article	13dc81fce2c73de67dbe3829a32ec23d663cec89
title	scGPT: toward building a foundation model for single-cell multi-omics using generative AI.
year	2024
url	https://www.semanticscholar.org/paper/13dc81fce2c73de67dbe3829a32ec23d663cec89
abstract	S2 TL;DR: This study has constructed a foundation model for single-cell biology, scGPT, based on a generative pretrained transformer across a repository of over 33 million cells, and illustrates that scGPT effectively distills critical biological insights concerning genes and cells.
author	Haotian Cui and Chloe X. Wang and Hassaan Maan and Kuan Pang and Fengning Luo and Nan Duan and Bo Wang
journal	Nature methods
volume	null
pages	null
doi	10.1038/s41592-024-02201-0
pmid	38409223
article	**889a6e66f6b9f851ff156bf8f7a47af8b5dc06f9**
title	Generative pretraining from large-scale transcriptomes for single-cell deciphering
year	2023
url	https://www.semanticscholar.org/paper/889a6e66f6b9f851ff156bf8f7a47af8b5dc06f9
abstract	S2 TL;DR: The result is that tGPT represents a new analytical paradigm for integrating and deciphering massive amounts of transcriptome data and it will facilitate the interpretation and clinical translation of single-cell transcriptomes.
author	Hong-xiu Shen and Jilei Liu and Jiani Hu and X. Shen and Chao Zhang and Dan Wu and Mengyao Feng and Meng Yang and Yang Li and Yichen Yang and Wei Wang and Qiang Zhang and Ji-Lan Yang and Kexin Chen and Xiangchun Li
journal	iScience
volume	26
pages	null
doi	10.1016/j.isci.2023.106536
pmid	37187700
article	**d5baf1912d43d0af0e8f683eb5532f5f5445430e**
title	A benchmark of batch-effect correction methods for single-cell RNA sequencing data
year	2020
url	https://www.semanticscholar.org/paper/d5baf1912d43d0af0e8f683eb5532f5f5445430e

abstract	S2 TL;DR: An in-depth benchmark study on available batch correction methods to determine the most suitable method for batch-effect removal and batch integration, with Harmony, LIGER, and Seurat 3 recommended as viable alternatives.
author	Hoa Tran and Kok Siong Ang and Marion Chevrier and Xiaomeng Zhang and N. Lee and Michelle Goh and Jinmiao Chen
journal	Genome Biology
volume	21
pages	null
doi	10.1186/s13059-019-1850-9
pmid	31948481
article	**da9807e1e4dc913b34d86529fea034e7240656fc**
title	Assessing the limits of zero-shot foundation models in single-cell biology
year	2023
url	https://www.semanticscholar.org/paper/da9807e1e4dc913b34d86529fea034e7240656fc
abstract	The advent and success of foundation models such as GPT has sparked growing interest in their application to single-cell biology. Models like Geneformer and scGPT have emerged with the promise of serving as versatile tools for this specialized field. However, the efficacy of these models, particularly in zero-shot settings where models are not fine-tuned but used without any further training, remains an open question, especially as practical constraints require useful models to function in settings that preclude fine-tuning (e.g., discovery settings where labels are not fully known). This paper presents a rigorous evaluation of the zero-shot performance of these proposed single-cell foundation models. We assess their utility in tasks such as cell type clustering and batch effect correction, and evaluate the generality of their pretraining objectives. Our results indicate that both Geneformer and scGPT exhibit limited reliability in zero-shot settings and often underperform compared to simpler methods. These findings serve as a cautionary note for the deployment of proposed single-cell foundation models and highlight the need for more focused research to realize their potential.2
author	Kasia Z. Kedzierska and Lorin Crawford and Ava P. Amini and Alex X. Lu
journal	bioRxiv
volume	null
pages	null
doi	10.1101/2023.10.16.561085
article	**20b9e73d7de14254a290467eff46716bf0604341**
title	scPerturb: Information Resource for Harmonized Single-Cell Perturbation Data
year	2022
url	https://www.semanticscholar.org/paper/20b9e73d7de14254a290467eff46716bf0604341

abstract	Recent biotechnological advances led to growing numbers of single-cell studies, which reveal molecular and phenotypic responses to large numbers of perturbations. However, analysis across diverse datasets is typically hampered by differences in format, naming conventions, data filtering and normalization. In order to facilitate development and benchmarking of computational methods in systems biology, we collect a set of 44 publicly available single-cell perturbation-response datasets with molecular readouts, including transcriptomics, proteomics and epigenomics. We apply uniform pre-processing and quality control pipelines and harmonize feature annotations. The resulting information resource enables efficient development and testing of computational analysis methods, and facilitates direct comparison and integration across datasets. Using these datasets, we demonstrate the application of E-distance for quantifying perturbation similarity and strength. This work provides an information resource and guide for researchers working with single-cell perturbation data and highlights conceptual considerations for new experiments. The data is publicly available at scperturb.org.
author	Stefan Peidli and Tessa D. Green and Ciyue Shen and T. Gross and Joseph K Min and J. Taylor-King and D. Marks and Augustin Luna and N. Blüthgen and C. Sander
article	**587547493b5cf221af4b929cf390ef81e9768937**
title	Cell2Sentence: Teaching Large Language Models the Language of Biology
year	2024
url	https://www.semanticscholar.org/paper/587547493b5cf221af4b929cf390ef81e9768937
abstract	We introduce Cell2Sentence (C2S), a novel method to directly adapt large language models to a biological context, specifically single-cell transcriptomics. By transforming gene expression data into “cell sentences,” C2S bridges the gap between natural language processing and biology. We demonstrate cell sentences enable the finetuning of language models for diverse tasks in biology, including cell generation, complex celltype annotation, and direct data-driven text generation. Our experiments reveal that GPT-2, when fine-tuned with C2S, can generate biologically valid cells based on cell type inputs, and accurately predict cell types from cell sentences. This illustrates that language models, through C2S finetuning, can acquire a significant understanding of single-cell biology while maintaining robust text generation capabilities. C2S offers a flexible, accessible framework to integrate natural language processing with transcriptomics, utilizing existing models and libraries for a wide range of biological applications.
author	Daniel Levine and Sacha Lévy and S. Rizvi and Nazreen Pallikkavaliyaveetil and Xingyu Chen and David Zhang and Sina Ghadermarzi and Ruiming Wu and Zihe Zheng and Ivan Vrkic and Anna Zhong and Daphne Raskin and Insu Han and Antonio Henrique de Oliveira Fonseca and J. O. Caro and Amin Karbasi and Rahul M. Dhodapkar and David van Dijk
journal	bioRxiv
volume	null
pages	null
doi	10.1101/2023.09.11.557287
article	**1cc337ca86b52d3ecd3ffc19c23dde8d2fd05e56**
title	Sfaira accelerates data and model reuse in single cell genomics
year	2020
url	https://www.semanticscholar.org/paper/1cc337ca86b52d3ecd3ffc19c23dde8d2fd05e56
abstract	S2 TL;DR: This work proposes an adaption of cross-entropy loss for cell type classification tailored to datasets annotated at different levels of coarseness and demonstrates the utility of sfaira by training models across anatomic data partitions on 8 million cells.

