PLOS Computational Biology

RESEARCH ARTICLE

Assessing scale and predictive diversity in models for single-cell transcriptomics based on Geneformer

Junfan Chen[1], **Fabian Schmidt**[1*], **Ricardo Henao**[2*]

1 Biomedical Sciences Division (BioMed), King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia, **2** Department of Biostatistics & Bioinformatics, Duke University, Durham, North Carolina, United States of America

* fabian.schmidt@kaust.edu.sa (FS); ricardo.henao@duke.edu (RH)

OPEN ACCESS

Citation: Chen J, Schmidt F, Henao R (2026) Assessing scale and predictive diversity in models for single-cell transcriptomics based on Geneformer. PLoS Comput Biol 22(7): e1013701. https://doi.org/10.1371/journal.pcbi.1013701

Editor: Michael Domaratzki, University of Western Ontario: Western University, CANADA

Received: November 2, 2025

Accepted: July 2, 2026

Published: July 30, 2026

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1013701

Abstract

Single-cell transcriptomic data provide critical insights into cellular states and disease mechanisms, and foundation models have recently emerged as powerful tools for learning gene–gene relationships from these data. However, current approaches often overlook key challenges, including the mismatch between model design and the rank-ordered structure of gene expression profiles, as well as the unclear benefits of large-scale pretraining for biological applications. Here, we present GF_{CAB}, a modified modeling framework designed to better capture the structural properties of ranked single-cell transcriptomic data. GF_{CAB} incorporates a cumulative assignment mechanism to suppress repeated gene predictions and a similarity-based regularization strategy to promote diversity in model outputs. Across multiple evaluation settings, including pretraining behavior, biologically relevant classification tasks, and cross-dataset analyzes, GF_{CAB} consistently reduces redundancy and enhances the recovery of low-frequency genes with known functional and disease relevance while maintaining or improving predictive accuracy. In downstream applications, including classification and zero-shot batch effect correction, the model achieves competitive or improved performance compared to existing approaches. We further show that indiscriminately increasing the pretraining data scale does not uniformly improve performance. Instead, models trained on substantially smaller datasets can match or exceed the performance of larger models and often demonstrate improved generalization across datasets. Together, these findings highlight the importance of aligning model design with the intrinsic structure of biological data and suggest that architectural innovation can reduce reliance on large-scale training data. GF_{CAB} provides a framework for developing more efficient and biologically informative models for single-cell analysis, with potential applications in disease characterization and precision biology.

Data availability statement: All datasets used in this study are publicly available in the Hugging Face repository at https://huggingface.co/datasets/JFLa/GF-CAB_Datasets. This includes the pretraining datasets at two scales, the evaluation datasets, the downstream classification datasets, and the zero-shot batch-effect correction evaluation datasets. The code scripts used for model training, evaluation, and analysis are available in both the GitHub repository at https://github.com/CHENJ-VV/GF-CAB.git and the Hugging Face repository at https://huggingface.co/JFLa/GF-CAB. The repository includes scripts for data processing, model training, and benchmarking results generation.

Funding: This work was supported by the Smart Health Initiative (KSHI6088 to FS, RH) and baseline-funding from the King Abdullah University of Science and Technology (KAUST). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

Single-cell analysis helps researchers understand how genes work together inside individual cells, and recent artificial intelligence models have shown strong potential for uncovering these patterns. However, many existing approaches do not fully account for how this data is structured, and often assume that using more training data will always improve performance. In this study, we introduce GF_{CAB}, a model designed to better match the way single-cell data are organized. By reducing repeated predictions and encouraging the model to consider a wider range of genes, GF_{CAB} produces more diverse and biologically meaningful results, including the identification of less common but important gene signals. We also show that simply increasing the amount of training data does not always lead to better outcomes. Instead, well-designed models trained on smaller datasets can perform just as well and often generalize better to new data. These findings highlight the importance of model design in developing efficient and reliable tools for single-cell research.

Introduction

Characterizing genetic relationships within single-cell transcriptomic data [1,2], and leveraging these representations for a broad spectrum of downstream biological tasks, has emerged as a central paradigm across modern computational biology and related disciplines [3–5]. To this end, a wide range of computational approaches has been developed to systematically learn latent genetic structures from such data and generalize these representations to predictive tasks, including perturbation response modeling, drug sensitivity prediction, disease classification, and cellular dynamics inference [6–9]. Early efforts adapted classical frameworks, such as multilayer perceptrons (MLPs) and differential expression testing (DET) [10], while subsequent work introduced specialized models tailored to single-cell data, including scVI, scANVI, and GEARS [11–13], collectively demonstrating the feasibility of extracting biological representations from sparse transcriptomic profiles. More recently, single-cell transcriptomic foundation models have been proposed as a unified representational framework, leveraging self-attention mechanisms [14] to capture complex gene–gene dependencies and enabling flexible adaptation to a wide range of downstream tasks through pretraining and fine-tuning. Emerging models such as scGPT, scFoundation, and SCmilarity [4,15,16] highlight the potential of this paradigm. In parallel, the Bidirectional Encoder Representations from Transformers (BERT) architecture [17] has been adapted to single-cell transcriptomics. An early example, scBERT [18], applied a BERT-style framework to single-cell data and demonstrated its utility for cell-type annotation, while the subsequent Geneformer (GF), a ranking-based BERT variant pretrained on approximately 30 million cells [19], extended this paradigm toward foundation-model applications across a broader range of downstream tasks.

Despite these advances, foundation models such as GF remain susceptible to fundamental mismatches between backbone design choices and the intrinsic ranking properties of single-cell transcriptomic data, an issue that has yet to be systematically examined. In particular, GF operates on rank-ordered gene expression profiles, yet directly inherits the masked language modeling (MLM) objective from natural language processing [20], where token repetition is permissible. This misalignment introduces undesirable behaviors: GF can repeatedly predict the same gene within a single cellular profile, despite the biological constraint that each gene is uniquely represented per cell. Consequently, the model exhibits a bias toward highly frequent and ubiquitously expressed genes, leading to reduced prediction diversity and diminished biological interpretability. In addition, recent studies in foundation models have highlighted that indiscriminate scaling of pretraining data does not necessarily yield improved generalization performance [21–23]. Within the context of single-cell transcriptomic modeling, this issue remains largely unexplored. Current approaches predominantly emphasize increasing dataset size, often aggregating ever-larger collections of cells, yet lack a principled assessment of whether such scaling is necessary or effective for representation learning in this domain.

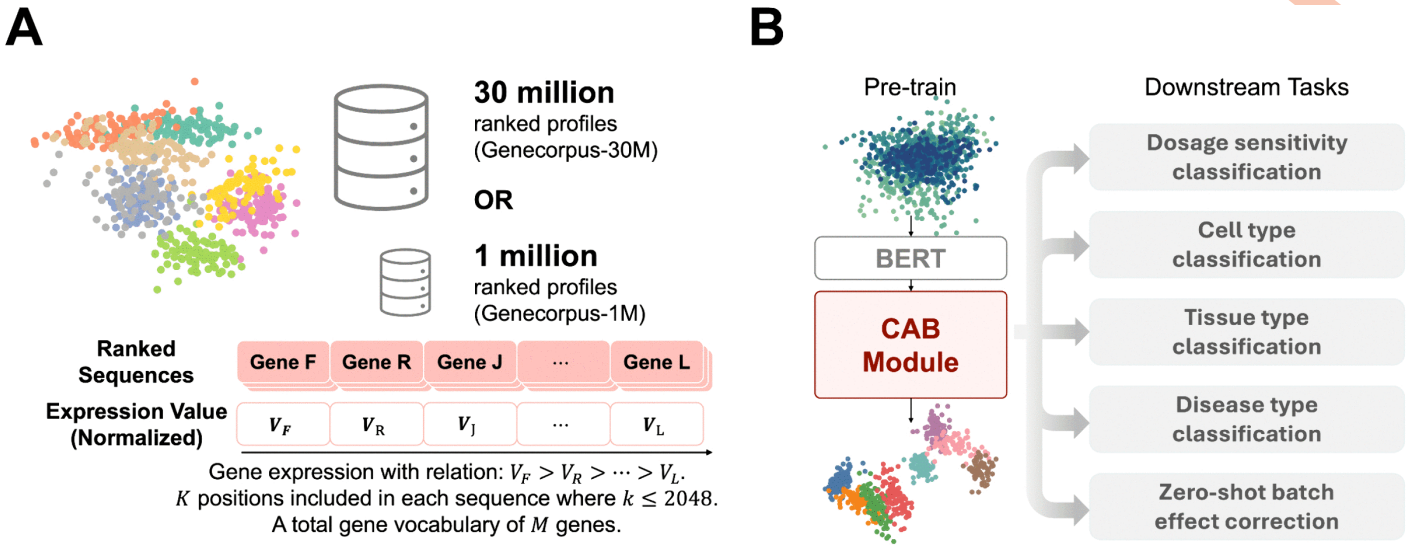
To address these underexplored limitations, we introduce GF_{CAB}, a modified GF architecture designed to better align model behavior with the rank-ordered structure of single-cell transcriptomic data. Central to this design is a cumulative assignment and balancing (CAB) module, implemented as a post-prediction processor. The CAB module incorporates a probability cumulative-assignment mechanism to propagate positional constraints across predictions, ensuring consistency with the non-redundant nature of gene rankings within each cell. In parallel, a similarity-based regularization term penalizes redundant or overly similar probability distributions across gene positions, thereby promoting diversity in predicted gene identities. Together, these components explicitly encode structural priors of the data into the prediction process, mitigating the mismatch introduced by conventional masked language modeling objectives. To further investigate the role of data scale in conjunction with architectural design, we additionally construct a reduced pretraining corpus (Genecorpus-1M) by uniformly subsampling one million profiles from the original Genecorpus-30M, enabling controlled comparisons across different data regimes (Fig 1A).

We show that integrating the CAB module effectively alleviates the aforementioned limitations, as evidenced by systematic improvements in masked gene prediction behavior, including reduced repetition and enhanced diversity. To evaluate downstream utility, we benchmark GF_{CAB} across a range of biologically relevant tasks, including gene dosage sensitivity prediction, cell type classification, tissue type classification, disease classification, and zero-shot batch effect correction (Fig 1B). Although ranking-based models remain inherently constrained in zero-shot generalization settings, GF_{CAB} consistently matches or outperforms the original GF across tasks. Furthermore, we examine the role of pretraining data scale by comparing models trained under varying dataset sizes. Our results indicate that increased data scale does not uniformly translate to performance gains. In contrast, targeted architectural modifications, such as the CAB module, can substantially narrow and, in some cases, eliminate performance gaps associated with data scale, achieving competitive or superior performance even under up to 30-fold reductions in pretraining data while maintaining stronger generalization capacity. Collectively, these findings establish GF_{CAB} as a principled and data-aligned extension of GF, demonstrating that incorporating structural constraints into model design can serve as an effective alternative to indiscriminate data scaling in single-cell transcriptomic foundation models.

