scDrugMap: Benchmarking Large Foundation Models for Drug Response Prediction

† The authors contributed equally to this work.
*Correspondence: qsong1@ufl.edu

Qing Wang[1,†], Yining Pan[1,†], Minghao Zhou[1,†],Zijia Tang[2], Yanfei Wang[1] ,,Guangyu Wang[3] ,Qianqian Song[1,*]

Department of Health Outcomes and Biomedical Informatics, University of Florida[1], Trinity College, Duke University[2], Center for Bioinformatics and Computational Biology, Houston Methodist Research Institute[3]

Code: https://github.com/QSong-github/scDrugMap
Web: https://scdrugmap.com

Introduction

Drug resistance remains a major challenge in cancer treatment. While single-cell profiling offers unprecedented resolution for uncovercovering resistance mechanisms, the potential of emerging foundation models for drug response prediction at the single-cell level is still largely unknown. Here, we introduce scDrugMap, a unified framework featuring both Python toolkits and an interactive web server for benchmarking and predicting drug responses with state-of-the-art foundation models. scDrugMap evaluates eight single-cell foundation models and two large language models across 495,000 cells from 60 datasets, spanning diverse tissues, drugs, cancer types, and treatment conditions. In pooled-data evaluation, scFoundation delivered the strongest performance, particularly in tumor tissue. In cross-data analysis, UCE performed best after fine-tuning, while in zero-shot settings, scGPT achieved the highest accuracy. Together, scDrugMap provides the first systematic benchmark of foundation models for single-cell drug response prediction and offers a powerful, user-friendly platform to accelerate drug discovery and translational precision oncology.