author	David S. Fischer and L. Dony and Martin König and A. Moeed and L. Zappia and Sophie Tritschler and Olle Holmberg and H. Aliee and Fabian J Theis
journal	Genome Biology
volume	22
pages	null
doi	10.1186/s13059-021-02452-6
pmid	34433466
article	**3d7e004b8a20467937231539a805e3d5c9258c15**
title	BIOFORMERS: A SCALABLE FRAMEWORK FOR EXPLORING BIOSTATES USING TRANSFORMERS
year	2023
url	https://www.semanticscholar.org/paper/3d7e004b8a20467937231539a805e3d5c9258c15
abstract	Generative pre-trained models, such as BERT and GPT, have demonstrated remarkable success in natural language processing and computer vision. Leveraging the combination of large-scale, diverse datasets, transformers, and unsupervised learning, these models have emerged as a promising method for understanding complex systems like language. Despite the apparent differences, human language and biological systems share numerous parallels. Biology, like language, is a dynamic, interconnected network where biomolecules interact to create living entities akin to words forming coherent narratives. Inspired by this analogy, we explored the potential of using transformer-based unsupervised model development for analyzing biological systems and proposed a framework that can ingest vast amounts of biological data to create a foundational model of biology using BERT or GPT. This framework focuses on the concept of a ‘biostate,’ defined as a high-dimensional vector encompassing various biological markers such as genomic, proteomic, transcriptomic, physiological, and phenotypical data. We applied this technique to a small dataset of single-cell transcriptomics to demonstrate its ability to capture meaningful biological insights into genes and cells, even without any pre-training. Furthermore, the model can be readily used for gene network inference and genetic perturbation prediction.
author	Siham Amara-Belgadi and Orion Li and D. Zhang and Ashwin Gopinath
journal	bioRxiv
volume	null
pages	null
doi	10.1101/2023.11.29.569320
article	**9b13ab48f6ca0c9a4ffe3157aa9e5b2a8d6e9e78**
title	Scalable querying of human cell atlases via a foundational model reveals commonalities across fibrosis-associated macrophages
year	2023
url	https://www.semanticscholar.org/paper/9b13ab48f6ca0c9a4ffe3157aa9e5b2a8d6e9e78
abstract	Single-cell RNA-seq (scRNA-seq) studies have profiled over 100 million human cells across diseases, developmental stages, and perturbations to date. A singular view of this vast and growing expression landscape could help reveal novel associations between cell states and diseases, discover cell states in unexpected tissue contexts, and relate in vivo cells to in vitro models. However, these require a common, scalable representation of cell profiles from across the body, a general measure of their similarity, and an efficient way to query these data. Here, we present SCimilarity, a metric learning framework to learn and search a unified and interpretable representation that annotates cell types and instantaneously queries for a cell state across tens of millions of profiles. We

	demonstrate SCimilarity on a 22.7 million cell corpus assembled across 399 published scRNA-seq studies, showing accurate integration, annotation and querying. We experimentally validated SCimilarity by querying across tissues for a macrophage subset originally identified in interstitial lung disease, and showing that cells with similar profiles are found in other fibrotic diseases, tissues, and a 3D hydrogel system, which we then repurposed to yield this cell state in vitro. SCimilarity serves as a foundational model for single cell gene expression data and enables researchers to query for similar cellular states across the entire human body, providing a powerful tool for generating novel biological insights from the growing Human Cell Atlas.
author	Graham S. Heimberg and Tony Kuo and D. DePianto and Tobias Heigl and N. Diamant and Omar Salem and Gabriele Scalia and Tommaso Biancalani and S. Turley and Jason Rock and H. C. Bravo and J. Kaminker and J. V. Heiden and A. Regev
journal	bioRxiv
volume	null
pages	null
doi	10.1101/2023.07.18.549537
article	**ea7b4dd0ba152eb86511503603718 56e63487167**
title	Deep learning does not outperform classical machine learning for cell-type annotation
year	2019
url	https://www.semanticscholar.org/paper/ea7b4dd0ba152eb8651150360371856e63487167
abstract	Deep learning has revolutionized image analysis and natural language processing with remarkable accuracies in prediction tasks, such as image labeling and semantic segmentation or named-entity recognition and semantic role labeling. Specifically, the combination of algorithmic and hardware advances with the appearance of large and well-labeled datasets has led up to seminal contributions in these fields. The emergence of large amounts of data from single-cell RNA-seq and the recent global effort to chart all cell types in the Human Cell Atlas has attracted an interest in deep-learning applications. However, all current approaches are unsupervised, i.e., learning of latent spaces without using any cell labels, even though supervised learning approaches are often more powerful in feature learning and the most popular approach in the current AI revolution by far. Here, we ask why this is the case. In particular we ask whether supervised deep learning can be used for cell annotation, i.e. to predict cell-type labels from single-cell gene expression profiles. After evaluating 10 classification methods across 14 datasets, we notably find that deep learning does not outperform classical machine-learning methods in the task. Thus, cell-type prediction based on gene-signature derived cell-type labels is potentially too simplistic a task for complex non-linear methods, which demands better labels of functional single-cell readouts.
author	Niklas D. Köhler and M. Büttner and Niry Andriamanga and Fabian J Theis
journal	bioRxiv
volume	null
pages	null
doi	10.1101/653907
article	**58ec472ddfdd86e34f573d35350e65391bd867c9**
title	Toward subtask decomposition-based learning and benchmarking for genetic perturbation outcome prediction and beyond
year	2024
url	https://www.semanticscholar.org/paper/58ec472ddfdd86e34f573d35350e65391bd867c9

