacher model transfers its learned knowledge to a compact Student model by sharing not just the final predictions, but the full probability distributions over all possible outputs. For instance, when classifying an image of a husky, the Teacher might output: husky (70%), wolf (20%), German shepherd (8%), other (2%). This probability distribution reveals that the model has learned meaningful visual similarities between related classes huskies do look more like wolves than like cats information that would be completely lost if we only provided the hard label "husky" with 100% confidence.

The effectiveness of knowledge distillation stems from this "dark knowledge" the implicit understanding of relationships between classes that emerges from extensive training. When the Student learns to mimic the Teacher's probability distributions, it inherits this *refined understanding without needing to see as many training examples or make the same mistakes.*

Our goal was to train a Student model with approximately 40-50% of the parameters of the original, using only $1/25^{th}$ of the compute resources, while maintaining comparable accuracy.

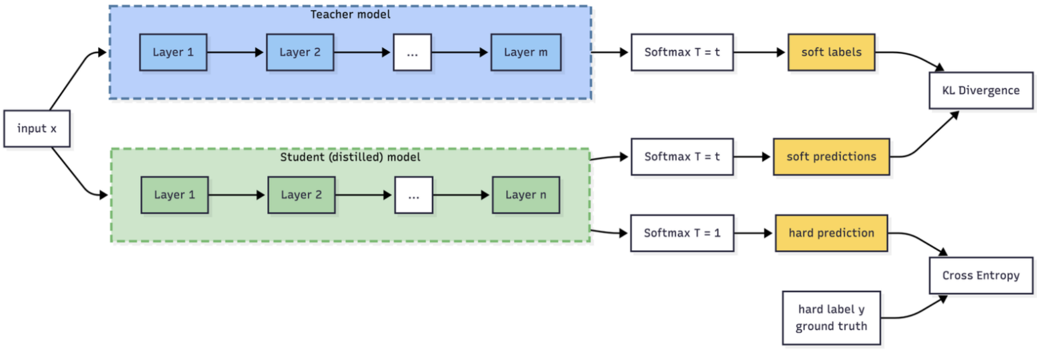


Figure 4: Schematic overview of the knowledge distillation process. The Student model learns from both the ground truth labels and the Teacher's probability distributions [2].

2.3 Distillation Methodology

To investigate the limits of model compression, we established a distillation framework involving a **frozen Teacher** and **multiple variable-sized Student** models [2].

1. **Teacher Model:** The original pretrained Geneformer (10M parameters). This model remains frozen during training to provide stable soft-label targets.

2. **Student Models:** We trained three distinct student variants to compare performance across different compression ratios:

    - **4.3M Parameters** (∼43% of Teacher size)
    - **3M Parameters** (∼30% of Teacher size)
    - **2M Parameters** (∼20% of Teacher size)

Parameter Counting Methodology: The reported parameter counts include all trainable weights in the model: token embeddings, positional encodings, transformer encoder blocks (Self-Attention and Feed-Forward networks), and the classification head.

To achieve these reductions, we decreased the network depth (number of transformer layers) and the embedding dimension from $d_{\text{model}} = 256$ to $d_{\text{model}} \in \{128, 96, 64\}$, as the total parameter count is dictated by the formula for BeRT style models:

$$N_{\text{params}} = 2 \cdot N_{\text{vocab}} \cdot d_{\text{model}} + N_{\text{layers}} \cdot \left(4d_{\text{model}}^2 + 2d_{\text{model}} \cdot d_{\text{ff}}\right) \tag{3}$$

where N_{vocab} is the vocabulary size, N_{layers} is the number of transformer layers, d_{model} is the embedding dimension, and d_{ff} is the feed-forward hidden dimension (typically $d_{\text{ff}} = 4 \cdot d_{\text{model}}$).