Results

GF_{CAB} is a model architecture aligned with rank-ordered single-cell transcriptomic profiles

GF_{CAB} is designed to address two key limitations of the original Geneformer (GF) architecture: repetitive gene predictions across masked positions within a single ranked cell profile and a systematic bias toward highly frequent, ubiquitously expressed genes at the expense of rarer but biologically informative signals. These issues arise from the mismatch between the masked language modeling objective and the non-redundant, rank-ordered structure of single-cell transcriptomic data. To mitigate this mismatch, GF_{CAB} operates on the post-BERT probability distributions and introduces a



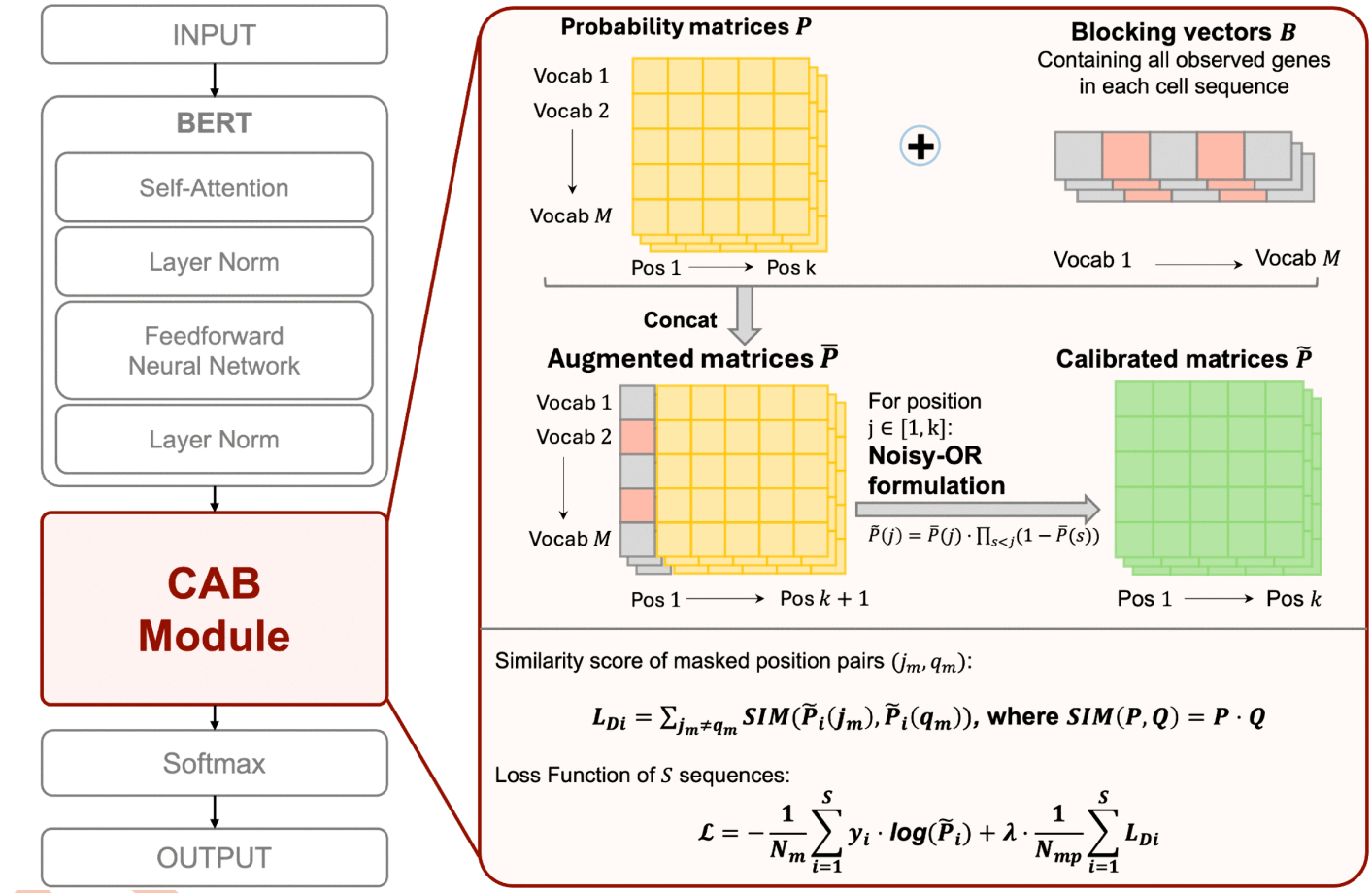


Fig 1. Modeling framework and evaluation pipeline overview. (A) Pretraining is performed on two data scales, 1 million (1M) and 30 million (30M) single-cell profiles, where each cell is represented as a rank-ordered gene sequence based on expression levels, with a vocabulary exceeding 20,000

genes. **(B)** Pretrained models are evaluated across multiple downstream tasks, including gene dosage sensitivity classification, cell type classification, tissue type classification, disease type classification, and zero-shot batch integration. **(C)** Illustration of the GF_{CAB} architecture, highlighting the cumulative-assignment suppression and similarity-based regularization components introduced during pretraining to reduce repetition and enhance predictive diversity. All icons and graphical elements in this figure were created by the authors or generated using Python scripts.

https://doi.org/10.1371/journal.pcbi.1013701.g001

cumulative assignment and balancing (CAB) module to enforce cross-position constraints and promote distributional diversity (Fig 1C; Methods).

Formally, for a cell profile with multiple masked positions, the BERT backbone produces a probability matrix over the gene vocabulary for each position, where each column represents a predicted distribution across all candidate genes. To reduce repetition across masked positions, we introduce a cumulative-assignment mechanism that explicitly propagates information across positions. Specifically, a blocking vector encoding observed (unmasked) genes is incorporated to suppress candidates that are already present within the same cell profile. The resulting probability distributions are then sequentially adjusted so that genes assigned high confidence at earlier positions are progressively down-weighted in subsequent predictions. This cumulative suppression, conceptually analogous to a Noisy-OR formulation [24], ensures that high-probability assignments are not repeatedly selected across positions, thereby enforcing non-redundant predictions within each cell.

In parallel, to enhance predictive diversity, we incorporate a similarity-based regularization term that penalizes overly similar probability distributions across masked positions. Concretely, the similarity between positional predictions is quantified using pairwise dot-product similarity, and the aggregated similarity across positions is minimized during training. This constraint discourages the model from concentrating probability mass on a limited subset of highly frequent genes. Instead, it encourages a broader exploration of candidate genes across positions. The final training objective combines the standard masked language modeling loss with this diversity regularization, balancing predictive accuracy and distributional diversity.

Together, these two components reshape the prediction landscape by explicitly encoding structural priors of rank-ordered transcriptomic data, enabling GF_{CAB} to generate more diverse and biologically informative gene predictions while maintaining consistency with the non-redundant nature of gene expression profiles.

GF_{CAB} improves masked gene prediction fidelity and reveals diminishing returns from data scaling

We first evaluated whether GF_{CAB} improves alignment with rank-ordered single-cell transcriptomic profiles. To systematically characterize predictive behavior, we quantified masked gene prediction accuracy, repetition ratio, and uniqueness (Methods), capturing model fidelity, redundancy across masked positions, and the diversity of predicted genes, respectively. We benchmarked both GF and GF_{CAB}, pretrained on 1M and 30M profiles, on an independent test set of 50,000 ranked cells (Fig 2A–2C, S1 Table). GF_{CAB} consistently achieved higher prediction accuracy, reduced repetition, and increased uniqueness across both data scales. These results indicate that improved architectural alignment with the rank-ordered structure directly translates into enhanced predictive performance and more biologically consistent outputs.

Notably, models trained on smaller datasets exhibited systematically higher diversity, characterized by reduced repetition and increased coverage of distinct genes (Fig 2B, 2C). This observation suggests that increasing pretraining data alone does not resolve the tendency of rank-based models to overproduce frequent genes and may instead reinforce such biases. One plausible explanation is that large-scale pretraining encourages the model to learn highly consistent and dominant gene–gene relationships, leading to sharper and more concentrated probability distributions over a limited set of high-frequency genes. In contrast, models trained on smaller datasets are exposed to fewer and less constrained co-expression patterns, resulting in smoother predictive distributions with higher entropy. This, in turn, promotes greater exploratory diversity during masked prediction, allowing lower-frequency genes to be more readily recovered. Such behavior is consistent with our empirical observations of increased uniqueness and reduced repetition in smaller-scale