abstract	Deciphering cellular responses to genetic perturbations is fundamental for a wide array of biomedical applications, ranging from uncovering gene roles and interactions to unraveling effective therapeutics. Accurately predicting the transcriptional outcomes of genetic perturbations is indispensable for optimizing experimental perturbations and deciphering cellular response mechanisms; however, three scenarios present principal challenges, i.e., predicting single genetic perturbation outcomes, predicting multiple genetic perturbation outcomes and predicting genetic outcomes across cell lines. In this study, we introduce SubTAsk decomposition Modeling for genetic Perturbation prediction (STAMP), a conceptually novel computational strategy for genetic perturbation outcome prediction and downstream applications. STAMP innovatively formulates genetic perturbation prediction as a subtask decomposition (STD) problem by resolving three progressive subtasks in a divide-and-conquer manner, i.e., identifying differentially expressed gene (DEG) postperturbations, determining the regulatory directions of DEGs and finally estimating the magnitudes of gene expression changes. In addition to facilitating perturbation prediction, STAMP also serves as a robust and generalizable benchmark guide for evaluating various genetic perturbation prediction models. As a result, STAMP exhibits a substantial improvement in terms of its genetic perturbation prediction ability over the existing approaches on three subtasks and beyond, including revealing the ability to identify key regulatory genes and pathways on small samples and to reveal precise genetic interactions. Overall, STAMP serves as a fundamentally novel and effective prediction and generalizable benchmarking strategy that can facilitate genetic perturbation prediction, guide the design of perturbation experiments, and broaden the understanding of perturbation mechanisms.
author	Yicheng Gao and Zhiting Wei and Kejing Dong and Jingya Yang and Guohui Chuai and Qi Liu
journal	bioRxiv
volume	null
pages	null
doi	10.1101/2024.01.17.576034
article	**69d5ed2b0836931bd41ded2b0e8452eb3baae4f4**
title	scHyena: Foundation Model for Full-Length Single-Cell RNA-Seq Analysis in Brain
year	2023
url	https://www.semanticscholar.org/paper/69d5ed2b0836931bd41ded2b0e8452eb3baae4f4
abstract	Single-cell RNA sequencing (scRNA-seq) has made significant strides in unraveling the intricate cellular diversity within complex tissues. This is particularly critical in the brain, presenting a greater diversity of cell types than other tissue types, to gain a deeper understanding of brain function within various cellular contexts. However, analyzing scRNA-seq data remains a challenge due to inherent measurement noise stemming from dropout events and the limited utilization of extensive gene expression information. In this work, we introduce scHyena, a foundation model designed to address these challenges and enhance the accuracy of scRNA-seq analysis in the brain. Specifically, inspired by the recent Hyena operator, we design a novel Transformer architecture called singe-cell Hyena (scHyena) that is equipped with a linear adaptor layer, the positional encoding via gene-embedding, and a {bidirectional} Hyena operator. This enables us to process full-length scRNA-seq data without losing any information from the raw data. In particular, our model learns generalizable features of cells and genes through pre-training scHyena using the full length of scRNA-seq data. We demonstrate the superior performance of scHyena compared to other benchmark methods in downstream tasks, including cell type classification and scRNA-seq imputation.
author	Gyutaek Oh and B. Choi and Inkyung Jung and Jong Chul Ye
journal	ArXiv
volume	abs/2310.02713
pages	null

doi	10.48550/arXiv.2310.02713
arxivid	2310.02713
article	**2749975ade22c31c4e35d97afcf62c749235e9f3**
title	scPerturb: Harmonized Single-Cell Perturbation Data
year	2023
url	https://www.semanticscholar.org/paper/2749975ade22c31c4e35d97afcf62c749235e9f3
abstract	Recent biotechnological advances led to growing numbers of single-cell perturbation studies, which reveal molecular and phenotypic responses to large numbers of perturbations. However, analysis across diverse datasets is typically hampered by differences in format, naming conventions, and data filtering. In order to facilitate development and benchmarking of computational methods in systems biology, we collect a set of 44 publicly available single-cell perturbation-response datasets with molecular readouts, including transcriptomics, proteomics and epigenomics. We apply uniform pre-processing and quality control pipelines and harmonize feature annotations. The resulting information resource enables efficient development and testing of computational analysis methods, and facilitates direct comparison and integration across datasets. In addition, we introduce E-statistics for perturbation effect quantification and significance testing, and demonstrate E-distance as a general distance measure for single cell data. Using these datasets, we illustrate the application of E-statistics for quantifying perturbation similarity and efficacy. The data and a package for computing E-statistics is publicly available at scperturb.org. This work provides an information resource and guide for researchers working with single-cell perturbation data, highlights conceptual considerations for new experiments, and makes concrete recommendations for optimal cell counts and read depth.
author	Stefan Peidli and Tessa D. Green and Ciyue Shen and T. Gross and Joseph K Min and Samuele Garda and Bo Yuan and L. Schumacher and J. Taylor-King and D. Marks and Augustin Luna and N. Blüthgen and C. Sander
journal	bioRxiv
volume	null
pages	null
doi	10.1101/2022.08.20.504663
pmid	38279009
article	**cec0cf40c589ecaa6d779b8433d96393fdcd9fdc**
title	Mapping single-cell data to reference atlases by transfer learning
year	2021
url	https://www.semanticscholar.org/paper/cec0cf40c589ecaa6d779b8433d96393fdcd9fdc
abstract	S2 TL;DR: Deep learning strategy for mapping query datasets on top of a reference called single-cell architectural surgery (scArches), which uses transfer learning and parameter optimization to enable efficient, decentralized, iterative reference building and contextualization of new datasets with existing references without sharing raw data.
author	M. Lotfollahi and Mohsen Naghipourfar and M. Luecken and Matin Khajavi and M. Büttner and Marco Wagenstetter and Žiga Avsec and Adam Gayoso and N. Yosef and M. Interlandi and Sergei Rybakov and A. Misharin and Fabian J Theis
journal	Nature Biotechnology

