nature portfolio

Peer Review File

CellFM: a large-scale foundation model pre-trained on transcriptomics of 100 million human cells

Corresponding Author: Professor Yuedong Yang

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Review - Major revisions

The manuscript titled “CellFM: a large-scale foundation model pre-trained on transcriptomics of 100 million human cells” submitted by Zeng et al. is a promising take on the rapidly emerging field of large language models (LLMs)/foundational models in single-cell analysis since it surpasses existing models in terms of size while providing interesting nuances through it is architecture. Progressing this sub-field of single-cell computational biology is paramount to harnessing the immense capabilities of foundational models, which can provide unparalleled analytical solutions compared to traditional machine learning and deep learning approaches after being trained on many datasets. I believe that this work exhibits apparent scientific novelty, but there are critical issues that should be addressed to adhere to the high standards of the esteemed Nature Communications journal.

An initial issue the authors should address is significantly improving the manuscript and its context in writing and phrasing. I advise the authors to consider the excellent paper by Fabian Theis regarding the foundational models that are taking the field by storm (https://doi.org/10.1038/s41592-024-02353-z) and provide a bit more context in the Introduction or the Discussion sections regarding CellFM’s position among the various other foundational models. For example, I suspect that CellFM can be construed as a model belonging to the third category of transformers (value projection) and not in the other two (ordering, value categorisation).
My following criticism relates to reshaping the perturbation modelling section, where some key aspects should be addressed:
please provide some references for the first introductory paragraph of this section, some definitions of perturbation modelling, and some insight as to why this is probably the most translationally significant type of single-cell analysis since it connects to computational pharmacology and AI-driven drug discovery or repurposing. To aid the authors, I would encourage them to provide this information as a supplement if they feel that the flow of the text in the Result section will be compromised
explain why you are focusing on the MSE as a metric to be minimised and not on some other metric (some references could be useful here) during the MLP stage
I believe that the comparative study with scGPT is narrow. It should involve at least one or two more foundational models (scElmo, scFoundation or GeneCompass indicatively) as well as 1 or 2 non-foundational models that I assume would derive from the group of recent generative AI models predicated on autoencoders (i.e., scGEN-inspired models like scVIDR or CellOT). This is something that the authors should heed because in several publications (https://doi.org/10.1038/s41592-024-02353-z, https://doi.org/10.1016/j.csbj.2024.04.058) there are concerns that the fledgling foundational models are exhibiting questionable performance compared to simple machine-learning or deep-learning models.
In Figure 3A, although it is not clearly described in the main text, based on the legend description, comparisons are made regarding how well the models can generalise to unseen events by comparing predicted gene expressions to ground-truths. Is this accurate? In that case maybe the authors could try to calculate the R2 (square) metric also which describes how much of the variance in actual gene expression can be explained by the predicted values
I am not sure what is happening in Supl. Figure 1. Is CellFM trained on the existing KOs of the Norman dataset and extrapolates to other perturbations? Does it perform digital KOs based on the prior knowledge? Are the genes depicted on each cluster represent the target of the simulated KO?
In silico reverse perturbation prediction: is this a term the authors are coining or is it something from the literature?

References are missing if this is the latter. This represent a quite ambitions approach of perturbation modelling which I think has immense translation impact. I would advise the authors to provide some additional insights from other datasets (e.g., sciPlex with high-throughput screening, control vs disease scRNA-seq data). If CellFM can reverse-engineer its way to original perturbation events that are causing the cellular phenotype in question, this should be researched further.
The authors are also using GEARS as part of this small comparison which comes completely out of the blue in the flow of the text, without any background information especially for the reader that is not an expert in single-cell perturbation biology. I would surmise that the authors are using GEARS because it shares some resemblences with the CellFM in terms of using leveraging some prior knowledge (a knowledge graph in the case of GEARS, the voluminous dataset the CellFM is trained upon) and an MLP component. In any case, background information should be provided for the interested reader to delineate this part of the analysis. Furthermore, it is a further attestation to my previous claim that simpler, non-foundational models should be included in the previous perturbation analysis
Hybrid approaches of foundational and non-foundational models are also being explored in single-cell perturbation modelling. For example, scELMo can provide cell and feature embeddings in task-specific models (e.g., causal factor analysis in CINEMA-OT, gene expression predictions for cell-level and gene-level perturbations with CPA and GEARS, respectively) to simulate the effects of perturbations, outperforming the original iterations of these models. I would advise the authors to explore such a combination if deemed feasible between CellFM and some other tool.
With regards to deciphering gene relationships with CellFM, I would encourage the authors to provide some more examples apart from SPI1 regarding the emerging explainability of the attention maps in terms of the genes that are being affected a given perturbation in the Adamson dataset. There have been some sceptisism regarding whether attention can reflect actual biological dynamics ("Attention scores and feature importance have been shown not always to correlate and are hence not necessarily reliable tools for identifyin which input elements are responsible for an output") (https://doi.org/10.1038/s41592-024-02353-z).
Finally, regarding the Github page and the code provided, no script is provided for the perturbation analyses. Furthermore, the scripts from the tutorials provided by the authors are provided without comments or some context, making the results' reproducibility extremely problematic. In the spirit of Open Science and FAIRness, the scripts should be rewritten and the missing code should be provided.

(Remarks on code availability)
The code provided is partial and does not allow for full reproducibility of the results. Moreover, the Jupyter notebooks lack any clear structure and have almost no comments to guide the AI practitioner in recreating results or experimenting with the pipeline. A characteristic example is the perturbation notebook that is simply absent! https://github.com/biomed-AI/CellFM/blob/main/tutorials/GenePertubation.ipynb

Reviewer #2

(Remarks to the Author)
Zeng et al. proposed CellFM as a novel large-scale foundation model and demonstrated it outperforms other large language model on multiple single-cell data analysis tasks. This study well summarized the progresses in this rapid-growing field and compared their differences and limits in many aspects including data preprocessing and model architecture. The codes in this study are also well-documented. However, the methodological improvement is unclear and lacks strong benchmarking evidence. Additionally, many figures are in low-quality and not well prepared, and the biological interpretation of the results are not well-justified.

Major:
1. The author is aware that there are different approaches of applying large language models to single-cell transcriptome, and there are already many existing tools. They may need to include more of these existing models into benchmarking. As the author argued that the rank-based language model discards the expression values, it would be useful to show such downsides in one of benchmark results. Therefore, the author may include more large language modes in benchmark and provide some extra insights about application of these models.
2. One valuable effort the author made in this study is reprocessing and unifying numerous single-cell datasets from multiple sources. However, these re-unified datasets are not available. It is impossible to verify that the author indeed collects this many datasets to train CellFM. Meanwhile, it could be more informative if the author provides more comprehensive summary on their data collection, instead of a mere summary on technique types and organs. For example, does this data collection have overwhelming immune cell population? Were majority of cells from healthy donors? Etc. Such information can be very useful for anyone who has intention to apply this pre-trained model.
3. The task of gene function prediction seems to be too simple. The task was set in a confined scenario. It is either or not. Indeed, CellFM may outperform other models in binary classification, but it will not say that CellFM will also be accurate in multi-category classification and indicate that CellFM accurately predicts gene function in real world. In real-world application, the model would be asked to predict potential gene functions without knowing a certain category first. Therefore, benchmark here should be framed as a problem of multi-category classification.
4. The author may consider including some baseline methods of cell type prediction into results of Figure 4 (See references https://genomebiology.biomedcentral.com/articles/10.1186/s13059-021-02480-2 and https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1795-z). Meanwhile, scBERT has demonstrated outstanding performance in cell type prediction. Does CellFM also achieve superior or at least comparable performance to scBERT in this task?
5. In Figure 3d, the author only showed the top1 perturbation prediction. First, it is not clear if the author did proper fine tuning of scGPT with this perturbation data. Second, top1 perturbation prediction seems to be stringent to model. It could be possible that ground truth perturbation is in top 3 or 5 prediction in scGPT. Third, if the author set different random seeds for

fine tuning, can CellFM always get the top1 prediction correct?
6. Prediction result of GEARS in Figure 3e looks so odd. The hit rate remains at 0.1 without increasing as the number of his goes up. It seems that GEARS just do some random prediction here. It leaves that doubt if the author has properly used GEARS in this case.
7. For the task of cell type annotation, I am skeptical on the re-evaluation of scGPT in human pancreas data. The annotation for this data is so accurate in the original report (https://www.nature.com/articles/s41592-024-02201-0). However, the re-evaluation of scGPT in the same dataset is so much worse. Could this be that the author didn't really use scGPT properly? The same question can go to if the poor performance of Geneformer is due to unproper use by the author. Besides only showing annotation out of from zero-shot learning, they will also benchmark cell annotation after model fine tuning.
8. Another key benchmarking task that is missed in current manuscript is batch correction. How well the cell embedding of CellFM addresses batch effects is not known.
9. The author presented the gene regulatory network among interleukin family with pre-trained model in Figure 5a. It indeed looks like a solid result with background knowledge in immunology. However, to further validate this finding, the author will have to fine-tune CellFM with extra datasets, say immune cells and other non-immune cells. Will this gene relationship structure maintain in immune cells after fine tuning and differs in non-immune cells?
10. The author citied study from Zou et al. (https://academic.oup.com/nar/article/50/W1/W175/6553688) to back up their finding on top 20 affected genes by SPI1 suppression. Zou et al. did not specifically provide results on SPI1 suppression. It is not clear how the author drew the conclusion from Zou et al. and tried to back up the identification by CellFM. They will need to have clarification in results presented in Figure 5d and e. Meanwhile, it is confusing why SPI1 in column has connectivity while the SPI1 in row doesn't in Figure 5d. Additionally, as suggested in study on SPI1 (PMID: 35871293), SPI1 represses ALAS2. Is SPI1 repressing those top 20 genes or is it possible to distinguish repression and activation based on the attention map?
11. It's unclear that whether the attention map in Figure 5c was calculated in the same way as described in the scGPT paper (https://www.nature.com/articles/s41592-024-02201-0). Compared to Figure 6b and 6c in scGPT's paper, why are the values of most of the gene pairs near zero in Figure 5d heatmap? Some genes, e.g., CCNB2, CCNA2, TOP2A, MKI67, KPNA2 and CENPF have highly relevant biological functions.
12. In Figure 5f, the KEGG pathway analysis for the top 20 most influenced genes doesn't make sense, most of the enriched pathways only have 2 genes, what's the background gene set used and what's the biological interpretation?
13. For data preprocessing, the existing single-cell foundation models use either raw count or just the rank of the gene. But in this study, the author did log1p normalization. What's the influence of log1p normalization on the model performance? After log1p normalization, the variance of genes is correlated with the mean values of genes and is likely influenced by the batch effects. As in CellFM, the initial representation of each cell is the l_max genes with largest expression values, will log1p normalization induce any bias? Could other normalization methods, e.g., scTransform (PMID: 31870423) and Pearson residuals (PMID: 34488842), which correct the variance-mean bias, improve the performance of single-cell foundation model? The author should give better justification on this step of data preprocessing.
14. In implementation detail, the author used 100 million cells and did pre-training over 2 epochs. What's the reasoning on training a model for over 2 epochs? The author needs to provide more quantitative justification on model training. In current manuscript, it is unknown if model was underfit or overfit within 2 epochs, considering the training size is enormous.
15. There are also lack of some clarity in method section. This study lacks clarity and strong benchmarking evidence regarding which parts of the model mainly contributed to its improved performance, especially in relation to RetNet. Additionally, it's unclear whether the inclusion of the second term in the loss function, L_cls, improved the performance or not. Regarding computational efficiency, although the use of MHA improved the computational complexity from O((l_max)^2 * d) to O(l_max * d^2), given that l_max is 2048 and d is 1536, the improvement is minimal.
16. Finally, is it appropriate to not include long-noncoding RNAs in the scRNA-seq data? Since the function of most of the long-noncoding RNAs are unknown, it would be helpful if the single-cell foundation model could prioritize and predict the biological functions of long-noncoding RNAs in a cell type-specific manner?

Minor:
1. Please be aware of a proper format for writing gene names and protein names.
2. In Figure 2c, "CellAI" should be "CellFM", and it's more appropriate to order the three methods using the order in Figure 2a and Figure 2b.
3. In Figure 3a and 3b, the y-axis range should start at 0 to objectively show the differences.
4. In Figure 3d and Figure S2, the gene name "UBASH3A" is not fully displayed.
5. In both Figure 4a and Figure 4b, the test dataset name "Pancrm" is defined in Table S4, but not in the main text when this figure is mentioned.
6. Figure 5e appears in low resolution due to snapshotting, the borders are not cleanly removed, and the font type is different from that used in other figures.
7. In Figure S4 and Figure S5, the river plot and the confusion matrix does not match with each other. For example, in Figure S4a, the predicted cell types are CD8+ T cells, Erythroid progenitors and Erythrocytes, but the labels are not the same in the Figure S4b. This error also exists for Figure 4c and 4d (scGPT-related plots).

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)
The codes are well documented.

Reviewer #4

(Remarks to the Author)
Overview:

In this paper, Zeng et al. introduce a new large foundation model for single cell RNA-sequencing data called CellFM, which is trained on 100 million human cells. The authors argue that existing foundation models for single-cell may have limited performance because of their training on multiple species. They train a large model on human only data, and show that it outperforms existing methods like scGPT and Geneformer on a set of benchmarks. On a few benchmarks the improvement over scGPT and Geneformer is noteworthy, however, the authors do not present a compelling argument or validation that their design choices caused this performance increase. Specifically, the authors do not validate their claim that single species training is superior to multi species training, do not evaluate model performance as a function of number of parameters or size of training dataset, and do not adequately compare the model to other existing models. Additionally, the model architecture, which essentially performs masked gene reconstruction, is not novel, and the authors do not present any novel capabilities of their model.

Major Comments:

1. Confounding Effects of Multi-Species Training: The authors make the claim that: “Single-cell Large Language Models (LLMs) have been crafted to bridge this gap yet exhibit limited performance on human cells. This shortfall may stem from the confounding effects of training data from diverse species, partly because of limited cells for the single species” (Abstract). However, they do not explain what these confounding effects are and fail to compare their model to others trained on multiple species. While the collection of a large human dataset is an important contribution, this dataset could also be utilized in training multi-species models. Furthermore, human data already comprises the majority of training data in existing multi-species methods like UCE. Given these points, the authors’ claim is counterintuitive and requires stronger evidence.

2. Scaling and Ablation Experiments: A main contribution of this work is the collection of a large number of human scRNA-seq data and the training of a large (800 million parameter) transformer model. To validate this contribution, the authors must include ablation experiments demonstrating both parameter scaling and data scaling in some benchmarks. Ablation experiments should be performed, and they can be on a much smaller total number of parameters (e.g., less than 50M) to reduce computational burden.

3. Benchmarks and Comparisons: The authors mention various methods with different architectures, data sizes, and parameter counts. They should include representatives of these methods in their benchmarks. For gene-based benchmarks (e.g., Figure 2), comparisons should be made to models like GenePT [1] / scELMO [2] and UCE, which use gene embeddings from large language or protein models. For cell-based benchmarks, especially the zero-shot cell type annotation task, multi-species models like UCE and GeneCompass should be included. Additionally, the authors should compare CellFM to scFoundation, which shares a similar masked autoencoding architecture. It would also be valuable to assess embedding quality for different organs and sequencing protocols.

4. Unclear Pre-training Setting: The authors should discuss the limitations of their model. In particular, the pre-training step, where gene IDs are padded and random expression values are added (Section 4.3), raises concerns. The pre-training task uses a masked language modeling loss, where 20% of the input is masked. How can the model learn meaningful compression if the input includes random values? The impact of excluding these cells from the 100M dataset should be investigated.

5. Benchmarks and Missing Details on Cell-type annotation: The cell-type annotation benchmark is unrealistic and should be replaced by benchmarks from the single cell integration benchmark scIB [3], scGraph [4], and out-of-domain settings [5]. A randomized train-test split for cells is an unrealistic setting and does not adequately measure the ability of a model to encode cell types or correct batches. Additionally, the presentation of these results should be improved: in figure 4ef, results from different dataset should not be shown in the same box plot. If the authors continue to use the same 70% validation split, they should perform multiple splits to characterize the variance of these metrics. Finally, it is not clear whether these are truly zero-shot metrics: were any of these datasets included in training of any model? Is the base-foundation model fully frozen? What model is used to make the cell type annotations / predictions?

6. Performance and Counterintuitive Interpretation on Perturbation Predictions: The reported improvement in performance on the Norman and Adamson perturbation datasets is very minor according to the authors (1.65% and 2%). How does such a minor improvement in performance lead to such significant changes in the hit rate when predicting in the reverse direction (Figure 4e)? GEARs should also be compared on this metric. Additionally, when predicting the expression of cells under perturbation, the output looks highly unrealistic (Supplementary Figure S1). Many perturbations have no effect, or similar

effects, yet the outputs all form tight clusters.

Minor Comments:

- Referring to foundation models as "large language models" (LLMs) is inappropriate since these models are not trained on natural language. A more suitable term would be "foundation models."

- The color scheme in Figure S3 is confusing.

- In Figure 2c, the label should read "CellFM" rather than "CellAI."

- In Figure S2, clarify which dataset the true box represents (e.g., test, validation, train, or unseen).

- The text in many figures is small and hard to read, requiring significant zooming. Subfigure order is inconsistent across figures, such as clockwise in some and counterclockwise in others.

- Clarify which tasks use fine-tuning and when LoRA is applied. The authors do not compare the performance of full fine-tuning versus LoRA.

- A computational comparison between ERetNet and a vanilla transformer, along with ablation studies on Simple Gated Linear Unit and Layer Normalization, would provide additional insights.

- Although the authors have made the curated data publicly available, they should provide a list of the training datasets and their sources.

- Provide a clear description of the limitations of their current model, potentially using a model card format as outlined in [6].

References

[1] GenePT: A Simple But Effective Foundation Model for Genes and Cells Built From ChatGPT, Chen & Zou, bioRxiv 2023
[2] scELMo: Embeddings from Language Models are Good Learners for Single-cell Data Analysis, Liu et al., bioRxiv 2023
[3] Benchmarking atlas-level data integration in single-cell genomics, Luecken et al. Nature Methods 2021
[4] Metric mirages in Cell embeddings, Wang, Leskovec & Regev, bioRxiv 2024
[5] Cell-ontology guided transcriptome foundation model, Yuan et al., arXiv 2024
[6] https://huggingface.co/docs/hub/en/model-cards

(Remarks on code availability)
We briefly reviewed the GitHub repository, which contains both the pre-training and downstream modules. However, we do not have the budgets to run pre-training. Pre-trained models and fine-tuning notebooks should be provided.

Reviewer #5

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)
See main review

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors have done a remarkable work in addressing all of my concerns. I would still like to see a more well-document GitHub in terms of notebooks. As a final comment I would like to see the code of the extra experiments they made with CellOT and GEARS as I was not able to see them in the GitHub (maybe I am missing something). A final check on the wording should also take place. Apart from these comments I am not against supporting a positive outcome for publication.

(Remarks on code availability)
I would like to see the benchmark experiments with CellOT and GEARS since I was not able to track them down in the repo.

Reviewer #2

(Remarks to the Author)
The writing and figure quality have significantly improved in the current version. However, based on the updated results, this study's novelty and the improvements in benchmarking performance remain relatively limited when compared to existing works.
Major:
1. The authors have addressed my question, and I have no more comments.
2. The authors have addressed my question. However, in "We identified the cell types for approximately 70 million cells in the training datasets, as the remaining cells were not classified in their original papers", are the authors saying, "We identified the cell types for approximately 70 million cells in the training datasets, as these cells were not classified in their original papers"? Additionally, a typo in Figure S1, is "Brest" "Breast"?
3. The authors have addressed my question. However, in Figure 2d, there are some shades in the CellFM legend.
4. The authors have addressed my question, and I have no more comments.
5. The authors have addressed my question, and I have no more comments.
6. The authors have addressed my question, and I have no more comments.
7. The authors have addressed my question, and I have no more comments.
8. The authors have addressed my question, and I have no more comments.
9. The authors have addressed my question, and I have no more comments.
10. The authors have addressed my question, and I have no more comments.
11. For "the current attention map fails to capture the gene relationships among CCNB2, CCNA2, TOP2A, MKI67, KPNA2, and CENPF due to their low attention scores. In the future, we plan to explore new explainability techniques to address this limitation." I do think this question should be addressed in this study, especially after authors' answer to question 15 showed that there is only minor improvements over existing commonly used model.
12. It's not that meaningful and persuasive to show the Figure 5f KEGG pathway enrichment analysis. I'd suggest the following two options: 1) remove Figure 5f, and directly mention specific target genes and their relevant studies in the manuscript, 2) add the detailed gene names to each enriched pathway name, to show whether these enriched pathways were contributed by the same genes or not. Additionally, Figure 4e is also not informative, the scores are quite large and close to each other. The authors need to provide the direct evidence from public SPI1 ChIP-seq datasets, e.g., snapshots of ChIP-seq BigWig files visualized in IGV genome browser. The same results should be provided for the transcription factor (TF) JUN in Figure S20 as well.
Besides, in Figure S20b, the label of the gene between FRG1 and KNPA2 is not shown.
I'd suggest building a TF-TF network based on the similarities of influenced genes (probably not just top influenced 20 genes), this analysis would be useful for biologists to understand the gene regulatory mechanisms in cells.
13. The authors have addressed my question, and I have no more comments.
14. The authors have addressed my question, and I have no more comments.
15. The updated results showed that there are minor improvements over the classic Transformer model. The novelty of this study and the improvements in benchmarking performance are relatively limited compared to existing studies.
A typo in "Based on the findings in the "Scaling of data and model size" section (Supplementary Note 4), we used CellFM with 8,000 model parameters (CellFM-80M) for cell annotation and batch effect correction tasks." "8,000" should be the wrong number.
16. The authors have addressed my question. However, miR-29a is micro-RNA, not lncRNA. The authors need to revise and simplify the corresponding descriptions.

Minor:
I have no more comments.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Reviewer #4

(Remarks to the Author)
Overview:

Following the first round of reviews the authors have made a number of improvements to their manuscript, including adding benchmarks, ablation studies and clarifying text. However, there remain significant concerns over which benchmarks are chosen and how their results are conveyed to the reader. Additionally, there remain a number of statements or conclusions that are not backed up by experiments that should be modified or removed.

Benchmarking Setup and Presentation

The benchmarks used in the paper are insufficient or not useful for understanding model performance. Benchmarks should measure model performance in realistic use cases. No one annotates cell types for a random proportion of a dataset, and then uses their annotations to annotate the rest of the data, except in the rare case of subsampling a very large dataset.

Newly collected data, when annotated, will come from new experiments and batches in almost all cases. Therefore, the random split cell type annotation task is unconnected to the real use cases of such models. In comparison, the more standard benchmark, scIB, measures overall embedding quality, using a number of different metrics. While not a perfect set of metrics by any means, SCIB more closely reflects the real-world use of such a model, which will be used for generating embeddings that are used in downstream tasks like clustering. The authors already modified the manuscript to include scIB scores for one single dataset, PBCM 10k, but did not benchmark the other datasets.

Using scIB as the primary benchmark for embedding quality will also improve other questions about the benchmarking setup, including the choice of a 3 layer MLP for cell type annotation, which is an overly complex network, again unreflective of most real world use cases.

For the additional batch effect benchmark added, the authors only tested one dataset, PBMC 10k. PBMC 10k is a poor choice for this test since it contains a small number of cell types and only two batches. For two other datasets, liver and pancreas, the authors only compare CellFM and scFoundation, and only use UMAP, which is not empirical. The authors should perform more scIB benchmarks on more complicated datasets with many batches and cell types, such as various atlas datasets like the Human Brain Cell Atlas. It would also be beneficial to check whole organism datasets like Tabula Sapiens to investigate cross-tissue performance.

For the cell type annotation benchmark, the authors did not include statistics representing the uncertainty of the results, which is critical for comparing models. Out of curiosity, I'm particularly interested in how this annotation framework works to distinguish between exhausted and activated CD8+ T cells, as discussed in [1][2].

[1] Pan-cancer mapping of single CD8+ T cell profiles reveals a TCF1:CXCR6 axis regulating CD28 co-stimulation and anti-tumor immunity, Cell reports Medicine 2024
[2] A Cancer Cell Program Promotes T Cell Exclusion and Resistance to Checkpoint Blockade, Cell 2024

Benchmarking Presentation

The presentation of benchmarking results in the paper is less satisfying.

The authors need to ensure that benchmarking results are presented in a straightforward and empirical manner. In Figures 4 ABEF the authors present more than 100 individual data points (15 experiments x 11 methods) and in Figure S12, Figure S13 dozens of data points. These experiments are incorrectly grouped into box plots by method, even though they are from totally different experiments. Each of these experiments individually also requires measurement of uncertainty (see previous section).

Overall, it is very difficult to compare results between methods. The authors should include a table with the actual numerical results of each benchmark, divided by dataset and method. When presenting these results, the authors need to make it clear in the description and on the figure how each method was assessed, specifically, if the underlying foundation model was finetuned or not (this is not clear in Figures S14, S15).

Finally, UMAP should not be used as a tool to compare embeddings from different models. Please remove this text from section 2.6 of the text and other sections which may mention it.

Single Species vs Multiple Species Training

The authors continue to claim that:

"While larger datasets spanning diverse species can enhance model generalization, the inherent inter-species differences may introduce noise and hinder model performance." (Abstract)

In their response, the authors claim that:

"In fact, our study indicates human dataset-based models (scFoundation and our model) tend to outperform diverse species models (GeneCompass and UCE) in human test sets. Concretely, when a model is trained on such heterogeneous data, it may inadvertently learn features that are not relevant to humans, leading to overfitting and bias."

These models are all trained on different data corpus, with different sizes and different architectures. In order to make this claim, the authors, or someone else, would need to run a proper ablation experiment testing specifically the impact of including other species' data in the training corpus. The authors of GeneCompass did this with their architecture and found that including data from both species improved results (GeneCompass figure 3A).

This claim should either be removed from the manuscript or confirmed empirically.

Minor Comments:

PBMC 10k is missing from the Table S4, or is labeled with a different name, this is important because it can have a different number of genes depending on where it is downloaded from.

(Remarks on code availability)

Reviewer #5

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Version 2:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
My questions regarding perturbations from the CellFM model have been answered.

(Remarks on code availability)
I was able to review the extra code that I have asked and it seems in order.

Reviewer #2

(Remarks to the Author)
The authors have addressed our comments very well this round. We really appreciate the authors took efforts to revised the manuscript given our comments.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Reviewer #4

(Remarks to the Author)
We acknowledge that the authors have addressed most of our previous comments. We have a few minor remaining points:

1. Benchmark Transparency: Please clarify, for each dataset, whether models were fine-tuned or used in a zero-shot setting, and indicate this in both the figure and its caption.

2. Score Reporting: Report all individual scIB metric scores, not just the averages (AvgBio), to support fair and reproducible evaluation.

3. Model Consistency: Use a consistent set of models (e.g., scBERT, scGPT) across all datasets and tasks to ensure comparability.

(Remarks on code availability)

Reviewer #5

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Reviewer #1 (Remarks to the Author):

Review - Major revisions

The manuscript titled “CellFM: a large-scale foundation model pre-trained on transcriptomics of 100 million human cells” submitted by Zeng et al. is a promising take on the rapidly emerging field of large language models (LLMs)/foundational models in single-cell analysis since it surpasses existing models in terms of size while providing interesting nuances through it is architecture. Progressing this sub-field of single-cell computational biology is paramount to harnessing the immense capabilities of foundational models, which can provide unparalleled analytical solutions compared to traditional machine learning and deep learning approaches after being trained on many datasets. I believe that this work exhibits apparent scientific novelty, but there are critical issues that should be addressed to adhere to the high standards of the esteemed Nature Communications journal.
Response: We sincerely appreciate the reviewer’s positive feedback on our manuscript. We are encouraged by the recognition of the scientific novelty and the potential impact of our work in advancing foundational models for single-cell analysis. We also appreciate the constructive comments and will carefully address the critical issues raised to meet the high standards of Nature Communications.

1. An initial issue the authors should address is significantly improving the manuscript and its context in writing and phrasing. I advise the authors to consider the excellent paper by Fabian Theis regarding the foundational models that are taking the field by storm (https://doi.org/10.1038/s41592-024-02353-z) and provide a bit more context in the Introduction or the Discussion sections regarding CellFM’s position among the various other foundational models. For example, I suspect that CellFM can be construed as a model belonging to the third category of transformers (value projection) and not in the other two (ordering, value categorisation).
Response: Thanks for your valuable suggestions. We have revised the manuscript to improve both its writing and phrasing in the “Introduction” section. In addition, we have categorized all single-cell foundational models trained from scratch into three distinct groups as described in the study (PMID: 39122952) and classified CellFM under the "value projection" category. Further details are available in the “Introduction” section, with a brief overview provided as follows:

“
CellFM is classified as a value-projection-based single-cell foundation model, as it seeks to recover the vector embeddings of masked genes through their linear projections, which are based on gene expression values.
”

2. My following criticism relates to reshaping the perturbation modelling section, where some key aspects should be addressed:

2. 1 please provide some references for the first introductory paragraph of this section, some definitions of perturbation modelling, and some insight as to why this is probably the most translationally significant type of single-cell analysis since it connects to computational

pharmacology and AI-driven drug discovery or repurposing. To aid the authors, I would encourage them to provide this information as a supplement if they feel that the flow of the text in the Result section will be compromised.

Response: Thanks for your valuable suggestions. We have provided definitions of perturbation modelling and some insight. Concretely, we first introduced them as follows:

Definitions of Perturbation Modeling:

Recent advancements in sequencing and gene editing have enabled large-scale experimental perturbation simulations to study changes in gene expression, cellular behavior, and phenotypic outcomes. These simulations are essential for understanding cellular responses to various stimuli and are increasingly applied to investigate drug effects, disease mechanisms, and therapeutic interventions. However, the vast combinatorial space of potential gene perturbations quickly exceeds the practical limits of experimental feasibility. To overcome this, single-cell foundation models adopt perturbation modeling, leveraging knowledge from known experimental perturbations to predict responses to unknown perturbations. By utilizing self-attention mechanisms over the gene dimension, these models can capture intricate gene interactions and accurately predict gene expression responses in unseen perturbations.

Translational Significance of Perturbation Modeling in Single-Cell Analysis:

The predictive power of perturbation modeling becomes especially valuable in AI-driven drug discovery, where it is used to forecast how existing drugs or genes will affect cellular processes, identify new drug targets, and explore drug repurposing opportunities. Additionally, these models offer deep insights into cellular heterogeneity, which is crucial for developing personalized medicine strategies. As computational pharmacology and AI-driven drug discovery continue to evolve, integrating single-cell perturbation models will increasingly bridge the gap between basic science and clinical applications. This enables the identification of novel biomarkers, prediction of drug responses, and optimization of therapeutic strategies, ultimately accelerating drug development and improving patient outcomes.

We have incorporated aforementioned background introduction into the Perturbation section as follows:

“

Recent advancements in sequencing and gene editing have enabled large-scale experimental perturbation simulations to study changes in gene expression, cellular behavior, and phenotypic outcomes. These simulations are essential for understanding cellular responses to various stimuli and are increasingly applied to investigate drug effects, disease mechanisms, and therapeutic interventions. However, the vast combinatorial space of potential gene perturbations quickly exceeds the practical limits of experimental feasibility. To overcome this, single-cell foundation models adopt perturbation modeling, leveraging knowledge from known experimental perturbations to predict responses to unknown perturbations. By utilizing self-attention mechanisms over the gene dimension, these models can capture intricate gene interactions and accurately predict gene expression responses in unseen perturbations.

”

“

The predictive power of perturbation modeling becomes especially valuable in AI-driven drug discovery, where it is used to forecast how existing drugs or genes will affect cellular processes, identify new drug targets, and explore drug repurposing opportunities.

”

2.2 explain why you are focusing on the MSE as a metric to be minimised and not on some other metric (some references could be useful here) during the MLP stage.

Response: Thanks for your valuable suggestions. In CellFM, we focus on minimizing the Mean Squared Error (MSE) as the primary metric because it effectively measures the discrepancy between the predicted and actual gene vector embeddings for masked genes. MSE is particularly suitable in this context as it penalizes larger errors more heavily, making it crucial for accurately recovering gene representations.

Additionally, MSE has been widely adopted in similar tasks involving gene expression prediction and representation learning. For example, scFoundation and GeneCompass employ MSE to optimize gene expression prediction in high-dimensional spaces, demonstrating its effectiveness in promoting precise modeling. These studies highlight MSE's ability to improve model performance in scenarios similar to ours, particularly in single-cell analysis. Detailed information regarding this can be found in the "Methods" section, as summarized below:

“

In CellFM, we focus on minimizing the Mean Squared Error (MSE) as the primary metric because it effectively measures the discrepancy between the predicted and actual gene vector embeddings for masked genes. MSE is particularly suitable in this context as it penalizes larger errors more heavily, making it crucial for accurately recovering gene representations. Additionally, MSE has been widely adopted in similar tasks involving gene expression prediction and representation learning. For example, scFoundation and GeneCompass employ MSE to optimize gene expression prediction in high-dimensional spaces, demonstrating its effectiveness in promoting precise modeling. These studies highlight MSE's ability to improve model performance in scenarios similar to ours, particularly in single-cell analysis.

”

2.3 I believe that the comparative study with scGPT is narrow. It should involve at least one or two more foundational models (scELMo, scFoundation or GeneCompass indicatively) as well as 1 or 2 non-foundational models that I assume would derive from the group of recent generative AI models predicated on autoencoders (i.e., scGEN-inspired models like scVIDR or CellOT). This is something that the authors should heed because in several publications (https://doi.org/10.1038/s41592-024-02353-z, https://doi.org/10.1016/j.csbj.2024.04.058) there are concerns that the fledgling foundational models are exhibiting questionable performance compared to simple machine-learning or deep-learning models.

Response: Thanks for your valuable suggestions. We have expanded our comparative analysis to include recent single-cell foundation models (such as scFoundation, GeneCompass, and scELMo),

alongside non-foundational models like GEARS on the gene perturbation datasets. For all single-cell foundation models, we followed your recommendation to combine them with the classic perturbation model GEARS. Concretely, we replaced GEARS' gene embeddings with those derived from CellFM (Figure 3a). As shown in Figure 3b-c, our model consistently outperformed all single-cell foundation models achieving improvements of 1% and 1.45% in average PCC and MSE compared to the second-ranked model scFoundation, respectively. Additionally, CellFM consistently surpassed GEARS with 4.75% and 7% improvement regarding average PCC and MSE values, respectively. For the reverse perturbation prediction, we maintained the original comparison for CellFM and only evaluated it against scGPT and GEARS since the newly added single-cell foundation models neither provided codes nor published corresponding results in their original studies (Supplementary Figure S3), and reimplementing these models would have required substantial programming effort and time.

On the other hand, we also evaluated the performance of CellFM on the drug perturbation data. Concretely, we integrated CellFM and CellOT by replacing the input cell representation of CellOT with the cell representation from CellFM (Supplementary Figure S9), as CellOT is specifically tailored for drug perturbation and does not require gene embeddings. In this setting, we compared CellFM with CellOT, scGEN, and Identity. As indicated in Figure S9, CellFM outperformed CellOT in perturbation prediction on four drugs, achieving improvements of 66.6% and 2.2% in average l_2 and PCC, respectively. These new results have been incorporated into the "Results" section as follows:

“

We evaluated CellFM on the perturbation task by combining CellFM with the classic perturbation model GEARS as suggested by scFoundation. Concretely, we replaced GEARS' gene embeddings with those derived from CellFM (Figure 3a). As shown in Figure 3b-c, our model consistently outperformed all single-cell foundation models achieving improvements of 1% and 1.45% in average PCC and MSE compared to the second-ranked model scFoundation, respectively. Additionally, CellFM consistently surpassed GEARS with 4.75% and 7% improvement regarding average PCC and MSE values, respectively.

”

“

To further validate the performance of CellFM on drug-perturbed data, we conducted additional evaluations using these datasets. Similarly, we also combined CellFM with a non-single foundation model CellOT [PMID: 37770709], which was a classic model for drug perturbation prediction. Concretely, we integrated CellFM and CellOT by replacing the input cell representation of CellOT with the cell representation from CellFM (Supplementary Figure S9), as CellOT is specifically tailored for drug perturbation and does not require gene embeddings. In this setting, we compared CellFM with CellOT, scGEN, and Identity. As indicated in Figure S9, CellFM outperformed CellOT in perturbation prediction on four drugs, achieving improvements of 66.6% and 2.2% in average l_2 and PCC, respectively.