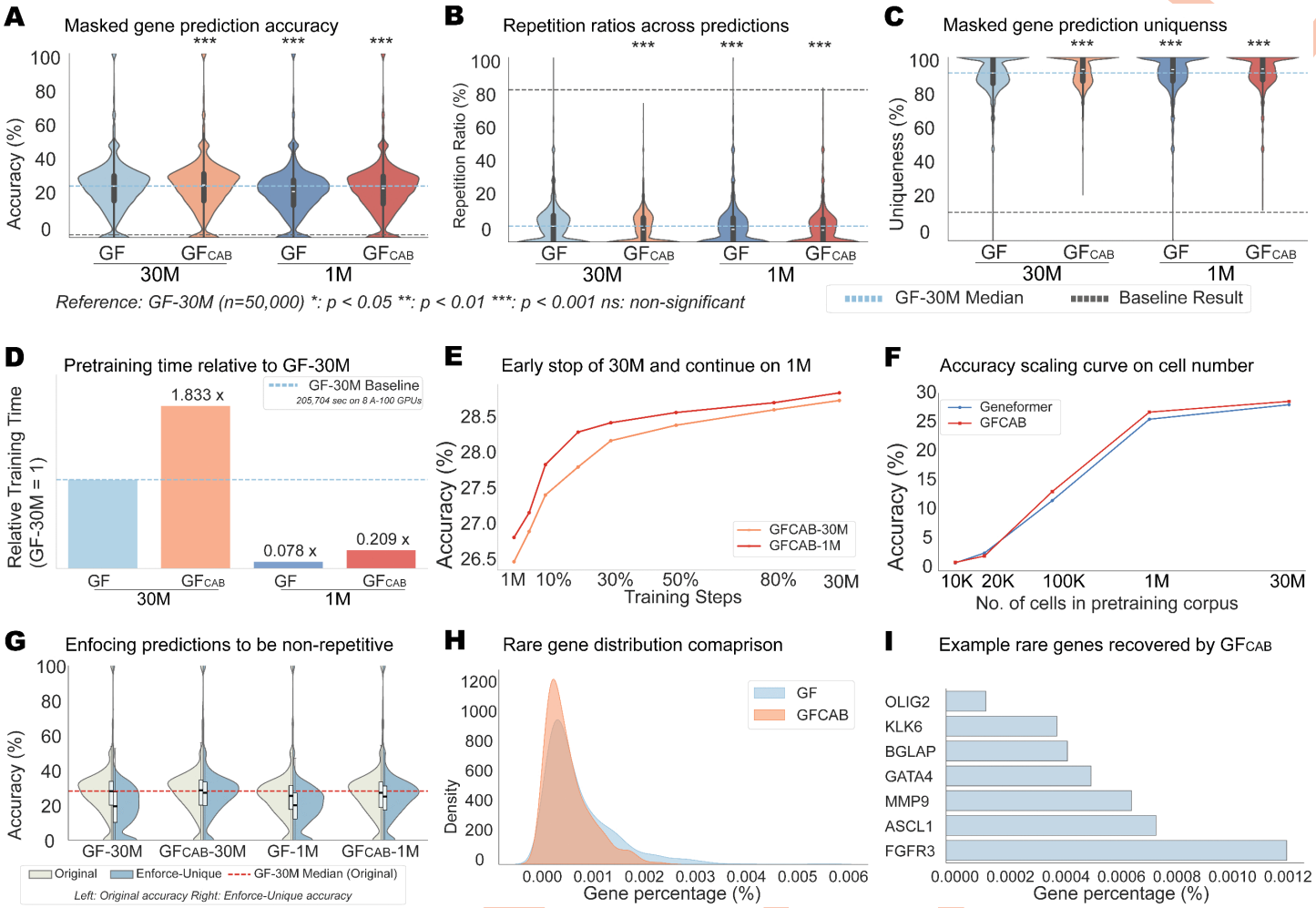


Fig 2. GF_{CAB} improves masked gene prediction behavior across accuracy, repetition, and diversity. (A) Distributions of masked token prediction accuracy for GF and GF_{CAB} at 30M and 1M pretraining scales. Wilcoxon tests were performed relative to GF-30M; white lines indicate medians and baseline accuracy. GF_{CAB} achieves consistently higher accuracy. **(B)** Distributions of repetition ratios across models and data scales, showing reduced redundancy in GF_{CAB}. **(C)** Distributions of uniqueness scores, demonstrating increased predictive diversity in GF_{CAB}. **(D)** Relative training time of GF and GF_{CAB}, normalized to GF trained on 30M cells (GF-30M). **(E)** Effect of training dynamics: comparison between continued training on 1M data and early stopping on 30M data. Performance gains are concentrated in early training stages, with diminishing returns as training progresses. **(F)** Scaling behavior of masked token prediction accuracy with increasing pretraining data size, showing diminishing marginal gains at larger scales above 1M cells. **(G)** Comparison between standard predictions and predictions under an enforced-uniqueness constraint. While enforcing uniqueness eliminates repetition, it leads to decreased accuracy, highlighting the trade-off addressed by GF_{CAB}. **(H)** Frequency distribution of low-prevalence genes predicted by GF and GF_{CAB}, showing that GF_{CAB} preferentially captures rarer genes. **(I)** Representative biologically relevant genes uniquely recovered by GF_{CAB}, including lineage-defining transcription factors (ASCL1, OLIG2, KLK6), tissue-specific markers (BGLAP, GATA4), and disease-associated regulators (FGFR3, MMP9).

https://doi.org/10.1371/journal.pcbi.1013701.g002

models. To contextualize these findings, we additionally report relative training times (Fig 2D), highlighting the trade-off between computational cost and marginal performance gains.

We next investigated the relationship between data scale and predictive performance. Tracking masked gene prediction accuracy across training steps revealed that performance gains are largely concentrated in the early phase of training, with most improvements achieved within the first ∼30% of total steps, followed by clear diminishing returns (Fig 2E, S1 Fig). Across both architectures, models trained on 1M profiles consistently matched or outperformed their

30M counterparts, indicating that increased data scale does not necessarily yield improved predictive capacity under fixed model capacity. Extending this analysis across a broader range of dataset sizes (10K–30M cells) further confirmed a saturating scaling behavior, with performance gains plateauing beyond the ∼1M regime (Fig 2F, S2 Fig). To further assess architectural scaling, we fixed the pretraining dataset at 30M profiles and systematically increased model capacity (S3 Fig). Performance improvements were primarily observed in lower-capacity, underparameterized regimes, suggesting that gains diminish once the model approaches saturation relative to the data scale. These results highlight the importance of coordinated scaling between model capacity and data size, rather than independent expansion of either factor. Notably, GF_{CAB} consistently outperformed baseline models across all data and model scales, underscoring the importance of architectural alignment over brute-force scaling. However, we acknowledge the risk that our subsampling strategy may benefit our smaller datasets in their diversity representation compared to curating datasets simply from a more limited number of sources.

To disentangle diversity improvements from potential metric artifacts, we compared naive predictions with those obtained under enforced uniqueness, where repeated predictions are explicitly prevented. Under this constraint, GF exhibited a substantial drop in predictive accuracy, whereas GF_{CAB} retained performance with only minor degradation (Fig 2G, S4 Fig). This result indicates that GF_{CAB} does not achieve improved diversity through post hoc redistribution alone, but rather through intrinsically better-structured predictive distributions.

Finally, we assessed the biological relevance of enhanced diversity by examining predictions across gene frequency strata. GF_{CAB} showed increased recovery of low-prevalence genes compared to GF (Fig 2H, 2I, S5 Fig), including lineage-associated transcription factors (ASCL1, OLIG2, KLK6) [25–27], tissue-specific markers (BGLAP, GATA4) [28,29], and disease-relevant signaling and remodeling mediators (FGFR3, MMP9) [30,31]. These findings suggest that improved diversity translates into greater sensitivity to biologically meaningful but underrepresented signals.

Collectively, these results demonstrate that GF_{CAB} enhances masked gene prediction fidelity by reducing redundancy and increasing diversity, while revealing that performance gains from either data scaling or model scaling are inherently limited. Instead, principled architectural alignment with data structure provides a more effective route to improving model behavior in single-cell transcriptomic foundation models.

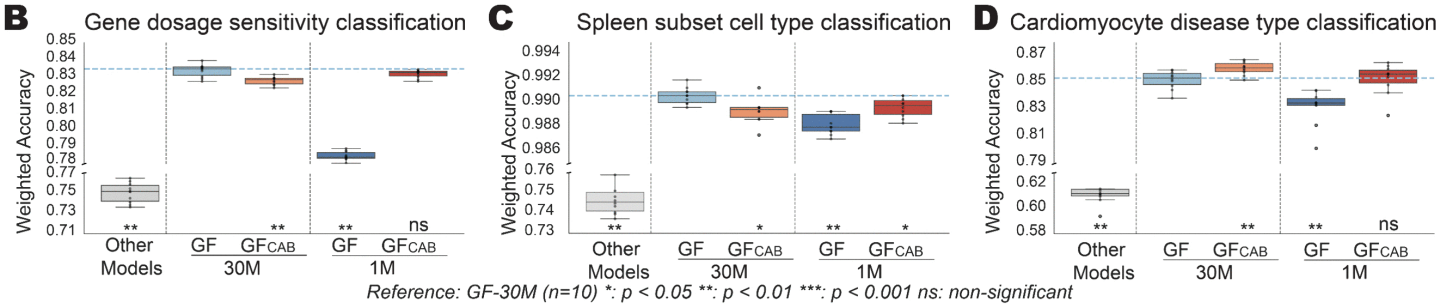
GF_{CAB} improves downstream task performance and preserves generalization under reduced data scaling

To evaluate whether improvements in pretraining behavior translate into downstream utility, we assessed GF_{CAB} across a diverse set of biologically relevant classification tasks and zero-shot batch effect correction benchmarks. Specifically, we considered three classification tasks, including gene dosage sensitivity prediction, cardiomyocyte disease-state classification, and tissue-specific cell type classification, and four zero-shot batch integration datasets, including Immune 330K [32], Pancreas 16K [33], PBMC 12K [12,34], and scCello out-of-distribution (OOD) cell type 487K [35] (Fig 3A). We benchmarked GF_{CAB} against the original Geneformer (GF) across both 1M and 30M pretraining regimes, alongside representative classical baselines including logistic regression [36], random forest [37], and support vector machine [38], as well as established reference methods including HVGs-based representations and scVI [12] for batch integration.

Across classification tasks, GF_{CAB} demonstrated competitive or improved predictive performance relative to GF (Fig 3B–3D, S2–S5 Table). For in-distribution tasks derived directly from the pretraining corpus, including gene dosage sensitivity and spleen cell type classification, GF_{CAB} showed slightly reduced performance compared to GF at the 30M scale but consistently outperformed GF under the 1M setting. This behavior can be attributed to the regularization introduced by the CAB module, which suppresses redundant predictions and reduces overly concentrated probability distributions across masked positions. While these mechanisms can improve predictive diversity and robustness, they may also reduce the extent to which the model exploits task-specific predictive signals in settings where the original GF model already performs strongly. Thus, the slight advantage of GF-30M in these in-distribution tasks may reflect differences in the trade-off between predictive diversity and task-specific optimization.

