nature

TRANSFER LEARNING ENABLES PREDICTIONS IN NETWORK BIOLOGY

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 & Patrick T. Ellinor

Mertzani Styliani
Voulgari Despoina

ABOUT THE PAPER

Core Idea: Demonstrating and validating **transfer learning** to boost predictive performance

The Model: Geneformer, a neural network built to leverage vast biological data.

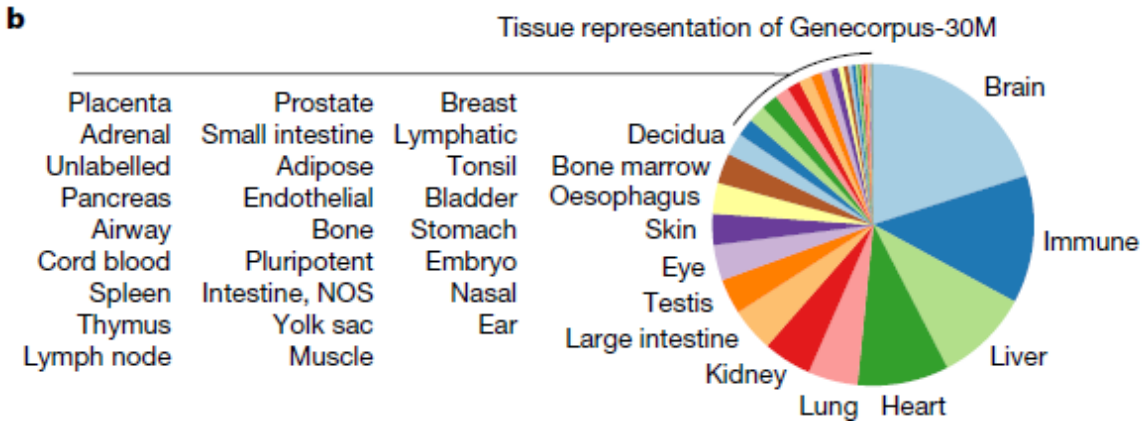
Key Advantage: Achieves **superior prediction accuracy** by fine-tuning a pre-trained model, outperforming models trained from scratch.

Leveraging Prior Knowledge: Pre-trained on a **massive corpus of 30 million single-cell transcriptomes**

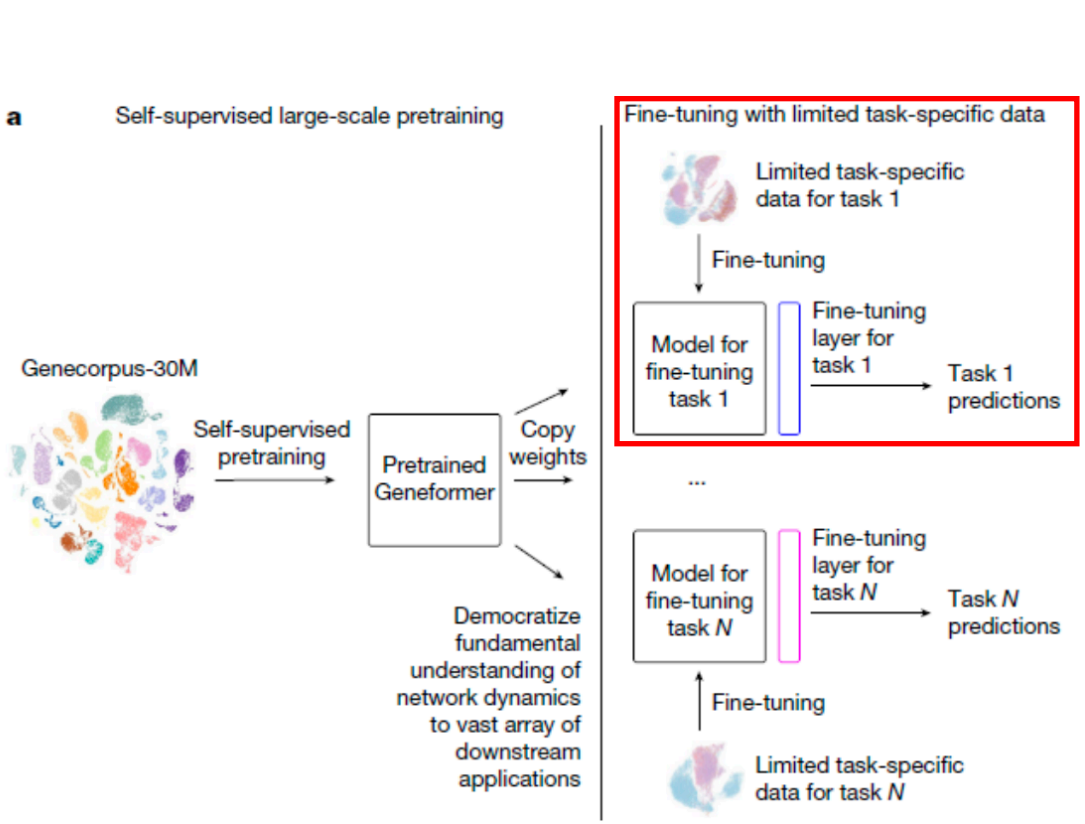
Broad Applications: Applicable across various network biology areas, including:

