Self-distillation contrastive learning enables clustering-free signature extraction and mapping to multimodal single-cell atlas of multimillion scale

Meng Yang (yangmeng1@mgi-tech.com)
MGI, BGI-Shenzhen

Yueyuxiao Yang
MGI, BGI-Shenzhen

Haiping Huang
MGI, BGI-Shenzhen

Chenxi Xie
MGI, BGI-Shenzhen

Huanming Yang
BGI-Shenzhen

Feng Mu
MGI, BGI-Shenzhen

Article

Keywords: single-cell multi-omics datasets, cell representations, query-to-reference mapping

Posted Date: November 9th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-841909/v1

Version of Record: A version of this preprint was published at Nature Machine Intelligence on August 25th, 2022. See the published version at https://doi.org/10.1038/s42256-022-00518-z.

Self-distillation contrastive learning enables clustering-free signature extraction and mapping to multimodal single-cell atlas of multimillion scale

Meng Yang[1,2]*, Yueyuxiao Yang[1], Haiping Huang[1], Chenxi Xie[1], Huanming Yang[3,4], Feng Mu[1]*

[1]MGI, BGI-Shenzhen, Shenzhen 518083, China.
[2]Department of Biology, University of Copenhagen, Copenhagen DK-2200, Denmark.
[3]BGI-Shenzhen, Shenzhen 518083, China.
[4]Guangdong Provincial Academician Workstation of BGI Synthetic Genomics, BGI-Shenzhen, Shenzhen, 518120, China.

*Correspondence to: yangmeng1@mgi-tech.com, mufeng@mgi-tech.com

Massively generated single-cell multi-omics datasets are revolutionizing biological studies of heterogenous tissues and organisms, which necessitate powerful computational methods to unleash the full potential of these tremendous data. Here, we present Concerto, stands for self-distillation contrastive learning of cell representations, a self-supervised representation learning framework optimized with asymmetric teacher-student configuration to analyze single-cell multi-omics datasets with scalability up to building 10 million-cell reference within 1.5 hour and querying 10k cells within 8 seconds. Concerto leverages dropout layer as minimal data augmentation to learn meaningful cell representations in a contrastive manner. The teacher module uses attention mechanism to aggregate contextualized gene embeddings within cellular context, while the student module uses simpler dense structure with discreate input. The learned task-agnostic representations can be adapted to a broad range of single-cell computation tasks. 1) Via supervised fine-tuning, Concerto enables automatic cell classification as well as novel cell-type discovery; 2) Attention weights provide model interpretability via automatically extracting specific molecular signatures at single-cell resolution without the needs of clustering; 3) Via source-aware training, Concerto supports efficient data integration by projecting all cells across multiple batches into a joint embedding space. 4) Via batch-aware inference or unsupervised fine-tuning, Concerto enables mapping query cells onto reference and accurately transferring annotations. Concerto can flexibly extend to multi-omics datasets simply through cross-modality summation operation to obtain unified cell embeddings. Using examples from human peripheral blood, human thymus, human pancreas, and mouse tissue atlas, Concerto shows superior performance benchmarking against other top-performing methods. We also demonstrate Concerto recapitulates detailed COVID-19 disease variation through query-to-reference mapping. Concerto can operate on all genes and represents a fully data-driven approach with minimum prior distribution assumptions, eliminating the needs of PCA-like or autoencoder-like dimensionality reduction, which significantly reforms the current best practice. Concerto is a simple, straightforward, robust, and scalable framework, offering a brand new perspective to derive cell representations and can effectively satisfy the emerging paradigm of query-to-reference mapping in the era of atlas-level single-cell multimodal analysis.

Recent advancements in high-throughput single-cell sequencing technologies have opened a new way to characterize heterogenous tissues or systems at single-cell resolution. Large-scale consortium efforts, such as Human Cell Atlas[1] and Tabula Muris Atlas[2], are accumulating profiling datasets of millions of cells. Such large and growing resources necessitate development of efficient analysis pipelines to exploit the wealth. Two mainstream well-recognized packages, Seurat[3] and SCANPY[4] (single-cell analysis in Python), offer diverse features covering various aspects of common single-cell RNA-seq (scRNA) analysis tasks. Starting from a quality controlled gene-count matrix, both methods follow similar workflows, including conducting normalization for each cell across genes and standardizing expression values for each gene across cells, followed by highly variable genes (HVGs) selection, optionally batch-effect correction, then performing dimensionality reduction on HVGs using principal component analysis (PCA)[10], constructing k-nearest neighbor (k-NN) graph on the PCA space, applying community detection algorithms to find clusters and annotate clusters with putative cell type labels. Several benchmark studies have been performed [6,7,8].

ScRNA-seq datasets contains abundant zero read counts, due to either technical drop-out events or biological effects. Current wide-recognized packages, including Seurat, SCANPY and Pegasus[5], are often not scalable to deal with entire gene count matrix and rely on feature selection to tackle with inherent noise and drop-out effects. The statistical distribution hypothesis of scRNA-seq data is still under debate. Several methods assume zero-inflated negative binomial distribution[9]. However, Svensson has reported that droplet-based scRNA-seq data is not zero-inflated. HVGs are selected based on estimation of relationship between candidate gene's mean and variance. These operations might bear the risk of information loss. We argue that HVGs selection might not be necessary, and it is of great interest to develop a scalable approach to operate on all genes to keep original signals as much as possible. Following HVG selection, typical pipelines perform PCA-like dimensionality reduction to identify the directions of greatest structured variation between cells. PCA is a linear technique, whose underlying assumption of modelling continuous multivariate Gaussian distributions is conflicted with scRNA-seq readouts. In addition, some principal components might represent unwanted source of variation related to technical noise or random fluctuations. To represent inherent non-linear structure of scRNA-seq datasets, several deep learning approaches, especially autoencoders, emerged recently to learn low-dimensional latent embeddings aiming for stronger explanatory power. Especially, variational autoencoder has shown decent performance through coupling probabilistic representation learning with downstream analysis (scVI)[11]. This series of methods mainly consist of an encoder structure and a reconstruction function parameterized by neural networks. We argue that the reconstruction part may not be necessary to learn high-level semantic representations. Considering the over-abundance of zeros, forcing the model to reconstruct those ambiguous null values without distinguishing biological or technical effects is questionable.

Contrastive learning has recently achieved great success in computer vision domain, such as SimCLR[12] and MoCo[14]. This type of method defines a pretext tasks and conducts self-supervised learning for large-scale unlabeled data via minimizing contrastive loss between different augmented views to obtain representations[13]. Classification result are then derived through supervised fine-tuning on few (10%) labelled samples from ImageNet datasets, outperforming standard supervised methods trained with all labels[15]. Inspired from this emerging paradigm, we anticipate that each cell can firstly learn from itself to obtain high-level features and then be assigned to certain predefined cell subtypes through fine-tuning. On the other hand, unsupervised clustering has been routinely conducted in scRNA-seq studies. Some popular clustering methods, such as K-means, spectral clustering, or community-based methods (Louvain or Leiden)[22] have been widely used to discover heterogeneous patterns from scratch. Then cell profiles are visualized by either t-distributed stochastic neighbor embedding (t-SNE)[16,17] or uniform manifold approximation and projection (UMAP)[18,19]. Motivated by deep embedding clustering (DEC)[20], scDeepCluster[21] conducts feature learning with clustering simultaneously and shows promising accuracy but instability to model initialization. We anticipate that contrastively learned embeddings can be also extended for clustering task. By this setting, feature learning and clustering are disentangled.

Besides, integrating scRNA-seq datasets of multiple batches across technologies, conditions and donors is of great importance to conduct integrative analysis or construct harmonized reference atlas. Several integration tools have been developed to disentangle biological signals from confounding effects. Seurat V3[25] applies canonical correlation analysis (CCA) to project cells into shared

correlation component space and performs data integration via mutual nearest neighbors (MNN)[23,24] defined as 'anchor cell pairs' between batches. MNN-based approach only allows integrating two batches at a time. Increasing number of cells will require exponentially boosted memory consumption when searching plausible anchor pairs, leading to poor scalability to process million-scale datasets. Harmony[26] iteratively uses fuzzy clustering and linear correction method until reaching stable assignment of cell clusters. Deep learning-based trVAE[27] leverages conditional VAE in the encoder module to correct batch effects. Ideal data integration method should scale to large datasets, enable integrating multiple batches simultaneously without over-correcting non-overlapping populations. The instance discrimination nature of contrastive learning has the potential to learn batch-invariant cell embeddings without compromising biological signal.

Finally, with the emergence multimillion-cell atlases, mapping query cells against reference is of high potential to enable fast interpretation of new single-cell studies and conducting comparative analysis, which eliminates the needs of *de novo* clustering or laborious manual annotations. In this scenario, it is unnecessary to load complete reference datasets or reimplement the full integration pipelines which will otherwise incur large computation burden and subject to legal restrictions of sharing private data. Alternatively, an ideal approach only utilizes the hidden knowledge in compressed representations from reference without sharing raw data. MARS[28] uses a meta-learning approach to learn cell landmarks as well as joint embedding of annotated and unannotated cells, enabling knowledge transfer across heterogenous tissues. However, MARS lacks domain adaptation module and thus requires batch-effect correction beforehand. Seurat V4[29] uses supervised principal component analysis (sPCA) to project a transformation onto the query and identify MNN-based anchors, and then transfers reference labels to query cells based on weighted anchors' voting. Symphony[30] uses mixture modelling framework to localize query cells within a stable reference embedding and transfer reference annotations. ScArches[31] explicitly models categorical batch labels using conditional autoencoder and maps query onto reference through fine-tuning. This mapping task is different from rigid annotation. Existing cell type labels might not fully define cell identity, oversimplifying the rich biological content within a cell to a single rigid concept from predefined categories, which might be updated over time. We consider query-to-reference mapping task as an unsupervised transfer representation learning problem, and query cells can be localized via direct inference or unsupervised fine-tuning. The obtained query embeddings can be used to derive voting-based annotation, infer cell trajectory, or complement values of missing modalities.

To address all abovementioned challenges, we propose Concerto, a novel unified framework for single-cell analysis. Concerto defines a contrastive instance discrimination pre-text task and learns task-agnostic cell representations by maximizing agreement between each cell's two augmented views in the latent space. Concerto leverages a novel dropout layer operation to generate minimal augmentations, which is well suitable for single-cell data format. Through comprehensive benchmark studies on real datasets, the learned embeddings can be fine-tuned to satisfy various downstream needs, covering automatic cell type classification, clustering, data integration with batch-effect correction and query-to-reference mapping. Concerto can flexibly handle multi-omics datasets simply through element-wise summation of each modality. As for each specific task, Concerto achieves state-of-the art performance against other published top-performing methods. Besides, we demonstrate Concerto can retain genuine biological signal via extensive bioinformatics

analysis for several specially designed cases, including fine-grained immune cell classification, cross-tissue cell type discovery, avoidance of over-correction, mapping unseen cell type against reference and inference of missing modality. Concerto can operate on all genes, and we show that using HVGs might lose certain biological nuance. Furthermore, we leverage Concerto to query a COVID-19 immune cell dataset against a healthy reference and recapitulate several differential immune features among patients with diverse disease status. In brief, Concerto is a robust, accurate, scalable representation learning framework for single-cell multi-omics analysis of ten-million scale.

Results

Overview of Concerto architecture

Concerto leverages a self-distillation contrastive learning framework, which is configured as an asymmetric teacher-student architecture. Teacher module aggregates distributional gene embeddings with attention mechanism[40] followed by several non-linear fully connected layers to obtain final cell embeddings, while the student module accepts discrete gene counts as input with dense structure. This asymmetric configuration injects imbalanced complexity presented as teacher and student. For model input, normalized gene count matrix will be transformed into index-value format, i.e., index refers to gene ID defined by a dictionary consisted of all genes of a certain species, and value refers to relevant count number within a cell. This encoding scheme is introduced to improve computational efficiency to tackle with spare high-dimensional input. The teacher module then scales each gene's embedding by its count value as input for subsequent operations. Through defining a pre-text task of predicting itself for each unlabeled cell, Concerto learns effective representations by maximizing agreement between each cell's different views using a contrastive loss in the latent space. Two augmented embeddings for the same cell are obtained via passing the same cell through student and teacher module with random dropout mask right before the output layer. The intuition behind this operation is that single cell data format is inherently discrete rather than continuously organized image pixels of computer vision domain. Common perturbation techniques, which explicitly exerts transformations or injects randomness to data input, might introduce unfavorable negative noise with risk of altering the intrinsic semantics. We simply use asymmetric teacher-student network and sample independent dropout[44] masks before the final layer to obtain paired semantic-invariant embeddings for each cell. Dropout mask can be regarded as a minimal data augmentation because the input never changed. Projected onto the unit hypersphere space[13], the contrastive loss explicitly compares pairs of cell embeddings to push apart different cells within a batch while pulling together teacher-student views of the same cell as positive pairs. The distance is measured by cosine similarity of L2-normalized embeddings using dot product operation. To process multi-omics data, simply element-wise summation of modal-specific attention output in teacher module or dense output in student module enables Concerto to generate unified cell embeddings. The contrastively learned embeddings are expected to capture high-level features to discriminate different cells, which can be fine-tuned in different ways for downstream analysis.