”

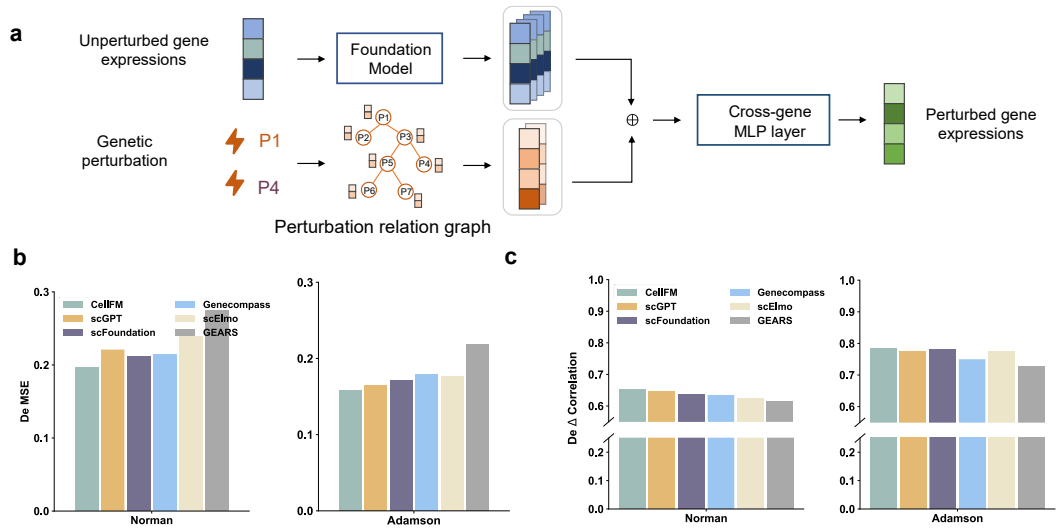


Fig. 3. Analysis of Perturbation Response and Reverse Perturbation Predictions. (a) A diagram showcasing the perturbation prediction model leveraging cell-specific gene embeddings derived from CellFM. (b) The mean square error (MSE) between predicted and actual post-gene expressions for the top 20 differentially expressed (DE) genes. (c) Comparison of CellFM with other single-cell foundation models and the perturbation prediction method GERAS. Pearson correlation coefficients between predicted and actual gene expression changes are reported for the top 20 differentially expressed (DE) genes.

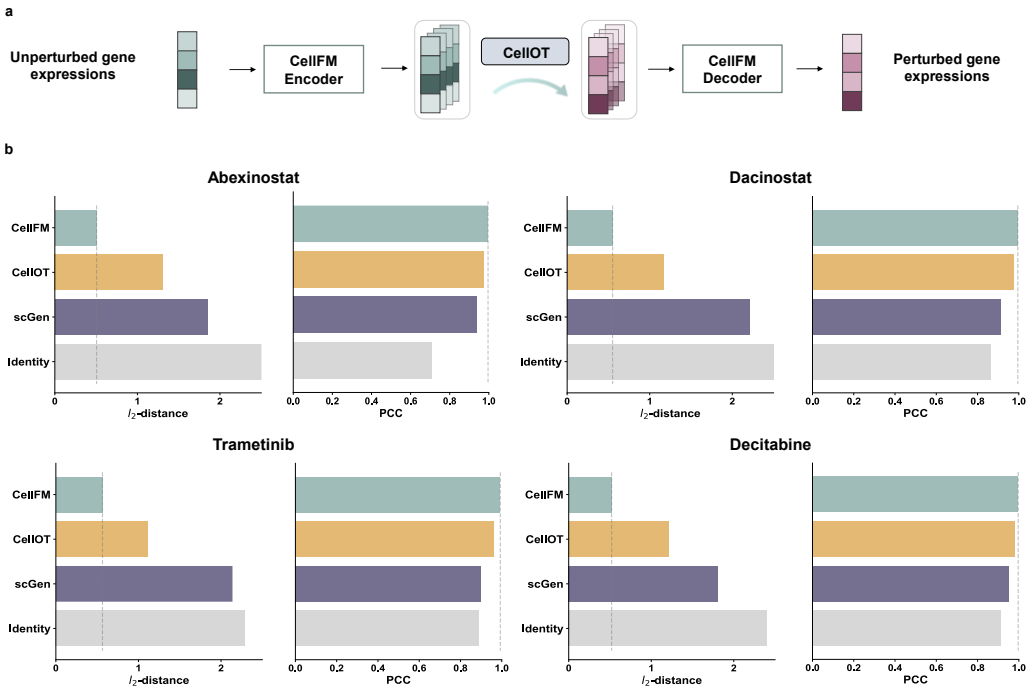


Fig. S9. Performance assessment of CellFM compared to different baselines with respect to l_2 distances and PCC between the observed perturbed cells and predicted responses from control cells. The trivial identity was included as a baseline, predicting treatment effects by simply returning the untreated states.

Models	Open soucre	Gene Function Prediciton		Cell type annotaion	Batch effect corrrection	Perturbation Prediciton	
		Binary classification	Multi-category classification			Predicting unseen gene perturbations	In silico reverse perturbation prediction
CelllFM	✓	✓	✓	✓	✓	✓	✓
scGPT	✓			✓	✓	✓	✓
Geneformer	✓	✓		✓			
scFoundation	✓			✓		✓	
GeneCompass	✓	✓		✓		✓	
UCE	✓			✓	✓		
scELMo	✓	✓		✓	✓	✓	
scBERT	✓			✓			

Fig. S3. The table provides hints for selecting a model based on task usability. Blank cells indicate that the selected models lack specific designs to meet the corresponding criteria.

2.4 In Figure 3A, although it is not clearly described in the main text, based on the legend description, comparisons are made regarding how well the models can generalise to unseen events by comparing predicted gene expressions to ground-truths. Is this accurate? In that case maybe the authors could try to calculate the R2 (square) metric also which describes how much of the variance in actual gene expression can be explained by the predicted values.

Response: Thanks for your valuable suggestions. It is true that we have trained our model on data with known gene perturbations and then used the trained CellFM to predict gene expression for unseen gene perturbations in control cells. In response to your recommendation, we have now included the R^2 metric to further evaluate CellFM's performance. As shown in Supplementary Figure S4, CellFM consistently outperformed all other single-cell foundation models, as well as the non-foundational model GEARS. CellFM achieved an R^2 value that was 1.3% higher than the second-best model scGPT in terms of average R^2. The results have been added to the "Results" section as follows:

"

As shown in Supplementary Figure S4, CellFM consistently outperformed all other single-cell foundation models, as well as the non-foundational model GEARS. CellFM achieved an R^2 value that was 1.3% higher than the second-best model scGPT in terms of average R^2.

"

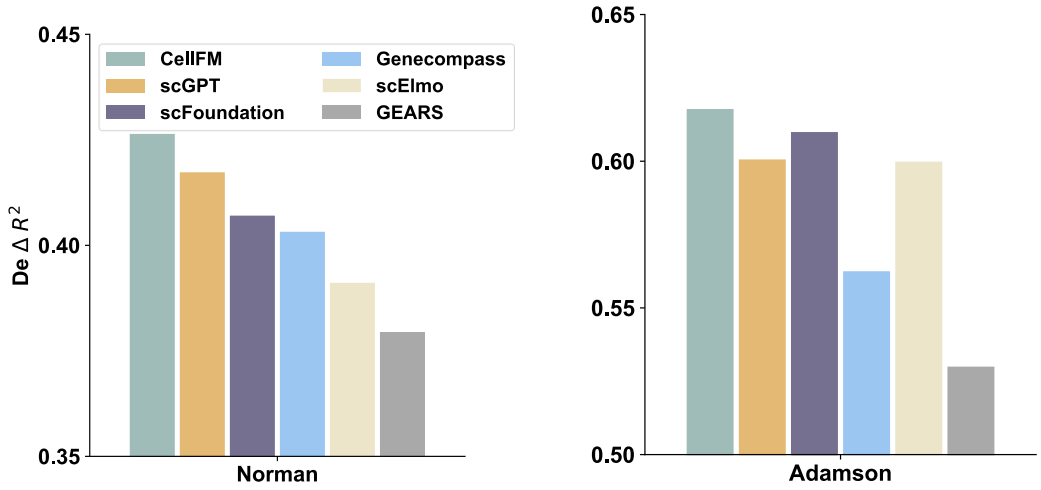


Fig. S4. In a zero-shot learning context, a comparative analysis of CellFM against other models on the Norman and Adamson datasets, emphasizing the R^2 between the predicted and actual changes in gene expression for the top differentially expressed genes.

2.5 I am not sure what is happening in Supl. Figure 1. Is CellFM trained on the existing KOs of the Norman dataset and extrapolates to other perturbations? Does it perform digital KOs based on the prior knowledge? Are the genes depicted on each cluster represent the target of the simulated KO?

Response: Thanks for your valuable comments. It is true that we trained CellFM on the existing knockouts (KOs) from the Norman dataset and extrapolated to other perturbations. However, we did not use prior knowledge explicitly during the in silico expansion of untested perturbations. CellFM was trained on the original Norman dataset, which covers 236 perturbations targeting 105 genes. We then used the fine-tuned model to predict and expand the responses to untested perturbation combinations in silico. These predictions were visualized as the response for each perturbation combination. As illustrated in Supplementary Figure S5 (Fig S1 in initial submission), the genes depicted in each cluster represent the "dominant gene" within the perturbation combinations. The results demonstrate a strong association between the clusters and their respective dominant genes. For example, the cluster associated with the *SET* gene indicates that the data points in this cluster correspond to combined perturbations involving *SET* and another gene (e.g., *SET+CLDN6*, *SET+MIDN*). The detailed information is provided in the response to Reviewer #4, question #6, where we describe the steps taken to replot Supplementary Figure S5 in detail. We updated the description in “Results” section as follows:

“

We trained CellFM on the existing knockouts (KOs) from the Norman dataset and extrapolated to other perturbations. CellFM was trained on the original Norman dataset, which covers 236 perturbations targeting 105 genes. We then used the fine-tuned model to predict and expand the responses to untested perturbation combinations in silico. These predictions were visualized as the response for each perturbation combination. As illustrated in Supplementary Figure S5, the genes depicted in each cluster represent the "dominant gene" within the perturbation combinations. The results have demonstrated a strong association between the clusters and their respective dominant

genes. For example, the cluster associated with the *SET* gene indicates that the data points in this cluster correspond to combined perturbations involving *SET* and another gene (e.g., *SET+CLDN6*, *SET+MIDN*).
”

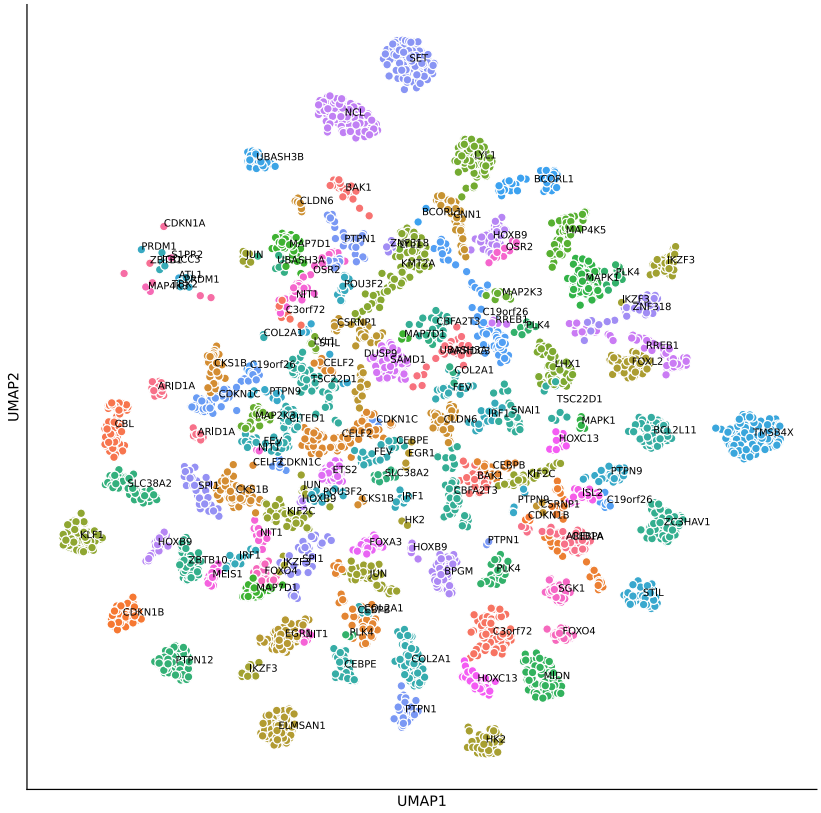


Fig. S5. The UMAP visualization of predicted gene expression profiles under various perturbation conditions. Each cluster in the UMAP plot is annotated with the predominant gene for each cluster.

2.6 In silico reverse perturbation prediction: is this a term the authors are coining or is it something from the literature? References are missing if this is the latter. This represent a quite ambitions approach of perturbation modelling which I think has immense translation impact. I would advise the authors to provide some additional insights from other datasets (e.g., sciPlex with high-throughput screening, control vs disease scRNA-seq data). If CellFM can reverse-engineer its way to original perturbation events that are causing the cellular phenotype in question, this should be researched further.

Response: Thanks for your valuable comments. The term “in silico reverse perturbation prediction” originates from the scGPT study (PMID: 38409223) and we have added this citation in our manuscript. Additionally, this concept is specifically suited for predicting original perturbations from gene perturbation data. The current study is limited to gene perturbation without considering

drug molecules, and thus our method doesn't apply to drug perturbation datasets such as sciPlex, since drug perturbation datasets require additional information specific to the drug molecules themselves.

Moreover, effective drug perturbation prediction often requires separate models tailored to each drug, as demonstrated by methods like CellOT, which trains an individual model for each drug perturbation. While this approach performs well on specific datasets, it lacks generalizability to novel drugs. Modifying the architecture of CellFM alone would likely be insufficient to accommodate these differences effectively. Therefore, in this version, we do not yet provide a reverse prediction function for drug perturbations within CellFM. In future, we will expand CellFM's capabilities to support drug perturbation datasets, which will involve adapting the model architecture to incorporate drug-specific molecular information. This would enable broader applicability for reverse prediction in a drug perturbation context. We have added this detailed information into "Discussion" section as follows:

"
We further performed "in silico reverse perturbation prediction," following the approach outlined in the scGPT study (PMID: 38409223).
"

"
The current study is limited to gene perturbation without considering drug molecules. Thus, our method doesn't perform reverse perturbation prediction in silico on the drug perturbation datasets such as sciPlex, since drug perturbation datasets require additional information specific to the drug molecules themselves. In the future, we will expand CellFM's capabilities to support drug perturbation datasets, which will involve adapting the model architecture to incorporate drug-specific molecular information.
"

2.7 The authors are also using GEARS as part of this small comparison which comes completely out of the blue in the flow of the text, without any background information especially for the reader that is not an expert in single-cell perturbation biology. I would surmise that the authors are using GEARS because it shares some resemblences with the CellFM in terms of using leveraging some prior knowledge (a knowledge graph in the case of GEARS, the voluminous dataset the CellFM is trained upon) and an MLP component. In any case, background information should be provided for the interested reader to delineate this part of the analysis. Furthermore, it is a further attestation to my previous claim that simpler, non-foundational models should be included in the previous perturbation analysis.

Response: Thanks for your valuable suggestions. We have added several background information about GEARS. Concretely, GEARS is a computational tool specifically designed for predicting single and multi-gene perturbations based on single-cell RNA sequencing (scRNA-seq) datasets. Its primary aim is to enhance our understanding of genetic interactions and their functional implications in various biological contexts. GEARS operates by integrating a gene-gene interaction network as

prior knowledge, which allows it to leverage existing biological information to improve prediction accuracy. The model employs a cross-gene neural network architecture in conjunction with a graph neural network, enabling it to capture complex relationships between genes and predict changes in gene expression following perturbations. This tool is particularly valuable for researchers studying single-cell perturbation biology, as it addresses the challenges of interpreting high-dimensional data from scRNA-seq experiments. Regarding non-single-cell foundational models, GEARS has demonstrated state-of-the-art performance in gene perturbation predictions, making it a leading choice in the field. For further details on GEARS and its applications, we recommend referring to the original publication (PMID: 37592036).

Additionally, since many single-cell foundational models use GEARS as the baseline for smaller models, we included a comparison between the non-foundational models GEARS and CellOT in the perturbation analysis. Detailed information can be found in the response to above response #2.3 and #2.4. We have added the information in section "Results" as follows:

"
GEARS is a computational tool specifically designed for predicting single and multi-gene perturbations based on scRNA-seq datasets. GEARS operates by integrating a gene-gene interaction network as prior knowledge, which allows it to leverage existing biological information to improve prediction accuracy. Regarding non-single-cell foundational models, GEARS has demonstrated state-of-the-art performance in gene perturbation predictions, making it a leading choice in the field.
"

2.8 Hybrid approaches of foundational and non-foundational models are also being explored in single-cell perturbation modelling. For example, scELMo can provide cell and feature embeddings in task-specific models (e.g., causal factor analysis in CINEMA-OT, gene expression predictions for cell-level and gene-level perturbations with CPA and GEARS, respectively) to simulate the effects of perturbations, outperforming the original iterations of these models. I would advise the authors to explore such a combination if deemed feasible between CellFM and some other tool.
Response: Thanks for your valuable suggestions. We have combined all single-cell foundation models including CellFM with GEARS to evaluate their performances through the $Pearson_{\mathrm{delta}}$, MSE, and R^2. Specifically, the original GEARS model used a Gene Ontology knowledge graph to represent unseen gene perturbations by learning from a combination of previously observed gene perturbation nodes, and a gene co-expression graph combined with perturbation information to predict the post-perturbation gene expression. Each node in the co-expression graph represented a gene with initially randomized embeddings, and edges connected to co-expressed genes. This graph was shared across all cells.

In our setting, we obtained gene context embeddings for each cell from each single-cell foundation models and set these embeddings as the nodes in the graph, resulting in a cell-specific gene co-expression graph for predicting perturbations (Figure 3a). As shown in Figure 3b-c, our model consistently outperformed all single-cell foundation models achieving improvements of 1% and 1.45% in average PCC and MSE compared to the second-ranked model scFoundation, respectively. Additionally, CellFM consistently surpassed GEARS with 4.75% and 7% improvement regarding average PCC and MSE values, respectively. For the reverse perturbation

prediction, we maintained the original comparison for CellFM and only evaluated it against scGPT and GEARS since the newly added single-cell foundation models neither provided code nor published corresponding results in their original studies (Supplementary Figure S3), and reimplementing these models would have required substantial programming effort and time. The detailed information can be found in the response to the above response #2.3 and #2.4. We have added these results into section "Results" as follows:

"

We have combined all single-cell foundation models including CellFM with GEARS and evaluated their performance through the $Pearson_{\mathrm{delta}}$, MSE, and R^2. Specifically, the original GEARS model used a Gene Ontology knowledge graph to represent unseen gene perturbations by learning from a combination of previously observed gene perturbation nodes, and a gene co-expression graph combined with perturbation information to predict the post-perturbation gene expression. Each node in the co-expression graph represented a gene with initially randomized embeddings, and edges connected to co-expressed genes. This graph was shared across all cells.

In our setting, we obtained gene context embeddings for each cell from each single-cell foundation models and set these embeddings as the nodes in the graph, resulting in a cell-specific gene co-expression graph for predicting perturbations (Figure 3a). As shown in Figure 3b-c, our model consistently outperformed all single-cell foundation models achieving improvements of 1% and 1.45% in average PCC and MSE compared to the second-ranked model scFoundation, respectively. Additionally, CellFM consistently surpassed GEARS with 4.75% and 7% improvement regarding average PCC and MSE values, respectively. For the reverse perturbation prediction, we maintained the original comparison for CellFM and only evaluated it against scGPT and GEARS since the newly added single-cell foundation models neither provided code nor published corresponding results in their original studies (Supplementary Figure S3), and reimplementing these models would have required substantial programming effort and time.
"

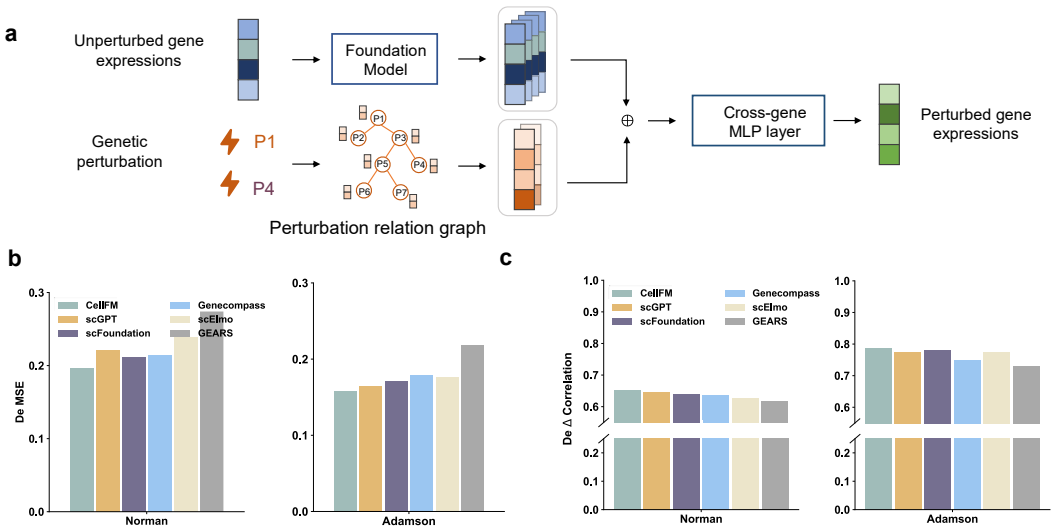


Fig. 3. Analysis of Perturbation Response and Reverse Perturbation Predictions. (a) A diagram showcasing the perturbation prediction model leveraging cell-specific gene embeddings derived from CellFM. (b) The mean square error (MSE) between predicted and actual post-gene expressions for the top 20 differentially expressed (DE) genes. (c) Comparison of CellFM with other single-cell

foundation models and the perturbation prediction method GERAS. Pearson correlation coefficients between predicted and actual gene expression changes are reported for the top 20 differentially expressed (DE) genes.

Models	Open soucre	Gene Function Prediciton		Cell type annotaion	Batch effect corrrection	Perturbation Prediciton	
		Binary classification	Multi-category classification			Predicting unseen gene perturbations	In silico reverse perturbation prediction
CelllFM	✓	✓	✓	✓	✓	✓	✓
scGPT	✓			✓	✓	✓	✓
Geneformer	✓	✓		✓			
scFoundation	✓			✓		✓	
GeneCompass	✓	✓		✓		✓	
UCE	✓			✓	✓		
scELMo	✓	✓		✓	✓	✓	
scBERT	✓			✓			

Fig. S3. The table provides hints for selecting a model based on task usability. Blank cells indicate that the selected models lack specific designs to meet the corresponding criteria.

2.9 With regards to deciphering gene relationships with CellFM, I would encourage the authors to provide some more examples apart from SPI1 regarding the emerging explainability of the attention maps in terms of the genes that are being affected a given perturbation in the Adamson dataset. There have been some sceptisism regarding whether attention can reflect actual biological dynamics ("Attention scores and feature importance have been shown not always to correlate and are hence not necessarily reliable tools for identify in which input elements are responsible for an output") (https://doi.org/10.1038/s41592-024-02353-z).

Response: Thanks for your valuable suggestions. We have added nine cases in Supplementary Figure S19. Across all nine case genes, among the top 20 genes most influenced by CellFM for each perturbation gene, an average of 18 were found in the ChIP-Atlas database. The results showed that CellFM can correctly identify influenced genes through attention scores. Additionally, we further conducted the pathway analysis on the case perturbation gene *JUN* to show that CellFM captured distinct pathway-activation patterns through the genes most influenced by perturbation gene JUN. As shown in Supplementary Figure S20, CellFM was able to capture distinct pathway-activation patterns. For example, the pathway Focal adhesion identified key genes such as *JUN, FAK, CCND3,* and *CTNNB1*, all of which play essential roles in cell adhesion and proliferation. Concretely, *JUN* encodes c-Jun, a component of the AP-1 complex that regulates gene expression and proliferation. It is closely linked to focal adhesion pathways, particularly through *FAK* (Focal Adhesion Kinase), which mediates cell attachment to the extracellular matrix (ECM) and influences migration and invasion [1]. *CCND3*, a regulator of the G1/S phase transition, is connected to *FAK* signaling, linking cell adhesion to cell cycle progression [2]. Additionally, *CTNNB1* (β-catenin) through the Wnt/β-catenin pathway promotes cell adhesion by interacting with cadherins and modulating *FAK* activity. This coordinated regulation of *JUN, FAK, CCND3*, and *CTNNB1* suggests their collaborative role in driving cancer development and progression through cell adhesion and proliferation pathways [3]. We have added these results into "Result" section as follows:

“

We have provided nine cases in Supplementary Figure S19. Across all nine case genes, among the top 20 genes most influenced by CellFM for each perturbation gene, an average of 18 were found in the ChIP-Atlas database. The results showed that CellFM can correctly identify influenced genes through attention scores. Additionally, we further conducted the pathway analysis on case perturbation genes *JUN* and *SPI1* to show that CellFM captured distinct pathway-activation patterns through the genes most influenced by perturbation gene*s*. As shown in Supplementary Figure S20, CellFM was able to capture distinct pathway-activation patterns. For example, the pathway Focal adhesion identified key genes such as *JUN, FAK, CCND3*, and *CTNNB1*, all of which play essential roles in cell adhesion and proliferation. Concretely, *JUN* encodes c-Jun, a component of the AP-1 complex that regulates gene expression and proliferation. It is closely linked to focal adhesion pathways, particularly through *FAK* (Focal Adhesion Kinase), which mediates cell attachment to the extracellular matrix (ECM) and influences migration and invasion [1]. *CCND3*, a regulator of the G1/S phase transition, is connected to *FAK* signaling, linking cell adhesion to cell cycle progression [2]. Additionally, *CTNNB1* (β-catenin) through the Wnt/β-catenin pathway promotes cell adhesion by interacting with cadherins and modulating FAK activity [3]. This coordinated regulation of *JUN, FAK, CCND3*, and *CTNNB1* suggests their collaborative role in driving cancer development and progression through cell adhesion and proliferation pathways.

”

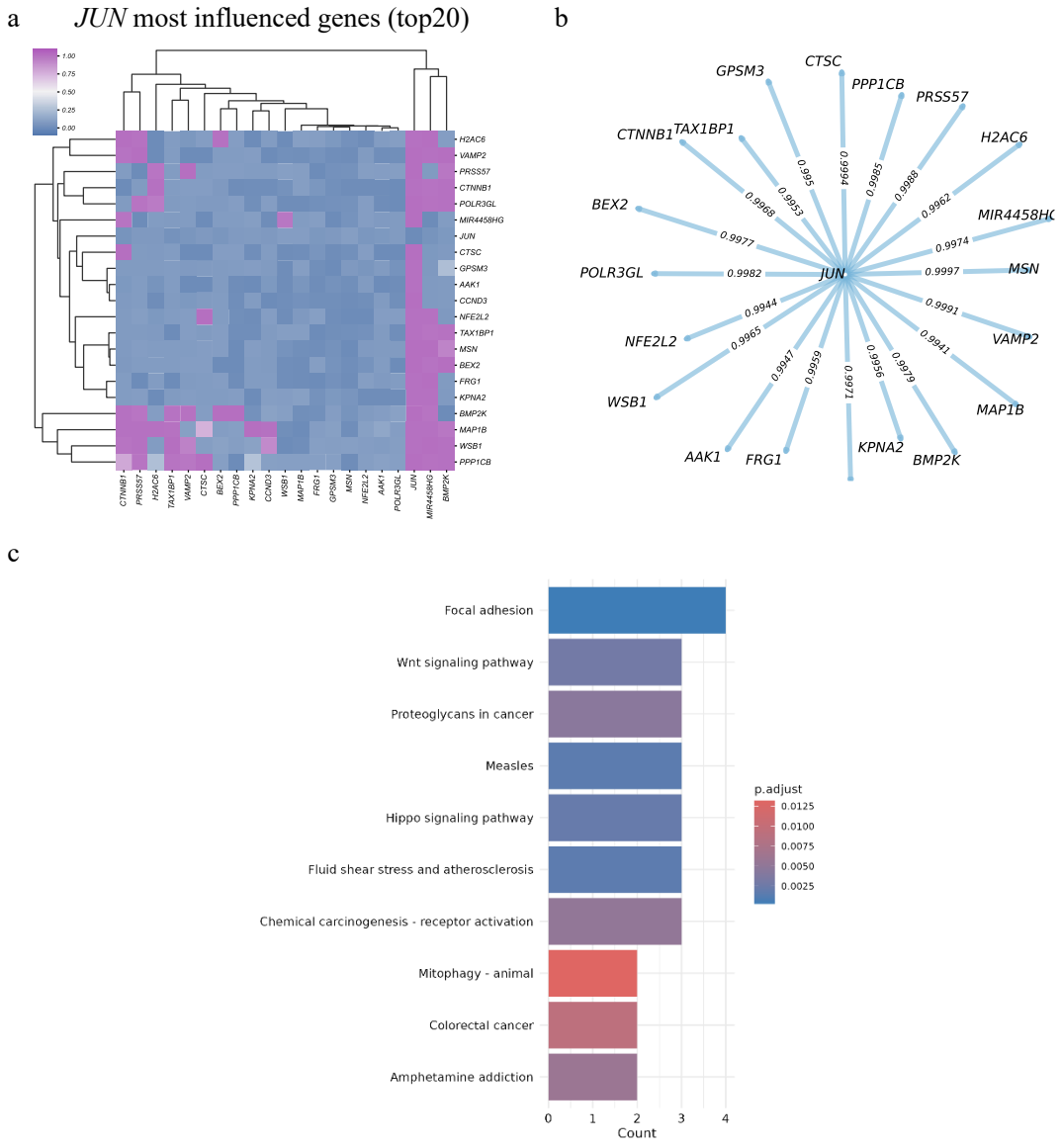


Fig. S20. (a) The heatmap displays the connectivity changes in the network of the top 20 genes most influenced by the gene *JUN*. (b) The network graph represents the top 20 genes, with ChIP-seq predicted targets validated in the ChIP-Atlas database, highlighted in light blue. (c) The heatmap illustrates the pathways mapped using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database, based on the top 20 genes most significantly impacted by *JUN*.

References:

[1] Wang, Y., Ma, J., Chen, L., Xie, X.-L., & Jiang, H. (2017). Inhibition of focal adhesion kinase on hepatic stellate-cell adhesion and migration. American Journal of the Medical Sciences, 353(1), 41-48. https://doi.org/10.1016/j.amjms.2016.11.020

[2] Yamamoto, D., Sonoda, Y., Hasegawa, M., Funakoshi-Tago, M., Aizu-Yokota, E., & Kasahara, T. (2003). FAK overexpression upregulates cyclin D3 and enhances cell proliferation via the PKC and PI3-kinase-Akt pathways. Cell Signalling, 15(6), 575–583. https://doi.org/10.1016/s0898-6568(02)00142-0

[3] Francalanci, P., Giovannoni, I., De Stefanis, C., Romito, I., Grimaldi, C., Castellano, A., D'Oria, V., Alaggio, R., & Alisi, A. (2020). Focal Adhesion Kinase (FAK) Over-Expression and Prognostic Implication in Pediatric Hepatocellular Carcinoma. International Journal of Molecular Sciences, 21(16), 5795. https://doi.org/10.3390/ijms21165795

2.10 Finally, regarding the Github page and the code provided, no script is provided for the perturbation analyses. Furthermore, the scripts from the tutorials provided by the authors are provided without comments or some context, making the results' reproducibility extremely problematic. In the spirit of Open Science and FAIRness, the scripts should be rewritten and the missing code should be provided.

Response: Thanks for your constructive comment. We have added the scripts for the perturbation analyses on our GitHub page (https://github.com/biomed-AI/CellFM/blob/main/tutorials/Perturbation/GenePerturbation.ipynb). Additionally, we have revised the tutorial scripts to include detailed comments and context, enhancing clarity and usability. The step-by-step tutorials enable the recovery of our main results.

Reviewer #1 (Remarks on code availability):

The code provided is partial and does not allow for full reproducibility of the results. Moreover, the Jupyter notebooks lack any clear structure and have almost no comments to guide the AI practitioner in recreating results or experimenting with the pipeline. A characteristic example is the perturbation notebook that is simply absent! https://github.com/biomed-AI/CellFM/blob/main/tutorials/GenePertubation.ipynb

Response: Thanks for your valuable feedback. We have updated our GitHub repository to include the missing perturbation notebook, which is now available at https://github.com/biomed-AI/CellFM/blob/main/tutorials/Perturbation/GenePerturbation.ipynb. Additionally, we have reorganized the existing Jupyter notebooks to enhance their structure and added comprehensive comments throughout the code.

Reviewer #2 (Remarks to the Author):

Zeng et al. proposed CellFM as a novel large-scale foundation model and demonstrated it outperforms other large language model on multiple single-cell data analysis tasks. This study well summarized the progresses in this rapid-growing field and compared their differences and limits in many aspects including data preprocessing and model architecture. The codes in this study are also well-documented. However, the methodological improvement is unclear and lacks strong benchmarking evidence. Additionally, many figures are in low-quality and not well prepared, and the biological interpretation of the results are not well-justified.

Response: We thank the reviewer for the positive feedback including recognizing that CellFM outperformed other single-cell foundation models in multiple single-cell data analysis tasks and the well-documented code. The main novelty of CellFM can be summarized in three key aspects: (1)

We compiled a standardized dataset of approximately 100 million human cells from public databases, twice as large as those used in the current largest single-species models; (2) Based on the efficient ERetNet architecture, the large dataset allows training CellFM with 800 million parameters, an eight-fold increase in size compared to prior single-species models; (3) CellFM consistently achieved the best performance among all single-cell foundation models across six representative tasks, demonstrating its superior robustness and accuracy. We have also carefully addressed the reviewer's concerns such as methodological improvement, figure quality, and biological interpretation as detailed in the following responses.

Major:

1. The author is aware that there are different approaches of applying large language models to single-cell transcriptome, and there are already many existing tools. They may need to include more of these existing models into benchmarking. As the author argued that the rank-based language model discards the expression values, it would be useful to show such downsides in one of benchmark results. Therefore, the author may include more large language modes in benchmark and provide some extra insights about application of these models.

Response: Thanks for your valuable comments. We have added several recent single-cell foundation models into our benchmarking including scFoundation, GeneCompass, UCE, scELMo, and scBERT. These models represent different approaches to single-cell analysis and we followed recent review article (Nature method, PMID: 39122952) to divide them based on their methodological focus. Specifically, models including CellFM, scFoundation, scELMo, and GeneCompass fall under the value projection category, while UCE, scGPT, and scBERT belong to the value categorization category, and Geneformer is placed in the ordering category.

In our manuscript, we did not intend to critique rank-based models or other methodologies but rather aimed to provide an objective categorization based on the framework described in Nature Methods (PMID: 39122952). Specifically, rank-based models, such as Geneformer, focus solely on gene ordering and do not incorporate expression values, while several models include expression data. This distinction reflects methodological differences rather than any inherent drawbacks, as each approach has its strengths and weaknesses depending on the application. We have revised the relevant descriptions in the "Introduction" section.

We have evaluated our model CellFM alongside these single-cell foundation models across six tasks (Seen in Supplementary Figure S3): cell type annotation, gene function prediction including binary classification and multi-category classification, cell perturbation prediction including unseen gene perturbations prediction and in silico reverse perturbation prediction, and batch effect correction. CellFM consistently achieved the best performance among all single-cell foundation models across six representative tasks, demonstrating its superior robustness and accuracy. The results for each task are presented succinctly below:

For the cell type annotation task, CellFM achieved an average accuracy of 94.02% on inter-datasets, outperforming the top three single-cell foundation models scFoundation, scBERT, and UCE by 2.62%, 4.22%, and 6.34% regarding average accuracy values, respectively. Detailed information on these results can be found in our responses to the following response # 4 and Reviewer #4 (Question 5).

For predicting unseen gene perturbations, CellFM achieved the highest performance, exceeding scFoundation and GeneCompass by 1% and 2.7% regarding average accuracy values, respectively. Detailed information on these results can be found in our responses to Reviewer #1 (Question 2.3). For in silico reverse perturbation prediction, CellFM achieved the highest performance, exceeding the second-best model scGPT 13.6% regarding average hit rates among the top 1-10 across two datasets. Detailed information on these results can be found in our responses to Reviewer #4 (Question 6).

For the batch effect correction task, the deep learning framework used for CellFM, MindSpore, does not provide the Gradient Reversal Layer (GRL) technique. Similarly, other single-cell foundation models with batch effect correction functionality also did not implement GRL. To ensure a fair comparison, we re-evaluated scGPT by removing the GRL loss while retaining its other loss functions. The results demonstrated that CellFM achieved comparable performance to scGPT and outperformed other single-cell foundation models by 4.6%. Detailed information about the results can be found in the following response #8.

For gene function prediction in the context of binary classification, CellFM consistently outperformed all single-cell foundation modes and achieved an average accuracy of 84.9%, which was 5.68% higher than the second-ranked model UCE. For gene function prediction in the context of multi-category classification, CellFM consistently outperformed all single-cell foundation modes and achieved an average AURP value of 73.7%, which was 1.6% higher than the second-ranked model GeneCompass. Detailed information on these results can be found in the following response #3. We didn't compare with scELMo on the gene function prediction tasks since scELMo employed large language models (LLMs) like GPT-3.5 as generators to generate embeddings from metadata descriptions in the training phase, which has included gene function information. We have added the detailed information to "Results" section as follows:

"

We have incorporated several recent single-cell foundation models into our benchmarking, including scFoundation, GeneCompass, UCE, scELMo, scBERT, scGPT, and Geneformer. These models represent different approaches to single-cell analysis, and we have categorized them based on their methodological focus. Specifically, models including CellFM, scFoundation, scELMo, and GeneCompass fall under the value projection category, while UCE, scGPT, and scBERT belong to the value categorization category, and Geneformer is placed in the ordering category.

"

2. One valuable effort the author made in this study is reprocessing and unifying numerous single-cell datasets from multiple sources. However, these re-unified datasets are not available. It is impossible to verify that the author indeed collects this many datasets to train CellFM. Meanwhile, it could be more informative if the author provides more comprehensive summary on their data collection, instead of a mere summary on technique types and organs. For example, does this data collection have overwhelming immune cell population? Were majority of cells from healthy donors? Etc. Such information can be very useful for anyone who has intention to apply this pre-trained model.