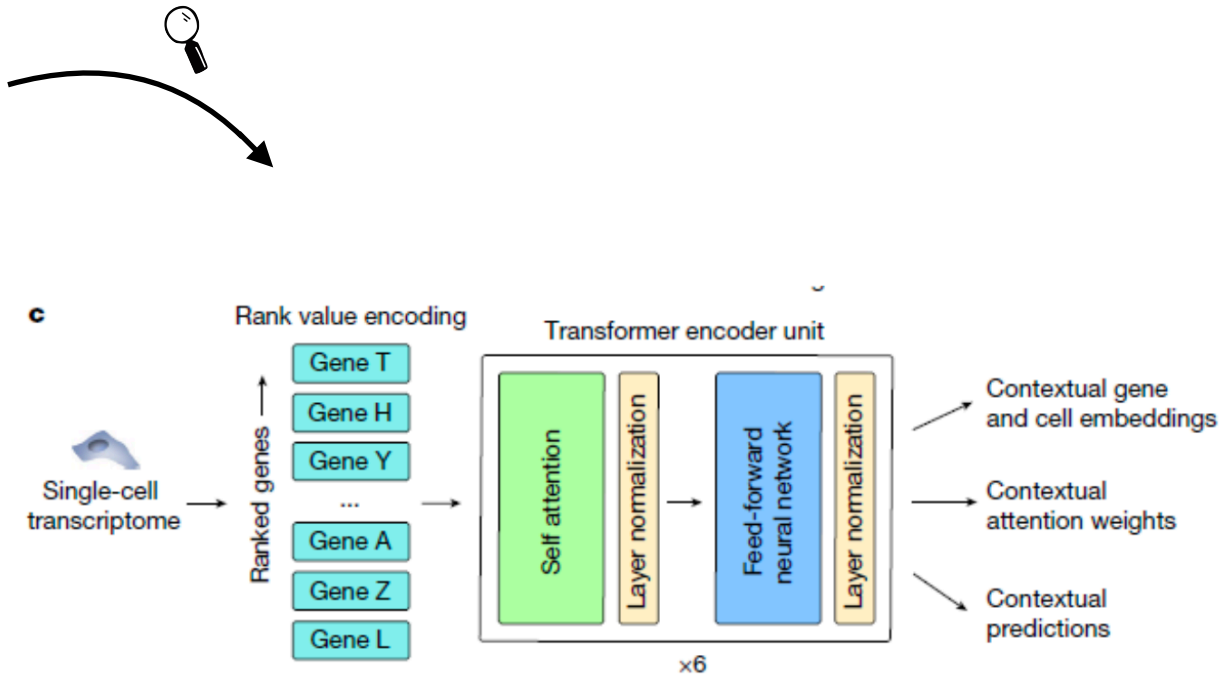
- Protein-protein interaction networks
- Gene regulatory networks
- Predicting gene function
- Identifying disease-associated genes