1) For automatic cell type identification, Concerto leverages contrastive learning as task-agnostic pre-training procedure followed by supervised fine-tuning using existing annotations. Pre-training on large unlabeled data and fine-tuning over a few labels has become a predominant

paradigm in natural language processing. For within-datasets prediction (intra-dataset), we fine-tune Concerto via an extra fully-connected classification layer with SoftMax operation over dimension of pre-defined categories. For cross-datasets prediction (inter-dataset), we conduct semi-supervised fine-tuning via adding a domain-adaptation module for cross-tissue prediction.

2) To group functionally similar cells into clusters, Concerto decouples cell representation learning and clustering into two stages, which is expected to be less sensitive to model initialization in contrast to other deep-learning based clustering approaches.
3) For *de novo* data integration, Concerto aims to learn batch-invariant embeddings, such as various tissues, species, technology platforms, experimental conditions, or sample status. These meta-data will be incorporated as model input to guide Concerto implementing source-specific Batch Normalization[43] within a training mini-batch. This simple configuration enables Concerto to extract batch-invariant biological signal and project all cells into a generalized embedding space after removal of unwanted confounding factors.
4) For query-to-reference mapping, it is assumed that a reference atlas has been constructed via above operations. Query cell embeddings are simply inferred via passing through the trained teacher network. In this case, users directly utilize the model weights and contextualize query cells onto a stable reference embedding space. Reference annotations can be easily transferred to query cells through a NN-voting scheme, which enables fast interpretation of query cells. It is noted that this operation is distinct from supervised rigid annotation as in part 1), i.e., the reference annotation label is never used in the training process. On the other hand, users can also leverage reference weights as initialization and implement unsupervised fine-tuning in a contrastive manner on query cells, which is equivalent to joint analysis. Concerto can be continuously updated in this convenient way to support constructing ever-expanded atlas.

Contrastively learned embeddings significantly boost performance of automatic cell annotations via fine-tuning and supports novel cell type discovery across tissues

To demonstrate the utility of cell embeddings learned by contrastive pre-training, existing annotations are used as training labels to implement supervised fine-tuning for Concerto. We use classical human peripheral blood mononuclear cells dataset composed of 31,021 cells (PBMC 45k)[37] derived from 7 different sequencing protocols applied on 2 samples to conduct benchmark study against several top-performing classifiers representing diverse roadmaps, including likelihood-based multinomial model inference, SciBet[33]; neural network-based generative method, Cell BLAST[34], iterative correlation to reference, SingleR[35]; support vector machine (SVM) with linear kernel, Moana[36]; and a meta-learning approach, MARS[28]. Concerto leverages a two-step approach with pre-training and fine-tuning, while all other methods are developed by end-to-end supervised training. We also evaluate the performance of the end-to-end version of Concerto (Concerto-E2E) via discarding the contrastive loss and training Concerto in a fully supervised manner. For intra-dataset evaluation, we apply 5-fold cross-validation within each batch from PBMC 45k (n=9). Training and testing folds are kept the same to make head-to-head evaluation across each classifier. F1-score is computed for each cell type and median F1-score across all cell types is used as evaluation metric. The result shows that Concerto achieves the highest median F1-score of 0.926 with the most stable performance across each fold (Figure 1a). It is noted that Concerto-E2E obtains lower score (=0.867), demonstrating the advantage of two-step training

paradigm. The performance of Concerto-E2E is comparable with singleR (=0.889) and SciBet (=0.881), largely surpassing SVM-based Moana (=0.688), which validates the utility of aggregating contextualized gene embeddings into cell representations as a stronger feature extractor (detailed heatmap results for each fold in Supplementary Figure 1). For inter-dataset assessment, which stands for a more realistic scenario, we intentionally leave cells from one sequencing protocol out as test set and use cells from other protocols as training set sampled by five times of bootstrapping. Concerto significantly outperforms all other methods in almost all train-test combinations (Figure 2b). When Seq-well dataset is held-out, all methods report declined performance. It is probably because Seq-well leverages a microwell-based library construction protocol different from other droplet-based methods, leading to more difficulties of model transfer when facing significant data distribution shift (overall comparison with detailed heatmap results for each fold in Supplementary Figure 2). To assess model's ability to tackle with more complex datasets with fine-grained annotations, we further mix Thymus scRNA-seq atlas[38] (thymus, n=107,969 cells) with PBMC 45k to construct a multi-hierarchical immune cell dataset. Incorporating high-resolution thymus datasets poses a greater challenge for classifiers to distinguish subtle state discrepancy along the trajectory of T-cell development. Concerto reaches the highest mean accuracy (ACC) of 0.939 (5-fold cross-validation), substantially outperforming singleR (mean ACC=0.718) and SciBet (mean ACC= 0.690) (Figure 2c). Concerto can successfully discriminate different developmental stages of double negative T-cells (DN), double positive T-cells (DP) and single positive T cells, even though these cells lie close along developmental trajectory with subtle transcripts discrepancy. We also use *Tabula Muris Senis* (TMS) cell atlas (n=148,116 cells from 23 heterogeneous mouse tissues) to train a tissue-wise classifier (intra-tissue prediction, five-fold cross-validation). Concerto exceeds SciBet in prediction accuracy for all tissues by a large margin (Figure 2f), achieving top mean ACC of each fold close to 1 for Bladder (=0.999), Brain Myeloid (=0.999) and Mammary Gland (=0.996). The largest absolute gain over SciBet measured by mean ACC is Tongue (+7.85%), Large Intestine (+7.69%), Brown Adipose Tissue (BAT) (+7.26%). (Detailed results in Supplementary Figure 3)

Next, we evaluate Concerto can support marking none-of-the-above (NOTA) cells as a rejection option if the test set contains certain cell subpopulations not existed in training samples. These cells cannot be accurately predicted and should be assigned as 'NOTA' when the classifier is not confident enough to annotate it with predefined labels. We download a multimodal PBMC CITE-seq atlas of 161,764 cells (PBMC 160k) with three levels of annotations. In this rejection study, only RNA counts are used as features. For different levels, we remove different granularities of T cells from the training set to form progressively increased difficulties. First, All T cells are removed, then only CD4 T cells are removed, then only CD4 Mem T cells are removed. 20% of training set is randomly selected as validation set. The test set only contains removed cell types at each level (Detailed mask setting in Supplementary Table 3). A qualified classifier should predict accurate labels for cells in the validation set while assign NOTA to cells of test set. As shown in Figure 2-d, Concerto can clearly separate confidence curve of validation and test set for level-1 and level-2 masking setting. Even for the hardest level-3 scenario, Concerto obtains a bimodal curve for test data with partial overlap with validation curve, while competing method almost misassigns unseen CD4 Mem T cells as other types (Detailed results of Concerto and SciBet seen in Supplementary Figure 4-6). To further demonstrate Concerto can make cross-tissue annotations, we benchmark Concerto against a meta-learning method, MARS, on TMS dataset. We follow the experimental design by leaving one

tissue out as unannotated, fine-tune Concerto on all other tissues and evaluate performance on test tissue. Concerto achieves improvement in adjusted Rand index (ARI) over MARS on 22 held-out tissues except Brain Myeloid precluded by MARS, ranging from the largest absolute gain for Spleen (+89.4% mean ARI) to smallest absolute gain for Bladder (+0.613% mean ARI) (Figure 2-g) (bootstrapping 3 times). The held-out tissue often contains specific cell types never seen in training tissues. Like MARS, Concerto also well support knowledge transfer to discover novel cell types across tissue. In particular, when using Limb Muscle as test tissue, Concerto obtains cell embeddings and visualize them by UMAP using Limb Muscle as an example (Figure 2-h). Concerto effectively places functional similar cell types from all other tissues closer to the true annotations of Limb Muscle, including six major cell types, i.e., B cell, T cell, Endothelial Cell, Macrophage, Mesenchymal Stem Cell and Skeletal Muscle Satellite Cell. Further illustrated by Sankey plot (Figure 2-i), Concerto successfully transfer annotations of Skeletal Muscle Satellite cells from other muscles to satellite cells from Limb Muscle, while MARS erroneously uses T cells to annotate satellite cells. Concerto also relates Mesenchymal stem cells from other adipose tissues with their counterparts from Limb Muscle. Annotations of General B cell, T cell, Endothelial Cell, Macrophages from other tissues are correctly transferred to Limb Muscle. This experiment demonstrates the superior transfer capability of Concerto empowered by an unsupervised domain adaptation module (Methods). Sankey plots for Spleen and Brain Non-myeloid are shown in Supplementary Figure 7. We further design a simulation study to benchmark the robustness of competing methods. Specifically, we use R package Splatter[39] to simulate scRNA-seq data to mimic various biological scenarios under different dropout count rates (defined as the proportion of expressed genes being knocked out) and different expression signal strength (defined as varied fold change levels of differential genes). The result shows Concerto can maintain the highest ACC value when decreasing intensities of differential expression at fixed dropout rate or increasing dropout rate at fixed expression variance (Supplementary Figure 8). To assess Concerto's competency to process multi-omics data, we use PBMC 160k[29] to train a multi-modal annotation model in three different scenarios, each with RNA, protein, or both as feature input, respectively. Concerto achieves mean ACC of 0.882, 0.857 and 0.890, correspondingly (5-fold cross validation, intra-dataset prediction), demonstrating that unifying multimodal features enables better cell annotations than unimodal input (Figure 2-e). Compared to Azimuth, a recently published multimodal annotation tool trained on the same dataset, Concerto outperforms Azimuth in all three scenarios, obtaining absolute improvement of 3.9% when using both RNA and protein as input. The non-linear nature of Concerto gives stronger explanatory power over linear methods.

Concerto can automatically extract specific molecular signatures from attention weights to define cell identity at single-cell resolution without the need of discrete clustering

The utility of discrete clustering approach is limited to certain level of resolution, which simplifies cell heterogeneity into partitioned classes and risks ignoring nuanced signal from smooth transitions between cell states. Cell-ID[42] reported a statistical method to extract per-cell gene signatures in a clustering-free manner. Unsupervised clustering often serves as a starting point for a novel study. We start by assessing whether Concerto embeddings match well with biological coarse-grained assignment and then demonstrate how Concerto can extract biologically meaningful signatures at single-cell resolution without clustering. To assess the quality of cell embeddings for de novo

clustering, we choose a subset from PBMC45k (n=11,377 cells, 10X-v2, v3) with minor batch-effects and leave cross-batch integration to be discussed in the following sections. Classical Seurat package recommends share nearest neighbor (SNN) graph construction followed by Louvain or Leiden clustering. We also include scDeepCluster as a representative of simultaneously learning representations and clustering. 2000 HVGs are used for all methods. We compute common clustering performance metrics at different resolutions to make thorough comparison, including normalized mutual information (NMI), adjusted Rand index (ARI) and silhouette score. Concerto embeddings with Leiden clustering (Concerto + Leiden) dramatically exceeds all other methods (set 5 different resolutions, resolution = 0.1,0.2,0.3,0.4,0.6 for Louvain or Leiden, k=7,8,9,10,11 for scDeepCluster) (mean NMI=0.750, ARI=0.646, silhouette score=0.332, Concerto + Leiden) (Figure 3-a). Better quality of cell embeddings leads to better aligned clustering. To further illustrate the representation effectiveness, we apply UMAP to visualize embedded cells in the 2D space to make comparative observation between clustering assignments with ground truth annotations (Figure 3-b, resolution=0.4 for Leiden, k=9 for scDeepCluster). Concerto clearly separates myeloid cells, including CD14 Monocyte (CD14 Mono), CD 16 Monocyte (CD16 Mono) and Dendritic Cell (DC) and draws clear boundary between CD4 T cell and Cytotoxic T cell. In contrast, other competing methods more or less mingle CD4 T cell with Cytotoxic T cell and include intermixed populations among CD14 Mono, CD16 Mono and DC. All results are shown in Supplementary Figure 9-11.