Response: Thanks for your valuable comments. We have provided a comprehensive dataset summary (Supplementary Figure S1) and shared a detailed list of the original sources for the training datasets used in CellFM (Supplementary_Metadata_information.xlsx). Concretely, 46.3 million cells were derived from normal donors, the other cells were from diseased donors such as 7.1 million cells from Viral infections donors and 3.5 million cells were from lung cancer donors. Most datasets were sequenced by 10x genomics 3′ containing 66.7 million cells. These cells were from diverse tissues including Brain (17.8 million), Blood (18.9 million), and Lung (6.5 million). We identified the cell types for approximately 70 million cells in the training datasets, as the remaining cells were not classified in their original papers. Specifically, the training dataset includes a diverse range of cell types, such as T cells (19.2 million), mononuclear phagocytes (7.01 million), neurons (6.29 million), and fibroblasts (3 million). We have added these results into the “Results” section as follows:

“

We have provided a comprehensive dataset summary (Supplementary Figure S1) and shared a detailed list of the original sources for the training datasets used in CellFM (Supplementary_Metadata_information.xlsx). Concretely, 46.3 million cells were derived from normal donors, the other cells were from diseased donors such as 7.1 million cells from Viral infections donors and 3.5 million cells were from lung cancer donors. Most datasets were sequenced by 10x genomics 3′ containing 66.7 million cells. These cells were from diverse tissues including Brain (17.8 million), Blood (18.9 million), and Lung (6.5 million). We identified the cell types for approximately 70 million cells in the training datasets, as the remaining cells were not classified in their original papers. Specifically, the training dataset includes a diverse range of cell types, such as T cells (19.2 million), mononuclear phagocytes (7.01 million), neurons (6.29 million), and fibroblasts (3 million).

”

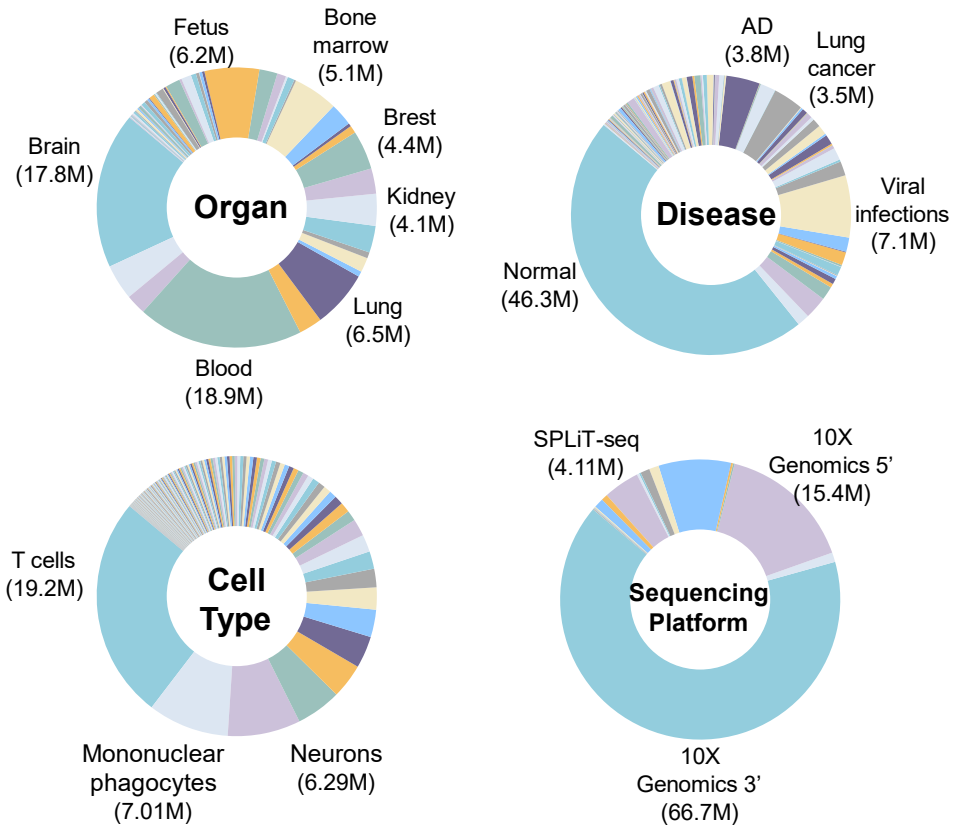


Figure S1. Statistical summaries of the datasets' metadata, including sequencing platform, organ, disease, and cell type, are illustrated through pie charts.

3. The task of gene function prediction seems to be too simple. The task was set in a confined scenario. It is either or not. Indeed, CellFM may outperform other models in binary classification, but it will not say that CellFM will also be accurate in multi-category classification and indicate that CellFM accurately predicts gene function in real world. In real-world application, the model would be asked to predict potential gene functions without knowing a certain category first. Therefore, benchmark here should be framed as a problem of multi-category classification.

Response: Thanks for your valuable suggestions. We have conducted a new multi-class classification analysis for gene function prediction using data from the Gene Ontology (GO). This dataset includes three major categories: biological process (BP), cellular component (CC), and molecular function (MF). Detailed information about the GO dataset can be found in the referenced study (PMID: 36964722). Given the complexity of predicting all gene functions (BP: 1578, CC: 253, MF: 299), we focused our evaluation on the top 10 most frequent functions within each category. This approach ensures a realistic yet manageable benchmark for model comparison. To guarantee fairness across methods, we intersected the gene sets from each foundational single-cell model and maintained consistent training, validation, and test sets.

As shown in Figure 2d, CellFM demonstrated superior average performance, outperforming the top two models GeneCompass and UCE by 1.6% and 1.94% in average AUPR, respectively. Other models such as scFoundation and Geneformer also delivered competitive results with scGPT

achieving an AUPR of 71.3%. We didn't compare with scELMo since scELMo employed large language models (LLMs) like GPT-3.5 as generators to generate embeddings from metadata descriptions in the training phase, which has included gene function information. These new results have been incorporated into the "Results" section of the manuscript as follows:

"

We have conducted a new multi-class classification analysis for gene function prediction using data from the Gene Ontology (GO). This dataset includes three major categories: biological process (BP), cellular component (CC), and molecular function (MF). Detailed information about the GO dataset can be found in the referenced study (PMID: 36964722). Given the complexity of predicting all gene functions (BP: 1578, CC: 253, MF: 299), we focused our evaluation on the top 10 most frequent functions within each category. This approach ensures a realistic yet manageable benchmark for model comparison. To guarantee fairness across methods, we intersected the gene sets from each foundational single-cell model and maintained consistent training, validation, and test sets.

As shown in Figure 2d, CellFM demonstrated superior average performance, outperforming the top two models GeneCompass and UCE by 1.6% and 1.94% in average AUPR, respectively. Other models such as scFoundation and Geneformer also delivered competitive results with scGPT achieving an AUPR of 71.3%. We didn't compare with scELMo since scELMo employed large language models (LLMs) like GPT-3.5 as generators to generate embeddings from metadata descriptions in the training phase, which has included gene function information.
"

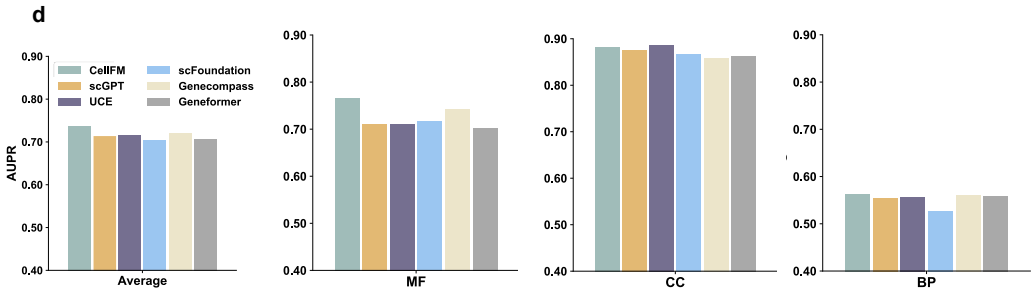


Figure 2d. AUPR values for multiple gene functions predicted by CellFM and other competing single-cell foundation models on GO data, where each gene is annotated with numerous functions. MF: Molecular Function, CC: Cellular Component, and BP: Biological Process.

4. The author may consider including some baseline methods of cell type prediction into results of Figure 4 (See references https://genomebiology.biomedcentral.com/articles/10.1186/s13059-021-02480-2 and https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1795-z). Meanwhile, scBERT has demonstrated outstanding performance in cell type prediction. Does CellFM also achieve superior or at least comparable performance to scBERT in this task?

Response: Thanks for your valuable suggestions. We have added two baseline methods—SVM and scmap—which were also suggested simultaneously in the referenced studies (PMID: 34503564 and PMID: 31500660), along with the single-cell pre-training model scBERT. As shown in Figure 4a-b, CellFM achieved accuracy values that were 1.97%, 26.86%, and 10.2% higher than those of SVM,

scmap, and scBERT, respectively when evaluating cell type annotation on the intra-datasets. On the other hand, CellFM outperformed SVM, scmap, and scBERT on inter-dataset cell annotation tasks, with average accuracy improvements of 1.73%, 20.95%, and 4.22%, respectively (Figure 4e-f). Additionally, we have newly included comparisons with several recent single-cell foundation models, such as scFoundation, GeneCompass, UCE, and scELMo, where CellFM consistently outperformed these models in cell annotation tasks based on average accuracy.

Notably, as suggested by Reviewer #4, we conducted model scaling experiments and found that CellFM with 80 million obtained the best performance for cell type annotation. Cell annotation accuracy improved with an increase in parameters from 8M to 200M but decreased at 800M likely because this cell-level task is relatively simpler without requiring larger parameter size. In contrast, more complex tasks, such as gene function prediction and perturbation prediction, continue to benefit from increases in both model size and training data since the larger model could efficiently capture intricate patterns in high-dimensional data. Interestingly, we found that fine-tuned versions of CellFM-800M achieved superior cell type annotation performance again, indicating the potential power of larger model. Based on these findings, we utilized CellFM-80M for cell annotation and batch effect correction tasks, while CellFM-800M was employed for gene function prediction and cell perturbation prediction tasks. The detailed information and discussion can be found in the response to Reviewer #4 of question #2.

Meanwhile, the fine-tuned performance of scGPT, Geneformer, and GeneCompass aligned with the results reported in their original studies and the benchmark study scEval (PMID: 38464157). Detailed information and further discussion can be found in our response to the question #7. Detailed results have been added to the "Results" section for further clarity:

“

We have also included two baseline methods—SVM and scmap—which were also suggested simultaneously in the referenced studies (PMID: 34503564 and PMID: 31500660), along with the single-cell pre-training model scBERT.

”

“

As shown in Figure 4a-b, CellFM achieved accuracy values that were 1.97%, 26.86%, and 10.2% higher than those of SVM, scmap, and scBERT, respectively when evaluating cell type annotation on the intra-datasets.

”

“

CellFM outperformed SVM, scmap, and scBERT on inter-dataset cell annotation tasks, with average accuracy improvements of 1.73%, 20.95%, and 4.22%, respectively (Figure 4e-f).

”

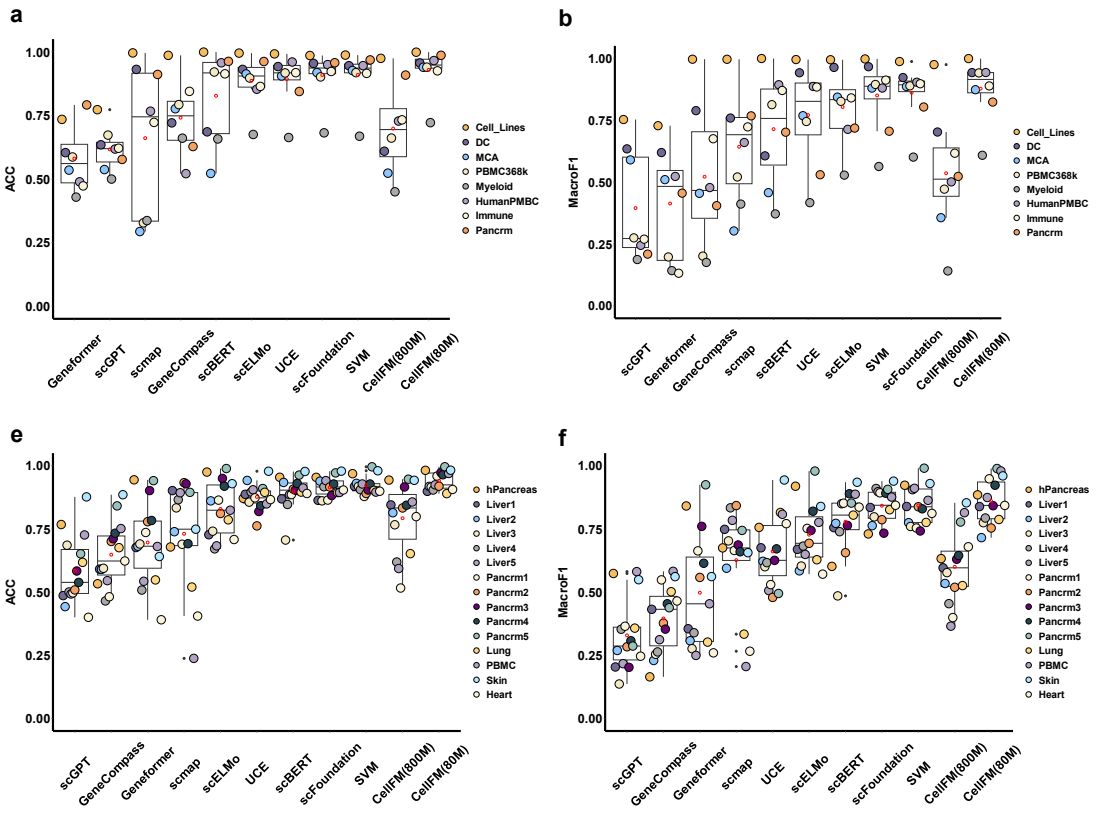


Figure 4. Zero-shot cell type annotation performance of each model. (a) Boxplots summarize the ACC scores for each method on n=8 biologically independent paired intra-datasets. (b) Boxplots summarize the Macro-F1 scores for each method on n=8 biologically independent paired intra-datasets. (e) Boxplots summarize the ACC scores for each method on n=15 biologically independent paired inter-datasets. (f) Boxplots summarize the Macro-F1 scores for each method on n=15 biologically independent paired inter-datasets.

5. In Figure 3d, the author only showed the top1 perturbation prediction. First, it is not clear if the author did proper fine tuning of scGPT with this perturbation data. Second, top1 perturbation prediction seems to be stringent to model. It could be possible that ground truth perturbation is in top 3 or 5 prediction in scGPT. Third, if the author set different random seeds for fine tuning, can CellFM always get the top1 prediction correct?

Response: Thanks for your valuable suggestions. We have followed the tutorial provided by scGPT to perform perturbation prediction. Specifically, for the reverse perturbation prediction task, we followed scGPT's approach by selecting 20 perturbation genes from the Norman dataset to construct perturbation cases for fine-tuning and testing. This combinatorial space consists of 210 one-gene or two-gene perturbation combinations. The subset was selected to maximize the representation of ground truth perturbation data across both training and testing cases, using a random train-test split. Since scGPT did not specify a particular seed for splitting, we utilized the system's default random seed. The resulting dataset contained 47 (about 22%) known perturbations (Fig. 3e), including 33 training cases, 3 validation cases, and 11 test cases, with the remaining perturbation cases left as unseen. Both scGPT and CellFM were evaluated on this newly split dataset, as we were unable to replicate the exact splits used in the original scGPT study. Although our repeated scGPT performance is not exactly the same as those reported in the scGPT study, the overall trend is

consistent.

To assess the robustness of CellFM, we performed additional evaluations by varying the random seeds during the dataset splitting process to generate new training and test cases. As shown in Figure S7b, CellFM outperformed scGPT and GEARS in terms of average hit rate accuracy. Specifically, CellFM achieved an average of 36.3% and 54.5% correctly identified perturbations in the top 3 and top 5 predictions, which were 18.1% and 18.2% higher than scGPT, respectively. To further examine whether the performance of CellFM was affected by random seeds during fine-tuning, we re-evaluated both CellFM, scGPT, and GEARS under different random seeds. The results demonstrated that CellFM consistently maintained comparable top 1 prediction performance, outperforming scGPT by 9.1% (Supplementary Figure S7a).

Finally, we did not include comparisons with other single-cell foundation models for the reverse perturbation prediction task because their original publications did not provide codes or results. Reimplementing these methods would require significant programming effort and time.

These new findings have been added to the “Results” section for further clarity and discussion as follows:

“

We followed scGPT selecting 20 perturbation genes from the Norman dataset to construct perturbation cases for fine-tuning and testing. This combinatorial space consists of 210 one-gene or two-gene perturbation combinations. The subset was selected to maximize the representation of ground truth perturbation data across both training and testing cases, using a random train-test split. Since scGPT did not specify a particular seed for splitting, we utilized the system's default random seed. The resulting dataset contained 47 (22%) known perturbations (Fig. 3d), including 33 training cases, 3 validation cases, and 11 test cases, with the remaining perturbation cases left as unseen. Both scGPT and CellFM were evaluated on this newly split dataset, as we were unable to replicate the exact splits used in the original scGPT study.

”

“

To assess the robustness of CellFM, we performed additional evaluations by varying the random seeds during the dataset splitting process to generate two new groups of training and test cases. As shown in Supplementary Figure S7b, CellFM outperformed scGPT and GEARS in terms of average hit rate accuracy. Specifically, CellFM achieved an average of 36.3% and 54.5% correctly identified perturbations in the top 3 and top 5 predictions, which were 18.1% and 18.2% higher than scGPT, respectively. To further examine whether the performance of CellFM was affected by random seeds during fine-tuning, we re-evaluated both CellFM and scGPT under different random seeds. The results demonstrated that CellFM consistently maintained comparable top 1 prediction performance, outperforming scGPT by 9.1% (Supplementary Figure S7a).

”

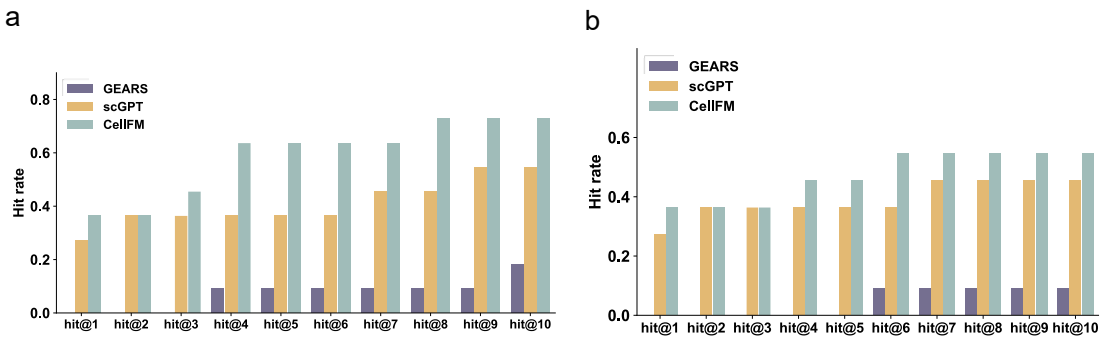


Fig. S7. The Top 1–10 accuracy of CellFM and scGPT after fine-tuning (a) for models with a different random seed and (b) for new test cases generated with a different random seed. The bar graph highlights the hit rates.

6. Prediction result of GEARS in Figure 3e looks so odd. The hit rate remains at 0.1 without increasing as the number of his goes up. It seems that GEARS just do some random prediction here. It leaves that doubt if the author has properly used GEARS in this case.

Response: Thanks for your valuable suggestions. This is consistent with the findings reported in the scGPT study [Nature Methods, PMID: 38409223], where the hit rates for top 1 to top 8 predictions also remained unchanged. This is likely due to the limitations of GEARS as a smaller model, which can't identify perturbation combinations effectively. Unlike CellFM and scGPT, which demonstrated increasing accuracy as the number of hits increased, GEARS failed to accurately predict most test cases, even at higher hit counts. To further validate this observation, we tested GEARS on new test cases generated using different random seeds (Supplementary Figure S7). The results were consistent, showing the same trend of fixed hit rates. We have updated the "Results" section to include additional details and explanations to clarify these findings:

"

This is consistent with the findings reported in the scGPT study [Nature Methods, PMID: 38409223], where the hit rates for top 1 to top 8 predictions also remained unchanged. This is likely due to the limitations of GEARS as a smaller model, which can't identify perturbation combinations effectively. Unlike CellFM and scGPT, which demonstrated increasing accuracy as the number of hits increased, GEARS failed to accurately predict most test cases, even at higher hit counts. To further validate this observation, we tested GEARS on new test cases generated using different random seeds (Supplementary Figure S7). The results were consistent, showing the same trend of fixed hit rates.

"

7. For the task of cell type annotation, I am skeptical on the re-evaluation of scGPT in human pancreas data. The annotation for this data is so accurate in the original report (https://www.nature.com/articles/s41592-024-02201-0). However, the re-evaluation of scGPT in the same dataset is so much worse. Could this be that the author didn't really use scGPT properly? The same question can go to if the poor performance of Geneformer is due to unproper use by the

author. Besides only showing annotation out of from zero-shot learning, they will also benchmark cell annotation after model fine tuning.

Response: Thanks for your valuable suggestions. The results are different from the original paper because this is a zero-shot prediction in our manuscript, where all models were evaluated without task-specific fine-tuning to ensure fair comparison between models. Now, we have added the fine-tuned result by CellFM, scGPT, Geneformer, and GeneCompass. This choice was driven by the significant memory and time requirements of the fine-tuning process. As shown in Supplementary Figure S12, the performance of fine-tuned scGPT aligns closely with the results reported in its original article and the benchmark study scEval (DOI:10.1101/2023.09.08.555192). For instance, fine-tuned scGPT achieved 92.2% cell type annotation accuracy on the human pancreas dataset, comparable to its original study and scEval. Similarly, fine-tuned Geneformer reached 85.3% accuracy on the same dataset, consistent with the performance reported in scEval, although its original paper did not provide results for cell type classification. The fine-tuned GeneCompass achieved comparable results with the fine-tuned CellFM. However, across all inter-datasets, CellFM (800M) demonstrated an average fine-tuning accuracy that was 12.8% and 15.92% higher than scGPT and Geneformer, respectively. The fine-tuned CellFM-80M achieved the best performance, with a similar trend observed in the zero-shot setting. Therefore, for the cell type annotation task, we used the 80M parameter version of CellFM. Detailed information can be found in our response to Reviewer #4, question #2. We have incorporated these findings and detailed explanations into the "Results" section as follows:

“

We also added the fine-tuned result by CellFM, scGPT, GeneCompass, and Geneformer. This choice was driven by the significant memory and time requirements of the fine-tuning process. As shown in Supplementary Figure S12, the performance of fine-tuned scGPT aligns closely with the results reported in its original article and the benchmark study scEval (DOI:10.1101/2023.09.08.555192). For instance, fine-tuned scGPT achieved 92.2% cell type annotation accuracy on the human pancreas dataset, comparable to its original study and scEval. Similarly, fine-tuned Geneformer reached 85.3% accuracy on the same dataset, consistent with the performance reported in scEval, although its original paper did not provide results for cell type classification. The fine-tuned GeneCompass achieved comparable results with the fine-tuned CellFM. However, across all inter-datasets, CellFM (800M) demonstrated an average fine-tuning accuracy that was 12.8% and 15.92% higher than scGPT and Geneformer, respectively.

”

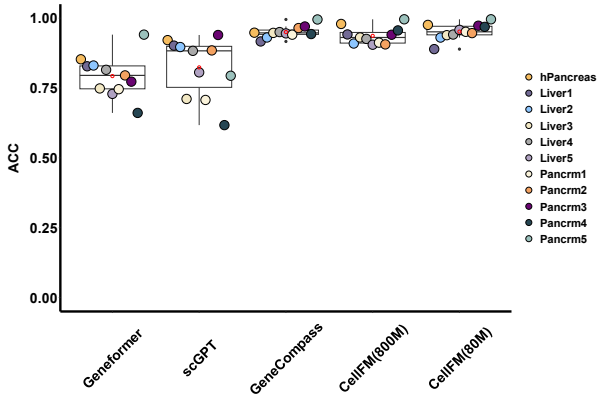


Fig. S12. The cell type annotation performance of Geneformer, scGPT, GeneCompass, and CellFM after fine-tuning with inter-datasets.

8. Another key benchmarking task that is missed in current manuscript is batch correction. How well the cell embedding of CellFM addresses batch effects is not known.

Response: Thanks for your valuable suggestions. We have added the batch effect removal experiments by conducting a comparison among three single-cell foundation models—scELMo, scGPT, and UCE—alongside three non-single-cell foundation batch correction methods used in scGPT: Harmony, Seurat, and scVI. Our comparison included scGPT, scELMo, and UCE as their original studies reported batch correction performance. Following the approach described in scGPT, we performed batch effect correction experiments on the PBMC 10k dataset. The deep learning framework used for CellFM, MindSpore, does not support the Gradient Reversal Layer (GRL) technique. Similarly, other single-cell foundation models with batch effect correction capabilities also lack GRL implementation. To ensure a fair comparison, we re-evaluated scGPT by removing the GRL loss while retaining its other loss functions. As shown in Supplementary Figure S15, CellFM achieved performance comparable to scGPT and Harmony while outperforming other batch correction models. Specifically, our model achieved an AvgBIO score of 78.9% (AvgBIO: average clustering performance; a detailed definition is provided in Supplementary Note 2), which is comparable to scGPT (78.2%) and Harmony (78.4%) and 4.6% and 3.6% higher than UCE and scVI, respectively. These results have been incorporated into the "Results" section as follows:

“

To evaluate the performance of CellFM on the batch effect correction, we have conducted a comparison among three single-cell foundation models—scELMo, scGPT, and UCE—alongside three non-single-cell foundation batch correction methods used in scGPT: Harmony, Seurat, and scVI. Our comparison included scGPT, scELMo, and UCE as their original studies reported batch correction performance. Following the approach described in scGPT, we performed batch effect correction experiments on the PBMC 10k dataset. The deep learning framework used for CellFM, MindSpore, does not support the Gradient Reversal Layer (GRL) technique. Similarly, other single-cell foundation models with batch effect correction capabilities also lack GRL implementation. To ensure a fair comparison, we re-evaluated scGPT by removing the GRL loss while retaining its other loss functions. As shown in Supplementary Figure S15, CellFM achieved performance comparable to scGPT and Harmony while outperforming other batch correction models. Specifically, our model

achieved an AvgBIO score of 78.9% (AvgBIO: average clustering performance; a detailed definition is provided in Supplementary Note 2), which is comparable to scGPT (78.2%) and Harmony (78.4%) and 4.6% and 3.6% higher than UCE and scVI, respectively.
”

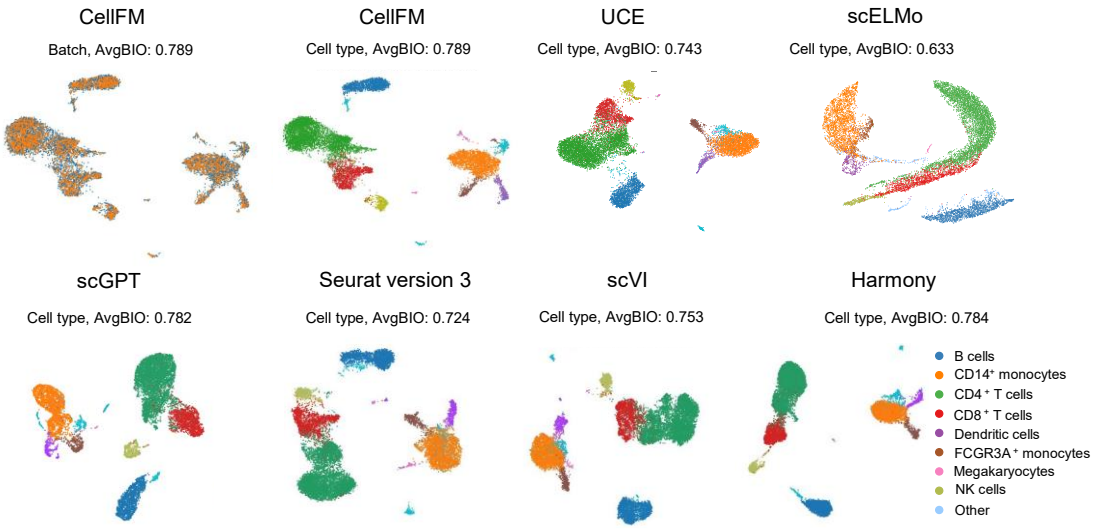


Figure. S15. A benchmark comparison of CellFM, scGPT, scELMo, and UCE on the PBMC 10k dataset for the cell type clustering task. The UMAP plot displays the learned cell embeddings, with colors representing different cell types. The results of methods Seurat version 3, scVI, and Harmony were from the study scGPT.

9. The author presented the gene regulatory network among interleukin family with pre-trained model in Figure 5a. It indeed looks like a solid result with background knowledge in immunology. However, to further validate this finding, the author will have to fine-tune CellFM with extra datasets, say immune cells and other non-immune cells. Will this gene relationship structure maintain in immune cells after fine tuning and differs in non-immune cells?

Response: Thanks for your valuable suggestions. To evaluate the gene relationships captured by our model, we fine-tuned CellFM using 32484 immune cells from the Immune data and about 200000 non-immune cells from human brain data. Specifically, we presented three gene relationship graphs in Supplementary Figure S18: (a) shows the pre-trained CellFM, (b) displays CellFM trained on immune cells, and (c) illustrates CellFM trained on non-immune cells. The results show that the relationships between genes *IL-2, IL-3,* and *IL-4,* observed in the pre-trained CellFM, were preserved when CellFM was fine-tuned on immune cells. However, these relationships were absent when CellFM was trained solely on non-immune cells. Previous studies have shown that *IL-2, IL-3*, and *IL-4* are involved in the JAK/STAT pathway. *IL-2*, secreted primarily by Th1 cells, promotes immune activation, while *IL-4,* secreted by Th2 cells, facilitates anti-inflammatory signaling. Together, they regulate immune responses by mediating cell proliferation, differentiation, and immune homeostasis. Additionally, *IL-3* stimulates the STAT5 pathway, which regulates cell proliferation, differentiation, and anti-apoptotic signaling [1-2]. In summary, CellFM effectively preserved biologically relevant immune gene relationships.

On the other hand, we observed that the gene relationship structures between *IL1RAP* and *IL1R1* were only present in the CellFM trained on immune cells. Previous studies have shown that *IL1RAP* and *IL1R1* form a functional receptor complex that mediates IL-1β signaling. This interaction plays a critical role in neutrophilic inflammation, exacerbating airway inflammation and contributing to worsened pulmonary obstruction. In conclusion, training CellFM on immune cells enabled it to uncover new, biologically meaningful gene relationships [3]. However, the gene relationships among *IL-2, IL-3, IL-4,* as well as between *IL1RAP* and *IL1R1*, were absent when CellFM was trained on non-immune cells. While we have corrected slight mapping errors in the previous code for building Fig 5a, the biological findings keep consistent. We have included these results in the "Results" section as follows:

“

To evaluate the gene relationships captured by our model, we fine-tuned CellFM using 32484 immune cells from Immune data and about 200000 non-immune cells from human brain data. Specifically, we present three gene relationship graphs in Supplementary Figure S18: (a) shows the pre-trained CellFM, (b) displays CellFM trained on immune cells, and (c) illustrates CellFM trained on non-immune cells. The results show that the relationships between genes *IL-2, IL-3,* and *IL-4,* observed in the pre-trained CellFM, were preserved when CellFM was fine-tuned on immune cells. However, these relationships were absent when CellFM was trained solely on non-immune cells. Previous studies have shown that *IL-2, IL-3*, and *IL-4* are involved in the JAK/STAT pathway. IL-2, secreted primarily by Th1 cells, promotes immune activation, while *IL-4,* secreted by Th2 cells, facilitates anti-inflammatory signaling. Together, they regulate immune responses by mediating cell proliferation, differentiation, and immune homeostasis. Additionally, *IL-3* stimulates the *STAT5* pathway, which regulates cell proliferation, differentiation, and anti-apoptotic signaling [1-2]. In summary, CellFM effectively preserved biologically relevant immune gene relationships.

On the other hand, we observed that the gene relationship structures between *IL1RAP* and *IL1R1* were only present in the CellFM trained on immune cells. Previous studies have shown that *IL1RAP* and *IL1R1* form a functional receptor complex that mediates IL-1β signaling [3]. This interaction plays a critical role in neutrophilic inflammation, exacerbating airway inflammation and contributing to worsened pulmonary obstruction. In conclusion, training CellFM on immune cells enabled it to uncover new, biologically meaningful gene relationships. However, the gene relationships among *IL-2, IL-3, IL-4,* as well as between *IL1RAP* and *IL1R1*, were absent when CellFM was trained on non-immune cells.

”

Reference:

[1] Lu Q, Hu Y, Nabi F, Li Z, Janyaro H, Zhu W, et al. Effect of Penthorum Chinense Pursh compound on AFB1-induced immune imbalance via JAK/STAT signaling pathway in spleen of broiler chicken. Vet Sci. 2023;10(8):521. doi:10.3390/vetsci10080521.

[2] Gündogdu MS, Liu H, Metzdorf D, et al. The haematopoietic GTPase RhoH modulates IL3 signalling through regulation of STAT activity and IL3 receptor expression. Mol Cancer. 2010;9:225. doi:10.1186/1476-4598-9-225.

[3] Evans MD, Esnault S, Denlinger LC, Jarjour NN. Sputum cell IL-1 receptor expression level is a marker of airway neutrophilia and airflow obstruction in asthmatic patients. J Allergy Clin Immunol. 2018;142(2):415-423. doi:10.1016/j.jaci.2017.09.035.

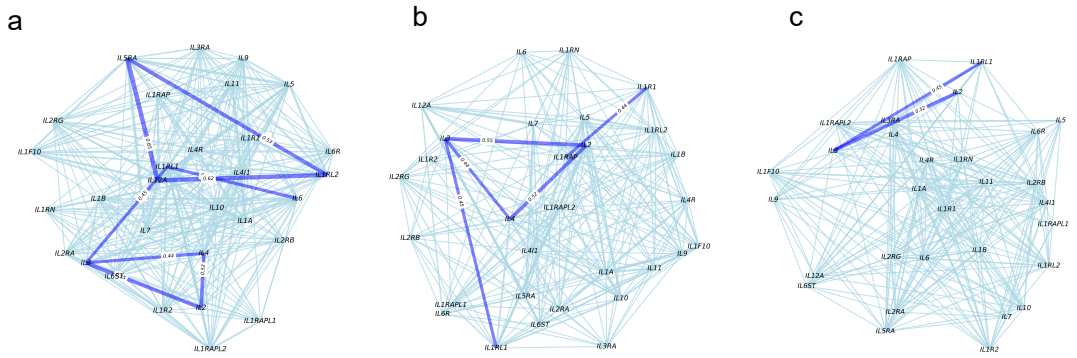


Figure S18. **Gene-Gene Relationships Unveiled by CellFM.** (a) A gene cluster comprising *IL-2, IL-3,* and *IL-4* was identified using the cosine similarity of gene embeddings generated by the zero-shot CellFM model. (b) When CellFM was fine-tuned with immune cell data, the relationships among *IL-2, IL-3,* and *IL-4* were preserved, and new relationships involving *IL1RAP* and *IL1R1* were discovered. (c) In contrast, when CellFM was trained on non-immune cell data, the relationships among *IL-2* and *IL-3* as well as those involving *IL1RAP* and *IL1R1* were no longer detected.

10. The author citied study from Zou et al. (https://academic.oup.com/nar/article/50/W1/W175/6553688) to back up their finding on top 20 affected genes by SPI1 suppression. Zou et al. did not specifically provide results on SPI1 suppression. It is not clear how the author drew the conclusion from Zou et al. and tried to back up the identification by CellFM. They will need to have clarification in results presented in Figure 5d and e. Meanwhile, it is confusing why SPI1 in column has connectivity while the SPI1 in row doesn't in Figure 5d. Additionally, as suggested in study on SPI1 (PMID: 35871293), SPI1 represses ALAS2. Is SPI1 repressing those top 20 genes or is it possible to distinguish repression and activation based on the attention map?

Response: Thanks for your valuable comments. We have followed the previous reference scGPT [Nature method, PMID: 38409223] to use the "suppression" term, since these are perturbation experiments in the Adamson dataset achieved through gene knockout. This refers to the nature of *SPI1's* perturbation, not to *SPI1* suppressing other genes. To avoid confusion, we have revised the manuscript to explicitly state this distinction. Concretely, we modeled the effects of perturbing *SPI1* by providing the model with the control cell expression profiles (non-perturbed) and explicitly indicating *SPI1* as the perturbed gene. The model then predicted how *SPI1* perturbation would affect the expression of other genes. Using the attention mechanism, we identified the 20 genes most influenced by *SPI1*. However, we do not claim that *SPI1* suppresses or activates these genes. We only observed that these 20 genes have known relationships with *SPI1*, as supported by evidence from the ChIP-Atlas database. The exact regulatory nature of these relationships—whether *SPI1* suppresses, activates, or interacts differently with these genes—cannot be conclusively determined from our current analysis.