Biological significance: Discovery of promising therapeutic agents and targets



THE MODEL

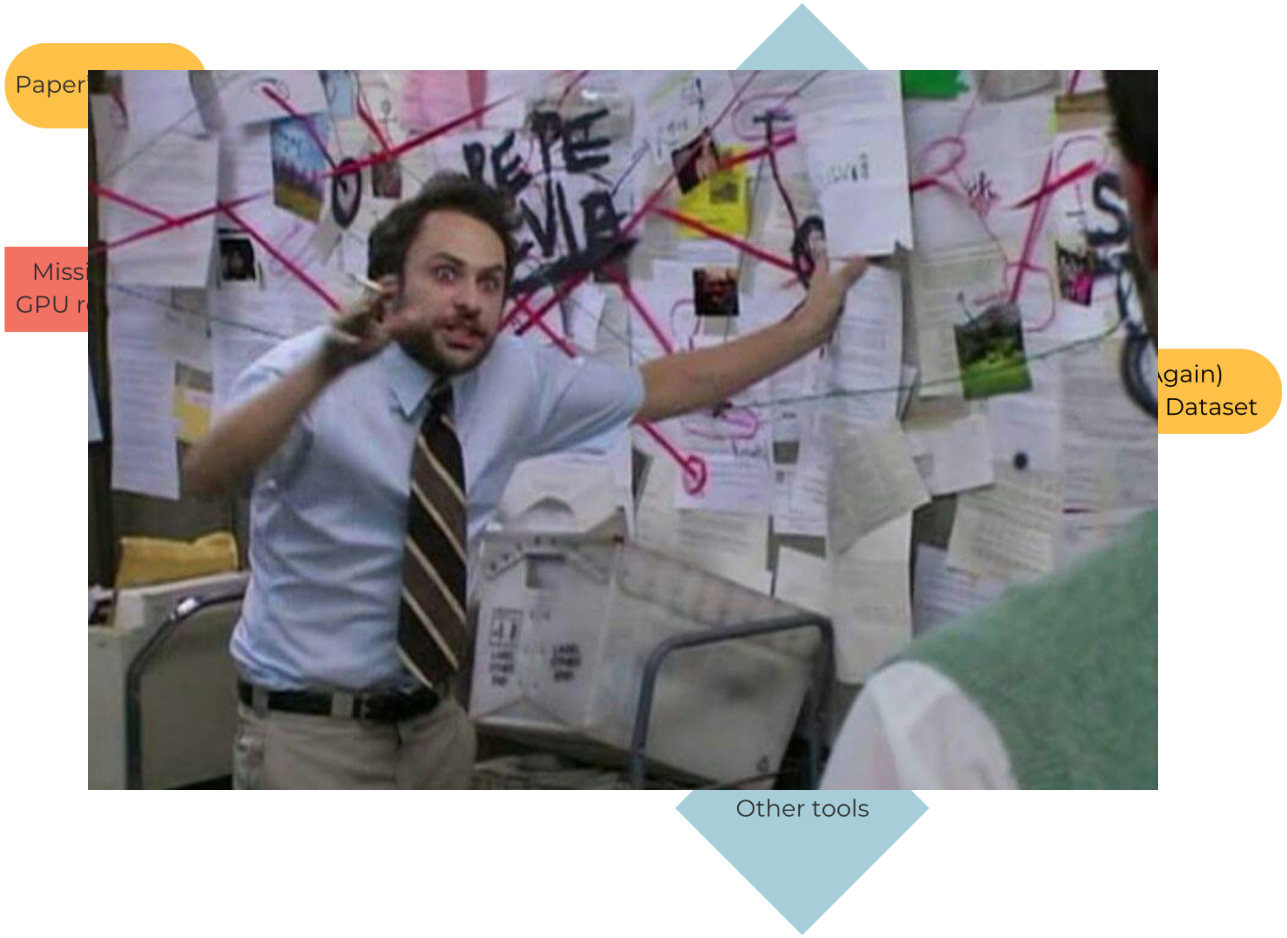

a
Self-supervised large-scale pretraining
Genecorpus-30M
Self-supervised pretraining
Pretrained Geneformer
Copy weights
Democratize fundamental understanding of network dynamics to vast array of downstream applications
Fine-tuning with limited task-specific data
Limited task-specific data for task 1
Fine-tuning
Model for fine-tuning task 1
Fine-tuning layer for task 1
Task 1 predictions
...
Model for fine-tuning task N
Fine-tuning layer for task N
Task N predictions
Fine-tuning
Limited task-specific data for task N


c
Rank value encoding
Transformer encoder unit
Single-cell transcriptome
Ranked genes
Gene T
Gene H
Gene Y
...
Gene A
Gene Z
Gene L
Self attention
Layer normalization
Feed-forward neural network
Layer normalization
×6
Contextual gene and cell embeddings
Contextual attention weights
Contextual predictions

