CytoRx AI: An Optimized Suite of Single Cell Foundation Model-Based Tools to Aid Drug Target Discovery

Aaryan Gupta*, Aditya Goel*, Amar Yadav*, Harshavarthan PK*, Joy Makwana*, Lavanya Nemani*, Nischal Gupta*, Param Shah*, Pratham Nagpure*, Sayak Dhar*, Shekhar Nath*, Navneet Kumar, Radhakrishna Bettadapura, Ramesh Hariharan, Badri Padhukasahasram;
Strand Life Sciences, Bangalore, India **Contributed equally*

VISIT OUR **WEBSITE**

strand

Contact

Badri Padhukasahasram
VP, Data Science

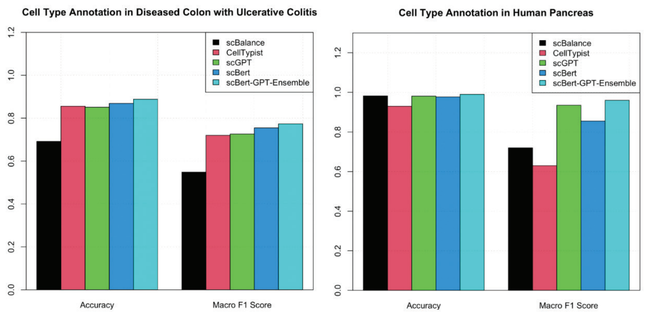
badri.p@strandls.com

Introduction

- Single-cell foundation models, inspired from powerful language models like GPT-4o and chatGPT, excel at learning the complex interactions and co-expression patterns of genes across various tissues.
- Such models are typically pre-trained on massive corpus of unlabelled single-cell transcriptomic data using transformer architecture to enable them to understand language and grammar of gene-expression.
- We introduce CytoRx AI, an AI platform that provides an efficient and optimized suite of tools leveraging recently published single-cell foundation models to aid drug-target discovery.
- We compared performance with conventional methods for common single-cell tasks.

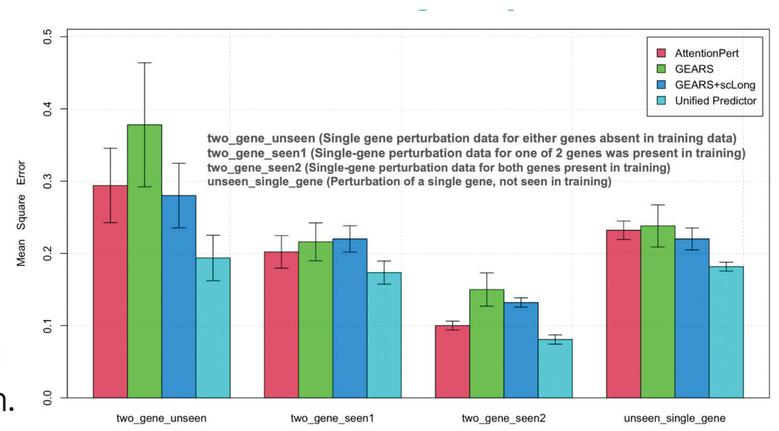
Cell-Type Annotation

Cell-type annotation is a fundamental task in single-cell analysis and information at individual cell-type level provides a more comprehensive picture of biological processes, disease and drug response. Thus, it can enable us to derive more detailed insights for target discovery via differential gene-expression analysis. Conventional approaches typically require dimensionality reduction which leads to loss of information. In contrast, single-cell foundation models directly take in gene-expression in an unbiased manner. We fine-tuned 2 single-cell foundation models scBERT and scGPT for cell-type annotation on a high quality curated dataset of Ulcerative Colitis (Smillie et al. 2019) as well as a well-studied pancreas dataset (Chen et al. 2023). Comparison with conventional ML approaches demonstrated improved cell-type classification accuracy. Ensemble models combining probabilities from multiple fine-tuned foundation models further enhanced accuracy.These methods can also be adapted for discovery of unknown cell types.



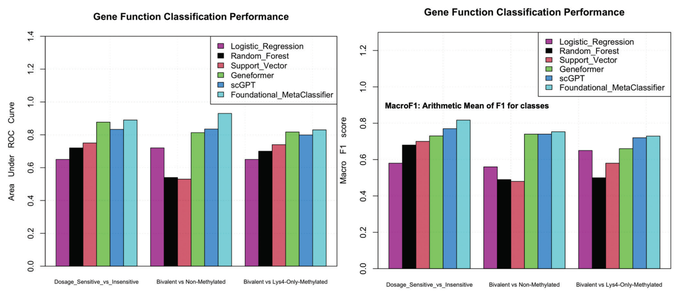
Accuracy: Fraction of cell labels that match ground truth Macro F1: Arithmetic mean of F1 scores for cell types

Predict Effects of Two-Gene Perturbations



Prediction of final gene-expression for two-gene perturbations can be improved through use of pre-trained single-cell foundation models. Accuracy is measured as mean square error between predicted and observed gene-expression for a given perturbation. Unified predictor uses new gene-embedding choices in AttentionPert as well as stacks predictions from individual methods (GEARS with gene2vec embedding, AttentionPERT with scGPT embedding and scLong with GEARS decoder) via xgboost to surpass previous best results for the Norman et al 2019 data. Accuracy also shown for unseen single gene perturbations. Such analyses can guide prioritization of targets for combination therapy for diseases like cancer.