On the other hand, the difference of *SPI1* in row and column was that the column of *SPI1* represented the attention score between the perturbated *SPI1* gene and other genes while the row of *SPI1* represented the attention score between the other genes to the *SPI1* gene. In the attention mechanism, the relationship is not symmetric because the attention scores are calculated based on the dot product of queries and keys, where the roles of queries and keys are distinct, leading to different scores for A attending to B versus B attending to A. We showed that the attention from A to B is not the same as from B to A due to the different roles of queries and keys as follows:

$$Attention(A \rightarrow B) = Softmax\left(\frac{Q_A K_B^T}{\sqrt{d_k}}\right)V_B$$

$$Attention(B \rightarrow A) = Softmax\left(\frac{Q_B K_A^T}{\sqrt{d_k}}\right)V_A$$

Where: Q_A and K_A are the query and key matrices for sequence A. Q_B and K_B are the query and key matrices for sequence B. V_A and V_B are the value matrices for sequences A and B, respectively. d_k is the dimension of the key vector. Additionally, it is difficult to distinguish repression and activation based on the attention map since the attention map used in our model only contained the scalar values for attention scores. We have added the detailed information into "Results" section and Supplementary Note 3 as follows:

"

Concretely, we modeled the effects of perturbing *SPI1* by providing the model with the control cell expression profiles (non-perturbed) and explicitly indicating *SPI1* as the perturbed gene. The model then predicted how *SPI1* perturbation would affect the expression of other genes. Using the attention mechanism, we identified the 20 genes most influenced by *SPI1*.
"

"

On the other hand, the difference of *SPI1* in row and column was that the column of *SPI1* represented the attention score between the perturbated *SPI1* gene and other genes while the row of *SPI1* represented the attention score between the other genes to the *SPI1* gene. In the attention mechanism, the relationship is not symmetric because the attention scores are calculated based on the dot product of queries and keys, where the roles of queries and keys are distinct, leading to different scores for A attending to B versus B attending to A. We showed that the attention from A to B is not the same as from B to A due to the different roles of queries and keys as follows:

$$Attention(A \rightarrow B) = Softmax\left(\frac{Q_A K_B^T}{\sqrt{d_k}}\right)V_B$$

$$Attention(B \rightarrow A) = Softmax\left(\frac{Q_B K_A^T}{\sqrt{d_k}}\right)V_A$$

Where: Q_A and K_A are the query and key matrices for sequence A. Q_B and K_B are the query and key matrices for sequence B. V_A and V_B are the value matrices for sequences A and B, respectively. d_k is the dimension of the key vector. Additionally, it is difficult to distinguish

repression and activation based on the attention map since the attention map used in our model only contained the scalar values for attention scores.

”

11. It’s unclear that whether the attention map in Figure 5c was calculated in the same way as described in the scGPT paper (https://www.nature.com/articles/s41592-024-02201-0). Compared to Figure 6b and 6c in scGPT’s paper, why are the values of most of the gene pairs near zero in Figure 5d heatmap? Some genes, e.g., CCNB2, CCNA2, TOP2A, MKI67, KPNA2 and CENPF have highly relevant biological functions.

Response: Thanks for your valuable comments. The attention map shown in Figure 5c was generated using the standard methodology employed in the classical transformer architectures. The observation that most gene pairs in the heatmap of Figure 5d have values near zero is due to our normalization strategy. Specifically, we ranked all genes in the perturbed dataset from Adamson based on their attention scores and recalculated the connection scores by dividing these ranked scores by the total number of genes. As a result, genes with lower ranks exhibit connection scores that approach zero. Our approach differs from scGPT as we use a more traditional transformer-based method by directly leveraging raw attention scores, sorting and normalizing them to identify the most-influenced genes. In contrast, scGPT introduces additional steps, such as integrating control attention and performing extra normalization. We chose our approach to stay closer to the standard interpretation of transformer attention, ensuring simplicity and focusing directly on the relationships captured by the model.

Although our model, CellFM, efficiently captures perturbation-related genes for multiple perturbed genes (Supplementary Figures S19–20; details provided in the response to Reviewer #1, Question 2.9), the current attention map fails to capture the gene relationships among *CCNB2*, *CCNA2*, *TOP2A*, *MKI67*, *KPNA2*, and *CENPF* due to their low attention scores. In the future, we plan to explore new explainability techniques to address this limitation. The detailed information has been added into “Result” and “Discussion” sections as follows:

“

The observation that most gene pairs in the heatmap of Figure 5d have values near zero is due to our normalization strategy. Specifically, we ranked all genes in the perturbed dataset from Adamson based on their attention scores and recalculated the connection scores by dividing these ranked scores by the total number of genes. As a result, genes with lower ranks exhibit connection scores that approach zero.

”

“

The attention map in CellFM was limited in capturing gene relationships related to static or global biological knowledge. In the future, we will explore new explainability techniques to overcome this challenge.

”

12. In Figure 5f, the KEGG pathway analysis for the top 20 most influenced genes doesn't make sense, most of the enriched pathways only have 2 genes, what's the background gene set used and what's the biological interpretation?

Response: Thanks for your valuable comments. For pathway enrichment analysis, we included all genes from the KEGG pathway database as the background gene set to provide a broader biological context. This approach was intentional to highlight the pathways enriched by the top 20 genes most influenced by *SPI1* perturbation. Since the input for enrichment analysis included only these 20 genes and *SPI1*, it is expected that some pathways are enriched by one or two genes, leading to pathways with smaller gene counts. Our aim was to determine whether *SPI1* and the most affected genes might jointly participate in specific biological pathways, potentially reflecting downstream regulatory effects of *SPI1* perturbation. The enriched pathways presented in Figure 5f emphasize those in which *SPI1* and its perturbed genes may exert a coordinated influence on important biological processes. This suggests that *SPI1* could interact with specific genes to drive regulatory mechanisms within these pathways. In summary, this analysis provides potential insights into the regulatory networks and pathways that might be influenced by *SPI1* perturbation, suggesting its role in modulating genes associated with key biological functions.

To substantiate the biological interpretation provided above, references can be gathered from existing literature, particularly focusing on *SPI1* and its regulatory network, as well as the pathways identified in the enrichment analysis. Here's a structured approach to support each pathway with references:

(1) Human T-cell leukemia virus 1 infection (HTLV-1)

In this analysis, the enrichment of *SPI1*, *CCNB2*, and *CCNA2* in the "Human T-cell leukemia virus 1 infection (HTLV-1)" pathway (hsa05166) holds particular significance. First, the *SPI1* gene plays a critical role in leukemia stem cell (LSC) self-renewal [1]. In HTLV-1 infection, *SPI1* activation may promote the expansion of leukemia stem cells, enhancing their self-renewal capacity and accelerating leukemia progression. This suggests that *SPI1* is not only active in the early stages of viral infection but also serves as a significant regulatory factor in leukemia transformation following infection. Second, the *CCNA2* gene primarily regulates cell cycle transitions. HTLV-1-encoded oncoproteins Tax and HBZ disrupt *CCNA2* expression, causing cell cycle dysregulation and increasing cell immortalization potential [2]. In HAM/TSP patients, observed downregulation of *CCNA2* may be due to IRF-1 suppression, which could prevent transformation into ATLL. These findings not only reveal the potential downstream regulatory impacts of SPI1 perturbation but also provide insight into the coordinated role that *SPI1* and *CCNA2* might play in pathways associated with leukemia progression.

(2) Acute myeloid leukemia (AML)

In this analysis, the enrichment of *SPI1* and *CCNA2* in the "Acute myeloid leukemia (AML)" pathway (hsa05221) holds particular significance. First, SPI1 is crucial for normal blood cell differentiation, and its dysregulation in AML leads to uncontrolled cell growth and blocked differentiation, promoting leukemia development [3]. In AML, *CCNA2* is abnormally overexpressed in certain subtypes, particularly in therapy-related AML (t-AML) [4]. This

overexpression, along with other cell cycle regulation genes like *CCNE2* and *CDC2*, is linked to poor prognosis and is often associated with chromosomal deletions of 5 and 7, which correlate with lower survival rates. These findings highlight the potential downstream regulatory effects of SPI1 perturbation and underscore the coordinated role of *SPI1* and *CCNA2* in pathways related to leukemia progression. This suggests that *SPI1* and *CCNA2* may work together to drive critical biological processes involved in leukemia development. We have added these results into Section "Results" as follows:

"

For pathway enrichment analysis, we included all genes from the KEGG pathway database as the background gene set to provide a broader biological context. This approach was intentional to highlight the pathways enriched by the top 20 genes most influenced by *SPI1* perturbation. Since the input for enrichment analysis included only these 20 genes and *SPI1*, it is expected that some pathways are enriched by one or two genes, leading to pathways with smaller gene counts. The enriched pathways presented in Figure 5f emphasize those in which *SPI1* and its perturbed genes may exert a coordinated influence on important biological processes. For instance, the Human T-cell leukemia virus 1 infection (HTLV-1) contained *SPI1*, *CCNB2*, and *CCNA2* holding particular significance. First, the SPI1 gene plays a critical role in leukemia stem cell (LSC) self-renewal [1]. In HTLV-1 infection, SPI1 activation may promote the expansion of leukemia stem cells, enhancing their self-renewal capacity and accelerating leukemia progression. Second, the *CCNA2* gene primarily regulates cell cycle transitions. HTLV-1-encoded oncoproteins Tax and HBZ disrupt *CCNA2* expression, causing cell cycle dysregulation and increasing cell immortalization potential [2]. In HAM/TSP patients, observed downregulation of *CCNA2* may be due to IRF-1 suppression, which could prevent transformation into ATLL. For the other pathway Acute myeloid leukemia (AML) enriched *SPI1* and *CCNA2* holding particular significance. First, SPI1 is crucial for normal blood cell differentiation, and its dysregulation in AML leads to uncontrolled cell growth and blocked differentiation, promoting leukemia development [3]. In AML, *CCNA2* is abnormally overexpressed in certain subtypes, particularly in therapy-related AML (t-AML) [4]. This overexpression, along with other cell cycle regulation genes like *CCNE2* and *CDC2*, is linked to poor prognosis and is often associated with chromosomal deletions of 5 and 7, which correlate with lower survival rates. This suggests that *SPI1* and *CCNA2* may work together to drive critical biological processes involved in leukemia development.
"

References:

[1] Hegde, S., Hankey, P., & Paulson, R. F. (2012). Self-renewal of leukemia stem cells in Friend virus-induced erythroleukemia requires proviral insertional activation of Spi1 and hedgehog signaling but not mutation of p53. Stem Cells, 30(2), 121–130.

[2] Saffari, M., Rahimzada, M., Mirhosseini, A., Ahmadi Ghezaldasht, S., Valizadeh, N., Moshfegh, M., Moradi, M.-T., & Rezaee, S. A. (2022). Coevolution of HTLV-1-HBZ, Tax, and proviral load with host IRF-1 and CCNA-2 in HAM/TSP patients. Infection, Genetics and Evolution, 103, 105337.

[3] Wang, X., Jin, P., Zhang, Y., & Wang, K. (2021). CircSPI1 acts as an oncogene in acute myeloid leukemia through antagonizing SPI1 and interacting with microRNAs. Cell Death & Disease, 12(4), 297.

[4] Qian, Z., Joslin, J. M., Tennant, T. R., Reshmi, S. C., Young, D. J., Stoddart, A., Larson, R. A., & Le Beau, M. M. (2010). Cytogenetic and genetic pathways in therapy-related acute myeloid leukemia. Chemico-Biological Interactions, 184(1-2), 50-57.

13. For data preprocessing, the existing single-cell foundation models use either raw count or just the rank of the gene. But in this study, the author did log1p normalization. What's the influence of log1p normalization on the model performance? After log1p normalization, the variance of genes is correlated with the mean values of genes and is likely influenced by the batch effects. As in CellFM, the initial representation of each cell is the l_max genes with largest expression values, will log1p normalization induce any bias? Could other normalization methods, e.g., scTransform (PMID: 31870423) and Pearson residuals (PMID: 34488842), which correct the variance-mean bias, improve the performance of single-cell foundation model? The author should give better justification on this step of data preprocessing.

Response: Thanks for your valuable comments. We have employed log1p normalization in CellFM due to its widespread use in single-cell RNA-seq data preprocessing and its ability to address key challenges, including the dynamic range of expression values, skewed distributions, and technical noise. Log1p normalization compresses the wide dynamic range of gene expression values, ensuring a more balanced representation of genes; it also reduces the impact of outliers by transforming skewed distributions into a more normal-like form and stabilizes low-expression values to mitigate technical noise, such as dropouts. While log1p normalization does induce a correlation between gene variance and mean expression levels, this characteristic is commonly observed in single-cell RNA-seq analysis and is generally acceptable as a trade-off for improved data robustness (PMID: 31530958). Regarding potential bias in selecting the top l_{max} genes after log1p normalization, we acknowledge that this approach might introduce slight variations in gene rankings. However, these effects are minimal compared to the overall benefits of stabilization and noise reduction, as demonstrated by the robust performance of CellFM. Apart from CellFM, the single-cell foundation model GeneCompass also employs the log1p normalization step, as noted in their preprocessing pipeline (https://github.com/xCompass-AI/GeneCompass/blob/main/preprocess/preprocess.py).

As suggested by the reviewer, we have tested Pearson residuals (PMID: 34488842), which corrects the variance-mean bias. Due to computational constraints, we performed these experiments using a smaller CellFM model with 80 million parameters instead of the original CellFM with 800 million parameters. As shown in Supplementary Figure S17a, the performance with Pearson residuals was slightly lower than with log1p normalization used in CellFM. These findings suggest that while Pearson residuals addresses variance-mean bias explicitly, it does not yield substantial improvements in CellFM's performance for the cell type annotation task evaluated. We have incorporated this detailed justification in "Results" section as follows:

"

To evaluate whether alternative normalization methods might further improve CellFM's performance, we tested Pearson residuals (PMID: 34488842), which corrects the variance-mean bias. Due to computational constraints, we performed these experiments using a smaller CellFM

model with 80 million parameters instead of the original 800 million. As shown in Supplementary Figure S17, the performance with Pearson residuals was slightly lower than with log1p normalization used in CellFM. These findings suggest that while Pearson residuals addresses variance-mean bias explicitly, it does not yield substantial improvements in CellFM's performance for the cell type annotation task evaluated.

”

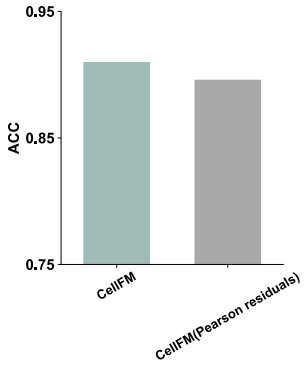


Fig. S17a. The ablation study of CellFM. (a) Performance of CellFM using datasets normalized by log1p (default) or Pearson residuals.

14. In implementation detail, the author used 100 million cells and did pre-training over 2 epochs. What’s the reasoning on training a model for over 2 epochs? The author needs to provide more quantitative justification on model training. In current manuscript, it is unknown if model was underfit or overfit within 2 epochs, considering the training size is enormous.

Response: Thanks for your valuable comments. The decision to train the CellFM model for two epochs was informed by standard practices in large-scale model training [1-2], where rapid convergence is typically observed within the initial epochs. Previously, when training the 800 million parameter CellFM model, we observed a similar trend, with the loss decreasing significantly within the first epoch and stabilizing in the second. Unfortunately, detailed logs from that training run were not retained. To validate this phenomenon, we conducted a new experiment using the 80-million-parameter version of CellFM on all training datasets. The results confirmed the same pattern: the loss dropped sharply—from 8 to below 1—during the first epoch, with only minimal changes in the second epoch. This behavior reflects the typical convergence dynamics of large-scale models and supports our choice to limit training to two epochs to balance efficiency and performance. We acknowledge the importance of providing more detailed quantitative justification and will incorporate additional monitoring and reporting in future studies to ensure greater transparency in our training process. The detailed information has been added into “Results” section as follows:

“

The decision to train the CellFM model for two epochs was informed by standard practices in large-scale model training [1-2], where rapid convergence is typically observed within the initial epochs. To validate this convergence of CellFM, we conducted the experiment using the 80-million-parameter version of CellFM on all training datasets. The results confirmed the same pattern: the loss dropped sharply—from 8 to below 1—during the first epoch, with only minimal changes in the second epoch. This behavior reflects the typical convergence dynamics of large-scale models and supports our choice to limit training to two epochs to balance efficiency and performance.

”

[1] Touvron H, Lavril T, Izacard G, et al. Llama: Open and efficient foundation language models[J]. arXiv preprint arXiv:2302.13971, 2023.
[2] Touvron H, Martin L, Stone K, et al. Llama 2: Open foundation and fine-tuned chat models[J]. arXiv preprint arXiv:2307.09288, 2023.

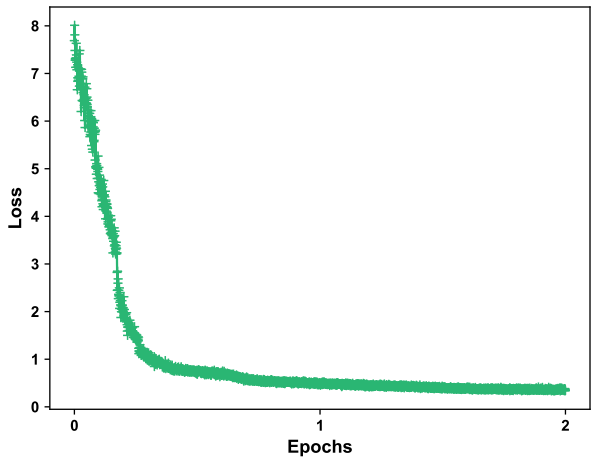


Fig. S21. The training loss of CellFM with 80 million parameters on all training datasets.

15. There are also lack of some clarity in method section. This study lacks clarity and strong benchmarking evidence regarding which parts of the model mainly contributed to its improved performance, especially in relation to RetNet. Additionally, it's unclear whether the inclusion of the second term in the loss function, L_cls, improved the performance or not. Regarding computational efficiency, although the use of MHA improved the computational complexity from O((l_max)^2 * d) to O(l_max * d^2), given that l_max is 2048 and d is 1536, the improvement is minimal.

Response: Thanks for your valuable comments. We evaluated two key modifications to RetNet: replacing the traditional feedforward network with a gated bilinear network, and substituting the pre-layer LayerNorm with the DeepNorm layer normalization technique. To further assess the improvements in CellFM, we also benchmarked it against the classic Transformer model. All ablation experiments were conducted using a newly trained CellFM model with 80 million parameters, with a focus on the cell type annotation task, due to limitations in time and computational resources. As shown in Supplementary Figure S17, removing the Simple Gated Linear Unit and DeepNorm resulted in decreases of 0.8% and 0.9%, respectively. Additionally, the removal of the L_{cls} loss led to a slight drop (0.4%) in performance. When compared to the classic Transformer, CellFM demonstrated a 1.2% improvement. In conclusion, these modifications collectively contributed to the robust performance of CellFM, as evidenced by the benchmarking results.

Regarding computational complexity, we apologize for the typographical errors. In fact, the implementation of Gated Multi-head Attention (MHA) improved the computational complexity from $O(l_{max}^2 d)$ to $O(l_{max}\frac{d^2}{h})$, where d was set to 1536 and the number of attention heads was set to 48. Consequently, the actual computational complexity of CellFM is $O(2048 \times \frac{1536^2}{48})$, which is

smaller than $O(2048^2 \times 1536)$. The formula derivation can be found in the Supplementary Note 1. We have updated this information in the "Results" section as follows:

"

We evaluated two key modifications to RetNet: replacing the traditional feedforward network with a gated bilinear network, and substituting the pre-layer LayerNorm with the DeepNorm layer normalization technique. To further assess the improvements in CellFM, we also benchmarked it against the classic Transformer model. All ablation experiments were conducted using a newly trained CellFM model with 80 million parameters, with a focus on the cell type annotation task, due to limitations in time and computational resources. As shown in Supplementary Figure S17, removing the Simple Gated Linear Unit and DeepNorm resulted in decreases of 0.8% and 0.9%, respectively. Additionally, the removal of the L_{cls} loss led to a slight drop (0.4%) in performance. When compared to the classic Transformer, CellFM demonstrated a 1.2% improvement. In conclusion, these modifications collectively contributed to the robust performance of CellFM, as evidenced by the benchmarking results.

"

"

The implementation of Gated Multi-head Attention (MHA) in CellFM improved the computational complexity from $O(l_{max}^2 d)$ to $O(l_{max} \frac{d^2}{h})$, where d was set to 1536 and the number of attention heads was set to 48. Consequently, the actual computational complexity of CellFM is $O(2048 \times \frac{1536^2}{48})$, which is smaller than $O(2048^2 \times 1536)$. The formula derivation can be found in the Supplementary Note 1.

"

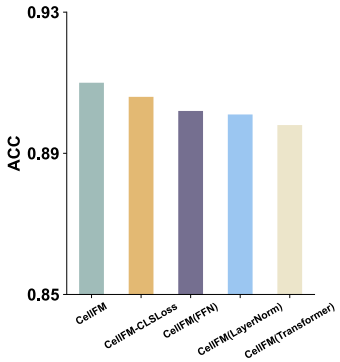


Fig. S17b. **The ablation study of CellFM.** Performance of CellFM with the removal of ClS_{loss}, the gated bilinear network, and DeepNorm, as well as when replacing the backbone with a Transformer.

16. Finally, is it appropriate to not include long-noncoding RNAs in the scRNA-seq data? Since the function of most of the long-noncoding RNAs are unknown, it would be helpful if the single-cell foundation model could prioritize and predict the biological functions of long-noncoding RNAs in a cell type-specific manner?

Response: Thanks for your valuable suggestions. In our study, we included long non-coding RNAs

(lncRNAs) in the original training of CellFM. To identify cell-type-specific lncRNAs, we leveraged the attention mechanism within CellFM to highlight lncRNAs that are most relevant to specific cell types. Specifically, CellFM uses the CLS token as a global representation for cell classification. By analyzing the attention scores between the CLS token and individual genes, we selected the top 100 genes with the highest attention scores for each cell type. Among these top genes, we identified lncRNAs that contribute significantly to the classification of specific cell types.

To demonstrate the capability of CellFM in identifying biologically relevant lncRNAs, we trained the model on PBMC data and identified cell-type-specific lncRNAs. Here, we presented two examples illustrating the CellFM can effectively identify key lncRNAs associated with distinct cell types. For cell type CD14+ Monocytes (CD14+ Mono), the lncRNA *HOTAIRM1* identified by CellFM as one of the top genes with the highest attention scores for this cell type. *HOTAIRM1* was a lncRNA specifically expressed in myeloid cells, which played a pivotal role in monocyte differentiation and identity maintenance by regulating pathways involving miR-3960 and *HOXA1*. Silencing *HOTAIRM1* results in reduced expression of CD14 and other monocyte markers, underscoring its critical regulatory function. Including *HOTAIRM1* in scRNA-seq annotation enhances the accurate identification of CD14+ Mono cells by addressing regulatory mechanisms beyond protein-coding genes. This highlights the value of incorporating lncRNAs in cell type classification [1]. For the cell type CD4+ T Cells, the lncRNA *miR-29a* was also identified as cell-type-specific with strong biological significance. *miR-29a* plays a key role in CD4+ T cell differentiation, particularly in the development of Th1 cells, by regulating markers such as T-bet and IFN-γ and interacting with the Notch signaling pathway. This positions *miR-29a* as a central regulator of Th1 cell differentiation, reflecting their functional state. Consequently, *miR-29a* serves not only as a marker of Th1 cell differentiation but also as a valuable biomarker for functional studies and immune profiling [2]. These examples underscore the utility of CellFM in prioritizing biologically significant lncRNAs, providing insights into their cell-type-specific functions and offering new opportunities for scRNA-seq-based analyses. We have provided the detailed tutorial for identifying cell-type-specific lncRNAs (https://github.com/biomed-AI/CellFM/blob/main/tutorials/IdentifyingCelltypelncRNAs.ipynb) and added these findings to the "Results" section as follows:

"

In our study, we included long non-coding RNAs (lncRNAs) in the original training of CellFM. To identify cell-type-specific lncRNAs, we leveraged the attention mechanism within CellFM to highlight lncRNAs that are most relevant to specific cell types. Specifically, CellFM uses the CLS token as a global representation for cell classification. By analyzing the attention scores between the CLS token and individual genes, we selected the top 100 genes with the highest attention scores for each cell type. Among these top genes, we identified lncRNAs that contribute significantly to the classification of specific cell types.

To demonstrate the capability of CellFM in identifying biologically relevant lncRNAs, we trained the model on PBMC data and identified cell-type-specific lncRNAs. Here, we presented two examples illustrating the CellFM can effectively identify key lncRNAs associated with distinct cell types. For cell type CD14+ Monocytes (CD14+ Mono), the lncRNA *HOTAIRM1* identified by

CellFM as one of the top genes with the highest attention scores for this cell type. *HOTAIRM1* was a lncRNA specifically expressed in myeloid cells, which played a pivotal role in monocyte differentiation and identity maintenance by regulating pathways involving *miR-3960* and *HOXA1*. Silencing *HOTAIRM1* results in reduced expression of CD14 and other monocyte markers, underscoring its critical regulatory function. Including *HOTAIRM1* in scRNA-seq annotation enhances the accurate identification of CD14+ Mono cells by addressing regulatory mechanisms beyond protein-coding genes. This highlights the value of incorporating lncRNAs in cell type classification [1]. For the cell type CD4+ T Cells, the lncRNA *miR-29a* was also identified as cell-type-specific with strong biological significance. *miR-29a* plays a key role in CD4+ T cell differentiation, particularly in the development of Th1 cells, by regulating markers such as T-bet and IFN-γ and interacting with the Notch signaling pathway. This positions *miR-29a* as a central regulator of Th1 cell differentiation, reflecting their functional state. Consequently, *miR-29a* serves not only as a marker of Th1 cell differentiation but also as a valuable biomarker for functional studies and immune profiling [2]. These examples underscore the utility of CellFM in prioritizing biologically significant lncRNAs, providing insights into their cell-type-specific functions and offering new opportunities for scRNA-seq-based analyses.
”

[1] Xin, J., Li, J., Yang, R, et.al. Downregulation of long noncoding RNA HOTAIRM1 promotes monocyte/dendritic cell differentiation through competitively binding to endogenous miR-3960. (2017) OncoTargets and Therapy, 10, 1307–1315.
[2] Chandiran, K., Lawlor, R., Pannuti, A, et.al.. Notch1 primes CD4 T cells for T helper type I differentiation through its early effects on miR-29. (2018) Journal Name, Volume(Issue), page numbers.

Minor:

1. Please be aware of a proper format for writing gene names and protein names.

Response: Thanks for your valuable suggestions. We carefully revised the manuscript to ensure proper formatting for gene and protein names. Specifically, we italicized all gene names and ensured capitalization. Protein names, on the other hand, were not italicized and were written in standard capitalization (e.g., TP53).

2. In Figure 2c, “CellAI” should be “CellFM”, and it’s more appropriate to order the three methods using the order in Figure 2a and Figure 2b.

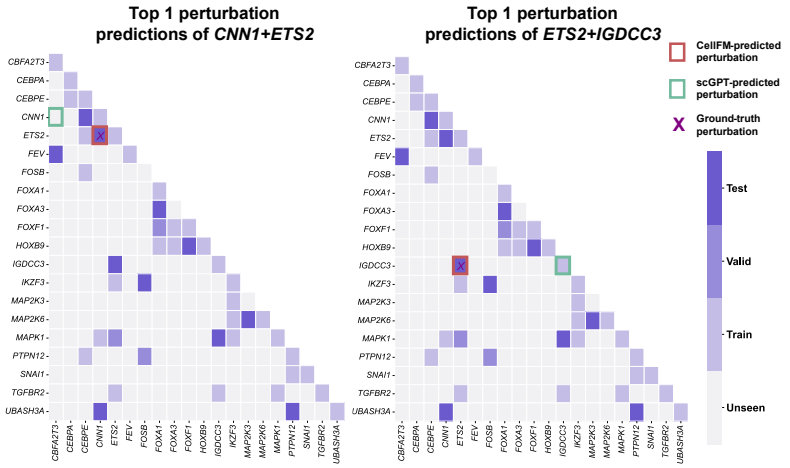
Response: Thanks for your valuable suggestions. We have updated the typographical errors and ordered the methods in Figure 2c following the order in Figure 2a.

3. In Figure 3a and 3b, the y-axis range should start at 0 to objectively show the differences.

Response: Thanks for your valuable suggestions. We have updated Figure 3c (Figures 3a and 3b in our initial submission) by setting the start of the y-axis to zero.

4. In Figure 3d and Figure S2, the gene name “UBASH3A” is not fully displayed.

Response: Thanks for your valuable suggestions. We have updated Figure 3e (Figure 3d in our initial submission) and Figure S6 (Fig S2 in initial submission) and fully displayed the gene *UBASH3A*.



5. In both Figure 4a and Figure 4b, the test dataset name “Pancrm” is defined in Table S4, but not in the main text when this figure is mentioned.

Response: Thanks for your valuable suggestions. Figures 4a and 4b presented the cell annotation performance of each model on intra-datasets, with "Pancrm" being one of these datasets. We did not report specific performance metrics for each dataset, opting instead to provide average results. In both intra- and inter-dataset settings, we used the Pancrm dataset, which contains 5 batches of data. For the intra-dataset setting, we divided each batch into training and testing sets using a 7:3 ratio. In the inter-dataset setting, we used a leave-one-batch-out strategy, where each batch was sequentially used as the test set, and the remaining batches served as the training set.

6. Figure 5e appears in low resolution due to snapshotting, the borders are not cleanly removed, and the font type is different from that used in other figures.

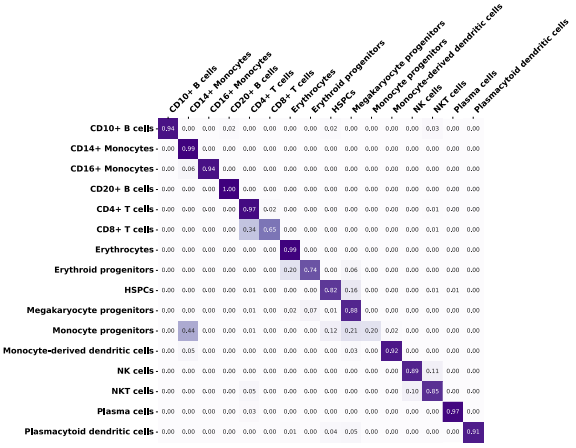
Response: Thanks for your valuable comments. We have updated Figure 5e using the font type Arial and provided a high-quality image without the borders.

7. In Figure S4 and Figure S5, the river plot and the confusion matrix does not match with each other. For example, in Figure S4a, the predicted cell types are CD8+ T cells, Erythroid progenitors and Erythrocytes, but the labels are not the same in the Figure S4b. This error also exists for Figure 4c and 4d (scGPT-related plots).

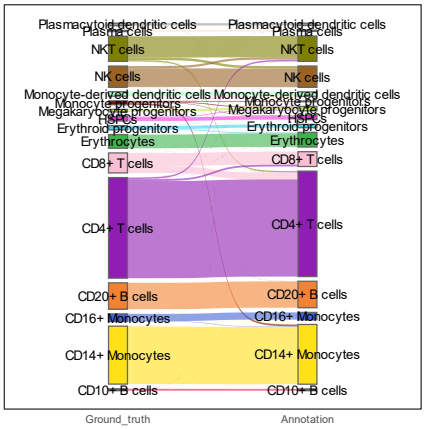
Response: Thanks for your valuable suggestions. We apologize for the discrepancies between the figures. After carefully reviewing the accuracy and Macro-F1 scores reported in Figures S10, S11, 4c-d, and 4g-h, we found that the typographical errors were caused by the mismatch between the string cell type labels and the one-hot encoded cell types during plotting the River plots and confusion matrices. We have since updated Figures S10, S11, 4c-d, and 4g-h, and thoroughly reviewed all other figures to ensure consistency throughout the manuscript.

CellFM

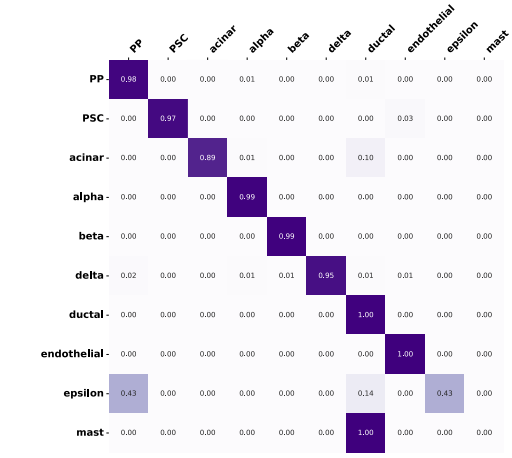
a


CD10+ B cells
CD14+ Monocytes
CD16+ Monocytes
CD20+ B cells
CD4+ T cells
CD8+ T cells
Erythrocytes
Erythroid progenitors
HSPCs
Megakaryocyte progenitors
Monocyte progenitors
Monocyte-derived dendritic cells
NK cells
NKT cells
Plasma cells
Plasmacytoid dendritic cells

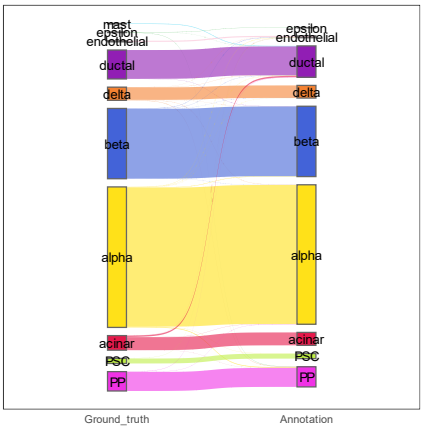
b


Plasmacytoid dendritic cells
Plasma cells
NKT cells
NK cells
Monocyte-derived dendritic cells
Monocyte progenitors
Megakaryocyte progenitors
HSPCs
Erythroid progenitors
Erythrocytes
CD8+ T cells
CD4+ T cells
CD20+ B cells
CD16+ Monocytes
CD14+ Monocytes
CD10+ B cells
Ground_truth
Annotation

c

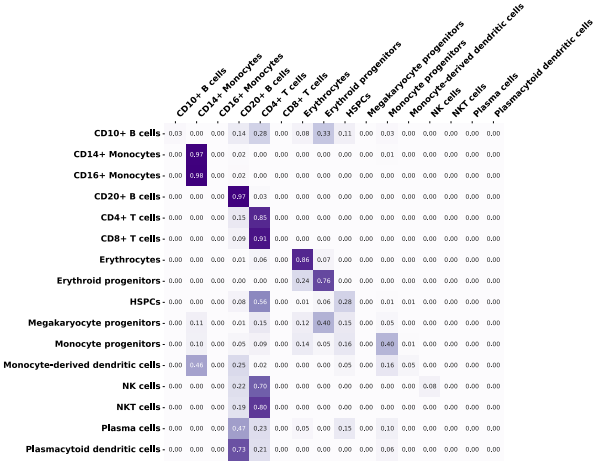

PP
PSC
acinar
alpha
beta
delta
ductal
endothelial
epsilon
mast

d

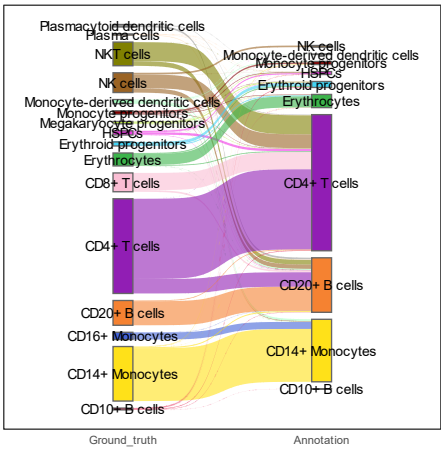

mast
epsilon
endothelial
ductal
delta
beta
alpha
acinar
PSC
PP
Ground_truth
Annotation

scGPT

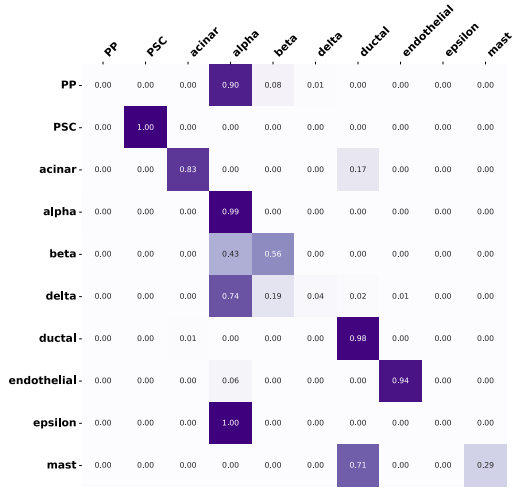
a


CD10+ B cells
CD14+ Monocytes
CD16+ Monocytes
CD20+ B cells
CD4+ T cells
CD8+ T cells
Erythrocytes
Erythroid progenitors
HSPCs
Megakaryocyte progenitors
Monocyte progenitors
Monocyte-derived dendritic cells
NK cells
NKT cells
Plasma cells
Plasmacytoid dendritic cells

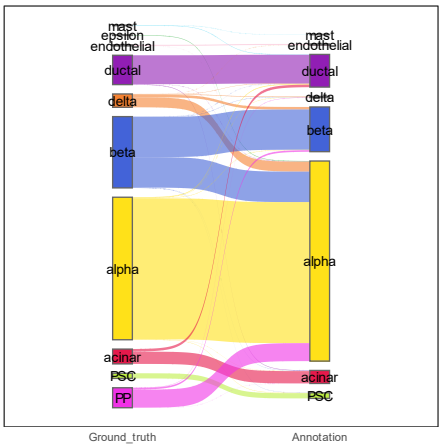
b


Plasmacytoid dendritic cells
Plasma cells
NKT cells
NK cells
Monocyte-derived dendritic cells
Monocyte progenitors
Megakaryocyte progenitors
HSPCs
Erythroid progenitors
Erythrocytes
CD8+ T cells
CD4+ T cells
CD20+ B cells
CD16+ Monocytes
CD14+ Monocytes
CD10+ B cells
NK cells
Monocyte-derived dendritic cells
Monocyte progenitors
HSPCs
Erythroid progenitors
Erythrocytes
CD4+ T cells
CD20+ B cells
CD14+ Monocytes
CD10+ B cells
Ground_truth
Annotation

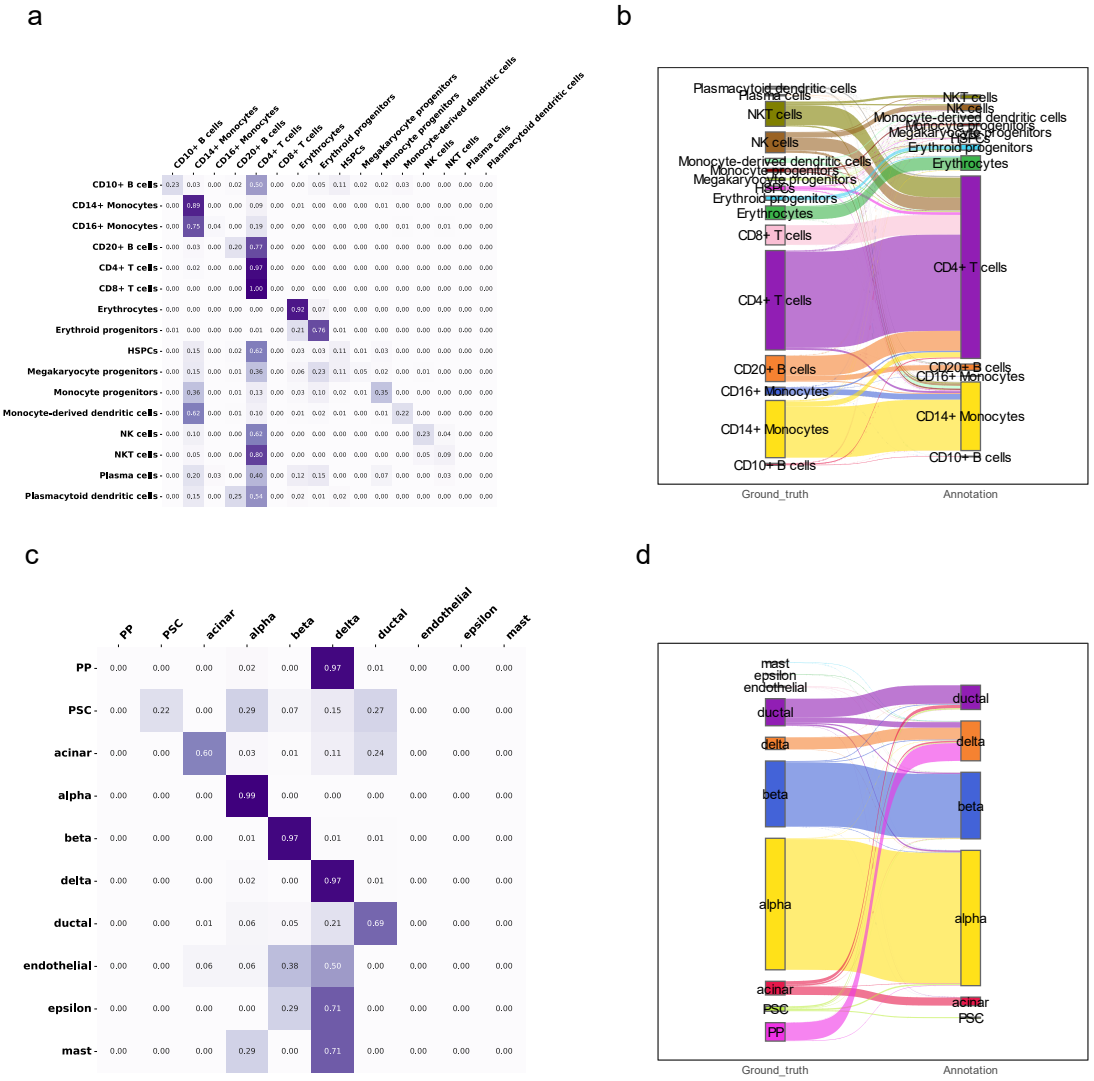
c


PP
PSC
acinar
alpha
beta
delta
ductal
endothelial
epsilon
mast

d


mast
epsilon
endothelial
ductal
delta
beta
alpha
acinar
PSC
PP
mast
endothelial
ductal
delta
beta
alpha
acinar
PSC
Ground_truth
Annotation

Geneformer



Reviewer #3 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #3 (Remarks on code availability):
The codes are well documented.