2.3.1 Loss Function Design

The Student is trained using a composite loss function that combines two distinct objectives [2]:

$$L_{total} = \alpha L_{CE} + (1 - \alpha) L_{KL} \tag{4}$$

Where:

- L_{CE} (Cross Entropy Loss): This measures the error between the Student's hard predictions ($p_{student}$) and the ground truth labels (y). It ensures the student learns the "right answer" by penalizing incorrect classifications.

$$L_{CE} = -\sum_{i=1}^{C} y_i \log(p_{student,i}) \tag{5}$$

- L_{KL} (Kullback-Leibler Divergence): This measures the distributional distance between the Student's soft predictions (P_S) and the Teacher's soft labels (P_T). Unlike Cross Entropy, which focuses on the single correct class, KL Divergence forces the Student to mimic the Teacher's uncertainty and secondary probabilities. This transfers the "dark knowledge" the structural relationships between classes (e.g., that Class A is more similar to Class B than to Class C) [2].

$$L_{KL}(P_T || P_S) = \sum_{i=1}^{C} p_{teacher,i} \log \left(\frac{p_{teacher,i}}{p_{student,i}} \right) \tag{6}$$

By learning from the Teacher's soft labels, the Student learns the "dark knowledge" the subtle relationships between classes that the Teacher has learned.

For the distillation process, we set the loss weighting coefficient $\alpha = 0.65$ to balance the contributions of soft predictions from the Teacher model and hard ground-truth labels. It should be noted that no systematic hyperparameter optimization was conducted in this study. We believe that *comprehensive hyperparameter tuning, including optimization of α, temperature T & learning rate would yield substantial improvements in Student model performance.*

2.4 Engineering Optimizations

To enable training on a single GPU, we implemented critical engineering optimizations to handle the computational constraints:

2.4.1 Dynamic Length-Grouped Sampling

Standard training pipelines typically sample batches randomly. However, as illustrated in Figure 5, the Genecorpus-30M dataset exhibits a highly skewed distribution of sequence lengths. While the maximum input size is 2,048 tokens, the median sequence length is only 234 tokens [2].

In a standard random batch, if a single long sequence (e.g., 2,048) is paired with many short sequences (e.g., 200), all short sequences must be padded with zeros to match the longest one. This results in the GPU spending the vast majority of its compute power **processing empty "pad" tokens.**

To solve this, **Dynamic Length-Grouped Sampling** was implemented:

1. **Sorting:** We gather a large "megabatch" of samples and sort them by sequence length.
2. **Grouping:** We construct minibatches using sequences of similar lengths.
3. **Dynamic Padding:** Each minibatch is padded only to the longest sequence *within that specific batch*, rather than the global maximum.

This strategy reduced padding overhead by 62% and resulted in a **29.4x speedup** compared to standard uniform sampling [2].