A

Tasks	Datasets	Models					Ranks (best:1)
		GF-30M	GF$_{CAB}$-30M	GF-1M	GF$_{CAB}$-1M	Other Model	
Classification tasks		**Accuracy achieved by each model**					
Gene dosage sensitivity	gene dosage 50K	0.834	0.828	0.785	0.832	0.750	
Cardiomyocyte disease	cardiomyocyte 580K	0.852	0.861	0.832	0.853	0.688	
Tissue-specific cell type	cell type 250K	0.907	0.904	0.896	0.901	0.493	
Zero-shot batch effect tasks		**Average Scores achieved by each model**					
Zero-shot batch effect integration	Immune 330K	0.450	0.471	0.507	0.566	0.700	
	Pancreas 16K	0.306	0.358	0.360	0.394	0.763	
	PBMC 12K	0.639	0.661	0.628	0.712	0.675	
	scCello cell type 487K	0.553	0.552	0.610	0.663	0.719	
		Average Rankings across Tasks					
***Average ranks (Classification)**		1.667	2.000	4.000	2.334	5.000	
***Average ranks (Batch-effect)**		4.250	4.250	3.500	1.750	1.250	
***Average ranks (Overall)**		3.143	3.286	3.714	2.000	2.857	



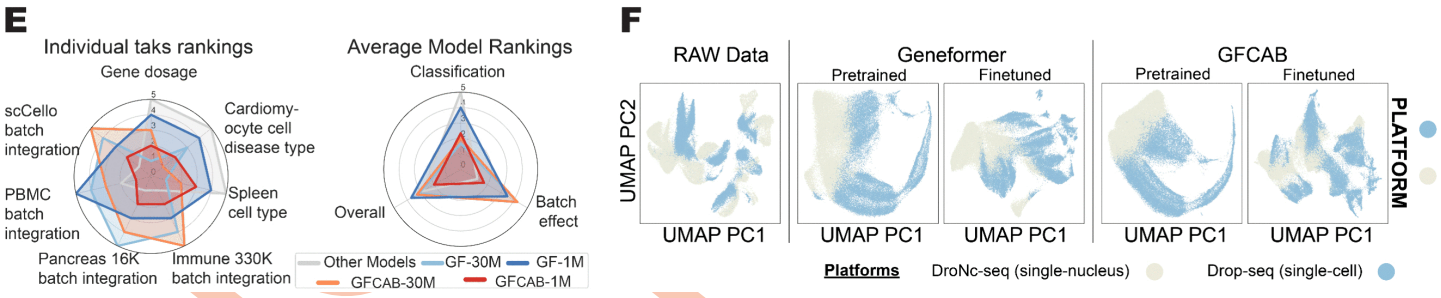


Fig 3. Overview of model performance across downstream tasks, model architectures, and pretraining data scales. (A) Summary of model performance across three classification tasks (gene dosage sensitivity, cardiomyocyte disease-state classification, and spleen cell type classification) and four zero-shot batch integration benchmarks, with numerical metrics displayed alongside color-coded ranking indicators. **(B–C)** Detailed classification results for each task, showing comparisons between foundation models and the best-performing classical baselines (SVM, Random Forest, Logistic Regression). **(E)** Per-task ranking and average ranking across all tasks and datasets, with a radar plot illustrating comparative performance across models and data scales. **(F)** Visualization of embedding quality on cardiomyocyte datasets, including both Drop-seq (single-cell) and DroNc-seq

(single-nucleus) modalities. Models are fine-tuned on single-cell data and evaluated in a zero-shot manner across modalities, demonstrating improved batch integration performance after fine-tuning.

https://doi.org/10.1371/journal.pcbi.1013701.g003

In contrast, for the out-of-distribution cardiomyocyte disease classification task, where we predict disease states (non-failing, hypertrophic, and dilated) from independently processed datasets, GF_{CAB} outperformed GF across both pretraining scales (Fig 3D). This reversal highlights the effect of regularization in mitigating overfitting to pretraining-specific patterns and improving generalization under distribution shift. These observations are consistent with a classical bias–variance trade-off, whereby increased regularization reduces overfitting at the expense of peak in-distribution performance while enhancing robustness to unseen data. Notably, GF_{CAB} pretrained on 1M data consistently matched or exceeded the performance of GF pretrained on 30M data, indicating that architectural improvements can offset, and in some cases surpass, gains from large-scale pretraining.

We next evaluated zero-shot batch effect correction as a proxy for model generalization, following established evaluation protocols [39]. GF_{CAB} consistently improved batch mixing performance relative to GF across datasets and scales (Fig 3A, S6 Table). Although ranking-based foundation models remain less competitive than dedicated integration methods [35,39,40], this improvement indicates that enhanced predictive structure translates into better cross-dataset generalization. Notably, models pretrained on 1M data exhibited stronger generalization than their 30M counterparts, suggesting that excessive data scaling may overfit dominant dataset-specific patterns and reduce transferability.

To provide a unified assessment, we aggregated performance across all tasks using a ranking-based summary (Fig 3A, 3E). GF_{CAB} pretrained on 1M data achieved the highest overall ranking when jointly considering classification accuracy and batch integration performance. This trend remained consistent when prioritizing biologically relevant classification tasks while grouping batch correction benchmarks as a separate evaluation axis, further highlighting the robustness of the model under reduced data scaling (S7 Table).

Finally, to assess generalization under cross-platform distribution shifts, we evaluated models on cardiomyocyte datasets spanning Drop-seq (single-cell) and DroNc-seq (single-nucleus) modalities [41]. Models were fine-tuned on single-cell data and evaluated in a zero-shot manner across modalities (Fig 3F). GF_{CAB} exhibited improved integration performance following fine-tuning, consistent with prior observations [19], further supporting its ability to generalize across technical and biological shifts.

Collectively, these results demonstrate that GF_{CAB} translates improved pretraining behavior into enhanced downstream performance, achieving competitive classification accuracy while consistently improving generalization in zero-shot settings. Importantly, these gains are maintained, and in some cases amplified, under reduced data scaling, highlighting that principled architectural alignment can preserve and even enhance the generalization capacity of single-cell foundation models without reliance on large-scale pretraining.

Component-wise contributions of GF_{CAB} and benchmarking against foundation models

To systematically dissect the contributions of the cumulative-assignment recalibration and similarity-based regularization in the CAB module, we conducted an ablation study comparing three model variants: the original Geneformer (GF), GF with cumulative-assignment recalibration only (GF+Recalib.), and the full GF_{CAB} model. These variants were evaluated across pretraining behavior, downstream classification, and zero-shot batch integration tasks under both 1M and 30M data scales (Fig 4A, S8 Table).

For masked gene prediction, the combination of cumulative-assignment and similarity-based regularization consistently improved both accuracy and predictive diversity relative to GF. In contrast, cumulative-assignment alone yielded limited gains and, in some cases, slightly underperformed GF, suggesting that suppressing redundancy without additional distributional regularization is insufficient to fully improve predictive behavior. The full GF_{CAB} model, however, achieved stable

A

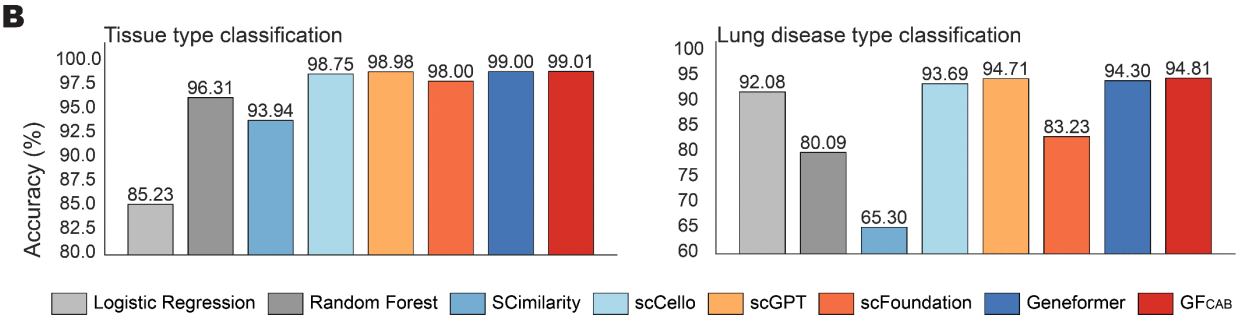
Models	Pretrain			Gene-dosage		Tissuecell type		Cardiomyocyte cell type	
	Acc	Repetition	Unique	Acc	Macro F1	Acc	Macro F1	Acc	Macro F1
30M models									
GF	27.251	9.786	0.902	0.903	0.834	0.946	0.907	0.852	0.830
GF+Recalib	27.828	10.156	0.898	0.899	0.828	0.955	0.904	0.852	0.831
GFCAB	26.354	4.510	0.955	0.851	0.796	0.954	0.902	0.851	0.834
1M models									
GF	24.672	9.263	0.908	0.849	0.785	0.952	0.896	0.832	0.810
GF+Recalib.	25.913	9.494	0.905	0.856	0.792	0.952	0.899	0.847	0.825
GFCAB	26.017	7.605	0.924	0.860	0.796	0.951	0.896	0.853	0.831

Ranks (best:1)

Models	Zero-shot batch-effect											
	Immune 330K			Pancreas 16K			PBMC 12K			scCello OOD Cell Type 487K		
	AvgBio	AvgBatch	AvgScore	AvgBio	AvgBatch	AvgScore	AvgBio	AvgBatch	AvgScore	AvgBio	AvgBatch	AvgScore
30M models												
GF	0.301	0.598	0.450	0.155	0.457	0.306	0.568	0.710	0.639	0.448	0.658	0.553
GF+Recalib.	0.242	0.564	0.403	0.221	0.495	0.358	0.600	0.686	0.643	0.332	0.615	0.474
GFCAB	0.326	0.615	0.471	0.160	0.488	0.324	0.628	0.694	0.661	0.448	0.656	0.552
1M models												
GF	0.384	0.630	0.507	0.215	0.505	0.360	0.570	0.686	0.628	0.547	0.672	0.610
GF+Recalib.	0.416	0.683	0.550	0.221	0.500	0.360	0.691	0.697	0.694	0.568	0.679	0.623
GFCAB	0.444	0.689	0.566	0.255	0.533	0.394	0.708	0.704	0.706	0.635	0.692	0.663

Ranks (best:1)

B



C

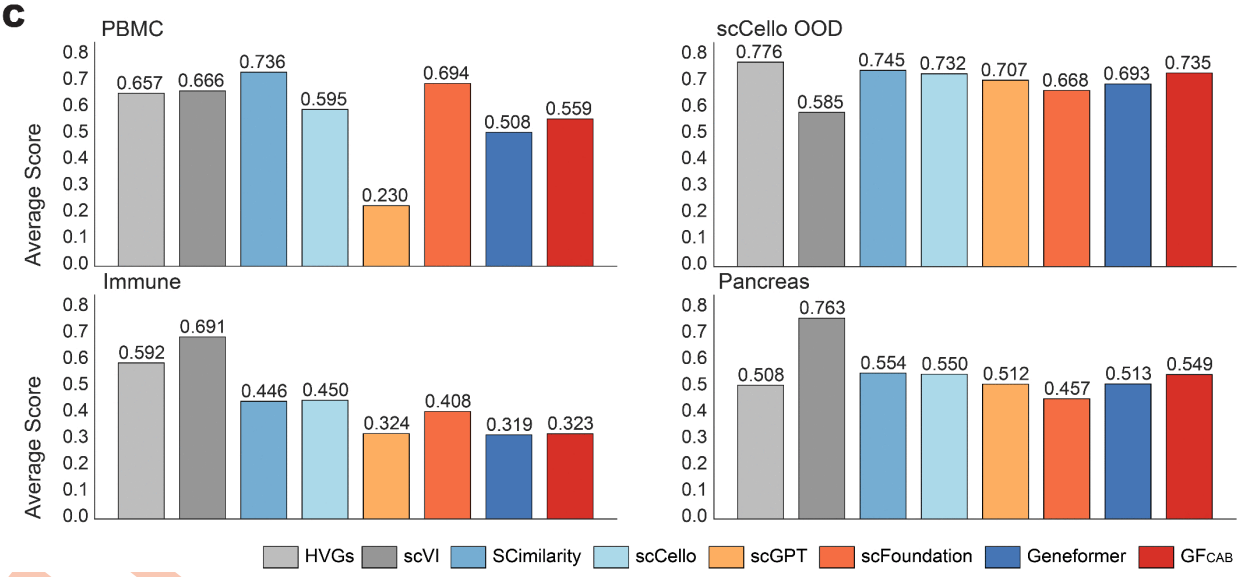


Fig 4. Ablation and benchmarking of GF_{CAB} across downstream classification and zero-shot batch integration tasks. (A) Ablation analysis of GF_{CAB}, evaluating the contributions of its components across three classification tasks (gene dosage sensitivity, cardiomyocyte disease-state