Response: Thanks for your involvement in the peer review process and for co-reviewing our manuscript as part of the Nature Communications initiative. We appreciate your effort to support Early Career Researchers in this important endeavor. We are glad to hear that the documentation of the codes was well received. We understand the importance of clear and comprehensive documentation for reproducibility, and we remain committed to enhancing the accessibility and usability of our resources.

Reviewer #4 (Remarks to the Author):

Overview:

In this paper, Zeng et al. introduce a new large foundation model for single cell RNA-sequencing data called CellFM, which is trained on 100 million human cells. The authors argue that existing foundation models for single-cell may have limited performance because of their training on multiple species. They train a large model on human only data, and show that it outperforms existing methods like scGPT and Geneformer on a set of benchmarks. On a few benchmarks the improvement over scGPT and Geneformer is noteworthy, however, the authors do not present a compelling argument or validation that their design choices caused this performance increase. Specifically, the authors do not validate their claim that single species training is superior to multi species training, do not evaluate model performance as a function of number of parameters or size of training dataset, and do not adequately compare the model to other existing models. Additionally, the model architecture, which essentially performs masked gene reconstruction, is not novel, and the authors do not present any novel capabilities of their model.

Response: We thank the reviewer for their valuable comments. As highlighted by Reviewer #1, CellFM exhibits scientific novelty and significant potential in advancing foundational models for single-cell analysis. The main novelty of CellFM can be summarized in three key aspects: (1) We compiled a standardized dataset of approximately 100 million human cells from public databases, twice as large as those used in the current largest single-species models; (2) Based on the efficient ERetNet architecture, the large dataset allows training CellFM with 800 million parameters, an eight-fold increase in size compared to prior single-species models; (3) CellFM consistently achieved the best performance among all single-cell foundation models across six representative tasks, demonstrating its superior robustness and accuracy. We have revised the manuscript to emphasize the importance of a larger training dataset and model parameter size. We then carefully addressed the reviewer's concerns, including the addition of recent single-cell foundation models in our comparisons, simulation experiments exploring the impact of data size and model scaling, and other aspects as discussed in the following responses.

Major Comments:

1. Confounding Effects of Multi-Species Training: The authors make the claim that: "Single-cell Large Language Models (LLMs) have been crafted to bridge this gap yet exhibit limited performance on human cells. This shortfall may stem from the confounding effects of training data

from diverse species, partly because of limited cells for the single species" (Abstract). However, they do not explain what these confounding effects are and fail to compare their model to others trained on multiple species. While the collection of a large human dataset is an important contribution, this dataset could also be utilized in training multi-species models. Furthermore, human data already comprises the majority of training data in existing multi-species methods like UCE. Given these points, the authors' claim is counterintuitive and requires stronger evidence.

Response: Thanks for your valuable comments. We have revised the abstract to indicate the potential advantages of a diverse species dataset. In fact, our study indicates human dataset-based models (scFoundation and our model) tend to outperform diverse species models (GeneCompass and UCE) in human test sets. Concretely, when a model is trained on such heterogeneous data, it may inadvertently learn features that are not relevant to humans, leading to overfitting and bias. This can result in performance that reflects these non-human-specific features rather than the true biology of human cells, complicating interpretability and making it difficult to identify relevant biological signals. Additionally, we also discussed combining other species for training next-generation single-cell foundation models in the "Discussion" section.

We have added single-cell foundation models trained on multiple species including UCE and GeneCompass in our comparison. While large human datasets have also been used to train multi-species models like UCE and GeneCompass, the number of human cells in these models did not exceed 50 million. In contrast, our model was trained on approximately 100 million human cells. Moreover, our model's parameter size is eight times larger than GeneCompass and 1.23 times larger than UCE. The results showed that CellFM consistently achieved the best performance among all single-cell foundation models across six representative tasks, demonstrating its superior robustness and accuracy (The corresponding description can also be found in the response to question #3 below). These results highlighted that our model, trained exclusively on human datasets, surpassed the performance of existing multi-species models. Detailed information has been added to "Abstract" and "Discussion" sections as follows:

"

While larger datasets spanning diverse species can enhance model generalization, the inherent inter-species differences may introduce noise and hinder model performance.

"

"

The current model is limited by the absence of multi-species data, which restricts its potential for broader biological contexts and cross-species comparisons.

"

2. Scaling and Ablation Experiments: A main contribution of this work is the collection of a large number of human scRNA-seq data and the training of a large (800 million parameter) transformer model. To validate this contribution, the authors must include ablation experiments demonstrating both parameter scaling and data scaling in some benchmarks. Ablation experiments should be performed, and they can be on a much smaller total number of parameters (e.g., less than 50M) to

reduce computational burden.

Response: Thanks for your valuable suggestions. To evaluate data scaling, we trained CellFM-80M containing 80 million parameters on varying numbers of cells: 10^5, 10^6, 10^7, 10^8. As shown in Supplementary Figure S16a, the cell annotation accuracy improved consistently as the number of cells increased, ranging from 88% at 10^5 cells to 91.2% at 10^8 cells. A similar trend was observed for GeneCompass, as well as for other tasks including cell perturbation and gene function prediction. These results highlight the importance of larger datasets in enhancing model performance.

To evaluate model scaling, we varied the model size from 8 million (8M) to 800 million (800M) parameters, with configurations at 8M, 80M, 200M, and 800M (the original model). As demonstrated in Supplementary Figure S16b, most tasks—except for cell type annotation at the zero-shot scenario—exhibited performance improvements with increasing model size. Cell annotation accuracy improved with an increase in parameters from 8M to 200M but decreased at 800M likely because this cell-level task is relatively simpler without requiring a larger parameter size. In contrast, more complex tasks, such as gene function prediction and perturbation prediction, continue to benefit from increases in both model size and training data since the larger model could efficiently capture intricate patterns in high-dimensional data. Interestingly, we found that fine-tuned versions of CellFM-800M achieved superior cell type annotation performance again, indicating the potential power of the larger model.

Based on these findings, we utilized CellFM-80M for cell annotation and batch effect correction tasks, while CellFM-800M was employed for gene function prediction and cell perturbation prediction tasks. We have incorporated these results into the "Results" section as follows:

“

To evaluate data scaling, we trained CellFM-80M containing 80 million parameters on varying numbers of cells: 10^5, 10^6, 10^7, 10^8. As shown in Supplementary Figure S16a, the cell annotation accuracy improved consistently as the number of cells increased, ranging from 88% at 10^5 cells to 91.2% at 10^8 cells. A similar trend was observed for GeneCompass, as well as for other tasks including cell perturbation and gene function prediction. These results highlight the importance of larger datasets in enhancing model performance.

To evaluate model scaling, we varied the model size from 8 million (8M) to 800 million (800M) parameters, with configurations at 8M, 80M, 200M, and 800M (the original model). As demonstrated in Supplementary Figure S16b, most tasks—except for cell type annotation at the zero-shot scenario—exhibited performance improvements with increasing model size. Cell annotation accuracy improved with an increase in parameters from 8M to 200M but decreased at 800M likely because this cell-level task is relatively simpler without requiring a larger parameter size. In contrast, more complex tasks, such as gene function prediction and perturbation prediction, continue to benefit from increases in both model size and training data since the larger model could efficiently capture intricate patterns in high-dimensional data. Interestingly, we found that fine-tuned versions of CellFM-800M achieved superior cell type annotation performance again, indicating the potential power of the larger model.

Based on these findings, we utilized CellFM-80M for cell annotation and batch effect correction tasks, while CellFM-800M was employed for gene function prediction and cell perturbation prediction tasks.
”

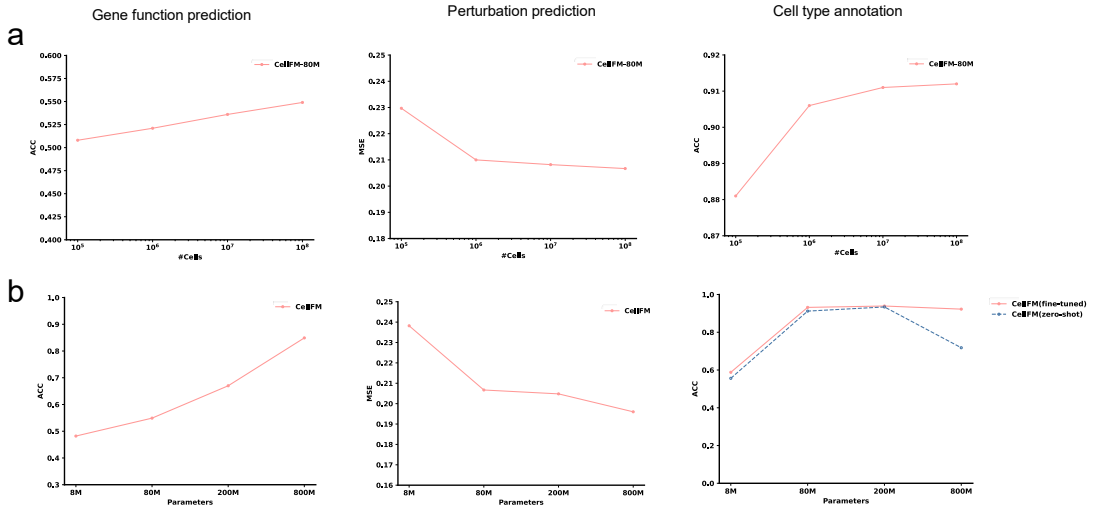


Fig. S16. The performance of CellFM when scaling training parameters and the number of training cells on downstream tasks, including gene function prediction (binary classification), perturbation prediction, and cell type annotation (Liver data) (a) Performance of CellFM with 80M training parameters trained on the human single-cell transcriptome corpus with varying cell counts. (b) Performance of CellFM with different training parameters trained on the fully human single-cell transcriptome corpus containing approximately 100M cells.

3. Benchmarks and Comparisons: The authors mention various methods with different architectures, data sizes, and parameter counts. They should include representatives of these methods in their benchmarks. For gene-based benchmarks (e.g., Figure 2), comparisons should be made to models like GenePT [1] / scELMo [2] and UCE, which use gene embeddings from large language or protein models. For cell-based benchmarks, especially the zero-shot cell type annotation task, multi-species models like UCE and GeneCompass should be included. Additionally, the authors should compare CellFM to scFoundation, which shares a similar masked autoencoding architecture. It would also be valuable to assess embedding quality for different organs and sequencing protocols.

Response: Thanks for your valuable suggestions. We have newly incorporated several recent single-cell foundation models into our benchmarking, including scFoundation, GeneCompass, UCE, scELMo, and scBERT. We have evaluated our model CellFM alongside these single-cell foundation models across six representative tasks (if these models include publicly available code or results in their manuscripts, seen in Supplementary Figure S3): cell type annotation, gene function prediction including binary classification and multi-category classification, cell perturbation prediction including unseen gene perturbations prediction and in silico reverse perturbation prediction, and batch effect correction. CellFM consistently achieved the best performance among all single-cell foundation models across six representative tasks, demonstrating its superior robustness and

accuracy. The results for each task are presented succinctly below:

For the cell type annotation task, CellFM achieved an average accuracy of 94.02% on inter-datasets, outperforming the top three single-cell foundation models scFoundation, scBERT, and UCE by 2.62%, 4.22%, and 6.34% regarding average accuracy values on inter-datasets, respectively. Detailed information on these results can be found in our responses to Reviewer #2 (Question 4) and the response to Question #5.

For predicting unseen gene perturbations, CellFM achieved the highest performance, exceeding scFoundation and GeneCompass by 1% and 2.7% regarding average accuracy values, respectively. Detailed information on these results can be found in our responses to Reviewer #1 (Question 2.3). For in silico reverse perturbation prediction, CellFM achieved the highest performance, exceeding the second-best model scGPT 13.6% regarding average hit rates among the top 1-10 across two datasets. Detailed information on these results can be found in our responses to the following response #6.

For the batch effect correction task, the deep learning framework used for CellFM, MindSpore, does not provide the Gradient Reversal Layer (GRL) technique. Similarly, other single-cell foundation models with batch effect correction functionality also did not implement GRL. To ensure a fair comparison, we re-evaluated scGPT by removing the GRL loss while retaining its other loss functions. The results demonstrated that CellFM achieved comparable performance to scGPT and outperformed other single-cell foundation models by 4.6%. Detailed information about the results can be found in the responses to Reviewer #2 of question #8.

For gene function prediction in the context of binary classification, As shown in Figure 2, CellFM consistently outperformed all single-cell foundation modes and achieved an average accuracy of 84.9%, which was 5.68% higher than the second-ranked model UCE. For gene function prediction in the context of multi-category classification, CellFM consistently outperformed all single-cell foundation modes and achieved an average AURP value of 73.7%, which was 1.6% higher than the second-ranked model GeneCompass. Detailed information on these results can be found in the responses to Reviewer #2 of question #3. We didn't compare with scELMo on the gene function prediction tasks since scELMo employed large language models (LLMs) like GPT-3.5 as generators to generate embeddings from metadata descriptions in the training phase, which has included gene function information.

To further evaluate the embedding quality of CellFM and scFoundation across different organs and sequencing protocols, we generated UMAP plots using embeddings extracted from both models. Concretely, we compared CellFM with scFoundation on the tissue human Liver sequenced by mCEL-Seq2, human Pancreas sequenced by CEL-Seq2, 10K PBMC sequenced by 10x, and the gene function prediction dataset. As shown in Supplementary Figure S14(a), CellFM more effectively grouped different cell types in UMAP space compared to scFoundation on the Liver and hPancreas data regarding cell type annotation task, which aligns with its superior cell type annotation performance. For example, in dataset hPancreas, CellFM efficiently grouped different cell types, while scFoundation separated alpha cell type into three groups. A similar trend was observed when analyzing the Liver dataset. CellFM and scFoundation achieved similar embeddings on the 10K PBMC data regarding the batch effect correction task, while CellFM achieved higher AvgBIO values (Supplementary Figure S14b). For the gene embeddings, CellFM more efficiently distinguished genes with different functions, while scFoundation struggled to separate them

(Supplementary Figure S14c). These results have been incorporated into the "Results" section as follows:

“

To further evaluate the embedding quality of CellFM and scFoundation across different organs and sequencing protocols, we generated UMAP plots using embeddings extracted from both models. Concretely, we compared CellFM with scFoundation on the tissue human Liver sequenced by mCEL-Seq2, human Pancreas sequenced by CEL-Seq2, 10K PBMC sequenced by 10x, and the gene function prediction dataset. As shown in Supplementary Figure S14(a), CellFM more effectively grouped different cell types in UMAP space compared to scFoundation on the Liver and hPancreas data regarding cell type annotation task, which aligns with its superior cell type annotation performance. For example, in dataset hPancreas, CellFM efficiently grouped different cell types, while scFoundation separated alpha cell type into three groups. A similar trend was observed when analyzing the Liver dataset. CellFM and scFoundation achieved similar embeddings on the 10K PBMC data regarding the batch effect correction task, while CellFM achieved higher AvgBIO values (Supplementary Figure S14b). For the gene embeddings, CellFM more efficiently distinguished genes with different functions, while scFoundation struggled to separate them (Supplementary Figure S14c).

”

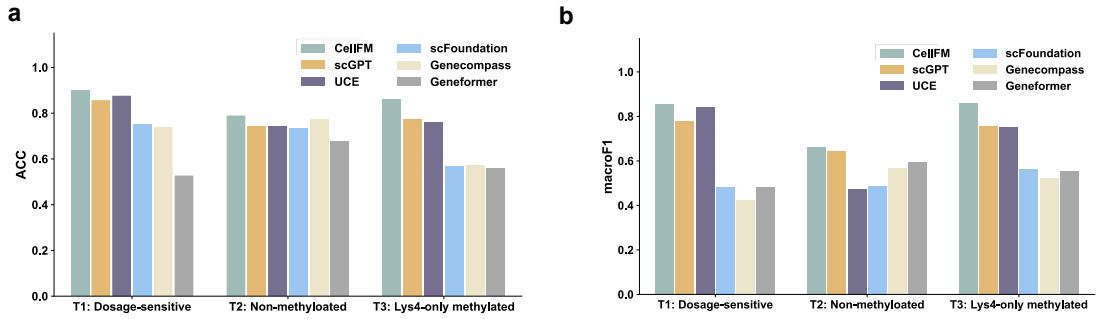


Fig. 2. Comparison of Gene Function Prediction performance in a zero-shot setting. (a) accuracy (ACC) scores and (b) Macro-F1 values for CellFM along with other competing single-cell foundation models.

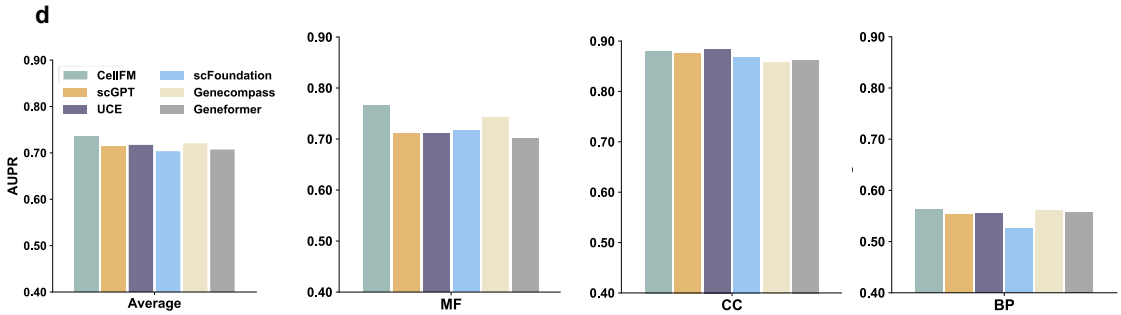


Figure 2d. AUPR values for multiple gene functions predicted by CellFM and other competing single-cell foundation models on GO data, where each gene is annotated with numerous functions. MF: Molecular Function, CC: Cellular Component, and BP: Biological Process.

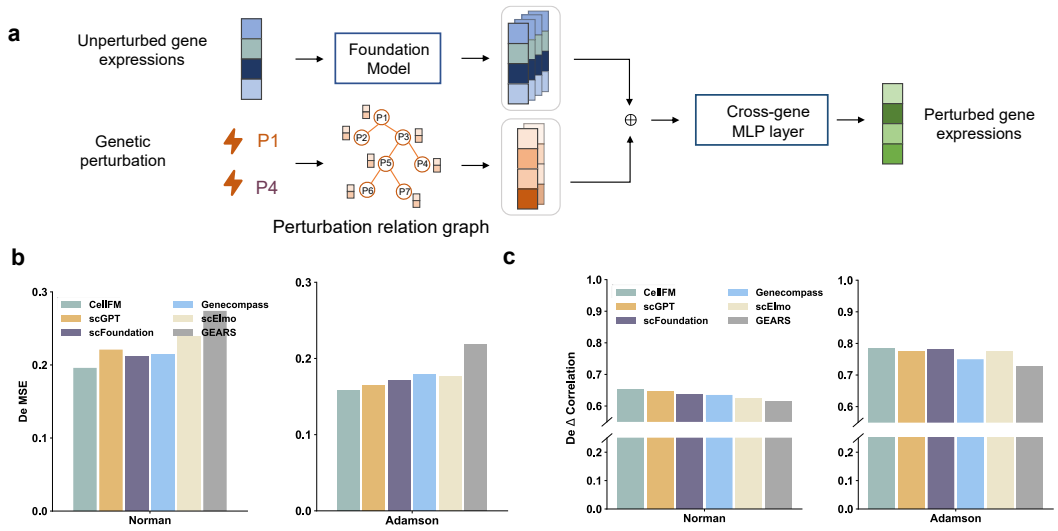


Fig. 3. Analysis of Perturbation Response and Reverse Perturbation Predictions. (a) A diagram showcasing the perturbation prediction model leveraging cell-specific gene embeddings derived from CellFM. (b) The mean square error (MSE) between predicted and actual post-gene expressions for the top 20 differentially expressed (DE) genes. (c) Comparison of CellFM with other single-cell foundation models and the perturbation prediction method GERAS. Pearson correlation coefficients between predicted and actual gene expression changes are reported for the top 20 differentially expressed (DE) genes.

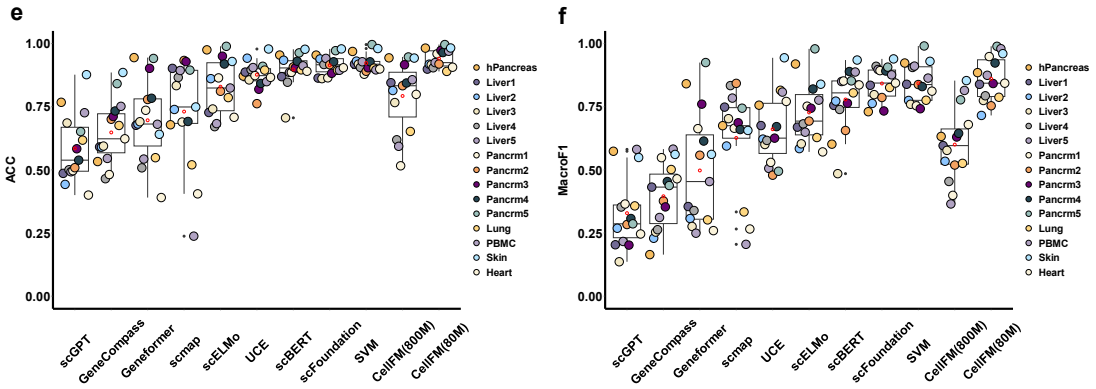


Figure 4. Zero-shot cell type annotation performance of each model. (e) Boxplots summarize the ACC scores for each method on n=15 biologically independent paired inter-datasets. (f) Boxplots summarize the Macro-F1 scores for each method on n=15 biologically independent paired inter-datasets.

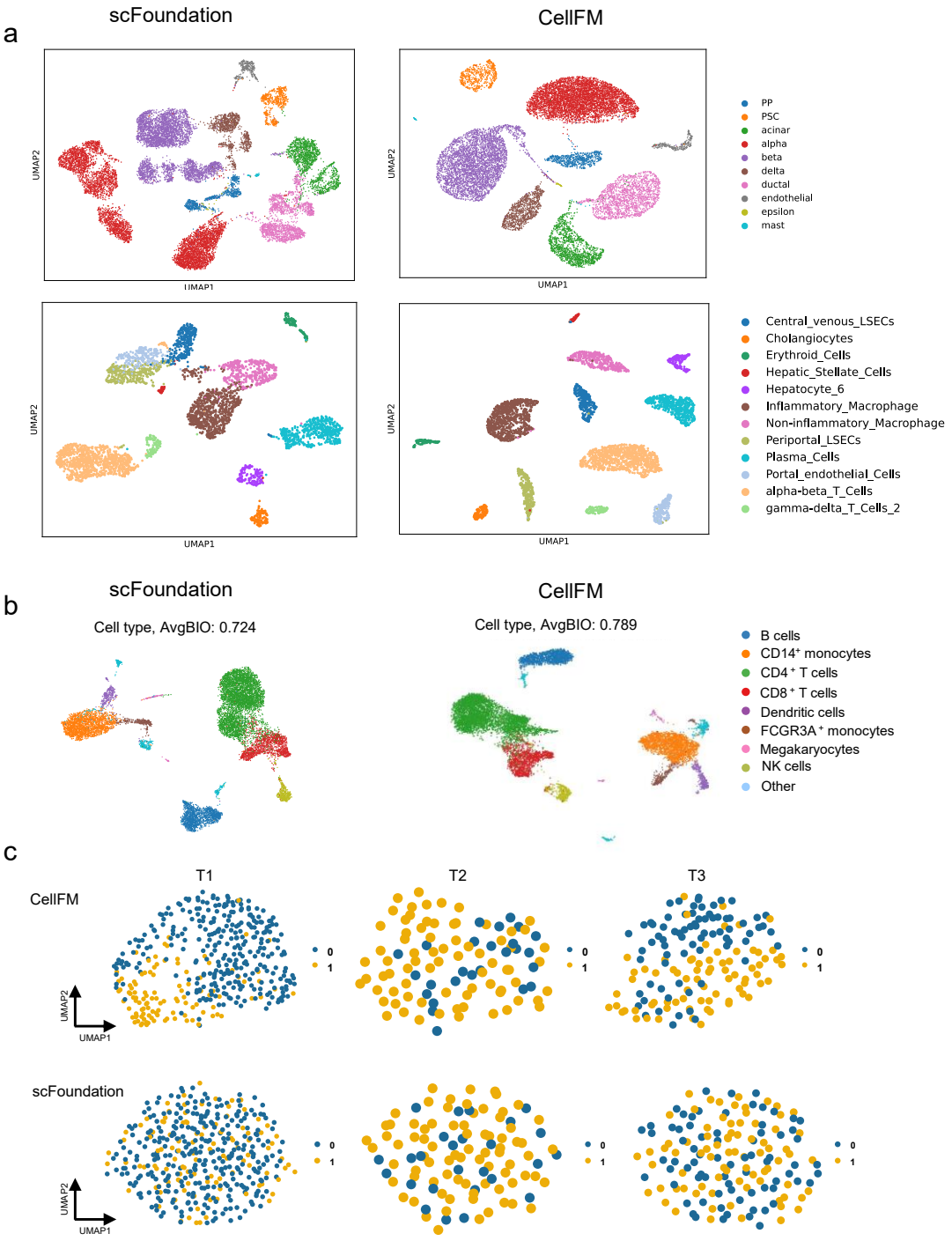


Fig. S14. (a) UMAP plots of CellFM and scFoundation using embeddings generated by CellFM and scFoundation on the hPancreas and Liver datasets for the cell type annotation task. (b) UMAP results plotted with embeddings from CellFM and scFoundation for the batch effect correction task on the PBMC 10K dataset. (c) UMAP results generated using embeddings from CellFM and scFoundation for the gene function prediction dataset.

Models	Open soucre	Gene Function Prediciton		Cell type annotaion	Batch effect corrrection	Perturbation Prediciton	
		Binary classification	Multi-category classification			Predicting unseen gene perturbations	In silico reverse perturbation prediction
CelllFM	✓	✓	✓	✓	✓	✓	✓
scGPT	✓			✓	✓	✓	✓
Geneformer	✓	✓		✓			
scFoundation	✓			✓		✓	
GeneCompass	✓	✓		✓		✓	
UCE	✓			✓	✓		
scELMo	✓	✓		✓	✓	✓	
scBERT	✓			✓			

Fig. S3. The table provides hints for selecting a model based on task usability. Blank cells indicate that the selected models lack specific designs to meet the corresponding criteria.

4. Unclear Pre-training Setting: The authors should discuss the limitations of their model. In particular, the pre-training step, where gene IDs are padded and random expression values are added (Section 4.3), raises concerns. The pre-training task uses a masked language modeling loss, where 20% of the input is masked. How can the model learn meaningful compression if the input includes random values? The impact of excluding these cells from the 100M dataset should be investigated.

Response: Thanks for your valuable suggestions. We simply set the padded values as zero (instead of random values) because the model architecture has a fixed length l_{max} for parallel computing. These padded values won't participate in the calculations during CellFM's training and thus don't influence our models. During gene masking, the 20% of genes masked during the pre-training task were exclusively selected from non-padded genes, ensuring that CellFM focused on reconstructing meaningful gene expression patterns using the remaining relevant gene information. This design allows the model to learn effective compression and meaningful representations without interference from padding values. We have updated the relevant description and discussed the limitations of CellFM in the "Methods" and "Discussion" sections as follows:

“

We simply set the padded values as zero because the model architecture has a fixed length l_{max} for parallel computing. These padded values won't participate in the calculations during CellFM's training and thus don't influence our models.

”

“ During gene masking, the 20% of genes masked during the pre-training task were exclusively selected from non-padded genes, ensuring that CellFM focused on reconstructing meaningful gene expression patterns using the remaining relevant gene information. This design allows the model to learn effective compression and meaningful representations without interference from padding values.

”

“

Despite the advances in CellFM, several limitations remain to be explored. Firstly, the attention map in CellFM was limited in capturing gene relationships related to static or global biological knowledge. In the future, we will explore new explainability techniques to overcome this challenge. Furthermore, the current model is limited by the absence of multi-species data, which restricts its potential for broader biological contexts and cross-species comparisons. Finally, the model's construction did not leverage existing biological prior knowledge, which could affect its depth and accuracy in interpreting biological phenomena.
”

5. Benchmarks and Missing Details on Cell-type annotation: The cell-type annotation benchmark is unrealistic and should be replaced by benchmarks from the single cell integration benchmark scIB[3], scGraph [4], and out-of-domain settings [5]. A randomized train-test split for cells is an unrealistic setting and does not adequately measure the ability of a model to encode cell types or correct batches. Additionally, the presentation of these results should be improved: in figure 4ef, results from different dataset should not be shown in the same box plot. If the authors continue to use the same 70% validation split, they should perform multiple splits to characterize the variance of these metrics. Finally, it is not clear whether these are truly zero-shot metrics: were any of these datasets included in training of any model? Is the base-foundation model fully frozen? What model is used to make the cell type annotations / predictions?

Response: Thanks for your valuable suggestions. The cell annotation benchmarks used in our study, including intra-dataset and inter-dataset evaluations, were based on the latest benchmarking framework, scEval, designed to evaluate single-cell foundation models on cell annotation tasks. Following prior studies such as scGPT and scEval, we utilized randomized train-test splits for intra-dataset evaluation to assess model performance under consistent experimental conditions. While intra-datasets were a part of our evaluation, we also performed inter-dataset testing, which better reflects real-world scenarios involving large batch effects. For inter-datasets, we partitioned the data by batch ID, iteratively using each batch as the test set while training on the remaining batches. For datasets like hPancreas, we directly adopted the train-test settings established in scGPT, ensuring training and testing data came from different batches. This cross-batch validation ensures a thorough and fair assessment of model robustness across batches. Results in Figures 4e and 4f reflect these batch-specific evaluations. We organized them using box plots, following the approach in scBERT [Nature Machine Intelligence, DOI: 10.1038/s42256-022-00534-z].

Importantly, the cell annotation datasets used in our study were not included during the pre-training of CellFM. For the zero-shot cell type annotation task, all foundation models, including ours, were fully frozen during training. Classifiers (e.g., MLP or CNN+MLP) were then trained on embeddings extracted from these frozen models using labeled training data and the cross-entropy loss function. These trained classifiers were subsequently used to predict cell types on test datasets. We used the default classifiers implemented in each single-cell foundation model. For models like UCE, which lack a predefined classifier, we implemented a multi-layer perceptron (MLP) approach, in line with standard foundation model practices. For our CellFM model, we followed it with a three-layer MLP classifier for the classification task.

In response to your suggestions, we have incorporated additional datasets from scIB, scGraph, and “out-of-domain settings” to expand our evaluations. Specifically, we included the Lung dataset from scIB, while the Pancreas and Immune datasets have already been used in this study. Other datasets from scIB, such as cross-species or cross-modal data (e.g., scATAC-seq), were excluded because our model is not currently capable of processing them. From scGraph, we strategically chose two datasets (Skin and Heart) that represent small and medium-sized data, taking into account the limitations on time and memory consumption. Furthermore, we opted to include only the PBMC data from out-of-domain settings since the other Lung dataset has been sourced from scIB. These newly introduced datasets, which encompass multiple batches, were assessed as inter-dataset comparisons. As depicted in the following Figure, our model outperformed the second-best model, scFoundation, by 1.75% in terms of average accuracy on these four datasets. A similar trend was observed when the Macro-F1 metric was employed. These results have been integrated into Figure 4e-f and the "Results" section as detailed below:

“

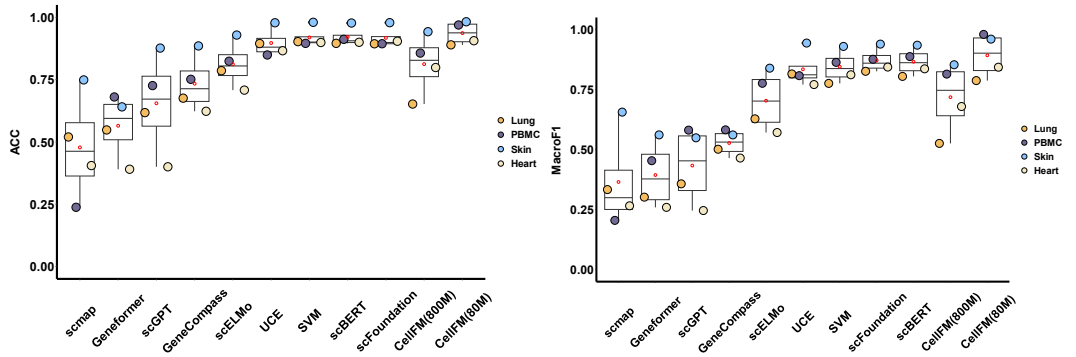
The cell annotation benchmarks used in our study, including intra-dataset and inter-dataset evaluations, were based on the latest benchmarking framework, scEval, designed to evaluate single-cell foundation models on cell annotation tasks. Following prior studies such as scGPT and scEval, we utilized randomized train-test splits for intra-dataset evaluation to assess model performance under consistent experimental conditions. While intra-datasets were a part of our evaluation, we also performed inter-dataset testing, which better reflects real-world scenarios involving large batch effects. For inter-datasets, we partitioned the data by batch ID, iteratively using each batch as the test set while training on the remaining batches. For hPancreas, we directly adopted the train-test settings established in scGPT, ensuring training and testing data came from different batches. This cross-batch validation ensures a thorough and fair assessment of model robustness across batches.