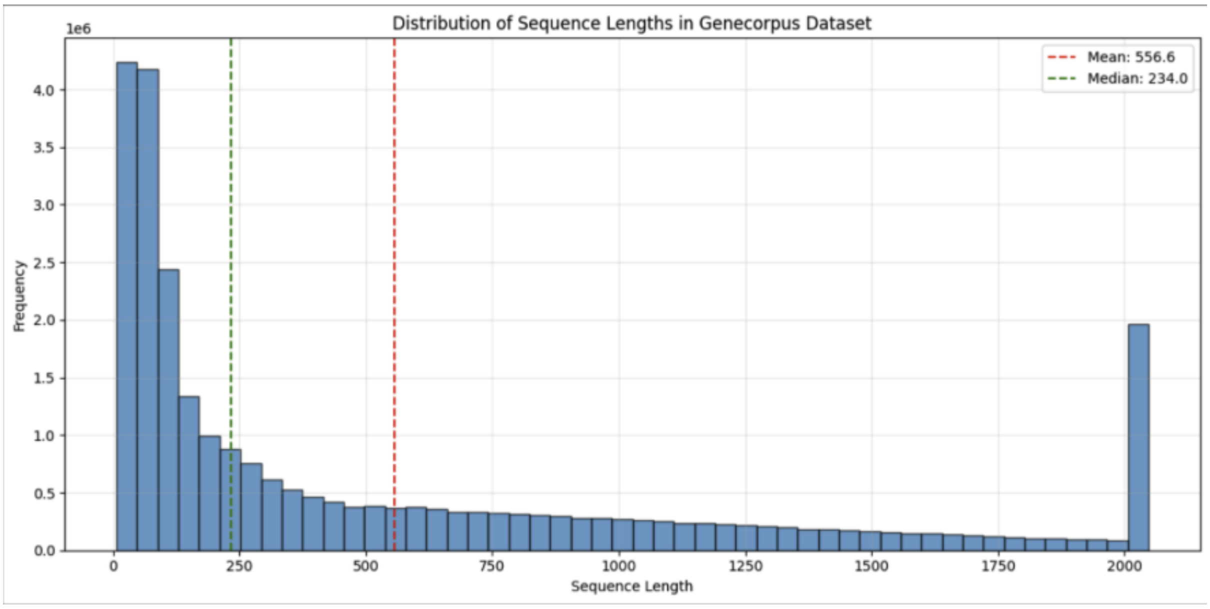


Figure 5: Distribution of Sequence Lengths in Genecorpus-30M. The stark contrast between the median length (234) and the maximum length (2048) highlights the inefficiency of random batching and the necessity of length-grouped sampling [2].

3 Results and Discussion

3.1 Initial Evaluation: Pretraining Metrics

Before fine-tuning on specific disease classification tasks, it was critical to verify that the Student model had successfully acquired the fundamental ”grammar” of gene regulation from the Teacher. We evaluated this using two key metrics on the held-out validation set:

- **Masked Language Modeling (MLM) Accuracy:** Measures how often the model correctly predicts a masked gene based on the context of the surrounding genes. High accuracy indicates a strong grasp of network topology.
- **Perplexity:** Measures the model’s uncertainty in predicting the next token (lower is better).

Evaluating these metrics ensures that the Knowledge Distillation process successfully transferred the generalizable biological knowledge, rather than just overfitting to a specific downstream task. As shown in Table 1, the distilled Student model (4.3M parameters) achieves performance comparable to the Teacher (10M parameters). The marginal gap in accuracy (0.04) and perplexity (6.08) confirms that the Student retains the core structural understanding required for biological inference [2].

Model Variant	Metric	Teacher	Student	Gap
4.3M Parameters	MLM Accuracy	0.2981	0.2490	-0.0491
(*Best Balance*)	Perplexity	16.76	24.25	+7.49
3M Parameters	MLM Accuracy	0.3049	0.2034	-0.1025
	Perplexity	16.32	34.32	+18.00
2M Parameters	MLM Accuracy	0.3049	0.1888	-0.1161
	Perplexity	15.77	40.30	+24.52

Table 1: Comparison of Pretraining Metrics across different Student model sizes. While the 4.3M model maintains a close proximity to the Teacher's perplexity, the 3M and 2M models show a drop in prediction capability, as indicated by the sharp rise in perplexity [2].

3.2 Experimental Setup: Cardiomyopathy Classification

Having validated the Student's general representational power, we validated our approach on a downstream task of classifying cardiomyocytes into three phenotypes [2]:

1. **Non-Failing (NF):** Healthy control tissue.
2. **Hypertrophic Cardiomyopathy (HCM):** Thickened heart muscle.
3. **Dilated Cardiomyopathy (DCM):** Enlarged heart chamber.

The dataset consisted of ∼132,000 cells [2]. To prevent data leakage and memorization, we split the data by Patient ID rather than by cell (Train: Patients A & B; Test: Patient C) [2].

3.3 Teacher Model Performance

The original Teacher model achieved high accuracy on this task. As shown in the confusion matrix and heatmap below (Figure 6), it correctly classifies the majority of HCM (88%) and DCM (84%) cases with high confidence.