volume	40
pages	121 - 130
doi	10.1038/s41587-021-01001-7
pmid	34462589
article	**c520d8a888355f7abb7728b2e2510fe7bc63f814**
title	Large Scale Foundation Model on Single-cell Transcriptomics
year	2023
url	https://www.semanticscholar.org/paper/c520d8a888355f7abb7728b2e2510fe7bc63f814
abstract	Large-scale pretrained models have become foundation models leading to breakthroughs in natural language processing and related fields. Developing foundation models in life science for deciphering the “languages” of cells and facilitating biomedical research is promising yet challenging. We developed a large-scale pretrained model scFoundation with 100M parameters for this purpose. scFoundation was trained on over 50 million human single-cell transcriptomics data, which contain high-throughput observations on the complex molecular features in all known types of cells. scFoundation is currently the largest model in terms of the size of trainable parameters, dimensionality of genes and the number of cells used in the pre-training. Experiments showed that scFoundation can serve as a foundation model for single-cell transcriptomics and achieve state-of-the-art performances in a diverse array of downstream tasks, such as gene expression enhancement, tissue drug response prediction, single-cell drug response classification, and single-cell perturbation prediction.
author	Minsheng Hao and Jing Gong and Xin Zeng and Chiming Liu and Yucheng Guo and Xingyi Cheng and Taifeng Wang and Jianzhu Ma and Leo T. Song and Xuegong Zhang
journal	bioRxiv
volume	null
pages	null
doi	10.1101/2023.05.29.542705
article	**f5e380b3b0534a5e89ba5338fb4dce4b84ec6d13**
title	Predicting cellular responses to complex perturbations in high-throughput screens
year	2023
url	https://www.semanticscholar.org/paper/f5e380b3b0534a5e89ba5338fb4dce4b84ec6d13
abstract	Recent advances in multiplexed single-cell transcriptomics experiments facilitate the high-throughput study of drug and genetic perturbations. However, an exhaustive exploration of the combinatorial perturbation space is experimentally unfeasible. Therefore, computational methods are needed to predict, interpret, and prioritize perturbations. Here, we present the compositional perturbation autoencoder (CPA), which combines the interpretability of linear models with the flexibility of deep-learning approaches for single-cell response modeling. CPA learns to in silico predict transcriptional perturbation response at the single-cell level for unseen dosages, cell types, time points, and species. Using newly generated single-cell drug combination data, we validate that CPA can predict unseen drug combinations while outperforming baseline models. Additionally, the architecture's modularity enables incorporating the chemical representation of the drugs, allowing the prediction of cellular response to completely unseen drugs. Furthermore, CPA is also applicable to genetic combinatorial screens. We demonstrate this by imputing in silico 5,329 missing combinations (97.6% of all possibilities) in a single-cell Perturb-seq experiment with diverse genetic interactions. We envision CPA will facilitate efficient experimental design and hypothesis generation by enabling in silico response prediction at the single-cell

	level and thus accelerate therapeutic applications using single-cell technologies.
author	M. Lotfollahi and Anna Klimovskaia Susmelj and C. De Donno and Leon Hetzel and Yuge Ji and Ignacio L Ibarra and Sanjay R Srivatsan and Mohsen Naghipourfar and R. Daza and Beth K. Martin and J. Shendure and José L. McFaline-Figueroa and Pierre Boyeau and F. A. Wolf and N. Yakubova and Stephan Günnemann and C. Trapnell and David Lopez-Paz and Fabian J Theis
journal	Molecular Systems Biology
volume	19
pages	null
doi	10.15252/msb.202211517
pmid	37154091
article	**f26ca7716b8e4c809901dfa4fa65e822ede08b3e**
title	Disentanglement of single-cell data with biolord.
year	2024
url	https://www.semanticscholar.org/paper/f26ca7716b8e4c809901dfa4fa65e822ede08b3e
abstract	null
author	Z. Piran and Niv Cohen and Yedid Hoshen and M. Nitzan
journal	Nature biotechnology
volume	null
pages	null
doi	10.1038/s41587-023-02079-x
pmid	38225466
article	**7d1e59ce254bea5228da634dbe7c5c4160df6f98**
title	Transfer learning enables predictions in network biology
year	2023
url	https://www.semanticscholar.org/paper/7d1e59ce254bea5228da634dbe7c5c4160df6f98
abstract	S2 TL;DR: A context-aware, attention-based deep learning model pretrained on single-cell transcriptomes enables predictions in settings with limited data in network biology and could accelerate discovery of key network regulators and candidate therapeutic targets.
author	Christina V. Theodoris and Ling Xiao and Anant Chopra and M. Chaffin and Z. A. Al Sayed and M. C. Hill and Helene Mantineo and Elizabeth M Brydon and Zexian Zeng and X. S. Liu and P. Ellinor
journal	Nature
volume	618

pages	616-624
doi	10.1038/s41586-023-06139-9
pmid	37258680
article	**b1e90b67675b6d7ae88b563a93cb4d375857cb15**
title	CellPLM: Pre-training of Cell Language Model Beyond Single Cells
year	2023
url	https://www.semanticscholar.org/paper/b1e90b67675b6d7ae88b563a93cb4d375857cb15
abstract	The current state-of-the-art single-cell pre-trained models are greatly inspired by the success of large language models. They trained transformers by treating genes as tokens and cells as sentences. However, three fundamental differences between single-cell data and natural language data are overlooked: (1) scRNA-seq data are presented as bag-of-genes instead of sequences of RNAs; (2) Cell-cell relations are more intricate and important than inter-sentence relations; and (3) The quantity of single-cell data is considerably inferior to text data, and they are very noisy. In light of these characteristics, we propose a new pre-trained model CellPLM, which takes cells as tokens and tissues as sentences. In addition, we leverage spatially-resolved transcriptomic data in pre-training to facilitate learning cell-cell relationships and introduce a Gaussian mixture prior distribution as an additional inductive bias to overcome data limitation. CellPLM is the first single-cell pre-trained transformer that encodes cell-cell relations and it consistently outperforms existing pre-trained and non-pre-trained models in diverse downstream tasks, with 100x times higher inference speed compared to existing pre-trained models.
author	Hongzhi Wen and Wenzhuo Tang and Xinnan Dai and Jiayuan Ding and Wei Jin and Yuying Xie and Jiliang Tang
journal	bioRxiv
volume	null
pages	null
doi	10.1101/2023.10.03.560734
article	**edb5495bc9081f2adfc8de51e0981510802e4090**
title	scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data
year	2022
url	https://www.semanticscholar.org/paper/edb5495bc9081f2adfc8de51e0981510802e4090
abstract	S2 TL;DR: A pretrained deep neural network-based model, single-cell bidirectional encoder representations from transformers (scBERT), to overcome the challenges of existing annotation methods, and is validated on cell type annotation, novel cell type discovery, robustness to batch effects and model interpretability.
author	Wenchuan Wang and Fan Yang and Yuejing Fang and Duyu Tang and Junzhou Huang and Hui Lu and Jianhua Yao
journal	Nature Machine Intelligence
volume	4
pages	852 - 866
doi	10.1038/s42256-022-00534-z