Cell type labels of above PBMC45k dataset are defined by differential RNA expression[41] followed by clustering, which might not be the real ground truth. Multimodal definition of cellular identity is expected to incorporate integrative signal beyond transcriptome[29]. To further demonstrate Concerto can effectively learn cell embeddings with intrinsic biological meaning. We revisit PBMC 160k CITE-seq dataset and implement Concerto on all cells with only RNA, only protein or both as model input. The learned embeddings are then visualized by UMAP and evaluated by original hierarchical annotations by the author (Figure 3-c). We use Level-1 categories for unimodality plot and Level-2 categories for dual-modality plot. CD4 T and CD8T can be well separated by protein alone but partially mixed by RNA alone. NK cells are partially mixed with CD8 T cell by RNA alone while lie between Other T cell and CD8 T cell by protein alone. DC are mingled within Monocytes by protein alone while forming a clear boundary by RNA alone. All these signals recapitulate Seurat V4's conclusions that protein is more informative than RNA to discriminate CD4 T cell with CD8 T cell and uncover subtle heterogeneity within NK cells. On the other hand, relationship of Mono-DC lineage can be better delineated by RNA than classical immunophenotype surface protein markers. For the bimodal plot, Concerto can display a smooth directional developmental manifold (shown as light blue arrow in Figure 3-c) for subpopulations within CD4 T cell lineage (CD4 Naïve, CD4 TCM and CD4 TEM), CD8 T cell lineage (CD8 Naïve, CD8 TCM and CD8 TEM) and B cell lineage (B naïve, B intermediate, B memory and Plasmablast). In addition, CD4+ regulatory T cells (Treg), MAIT cells and subpopulations of γδ T cells (gdT) can be identified via leveraging both modalities jointly. Concerto can tackle any number of expanded modalities simply by element-wise summation before the Batch Normalization Layer, which render it as a very concise learning-based multi-modal integration protocol compared against weighted nearest neighbor (WNN)-based discriminative approach from Seurat V4, which includes tailored processing step for different modalities. Concerto can facilitate synthesizing multi-modal information and preserving genuine biological signal towards a unified view of a cell at finer resolution.

Concerto introduces the cellular context vector to aggregate gene embeddings, and we further check whether Concerto's attention weights[40] can offer interpretability via reproducing molecular signatures previously established for well-known cell types. The heatmap (Figure 3-d) shows the relative normalized attention contributions of characteristic features to define cell identity, which successfully recovers canonical modal-specific markers for each cell type (Methods). CD4 T cell discriminates itself from CD8 T cell appeared as obvious divergence of attention pattern for CD4 and CD8 protein marker while the corresponding CD4 or CD8 RNA marker is not significant, which further recapitulates the particular utility of protein data to defining these T cells. For B cells, CD19 emerges as a key protein marker while MS4A1 (RNA form of CD20 protein) as RNA marker. Concerto extracts CD16 as protein marker for activated NK cells with significant cytotoxic transcripts (GZMB, GNLY). By this means, modal-specific signatures are automatically extracted at single-cell resolution and match well with biological implications, though all these signals are obtained without any manual labels or discrete clustering operation. The only learning signal is from each cell itself. We propose this novel attention-based self-supervised marker identification approach, offering intrinsic interpretability to define cell identity at single-cell resolution and potentially revolutionizing current practice of differential test followed by clustering.

Concerto enables *de novo* data integration via removing unwanted batch-effects and well supports integrating partially-overlapping datasets

Facing the need of integrating different data sources into reference atlas, batch effects should be removed. We first assess data integration performance of Concerto on well-curated multi-donor Human Pancreatic Islet (HP) dataset (8 batches of 5 technologies, n=14,890 cells)[45,46,47,48,49] against other top-performing methods, including Seurat V3, Harmony, trVAE, and naïve PCA with no correction module as a baseline. We quantify the batch mixing performance using k-nearest neighbor batch-effect Test (kBET)[50] and evaluate conservation of cell type purity using average silhouette width (ASW)[51], all calculated in the same 128-dimension embedding space. Harmony and Seurat V3 operate on principal components while trVAE directly uses fully-connected layer to compress the input. Concerto's encoding scheme can easily scale to all genes as input and its attention operation within teacher module avoids direct autoencoder-based dimension reduction. We use the top 2000, 5000, 10000, 15000, 20000 most variable genes and all genes to compose six scenarios to benchmark all methods. UMAP visualization of 2000 HVGs and all genes for 5 methods are shown in Figure 3-e colored by batch and cell type labels (other 4 HVG scenarios in Supplementary Figure 12). Compared with naïve-PCA, which results in nearly no integration effect where cell subgroups are separated purely by batch sources, all integration methods can mix biological subpopulations from different sources together to diverse extent. As measured by ASW, Concerto outperforms all other competing methods at all 6 levels of input HVGs by a large margin (ASW: 0.533 for 2000 HVGs; ASW: 0.305 for all genes) (Figure 3-g). Higher ASW indicates larger distances between cell subpopulations and lower distance within cell types. All methods present decreasing trend when accepting more genes as model input. The underlying reason might come from the labels used to calculate ASW, whose annotations might not be the real ground truth but derive from PCA of 2000 HVGs followed by clustering and manual inspection of cluster-specific marker genes, more resembling the protocols used by Seurat and Harmony. Cluster labels annotated via 2000 HVGs might lose nuanced biological variations, especially for high-resolution

heterogenous subtypes or perturbed substates. Although Concerto obtains the lowest kBET score (=0.10), Concerto successfully achieves data integration across 8 sources evaluated visually or numerically with higher ASW. We argue that kBET has certain sweet point threshold, above which no further mixing is necessary as long as biological signal converges together rather than confounded by batch effects. Over-pursuing larger kBET might aggressively misalign distinct cell populations together, in other words, over-correction. To further validate this hypothesis, we design a more complex scenario of integrating partial overlapping datasets by manually removing beta cells from all 5 technologies except for Cel-Seq2. UMAP visualization shows Harmony and Seurat mix up beta cells (colored in red) with other cell types to some extent, implying over-correction with lower beta-cell ASW though larger kBET (kBET: 0.32, Beta-ASW: 0.29 for Harmony; kBET: 0.39, Beta-ASW: 0.05 for Seurat V3). Concerto clearly separates beta cells from others, obtaining larger beta-ASW (Beta-ASW: 0.34 for Concerto), although confers comparably lower kBET (kBET: 0.03 for Concerto). In this task, Concerto exerts minimal acceptable batch-correction and strikes a just suitable balance between ASW and kBET scores. Concerto's contrastive learning objective is immune to risk of over-correction and enables preserving biological variation without merging non-matching subpopulations, which is suitable for building a high-resolution reference atlas.

Concerto achieves state-of-the-art accuracy for query-to-reference mapping and supports projecting unseen cell types in the reference

After building the reference, we further evaluate Concerto's utility of mapping query cells onto harmonized reference embeddings. This task is different from rigid annotation described above, while we first calculate query embeddings according to the pre-trained weights by reference and locate query cell near its most similar reference cells. If annotation is necessary, a simple k-NN (k =5 for most cases) voting classifier can be used to transfer reference annotations to queries. We design two experiments, one is cross-technology transfer with inDrop Baron dataset (n= 8,569 cells) as query while all other 4 technologies (n=6,321 cells) as reference (HP→inDrop), and another one is cross-species transfer using the same reference but Mouse Pancreatic Islet (MP) (n=1,880 cells) as query (HP→MP). We compare against another top three query-to-reference methods, including scArches, Symphony and Seurat V4, each corresponds to its reference building method, i.e., trVAE, Harmony and Seurat V3, respectively. 2000 HVGs are used as input for fair comparison. Concerto achieves the highest performance among all methods measured by mean overall ACC (repeated 5 times) for both scenarios (0.981, HP→inDrop; 0.927, HP→MP) (Figure 4-a). As shown in the confusion matrix (Figure 4-b), Concerto is able to transfer labels annotated by different technologies and map analogous cell types from human to mouse with high accuracy for the majority of cases.

As described above, HVGs filtration might risk unrecognizing nuanced biological state. Concerto can leverage information from all genes by automatically extracting molecular signatures which discriminate each cell from others. We design a more challenging scenario using multimodal PBMC 160k dataset by assigning cells from one sample (labelled as 'P3') as query and building a reference from other 7 samples after intentionally removing all CD8 T cells within. We aim to test Concerto's ability to project unseen cell types (CD8 T cell) and evaluate whether incorporating all genes will bring benefits. Concerto operating on all genes obtains significantly higher ACC (0.988 for all gene, 0.772 for 2000 HVGs). UMAP visualization in Figure 4-c shows Concerto operating on all genes

can precisely localize all query cells onto corresponding reference coordinates, including B, DC, Mono, NK and CD4 T cells, and particularly place CD8 T cells at positions between NK cells and CD4 T cells even though Concerto has never directly utilized the supervised signal from CD8 T cell. The heatmap shows that protein marker of CD8a is enriched at CD8 T cell positions assigned by Concerto. Operating on all genes is expected to better identify fine-grained subtypes or heterogenous states. Enrichment region of cytotoxic RNA marker (GZMA) is closer to NK cells while naïve/memory-like marker (CCR7) locates further away from NK cell, which is consistent with CD8 T cell's developmental direction. We further evaluate whether smooth transcriptional gradients of canonical marker genes correlate well with relative distance between various CD8 T subtypes with NK cells. In 'all genes' scenario, CD8 Naïve cells, which are further away from NK cells show lower cytotoxic signatures (measured by expression value of GNLY and NKG7) than proliferating and effector cells whose locations are closer to NK cells (Figure 4-d). GNLY and NKG7 obtain negative Pearson correlation coefficient of -0.327 (p-value=$1.01*10^{-41}$) and -0.639 (p-value=$9.13*10^{-187}$), respectively. Expression of naïve signatures (CCR7) of CD8 T cells shows positive correlation with distance away from NK cells (Pearson correlation coefficient = 0.253, p-value= $4.92*10^{-25}$). These signals demonstrate Concerto's ability to project unseen cell subtypes along a biological meaningful continuum even though the reference excludes those information. To further demonstrate Concerto can recapitulate genuine quantitative biological signal, we leverage 80% of samples from PBMC160k dataset to build a bimodal reference and use remaining 20% with only mRNA count as query. We expect a powerful mapping method can accurately infer missing modality value of query cells. The measured protein expression will be used as ground truth for validation. Concerto follows similar smoothing protocol as in Symphony and achieves consistent prediction result. (Top 20 prediction, Pearson r: 0.966-0.998, Figure 4-e). Inferred protein expression paired with ground truth of CLEC12A, CD155, and CD11c and CD14 are visualized by UMAP (Figure 4-h). Concerto can be used to infer unmeasured modalities of single-cell data, showing great potential to uncover missing signal towards a more holistic view of cell identity.

Concerto can efficiently scale to 10 million-cell atlas construction and reference mapping

To demonstrate Concerto's scalability to build large reference and implement efficient mapping, we design a study by simulating virtual reference datasets of 50k, 100k, 500k, 1 million, 2 million and 10 million cells. Then we map same size of query against corresponding reference. We measure total runtime for each scenario. Concerto's learning objective is naturally suitable for parallelized computing and only replies on mini-batch size within a training epoch. Hence, Concerto can be scalable to extra-large datasets through dividing the whole task into multiple processing batches to orchestrated GPUs. Via distributed training on 8 GPUs (NVIDIA Quadro RTX 6000), Concerto can build a 1 million-cell reference in 585 seconds (less than 10 minutes), 10 million-cell reference in 5,133 seconds (less than 1.5 hour) (Figure 4-g). Reference only needs to be built once and easily shared via model weights. Researchers can simply download pre-trained weights from Internet and make direct inference or update Concerto with in-house data through unsupervised fine-tuning. In this way, Concerto can be updated iteratively for continuous learning. For instance, mapping a million query against million reference takes 168 seconds (less than 3 minutes) (Figure 4-g). The peak memory usage is set to 6G per CPU (Intel Xeon Gold 6226R) and 2.5G per GPU (NVIDIA Quadro RTX 6000). These results show that Concerto can efficiently scale to building multi-million

cell reference, enabling rapid mapping within several minutes. In summary, Concerto achieves the best overall performance against another three well-recognized tools, i.e., nearest neighbor-based Seurat, soft clustering-based Harmony/Symphony and VAE-based scArches, evaluated by various qualitative and quantitative metrics (Figure 4-h). Concerto is most scalable, interpretable, does not require PCA or scaling, can operate on all genes and supports multi-modal integration.