TASK 1

CELL CLASSIFICATION

WORKFLOW



Functional Human Thymus (GSE159745)

- mTECs
- cTECs

FACS-isolation to separate the 2 distinct epithelial cell populations

Processed on the 10x Genomics Chromium Platform

Sequenced using Illumina HiSeq 4000

```
AnnData object with n_obs × n_vars = 11176 × 23937
    obs: 'sample_id', 'n_genes', 'cell_type', 'n_counts'
    var: 'gene_symbols', 'feature_types', 'n_cells', 'ensembl_id'

AnnData object prepared for Geneformer and saved to ../data/human_thymus_for_geneformer.h5ad
```

WORKFLOW (II)

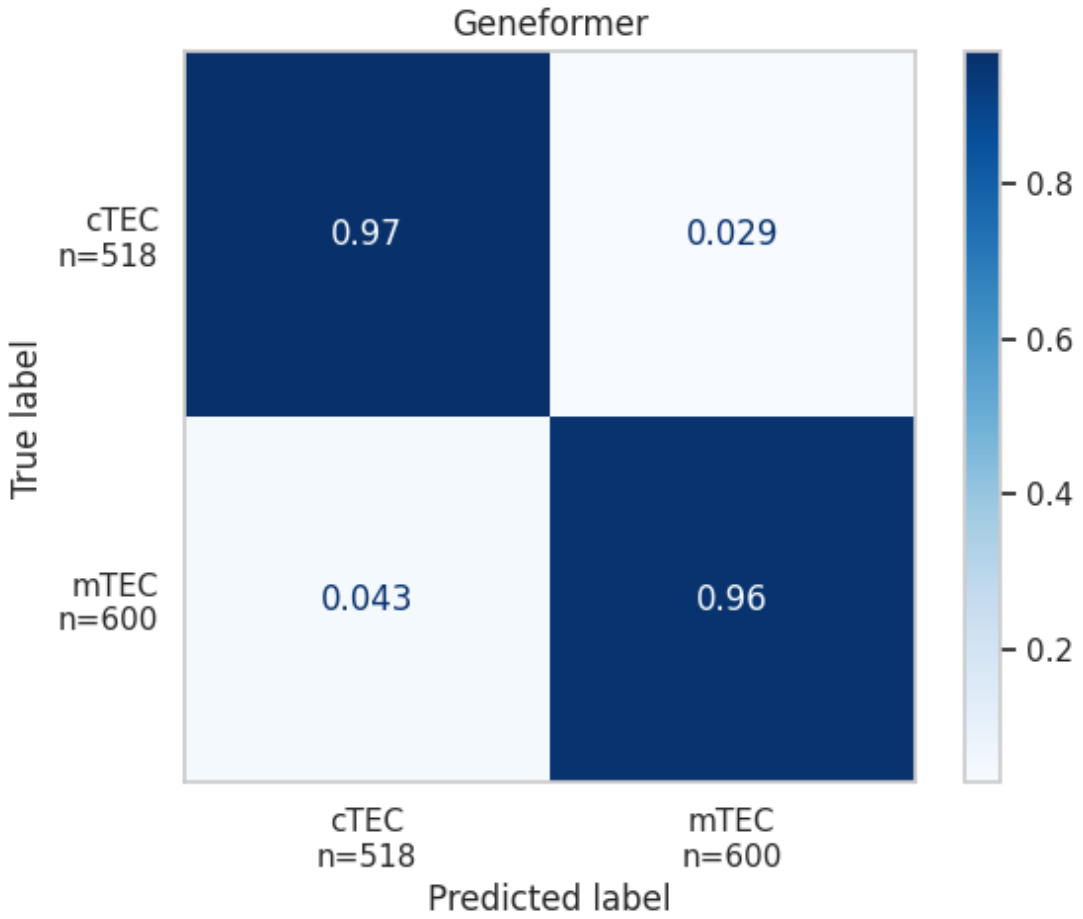
- Tokenization (tk.tokenize_data)
  - 2048 or 4096 models (depending on the model)
- classifier.prepare
  - 70% training, 15% validation, 15% testing
- classifier.validate