Gene-Function Prediction



Both foundational models (Geneformer, scGPT) as well as a meta classifier stacking scGPT and Geneformer probabilities with an optimized classification threshold demonstrated improved performance over baseline ML methods for gene-function classification tasks to distinguish i) Dosage Sensitive vs Insensitive genes ii) Bivalent (histone modifications H3K4me3 and H3K27me3) vs Non-Methylated genes iii) Bivalent genes vs Lys4-only-Methylated (histone modification H3K4) genes.

Multimodal Integration

Foundational models can perform multi-modal integration of scRNA-seq and scATAC-seq data and enable downstream applications based on joint embedding such as cell clustering and trajectory inference. A foundation model like approach scMVP was compared with a conventional method Seurat v5 for multimodal integration and subsequent cell type annotation. Ensemble approach combining Seurat and scMVP demonstrated improvements over a conventional method.

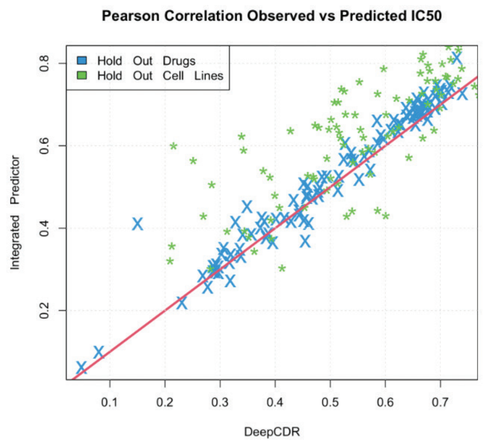
Dataset	Method	#Cells	Adjusted Rand Index	Normalized Mutual Information
10x PBMC	scMVP	11,909	0.717	0.642
10x PBMC	scMVP + Seurat V5 ensemble model	11,909	0.934	0.796
10x PBMC	Seurat V5	11,909	0.764	0.669
10x Lymph Node	scMVP	14,645	0.433	0.367
10x Lymph Node	Seurat V5	14,645	0.405	0.351

Summary & Achievements

- Single-cell foundational models are versatile tools that demonstrated impressive performance for diverse analyses tasks.
- Unified approaches blending multiple model outputs through stacked generalization provided further improvements and outperformed previous methods for cell-type annotation (~7%,33% improvements in macroF1), perturbation prediction (~10-20% decrease in mean square error), gene-function classification (~15% improvement in macroF1) and cancer drug response prediction (~15% improvement in Pearson Correlation) tasks.
- Foundational models either alone or as part of ensembles can comprehensively capture complex gene-gene relationships and lead to more accurate approaches for target discovery.

Cancer Drug Response Prediction

An integrated predictive model that combines embeddings from a cancer-specific foundation model (CancerFoundation), a large single-cell foundation model (scLong) and drug chemical structure was trained to predict a drug response metric (IC50) for 104 cancer drugs across 884 cell lines. New predictor outperformed previous published method DeepCDR. Such predictive models can eventually enable selection of drugs that are most likely to provide response for a given patient based on their basal gene-expression patterns. Our tool also enables computation of attention scores to help identify most important genes mediating drug response and guide new target discovery.



Top 10 genes by attention scores

Feature	Importance
RGS16	0.7445788575
SLC6A2	0.7436331565
FANCM	0.743502708
GFOD1	0.7426553698
TBC1D22B	0.7426492884
LACRT	0.7418029877
SRD5A2	0.7396675576
SCAND2P	0.739620516
UGP2	0.739456025
PRKRIP1	0.7391781338

Computational Resources and RunTimes

Task	Running Time	Hardware and memory
Cell-type Annotation scBERT fine-tune Cell-type Annotation scGPT fine-tune	1100 min 240 min	4× NVIDIA RTX 3090 (96GB VRAM) 4x NVIDIA RTX 4090 (24GB VRAM)
Perturbation prediction training GEARS + gene2vec initialization scLong + GEARS decoder AttentionPERT + scGPT embedding initialization	12 hr 36 min 17 hr 3 min 13 hr 50 min	NVIDIA RTX A4000 NVIDIA GeForce RTX 4070 Ti NVIDIA GeForce RTX 4070 Ti
Gene-function classification fine-tune scGPT 35000 samples, 1200 genes Geneformer 35000 samples, 1200 genes	210 min 50 min	1x NVIDIA RTX 3090, 10 Gb RAM 1x NVIDIA T4 GPU, 5 Gb RAM
Cancer Drug Response Prediction Holdout cell line prediction training Holdout drug prediction training	30 min 24 hrs	1x NVIDIA TX4090 with 24Gb RAM 1x NVIDIA TX4090 with 24Gb RAM
Multi-omic Integration with scMVP 10x PBMC scRNA-seq + scATAC-seq 10x LymphNode scRNA-seq + scATAC-seq	54 min 4 min	15 Gb RAM, 12th gen Intel i5-1235U CPU 11 Gb RAM, 12th gen Intel i5-1235U CPU