Hierarchical query-to-reference mapping preserves differential immune response of COVID-19 patients

Massive single-cell studies have been conducted to investigate differential immune response of COVID-19 patients across different levels of disease severities. We design a study to exploit the PBMC 160k dataset (RNA only) to enable fast interpretation of scRNA-seq data collected from COVID-19 patients. Specifically, we initialize Concerto with weights trained by PBMC 160k and project a recently published PBMC scRNA-seq data from Jonas et al. (n= 99,049 cells, EGAS00001004571)[52] onto a joint latent space by unsupervised fine-tuning. It is noted that, pre-trained weights of Concerto can be easily shared without compromising data privacy. In addition, updating weights is much more convenient than sharing count matrix.

We propose a hierarchical mapping approach; firstly, mapping all query cells on top of reference to obtain coarse-grained Level-1 annotations using nearest neighbor information calculated from all reference cells, then projecting grouped query cells to corresponding reference subgroup to yield Level- 2 annotations. Original Jonas et al. paper conducted comprehensive analysis on myeloid cells, and we complement its analysis on lymphoid cells and recapitulate differential immune signals reported in another COVID-19 multi-omics study (Yapeng Su et al.)[53]. UMAP visualization shows that Concerto successfully interprets query cells in a clustering-free approach and obtains aligned geometric distribution of Level-1 subpopulations against the reference (Figure 5-a). Through Level-2 reference mapping, Concerto is able to identify perturbed pathological states. For all annotated CD8 T cells, Concerto discriminates divergent compositions of subtypes from healthy controls, mild and severe patients, dividing CD8 T cells into naïve, proliferating, memory, and effector states according to transferred Level-2 annotations (Figure 5-b). Particularly, through calculating exhaustion score (Methods), Concerto identifies emerging exhausted T cells from COVID-19 patients even though disease states are absent in the reference. Expression heatmap of canonical state-specific signatures is shown in Figure 5-b. Through differential expression analysis, naïve markers of annotated CD8 Naïve cells manifest up-regulation against other subtypes, such as CCR7, LEF1, TCF7, SELL (log_2Fold-Change (log_2FC) =0.89, 0.60, 0.45 and 0.35; adjusted p-value (adj p-value) =$2.8*10^{-224}$, $9.9*10^{-174}$, $3.3*10^{-8}$ and $2.15*10^{-84}$, respectively; Wilcoxon rank sum test, false discovery rate (FDR)-adjusted). Cells annotated as CD8 T effector memory cells (TEM) show up-regulated cytotoxic transcripts against others, such as PRF1(log_2FC=0.473, adj p-value =$4.1*10^{-18}$) and GNLY (log_2FC=0.451, adj p-value =$7.6*10^{-5}$). TYMS is significantly upregulated in CD8 proliferation T cells compared to CD8 naïve T cells (log_2FC=3.044, adj p-value =0.0012). Visualized by UMAP, naive-like markers (CCR7) are enriched in bottom region and the relative percentages of CD8 naïve T cells significantly reduce in patients than healthy controls (Figure 5-b, boxplot). Effector-like markers (GZMA) reside mainly at the top-right region indicating elevated cytotoxicity in patients. Exhausted CD8 T cells evolve in infected samples (Figure 5-b, boxplot)

with LAG3 appearing in the patient region. Yapeng Su et al reported a hybrid proliferative-exhausted CD8 T phenotype with upregulated exhaustion transcripts (LAG3), proliferative marker (MK167) and cytotoxic signature (GZMA) without fully losing naïve (CCR7) feature, which we validate its presence visualized by a cell subgroup co-expressing these markers. For NK cells, Concerto identifies the presence of proliferative NK cells in the patient region shown in Figure 5-c. The less-differentiated NK $CD56^{bright}$ subpopulation appears in regions overlapped with elevated CD27 expression, whose relative percentages significantly decrease in patients compared to healthy controls, reflecting NK cells' differentiation towards effector phenotypes at disease state (Figure 5-c). The $CD56^{dim}CD16^{bright}$ subpopulations are significantly activated in severe patients, validated in a flow cytometry study[54]. $CD56^{dim}CD16^{bright}$ NK cells displaying high levels of cytotoxic expression (PRF1 and GZMB) with increased exhaustion marker (LAG3) and terminal-differentiation marker (CD57) as well as an inflammatory interferon-inducible chemokine receptor marker (CXCR3) are enriched in severe patients (Figure 5-c). For monocytes, Concerto clearly separate healthy, moderate and severe ones (Figure 5-d). Non-classical monocytes ($CD14^{low}CD16^{high}$) are enriched in healthy samples while depleted in infected samples. Among the classical monocytes ($CD14^{high}CD16^{low}$), Concerto identifies dysfunctional HLA-$DR^{lo}S100A^{hi}$ CD14+ subtypes enriched in severe patients, verifying the original discovery from Jonas et al. This inflammatory phenotype with deficiency of antigen presentation is also reported in another study[55]. Overall, Concerto successfully separates disease from healthy state through hierarchical mapping without applying any clustering operation. The learned embedding space can both visually and biologically capture pathological cell states, preserve nuanced status-specific variation, and identify emergence of novel cell subtypes even though disease states are unseen in the reference.

Discussion

Inspired from the concept that 'each cell is different', we present Concerto, a novel contrastive learning framework to learn cell representations. High quality of cell embeddings is of vital importance for downstream analysis. Concerto's learning objective is to discriminate each cell from all others through aligning augmented views while promoting distributional uniformity of distinct cells in the latent space. Each cell learns from itself, and the obtained representations are expected to preserve genuine biological signal to the greatest extent. Contrastive setting help alleviate the interference of high-expressed genes to define cell identity, which might otherwise become dominant factors then distort cell embeddings and mask meaningful signal from other relatively low-expressed genes. Concerto obtains low-dimensional embedding distinct from discriminative dimension reduction like PCA or generative reconstruction like VAE. The quality of Concerto's embeddings is verified either through supervised fine-tuning, or unsupervised clustering or multi-modal integration, showing advantages over PCA or VAE-based embeddings. Another key contribution of Concerto is innovating a new perspective to augment omics data without changing the underlying semantics or biological meaning. When we wrote our manuscript, a symmetric MoCo-based method has been reported to learn cell representations evaluated by clustering performance[32]. They use a computer-vison like transformations and thus bear unneglectable risk of obtaining negative instances during data augmentations. However, Concerto introduces dropout layer to generate different views of the same input without altering original readouts. Concerto's augmentation is imposed on model instead of data, which stands for a minimal augmentation

strategy with promising broad applicability in the domain of omics data.

The asymmetric teacher-student configuration helps to co-distill knowledge from model itself, equipping Concerto with another superior advantage of interpretability power. Molecular features, either in protein or RNA modality, perform certain functions under cellular context. Concerto leverages the intrinsic characteristics of contextualized attention weights to automatically extract meaningful signatures without using any labels. The attention weights are evaluated at single-cell resolution, eliminating the needs of discrete clustering followed by differential test, which might limit the interpretation granularity and sacrifice the utility of single-cell readouts. Concerto's encoding scheme can efficiently process all genes, casting doubt on the needs of HVGs selection. We argue that manually removing features according to certain prior distribution hypothesis might oversimply signals. We preliminary demonstrate this argument via designing challenging tasks including integrating partially-overlapping datasets and projecting unseen cells against reference, showing that operating on all genes can lead to a more continuous manifold and thus uncover subtle biological signals. Concerto offers a very simple way to integrate multimodal measurement. Different levels of omics data might appear with different forms of diverse distributions. Concerto uses a straightforward element-wise 'add' operation and let the model learn relative importance of each modality automatically. We demonstrate this by running Concerto on a multi-modal PBMC atlas and conclude how each modality compensate another to determine cell identity.

For heterogeneous data-integration, Concerto simply introduces a source-aware training strategy without explicitly changing the model architecture, effectively overcoming batch-effects at a minimal acceptable level. Concerto is immune to over-correction which might otherwise distort real biological signal and merge unrelated cell types together. Query-to-reference mapping has become a new paradigm in single-cell study when more and more atlases accumulate. Concerto supports two modes of mapping, one is to make direct inference using trained weights of teacher module, and another is to conduct unsupervised fine-tuning if certain domain shift exists. In either case, only model weights will be shared instead of original data, which offers great advantages of protecting privacy and facilitating collaborative research. We further demonstrate the discovery utility of Concerto via projecting a COVID-19 PBMC dataset (Jonas et al.) onto a healthy atlas and recapitulate divergent responses of lymphoid cells reported in another multi-omics study as well as reproducing presence of dysregulated monocytes as in original study. All analysis is implemented at single-cell level, manifesting superior advantages of discovering biological nuance. We envision one major direction for further development. Concerto can be used in perturbation analysis to identify subtle state transition at single-cell resolution before and after stimulation and pinpointing most relevant molecular signatures simultaneously. We believe this paradigm is essential to quantify perturbation effects since traditional clustering method creates somewhat arbitrary boundaries irrelevant to divergent response among distinct physiological states.

In summary, we have shown that Concerto leverages self-distillation contrastive learning to enable interpretable representations for single-cell analysis, offering a brand-new perspective to conduct investigation at single-cell resolution. Concerto is easily scalable to 10-million reference building within 1.5 hours and query 50k cells within 8 seconds. With the emergence of atlases for more tissues and conditions, we expect Concerto to enable promising scientific or translational discovery.

Reference

1. Rozenblatt-Rosen, O., Stubbington, M. J., Regev, A., & Teichmann, S. A. (2017). The Human Cell Atlas: from vision to reality. *Nature News*, *550*(7677), 451.
2. Tabula Muris Consortium. (2020). A single cell transcriptomic atlas characterizes aging tissues in the mouse. *Nature*, *583*(7817), 590.
3. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., & Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nature biotechnology*, *33*(5), 495-502.
4. Wolf, F. A., Angerer, P., & Theis, F. J. (2018). SCANPY: large-scale single-cell gene expression data analysis. *Genome biology*, *19*(1), 1-5.
5. Li, B., Gould, J., Yang, Y., Sarkizova, S., Tabaka, M., Ashenberg, O., ... & Regev, A. (2020). Cumulus provides cloud-based data analysis for large-scale single-cell and single-nucleus RNA-seq. *Nature methods*, *17*(8), 793-798.
6. Kiselev, V. Y., Andrews, T. S., & Hemberg, M. (2019). Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews Genetics*, *20*(5), 273-282.
7. Tran, H. T. N., Ang, K. S., Chevrier, M., Zhang, X., Lee, N. Y. S., Goh, M., & Chen, J. (2020). A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome biology*, *21*(1), 1-32.
8. Abdelaal, T., Michielsen, L., Cats, D., Hoogduin, D., Mei, H., Reinders, M. J., & Mahfouz, A. (2019). A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome biology*, *20*(1), 1-19.
9. Svensson, V. (2020). Droplet scRNA-seq is not zero-inflated. *Nature Biotechnology*, *38*(2), 147-150.
10. Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, *2*(4), 433-459.
11. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., & Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. *Nature methods*, *15*(12), 1053-1058.
12. Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020, November). A simple framework for contrastive learning of visual representations. In *International conference on machine learning* (pp. 1597-1607). PMLR.
13. Wang, T., & Isola, P. (2020, November). Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning* (pp. 9929-9939). PMLR.
14. He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9729-9738).
15. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., & Hinton, G. (2020). Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*.
16. Kobak, D., & Berens, P. (2019). The art of using t-SNE for single-cell transcriptomics. *Nature communications*, *10*(1), 1-14.
17. Linderman, G. C., Rachh, M., Hoskins, J. G., Steinerberger, S., & Kluger, Y. (2019). Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. *Nature methods*, *16*(3), 243-245.
18. McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