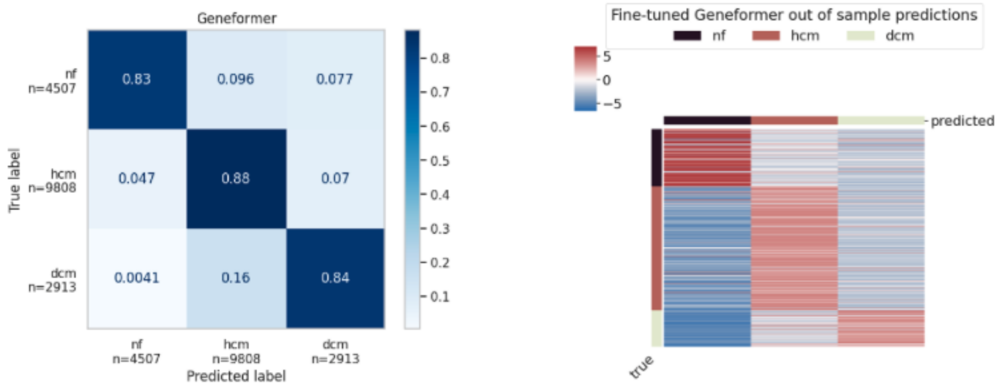


Figure 6: Teacher Model Results. The confusion matrix (left) confirms high precision, particularly for HCM (96%). The heatmap (right) shows confident separation between cell states [2].

3.4 Student Model (Distilled) Performance

The distilled 4.3M parameter Student model demonstrated exceptional performance retention. While there is slightly more overlap in the predictions compared to the teacher, the model correctly identifies most healthy patients (Figure 7).

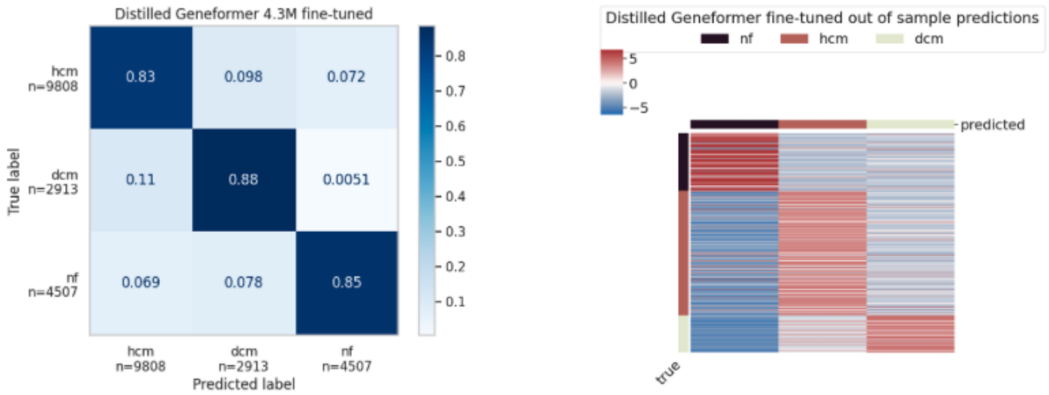


Figure 7: Student Model Results. The distilled model shows slightly more overlap in the center of the heatmap, indicating difficulty differentiating the specific signatures of the two cardiomyopathies, but maintains high accuracy for Non-Failing (NF) predictions [2].

3.4.1 Pretraining Metrics and Class-wise Analysis

Quantitatively, the Student model demonstrates remarkable robustness to class imbalance, a common challenge in biological datasets where disease samples are often rare compared to healthy controls.

As detailed in Figure 8, the model achieves an 88.4% Recall on the minority class (Dilated

Cardiomyopathy - DCM). This is a critical result because it indicates the model rarely misses a true DCM case, a high priority for diagnostic screening tools. However, the Precision for DCM is lower (66.3%), suggesting that while the model catches almost all DCM cases, it occasionally misclassifies other phenotypes (likely HCM, given the biological overlap) as DCM.

Conversely, for the Non-Failing (NF) healthy control group, the model maintains high Precision (84.2%) and *Recall (85.3%)*. This balance ensures trustworthiness; the model effectively screens out healthy patients without raising excessive false alarms. The high F1-scores across all three classes (87.6% for HCM, 75.8% for DCM, 84.7% for NF) validate that the distillation process preserved the Teacher's ability to distinguish subtle transcriptomic signatures across diverse biological states.