article	895ff8cd086b5c6238d66604dec6a1ede1e15db7
title	Evaluating the Utilities of Foundation Models in Single-cell Data Analysis
year	2024
url	https://www.semanticscholar.org/paper/f342713d3f5ac5530ff13fe1ff168aa8a6d28c94
abstract	Foundation Models (FMs) have made significant strides in both industrial and scientific domains. In this paper, we evaluate the performance of FMs in single-cell sequencing data analysis through comprehensive experiments across eight downstream tasks pertinent to single-cell data. By comparing ten different single-cell FMs with task-specific methods, we found that single-cell FMs may not consistently excel in all tasks than task-specific methods. However, the emergent abilities and the successful applications of cross-species/cross-modality transfer learning of FMs are promising. In addition, we present a systematic evaluation of the effects of hyper-parameters, initial settings, and stability for training single-cell FMs based on a proposed scEval framework, and provide guidelines for pre-training and fine-tuning. Our work summarizes the current state of single-cell FMs and points to their constraints and avenues for future development.
author	Tianyu Liu and Kexing Li and Yuge Wang and Hongyu Li and Hongyu Zhao
journal	bioRxiv
volume	null
pages	null
doi	10.1101/2023.09.08.555192
article	**e60fe8a9b9ca2210b15b92be5fb364798de73330**
title	Transformer for one stop interpretable cell type annotation
year	2023
url	https://www.semanticscholar.org/paper/e60fe8a9b9ca2210b15b92be5fb364798de73330
abstract	S2 TL;DR: TOSICA is a multi-head self-attention deep learning model based on Transformer that enables interpretable cell type annotation using biologically understandable entities, such as pathways or regulons, and its performance on large or atlas datasets is demonstrated.
author	Jiawei Chen and Hao Xu and Wanyu Tao and Zhaoxiong Chen and Yuxuan Zhao and Jing-Dong J. Han
journal	Nature Communications
volume	14
pages	null
doi	10.1038/s41467-023-35923-4
pmid	36641532
article	**e6411c3f02401c4a1712f8cd9b1947c226ead48c**
title	Evaluation of some aspects in supervised cell type identification for single-cell RNA-seq: classifier, feature selection, and reference construction
year	2021

url	https://www.semanticscholar.org/paper/e6411c3f02401c4a1712f8cd9b1947c226ead48c
abstract	S2 TL;DR: This work performs extensive real data analyses to systematically evaluate feature selection, prediction method, and choice of the reference dataset strategies in supervised cell identification and provides guidelines for using supervised cell typing methods.
author	Wenjing Ma and Kenong Su and Hao Wu
journal	Genome Biology
volume	22
pages	null
doi	10.1186/s13059-021-02480-2
pmid	34503564
article	**424132ec245c3173685751ac1101c3be6cc55a67**
title	xTrimoGene: An Efficient and Scalable Representation Learner for Single-Cell RNA-Seq Data
year	2023
url	https://www.semanticscholar.org/paper/424132ec245c3173685751ac1101c3be6cc55a67
abstract	The advances in high-throughput sequencing technology have led to significant progress in measuring gene expressions in single-cell level. The amount of publicly available single-cell RNA-seq (scRNA-seq) data is already surpassing 50M records for human with each record measuring 20,000 genes. This highlights the need for unsupervised representation learning to fully ingest these data, yet classical transformer architectures are prohibitive to train on such data in terms of both computation and memory. To address this challenge, we propose a novel asymmetric encoder-decoder transformer for scRNA-seq data, called xTrimoGene, which leverages the sparse characteristic of the data to scale up the pre-training. This scalable design of xTrimoGene reduces FLOPs by one to two orders of magnitude compared to classical transformers while maintaining high accuracy, enabling us to train the largest transformer models over the largest scRNA-seq dataset today. Our experiments also show that the performance of xTrimoGene improves as we increase the model sizes, and it also leads to SOTA performance over various downstream tasks, such as cell classification, perturb-seq effect prediction, and drug combination prediction.
author	Jing Gong and Minsheng Hao and Xin Zeng and Chiming Liu and Jianzhu Ma and Xingyi Cheng and Taifeng Wang and Xuegong Zhang and Leo T. Song
journal	bioRxiv
volume	null
pages	null
doi	10.1101/2023.03.24.534055
arxivid	2311.15156
article	**78f6f540df4d86d081c5d203db90d62c37b89623**
title	Deep identifiable modeling of single-cell atlases enables zero-shot query of cellular states
year	2023
url	https://www.semanticscholar.org/paper/78f6f540df4d86d081c5d203db90d62c37b89623

abstract	With the emerging single-cell RNA-seq datasets at atlas levels, the potential of a universal model built on existing atlas that can extrapolate to new data remains unclear. A fundamental yet challenging problem for such a model is to identify the underlying biological and batch variations in a zero-shot manner, which is crucial for characterizing scRNA-seq datasets with new biological states. In this work, we present scShift, a mechanistic model that learns batch and biological patterns from atlas-level scRNA-seq data as well as perturbation scRNA-seq data. scShift models genes as functions of latent biological processes, with sparse shifts induced by batch effects and biological perturbations, leveraging recent advances of causal representation learning. Through benchmarking in holdout real datasets, we show scShift reveals unified cell type representations as well as underlying biological variations for query data in zero-shot manners, outperforming widely-used atlas integration, batch correction, and perturbation modeling approaches. scShift enables mapping of gene expression profiles to perturbation labels, and predicts meaningful targets for exhausted T cells as well as a list of diseases in the CellxGene blood atlas.
author	Mingze Dong and Y. Kluger
journal	bioRxiv
volume	null
pages	null
doi	10.1101/2023.11.11.566161
pmid	38014345
article	**08d081e969f0fc00944cb1be98aa5ed08cf992d3**
title	GenePT: A Simple But Hard-to-Beat Foundation Model for Genes and Cells Built From ChatGPT
year	2023
url	https://www.semanticscholar.org/paper/08d081e969f0fc00944cb1be98aa5ed08cf992d3
abstract	There has been significant recent progress in leveraging large-scale gene expression data to develop foundation models for single-cell transcriptomes such as Geneformer [1], scGPT [2], and scBERT [3]. These models infer gene functions and interrelations from the gene expression profiles of millions of cells, which requires extensive data curation and resource-intensive training. Here, we explore a much simpler alternative by leveraging ChatGPT embeddings of genes based on literature. Our proposal, GenePT, uses NCBI text descriptions of individual genes with GPT-3.5 to generate gene embeddings. From there, GenePT generates single-cell embeddings in two ways: (i) by averaging the gene embeddings, weighted by each gene's expression level; or (ii) by creating a sentence embedding for each cell, using gene names ordered by the expression level. Without the need for dataset curation and additional pretraining, GenePT is efficient and easy to use. On many downstream tasks used to evaluate recent single-cell foundation models — e.g., classifying gene properties and cell types — GenePT achieves comparable, and often better, performance than Geneformer and other methods. GenePT demonstrates that large language model embedding of literature is a simple and effective path for biological foundation models.
author	Yiqun T. Chen and James Zou
journal	bioRxiv
volume	null
pages	null
doi	10.1101/2023.10.16.562533
pmid	37905130
article	**ab3d8119deba91f0b78e3ccdb86913675668e774**