classification, and spleen cell type classification) and four zero-shot batch integration benchmarks. Numerical performance metrics are shown alongside color-coded rankings. **(B)** Classification accuracy benchmarking of GF_{CAB} against classical baselines (logistic regression, random forest) and foundation model counterparts (scCello, scGPT, scFoundation, SCimilarity, and GF) under a unified data processing and evaluation framework, across tissue type classification and lung disease classification tasks. **(C)** Zero-shot batch integration performance, reported as average scores ($0.6 \times \text{AvgBIO} + 0.4 \times \text{AvgBatch}$), comparing GF_{CAB} with representative methods (highly variable genes (HVGs), scVI) and foundation models (scCello, scGPT, scFoundation, SCimilarity, and GF) across four datasets under the same evaluation pipeline.

https://doi.org/10.1371/journal.pcbi.1013701.g004

improvements across scales, highlighting the complementary roles of the two components in shaping well-calibrated and diverse prediction distributions.

In downstream classification tasks, both GF+Recalib. and GF_{CAB} generally matched or exceeded the performance of GF, although gains were task-dependent. The addition of similarity-based regularization did not consistently increase classification accuracy beyond cumulative reassignment alone but preserved comparable performance while enhancing prediction diversity (S8 Table). This indicates that the similarity constraint acts primarily as a regularizing factor, reducing redundancy without compromising discriminative capacity. Notably, given the already strong baseline performance of GF, these improvements are incremental but consistent.

The most pronounced improvements were observed in zero-shot batch integration. Across all four datasets, both cumulative reassignment and similarity-based regularization progressively improved generalization compared to GF, with GF_{CAB} achieving the strongest overall performance (Fig 4A). These results indicate that enforcing cross-position constraints and distributional diversity can partially mitigate the known limitations of rank-based models in handling cross-dataset variability.

To contextualize these gains, we further benchmarked GF_{CAB} against other foundation models, including scCello, scGPT, scFoundation, SCimilarity, and GF, under a unified data processing and evaluation framework (Fig 4, S6 Fig) [4,15,35,42]. Across classification tasks (tissue type and lung disease classification), GF_{CAB} achieved the highest accuracy, further extending the strong discriminative performance of ranking-based model GF (Fig 4B). These results reinforce that architectural alignment within the rank-based modeling paradigm can yield state-of-the-art performance for classification tasks.

We next extended this comparison to zero-shot batch integration (Fig 4C). Consistent with prior studies [35,39,40,43], ranking-based foundation models, including GF and GF_{CAB}, remain less competitive than dedicated integration methods such as HVG-based representations and scVI. Among foundation models, scFoundation and SCimilarity, both characterized by relatively simpler architectures, exhibited stronger generalization, while scCello outperformed GF by incorporating cell ontology information to guide representation learning. In this context, GF_{CAB} consistently improved over GF within the naive ranking-based family, but did not fully close the performance gap with non-ranking-based approaches. This suggests that while architectural refinement can alleviate some limitations, intrinsic constraints of rank-based representations remain a key factor in zero-shot generalization.

Collectively, these results demonstrate that the CAB module provides complementary improvements in predictive behavior and downstream performance. While these gains further strengthen the already high discriminative capacity of ranking-based foundation models, they also highlight the persistent gap in zero-shot batch integration, underscoring the need for future advances beyond architectural refinement alone.

Discussion

Model architectures that are explicitly aligned with the structural properties of single-cell transcriptomic data can yield consistent and meaningful performance gains. In this work, GF_{CAB} addresses a fundamental limitation of rank-based modeling by mitigating repetitive predictions in masked gene recovery, while improving overall predictive accuracy and sensitivity to biologically relevant signals. By integrating cumulative-assignment recalibration and similarity-based regularization, GF_{CAB} produces more

diverse and well-calibrated predictions, enabling the recovery of low-prevalence genes with functional significance. These findings demonstrate that architectural alignment with the rank-ordered nature of gene expression profiles can offset, and in some cases surpass, the benefits of large-scale pretraining, highlighting model design as a critical determinant of performance.

A second key contribution of this work is the systematic evaluation of data scaling in single-cell foundation models. Contrary to the prevailing assumption that larger pretraining datasets uniformly improve performance, we show that gains in both predictive behavior and downstream tasks exhibit clear diminishing returns, with the majority of improvements occurring early in training. Notably, under an improved architecture such as GF_{CAB}, models pretrained on substantially smaller datasets can match or exceed the performance of models trained on datasets up to 30-fold larger. Furthermore, smaller-scale models consistently exhibit stronger generalization in zero-shot settings, suggesting that excessive scaling may reinforce dataset-specific biases at the expense of transferability.

Collectively, this study addresses two underexplored challenges in the development of single-cell foundation models: the mismatch between model architectures and rank-ordered transcriptomic representations, and the limited understanding of data scaling effects in this domain. Our results motivate a shift in design principles, emphasizing that scaling alone is insufficient without architectures that respect the intrinsic structure of biological data. While rank-based representations provide a powerful framework for capturing gene expression patterns, they also impose inherent constraints, particularly in settings requiring robust cross-dataset generalization. Future progress will likely depend on jointly advancing model architectures and representation strategies to better reconcile predictive accuracy, diversity, and generalization in single-cell learning systems.

Methods

Datasets

For full-scale pre-training, we followed the protocol established in Geneformer (GF) and used the Genecorpus-30M dataset [19], which contains 29,900,531 single-cell transcriptomic profiles, where the genes in each cell were ranked by normalized expression values. To examine the impact of reduced data scale, we constructed Genecorpus-1M by randomly sampling one million profiles from the original corpus. In addition, we curated a non-overlapping dataset of five million profiles (Genecorpus-5M), which was reserved exclusively for the unbiased evaluation of the masked pretraining objective.

Downstream benchmarking relied on datasets that were previously used in the GF study to ensure comparability across tasks. The gene dosage-sensitivity task was defined using 122 dosage-sensitive and 368 dosage-insensitive genes reported in previous studies [44–46], along with 50,000 ranked profiles curated by Theodoris et al. [19] (Genecorpus gene dosage 50K). Multi-tissue cell type classification was performed on 249,556 profiles spanning 59 cell types across nine tissues (spleen, kidney, lung, liver, brain, placenta, large intestine, immune system, and pancreas) [19] (Genecorpus cell type 250K). For cardiomyocyte disease-state classification, we used the profiles from Chaffin et al. [47], which included cardiomyocytes from non-failing hearts, as well as from patients with hypertrophic cardiomyopathy (HCM) [48] and dilated cardiomyopathy (DCM) [49]. This dataset encompassed 580,000 ranked profiles from 42 individuals, annotated with metadata including age, sex, and cell type, producing a three-class classification task (Chaffin cardiomyocyte 580K). To evaluate the finetuned batch effect integration of pretrained models on biologically relevant dataset, we conducted research on the dual platform sequencing dataset on cardiomyocyte [41], with both the Drop-seq (single-cell) sequencing data and the DorNc-seq (single-nucleus) data.

To evaluate the quality of learned cell embeddings and assess batch integration under zero-shot conditions, we incorporated four external datasets following prior work [35,39]: Pancreas 16K [33], PBMC 12K [12,34], Immune 330K [32], and scCello OOD cell type 487K [35].

The expanded benchmark includes two classification tasks and zero-shot batch integration across four datasets. The classification tasks consist of (i) a multi-tissue classification task using eight tissues (bone marrow, eye, heart, kidney,

large intestine, small intestine, liver, and muscle) from the Tabula Sapiens dataset [50], and (ii) a lung disease classification task (COPD, IPF, control) using dataset GSE136831 [51]. For batch integration, we evaluated models on Pancreas 16K [33], PBMC 12K [12,34], Immune 330K [32], and the scCello OOD dataset (487K cells) [35] with re-processing from scratch to synchronize evaluation frameworks.

Enhancing prediction diversity with $\mathbf{GF}_{\mathbf{CAB}}$

Geneformer pioneered the application of language models (LMs) for single-cell RNA-seq by representing gene co-expression patterns as rank-ordered sequences [19]. Although effective, its masked language modeling (MLM) framework exhibits three limitations: (1) repeated assignment of the same gene across multiple masked positions within a profile, (2) systematic under-representation of low-expression genes, and (3) frequency-driven bias toward highly expressed, ubiquitous genes. To address these issues, we propose $\mathbf{GF}_{\mathbf{CAB}}$ (Geneformer with **C**umulative-assignment on the prediction **A**djustment and **B**oosting on the similarity-based diversity), which augments GF with two complementary mechanisms: cumulative-assignment prediction adjustment and similarity-based regularization.

Cumulative-assignment prediction adjustment. In standard MLM, masked positions are predicted independently, allowing the same gene to be assigned repeatedly across multiple sites. To mitigate this redundancy, we introduce a cumulative assignment-and-blocking mechanism that integrates one-hot substitution with cumulative Noisy-OR suppression [24].