19. Becht, E., McInnes, L., Healy, J., Dutertre, C. A., Kwok, I. W., Ng, L. G., ... & Newell, E. W. (2019). Dimensionality reduction for visualizing single-cell data using UMAP. *Nature biotechnology*, *37*(1), 38-44.
20. Xie, J., Girshick, R., & Farhadi, A. (2016, June). Unsupervised deep embedding for clustering analysis. In *International conference on machine learning* (pp. 478-487). PMLR.
21. Tian, T., Wan, J., Song, Q., & Wei, Z. (2019). Clustering single-cell RNA-seq data with a model-based deep learning approach. *Nature Machine Intelligence*, *1*(4), 191-198.
22. Traag, V. A., Waltman, L., & Van Eck, N. J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. *Scientific reports*, *9*(1), 1-12.
23. Haghverdi, L., Lun, A. T., Morgan, M. D., & Marioni, J. C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature biotechnology*, *36*(5), 421-427.
24. Butler, A., Hoffman, P., Smibert, P., Papalexi, E., & Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology*, *36*(5), 411-420.
25. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck III, W. M., ... & Satija, R. (2019). Comprehensive integration of single-cell data. *Cell*, *177*(7), 1888-1902.
26. Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., ... & Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature methods*, *16*(12), 1289-1296.
27. Lotfollahi, M., Naghipourfar, M., Theis, F. J., & Wolf, F. A. (2019). Conditional out-of-sample generation for unpaired data using trVAE. *arXiv preprint arXiv:1910.01791*.
28. Brbić, M., Zitnik, M., Wang, S., Pisco, A. O., Altman, R. B., Darmanis, S., & Leskovec, J. (2020). MARS: discovering novel cell types across heterogeneous single-cell experiments. *Nature methods*, *17*(12), 1200-1206.
29. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck III, W. M., Zheng, S., Butler, A., ... & Satija, R. (2021). Integrated analysis of multimodal single-cell data. *Cell*.
30. Kang, J. B., Nathan, A., Zhang, F., Millard, N., Rumker, L., Moody, D. B., ... & Raychaudhuri, S. (2021). Efficient and precise single-cell reference atlas mapping with Symphony. *bioRxiv*, 2020-11.
31. Lotfollahi, M., Naghipourfar, M., Luecken, M. D., Khajavi, M., Buttner, M., Avsec, Z., ... & Theis, F. J. (2020). Query to reference single-cell integration with transfer learning. *bioRxiv*.
32. Han, W., Cheng, Y., Chen, J., Zhong, H., Hu, Z., Chen, S., ... & Li, Y. (2021). Self-supervised contrastive learning for integrative single cell RNA-seq data analysis. *bioRxiv*.
33. Li, C., Liu, B., Kang, B., Liu, Z., Liu, Y., Chen, C., ... & Zhang, Z. (2020). SciBet as a portable and fast single cell type identifier. *Nature communications*, *11*(1), 1-8.
34. Cao, Z. J., Wei, L., Lu, S., Yang, D. C., & Gao, G. (2020). Searching large-scale scRNA-seq databases via unbiased cell embedding with Cell BLAST. *Nature communications*, *11*(1), 1-13.
35. Aran, D., Looney, A. P., Liu, L., Wu, E., Fong, V., Hsu, A., ... & Bhattacharya, M. (2019). Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature immunology*, *20*(2), 163-172.
36. Wagner, F., & Yanai, I. (2018). Moana: a robust and scalable cell type classification framework for single-cell RNA-Seq data. *BioRxiv*, 456129.

37. Ding, J., Adiconis, X., Simmons, S. K., Kowalczyk, M. S., Hession, C. C., Marjanovic, N. D., ... & Levin, J. Z. (2020). Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nature biotechnology*, *38*(6), 737-746.
38. Park, J. E., Botting, R. A., Conde, C. D., Popescu, D. M., Lavaert, M., Kunz, D. J., ... & Teichmann, S. A. (2020). A cell atlas of human thymic development defines T cell repertoire formation. *Science*, *367*(6480).
39. Zappia, L., Phipson, B., & Oshlack, A. (2017). Splatter: simulation of single-cell RNA sequencing data. *Genome biology*, *18*(1), 1-15.
40. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998-6008).
41. Soneson, C., & Robinson, M. D. (2018). Bias, robustness and scalability in single-cell differential expression analysis. *Nature methods*, *15*(4), 255-261.
42. Cortal, A., Martignetti, L., Six, E., & Rausell, A. (2021). Gene signature extraction and cell identity recognition at the single-cell level with Cell-ID. *Nature Biotechnology*, 1-8.
43. Ioffe, S., & Szegedy, C. (2015, June). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning* (pp. 448-456). PMLR.
44. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, *15*(1), 1929-1958.
45. Baron, M., Veres, A., Wolock, S. L., Faust, A. L., Gaujoux, R., Vetere, A., ... & Yanai, I. (2016). A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. *Cell systems*, *3*(4), 346-360.
46. Segerstolpe, Å., Palasantza, A., Eliasson, P., Andersson, E. M., Andréasson, A. C., Sun, X., ... & Sandberg, R. (2016). Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell metabolism*, *24*(4), 593-607.
47. Lawlor, N., George, J., Bolisetty, M., Kursawe, R., Sun, L., Sivakamasundari, V., ... & Stitzel, M. L. (2017). Single-cell transcriptomes identify human islet cell signatures and reveal cell-type–specific expression changes in type 2 diabetes. *Genome research*, *27*(2), 208-222.
48. Grün, D., Muraro, M. J., Boisset, J. C., Wiebrands, K., Lyubimova, A., Dharmadhikari, G., ... & van Oudenaarden, A. (2016). De novo prediction of stem cell identity using single-cell transcriptome data. *Cell stem cell*, *19*(2), 266-277.
49. Muraro, M. J., Dharmadhikari, G., Grün, D., Groen, N., Dielen, T., Jansen, E., ... & van Oudenaarden, A. (2016). A single-cell transcriptome atlas of the human pancreas. *Cell systems*, *3*(4), 385-394.
50. Büttner, M., Miao, Z., Wolf, F. A., Teichmann, S. A., & Theis, F. J. (2019). A test metric for assessing single-cell RNA-seq batch correction. *Nature methods*, *16*(1), 43-49.
51. Batool, F., & Hennig, C. (2021). Clustering with the average silhouette width. *Computational Statistics & Data Analysis*, *158*, 107190.
52. Schulte-Schrepping, J., Reusch, N., Paclik, D., Baßler, K., Schlickeiser, S., Zhang, B., ... & Ziebuhr, J. (2020). Severe COVID-19 is marked by a dysregulated myeloid cell compartment. *Cell*, *182*(6), 1419-1440.
53. Su, Y., Chen, D., Yuan, D., Lausted, C., Choi, J., Dai, C. L., ... & Heath, J. R. (2020). Multi-omics resolves a sharp disease-state shift between mild and moderate COVID-19. *Cell*, *183*(6), 1479-1495.

54. Maucourant, C., Filipovic, I., Ponzetta, A., Aleman, S., Cornillet, M., Hertwig, L., ... & Björkström, N. K. (2020). Natural killer cell immunotypes related to COVID-19 disease severity. *Science immunology*, *5*(50).
55. Ren, X., Wen, W., Fan, X., Hou, W., Su, B., Cai, P., ... & Zhang, Z. (2021). COVID-19 immune features revealed by a large-scale single-cell transcriptome atlas. *Cell*, *184*(7), 1895-1913.

Figures

Figure 1 Overview of Concerto architecture. (a) Asymmetric teacher-student configuration to generate different views of the same cell via self-distillation. Teacher network accepts distributional embeddings while student network uses discreate input. Attention weights under cellular context are used to extract molecular signatures automatically. Dropout layer with random mask is used to implement model-level augmentations. **(b)** Concerto implements contrastive learning via pulling together positive pairs while pushing apart different cells in a unit hypersphere space. The learned embeddings can be adapted to various downstream tasks (rigid annotation, clustering, data-integration, and query-to-reference mapping). **(c-f)** Illustration of workflow for 4 downstream tasks.

Figure 2 Contrastively learned embeddings significantly boost performance of automatic cell annotations via fine-tuning and supports novel cell type discovery across tissues. (a-b) Comparing performance of **(a)** intra-dataset (5-fold cross-validation for each of 9 batches) and **(b)** inter-dataset (Bootstrapping 5 times) **(b)** prediction measured by median F1-socre of cell labels on PBMC45k scRNA-seq dataset (n= 31,021 cells) against MARS, Cell BLAST, Moana, SingleR and SciBet. Boxplot shows medium, upper, lower quintiles with highest and lowest values. Concerto-E2E stands for end-to-end supervised training using the same model architecture without contrastive loss. Inter-dataset heatmap, minimum F1-score (red) to maximum (blue). **(c)** Performance benchmark measured by accuracy (ACC) on a challenging dataset composed of thymus scRNA-seq data (n= 107,969 cells) and PBMC45k. (5-fold cross-validation) **(d)** Rejection option study to test the ability to assign low confidence score for non-existed cell labels in training set on PBMC160k dataset (RNA only) (n=161,764 cells) against SciBet. (5-fold cross-validation) **(e)** Performance comparison for multi-modal annotation on PBMC160k dataset (RNA, Protein and RNA + Protein, respectively) against Azimuth. (5-fold cross-validation) **(f)** Benchmarking intra-tissue prediction accuracy on TMS dataset (n= 148,116 cells). **(g)** Benchmarking cross-tissue prediction by adjusted Rand index (ARI) against MARS. (Bootstrapping 3 times) **(h)** UMAP visualization of true label versus Concerto prediction for held-out Limb Muscle tissue. **(i)** Sankey plot to illustrate model utility of identifying relevant cells across tissues (Limb Muscle tissue as an example).

Figure 3 Concerto can automatically extract specific molecular signatures from attention weights and enables de novo data integration. (a) Benchmarking Concerto embedding's clustering performance against PCA, Seurat, scDeepCluster measured by Normalization mutual information (NMI), ARI and Silhouette score on PBMC45k dataset (10X-v2, v3 only, n=11,377 cells). Clustering methods include Louvain and Leiden for Concerto, PCA and Seurat. (Set 5 different resolutions, resolution = 0.1,0.2,0.3,0.4,0.6 for Louvain or Leiden, k=7,8,9,10,11 for scDeepCluster) **(b)** UMAP visualization of truth cell type labels versus cluster assignment (resolution=0.4 for Leiden, k=9 for scDeepCluster). **(c)** UMAP visualization of Concerto learned embeddings for PBMC 160k dataset (RNA, Protein, RNA+ Protein) labelled by Azimuth (Level 1 and Level 2 annotations). Blue arow shows continuous manifold of CD8 T cells, CD4 T cells and B cells. **(d)** Heatmap illustration of how attention weights relate to canonical modal-specific markers for major immune cell types. (Left, RNA marker; Right, Protein marker) Normalized heatmap, minimum (dark) to maximum (yellow green). **(e)** UMAP visualization colored by cell type label

and batch label for integrating Human Pancreatic Islet scRNA-seq dataset (n=14,890 cells, 8 batches of 5 technologies). Top, nHVG=2000; Bottom, All gene. Benchmark methods include Seurat V3 (CCA), Harmony, trVAE and uncorrected baseline (PCA only). **(f)** Over-correction analysis via removing all beta-cells except Cel-Seq2 batch, illustrated by UMAP visualization. **(g)** Comparison of batch-correction metric by kBET and ASW for 6 HVG scenarios (nHVG=2,000, 5,000, 10,000, 15,000, 20,000 and all gene). ASW, average silhouette score.

Figure 4 Concerto achieves state-of-the-art accuracy for query-to-reference mapping and supports projecting unseen cell types in the reference with scalability of multimillion. (a) Benchmarking accuracy of transferred annotations against Symphony, scArches, Seurat V4 (Left, HP→inDrop n= 8,569 cells, reference is HP dataset except inDrop as query; Right, HP→MP, reference is HP dataset except inDrop and Mouse Pancreatic Islet (MP) as query, n= 1,880 cells). (Repeated 5 times) **(b)** Confusion matrix of Concerto prediction for left, HP→inDrop and right, HP→MP. ACC, minimum (yellow) to maximum (dark blue). **(c)** Illustration of Concerto's ability of projecting unseen cell types onto reference via operating on all genes, evaluated on mulit-modal PBMC 160k dataset (RNA+ Protein). All CD 8 T cells in 20% query set is removed. Upper, all gene; Bottom, nHVG=2000. **(d)** Top, Heatmap shows Concerto can successfully identify CD 8 T cell masked in the reference, expressing canonical CD8 Protein marker, cytotoxic GZMA, CCR7 RNA marker enriched in annotated CD 8 T cell region. Heatmap, minimum (dark) to maximum (yellow green). Bottom, Concerto preserve biological signals consistent with the distance of annotated CD 8 T cells away from NK cells, including negative correlated cytotoxic marker GMLY, NKG7 (increased expression with closer distance) and positive correlated naïve/memory marker CCR7. **(e)** Pearson correlation coefficient of inferred protein expression with ground truth (top 20 proteins). Bar height represents the average correlation for each protein. **(f)** Heatmap of ground truth protein expression versus Concerto prediction for CLEC12A, CD155, CD11C and CD14 visualized by UMAP. (5 nearest neighbors (5-NN) are used to smooth and infer protein expression). Heatmap, minimum (dark) to maximum (yellow green). **(g)** Scalability of Concerto measured by elapsed time for reference building and querying cells of the same size (n=50k, 100k, 500k, 1M, 2M and 10M, respectively). **(h)** Comparison of overall performance among Concerto and 3 other packages.