- classifier.evaluate_saved_model
- Plots
  - Confusion Matrix
  - Metrics
  - ROC curve
  - UMAP

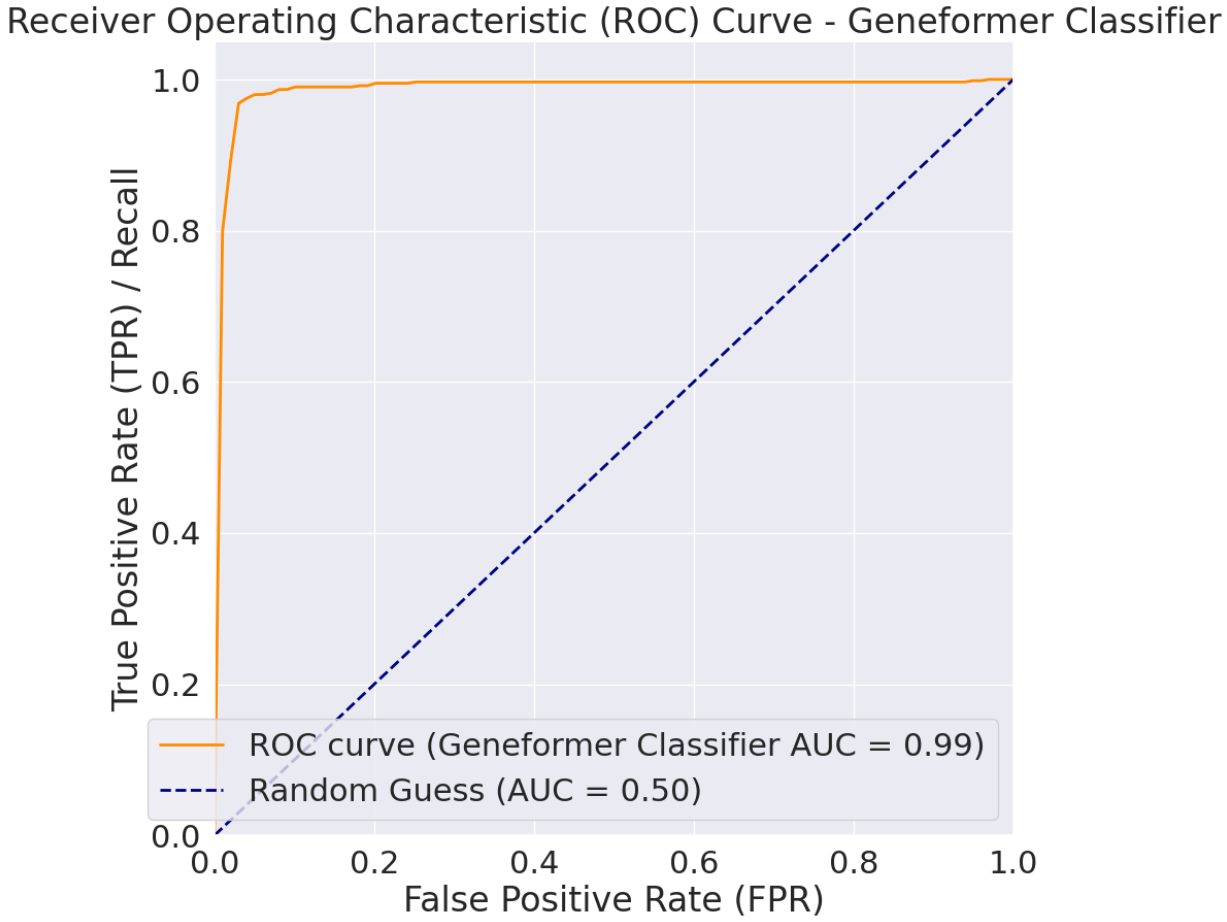
[2236/2236 03:26, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Accuracy	Macro F1
1	0.310500	0.326618	0.899821	0.899551
2	0.155900	0.132791	0.957961	0.957936

RESULTS (I)