title	Decoding cell identity with multi-scale explainable deep learning
year	2024
url	https://www.semanticscholar.org/paper/ab3d8119deba91f0b78e3ccdb86913675668e774
abstract	Cells are the fundamental structural and functional units of life. Studying the definition and composition of different cell types can help us understand the complex mechanisms underlying biological diversity and functionality. The increasing volume of extensive single-cell omics data makes it possible to provide detailed characterisations of cell types. Recently, there has been a rise in deep learning-based approaches that generate cell type labels solely through mapping query data to reference data. However, these approaches lack multi-scale descriptions and interpretations of identified cell types. Here, we propose Cell Decoder, a biological prior knowledge informed model to achieve multi-scale representation of cells. We implemented automated machine learning and post-hoc analysis techniques to decode cell identity. We have shown that Cell Decoder compares favourably to existing methods, offering multi-view interpretability for decoding cell identity and data integration. Furthermore, we have showcased its applicability in uncovering novel cell types and states in both human bone and mouse embryonic contexts, thereby revealing the multi-scale heterogeneity inherent in cell identities.
author	Jun Zhu and Zeyang Zhang and Yujia Xiang and Beini Xie and Xinwen Dong and Linhai Xie and Peijie Zhou and Rongyan Yao and Xiaowen Wang and Yang Li and Fuchu He and Wenwu Zhu and Ziwei Zhang and Cheng Chang
journal	bioRxiv
volume	null
pages	null
doi	10.1101/2024.02.05.578922
article	**88dabc93c9503c18ca1811b9fab3a968ccf5b010**
title	Optimal distance metrics for single-cell RNA-seq populations
year	2023
url	https://www.semanticscholar.org/paper/88dabc93c9503c18ca1811b9fab3a968ccf5b010
abstract	In single-cell data workflows and modeling, distance metrics are commonly used in loss functions, model evaluation, and subpopulation analysis. However, these metrics behave differently depending on the source of variation, conditions and subpopulations in single-cell expression profiles due to data sparsity and high dimensionality. Thus, the metrics used for downstream tasks in this domain should be carefully selected. We establish a set of benchmarks with three evaluation measures, capturing desirable facets of absolute and relative distance behavior. Based on seven datasets using perturbation as ground truth, we evaluated 16 distance metrics applied to scRNA-seq data and demonstrated their application to three use cases. We find that linear metrics such as mean squared error (MSE) performed best across our three evaluation criteria. Therefore, we recommend the use of MSE for comparing single-cell RNA-seq populations and evaluating gene expression prediction models.
author	Yuge Ji and Tessa D. Green and Stefan Peidli and Mojtaba Bahrami and Meiqi Liu and L. Zappia and Karin Hrovatin and C. Sander and F. Theis
journal	bioRxiv
volume	null
pages	null
doi	10.1101/2023.12.26.572833
article	**bfd2b76998a0521c12903ef5ced517adf70ad2ba**

title	HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution
year	2023
url	https://www.semanticscholar.org/paper/bfd2b76998a0521c12903ef5ced517adf70ad2ba
abstract	Genomic (DNA) sequences encode an enormous amount of information for gene regulation and protein synthesis. Similar to natural language models, researchers have proposed foundation models in genomics to learn generalizable features from unlabeled genome data that can then be fine-tuned for downstream tasks such as identifying regulatory elements. Due to the quadratic scaling of attention, previous Transformer-based genomic models have used 512 to 4k tokens as context (<0.001% of the human genome), significantly limiting the modeling of long-range interactions in DNA. In addition, these methods rely on tokenizers to aggregate meaningful DNA units, losing single nucleotide resolution where subtle genetic variations can completely alter protein function via single nucleotide polymorphisms (SNPs). Recently, Hyena, a large language model based on implicit convolutions was shown to match attention in quality while allowing longer context lengths and lower time complexity. Leveraging Hyenas new long-range capabilities, we present HyenaDNA, a genomic foundation model pretrained on the human reference genome with context lengths of up to 1 million tokens at the single nucleotide-level, an up to 500x increase over previous dense attention-based models. HyenaDNA scales sub-quadratically in sequence length (training up to 160x faster than Transformer), uses single nucleotide tokens, and has full global context at each layer. We explore what longer context enables - including the first use of in-context learning in genomics for simple adaptation to novel tasks without updating pretrained model weights. On fine-tuned benchmarks from the Nucleotide Transformer, HyenaDNA reaches state-of-the-art (SotA) on 12 of 17 datasets using a model with orders of magnitude less parameters and pretraining data. On the GenomicBenchmarks, HyenaDNA surpasses SotA on all 8 datasets on average by +9 accuracy points.
author	Eric D Nguyen and Michael Poli and Marjan Faizi and A. Thomas and C. Birch-sykes and Michael Wornow and Aman Patel and Clayton M. Rabideau and Stefano Massaroli and Y. Bengio and Stefano Ermon and S. Baccus and Christopher Ré
journal	ArXiv
volume	null
pages	null
doi	10.48550/arXiv.2306.15794
pmid	37426456
arxivid	2306.15794
article	**739e40fd65468ba8422ff30d3905798e505771a1**
title	scFormer: A Universal Representation Learning Approach for Single-Cell Data Using Transformers
year	2022
url	https://www.semanticscholar.org/paper/739e40fd65468ba8422ff30d3905798e505771a1
abstract	Single-cell sequencing has emerged as a promising technique to decode cellular heterogeneity and analyze gene functions. With the high throughput of modern techniques and resulting large-scale sequencing data, deep learning has been used extensively to learn representations of individual cells for downstream tasks. However, most existing methods rely on fully connected networks and are unable to model complex relationships between both cell and gene representations. We hereby propose scFormer, a novel transformer-based deep learning framework to jointly optimize cell and gene embeddings for single-cell biology in an unsupervised manner. By drawing parallels between natural language processing and genomics, scFormer applies self-attention to learn salient gene and cell embeddings through masked gene modelling. scFormer provides a unified framework to readily address a variety of downstream tasks such as data integration, analysis of gene function, and perturbation response prediction. Extensive experiments using scFormer show state-of-the-art performance on seven datasets across the relevant tasks. The scFormer model implementation is available at https://github.com/bowang-lab/scFormer.