Given a masked sequence $\tilde{C}_i = [g_1^i,\ \texttt{<MASK>},\ g_3^i,\ \ldots,\ \texttt{<MASK>},\ \ldots,\ g_k^i]$ of k tokens, the BERT model produces a $k \times V$ probability matrix P_i, where $V \in \mathbb{R}$ is the vocabulary size. Observed positions are replaced with one-hot vectors: $P_i'(j, g) = 1$ if $g = g_j^i$, and 0 otherwise. A blocking vector $B_i \in \{0, 1\}^V$ is then defined such that $B_i(g) = 1$ if $g \in G_{\text{observed}}$ (observed gene set within a profile), and it is concatenated with P_i' to form the augmented matrix $\bar{P}_i$, as shown in Eq (1).

$$\bar{P}_i = \begin{bmatrix} B_i \\ P_i' \end{bmatrix} \in \mathbb{R}^{(k+1)\times V}. \tag{1}$$

After complement $\bar{P}_i' = 1 - \bar{P}_i$, Noisy-OR normalization produces adjusted probabilities $\tilde{P}_i(j, g)$, as shown in Eq (2), which enforces cumulative suppression across positions. This mechanism mitigates duplicate predictions within the same cell profile, thereby improving biological plausibility.

$$\tilde{P}_i(j, g) = \bar{P}_i(j, g) \cdot \prod_{s<j} \bar{P}_i'(s, g). \tag{2}$$

Similarity-based regularization. Despite the redundancy observed within individual cell profiles, the predictions made by the GF remain strongly biased toward highly expressed, ubiquitous genes. This pattern suggests an overreliance on global expression trends, with limited utilization of local, context-specific signals that are essential for meaningful biological interpretation.

To address this frequency bias limitation, we introduce a similarity-based regularization strategy, inspired by Ayinde *et al.* [52], which is designed to penalize overly similar predictions across masked positions and encourage context-aware inference. This approach directly counteracts the model's tendency to favor frequent and highly expressed genes, thus improving its ability to capture rare, condition-specific signals and consequently improving predictive diversity.

For masked positions (j_m, q_m) in sequence $\tilde{C}_i$, we compute pairwise similarity between adjusted distributions, as shown in Eq (3).

$$\text{sim}\left(\tilde{P}_i(j_m), \tilde{P}_i(q_m)\right) = \tilde{P}_i(j_m) \cdot \tilde{P}_i(q_m). \tag{3}$$

Summing across all N_{mp} distinct masked-pair combinations produces a diversity loss, Eq (4), which discourages uniform predictions across masked sites by penalizing distributional redundancy.

$$L_{D_i} = \sum_{j_m \neq q_m} \mathrm{sim}(\tilde{P}_i(j_m), \tilde{P}_i(q_m)) \,. \tag{4}$$

Finally, we incorporate this similarity penalty into the masked sequence pretraining objective Eq (5) by combining it with the standard cross-entropy loss. Here, $S \in \mathbb{R}$, $N_m \in \mathbb{R}$, and $y_i \in \mathbb{R}^V$ denote the number of sequences, the number of masked positions, and the one-hot representation of the ground-truth label for sequence i, respectively. The hyperparameter $\lambda \in \mathbb{R}$ controls the strength of diversity regularization. This penalty discourages over-concentration on frequent genes, promoting rare and context-sensitive predictions that are critical for biological discovery.

$$\mathcal{L} = -\frac{1}{N_m} \sum_{i=1}^{S} y_i \cdot \log \tilde{P}_i + \lambda \cdot \frac{1}{N_{mp}} \sum_{i=1}^{S} L_{D_i} \,. \tag{5}$$

Baseline models

We compared $\mathrm{GF}_{\mathrm{CAB}}$ against GF in three dimensions: pre-training objectives, downstream classification, and zero-shot batch-effect mitigation. In classification tasks, we also included Support Vector Machine (SVM) [38], Random Forest (RF) [37], and Logistic Regression (LR) as baselines. For zero-shot settings, GF and $\mathrm{GF}_{\mathrm{CAB}}$ were compared with highly variable genes (HVG) and scVI [53], a variational autoencoder framework that explicitly models batch effects.

Evaluation metrics

For pretraining evaluation, models were evaluated on the masked gene prediction objective using metrics that jointly capture predictive accuracy and diversity. Masked prediction accuracy quantifies the proportion of correctly reconstructed masked tokens across the validation set, reflecting the model's ability to recover contextually appropriate genes. Repetition ratios measure redundancy in predictions and are reported in three forms: the overall repetition ratio (fraction of all duplicate predictions), the unmasked repetition ratio (proportion of predictions overlapping with unmasked input tokens), and the masked repetition ratio (fraction of repeated predictions restricted to masked positions). To further characterize predictive diversity, we introduced a uniqueness score, inspired by entropy-based measures such as the Kullback–Leibler divergence [54]. This score reflects the proportion of distinct gene predictions at masked sites, capturing the model's ability to avoid mode collapse and explore less frequent, yet biologically informative genes. Higher uniqueness values indicate more diverse and context-sensitive predictions.

For downstream tasks, we used classification, clustering, and batch integration metrics to jointly assess biological relevance and technical robustness. Classification performance was evaluated by accuracy, macro-averaged AUC [55,56], macro and weighted F1 scores [57,58], and recall [59], complemented by confusion matrices for error-structure visualization [60]. Cluster quality, following established benchmarking frameworks [33,39], was quantified using normalized mutual information (NMI) [61], adjusted Rand index (ARI) [62], and average silhouette width ($\mathrm{ASW}_{\mathrm{label}}$), which were aggregated into the AvgBIO metric [4] to summarize cluster separability. Batch integration was assessed through the average silhouette width with respect to batch ($\mathrm{ASW}_{\mathrm{batch}}$) [63] and graph connectivity (GraphCon) [35], which were combined into an AvgBatch measure. To provide a balanced evaluation, an overall score (AvgScore) [4] was reported, defined as the mean of AvgBIO and AvgBatch [35]. For benchmarking against other foundation models, AvgScore was alternatively computed as $0.6 \times \mathrm{AvgBIO} + 0.4 \times \mathrm{AvgBatch}$ to align with the practice of Kedzierska et al. [39], thereby integrating biological fidelity with batch mixing quality.

Implementation details

We implemented the proposed architecture, GF_{CAB}, as an extension of the masked language model GF. The model retains GF's core transformer design (6 encoder layers, 4 attention heads, 256-dimensional embeddings, 512-dimensional feedforward layers, and a maximum input length of 2,048) while introducing cumulative prediction adjustment and a similarity-based loss penalty for diversity boosting. Hyperparameters, training schedules, and preprocessing strategies were aligned with GF to enable direct comparison (see S9 Table). Input preparation followed the standard masking scheme through a modified Hugging Face Trainer framework [64], with 15% of tokens selected for perturbation and custom batching to account for variable gene counts per cell. Pre-training was conducted at two scales: a 1M-cell dataset (Genecorpus-1M) with a batch size of 12 for rapid iteration and a 30M-cell dataset (Genecorpus-30M) with a batch size of 8 for full-scale training, both run on 8 NVIDIA A100 GPUs. GF was pretrained under identical conditions to ensure a controlled baseline.

For downstream evaluation, we fine-tuned GF, GF_{CAB}, and classical machine learning baselines (SVM, random forest, and logistic regression) on three classification tasks that reflect distinct biological challenges. Depending on task requirements, we used either Hugging Face's `BertForTokenClassification` for gene-level prediction or `BertForSequenceClassification` for cell profile–level tasks, initializing with pre-trained weights and applying GF's fine-tuning conventions. Hyperparameters such as frozen layers, learning rate schedules, warmup steps, and epochs were set according to the GF's released scripts, with minor task-specific adjustments (*e.g.*, different frozen layer numbers and learning rates for cardiomyocyte classification). Baseline classifiers were implemented using scikit-learn [65] with default parameters, except for training epochs, which were harmonized with task objectives.

To test zero-shot transferability, we compared the embeddings from GF_{CAB} and GF with those derived from highly variable genes (HVG) and scVI, following the protocol established by Kedzierska *et al.* [39]. For the transformer models, cell embeddings were computed from unmasked gene inputs without further masking or fine-tuning, ensuring strict zero-shot conditions. HVG features provided a conventional dimensionality reduction baseline, while scVI embeddings were drawn directly from the latent space. All embeddings were benchmarked using cluster- and integration-based metrics, allowing for a direct comparison between foundation models and established single-cell representation methods.

Statistical analysis

Pre-training results are reported as the mean $\pm$ standard deviation across 50,000 independent ranked single-cell profiles. For downstream tasks, accuracy and macro-F1 values were averaged over 10 independent cross-validation runs (where applicable). Statistical significance of performance differences was evaluated using the paired Wilcoxon signed-rank test, with each model compared pairwise against GF-30M as the baseline, unless otherwise noted.

Reproducibility

All experiments were performed using PyTorch (`v2.6.0`) with Hugging Face Transformers (`v4.38.2`), CUDA (`v11.8`), and Python (`v3.11.7`). Training was distributed across 8 NVIDIA A100 GPUs with mixed precision enabled. Random seeds were fixed across NumPy, PyTorch, and CUDA to ensure replicability. The complete codebase, pretrained weights, evaluation scripts, and datasets incorporated in this study are available at the following repositories.

Supporting information

S1 Table. Pretraining performance of models on 50,000 profiles from Genecorpus-5M. Pretraining results are reported for Geneformer and GF_{CAB} models with varying similarity regularization coefficients (λ), trained on different scales of pretraining data and evaluated on 50,000 test profiles randomly sampled from Genecorpus-5M. Metrics include masked prediction accuracy (mean $\pm$ standard deviation), repetition ratios (overall, unmasked, and masked), and uniqueness scores.
(XLSX)

S2 Table. Gene dosage-sensitivity prediction performance of Geneformer and GF_{CAB} models. Macro F1-scores and accuracies are reported for Geneformer and GF_{CAB} variants with different similarity regularization coefficients (λ), pretrained on either Genecorpus-30M or Genecorpus-1M. Bold values indicate the best performance, while underlined values indicate the second-best performance.
(XLSX)

S3 Table. Cardiomyocyte cell type prediction performance of Geneformer and GF_{CAB} models. Macro F1-scores and accuracies are reported for Geneformer and GF_{CAB} variants with different similarity regularization coefficients (λ), pretrained on either Genecorpus-30M or Genecorpus-1M. Bold values indicate the best performance, while underlined values indicate the second-best performance.
(XLSX)

S4 Table. Tissue-specific cell type prediction performance of Geneformer and GF_{CAB} models. Macro F1-scores and accuracies are reported for individual tissues using Geneformer and GF_{CAB} variants with different similarity regularization coefficients (λ), pretrained on either Genecorpus-30M or Genecorpus-1M. Bold values denote the best performance, while underlined values denote the second-best performance.
(XLSX)

S5 Table. Spleen-specific cell type prediction performance of Geneformer and GF_{CAB} models. Macro F1-scores and accuracies are reported for spleen-derived cells using Geneformer and GF_{CAB} variants with different similarity regularization coefficients (λ), pretrained on either Genecorpus-30M or Genecorpus-1M. Bold values denote the best performance, while underlined values denote the second-best performance.
(XLSX)