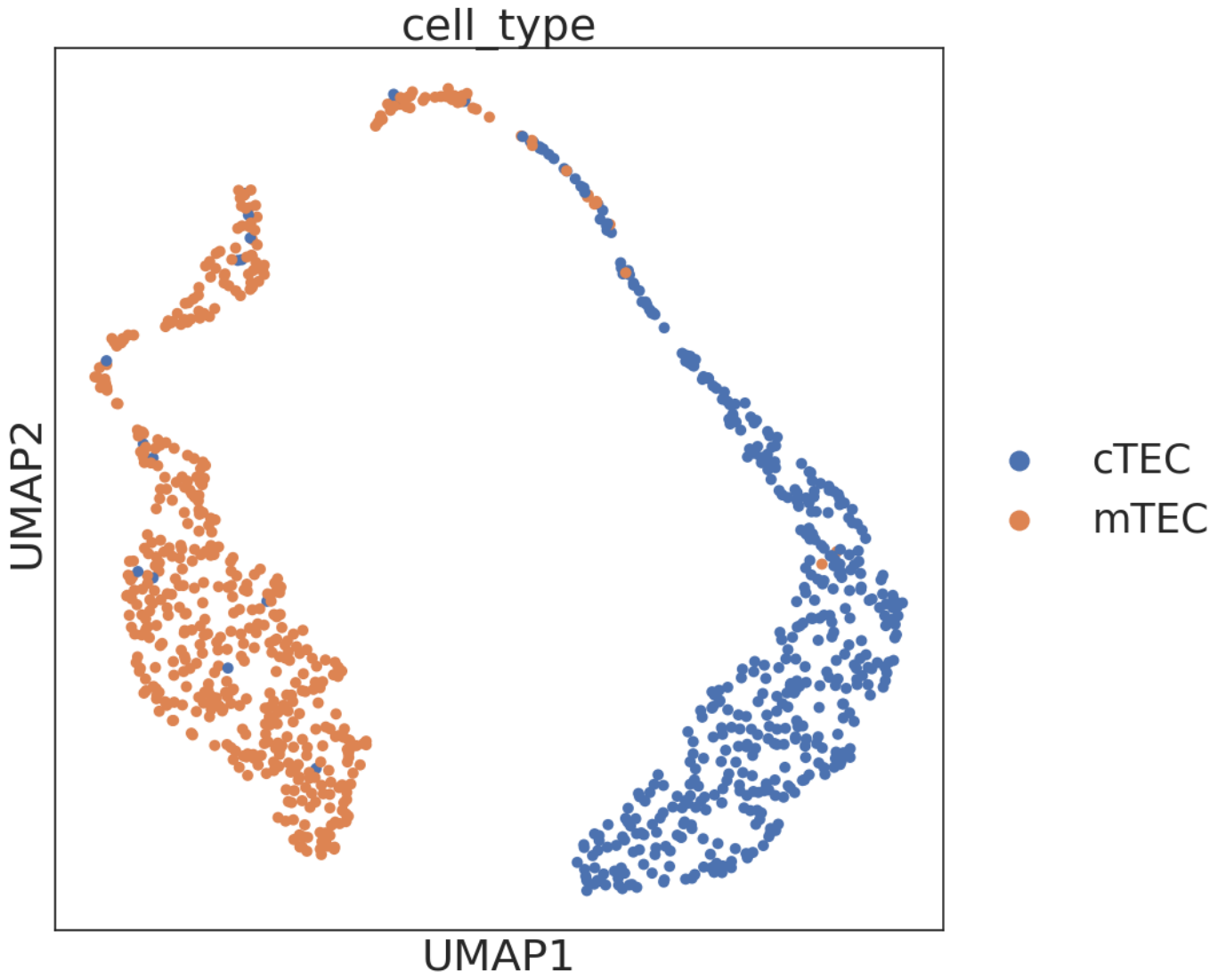



```
{'conf_matrix':       cTEC  mTEC
 cTEC   503    15
 mTEC    26   574,
 'macro_f1': np.float64(0.963178869018213),
 'acc': 0.9633273703041145,
```



RESULTS (II)

- Clear separation between populations
- The thymus has distinct cortical (cTEC-rich) and medullary (mTEC-rich) regions



RESULTS (III)

Original claim: Geneformer outperforms models trained from scratch

Explanation: Well separated clusters?