author	Haotian Cui and Chloe X. Wang and Hassaan Maan and Nan Duan and Bo Wang
journal	bioRxiv
volume	null
pages	null
doi	10.1101/2022.11.20.517285
article	**275fb93244b5a465d7e30fc6111e3403b47557be**
title	scGPT: Towards Building a Foundation Model for Single-Cell 1 Multi-omics Using Generative AI
year	2023
url	https://www.semanticscholar.org/paper/88399c4c6574be850918a1bce2dede2c87ef8241
abstract	10 Generative pre-trained models have achieved remarkable success in various domains such as nat-11 ural language processing and computer vision. Specifically, the combination of large-scale diverse 12 datasets and pre-trained transformers has emerged as a promising approach for developing founda-13 tion models. While texts are made up of words, cells can be characterized by genes. This analogy 14 inspires us to explore the potential of foundation models for cell and gene biology. By leveraging the
author	Haotian Cui and Chloe X. Wang and Hassaan Maan and Bo Wang and C. E. D. Masked-Attention
journal	bioRxiv
volume	null
pages	null
doi	10.1101/2023.04.30.538439
article	**c290247b923304154fbc2842d97914c613ef66f5**
title	GeneCompass: Deciphering Universal Gene Regulatory Mechanisms with Knowledge-Informed Cross-Species Foundation Model
year	2023
url	https://www.semanticscholar.org/paper/c290247b923304154fbc2842d97914c613ef66f5
abstract	Deciphering the universal gene regulatory mechanisms in diverse organisms holds great potential to advance our knowledge of fundamental life process and facilitate research on clinical applications. However, the traditional research paradigm primarily focuses on individual model organisms, resulting in limited collection and integration of complex features on various cell types across species. Recent breakthroughs in single-cell sequencing and advancements in deep learning techniques present an unprecedented opportunity to tackle this challenge. In this study, we developed GeneCompass, the first knowledge-informed, cross-species foundation model pre-trained on an extensive dataset of over 120 million single-cell transcriptomes from human and mouse. During pre-training, GeneCompass effectively integrates four types of biological prior knowledge to enhance the understanding of gene regulatory mechanisms in a self-supervised manner. Fine-tuning towards multiple downstream tasks, GeneCompass outperforms competing state-of-the-art models in multiple tasks on single species and unlocks new realms of cross-species biological investigation. Overall, GeneCompass marks a milestone in advancing knowledge of universal gene regulatory mechanisms and accelerating the discovery of key cell fate regulators and candidate targets for drug development.

author	Xiaodong Yang and Guole Liu and Guihai Feng and Dechao Bu and Pengfei Wang and Jie Jiang and Shubai Chen and Qinmeng Yang and Yiyang Zhang and Zhenpeng Man and Zhongming Liang and Zichen Wang and Yaning Li and Zheng Li and Yana Liu and Yao Tian and Ao Li and Jingxi Dong and Zhilong Hu and Chen Fang and Hefan Miao and Lina Cui and Zixu Deng and Haiping Jiang and Wentao Cui and Jiahao Zhang and Zhaohui Yang and Handong Li and Xingjian He and Liqun Zhong and Jiaheng Zhou and Zijian Wang and Qingqing Long and Ping Xu and Hongmei Wang and Z. Meng and Xuezhi Wang and Yangang Wang and Yong Wang and Shihua Zhang and Jingtao Guo and Yi Zhao and Yuanchun Zhou and Fei Li and Jing Liu and Yiqiang Chen and Ge Yang and Xin Li
journal	bioRxiv
volume	null
pages	null
doi	10.1101/2023.09.26.559542
article	**2ae47c27b0d6bc2a52b69359f3f7810fb371ffa0**
title	Benchmarking atlas-level data integration in single-cell genomics
year	2021
url	https://www.semanticscholar.org/paper/2ae47c27b0d6bc2a52b69359f3f7810fb371ffa0
abstract	S2 TL;DR: It is shown that highly variable gene selection improves the performance of data integration methods, whereas scaling pushes methods to prioritize batch removal over conservation of biological variation.
author	M. Luecken and M. Büttner and Kridsadakorn Chaichoompu and A. Danese and M. Interlandi and M. Mueller and D. Strobl and L. Zappia and M. Dugas and M. Colomé-Tatché and F. Theis
journal	Nature Methods
volume	19
pages	41 - 50
doi	10.1038/s41592-021-01336-8
pmid	34949812
article	**3be1dd73f04421743568dfcc9edd143ca511a8d0**
title	Online single-cell data integration through projecting heterogeneous datasets into a common cell-embedding space
year	2021
url	https://www.semanticscholar.org/paper/3be1dd73f04421743568dfcc9edd143ca511a8d0
abstract	S2 TL;DR: SCALEX substantially outperforms online iNMF and other state-of-the-art non-online integration methods on benchmark single-cell datasets of diverse modalities, especially for datasets with partial overlaps, accurately aligning similar cell populations while retaining true biological differences.
author	Lei Xiong and K. Tian and Yuzhe Li and Q. Zhang
journal	Nature Communications
volume	13