Figure 5 Hierarchical query-to-reference mapping preserves differential immune response of COVID-19 patients. (a) Illustration of mapping Jonus et al PBMC scRNA dataset (n=99,049 cells) against PBMC 160k (RNA only). **(b)** UMAP visualization of annotated CD8 T cells divided into 5 subtypes and differential distribution among healthy controls (HC), moderate (M) and severe (S) disease status. Expression heatmap shows canonical markers for CD8 naïve, proliferating, memory, effector, and exhaustion states. Bottom, boxplot shows relative percentages of CD8 T subtypes among CD8 T cells across different status. **(c)** UMAP visualization of annotated NK cells divided into 3 subtypes with differential appearance (HC, M, S). Boxplot shows relative percentages of NK $CD56^{dim}CD16^{bright}$ and NK $CD56^{bright}$ cells among annotated NK cells at different stages. Expression heatmap of canonical markers is visualized by UMAP. **(d)** Top, Concerto separates classical (CD14+) and non-classical (CD 16+) monocytes across disease stages with annotations visualized by UMAP. Middle, expression heatmap of CD14 and CD16. Bottom, deficiency of antigen presentation marker

(HLA-DR) and enrichment of inflammatory marker (S100A) co-locate in the upper left area of annotated monocytes visualized by UMAP, marked by red circle. Boxplot, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, **** $p < 0.0001$. Heatmap, minimum (dark) to maximum (yellow green).

Figure 1

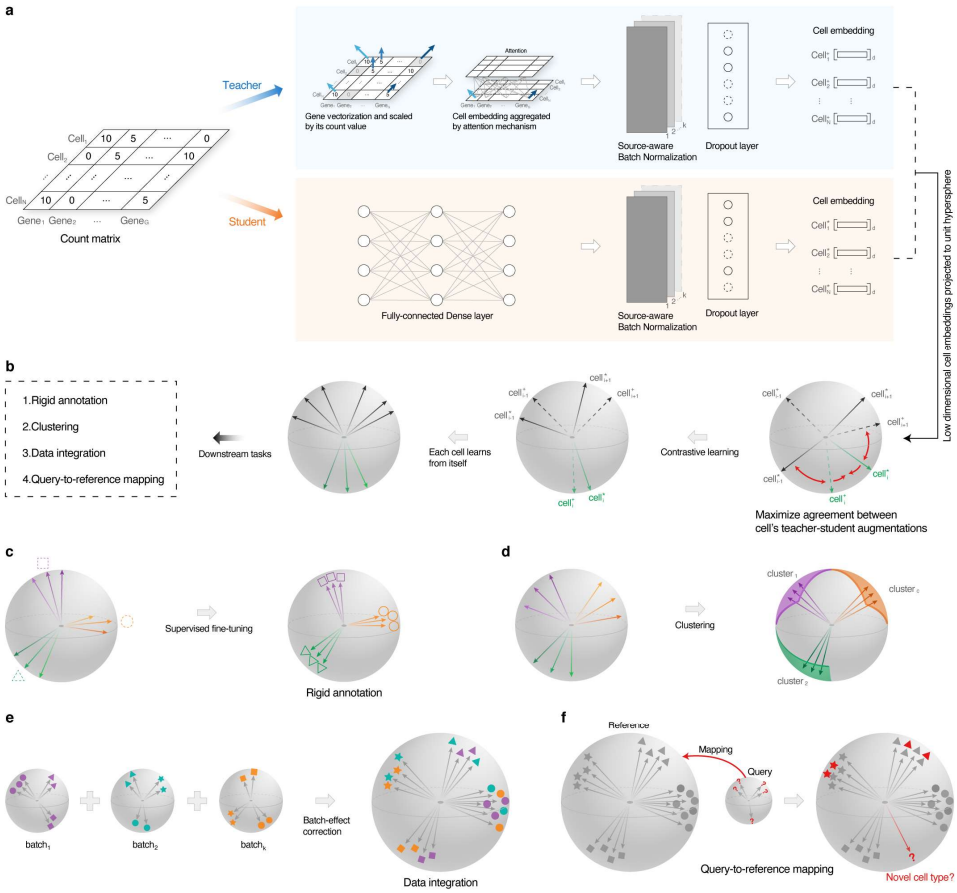

a
Teacher
Student
Count matrix
Gene vectorization and scaled by its count value
Cell embedding aggregated by attention mechanism
Source-aware Batch Normalization
Dropout layer
Cell embedding
Fully-connected Dense layer
Low dimensional cell embeddings projected to unit hypersphere
b
1.Rigid annotation
2.Clustering
3.Data integration
4.Query-to-reference mapping
Downstream tasks
Each cell learns from itself
Contrastive learning
Maximize agreement between cell's teacher-student augmentations
c
Supervised fine-tuning
Rigid annotation
d
Clustering
e
Batch-effect correction
Data integration
f
Reference
Mapping
Query
Query-to-reference mapping
Novel cell type?

Figure 2

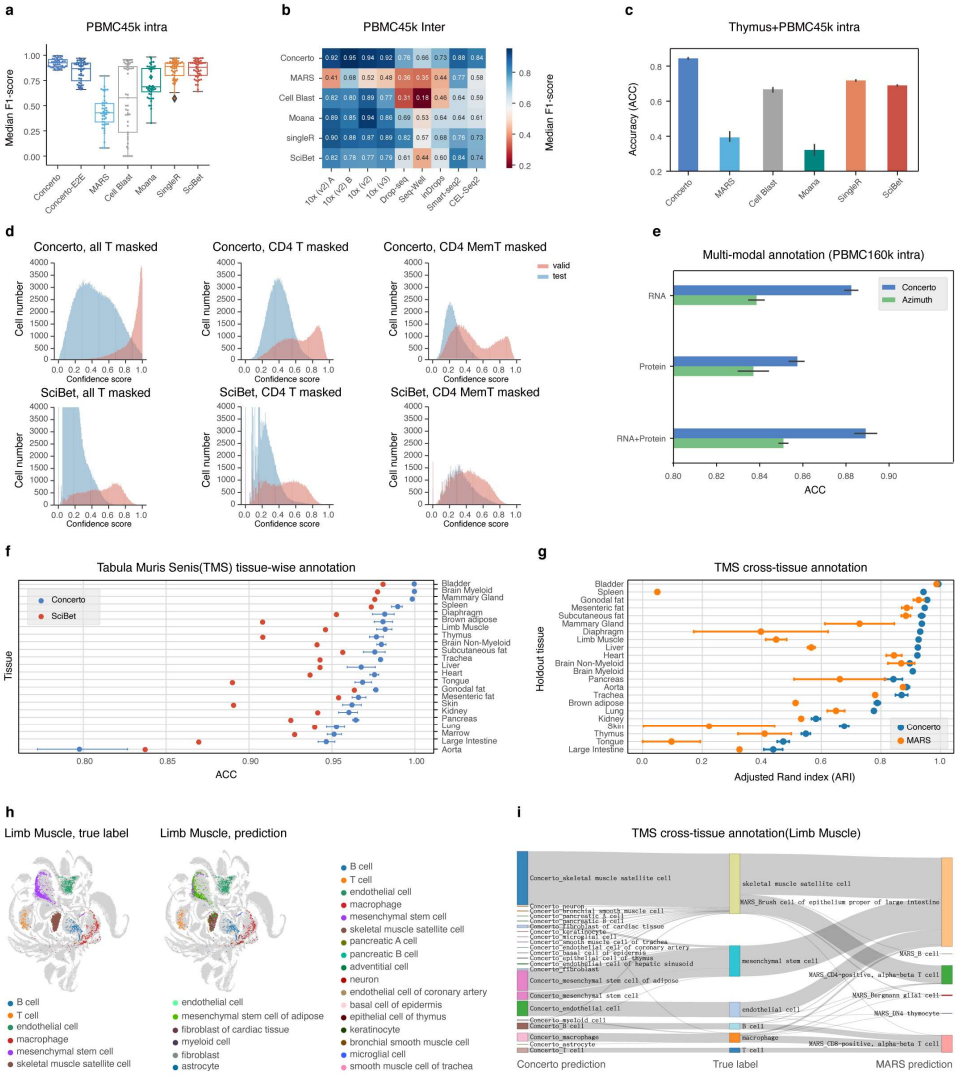

a
PBMC45k intra
Median F1-score
Concerto
Concerto-EEE
MARS
Cell Blast
Moana
SingleR
SciBet
b
PBMC45k Inter
Concerto
MARS
Cell Blast
Moana
singleR
SciBet
10x (v2) A
10x (v2) B
10x (v3)
10x (v3)
Drop-seq
Seq-Well
inDrops
Smart-seq2
CEL-Seq2
Median F1-score
c
Thymus+PBMC45k intra
Accuracy (ACC)
d
Concerto, all T masked
Concerto, CD4 T masked
Concerto, CD4 MemT masked
SciBet, all T masked
SciBet, CD4 T masked
SciBet, CD4 MemT masked
Cell number
Confidence score
valid
test
e
Multi-modal annotation (PBMC160k intra)
RNA
Protein
RNA+Protein
Concerto
Azimuth
ACC
f
Tabula Muris Senis(TMS) tissue-wise annotation
Concerto
SciBet
Tissue
ACC
g
TMS cross-tissue annotation
Holdout tissue
Concerto
MARS
Adjusted Rand index (ARI)
h
Limb Muscle, true label
Limb Muscle, prediction
B cell
T cell
endothelial cell
macrophage
mesenchymal stem cell
skeletal muscle satellite cell
pancreatic A cell
pancreatic B cell
adventitial cell
neuron
endothelial cell of coronary artery
basal cell of epidermis
epithelial cell of thymus
keratinocyte
bronchial smooth muscle cell
microglial cell
smooth muscle cell of trachea
mesenchymal stem cell of adipose
fibroblast of cardiac tissue
myeloid cell
fibroblast
astrocyte
i
TMS cross-tissue annotation(Limb Muscle)
Concerto prediction
True label
MARS prediction