S6 Table. Zero-shot batch-effect mitigation performance of Geneformer and GF_{CAB} models. Batch-effect mitigation performance is reported for Geneformer and GF_{CAB} variants with different similarity regularization coefficients (λ), pretrained on either Genecorpus-30M or Genecorpus-1M. Evaluation metrics include cell type clustering measures (NMI, ARI, ASW_{label}, AvgBio), batch clustering measures (Graph connectivity score, ASW_{batch}, AvgBatch), and the overall score (average of AvgBio and AvgBatch).
(XLSX)

S7 Table. Ranking results across three classification tasks and zero-shot batch integration evaluated on four datasets. Overall rankings are computed using two strategies: (i) treating each batch integration dataset as an independent task; and (ii) aggregating all batch integration results into a single task before averaging with the three classification tasks. Across both strategies, the smaller-scale pretrained GF_{CAB} model achieves the strongest overall performance, supporting the conclusion that a reduced pretraining scale combined with the proposed architectural modifications can maintain competitive performance while improving generalization to unseen datasets.
(XLSX)

S8 Table. Ablation study results for Geneformer and GF_{CAB} models. Ablation results are reported for Geneformer and GF_{CAB} variants with different similarity regularization coefficients (λ), pretrained on either Genecorpus-30M or Genecorpus-1M. Evaluations span pretraining performance, gene dosage-sensitivity prediction, tissue-specific and cardiomyocyte cell type classification, and zero-shot batch-effect mitigation across four datasets using the metrics described above. Bold values denote the best performance, while underlined values denote the second-best performance.
(XLSX)

S9 Table. Pretraining hyperparameters for Geneformer and GF_{CAB} models. Detailed hyperparameter configurations used for Geneformer and GF_{CAB} models benchmarked in this study are reported.
(XLSX)

S1 Fig. Comparison between elongating training on 1M data and early stopping on 30M data to investigate the net effect brought by the remaining 29M data. (A) Pretraining masked token prediction accuracy, repetition ratios, and uniqueness across different data scales and different training steps. (B) Tissue type classification accuracy, macro F1 score, and weighted F1 score across data scales and different training steps.
(PDF)

S2 Fig. Scaling behavior of GF and GF_{CAB} across different pretraining data sizes (10K, 100K, 1M, and 30M cells). (A) Pretraining performance, including masked token prediction accuracy, repetition ratio, and uniqueness. (B) Downstream tissue-type classification performance, measured by accuracy, macro F1, and weighted F1 scores. Both models show performance improvements with increasing data scale, with diminishing returns observed beyond 1M cells.
(PDF)

S3 Fig. Effect of model scaling on performance for GF and GF_{CAB}. Model scale is measured by the number of trainable parameters (ranging from 4.35M to 26.97M). (A) Pretraining performance, including masked token prediction accuracy, repetition ratio, and uniqueness. (B) Tissue-type classification performance (accuracy, macro F1, weighted F1). (C) Lung disease classification performance (COPD, IPF, control; accuracy, macro F1, weighted F1). Moderate increases in model size improve performance, while excessive scaling leads to diminishing or negative returns, likely due to data–model mismatch. GF_{CAB} consistently outperforms GF across all model scales.
(PDF)

S4 Fig. Comparison between original model predictions and predictions under an enforced-uniqueness constraint across masked positions. (A) Cross-model and cross-scale comparison showing prediction accuracy for Geneformer and GF_{CAB} variants trained on different data scales. (B) Direct comparison of Geneformer and GF_{CAB} (1M and 30M pretraining scales), illustrating the impact of enforcing uniqueness on prediction accuracy. Enforcing uniqueness eliminates repetition (repetition ratio = 0) but leads to a noticeable accuracy drop, particularly for Geneformer, whereas GF_{CAB} demonstrates improved robustness under this constraint.
(PDF)

S5 Fig. Characterization of low-frequency genes uniquely recovered by GF_{CAB}. (A) Distribution of sequence-level occurrence, measured as the proportion of sequences in Genecorpus-30M containing each gene uniquely predicted by GF_{CAB}. (B) Distribution of overall gene frequency in Genecorpus-30M for these genes, demonstrating their low prevalence in the pretraining dataset. (C) Representative examples of biologically relevant genes recovered by GF_{CAB} but not by Geneformer, including lineage-defining transcription factors (ASCL1, OLIG2, KLK6), tissue-specific markers (BGLAP, GATA4), and disease-relevant signaling and remodeling mediators (FGFR3, MMP9). (D) Comparison of the frequency distribution of low-prevalence genes predicted by Geneformer and GF_{CAB}, showing that GF_{CAB} preferentially captures genes with lower dataset frequency, consistent with enhanced predictive diversity.
(PDF)

S6 Fig. Benchmark comparison of rank-based and non–rank-based foundation models under a unified evaluation framework. (A) Tissue classification performance (accuracy and macro F1) across eight tissues from Tabula Sapiens. (B) Lung disease classification performance (COPD, IPF, control) evaluated using dataset GSE136831. (C) Zero-shot batch integration performance measured by AvgScore ($0.6 \times$AvgBio $+ 0.4 \times$AvgBatch) across four datasets (Pancreas 16K, PBMC 12K, Immune 330K, and scCello OOD). Models include GF, GF_{CAB}, scGPT, scFoundation, SCimilarity, and scCello,

with classical baselines (logistic regression and random forest) for classification and HVG-based representation and scVI for batch integration. GF_{CAB} consistently improves over GF while remaining competitive within the rank-based modeling paradigm.
(PDF)

Acknowledgments

We would like to acknowledge Xin Gao and Jesper Tegnér for their valuable feedback and constructive discussions that helped strengthen this study. We would also like to thank the Department of Biostatistics and Bioinformatics at Duke University for their collaborative insights and the KAUST Supercomputing Core Laboratory for providing essential computational resources used in model pretraining and large-scale analysis.

Author contributions

Conceptualization: Ricardo Henao.

Data curation: Junfan Chen.

Formal analysis: Junfan Chen.

Funding acquisition: Fabian Schmidt.

Investigation: Junfan Chen.

Methodology: Junfan Chen, Ricardo Henao.

Project administration: Fabian Schmidt, Ricardo Henao.

Resources: Fabian Schmidt.

Software: Junfan Chen.

Supervision: Fabian Schmidt, Ricardo Henao.

Validation: Junfan Chen.

Visualization: Junfan Chen.

Writing – original draft: Junfan Chen.

Writing – review & editing: Fabian Schmidt, Ricardo Henao.

References

1. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods. 2009;6(5):377–82. Available from: https://doi.org/10.1038/nmeth.1315
2. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, et al. Tumour evolution inferred by single-cell sequencing. Nature. 2011;472(7341):90–4. https://doi.org/10.1038/nature09807 PMID: 21399628
3. Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, et al. On the Opportunities and Risks of Foundation Models. arXiv. 2021. Available from: https://doi.org/10.48550/ARXIV.2108.07258
4. Cui H, Wang C, Maan H, Pang K, Luo F, Duan N, et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. Nat Methods. 2024;21(8):1470–80. https://doi.org/10.1038/s41592-024-02201-0 PMID: 38409223
5. Ramesh A, Pavlov M, Goh G, Gray S, Voss C, Radford A, et al. Zero-Shot Text-to-Image Generation. arXiv. 2021. Available from: https://doi.org/10.48550/ARXIV.2102.12092
6. Wu Y, Wershof E, Schmon SM, Nassar M, Osiński B, Eksi R, et al. PerturBench: Benchmarking Machine Learning Models for Cellular Perturbation Analysis. arXiv. 2024. Available from: https://doi.org/10.48550/ARXIV.2408.10609
7. Wang Y, Liu X, Fan Y, Xie B, Cheng J, Wong KC, et al. Predicting drug responses of unseen cell types through transfer learning with foundation models. Nat Comput Sci. 2026;6(1):39–52. https://doi.org/10.1038/s43588-025-00887-6 PMID: 41044387