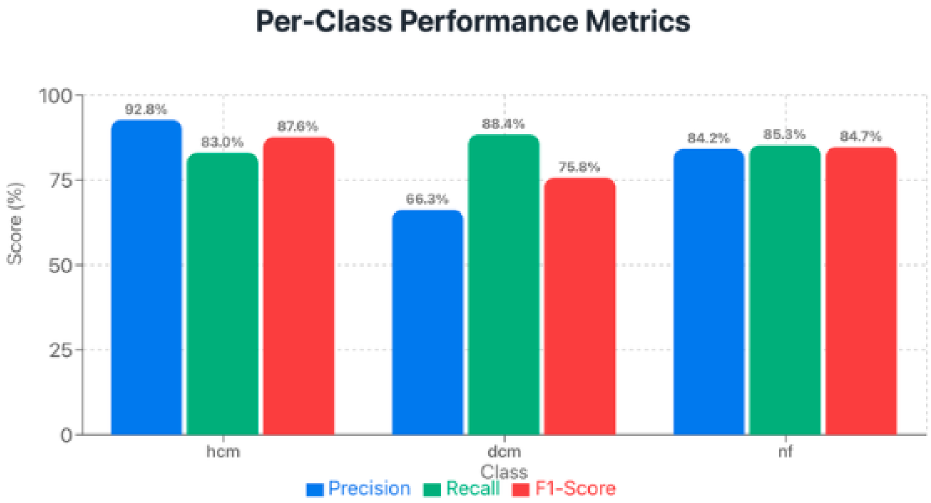


Figure 8: Per-Class Performance Metrics for the 4.3M Student Model. The model demonstrates robustness to class imbalance, achieving high recall on the minority DCM class [2].

3.4.2 Downscaling Analysis

We compared the Teacher model against varying sizes of the Student model (4.3M, 3M, and 2M parameters). As shown in Figure 9, the 4.3M parameter model retains **98%** of the Teacher's performance (84.53% vs 86.14%). Even the tiny 2M parameter model remains viable at 79.1% accuracy, proving that massive compute is not strictly necessary for this task.

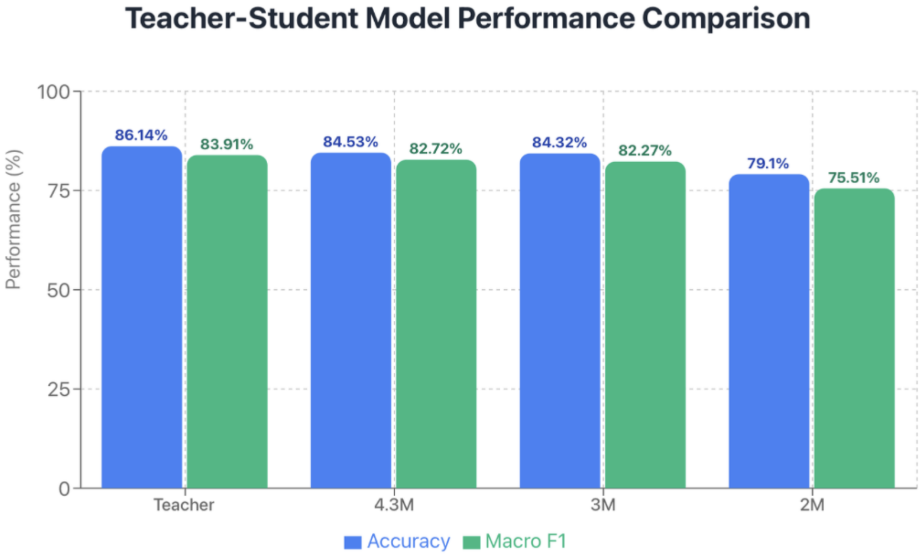


Figure 9: Teacher vs. Student Performance Comparison across model sizes. The 4.3M Student model performs nearly identically to the Teacher, validating the efficiency of the distillation process [2].

4 Discussion

Our reproduction and extension of Geneformer revealed significant insights into both the model's capabilities and the practical challenges of deploying genomic foundation models.