”

“

Importantly, the cell annotation datasets used in our study were not included during the pre-training of CellFM. For the zero-shot cell type annotation task, all foundation models, including ours, were fully frozen during training. Classifiers (e.g., MLP or CNN+MLP) were then trained on embeddings extracted from these frozen models using labeled training data and the cross-entropy loss function. These trained classifiers were subsequently used to predict cell types on test datasets. We used the default classifiers implemented in each single-cell foundation model. For models like UCE, which lack a predefined classifier, we implemented a multi-layer perceptron (MLP) approach, in line with standard foundation model practices. For our CellFM model, we followed it with a three-layer MLP classifier for the classification task.

”



The performance of each model on the Lung, PBMC, Skin, and Herat datasets under the zero-shot setting.

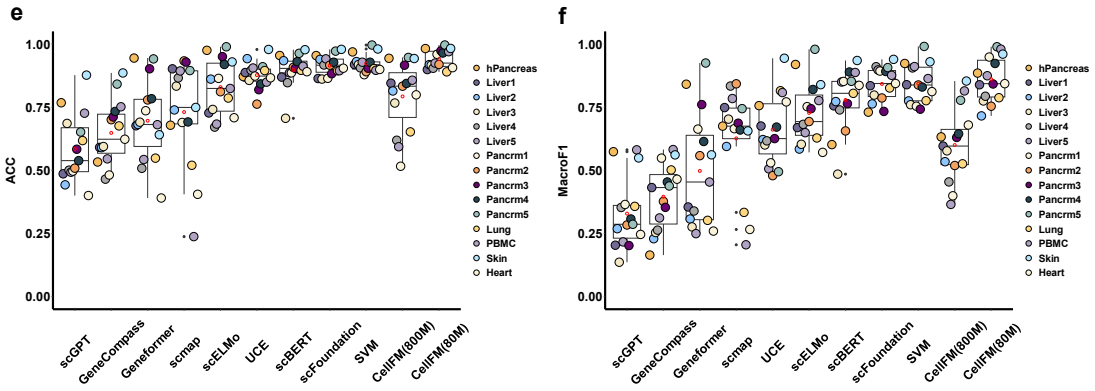


Figure 4. Zero-shot cell type annotation performance of each model. (e) Boxplots summarize the ACC scores for each method on n=15 biologically independent paired inter-datasets. (f) Boxplots summarize the Macro-F1 scores for each method on n=15 biologically independent paired inter-datasets.

6. Performance and Counterintuitive Interpretation on Perturbation Predictions: The reported improvement in performance on the Norman and Adamson perturbation datasets is very minor according to the authors (1.65% and 2%). How does such a minor improvement in performance lead to such significant changes in the hit rate when predicting in the reverse direction (Figure 4e)? GEARs should also be compared on this metric. Additionally, when predicting the expression of cells under perturbation, the output looks highly unrealistic (Supplementary Figure S1). Many perturbations have no effect, or similar effects, yet the outputs all form tight clusters.

Response: Thanks for your valuable suggestions. Our results are consistent with scGPT that there are big changes in hit rate but small changes in PCC, likely because hit rates are relatively more sensitive than PCC in measuring model performances. Specifically, the observed differences between forward and reverse perturbation predictions primarily arise from differences in evaluation metrics and the test datasets, as well as the nature of the tasks. For forward perturbation prediction, we followed scGPT and GEARS, evaluating performance using the top 20 most differential genes. In contrast, for reverse perturbation prediction, we followed scGPT's approach by selecting 20 genes from the Norman dataset to construct fine-tuning and testing use cases (The detailed construction steps are provided in the response to Reviewer #2, Question #5).

To further assess the robustness of CellFM, we performed additional evaluations by varying the random seeds during the dataset splitting process to generate new training and test cases. As shown in Supplementary Figure S7, our model consistently outperformed scGPT and GEARS. For example, CellFM achieved comparable performance at top 3 predictions. For the top 5 prediction, CellFM achieved 45.5% correctly identified perturbations, which was 9% higher than scGPT. Furthermore, we re-evaluated scGPT, CellFM, and GEARS under different random seeds for fine-tuning. The results demonstrated that CellFM consistently maintained similar top 1-10 prediction performance (Supplementary Figure S7). We did not include comparisons with UCE, Geneformer, GeneCompass, scFoundation, and scELMo on the reverse perturbation prediction task since their original publications did not provide code or results, and reimplementing these methods would require significant programming effort and time.

Previously, the expression patterns of perturbed cells appeared dispersed because we didn't exclude the perturbed genes that separate all cells. Now we excluded all perturbed genes and replot the UMAP graph. As shown in Supplementary Figure S5 (Figure S1 in the initial submission), the clusters exhibited overlap in certain regions while remaining distinct in others, consistent with the expectation that several perturbations either have no effect or produce similar effects. Specifically, after predicting perturbed gene expression using our model, we removed the perturbation genes. Following the scGPT framework, we calculated the average expression for each perturbation based on its predicted perturbed expression relative to the corresponding control cells. To identify domain genes, we employed a K-Nearest Neighbors (KNN)-based approach: for each perturbation, we determined its 10 most similar neighbors and defined any shared genes as domain genes if at least 3 neighbors contained the same gene. By iterating this process across all perturbation combinations, we identified the domain genes and generated the final UMAP plot. We have added these results into the "Result" section as follows:

"

We excluded all perturbed genes and plotted the UMAP graph. As shown in Supplementary Figure S5, the clusters exhibited overlap in certain regions while remaining distinct in others, consistent with the expectation that several perturbations either have no effect or produce similar effects.

"

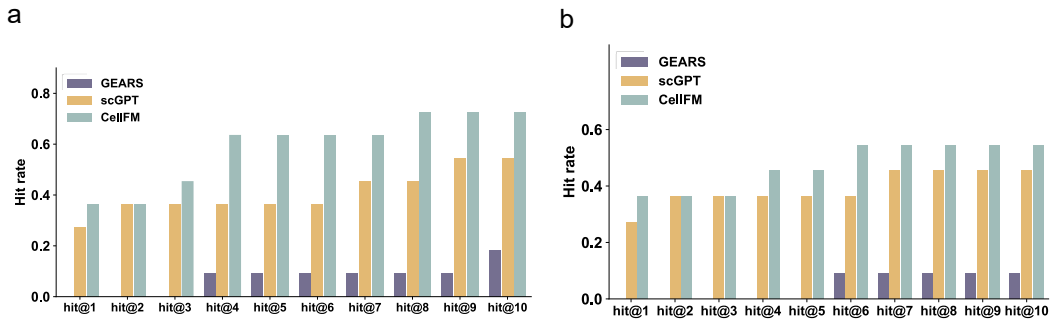


Fig. S7. (a) The Top 1–10 accuracy of CellFM and scGPT after fine-tuning with a new random seed. (b) The Top 1–10 accuracy of CellFM and scGPT on newly generated test cases. The bar graph highlights the hit rates.

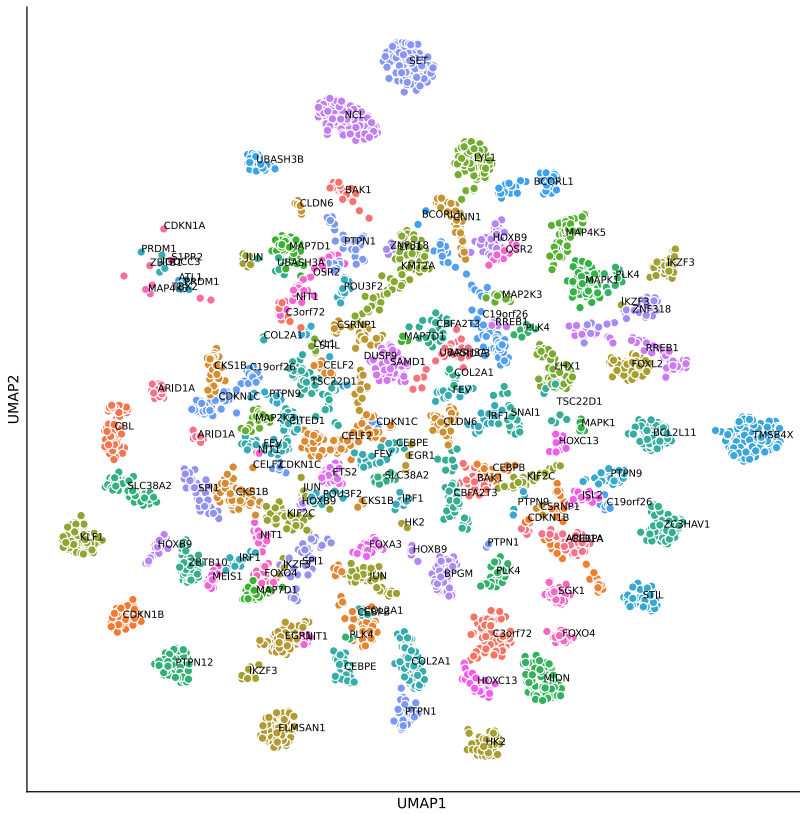


Fig. S5. The UMAP visualization of predicted gene expression profiles under various perturbation conditions. Each cluster in the UMAP plot is annotated with the predominant gene for each cluster.

Models	Open soucre	Gene Function Prediciton		Cell type annotaion	Batch effect corrrection	Perturbation Prediciton	
		Binary classification	Multi-category classification			Predicting unseen gene perturbations	In silico reverse perturbation prediction
CelllFM	✓	✓	✓	✓	✓	✓	✓
scGPT	✓			✓	✓	✓	✓
Geneformer	✓	✓		✓			
scFoundation	✓			✓		✓	
GeneCompass	✓	✓		✓		✓	
UCE	✓			✓	✓		
scELMo	✓	✓		✓	✓	✓	
scBERT	✓			✓			

Fig. S3. The table provides hints for selecting a model based on task usability. Blank cells indicate that the selected models lack specific designs to meet the corresponding criteria.

Minor Comments:

- Referring to foundation models as "large language models" (LLMs) is inappropriate since these models are not trained on natural language. A more suitable term would be "foundation models."
Response: Thanks for your valuable suggestions. We have replaced all instances of "large language models" with "foundation models" to accurately describe these models. Additionally, we have carefully reviewed other related expressions in the text to ensure overall consistency.

- The color scheme in Figure S3 is confusing.
Response: Thanks for your valuable suggestions. In Figure S8 (Fig S3 in initial submission), the colors dark purple, orange, and green were used to represent the hit rates of perturbation combination genes in the top-k predictions for GEARS, scGPT, and CellFM, respectively. The tight purple color indicated relevant predictions, signifying that at least one gene from the perturbation combination was included in the top-k predictions. We have revised the figure caption of Figure S8 for better readability.

- In Figure 2c, the label should read "CellFM" rather than "CellAI."
Response: Thanks for your valuable comments. We have meticulously corrected the typographical errors and thoroughly reviewed the manuscript to prevent the recurrence of similar typographical mistakes.

- In Figure S2, clarify which dataset the true box represents (e.g., test, validation, train, or unseen).
Response: Thanks for your valuable suggestions. To demonstrate the effectiveness of reverse perturbation prediction, we utilized a subset of the Norman dataset that focuses on perturbations involving 20 genes as shown in Supplementary Figure S6 (Fig S2 in initial submission). This combinatorial space comprises a total of 210 one-gene and two-gene perturbation combinations. In Figure S6, the color bar on the right indicates the categories of test, validation, train, or unseen data. Specifically, the boxes are colored as follows: dark purple for test, light purple for validation, medium purple for train, and gray for unseen. The detailed construction steps of the data are provided in the response to Reviewer #2, Question #5. We have included these details in the captions of Figures 3d and S6 as follows:

“
The boxes are colored as follows: dark purple for test, light purple for validation, medium purple for train, and gray for unseen.
”

- The text in many figures is small and hard to read, requiring significant zooming. Subfigure order is inconsistent across figures, such as clockwise in some and counterclockwise in others.
Response: Thanks for your valuable suggestions. We have improved the legibility and resolution of all figures by increasing the font size and adjusting the image resolution. To ensure consistency, we have also standardized the arrangement of subfigures in a clockwise direction across all figures.

- Clarify which tasks use fine-tuning and when LoRA is applied. The authors do not compare the

performance of full fine-tuning versus LoRA.

Response: Thanks for your valuable comments. We applied LoRA during the fine-tuning phase for the cell type annotation task. As shown in Supplementary Figure S13, the performance of CellFM showed minimal changes when LoRA was applied compared to when it was not, across the inter-datasets. However, using LoRA reduced the time required for fine-tuning. Based on these findings, we recommend applying LoRA during the fine-tuning of CellFM to enhance efficiency without sacrificing performance. The detailed information has been added into "Results" section as follows:

"

We applied LoRA during the fine-tuning phase for the cell type annotation task. As shown in Supplementary Figure S13, the performance of CellFM showed minimal changes when LoRA was applied compared to when it was not, across the inter-datasets. However, using LoRA reduced the time required for fine-tuning. Based on these findings, we recommend applying LoRA during the fine-tuning of CellFM to enhance efficiency without sacrificing performance.

"

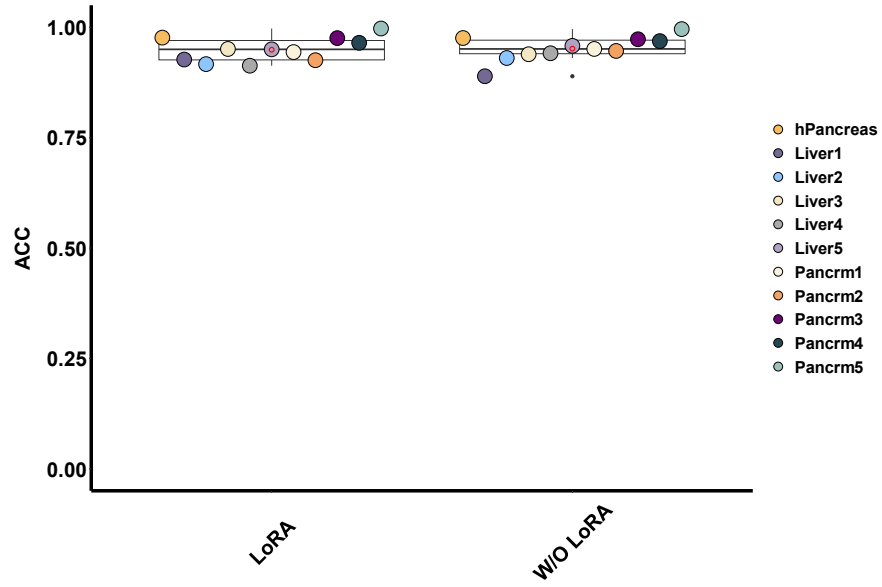


Fig. S13. The cell type annotation performance of CellFM after fine-tuning with or without LoRA on inter-datasets.

- A computational comparison between ERetNet and a vanilla transformer, along with ablation studies on Simple Gated Linear Unit and Layer Normalization, would provide additional insights.

Response: Thanks for your valuable suggestions. We have conducted the proposed ablation experiments comparing ERetNet with a vanilla transformer, as well as examining the effects of the Simple Gated Linear Unit and Layer Normalization. All ablation experiments were conducted using a newly trained CellFM model with 80 million parameters, with a focus on the cell type annotation task, due to limitations in time and computational resources. As shown in Supplementary Figure S17, removing the Simple Gated Linear Unit and DeepNorm resulted in decreases of 0.8% and 0.9%, respectively. Additionally, the removal of the L_cls loss led to a slight drop (0.4%) in performance. When compared to the classic Transformer, CellFM demonstrated a 1.2% improvement. In

conclusion, these modifications collectively contributed to the robust performance of CellFM, as evidenced by the benchmarking results. We have added these ablation results into "Results" section as follows:

"

We have conducted the ablation experiments comparing ERetNet with a vanilla transformer, as well as examining the effects of the Simple Gated Linear Unit and Layer Normalization. All ablation experiments were conducted using a newly trained CellFM model with 80 million parameters, with a focus on the cell type annotation task, due to limitations in time and computational resources. As shown in Supplementary Figure S17, removing the Simple Gated Linear Unit and DeepNorm resulted in decreases of 0.8% and 0.9%, respectively. Additionally, the removal of the L_cls loss led to a slight drop (0.4%) in performance. When compared to the classic Transformer, CellFM demonstrated a 1.2% improvement. In conclusion, these modifications collectively contributed to the robust performance of CellFM, as evidenced by the benchmarking results.

"

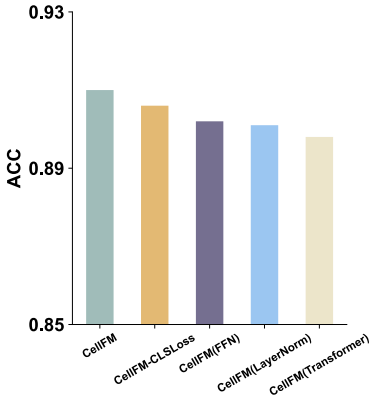


Fig. S17. The ablation study of CellFM. Performance of CellFM with the removal of ClS_{loss}, the gated bilinear network, and DeepNorm, as well as when replacing the backbone with a Transformer.

- Although the authors have made the curated data publicly available, they should provide a list of the training datasets and their sources.

Response: Thanks for your valuable suggestions. We have included a comprehensive list of the training datasets used in our study, along with their sources and additional metadata, in the supplementary file (Supplementary_Metadata_Information.xlsx). Additionally, a description of the datasets used is provided in Supplementary Figure S1.

- Provide a clear description of the limitations of their current model, potentially using a model card format as outlined in [6].

Response: Thanks for your valuable suggestions. We have uploaded our model (the code is available on GitHub: https://github.com/biomed-AI/CellFM.) on Hugging Face and included a comprehensive model card, which can be found here: (https://huggingface.co/ShangguanNingyuan/CellFM). The model card includes the following: 1)

Model Details: We have specified the model's name, developers, and release date to provide context for users. 2) Model Description: A detailed account of the model's architecture, training data, and process has been included to enhance transparency. 3) Installation: Instructions on how to install the model are provided. 4) Application: We have reported the model's applications across various tasks. 5) Limitations: We have thoroughly described the model's limitations.

References

[1] GenePT: A Simple But Effective Foundation Model for Genes and Cells Built From ChatGPT, Chen & Zou, bioRxiv 2023
[2] scELMo: Embeddings from Language Models are Good Learners for Single-cell Data Analysis, Liu et al., bioRxiv 2023
[3] Benchmarking atlas-level data integration in single-cell genomics, Luecken et al. Nature Methods 2021
[4] Metric mirages in Cell embeddings, Wang, Leskovec & Regev, bioRxiv 2024
[5] Cell-ontology guided transcriptome foundation model, Yuan et al., arXiv 2024
[6] https://huggingface.co/docs/hub/en/model-cards

Reviewer #4 (Remarks on code availability):

We briefly reviewed the GitHub repository, which contains both the pre-training and downstream modules. However, we do not have the budgets to run pre-training. Pre-trained models and fine-tuning notebooks should be provided.
Response: Thanks for your valuable suggestions. To improve accessibility and usability, we have updated the repository to include pre-trained models (https://huggingface.co/ShangguanNingyuan/CellFM/tree/main) and fine-tuning notebooks (https://github.com/biomed-AI/CellFM/tree/main/tutorials). This update will allow users to effectively utilize our models without the need for extensive pre-training, making it more convenient for everyone.

Reviewer #5 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #5 (Remarks on code availability):

See main review

Response: Thanks for your valuable contribution as a co-reviewer of our manuscript as part of the Nature Communications initiative. We appreciate your efforts to support Early Career Researchers in their development within the peer review process. We have carefully reviewed and addressed the issues raised in the main review, ensuring that we provided thorough responses to the reviewers' questions. We appreciate your feedback and are committed to improving the manuscript based on all the insights provided.

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

The authors have done a remarkable work in addressing all of my concerns. I would still like to see a more well-document GitHub in terms of notebooks. As a final comment I would like to see the code of the extra experiments they made with CellOT and GEARS as I was not able to see them in the GitHub (maybe I am missing something). A final check on the wording should also take place. Apart from these comments I am not against supporting a positive outcome for publication.
Response: Thanks for your positive feedback and valuable suggestions. We have added the codes for the CellOT and GEARS experiments to ensure clarity. Additionally, we have thoroughly revised the manuscript to refine the language and improve overall readability. We greatly appreciate your support and constructive input, which have significantly strengthened the quality of our work.

Reviewer #1 (Remarks on code availability):

I would like to see the benchmark experiments with CellOT and GEARS since I was not able to track them down in the repo.
Response: Thanks for your valuable comments. The benchmark experiments for CellFM+GEARS are available at https://github.com/biomed-AI/CellFM/blob/main/tutorials/Perturbation/GenePerturbation.ipynb , which implements the combined workflow for CellFM and GEARS as defined in https://github.com/biomed-AI/CellFM/blob/main/tutorials/Perturbation/gears/model.py . To run GEARS independently, you can set "pretrained_model=None" in the model configuration (https://github.com/biomed-AI/CellFM/blob/main/tutorials/Perturbation/gears/model.py). Additionally, we have included new benchmark experiments for CellFM+CellOT, accessible at https://github.com/biomed-AI/CellFM/blob/main/tutorials/ChemicalPerturbation .

Reviewer #2 (Remarks to the Author):

The writing and figure quality have significantly improved in the current version. However, based on the updated results, this study's novelty and the improvements in benchmarking performance remain relatively limited when compared to existing works.
Response: Thanks for your positive feedback on the improved writing and figure quality. We would like to highlight that our method has comprehensively made quotative analysis of existing single cell foundation models: seven models (while scGPT, geneCompass, and scFoundation only compared 1-3 baseline models in their works) over six benchmark tasks (only 3-4 quantitative tasks by other works), and our method consistently outperformed other models because of the large-scale human training datasets, and large model parameters. We sincerely appreciate the reviewer's constructive feedback and have revised the manuscript to more clearly articulate these key contributions.

Major:

1. The authors have addressed my question, and I have no more comments.

Response: We sincerely thank the reviewer for their valuable feedback and support. We are glad that our revisions addressed your concerns.

2. The authors have addressed my question. However, in "We identified the cell types for approximately 70 million cells in the training datasets, as the remaining cells were not classified in their original papers", are the authors saying, "We identified the cell types for approximately 70 million cells in the training datasets, as these cells were not classified in their original papers"? Additionally, a typo in Figure S1, is "Brest" "Breast"?

Response: Thanks for your valuable comments. We didn't annotate the cell types, and all cell types were annotated by the provided datasets. As a result, about 70 out of 100 million cells have been annotated by the original datasets. We corrected the typo, and have clarified it as:

"

Among the dataset, approximately 70 million cells had annotated cell types.

"

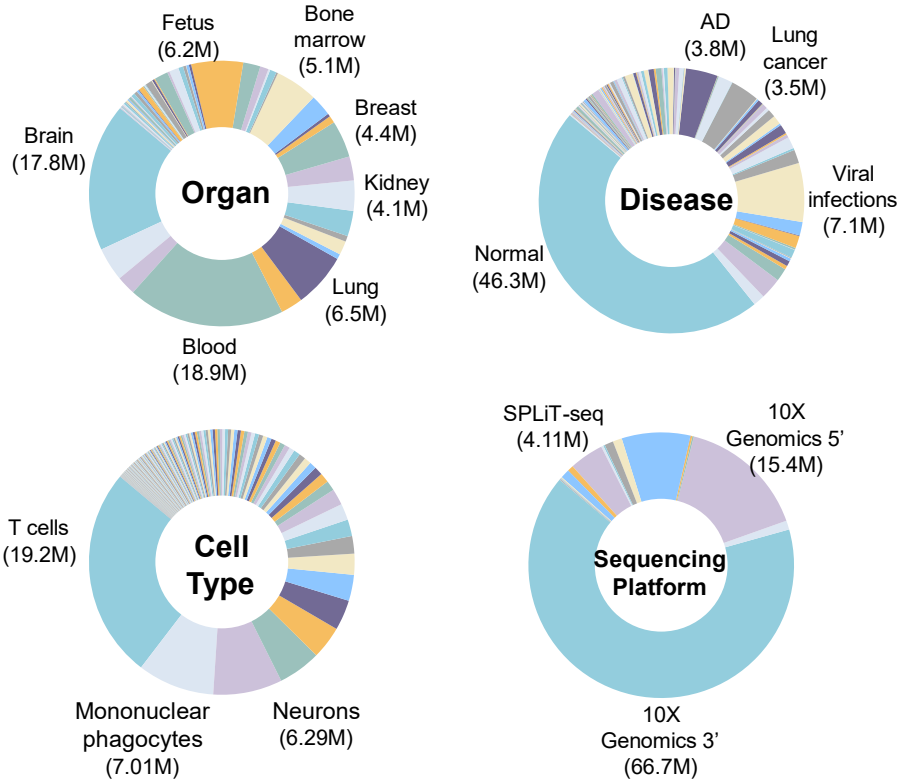


Fig. S1 Statistical summaries of the datasets' metadata, including sequencing platform, organ, disease, and cell type, are illustrated through pie charts.

3. The authors have addressed my question. However, in Figure 2d, there are some shades in the CellFM legend.

Response: Thanks for your valuable suggestions. We have removed the shades in Figure 2d.

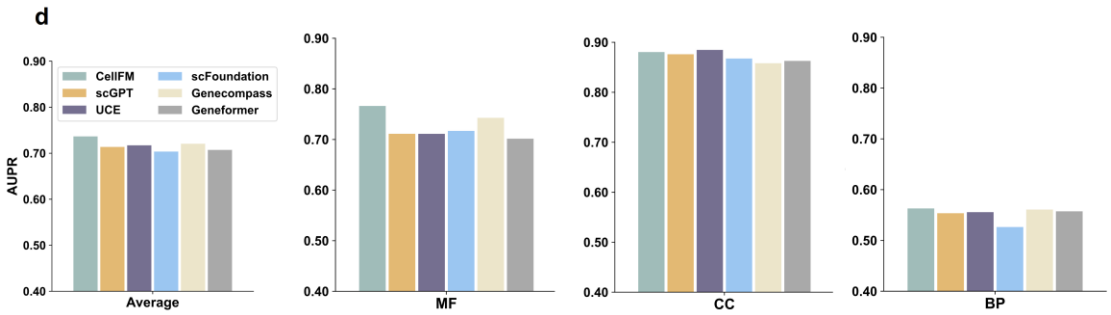


Fig. 2d. AUPR values for multiple gene functions predicted by CellFM and other competing single-cell foundation models on GO data, where each gene is annotated with numerous functions. MF: Molecular Function, CC: Cellular Component, and BP: Biological Process.

4. The authors have addressed my question, and I have no more comments.

Response: We sincerely thank the reviewer for their valuable feedback and support. We are glad that our revisions addressed your concerns.

5. The authors have addressed my question, and I have no more comments.

Response: We sincerely thank the reviewer for their valuable feedback and support. We are glad that our revisions addressed your concerns.

6. The authors have addressed my question, and I have no more comments.

Response: We sincerely thank the reviewer for their valuable feedback and support. We are glad that our revisions addressed your concerns.

7. The authors have addressed my question, and I have no more comments.

Response: We sincerely thank the reviewer for their valuable feedback and support. We are glad that our revisions addressed your concerns.

8. The authors have addressed my question, and I have no more comments.

Response: We sincerely thank the reviewer for their valuable feedback and support. We are glad that our revisions addressed your concerns.

9. The authors have addressed my question, and I have no more comments.

Response: We sincerely thank the reviewer for their valuable feedback and support. We are glad that our revisions addressed your concerns.

10. The authors have addressed my question, and I have no more comments.

Response: We sincerely thank the reviewer for their valuable feedback and support. We are glad that our revisions addressed your concerns.

11. For “the current attention map fails to capture the gene relationships among CCNB2, CCNA2, TOP2A, MKI67, KPNA2, and CENPF due to their low attention scores. In the future, we plan to explore new explainability techniques to address this limitation.” I do think this question should be addressed in this study, especially after authors’ answer to question 15 showed that there is only minor improvements over existing commonly used model.

Response: Thanks for your valuable comments. The partial relationships between CCNB2, CCNA2, TOP2A, MKI67, KPNA2, and CENPF weren’t captured likely because Figure 5d focused on capturing relationships with SPI1. When not designating perturbation targets in CellFM (i.e., during regular gene recovery tasks), we could observe relations between these genes as Supplementary Figure S24a. In addition, based on gene embeddings derived from CellFM, the cosine similarity scores of these genes (Supplementary Figure S24b) were significantly higher compared to those of other genes, indicating strong intrinsic relationships among them can be learned by CellFM. We have added these results into “Results” section as follows:

“

The partial relationships between CCNB2, CCNA2, TOP2A, MKI67, KPNA2, and CENPF weren’t captured likely because Figure 5d focused on capturing relationships with SPI1. When not designating perturbation targets in CellFM (i.e., during regular gene recovery tasks), we could observe relations between these genes as Supplementary Figure S24. In addition, based on gene embeddings derived from CellFM, the cosine similarity scores of these genes were significantly higher compared to those of other genes, indicating strong intrinsic relationships among them can be learned by CellFM.

”

(a) (b)

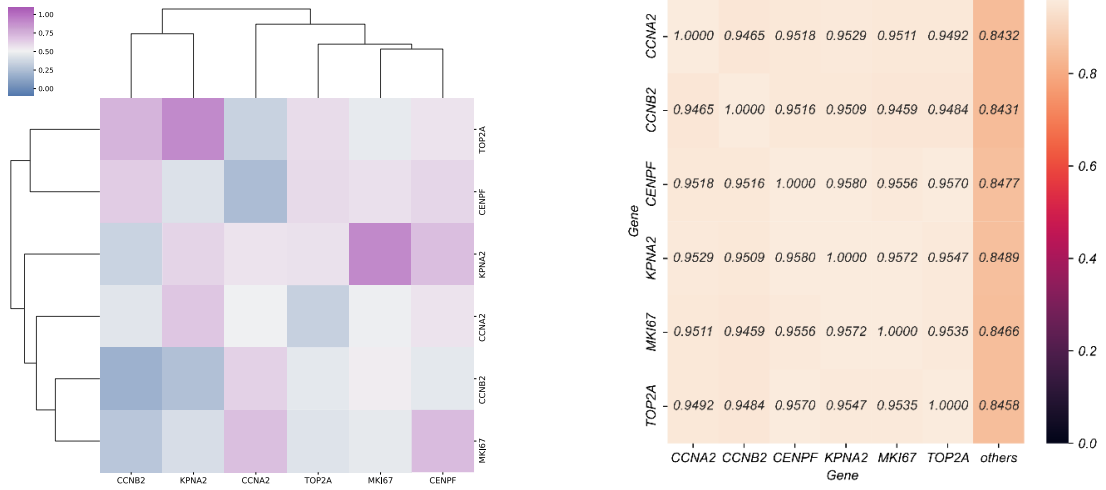


Fig. S24. (a) Attention scores of the genes CCNB2, CCNA2, TOP2A, MKI67, KPNA2, and CENPF, identified through gene recovery in CellFM without prior specification of the perturbation gene SPI1. (b) Based on gene embeddings derived from CellFM, the cosine similarity scores among CCNB2, CCNA2, TOP2A, MKI67, KPNA2, and CENPF. The "others" category denotes the cosine

similarity of the remaining genes excluding these six.

12. It's not that meaningful and persuasive to show the Figure 5f KEGG pathway enrichment analysis. I'd suggest the following two options: 1) remove Figure 5f, and directly mention specific target genes and their relevant studies in the manuscript, 2) add the detailed gene names to each enriched pathway name, to show whether these enriched pathways were contributed by the same genes or not. Additionally, Figure 4e is also not informative, the scores are quite large and close to each other. The authors need to provide the direct evidence from public SPI1 ChIP-seq datasets, e.g., snapshots of ChIP-seq BigWig files visualized in IGV genome browser. The same results should be provided for the transcription factor (TF) JUN in Figure S20 as well. Besides, in Figure S20b, the label of the gene between FRG1 and KNPA2 is not shown. I'd suggest building a TF-TF network based on the similarities of influenced genes (probably not just top influenced 20 genes), this analysis would be useful for biologists to understand the gene regulatory mechanisms in cells.

Response: Thanks for your valuable suggestions. We have incorporated detailed gene names for each enriched pathway (Figure 5f), explicitly demonstrating whether these pathways are driven by common or distinct gene sets. Specifically, the presence of common genes, such as *CCNB2* and *CCNA2*, across multiple pathways including the cell cycle, cellular senescence, and progesterone-mediated oocyte maturation, suggests functional overlap in cellular proliferation and regulation. This overlap indicates that these genes may play multifaceted roles in processes related to cell growth, division, and differentiation. Conversely, certain pathways, such as platinum drug resistance and osteoclast differentiation, are characterized by unique gene contributions. For instance, *TOP2A* and *BIRC5* are specifically associated with platinum drug resistance, while *JUND* is uniquely implicated in osteoclast differentiation.

To further directly demonstrate the relationships between *SPI1* and the *SPI1* perturbed genes, we obtained the *SPI1* ChIP-seq BigWig files (SRX2770855) from the public ChIP-Atlas database (https://chip-atlas.org/peak_browser) and visualized them using the Integrative Genomics Viewer (IGV). To select the top 5 genes with the strongest *SPI1* binding, we identified those with the highest MACS2 scores, indicating the most significant *SPI1* binding at their genomic loci. These genes were prioritized based on peak intensity and the extent of overlap with key genomic regions. The IGV snapshots allowed us to visually confirm the presence and strength of *SPI1* binding at these loci, providing direct evidence of *SPI1's* regulatory impact on these genes. As shown in Supplementary Figure S25, the results clearly show the binding peaks of *SPI1* at the genomic loci of the target genes, indicating potential regulatory interactions.

We also obtained the ChIP-seq BigWig files (SRX10976151) for the *JUN* transcription factor. The ChIP-seq data for *JUN* were also obtained from publicly available datasets and visualized using IGV. Representative snapshots of *JUN* binding sites across various genomic regions are provided in the Supplementary Figure S28. These visualizations highlight significant ChIP-seq peaks, offering direct evidence of *JUN*'s binding to specific genomic loci.

Additionally, we have revised Figure S27 (previously Figure S20) and ensured that all gene labels

are now clearly displayed in the updated version.

To construct a comprehensive gene-transcription factor (TF) interaction network, we utilized the TRRUST database (PMID:29087512) (http://www.grnpedia.org/trrust) to identify the top 300 genes most influenced by *SPI1*. As shown in Supplementary Figure S26, the network diagram illustrates the visualization of these interactions, where nodes represent genes or transcription factors, and edges denote the interactions between them. Red nodes represent transcription factors, while blue nodes represent genes. Notably, genes such as *MYC* emerged as hub genes, interacting with multiple key TFs, including *CTNNB1*, *JUNB*, and *CCNB1* (PMID: 20300972, 18838534, 11522645). This network analysis provides valuable insights into the regulatory mechanisms underlying *SPI1*-mediated gene expression, highlighting potential transcription factor interactions that may drive downstream cellular processes.

These results have been added into the “Results” section as follows:

“

To further directly demonstrate the relationships between *SPI1* and the *SPI1* perturbed genes, we obtained the *SPI1* ChIP-seq BigWig files (SRX2770855) from the public ChIP-Atlas database (https://chip-atlas.org/peak_browser) and visualized them using the Integrative Genomics Viewer (IGV). To select the top 5 genes with the strongest *SPI1* binding, we identified those with the highest MACS2 scores, indicating the most significant *SPI1* binding at their genomic loci. These genes were prioritized based on peak intensity and the extent of overlap with key genomic regions. The IGV snapshots allowed us to visually confirm the presence and strength of *SPI1* binding at these loci, providing direct evidence of *SPI1's* regulatory impact on these genes. As shown in Supplementary Figure S25, the results clearly show the binding peaks of *SPI1* at the genomic loci of the target genes, indicating potential regulatory interactions.

”

“

We also obtained the ChIP-seq BigWig files (SRX10976151) for the *JUN* transcription factor. The ChIP-seq data for *JUN* were also obtained from publicly available datasets and visualized using IGV. Representative snapshots of *JUN* binding sites across various genomic regions are provided in the Supplementary Figure S28. These visualizations highlight significant ChIP-seq peaks, offering direct evidence of *JUN*'s binding to specific genomic loci.

”

“

To construct a comprehensive gene-transcription factor (TF) interaction network, we utilized the TRRUST database (PMID:29087512) (http://www.grnpedia.org/trrust) to identify the top 300 genes most influenced by *SPI1*. As shown in Supplementary Figure S26, the network diagram illustrates

the visualization of these interactions, where nodes represent genes or transcription factors, and edges denote the interactions between them. Red nodes represent transcription factors, while blue nodes represent genes. Notably, genes such as *MYC* emerged as hub genes, interacting with multiple key TFs, including *CTNNB1*, *JUNB*, and *CCNB1* (PMID: 20300972, 18838534, 11522645). This network analysis provides valuable insights into the regulatory mechanisms underlying *SPI1*-mediated gene expression, highlighting potential transcription factor interactions that may drive downstream cellular processes.