FINAL RESULTS:

Model	Accuracy
Logistic Regression (baseline)	0.969
Logistic Regression (tuned)	0.969
Random Forest (baseline)	0.976
Random Forest (tuned)	0.972
XGBoost (baseline)	0.983
XGBoost (tuned)	0.984

Conclusion: Geneformer achieved indeed high accuracy score, but maybe it is better for more complex tasks.

TASK 2

GENE CLASSIFICATION

INITIAL GOALS VS. ACHIEVED OUTCOMES

Goals:

- Examine dosage sensitivity
- Run Geneformer
- Run SVM, RF, LR
- Compare Results between Geneformer and other classifiers
- Find best performing classifier for binary classification of gene dosage sensitivity

Status:

- Classified 2033 genes in terms of dosage sensitivity
- Pretrained and finetuned SVM, RF, LR AND GNB, XGBoost
- RF and XGBoost emerged best performing classifiers for binary classification of gene dosage sensitivity

WORKFLOW (I)

Data Transformation

- Gene Expression file -> shape: (26874, 10281)
- Highly Variable Genes list -> 2033 genes
- AnnData object created and filtered down to 2033 HVG
- shape: (2033, 10281) (Genes x Cells)

Dosage Sensitivity

- Created Label Dataframe: df_labels
- Merge with AnnData object
- 0 for dosage sensitive 21
- 1 for dosage insensitive 19
- NaN for genes that had no match in the df_labels 1933

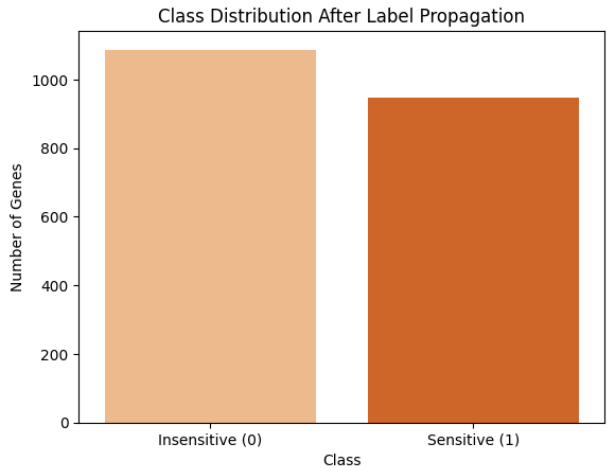
1

2

3

Pretraining & Label Propagation

- Self-Training on gene expression levels
- Train RF classifier on labeled data
- Predict labels iteratively



WORKFLOW (II)

Hyperparameter Tuning

4

- 5-fold Cross Validation
- Train tuned models

5

Model Analysis

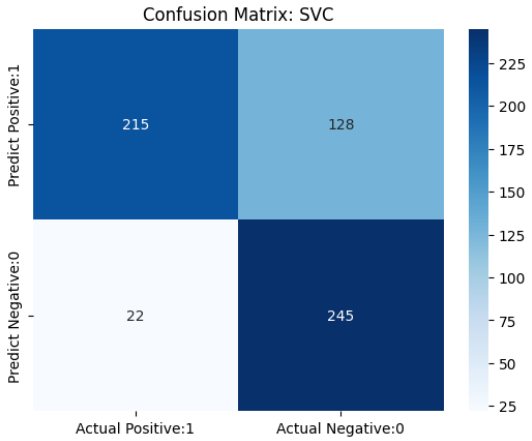
- Test Tuned Models (test set size: 30% of the original)
- Best Performing Classifier -> RandomForest
- Train tuned RandomForest classifier on whole dataset

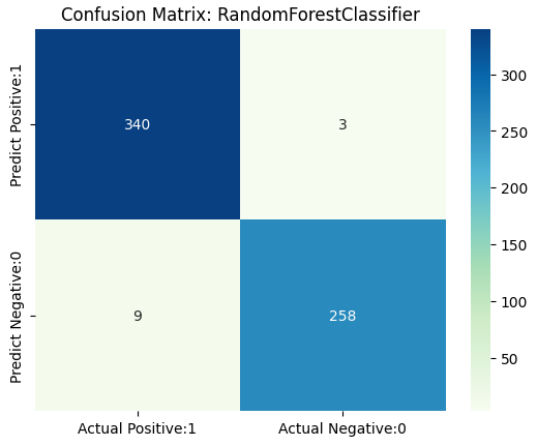
Model	Best AUC	Best Param Combo
Logistic Regression	0.7128	C=0.01, l1_ratio=0.0
Gaussian Naïve Bayes	0.6714	var_smoothing=0.01
Support Vector Machines	0.7654	kernel='rbf', C=1, gamma='scale'
Random Forests	0.9934	n_estimators=200, max_depth=20, min_samples_split=5, min_samples_leaf=1
XGBoost	0.9925	n_estimators=100, max_depth=6, learning_rate=0.1, subsample=1.0, colsample_bytree=0.8