pages	null
doi	10.1038/s41467-022-33758-z
pmid	36253379
article	**38fe1305e704bdd5852226b8c109b7b6eb4253b1**
title	Construction of continuously expandable single-cell atlases through integration of heterogeneous datasets in a generalized cell-embedding space
year	2021
url	https://www.semanticscholar.org/paper/38fe1305e704bdd5852226b8c109b7b6eb4253b1
abstract	Single-cell RNA-seq and ATAC-seq analyses have been widely applied to decipher cell-type and regulation complexities. However, experimental conditions often confound biological variations when comparing data from different samples. For integrative single-cell data analysis, we have developed SCALEX, a deep generative framework that maps cells into a generalized, batch-invariant cell-embedding space. We demonstrate that SCALEX accurately and efficiently integrates heterogenous single-cell data using multiple benchmarks. It outperforms competing methods, especially for datasets with partial overlaps, accurately aligning similar cell populations while r,etaining true biological differences. We demonstrate the advantages of SCALEX by constructing continuously expandable single-cell atlases for human, mouse, and COVID-19, which were assembled from multiple data sources and can keep growing through the inclusion of new incoming data. Analyses based on these atlases revealed the complex cellular landscapes of human and mouse tissues and identified multiple peripheral immune subtypes associated with COVID-19 disease severity.
author	Lei Xiong and K. Tian and Yuzhe Li and Q. Zhang
journal	bioRxiv
volume	
pages	null
doi	10.21203/RS.3.RS-398163/V1
article	**ed7c2fd0daf57276f9f4aa8f6d2c4fa767ee85c2**
title	Contrastive learning enables rapid mapping to multimodal single-cell atlas of multimillion scale
year	2022
url	https://www.semanticscholar.org/paper/ed7c2fd0daf57276f9f4aa8f6d2c4fa767ee85c2
abstract	S2 TL;DR: Contrastive learning of cell representations, Concerto, is presented, which leverages a self-supervised distillation framework to model multimodal single-cell atlases and substantially outperforms competing methods.
author	Mengcheng Yang and Yueyuxiao Yang and Chenxi Xie and Ming Ni and Jian Liu and Huanming Yang and Feng Mu and J. Wang
journal	Nature Machine Intelligence
volume	4
pages	696 - 709
doi	10.1038/s42256-022-00518-z
article	**e2fde6d6f5ddb0cc14aca91203d195f65e769176**

title	Batch alignment of single-cell transcriptomics data using deep metric learning
year	2023
url	https://www.semanticscholar.org/paper/e2fde6d6f5ddb0cc14aca91203d195f65e769176
abstract	S2 TL;DR: Comprehensive evaluations spanning different species and tissues demonstrated that scDML can remove batch effect, improve clustering performance, accurately recover true cell types and consistently outperform popular methods such as Seurat 3, scVI, Scanorama, BBKNN, Harmony et al.
author	Xiaokang Yu and Xinyi Xu and Jingxiao Zhang and Xiangjie Li
journal	Nature Communications
volume	14
pages	null
doi	10.1038/s41467-023-36635-5
pmid	36810607
article	**38488a6e8872db54a6ef070429080f6c95b4c337**
title	Biological representation disentanglement of single-cell data
year	2023
url	https://www.semanticscholar.org/paper/38488a6e8872db54a6ef070429080f6c95b4c337
abstract	Due to its internal state or external environment, a cell's gene expression profile contains multiple signatures, simultaneously encoding information about its characteristics. Disentangling these factors of variations from single-cell data is needed to recover multiple layers of biological information and extract insight into the individual and collective behavior of cellular populations. While several recent methods were suggested for biological disentanglement, each has its limitations; they are either task-specific, cannot capture inherent nonlinear or interaction effects, cannot integrate layers of experimental data, or do not provide a general reconstruction procedure. We present biolord, a deep generative framework for disentangling known and unknown attributes in single-cell data. Biolord exposes the distinct effects of different biological processes or tissue structure on cellular gene expression. Based on that, biolord allows generating experimentally-inaccessible cell states by virtually shifting cells across time, space, and biological states. Specifically, we showcase accurate predictions of cellular responses to drug perturbations and generalization to predict responses to unseen drugs. Further, biolord disentangles spatial, temporal, and infection-related attributes and their associated gene expression signatures in a single-cell atlas of Plasmodium infection progression in the mouse liver. Biolord can handle partially labeled attributes by predicting a classification for missing labels, and hence can be used to computationally extend an infected hepatocyte population identified at a late stage of the infection to earlier stages. Biolord applies to diverse biological settings, is implemented using the scvi-tools library, and is released as open-source software at https://github.com/nitzanlab/biolord.
author	Z. Piran and Niv Cohen and Yedid Hoshen and M. Nitzan
journal	bioRxiv
volume	null
pages	null
doi	10.1101/2023.03.05.531195
article	**c4b7abadab4d34652e64f7790a0f3f93424620ef**

title	scAnnotatR: framework to accurately classify cell types in single-cell RNA-sequencing data
year	2022
url	https://www.semanticscholar.org/paper/c4b7abadab4d34652e64f7790a0f3f93424620ef
abstract	null
author	Vy Nguyen and J. Griss
journal	BMC Bioinformatics
volume	23
pages	null
doi	10.1186/s12859-022-04574-5
pmid	35038984
article	**7f90939656b4291602db13f90dcb96f3401ed97a**
title	Query to reference single-cell integration with transfer learning
year	2020
url	https://www.semanticscholar.org/paper/7f90939656b4291602db13f90dcb96f3401ed97a
abstract	Large single-cell atlases are now routinely generated with the aim of serving as reference to analyse future smaller-scale studies. Yet, learning from reference data is complicated by batch effects between datasets, limited availability of computational resources, and sharing restrictions on raw data. Leveraging advances in machine learning, we propose a deep learning strategy to map query datasets on top of a reference called single-cell architectural surgery (scArches, https://github.com/theislab/scarches). It uses transfer learning and parameter optimization to enable efficient, decentralized, iterative reference building, and the contextualization of new datasets with existing references without sharing raw data. Using examples from mouse brain, pancreas, and whole organism atlases, we showcase that scArches preserves nuanced biological state information while removing batch effects in the data, despite using four orders of magnitude fewer parameters compared to de novo integration. To demonstrate mapping disease variation, we show that scArches preserves detailed COVID-19 disease variation upon reference mapping, enabling discovery of new cell identities that are unseen during training. We envision our method to facilitate collaborative projects by enabling the iterative construction, updating, sharing, and efficient use of reference atlases.
author	M. Lotfollahi and Mohsen Naghipourfar and Malte D. Luecken and Matin Khajavi and M. Büttner and Žiga Avsec and A. Misharin and Fabian J Theis
journal	bioRxiv
volume	null
pages	null
doi	10.1101/2020.07.16.205997
article	**d81bdaf1abed1682e6eaeed465f9b2b50d3b441c**
title	A comparison of automatic cell identification methods for single-cell RNA sequencing data
year	2019
url	https://www.semanticscholar.org/paper/d81bdaf1abed1682e6eaeed465f9b2b50d3b441c

abstract	S2 TL;DR: It is found that most classifiers perform well on a variety of datasets with decreased accuracy for complex datasets with overlapping classes or deep annotations, but the general-purpose support vector machine classifier has overall the best performance across the different experiments.
author	T. Abdelaal and Lieke Michielsen and D. Cats and Dylan Hoogduin and H. Mei and M. Reinders and A. Mahfouz
journal	Genome Biology
volume	20
pages	null
doi	10.1186/s13059-019-1795-z
pmid	31500660