”

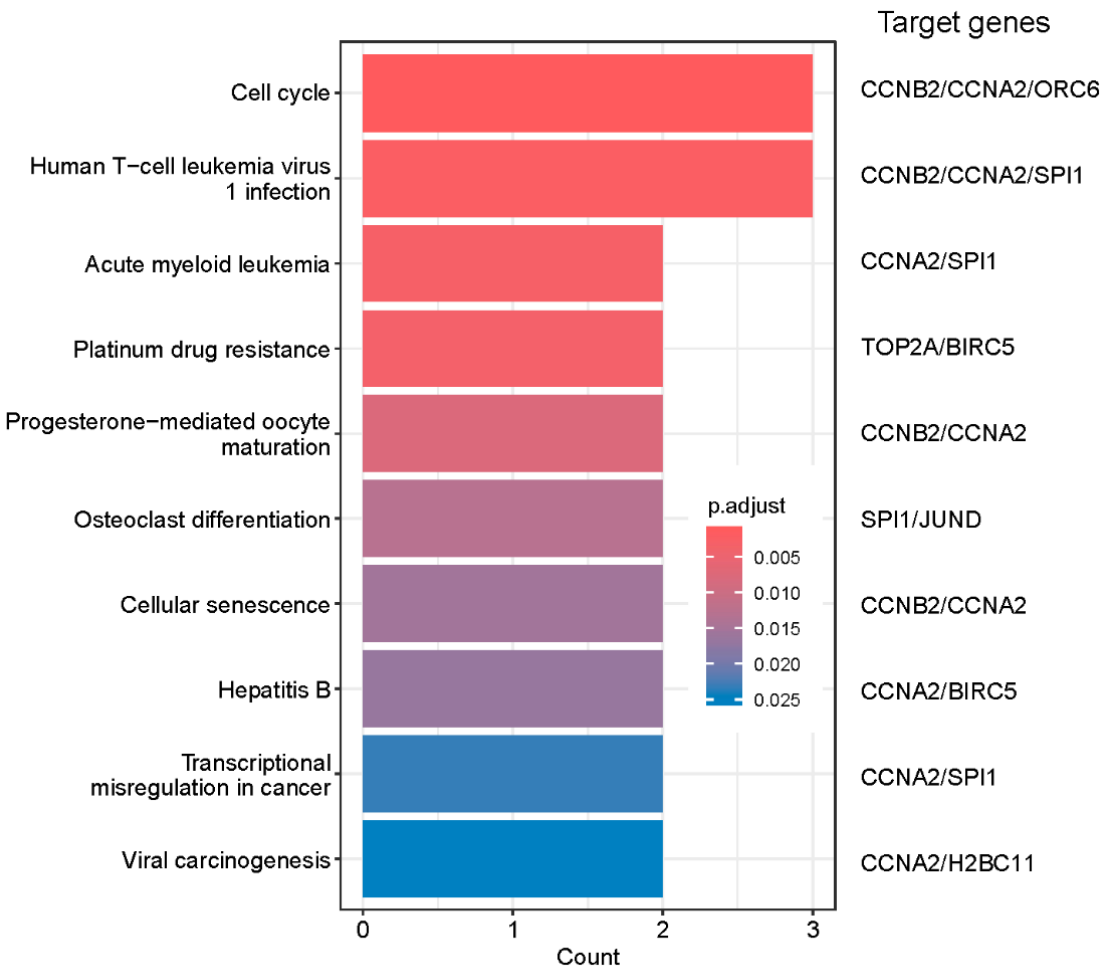


Fig. 5f. The heatmap illustrates the pathways mapped using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database, based on the top 20 genes most significantly impacted by *SPI1*.

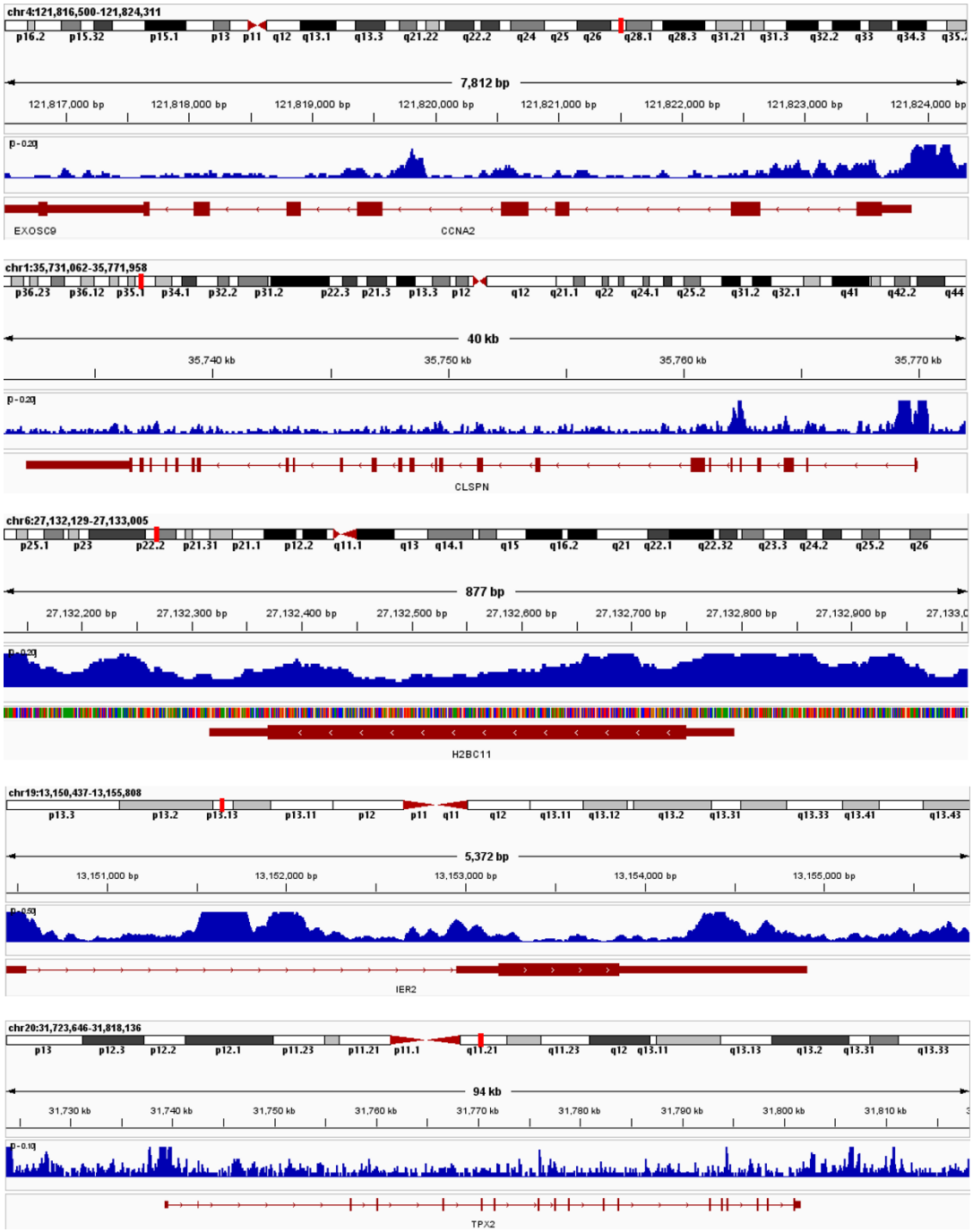


Fig. S25. **Visualization of *SPI1* Binding Sites in the Top 5 Genes Based on ChIP-seq Data.** The figure shows the *SPI1* ChIP-seq data visualized in the IGV for the top 5 genes identified based on the highest *SPI1* binding intensity. Each panel displays a genomic region with *SPI1* binding peaks, indicated by blue bars, representing the intensity of *SPI1* binding at specific loci.

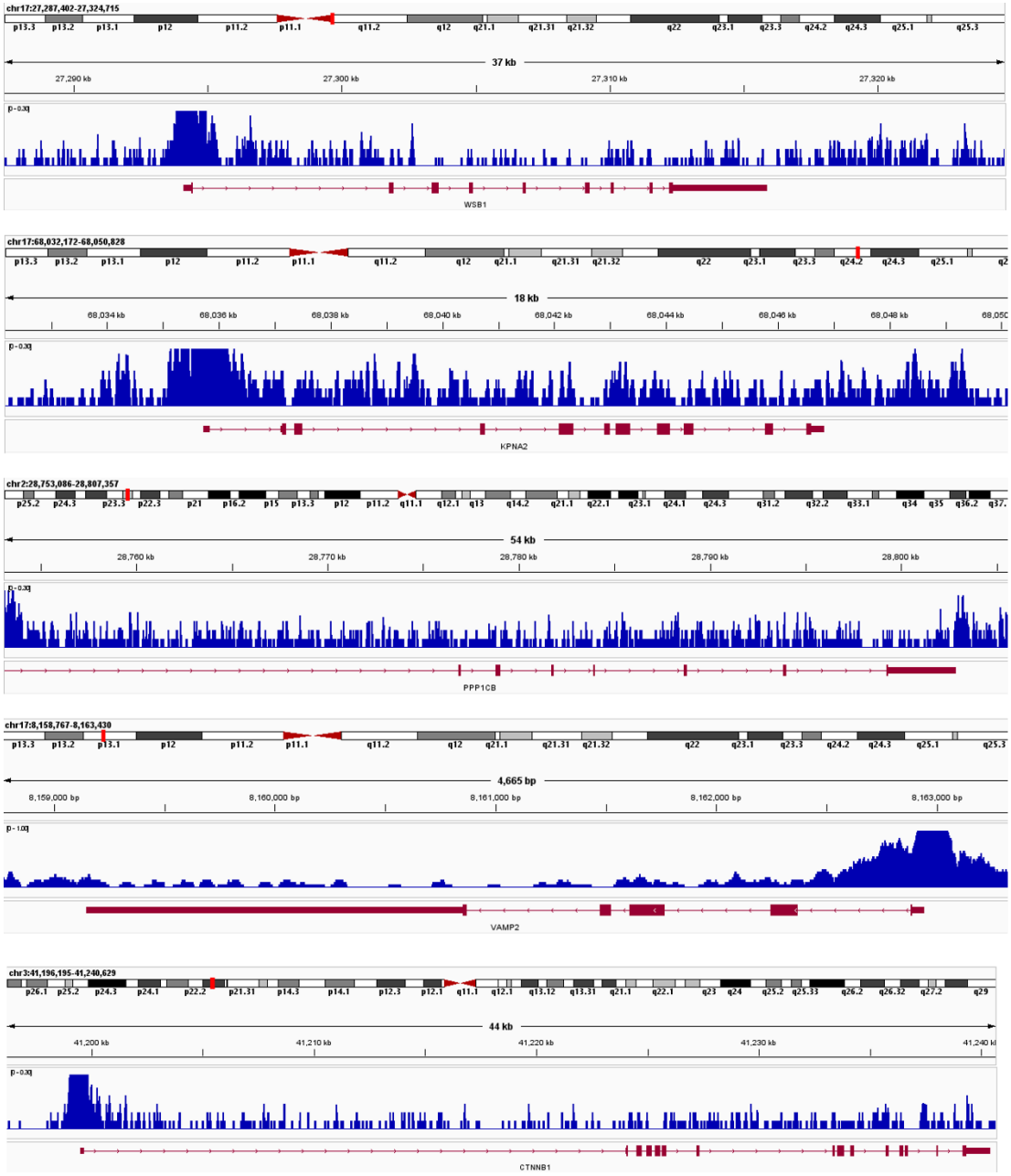


Fig. S28. **Visualization of *JUN* Binding Sites in the Top 5 Genes Based on ChIP-seq Data.** The figure shows the *JUN* ChIP-seq data visualized in the IGV for the top 5 genes identified based on the highest *JUN* binding intensity. Each panel displays a genomic region with *JUN* binding peaks, indicated by blue bars, representing the intensity of *JUN* binding at specific loci.

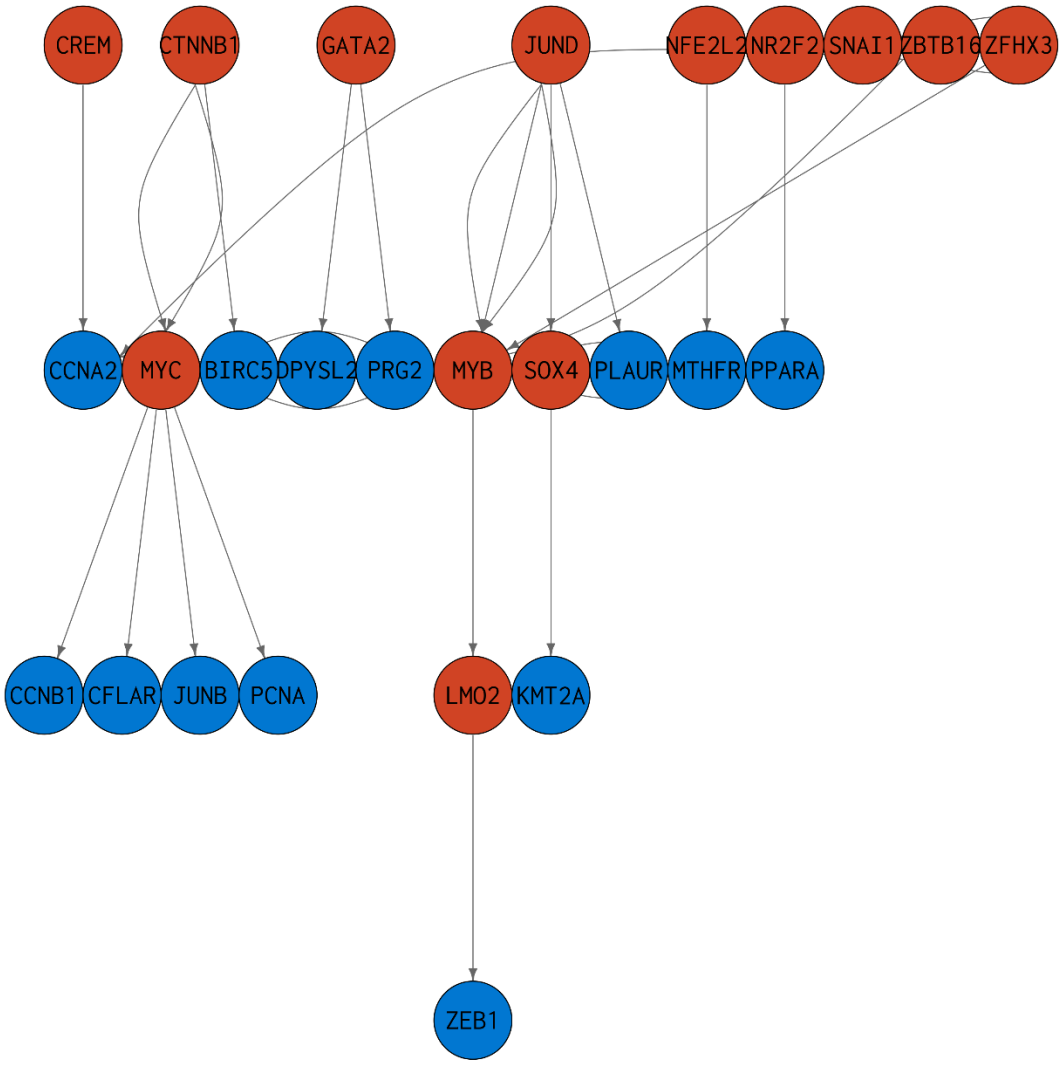


Fig. S26. **TF-TF Regulatory Network of the Top 300 Genes Affected by *SPI1* Perturbation.** Nodes represent either genes or transcription factors, with edges indicating the regulatory interactions between them. Red nodes represent transcription factors, while blue nodes represent target genes.

13. The authors have addressed my question, and I have no more comments.

Response: We sincerely thank the reviewer for their valuable feedback and support. We are glad that our revisions addressed your concerns.

14. The authors have addressed my question, and I have no more comments.

Response: We sincerely thank the reviewer for their valuable feedback and support. We are glad that our revisions addressed your concerns.

15. The updated results showed that there are minor improvements over the classic Transformer model. The novelty of this study and the improvements in benchmarking performance are relatively limited compared to existing studies.

A typo in “Based on the findings in the ”Scaling of data and model size” section (Supplementary Note 4), we used CellFM with 8,000 model parameters (CellFM-80M) for cell annotation and batch effect correction tasks.” “8,000” should be the wrong number.

Response: Thanks for your valuable comments. We would like to highlight that our method has comprehensively made quotative analysis of existing single cell foundation models: seven models (while scGPT, geneCompass, and scFoundation only compared 1-3 baseline models in their works) over six benchmark tasks (only 3-4 quantitative tasks by other works), and our method consistently outperformed other models because of the large-scale human training datasets, and large model parameters. Additionally, we have corrected the typo in the manuscript as follows:

"

Based on the findings in the 'Scaling of data and model size' section (Supplementary Note 4), we used CellFM with 80 million model parameters (CellFM-80M) for cell annotation and batch effect correction tasks.

"

16. The authors have addressed my question. However, miR-29a is micro-RNA, not lncRNA. The authors need to revise and simplify the corresponding descriptions.

Response: Thanks for your valuable comments. We have removed the related description of miR-29a and simplified the corresponding descriptions in the revised manuscript.

Minor:

I have no more comments.

Response: Thanks for your valuable comments. We appreciate your efforts in reviewing and providing helpful comments.

Reviewer #3 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Response: We sincerely thanks for your time and effort in co-reviewing our manuscript. We appreciate Nature Communications' initiative to support early career researchers in peer review, and we appreciate your contribution to the review process.

Reviewer #4 (Remarks to the Author):

Overview:

Following the first round of reviews the authors have made a number of improvements to their manuscript, including adding benchmarks, ablation studies and clarifying text. However, there remain significant concerns over which benchmarks are chosen and how their results are conveyed to the reader. Additionally, there remain a number of statements or conclusions that are not backed up by experiments that should be modified or removed.

Response: Thanks for your feedback and for acknowledging the improvements made to the manuscript. We have carefully revised the benchmarks to improve the clarity of the results.

Benchmarking Setup and Presentation

The benchmarks used in the paper are insufficient or not useful for understanding model performance. Benchmarks should measure model performance in realistic use cases. No one annotates cell types for a random proportion of a dataset, and then uses their annotations to annotate the rest of the data, except in the rare case of subsampling a very large dataset.

Newly collected data, when annotated, will come from new experiments and batches in almost all cases. Therefore, the random split cell type annotation task is unconnected to the real use cases of such models. In comparison, the more standard benchmark, scIB, measures overall embedding quality, using a number of different metrics. While not a perfect set of metrics by any means, SCIB more closely reflects the real-world use of such a model, which will be used for generating embeddings that are used in downstream tasks like clustering. The authors already modified the manuscript to include scIB scores for one single dataset, PBCM 10k, but did not benchmark the other datasets.

Reviewer #4 comment (reproduced verbatim): Newly collected data, when annotated, will come from new experiments and batches in almost all cases. Therefore, the random split cell type annotation task is unconnected to the real use cases of such models.

Response: Thanks for your valuable suggestions. We would like to clarify that our evaluation of CellFM primarily focuses on cross-batch cell type annotation (Figure 4e-f), aligning with real-world applications. By comparison, the intra-dataset experiments serve as a supplementary evaluation to provide a more comprehensive assessment of the model's capabilities. This is common in current studies as also used in scGPT (Nature Methods) and scEval (Benchmark for evaluating single-cell foundation models), since intra-dataset scenarios represent a more controlled setting that allow for more comprehensive comparison with existing methods. Therefore, we have retained these results in our manuscript to present a complete evaluation of CellFM's performance.

Reviewer #4 comment (reproduced verbatim): The authors already modified the manuscript to include scIB scores for one single dataset, PBCM 10k, but did not benchmark the other datasets.

Response: Thanks for your valuable suggestions. Following your suggestion, we have expanded our evaluation by calculating the scIB scores (AvgBio) for CellFM and the top three single-cell foundation models in the classification task on the inter-datasets. As illustrated in Supplementary Figure S15, CellFM achieved the highest average AvgBio scores, outperforming these competing models. These results further validate its superior performance in cell classification tasks. We have incorporated these results into the "Results" section as follows:

“

We also evaluated the embedding quality of all single-cell foundation models on inter-datasets using the AvgBio metric. As shown in Supplementary Figure S15, CellFM outperformed competing models in AvgBio scores, aligning with its enhanced cell classification performance.

”

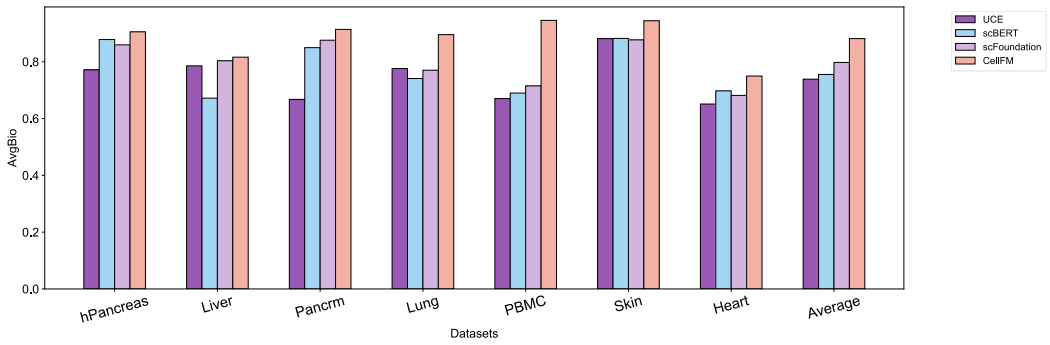


Fig. S15. A comparative analysis of AvgBio scores was performed, which were calculated based on the embeddings of test data generated by CellFM and the top three competing single-cell foundation models for classification tasks across inter-datasets.

Using scIB as the primary benchmark for embedding quality will also improve other questions about the benchmarking setup, including the choice of a 3 layer MLP for cell type annotation, which is an overly complex network, again unreflective of most real world use cases.

Response: Thanks for your valuable suggestions. We incorporated the scIB metric to evaluate the impact of MLP layers in CellFM and found minimal performance variation when adjusting layers from one to three. Specifically, we measured AvgBio scores and classification accuracy across three case inter-datasets (BPMC, Liver, and Pancrm). Supplementary Figure S16 shows a positive correlation between classification accuracy and AvgBio scores, but both metrics exhibited only minor changes across layer configurations. Since most single-cell foundation models use similar settings, we retained the default settings. These results have been added to the “Results” section.

“

We incorporated the scIB metric to evaluate the impact of MLP layers in CellFM and found minimal performance variation when adjusting layers from one to three. Specifically, we measured AvgBio scores and classification accuracy across three case inter-datasets (BPMC, Liver, and Pancrm). Supplementary Figure S16 shows a positive correlation between classification accuracy and AvgBio scores, but both metrics exhibited only minor changes across layer configurations.

”

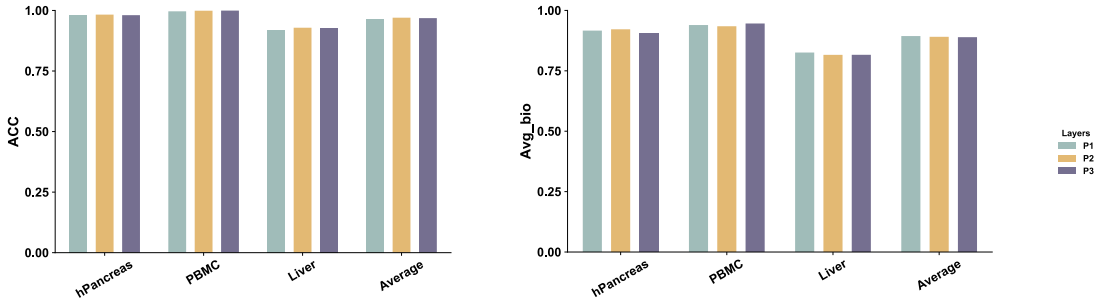


Fig. S16. Performance evaluation of CellFM with varying MLP layers for cell type annotation on three case inter-datasets (BPMC, Liver, and Pancrm). ACC and AvgBio scores are reported for configurations with 1, 2, and 3 MLP layers, denoted as P1, P2, and P3, respectively.

For the additional batch effect benchmark added, the authors only tested one dataset, PBMC 10k. PBMC 10k is a poor choice for this test since it contains a small number of cell types and only two batches. For two other datasets, liver and pancreas, the authors only compare CellFM and scFoundation, and only use UMAP, which is not empirical. The authors should perform more scIB benchmarks on more complicated datasets with many batches and cell types, such as various atlas datasets like the Human Brain Cell Atlas. It would also be beneficial to check whole organism datasets like Tabula Sapiens to investigate cross-tissue performance.

Response: Thanks for your valuable suggestions. We followed scGPT to use the PBMC 10k dataset. Now, we included the human brain cell atlas (DOI: 10.1126/science.add7046) and Tabula Sapiens including two versions (version1 from DOI: 10.1126/science.abl4896 and version2 from doi.org/10.1101/2024.12.03.626516) to further assess CellFM's performance and competing single-cell foundation models. As shown in Supplementary Figure S19, CellFM achieved the highest average AvgBio score of 0.60 across these datasets, outperforming the second-best method, UCE, by 2.1%. These results have been added to the "Results" section as follows:

“We evaluated CellFM across multiple datasets, including PBMC 10k, the human brain cell atlas, and two versions of Tabula Sapiens. As demonstrated in Supplementary Figure S19, CellFM achieved the highest average AvgBio scores on these datasets, outperforming the second-best method, UCE, by 2.1%.”

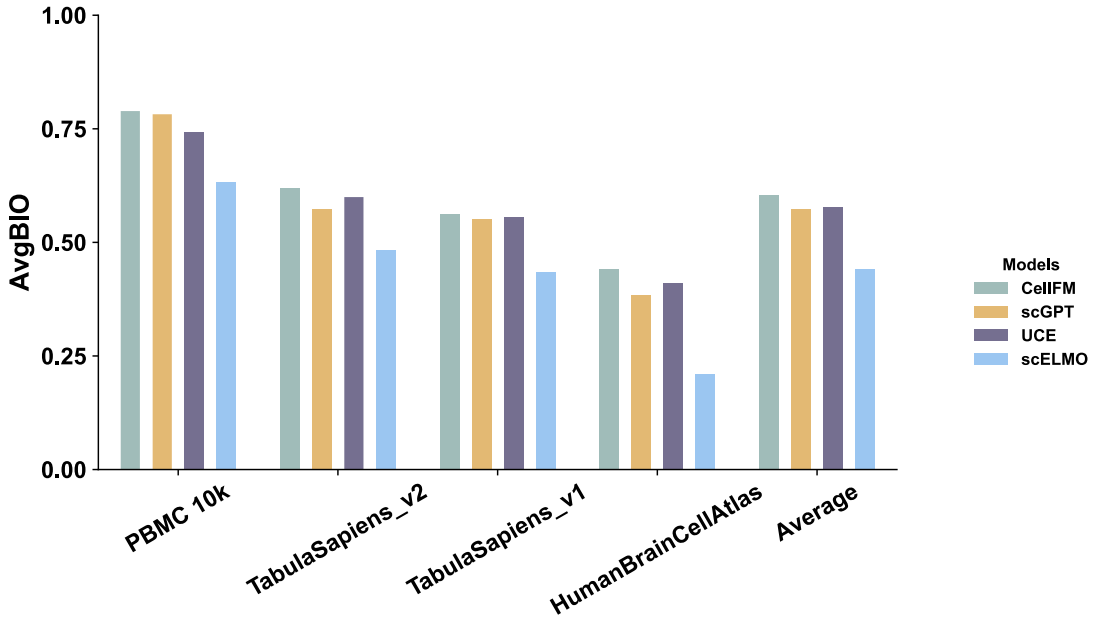


Fig. S19. Batch effect benchmark comparison of CellFM, scGPT, scELMo, and UCE on the PBMC 10k dataset, human brain cell atlas, and Tabula Sapiens. Performance was evaluated using the AvgBio score.

For the cell type annotation benchmark, the authors did not include statistics representing the uncertainty of the results, which is critical for comparing models. Out of curiosity, I'm particularly interested in how this annotation framework works to distinguish between exhausted and activated CD8+ T cells, as discussed in [1][2].

[1] Pan-cancer mapping of single CD8+ T cell profiles reveals a TCF1:CXCR6 axis regulating CD28 co-stimulation and anti-tumor immunity, Cell reports Medicine 2024

[2] A Cancer Cell Program Promotes T Cell Exclusion and Resistance to Checkpoint Blockade, Cell 2024

Response: Thanks for your valuable suggestions. To represent the uncertainty of the cell type annotation results, we have now included the standard deviation (SD) for the classification accuracy across multiple runs on the more challenging inter-datasets (Supplementary Figure S12). Specifically, we re-ran all single-cell foundation models on each inter-dataset four times. The updated results, along with the original results reported in our manuscript, were used to replot the detailed average scores with SD in Supplementary Figure S12 and the average performance in Figure 4e-f. The results demonstrate that CellFM consistently outperformed competing models.

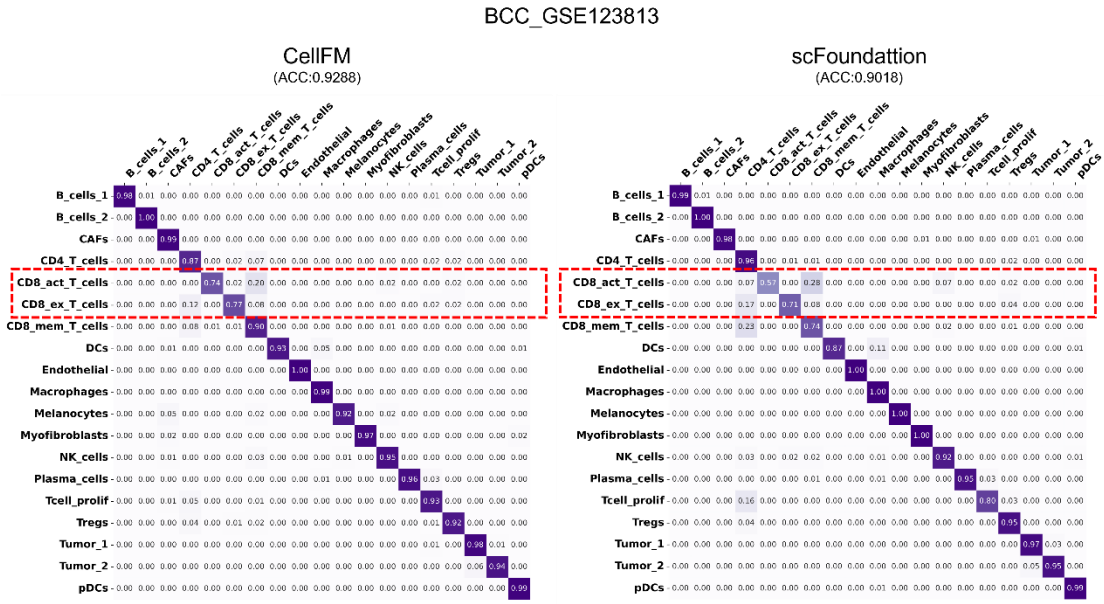
To further assess CellFM's capability to distinguish exhausted and activated CD8+ T cells, we evaluated it on the basal cell carcinoma (BCC) dataset (GSE123813) and the liver hepatocellular carcinoma (LIHC) dataset (GSE140228), both obtained from a database article (Nucleic Acids Res, PMID: 36321662) due to their availability in h5ad format and detailed cell type annotations. We did not use the datasets suggested by the reviewer [1][2] because the publicly available data from these studies does not provide sub-type annotations for exhausted and activated CD8+ T cells.

As shown in Supplementary Figures S13-S14, CellFM consistently outperformed other models across all cell types, achieving an average accuracy 2.3% higher than the second-best model, UCE. Notably, CellFM demonstrated exceptional performance in distinguishing exhausted and activated CD8+ T cells, surpassing UCE by an average of 6.5%. On the BCC_GSE123813 dataset, CellFM achieved accuracy scores of 77% and 74% for exhausted and activated CD8+ T cells, respectively, outperforming UCE by 6% and 7%. A similar trend was observed on the LIHC_GSE140228 dataset, further confirming CellFM's robustness in identifying these cell states. These results have been added to the “Results” section as follows:

“

To further assess CellFM's capability to distinguish exhausted and activated CD8+ T cells, we evaluated it on the basal cell carcinoma (BCC) dataset (GSE123813) and the liver hepatocellular carcinoma (LIHC) dataset (GSE140228), both obtained from a database article (Nucleic Acids Res, PMID: 36321662) due to their availability in h5ad format and detailed cell type annotations. As shown in Supplementary Figures S13-S14, CellFM consistently outperformed other models across all cell types, achieving an average accuracy 2.3% higher than the second-best model, UCE. Notably, CellFM demonstrated exceptional performance in distinguishing exhausted and activated CD8+ T cells, surpassing UCE by an average of 6.5%. On the BCC_GSE123813 dataset, CellFM achieved accuracy scores of 77% and 74% for exhausted and activated CD8+ T cells, respectively, outperforming UCE by 6% and 7%. A similar trend was observed on the LIHC_GSE140228 dataset, further confirming CellFM's robustness in identifying these cell states.

”



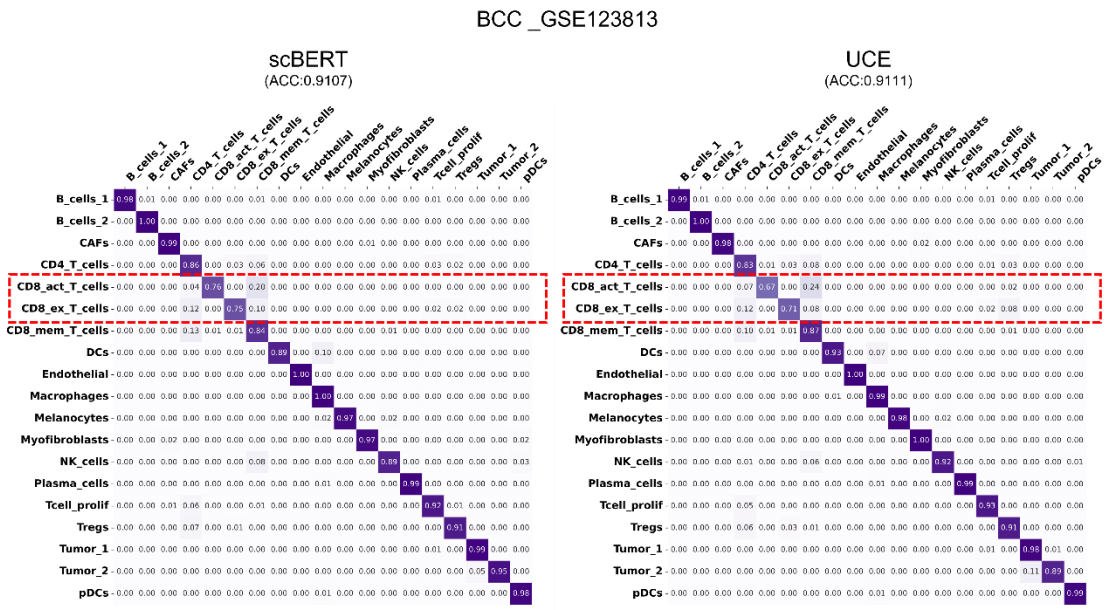
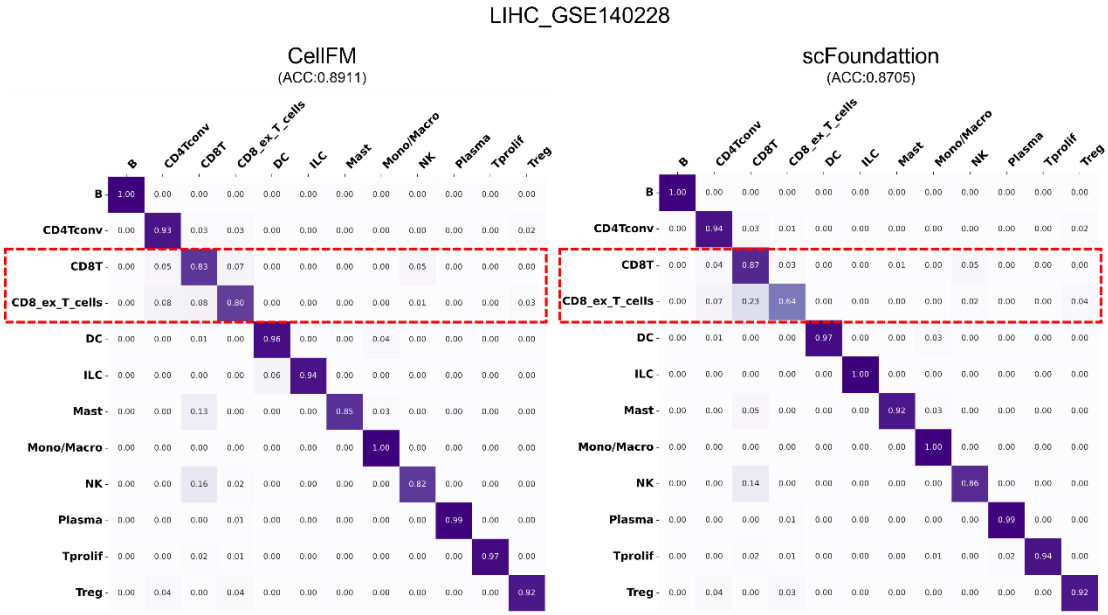


Fig. S13. Accuracy confusion matrices of CellFM and the top three competing methods—scFoundation, scBert, and UCE—on the BCC_GSE123813 dataset, which includes exhausted and activated CD8+ T cells. The overall accuracy (ACC) reflects the performance across all cell types, while the dotted red box highlights the classification performance specifically for exhausted and activated CD8+ T cells.