Figure 3

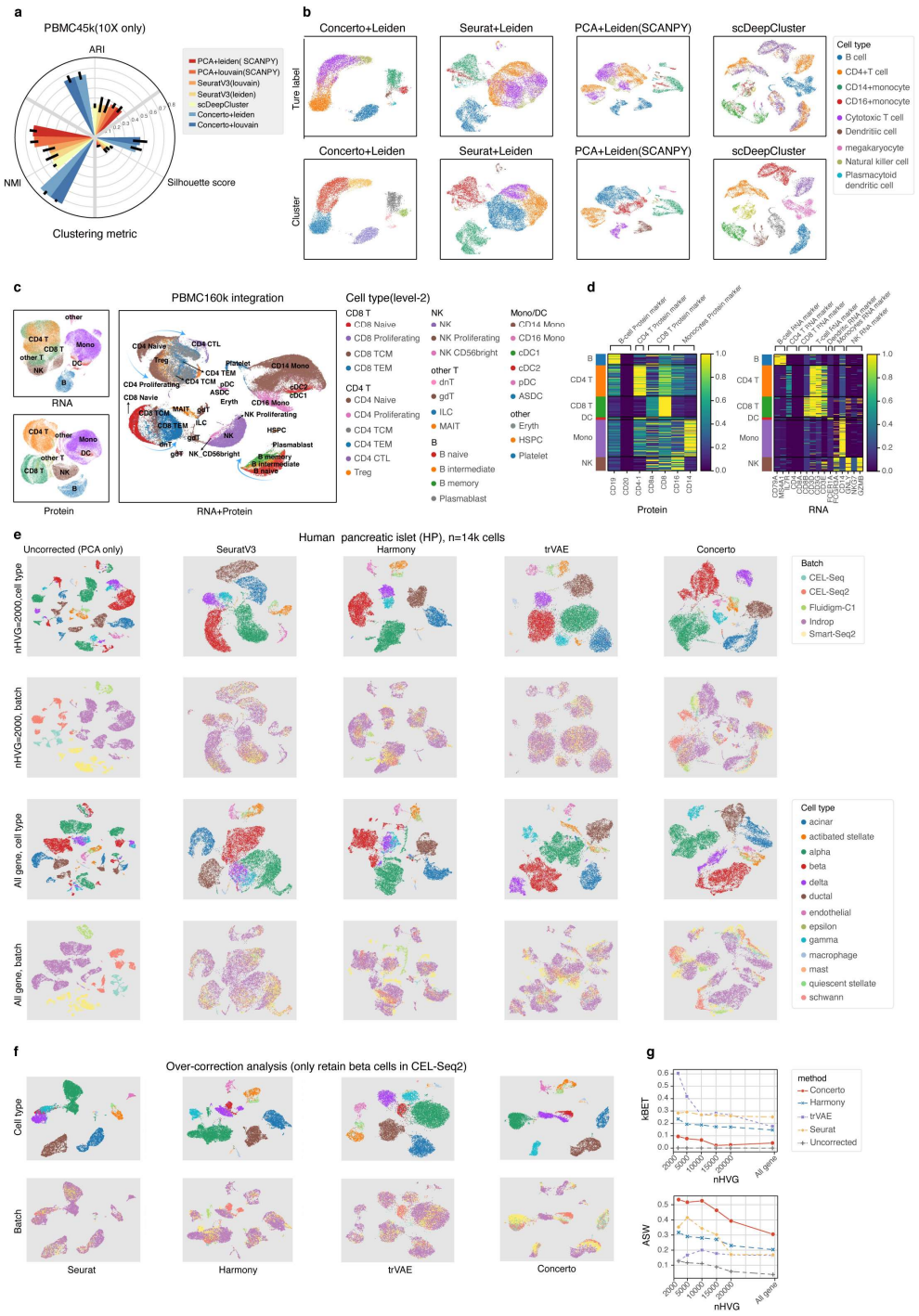

a
PBMC45k(10X only)
ARI
NMI
Silhouette score
Clustering metric
PCA+leiden(SCANPY)
PCA+louvain(SCANPY)
SeuratV3(louvain)
SeuratV3(leiden)
scDeepCluster
Concerto+leiden
Concerto+louvain
b
Concerto+Leiden
Seurat+Leiden
PCA+Leiden(SCANPY)
scDeepCluster
Ture label
Cluster
Cell type
B cell
CD4+T cell
CD14+monocyte
CD16+monocyte
Cytotoxic T cell
Dendritiic cell
megakaryocyte
Natural killer cell
Plasmacytoid dendritic cell
c
PBMC160k integration
RNA
Protein
RNA+Protein
Cell type(level-2)
CD8 T
CD8 Naive
CD8 Proliferating
CD8 TCM
CD8 TEM
CD4 T
CD4 Naive
CD4 Proliferating
CD4 TCM
CD4 TEM
CD4 CTL
Treg
NK
NK
NK Proliferating
NK CD56bright
other T
dnT
gdT
ILC
MAIT
B
B naive
B intermediate
B memory
Plasmablast
Mono/DC
CD14 Mono
CD16 Mono
cDC1
cDC2
pDC
ASDC
other
Eryth
HSPC
Platelet
d
B-cell Protein marker
CD4 T Protein marker
CD8 T Protein marker
Monocytes Protein marker
B-cell RNA marker
CD4 T RNA marker
CD8 T RNA marker
Cytotoxic RNA marker
Monocytes RNA marker
NK RNA marker
Protein
RNA
e
Human pancreatic islet (HP), n=14k cells
Uncorrected (PCA only)
SeuratV3
Harmony
trVAE
Concerto
nHVG=2000,cell type
nHVG=2000, batch
All gene, cell type
All gene, batch
Batch
CEL-Seq
CEL-Seq2
Fluidigm-C1
Indrop
Smart-Seq2
Cell type
acinar
actibated stellate
alpha
beta
delta
ductal
endothelial
epsilon
gamma
macrophage
mast
quiescent stellate
schwann
f
Over-correction analysis (only retain beta cells in CEL-Seq2)
Cell type
Batch
Seurat
Harmony
trVAE
Concerto
g
kBET
ASW
nHVG
All gene
method
Concerto
Harmony
trVAE
Seurat
Uncorrected

Figure 4

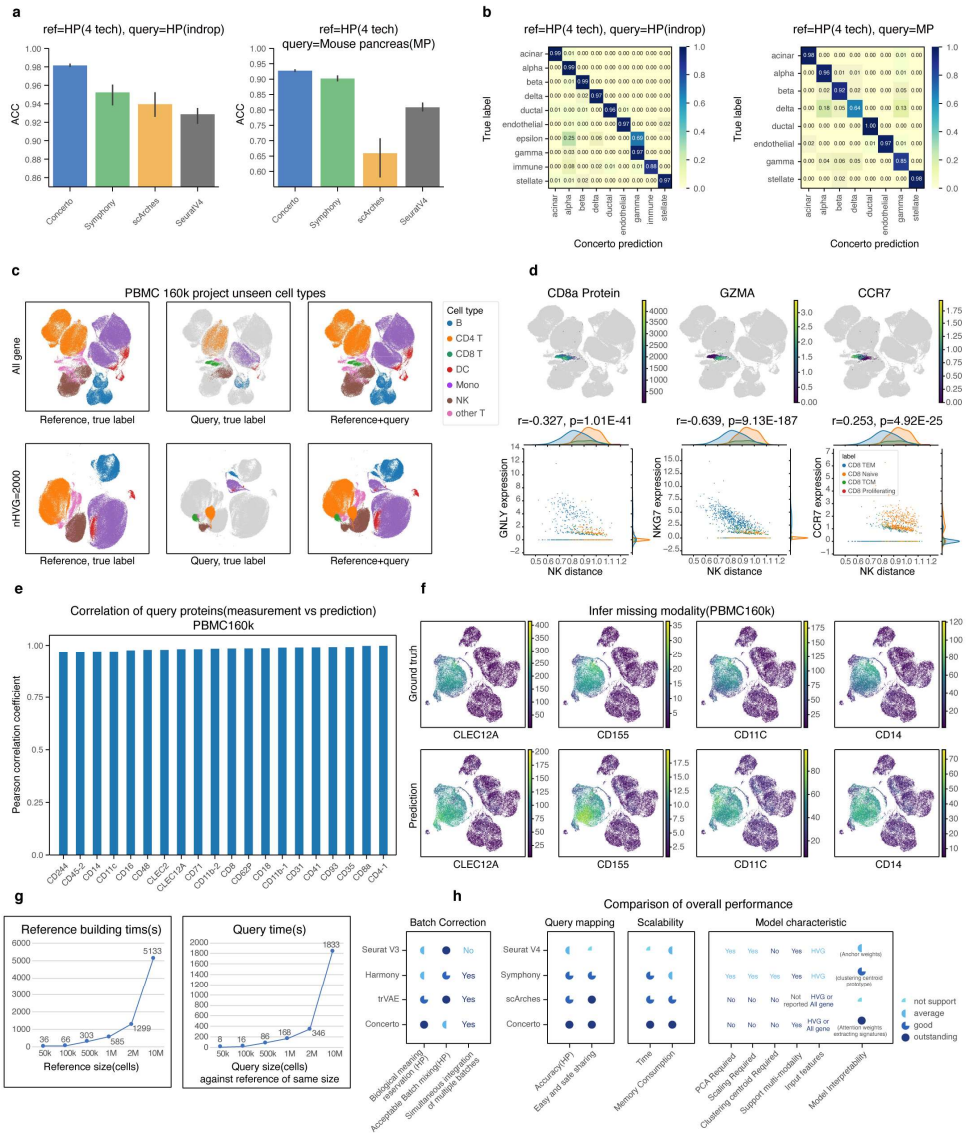
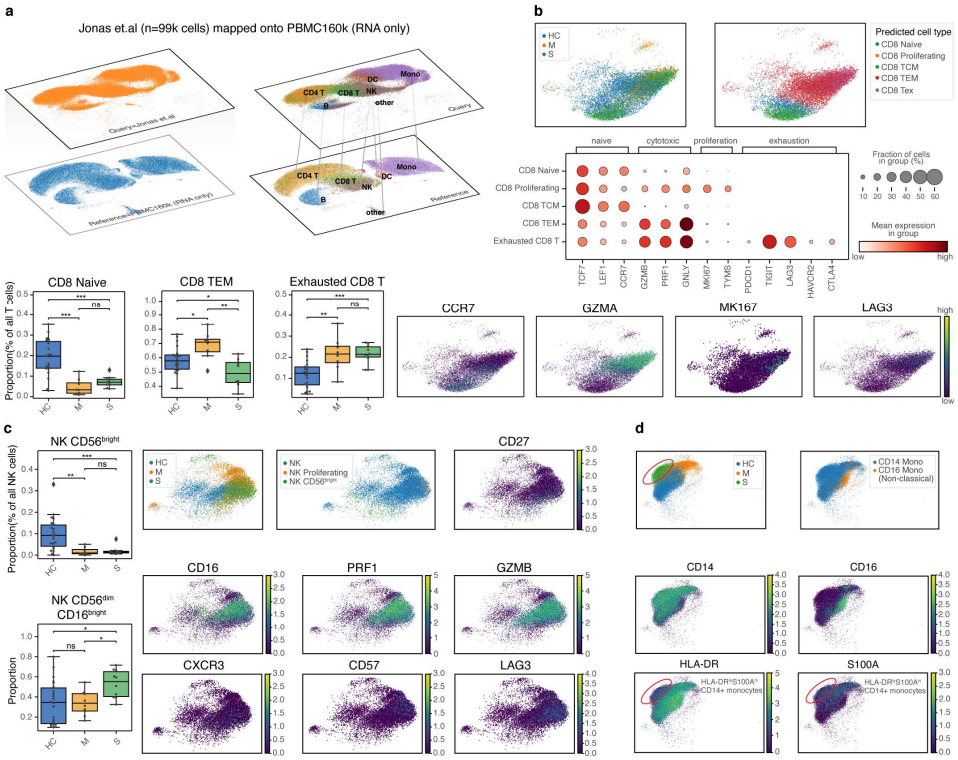

a
ref=HP(4 tech), query=HP(indrop)
ref=HP(4 tech) query=Mouse pancreas(MP)
ACC
Concerto
Symphony
scArches
SeuratV4
b
ref=HP(4 tech), query=HP(indrop)
ref=HP(4 tech), query=MP
True label
Concerto prediction
c
PBMC 160k project unseen cell types
All gene
nHVG=2000
Reference, true label
Query, true label
Reference+query
Cell type
B
CD4 T
CD8 T
DC
Mono
NK
other T
d
CD8a Protein
GZMA
CCR7
r=-0.327, p=1.01E-41
r=-0.639, p=9.13E-187
r=0.253, p=4.92E-25
GNLY expression
NKG7 expression
CCR7 expression
NK distance
label
CD8 TEM
CD8 Naive
CD8 TCM
CD8 Proliferating
e
Correlation of query proteins(measurement vs prediction)
PBMC160k
Pearson correlation coefficient
CD244
CD45-2
CD14
CD11c
CD16
CD48
CLEC2
CLEC12A
CD71
CD11b-2
CD8
CD62P
CD16
CD11b
CD31
CD41
CD62
CD36
CD8a
CD4-1
f
Infer missing modality(PBMC160k)
Ground truth
Prediction
CLEC12A
CD155
CD11C
CD14
g
Reference building tims(s)
Reference size(cells)
Query time(s)
Query size(cells)
against reference of same size
h
Comparison of overall performance
Batch Correction
Query mapping
Scalability
Model characteristic
Seurat V3
Harmony
trVAE
Concerto
Seurat V4
Symphony
scArches
Biological meaning reservation (HP)
Acceptable Batch mixing(HP)
Simultaneous integration of multiple batches
Accuracy(HP)
Easy and safe sharing
Time
Memory Consumption
PCA Required
Scaling Required
Clustering centroid Required
Support multi-modality
Input features
Model Interpretability
not support
average
good
outstanding

Figure 5


a
Jonas et.al (n=99k cells) mapped onto PBMC160k (RNA only)
Query-Jonas et.al
Reference-PBMC160k (RNA only)
CD4 T
CD8 T
NK
DC
Mono
B
other
Query
Reference
b
HC
M
S
Predicted cell type
CD8 Naive
CD8 Proliferating
CD8 TCM
CD8 TEM
CD8 Tex
naive
cytotoxic
proliferation
exhaustion
Fraction of cells in group (%)
Mean expression in group
low
high
TCF7
LEF1
CCR7
GZMB
PRF1
GNLY
MKI67
TYMS
PDCD1
TIGIT
LAG3
HAVCR2
CTLA4
CD8 Naive
CD8 TEM
Exhausted CD8 T
Proportion(% of all T cells)
CCR7
GZMA
MK167
LAG3
c
NK CD56bright
NK CD56dim CD16bright
Proportion(% of all NK cells)
Proportion
NK
NK Proliferating
NK CD56bright
CD27
CD16
PRF1
GZMB
CXCR3
CD57
LAG3
d
CD14 Mono
CD16 Mono
(Non-classical)
CD14
CD16
HLA-DR
S100A
HLA-DRhiS100Ahi
CD14+ monocytes

Methods

Concerto uses both simulated and real world single cell RNA-seq and CITE-seq data and implements several algorithms. A full description of the data source is in Supplementary Table 1, and pre-processing of individual data is discussed in the Supplementary Table 2. Here we describe the computational methods for general data filtering, normalization, input encoding, network backbone composed of teacher and student, data augmentation with dropout layer, contrastive loss function, multi-modality representation, supervised fine-tuning, data integration and query-to-reference mapping. Concerto is released in python package (https://github.com/melobio/Concerto), including additional functions mentioned below.