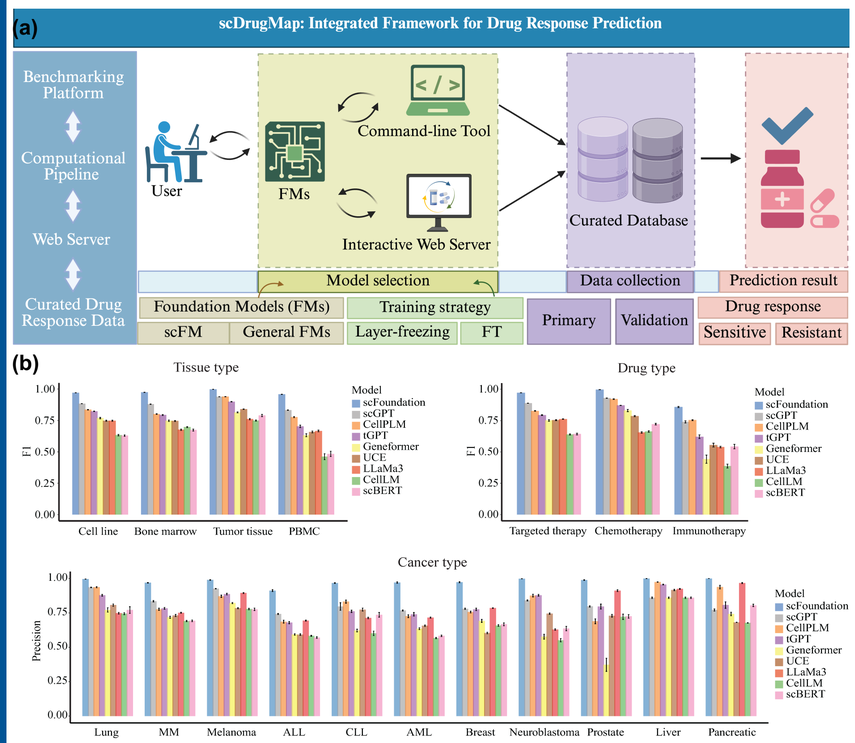
Methods

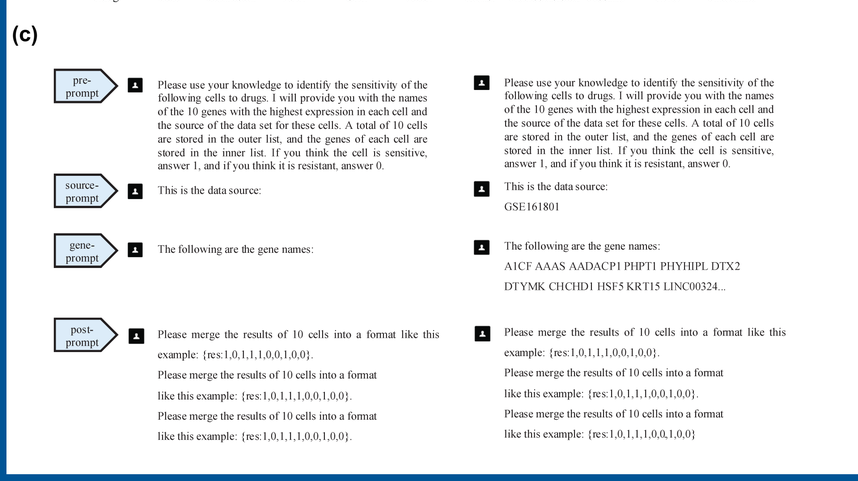
The study systematically benchmarks ten foundation models using two evaluation settings—pooled-data and cross-data—and two training strategies: layer-freezing and LoRA-based fine-tuning. Gene expression profiles are tokenized according to each model's input format, embeddings are extracted or partially fine-tuned, and a unified MLP classifier is trained to predict binary drug response. All models are evaluated with consistent 10-fold cross-validation, standardized preprocessing, and multiple metrics (F1, AUROC, precision, recall), enabling fair and comprehensive comparison across architectures.

Data collection

The authors curated a large single-cell drug response corpus comprising 36 primary datasets (326,751 cells) and 24 external validation datasets (168,486 cells), totaling over 495,000 cells with cell-level sensitive/resistant labels. The datasets span 14 cancer types, 3 therapy categories, multiple tissue sources (tumor, PBMC, bone marrow, cell lines, organoids), and diverse sequencing platforms. Uniform quality control, annotation harmonization, and batch-correction protocols were applied to ensure consistency and robust model evaluation.

Results





(a) Schematic of the scDrugMap framework. scDrugMap integrates a benchmarking platform, computational analysis pipeline, interactive web server, and curated single-cell drug response datasets. Users can choose from multiple foundation models—including single-cell–specific models and general-purpose language models—and apply either layer-freezing or fine-tuning strategies to predict drug sensitivity or resistance.

(b) Model performance in predicting drug response using primary single-cell data. F1 scores across tissue, drug, and cancer types under the layer-freezing training setting. Bars show mean ± s.e.m. from 10-fold cross-validation; dots indicate individual fold results. Error bars represent the standard deviation of F1 scores across folds. Tissue abbreviations: PBMC, peripheral blood mononuclear cells; MM, multiple myeloma; ALL, acute lymphoblastic leukemia; CLL, chronic lymphocytic leukemia; AML, acute myeloid leukemia.

(c) Prompt display used for GPT4o-mini. For the prompt word template, we first use a technique like the thought chain to prompt the model how it should think about the output, then tell the model the data source and sequence information, and finally we repeat telling the model and give an output template to ensure the consistent of the output format. A complete input example with prompt is also showed.

(d) Model performance in cross-data evaluation using primary single-cell data. F1 scores across tissue, drug, and cancer types using layer-freezing and fine-tuning. Radar plots show mean F1 values (0–1 scale) for tissue, drug type, and regimen based on 10-fold cross-validation. Circular bar charts summa

-rize the mean F1 score of each category. Box/violin plots display the distribution of F1 scores for each cancer type, with boxplots indicating median, quartiles, and whiskers (1.5× IQR). All replicates correspond to 10-fold technical replicates. Abbreviations: MM, multiple myeloma; ALL, acute lymphoblastic leukemia; CLL, chronic lymphocytic leukemia.

Conclusions

1) scDrugMap benchmarks ten foundation models for single-cell drug response prediction using large, diverse datasets spanning many cancer types and treatments.
2) Foundation models outperform traditional methods, but fine-tuning is essential and cross-study generalization remains limited, especially under class imbalance.
3) Future progress requires better robustness, multimodal integration, and improved interpretability to enable real-world clinical application.

Acknowledgements: Q.S. is supported by the National Institute of General Medical Sciences of the National Institutes of Health (R35GM151089).

Contact info: wang.qing@ufl.edu/ qsong1@ufl.edu