4.1 Critique of the Original Geneformer

Strengths:

- **Data Curation:** The aggregation and rigorous filtering of the 30M-cell corpus is a foundational achievement, providing a robust data foundation for transfer learning [2].
- **Rank Value Encoding:** This non-parametric approach effectively normalizes technical noise and deprioritizes housekeeping genes, forcing the model to focus on biologically relevant signal (transcription factors) [2].
- **Masking Strategy:** The 15% masking objective successfully forces the model to learn network topology, as evidenced by the high accuracy in downstream tasks [2].
- **In Silico Perturbation:** The framework enables therapeutic discovery without wet-lab experimentation, a powerful tool for drug repurposing [2].

Weaknesses:

- **Data Availability:** While the pre-tokenized data is provided, the raw transcriptome data is not easily accessible, limiting validation of the preprocessing steps and potential extension of the research. [2]
- **Input Representation:** Rank value encoding, while robust, discards precise expression magnitude information which might be crucial for subtle phenotype differentiation [2].
- **Data Bias:** The Genecorpus-30M is heavily skewed towards specific organs (e.g., fetal tissue), potentially biasing predictions for underrepresented tissues [2].
- **Compute Requirements:** The original training required massive computational resources (12 V100 GPUs), making it inaccessible for most academic labs to retrain or extend [2].

4.2 Reproducibility Experience

Reproducing the original results presented significant engineering hurdles:

- **Data Scale:** Handling 30M rows of pre-tokenized data required specialized memory optimization. The original dataset utilized ‘int64‘ integers; we converted this to ‘int16‘ format, which drastically reduced memory consumption and prevented RAM overflow during loading [2].
- **Metadata Complexity:** Filtering for specific cell types (e.g., Cardiomyocytes) required complex parsing of heterogeneous metadata across datasets [2].
- **Hardware Constraints:** Training the 10M parameter model was infeasible on consumer hardware, necessitating the development of our custom data collator and the adoption of Knowledge Distillation [2].

4.3 Future Work

To address these limitations and further democratize genomic AI, future research should focus on:

1. **Distillation Hyperparameter Optimization:** We used standard hyper-parameters for distillation, we believe results can be even better if we do hyper parameter optimization.

2. **Bias Mitigation:** Collecting more diverse datasets to balance organ representation in the pretraining corpus would improve generalization to rare tissues.

3. **V2 Dataset:** The authors have mentioned a 104M-row V2 dataset. Applying our distillation techniques to this larger corpus could yield even more powerful lightweight models [2].

5 Conclusion

In this project, we successfully reproduced the Geneformer pipeline and introduced a **Geneformer Distilled** model that democratizes genomic AI. Despite significant hardware constraints, we compressed the model to 4.3M, 3M and 2M parameters, achieving accuracy comparable to the original 10M-parameter Teacher while utilizing only **1/25th of the compute** and **1/20th of the data** [2].

Our analysis reveals a functional trade-off: while the Teacher acts as an "Aggressive Disease Detector" with higher recall (88% for HCM), the Student serves as a "Robust Healthy Screener," maintaining high precision for non-failing hearts (85%) despite slightly reduced granularity in distinguishing specific cardiomyopathies. Ultimately, this work demonstrates that lightweight models can effectively retain the grammar of gene regulation, significantly lowering the barrier to entry for therapeutic discovery in network biology.

References

[1] C. V. Theodoris et al., “Transfer learning enables predictions in network biology,” *Nature*, vol. 618, no. 7965, pp. 616–624, 2023.

[2] K. Agarwalla and A. Saini, “Seminar on (large) language model applications for text and biological data: Transfer learning enables prediction in network biology,” *Presentation Slides, LMU Munich*, 2023.

[3] M. Chaffin et al., “Single-nucleus profiling of human dilated and hypertrophic cardiomyopathy,” *Nature*, vol. 608, pp. 174–180, 2022.

[4] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” *arXiv preprint arXiv:1503.02531*, 2015.

[5] Stanford University, *Cs336: Language modeling from scratch*, `https://stanford-cs336.github.io/spring2025/`, Accessed: 2026-01-14, 2025.