Model	Best AUC
Logistic Regression	0.6869
Gaussian Naïve Bayes	0.6346
Support Vector Machines	0.7722
Random Forests	0.9788
XGBoost	0.9732

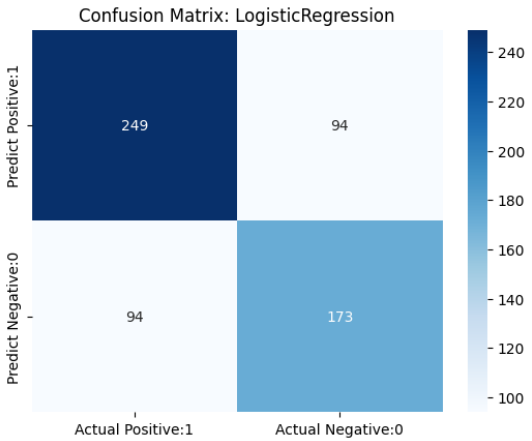
Random Forests	0.9788	(95% CI: 0.9656-0.9896)
XGBoost	0.9732	(95% CI: 0.9582-0.9852)

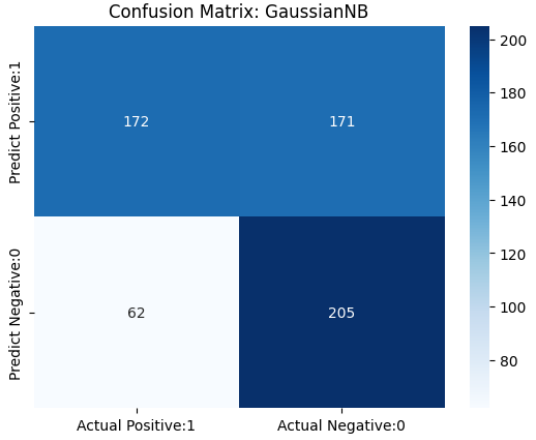
RESULTS

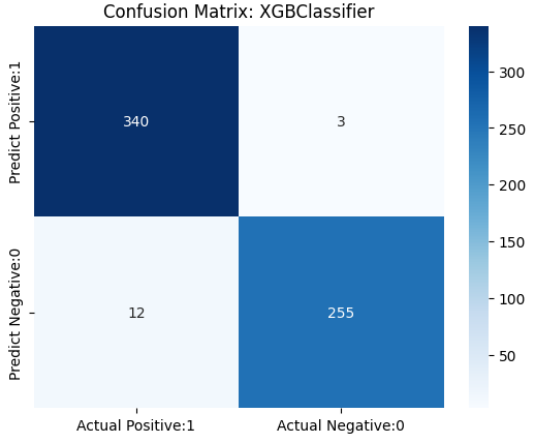


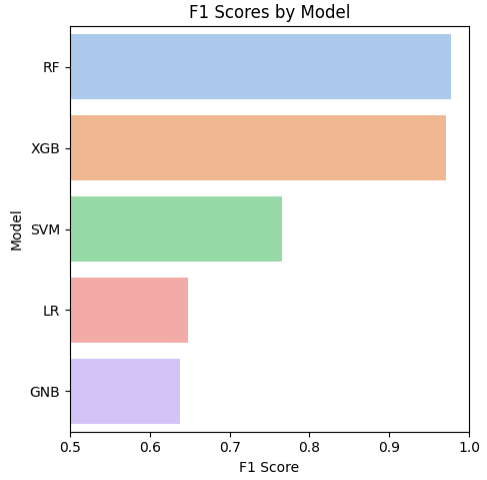


Their approach:









NEXT STEPS

- Run Geneformer with new raw data
- Compute AUC, F1 metrics and produce confusion matrix
- Compare Results between Geneformer and other classifiers

THANK YOU!