8. Theodoris CV, Li M, White MP, Liu L, He D, Pollard KS, et al. Human Disease Modeling Reveals Integrated Transcriptional and Epigenetic Mechanisms of NOTCH1 Haploinsufficiency. Cell. 2015;160(6):1072–86. Available from: https://doi.org/10.1016/j.cell.2015.02.035
9. Jin Y, He Y, Liu B, Zhang X, Song C, Wu Y, et al. Single-cell RNA sequencing reveals the dynamics and heterogeneity of lymph node immune cells during acute and chronic viral infections. Front Immunol. 2024;15:1341985. https://doi.org/10.3389/fimmu.2024.1341985 PMID: 38352870
10. Soneson C, Robinson MD. Bias, robustness and scalability in single-cell differential expression analysis. Nat Methods. 2018;15(4):255–61. https://doi.org/10.1038/nmeth.4612 PMID: 29481549
11. Xu C, Lopez R, Mehlman E, Regier J, Jordan MI, Yosef N. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. Mol Syst Biol. 2021;17(1):e9620. https://doi.org/10.15252/msb.20209620 PMID: 33491336
12. Gayoso A, Lopez R, Xing G, Boyeau P, Valiollah Pour Amiri V, Hong J, et al. A Python library for probabilistic analysis of single-cell omics data. Nat Biotechnol. 2022;40(2):163–6. https://doi.org/10.1038/s41587-021-01206-w PMID: 35132262
13. Roohani Y, Huang K, Leskovec J. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. Nat Biotechnol. 2024;42(6):927–35. https://doi.org/10.1038/s41587-023-01905-6 PMID: 37592036
14. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need. arXiv. 2017. Available from: https://doi.org/10.48550/ARXIV.1706.03762
15. Hao M, Gong J, Zeng X, Liu C, Guo Y, Cheng X, et al. Large-scale foundation model on single-cell transcriptomics. Nat Methods. 2024;21(8):1481–91. https://doi.org/10.1038/s41592-024-02305-7 PMID: 38844628
16. Heimberg G, Kuo T, DePianto D, Heigl T, Diamant N, Salem O, et al. Scalable querying of human cell atlases via a foundational model reveals commonalities across fibrosis-associated macrophages. bioRxiv. 2023. Available from: https://doi.org/10.1101/2023.07.18.549537
17. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language Models are Few-Shot Learners. arXiv. 2020. Available from: https://doi.org/10.48550/ARXIV.2005.14165
18. Yang F, Wang W, Wang F, Fang Y, Tang D, Huang J, et al. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. Nat Mach Intell. 2022;4(10):852–66. https://doi.org/10.1038/s42256-022-00534-z
19. Theodoris CV, Xiao L, Chopra A, Chaffin MD, Al Sayed ZR, Hill MC, et al. Transfer learning enables predictions in network biology. Nature. 2023;618(7965):616–24. https://doi.org/10.1038/s41586-023-06139-9 PMID: 37258680
20. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv. 2018. Available from: https://doi.org/10.48550/arXiv.1810.04805
21. Chen Z, Wang S, Xiao T, Wang Y, Chen S, Cai X, et al. Revisiting Scaling Laws for Language Models: The Role of Data Quality and Training Strategies. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vienna, Austria: Association for Computational Linguistics; 2025. p. 23881–99.
22. Springer JM, Goyal S, Wen K, Kumar T, Yue X, Malladi S, et al. Overtrained Language Models Are Harder to Fine-Tune. arXiv. 2025. Available from: https://doi.org/10.48550/ARXIV.2503.19206
23. Mizrahi D, Larsen ABL, Allardice J, Petryk S, Gorokhov Y, Li J, et al. Language Models Improve When Pretraining Data Matches Target Tasks. arXiv. 2025. Available from: https://doi.org/10.48550/ARXIV.2507.12466
24. Praveen P, Fröhlich H. Boosting probabilistic graphical model inference by incorporating prior knowledge from multiple sources. PLoS One. 2013;8(6):e67410. https://doi.org/10.1371/journal.pone.0067410 PMID: 23826291
25. Paez-Beltran LE, Chen H, Liyanapathirana M, Villicana E, Myers BL, Duan J, et al. ASCL1 and OLIG2 Expression Dynamics Control Glial Cell Fate and Regional Diversity in the Dorsal Forebrain. bioRxiv. 2026. Available from: https://doi.org/10.64898/2026.02.02.703158
26. Raabe FJ, Stephan M, Waldeck JB, Huber V, Demetriou D, Kannaiyan N, et al. Expression of Lineage Transcription Factors Identifies Differences in Transition States of Induced Human Oligodendrocyte Differentiation. Cells. 2022;11(2):241. https://doi.org/10.3390/cells11020241
27. Zhao K, Gao M, Lin M. KLK6 Functions as an Oncogene and Unfavorable Prognostic Factor in Bladder Urothelial Carcinoma. Dis Markers. 2022;2022:3373851. https://doi.org/10.1155/2022/3373851 PMID: 36193495
28. Nowicki JK, Jakubowska-Pietkiewicz E. Osteocalcin: Beyond Bones. Endocrinol Metabol. 2024;39(3):399–406. https://doi.org/10.3803/EnM.2023.1895
29. Yang Y, Song S, Li S, Kang J, Li Y, Zhao N, et al. GATA4 regulates the transcription of MMP9 to suppress the invasion and migration of breast cancer cells via HDAC1-mediated p65 deacetylation. Cell Death Dis. 2024;15(4):289. https://doi.org/10.1038/s41419-024-06656-z PMID: 38653973
30. Wang Y, Liu Z, Liu Z, Zhao H, Zhou X, Cui Y, et al. Advances in research on and diagnosis and treatment of achondroplasia in China. Intractable Rare Dis Res. 2013;2(2):45–50. https://doi.org/10.5582/irdr.2013.v2.2.45 PMID: 25343101
31. Wang J, Tsirka SE. Neuroprotection by inhibition of matrix metalloproteinases in a mouse model of intracerebral haemorrhage. Brain. 2005;128(Pt 7):1622–33. https://doi.org/10.1093/brain/awh489 PMID: 15800021
32. Domínguez Conde C, Xu C, Jarvis LB, Rainbow DB, Wells SB, Gomes T, et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans. Science. 2022;376(6594):eabl5197. https://doi.org/10.1126/science.abl5197 PMID: 35549406
33. Luecken MD, Büttner M, Chaichoompu K, Danese A, Interlandi M, Mueller MF, et al. Benchmarking atlas-level data integration in single-cell genomics. Nat Methods. 2022;19(1):41–50. https://doi.org/10.1038/s41592-021-01336-8 PMID: 34949812

34. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8:14049. https://doi.org/10.1038/ncomms14049 PMID: 28091601
35. Yuan X, Zhan Z, Zhang Z, Zhou M, Zhao J, Han B, et al. Cell-ontology guided transcriptome foundation model. arXiv. 2024. Available from: https://doi.org/10.48550/ARXIV.2408.12373
36. Tamir TS, Xiong G, Li Z, Tao H, Shen Z, Hu B, et al. Traffic Congestion Prediction using Decision Tree, Logistic Regression and Neural Networks. IFAC-PapersOnLine. 2020;53(5):512–7. https://doi.org/10.1016/j.ifacol.2021.04.138
37. Breiman L. Random Forests. Mach Learn. 2001;45(1):5–32. https://doi.org/10.1023/A:1010933404324
38. Boser BE, Guyon IM, Vapnik VN. A Training Algorithm for Optimal Margin Classifiers. In: Proceedings of the 5th Annual Workshop on Computational Learning Theory (COLT '92). Pittsburgh (PA): ACM Press; 1992. p. 144–52.
39. Kedzierska KZ, Crawford L, Amini AP, Lu AX. Zero-shot evaluation reveals limitations of single-cell foundation models. Genome Biol. 2025;26(1):101. https://doi.org/10.1186/s13059-025-03574-x PMID: 40251685
40. Kedzierska KZ, Crawford L, Amini AP, Lu AX. Assessing the limits of zero-shot foundation models in single-cell biology. bioRxiv. 2023. Available from: https://doi.org/10.1101/2023.10.16.561085
41. Selewa A, Dohn R, Eckart H, Lozano S, Xie B, Gauchat E, et al. Systematic Comparison of High-throughput Single-Cell and Single-Nucleus Transcriptomes during Cardiomyocyte Differentiation. Sci Rep. 2020;10(1):1535. https://doi.org/10.1038/s41598-020-58327-6 PMID: 32001747
42. Heimberg G, Kuo T, DePianto DJ, Salem O, Heigl T, Diamant N, et al. A cell atlas foundation model for scalable search of similar human cells. Nature. 2025;638(8052):1085–94. https://doi.org/10.1038/s41586-024-08411-y PMID: 39566551
43. Wang H, Leskovec J, Regev A. Limitations of cell embedding metrics assessed using drifting islands. Nat Biotechnol. 2026;44(4):574–7. https://doi.org/10.1038/s41587-025-02702-z PMID: 40500472
44. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536(7616):285–91. https://doi.org/10.1038/nature19057 PMID: 27535533
45. Shihab HA, Rogers MF, Campbell C, Gaunt TR. HIPred: an integrative approach to predicting haploinsufficient genes. Bioinformatics. 2017;33(12):1751–7. Available from: https://doi.org/10.1093/bioinformatics/btx028
46. Ni Z, Zhou X-Y, Aslam S, Niu D-K. Characterization of Human Dosage-Sensitive Transcription Factor Genes. Front Genet. 2019;10:1208. https://doi.org/10.3389/fgene.2019.01208 PMID: 31867040
47. Chaffin M, Papangeli I, Simonson B, Akkad A-D, Hill MC, Arduini A, et al. Single-nucleus profiling of human dilated and hypertrophic cardiomyopathy. Nature. 2022;608(7921):174–80. https://doi.org/10.1038/s41586-022-04817-8 PMID: 35732739
48. Liu X, Ma Y, Yin K, Li W, Chen W, Zhang Y, et al. Long non-coding and coding RNA profiling using strand-specific RNA-seq in human hypertrophic cardiomyopathy. Sci Data. 2019;6(1):90. https://doi.org/10.1038/s41597-019-0094-6 PMID: 31197155
49. Sweet ME, Cocciolo A, Slavov D, Jones KL, Sweet JR, Graw SL, et al. Transcriptome analysis of human heart failure reveals dysregulated cell adhesion in dilated cardiomyopathy and activated immune pathways in ischemic heart failure. BMC Genomics. 2018;19(1):812. https://doi.org/10.1186/s12864-018-5213-9 PMID: 30419824
50. Pisco A. Tabula Sapiens v2. figshare. 2024. Available from: https://doi.org/10.6084/M9.FIGSHARE.27921984
51. Adams TS, Schupp JC, Poli S, Ayaub EA, Neumark N, Ahangari F, et al. Single-cell RNA-seq reveals ectopic and aberrant lung-resident cell populations in idiopathic pulmonary fibrosis. Sci Adv. 2020;6(28):eaba1983. https://doi.org/10.1126/sciadv.aba1983 PMID: 32832599
52. Ayinde BO, Inanc T, Zurada JM. Regularizing Deep Neural Networks by Enhancing Diversity in Feature Extraction. IEEE Trans Neural Netw Learn Syst. 2019;30(9):2650–61. https://doi.org/10.1109/TNNLS.2018.2885972 PMID: 30624232
53. Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. Nat Methods. 2018;15(12):1053–8. https://doi.org/10.1038/s41592-018-0229-2 PMID: 30504886
54. Kullback S, Leibler RA. On Information and Sufficiency. Ann Math Stat. 1951;22(1):79–86. Available from: https://doi.org/10.1214/aoms/1177729694
55. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143(1):29–36. Available from: https://doi.org/10.1148/radiology.143.1.7063747
56. Junge MRJ, Dettori JR. ROC Solid: Receiver Operator Characteristic (ROC) Curves as a Foundation for Better Diagnostic Tests. Global Spine J. 2018;8(4):424–9. https://doi.org/10.1177/2192568218778294 PMID: 29977728
57. Opitz J, Burst S. Macro F1 and Macro F1. arXiv. 2019. Available from: https://doi.org/10.48550/ARXIV.1911.03347
58. Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med Imaging. 2015;15:29. https://doi.org/10.1186/s12880-015-0068-x PMID: 26263899
59. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015;10(3):e0118432. https://doi.org/10.1371/journal.pone.0118432 PMID: 25738806
60. Stehman SV. Selecting and interpreting measures of thematic classification accuracy. Remote Sens Environ. 1997;62(1):77–89. https://doi.org/10.1016/s0034-4257(97)00083-7

61. Vinh NX, Epps J, Bailey J. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In: Proceedings of the 26th Annual International Conference on Machine Learning. Montreal Quebec Canada: ACM; 2009. p. 1073–80.
62. Hubert L, Arabie P. Comparing partitions. J Classif. 1985;2(1):193–218. https://doi.org/10.1007/bf01908075
63. Tran HTN, Ang KS, Chevrier M, Zhang X, Lee NYS, Goh M, et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. Genome Biol. 2020;21(1):12. https://doi.org/10.1186/s13059-019-1850-9 PMID: 31948481
64. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. HuggingFace's Transformers: State-of-the-art Natural Language Processing. arXiv. 2019. Available from: https://doi.org/10.48550/ARXIV.1910.03771
65. Buitinck L, Louppe G, Blondel M, Pedregosa F, Mueller A, Grisel O, et al. API design for machine learning software: experiences from the scikit-learn project. arXiv. 2013. Available from: https://doi.org/10.48550/ARXIV.1309.0238