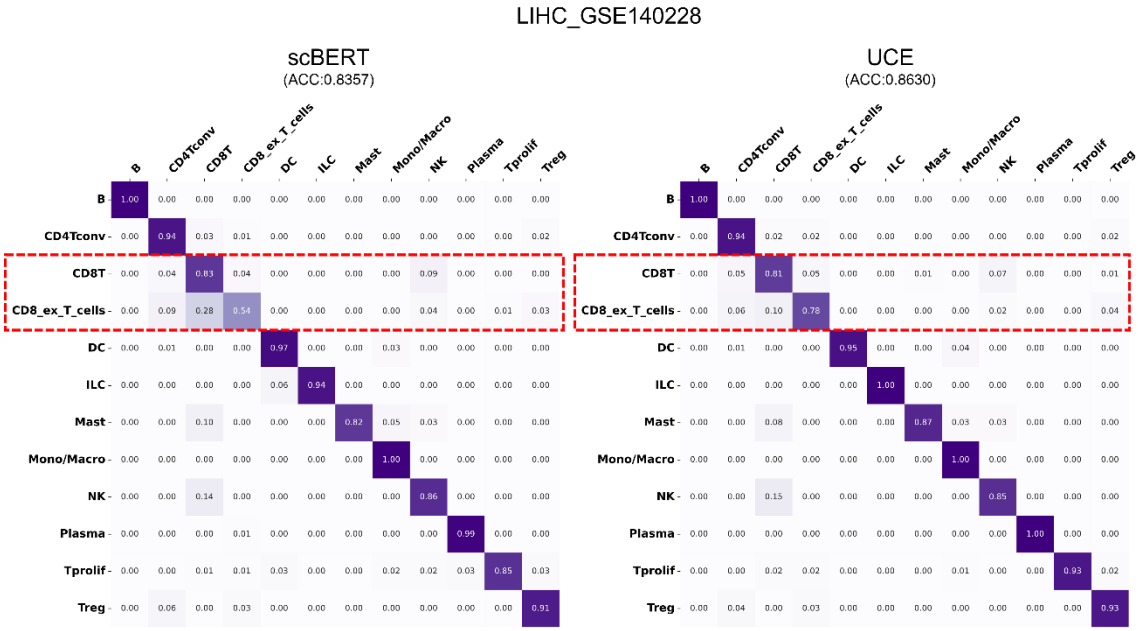


Fig. S14. Accuracy confusion matrices of CellFM and the top three competing methods—scFoundation, scBert, and UCE—on the LIHC_GSE140228 dataset, which includes exhausted and activated CD8+ T cells. The overall accuracy (ACC) reflects the performance across all cell types, while the dotted red box highlights the classification performance specifically for exhausted and activated CD8+ T cells.

Benchmarking Presentation

The presentation of benchmarking results in the paper is less satisfying.

The authors need to ensure that benchmarking results are presented in a straightforward and empirical manner. In Figures 4 ABEF the authors present more than 100 individual data points (15 experiments x 11 methods) and in Figure S12, Figure S13 dozens of data points. These experiments are incorrectly grouped into box plots by method, even though they are from totally different experiments. Each of these experiments individually also requires measurement of uncertainty (see previous section).

Overall, it is very difficult to compare results between methods. The authors should include a table with the actual numerical results of each benchmark, divided by dataset and method. When presenting these results, the authors need to make it clear in the description and on the figure how each method was assessed, specifically, if the underlying foundation model was finetuned or not (this is not clear in Figures S14, S15).

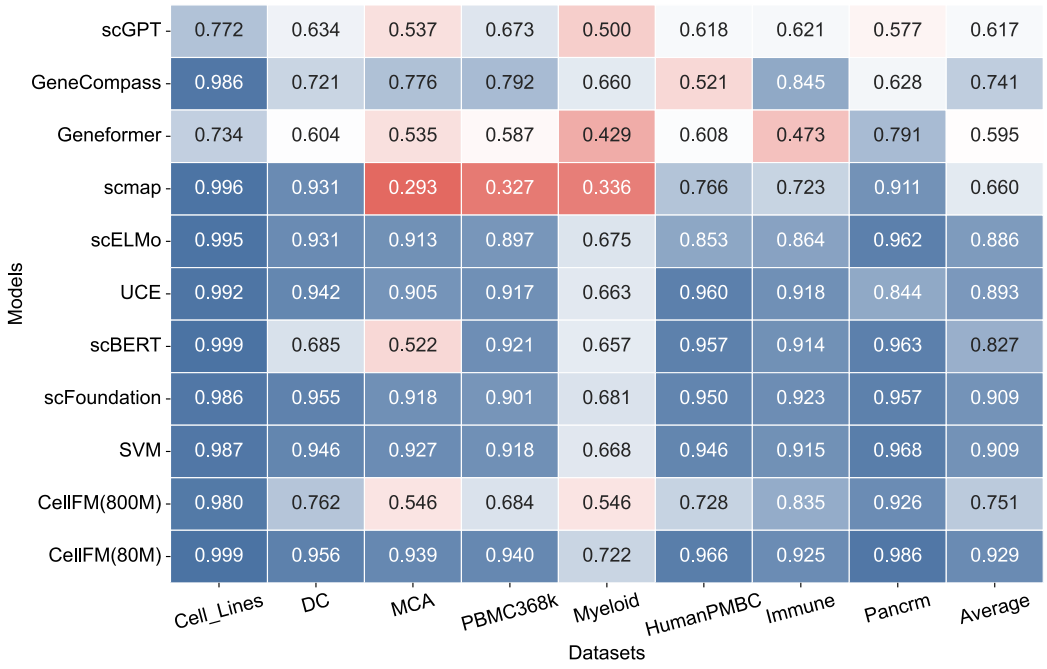
Finally, UMAP should not be used as a tool to compare embeddings from different models. Please remove this text from section 2.6 of the text and other sections which may mention it.

Response: Thanks for your valuable suggestions. The boxplot was used in the same form as many studies like scBert. Following your suggestion, we have made several revisions:

First, we have incorporated the standard deviation (SD) of classification accuracy across multiple runs for each inter-dataset to ensure a more robust and reliable representation of the results (Supplementary Figure S12 and Figure 4e-f). Specifically, we re-ran all single-cell foundation models on each inter-dataset four times. The updated results, combined with the original results reported in the manuscript, were used to generate detailed average scores with SD in Supplementary Figure S12 and to update the average performance metrics in Figure 4e-f. Additionally, the original Supplementary Figures S12-S13 have been revised and replotted as bar plots, now presented in Supplementary Figures S17-S18.

Second, we have replaced the UMAP visualizations (original Figures S14 and S15) with bar plots (Figure S19) to present the results more effectively. These results were generated in a zero-shot manner without fine-tuning the underlying foundation models. We have clarified it in the revised manuscript. In line with your suggestion, we have also removed all UMAP-related results (including Figures S14 and S15) and associated discussions from section 2.6 and other relevant sections of the manuscript.

(a)



(b)

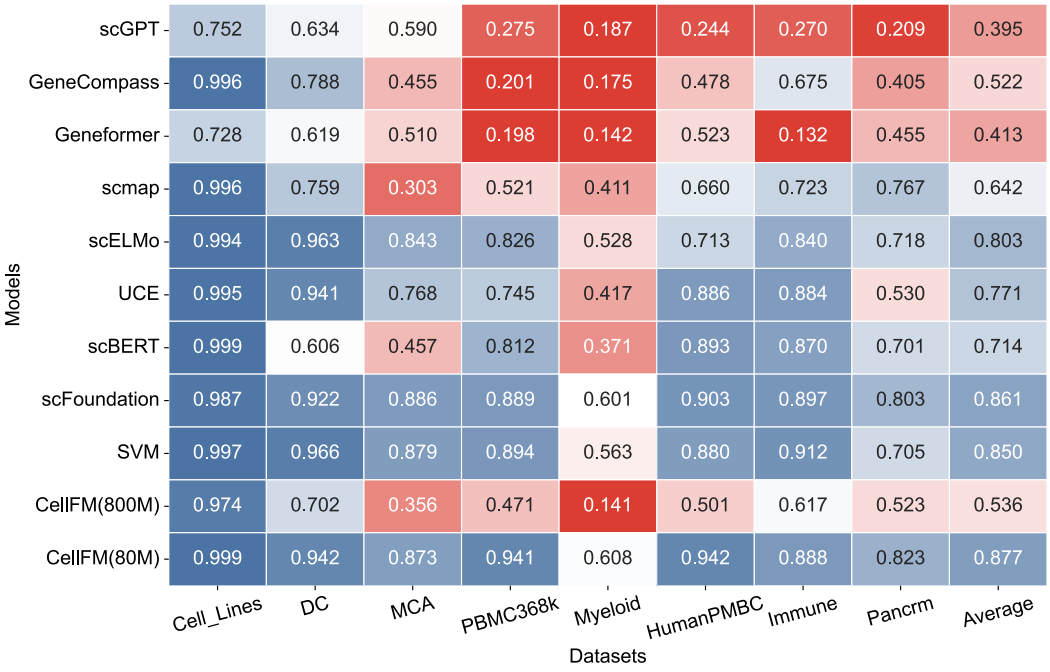
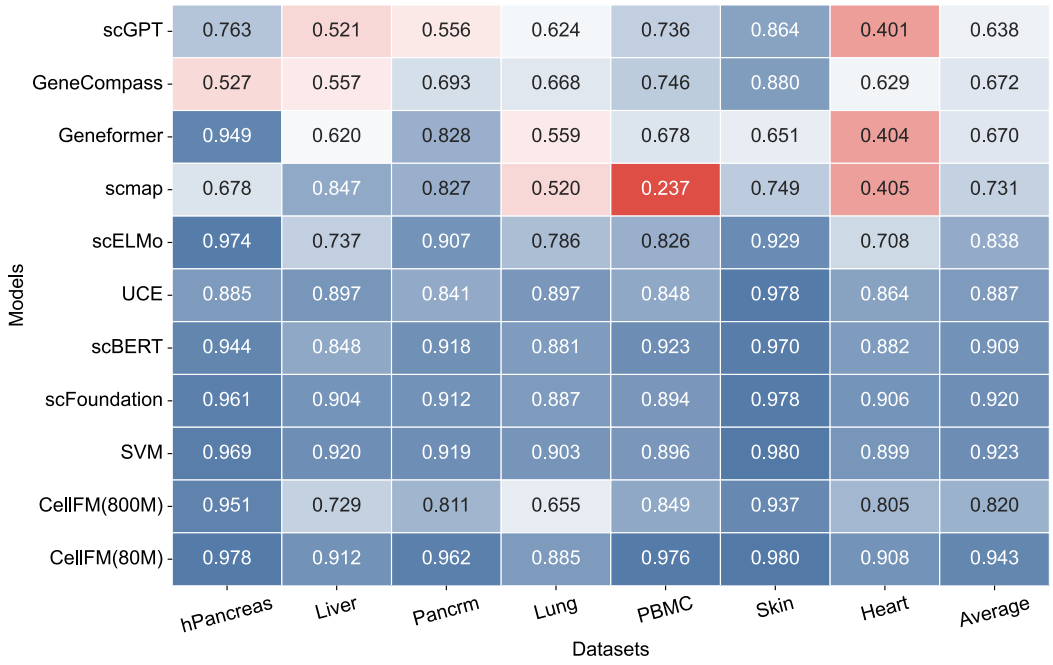


Fig. 4a-b. Performance comparison of CellFM and competing models using heatmap visualization. (a) Heatmap displaying classification accuracy across intra-datasets. (b) Heatmap showing Macro-F1 scores across the same intra-datasets.

(e)



(f)

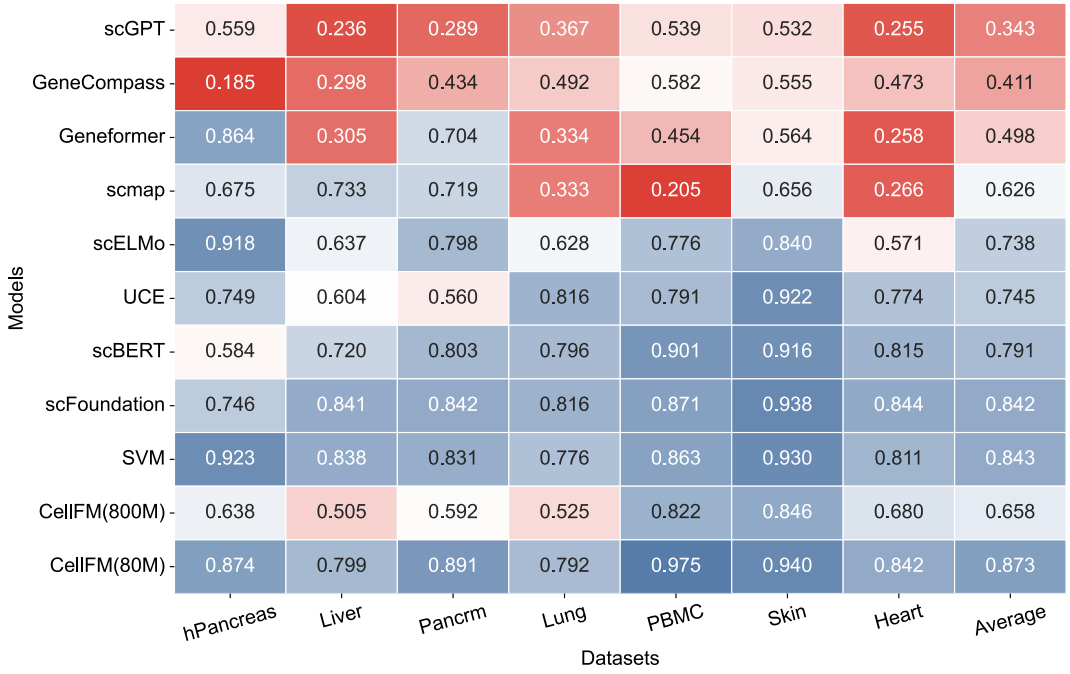
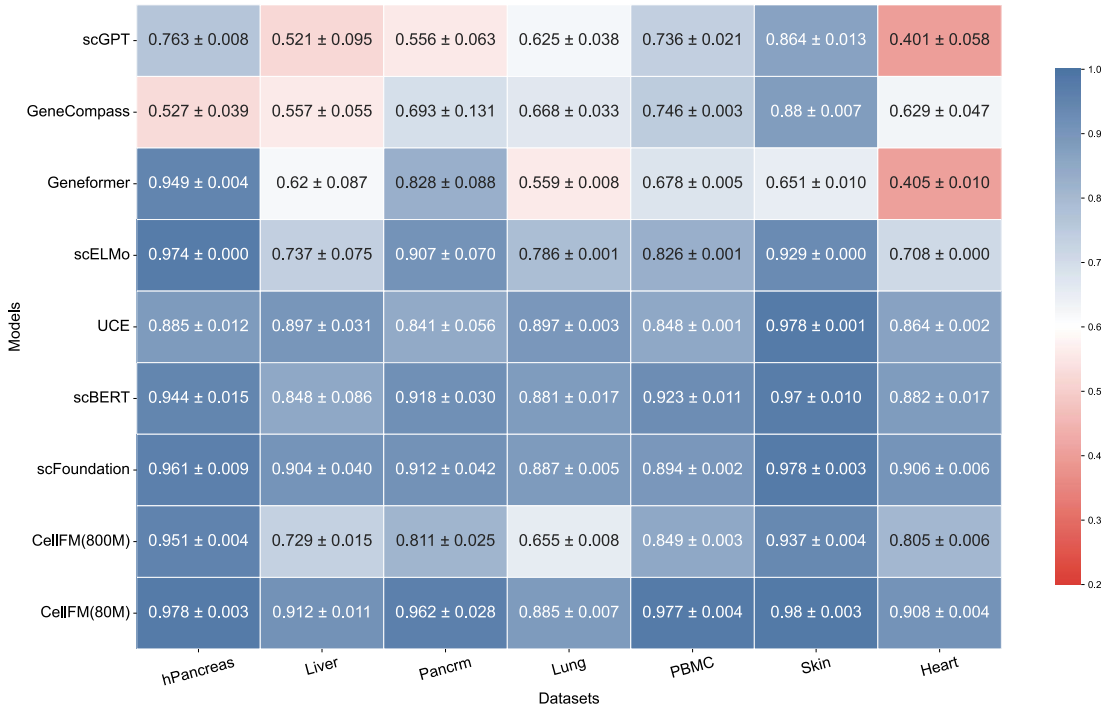


Fig. 4e-f. Performance comparison of CellFM and competing models using heatmap visualization. (e) Heatmap displaying classification accuracy across inter-datasets, with each value representing the average accuracy derived from five independent runs using different random seeds. (f) Heatmap illustrating Macro-F1 scores across the same inter-dataset, with each value representing the average accuracy derived from five independent runs using different random seeds.

(a)



(b)

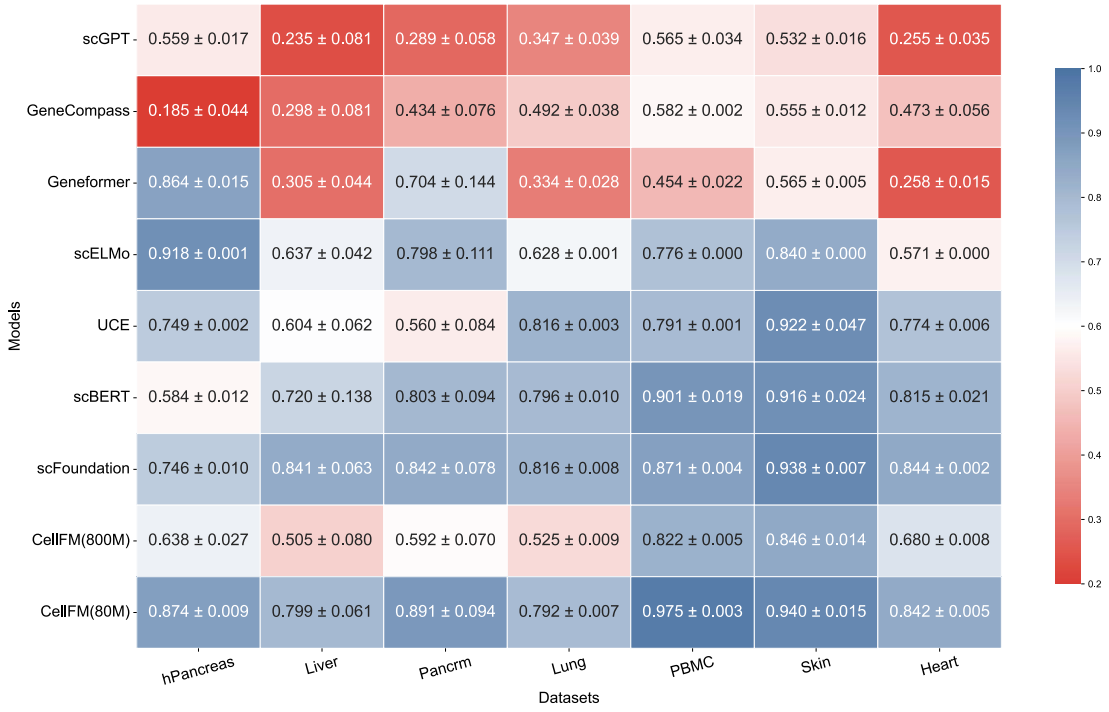


Fig S12. Comparative performance of each method on inter-datasets. (a) Accuracy scores and (b) Macro-F1 scores, with each value representing the mean and standard deviation (SD) calculated from five independent runs using different random seeds.

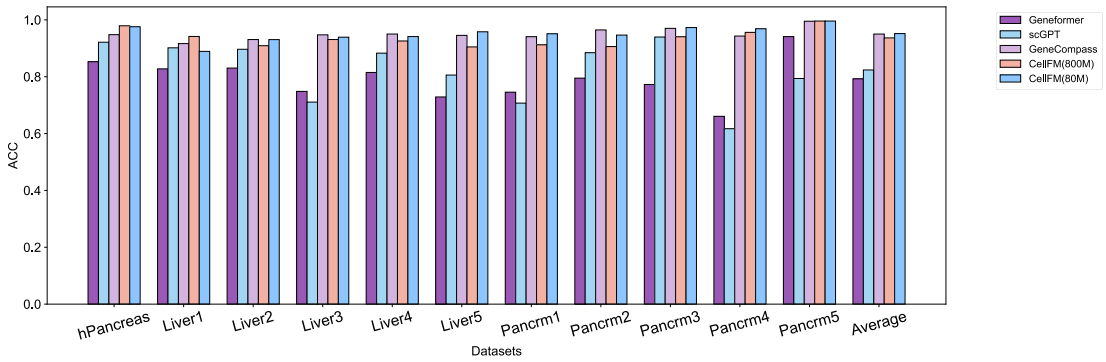


Fig. S17. The cell type annotation performance of Geneformer, scGPT, GeneCompass, and CellFM after fine-tuning with inter-datasets.

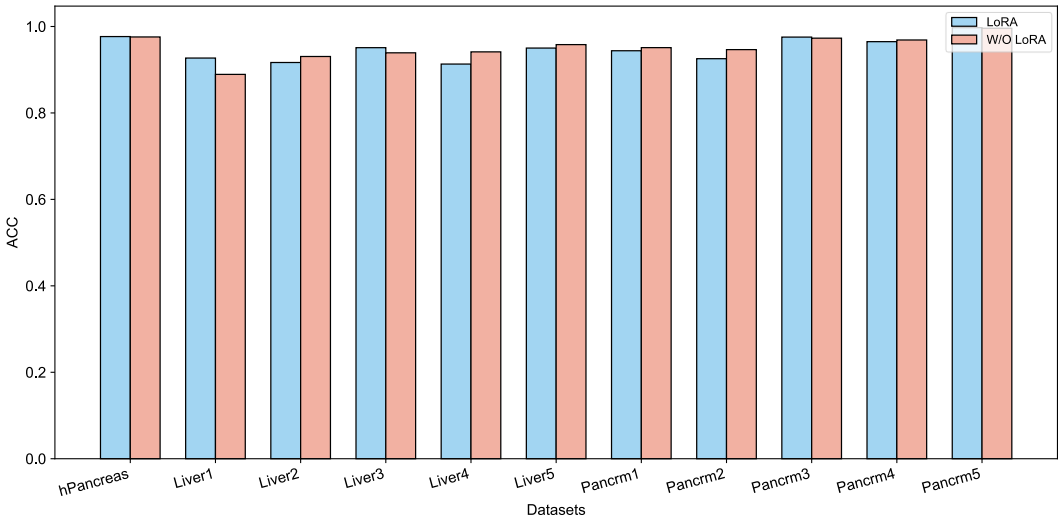


Fig. S18. The cell type annotation performance of CellFM after fine-tuning with or without LoRA on inter-datasets.

Single Species vs Multiple Species Training

The authors continue to claim that:

“While larger datasets spanning diverse species can enhance model generalization, the inherent inter-species differences may introduce noise and hinder model performance.” (Abstract)

In their response, the authors claim that:

“In fact, our study indicates human dataset-based models (scFoundation and our model) tend to outperform diverse species models (GeneCompass and UCE) in human test sets. Concretely, when a model is trained on such heterogeneous data, it may inadvertently learn features that are not relevant to humans, leading to overfitting and bias.”

These models are all trained on different data corpus, with different sizes and different architectures. In order to make this claim, the authors, or someone else, would need to run a proper ablation experiment testing specifically the impact of including other species' data in the training corpus. The authors of GeneCompass did this with their architecture and found that including data from both species improved results (GeneCompass figure 3A).

This claim should either be removed from the manuscript or confirmed empirically.

Reviewer #4 comment (reproduced verbatim): These models are all trained on different data corpus, with different sizes and different architectures. In order to make this claim, the authors, or someone else, would need to run a proper ablation experiment testing specifically the impact of including other species' data in the training corpus. The authors of GeneCompass did this with their architecture and found that including data from both species improved results (GeneCompass figure 3A). This claim should either be removed from the manuscript or confirmed empirically.
Response: Thanks for your valuable suggestions. We now removed the statement over using multiple-species dataset, and changed it as:

"
Although the endeavor, the potential of foundation models trained exclusively on 100 million human cells has yet to be fully explored.
"

Minor Comments:

PBMC 10k is missing from the Table S4, or is labeled with a different name, this is important because it can have a different number of genes depending on where it is downloaded from.
Response: Thanks for your valuable suggestions. We have updated the detailed information for the PBMC 10K in the Table S4, which was obtained from scGPT (https://github.com/bowang-lab/scGPT/tree/main/data).

Reviewer #5 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts
Response: We sincerely thank you for your time and effort in co-reviewing our manuscript. We appreciate Nature Communications' initiative to support early career researchers in peer review, and

we appreciate your contribution to the review process.

REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Author)

My questions regarding perturbations from the CellFM model have been answered.
Response: Thanks for your valuable comments. We are pleased to hear that your concerns have been addressed. We appreciate your insightful comments, which have contributed to improving the manuscript.

(Remarks on code availability)

I was able to review the extra code that I have asked and it seems in order.
Response: Thanks for your valuable feedback. We are glad to hear that your concerns have been addressed. We appreciate your time and effort in reviewing the additional code.

Reviewer #2 (Remarks to the Author)

The authors have addressed our comments very well this round. We really appreciate the authors took efforts to revised the manuscript given our comments.
Response: Thanks for your positive feedback. We sincerely appreciate your thoughtful review and constructive comments, which have helped us improve the manuscript.

Reviewer #3 (Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.
Response: Thanks for your valuable comments. We appreciate the efforts of both reviewers in providing thoughtful and constructive feedback. Your insights have been invaluable in improving our manuscript.

Reviewer #4 (Remarks to the Author)

We acknowledge that the authors have addressed most of our previous comments. We have a few minor remaining points:

1. Benchmark Transparency: Please clarify, for each dataset, whether models were

fine-tuned or used in a zero-shot setting, and indicate this in both the figure and its caption.

Response: Thanks for your valuable comments. We have updated the figure captions to explicitly indicate whether models were fine-tuned or used in a zero-shot setting for each dataset. We kept the figures unchanged because adding settings in figures will introduce visual clutter or redundancy alongside the performance metrics already displayed. The revised captions now provide a concise yet complete description, following common practices in the field to balance transparency with effective visual presentation.

2. Score Reporting: Report all individual scIB metric scores, not just the averages (AvgBio), to support fair and reproducible evaluation.

Response: Thanks for your valuable comments. We have now reported all individual scIB metrics, including batch effect removal, biological information preservation, and the overall metric. Specifically, we present the Graph Connection (batch effect removal), ARI, NMI, ASW (biological information preservation), and AvgBio (overall metric) as representative indicators. As shown in Supplementary Figures S15,16, and 19, our CellFM model outperformed the baseline models across all these metrics. We have added these results into “Results” section as follows:

“

As shown in Supplementary Figure S15, CellFM achieves the highest average values across all scIB metrics.

”

“

As shown in Supplementary Figure S16, CellFM demonstrated consistent performance across all scIB metrics when varying the number of MLP layers for cell type annotation.

”

“

As shown in Supplementary Figure S19, CellFM achieves the highest average values across all scIB metrics.

”

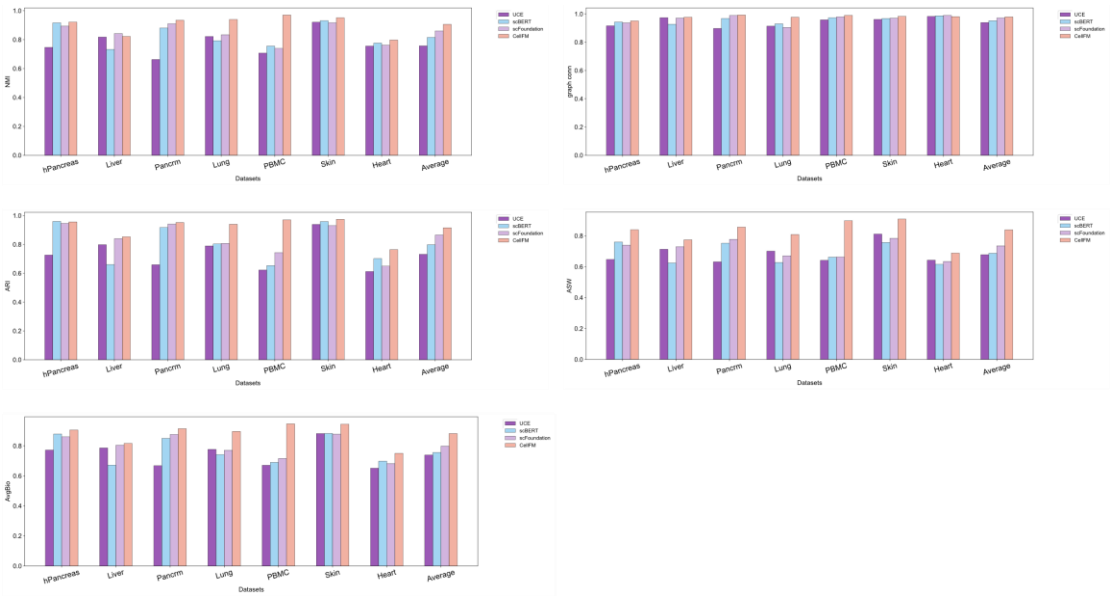


Fig. S15. A comparative analysis of scIB metric scores was performed, which were calculated based on the embeddings generated by CellFM and the top three competing single-cell foundation models for classification tasks across inter-datasets in a zero-shot setting.

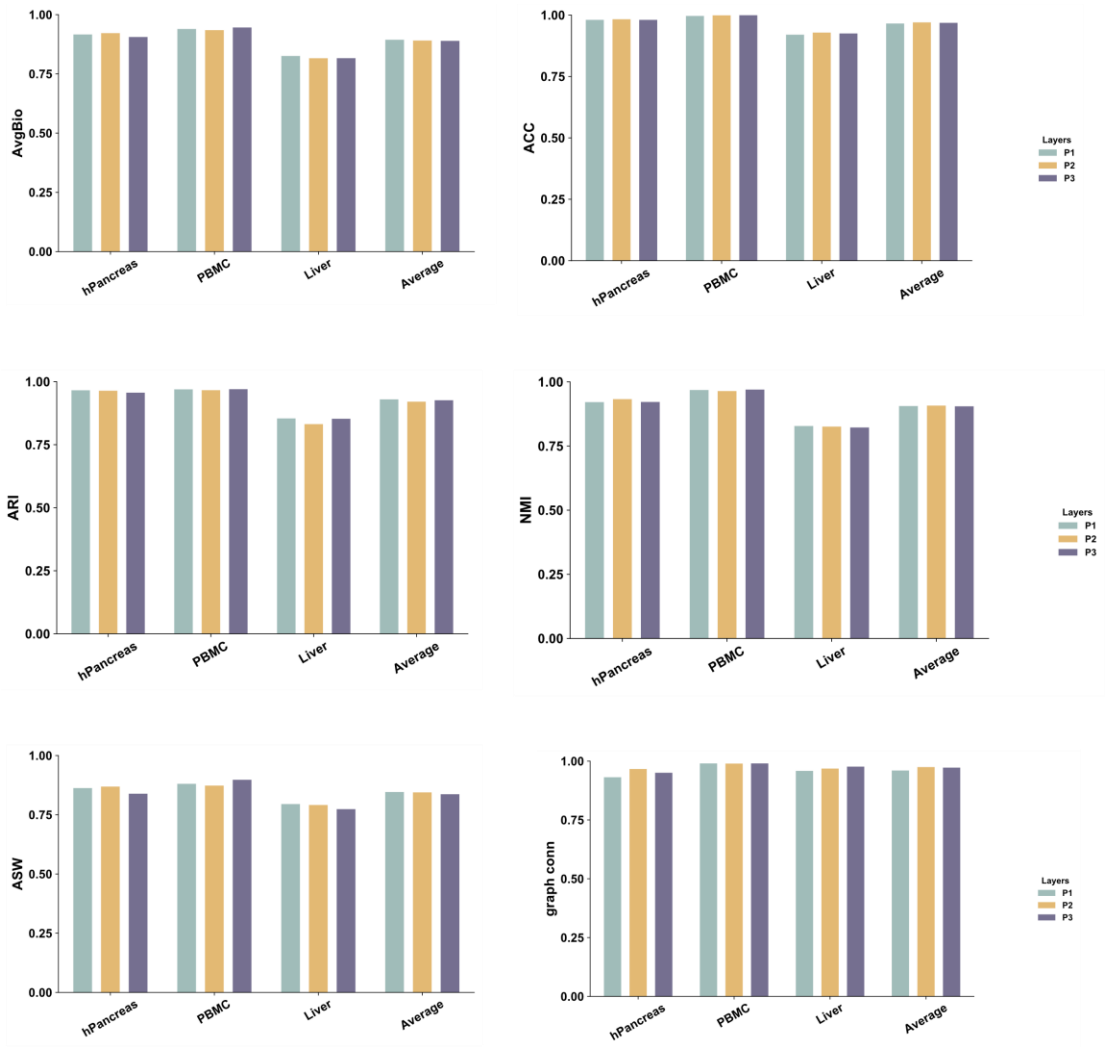


Fig. S16. Performance evaluation of CellFM with varying MLP layers for cell type annotation on three inter-datasets (BPMC, Liver, and Pancrm) in a zero-shot setting. ACC and scIB metric scores are reported for configurations with 1, 2, and 3 MLP layers, denoted as P1, P2, and P3, respectively.

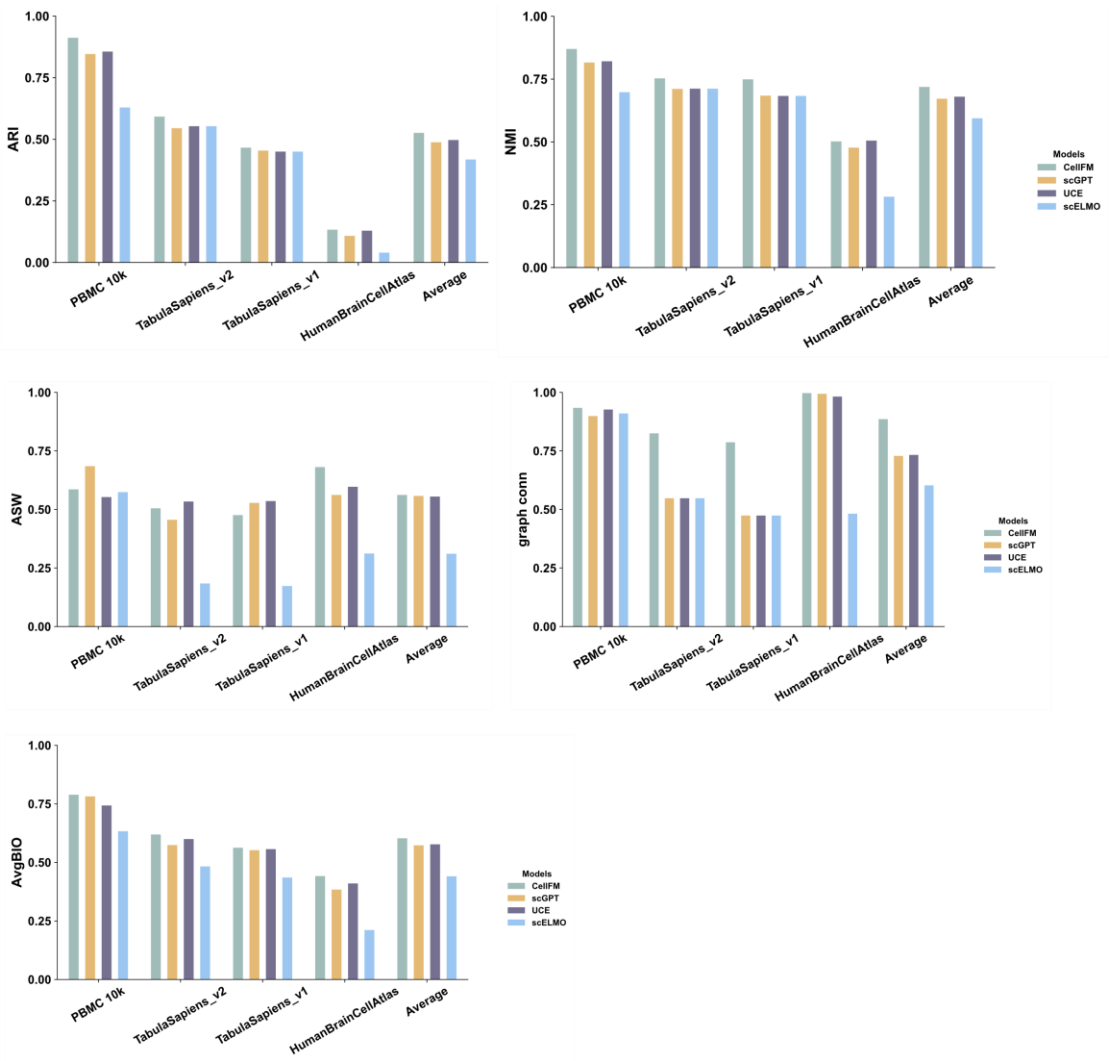


Fig. S19. Batch effect benchmark comparison of CellFM, scGPT, scELMo, and UCE on the PBMC 10k dataset, human brain cell atlas, and Tabula Sapiens dataset in a fine-tuning setting. Performance was evaluated using the scIB metric scores.

3. Model Consistency: Use a consistent set of models (e.g., scBERT, scGPT) across all datasets and tasks to ensure comparability.

Response: Thanks for your valuable comments. As shown in Supplementary Figure S3, for each task, we have compared all foundation models that could be performed in the task. We have no way to compare all models since the original authors didn't test the task and we have no corresponding codes and optimal hyperparameters. Our in-house implementations of these foundation models may be unfair to the models. This approach ensures that our comparisons are grounded in the intended use cases of each model, maintaining relevance and consistency in our evaluation.

Reviewer #5 (Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Response: Thanks for your valuable comments. We appreciate the efforts of both reviewers in providing thoughtful and constructive feedback. Your insights have been invaluable in improving our manuscript.