Filtering, Pre-processing, and Normalization

For real scRNA-seq data, we delete mitochondrial genes ('ERCC','MT-','mt-'), then we discard low-quality cells expressing fewer than 600 genes and removed genes expressed in fewer than three cells. To correct the effect of sequencing depth, we use SCANPY (1.7.1) to normalize each cell to 10,000 read counts and apply logarithmic transformation with 'log (count + 1)' operation.

HVGs selection

Concerto supports operating on all genes. Except 'all gene' scenario, we use SCANPY (1.7.1) function to conduct HVGs selection. For 'all genes' scenario, Concerto keeps consistent number of input genes within a batch. For 'HVG' scenario, only selected HVGs are used to generate index and value.

Homolog alignment

In the HP→ MP transfer task, the orthologous genes based on the Mouse Genome Informatics database are used as common genes of two species. The intersection of orthologous genes and genes (http://www.informatics.jax.org/downloads/reports/HOM_MouseHumanSequence.rpt. accessed 15 Aug 2020)

Input encoding scheme

Concerto leverages TF-record file format to encode normalized gene-count matrix. TF-record is a convenient way for sharding file in TensorFlow. A TF-record is a binary file that contains sequences of byte-strings. Data is serialized (encoded as a byte-string) before being written into a TF-record. Concerto encapsulates 'gene index' and 'count value' into TF-record file. Teacher network accepts both 'gene index' and 'count value' from TF-record files, while student network only reads 'count value' file.

Teacher network

Teacher network accepts $\mathrm{X}_{\mathrm{index}} \in \mathbb{R}^{\mathrm{N}\times\mathrm{G}}$ and $X_{counts} \in \mathbb{R}^{N\times G}$, where N denotes number of cells and G denotes number of genes, X_{index} represents gene index, index refers to gene ID defined by a dictionary consisted of all genes of a certain species. X_{counts} represents value of gene counts. Each gene within a cell is represented by $g_i(\, i \in \mathbb{R}^G)$, and count of certain gene is represented by $v_i(\, i \in \mathbb{R}^G)$. Firstly, g_i will be embedded into a d-dimension vector space, $e_{it} \in \mathbb{R}^{G\times d}(1)$, d is set to be 128 as default. Cross product of e_i and v_i outputs weighted hidden vector $h_{it} \in$

$\mathbb{R}^{G\times d}$ (2).

$$e_{it} = Embedding(g_i) \qquad i \in \mathbb{R}^G \quad t \in \mathbb{R}^d \qquad (1)$$

$$h_{it} = e_{it} \times v_i \qquad h_{it} \in \mathbb{R}^{G\times d} \qquad (2)$$

Concerto then uses attention mechanism to aggregate gene embeddings. h_{it} is firstly fed into a MLP with one hidden layer and non-linear $tanh$ transformation to obtain hidden vector $u_{it} \in \mathbb{R}^{G\times d}$(3). Then a cellular context vector u_w will be dot product with u_{it} with $softmax$ operation to obtain attention weights $\alpha_i \in \mathbb{R}^G$ (4); then aggregation is implemented on all genes' vectors h_{it} through weighted summation by attention weights α_i, obtaining aggregated vectors, $s_t \in \mathbb{R}^d$(5). MLP stands for multi-layer perceptron.

$$u_{it} = \tanh(W_{\omega 1} h_{it} + b_{\omega 1}) \qquad (3)$$

$$\alpha_i = \frac{\exp(u_{it}^T u_w)}{\sum_t \exp(u_{it}^T u_w)} \qquad (4)$$

$$s_t = \sum_i(\alpha_i \times h_{it}) \qquad (5)$$

Attention layer output will be fed into a Batch Normalization layer followed by Dropout layer, and then Dense layer with ReLU activation, leads to the final output of teacher network, $z_t \in \mathbb{R}^d$ (6)

$$Z_t = \text{ReLU}(W_{\omega 2} s_t + b_{\omega 2}) \qquad (6)$$

Student network

Student network accepts only $X_{counts} \in \mathbb{R}^{N\times G}$, then going through Batch Normalization layer followed by Dropout layer, and then Dense layer with ReLU activation, leads to the final output of student network, $Z_s \in \mathbb{R}^d$.

Data augmentation with dropout layer

Dropout layer operation is used as a minimal data augmentation strategy to obtain different views of the same cell. Via randomly masking neural units with certain probability (parameters, *dropout rate*= 0.2) before the final dense layer, different embeddings of positive cell pairs will be generated for subsequent contrastive learning.

Contrastive loss (NT-Xent loss)

Contrastive learning is conducted in a unit hypersphere space and explicitly compares pairs of cell embeddings of *d* dimension (*d*=128 as default). Through contrastive learning, Concerto pushes apart different cells within a batch while pulling together teacher-student views of the same cell as positive pairs. Assume a positive pair as $i \in Z_t$ and $j \in Z_s$, the distance is measured by cosine similarity of L2-normalized embeddings using dot product operation (7). NT-Xent loss stands for normalized temperature-scaled cross entropy loss, as formalized in (8). τ stands for adjustable temperature coefficient, which can be used to scale the degree of pushing apart negative samples. We randomly sample a minibatch of N cells and compute NT-Xent loss on pairs of augmented examples derived from the minibatch, resulting in 2N data points (9). Given a positive pair, the other 2(N − 1) augmented examples within a minibatch are treated as negative examples.

$$s_{i,j} = z_i^T z_j / \tau \parallel z_i \parallel \parallel z_j \parallel \qquad (7)$$

$$\ell_{i,j} = -log \frac{\exp(s_{i,j})}{\sum_{k=1}^{2N} \mathbb{1}_{[k\neq i]} \exp(s_{i,k})} \qquad (8)$$

$$\mathcal{L} = \frac{1}{2N}\sum_{k=1}^{N}[\ell(2k-1,2k) + \ell(2k,2k-1)] \quad (9)$$

Multimodal integration

Concerto support convenient multi-modal integration. To process multi-omics datasets, simply element-wise summation of modal-specific attention output in teacher module or dense output in student module enables Concerto to generate unified cell embeddings. In the case of two modalities (RNA and protein), we illustrate corresponding operations as in (10) and (11).

$$Z_s^{multi} = Add\left(Z_s^{RNA}, Z_s^{protein}\right) \quad Z_s^{multi} \in \mathbb{R}^d \quad (10)$$
$$Z_t^{multi} = Add(Z_t^{RNA}, Z_t^{protein}) \quad Z_t^{multi} \in \mathbb{R}^d \quad (11)$$

Supervised fine-tuning

For rigid annotation, Concerto leverages contrastive learning as task-agnostic pre-training procedure followed by supervised fine-tuning using manually annotated labels. For within-datasets prediction (intra-dataset), we fine-tune Concerto via an extra fully-connected classification layer with *softmax* operation over dimension of pre-defined cell type categories. The loss function is classical supervised cross entropy loss (12). For cross-datasets prediction (inter-dataset), we conduct semi-supervised fine-tuning via adding a domain-adaptation module to derive cross-tissue or cross-species predictions. We also validate the inferior performance of end-to-end training (Concerto-E2E) by discarding the contrastive loss without self-supervised training procedure while only retaining model backbone to conduct fully supervised training.

$$min\ \mathcal{J}(\theta) = \mathbb{E}_{x\sim PL(x)}[CE(p_\theta(x) \parallel x^*)] \quad (12)$$

Where $x \in Z_t^{source}$ stands for pre-trained source embeddings by Concerto; $p_\theta(x)$ stands for predicted classification probability. θ stands for parameters of final classification layer, x^* stands for true labels of source data, and $PL(x)$stands for distribution of source data.

Domain-adaptation module

For inter-dataset annotation, we add a domain adaptation module to adapt the target labels to the source distribution. Besides the supervised cross-entropy loss, we add a unsupervised consistency training loss (13).

$$\min\ \mathcal{J}(\theta) = \lambda\mathbb{E}_{y\sim PU(y)}\mathbb{E}_{\hat{y}\sim PU(\hat{y})}[CE(p_{\tilde{\theta}}(y|x) \parallel p_\theta(\hat{y}|x))] \quad (13)$$

Where $y \in Z_t^{target}$ stands for embeddings for unlabeled target cells to be predicted, and $p_\theta(y|x)$ stands for predicted classification probability. $\tilde{\theta}$ is a fixed copy of the current parameters θ indicating the gradient is not propagated through $\tilde{\theta}$, while $\hat{y} \in Z_s^{target}$ stands for augmented embeddings for unlabeled target cells. $PU(y)$ stands for distribution of target data. We set λ to 1 and add (12) and (13) to train Concerto in a semi-supervised setting for inter-dataset prediction.

Source-aware Batch Normalization for data-integration

For data integration, Concerto aims to learn batch-invariant embeddings to integrate heterogeneous

single-cell datasets from different sources and overcome batch-effects. Meta-data of source (batch) information will be incorporated as data Source ID and guide Concerto implementing source-specific Batch Normalization when cells pass through the network within a training mini-batch. This training strategy ensures that Batch Normalization is only conducted within cells from the same source defined by the investigator.

Query-to-reference mapping

For query-to-reference mapping, it is assumed that a reference atlas has been constructed via above operations. Cell embeddings of smaller-scale query cells are simply inferred via passing through the trained teacher network. In this case, users directly utilize the model weights and contextualize query cells onto a stable reference embedding space. On the other hand, users can also leverage existing model weights as initialization and implement unsupervised fine-tuning in a contrastive manner on query cells, which is equivalent to joint analysis of reference and query cells. Z_t^{ref} stands for the contrastively learned embeddings for reference data and Z_t^{query} stands for query embeddings either from direct inference or unsupervised fine-tuning.

NN-voting classifier

To transfer annotations from reference cells to query cells after obtaining all cell embeddings, Concerto uses a simple k-NN voting classifier to annotate query cells. The NN-voting classifier will assign k nearest neighbors for query cells, $y \in Z_t^{query}$. Cosine similarity is used to calculate the distance between neighbors, D_{y,N_y}(14) and normalization of D_{y,N_y} is implemented as in (15) to calculate the probability of assigning reference annotations (x^*) to query cells y. Where $x^{*(i)}$ is the annotation label of the *i-th* neighbor and y' is the transferred annotation with the maximum probability (16). We set k=5 for most cases, while k can be tuned accordingly.

$$D_{y,n,N_y} = cosine(y,n) \tag{14}$$

$$p(x^*,y) = \frac{\sum_{i\epsilon N_y} I(x^{*(i)}=x^*) D_{y,n_i,N_y}}{\sum_{j\epsilon N_y} D_{y,n_j,N_y}} \tag{15}$$

$$y' = argmax(p(x^*,y)) \tag{16}$$

UMAP Visualization

Cell embeddings are visualized by UMAP from SCANPY (1.7.1). The number of neighbors, *n_neighbors*, is set to 15. Other parameters are set as use_rep='X', metric= 'euclidean'. We then plot umap using scanpy.pl.umap function with default parameter.

Hyperparameters

Learning rate in the contrastive learning procedure is set to varied vale from 1e-4 to 1e-6 using Adam optimizer training for 3 epochs. For fine-tuning stage, learning rate is set to be 1e-3 using Adam optimizer training for 1 epoch. Temperature coefficient in NT-Xent loss is set to 0.1. Mini-batch size equals to 32. Number of all dense layer units (except final classification layer) is 128.

Exhausted T Annotation in COVID-19 analysis

Since CD8 exhausted T cell does not exist in PBMC160k dataset, we calculate exhaustion signature score using PDCD1, TIGIT, LAG3, HAVCR2, CTLA4 by summing the values (scaled to 0 to 1) as

an exhaustion gene set. CD8 T cells with exhausted score more than 0 are annotated as exhausted T cell.

Attention weight extraction

Attention weights α_i are extracted from pre-trained teacher network. Firstly, we load the pre-training weights from the teacher network as X_{index} and X_{counts}, and pass through a forward propagation. Then we obtain the attention weights α_i output by softmax after linear transformation and dot product with cellular context vector u_w according to the attention mechanism. All calculations are conducted in 128-dimensional space.

Robustness analysis

Simulated datasets for robustness analysis are generated using the Splatter R package. For differential expression (DE) simulation, following parameters are used in the splatSimulate() R function: groupCells = 5, nGenes = 2500, dropout.present = TRUE, dropout.shape -1, dropout.mid = 1, de.scale = 0.15, 0.2, 0.25, 0.3 respectively. For dropout rate simulation, following parameters are used in the splatSimulate() R function: groupCells = 5, nGenes = 2500, dropout.present = TRUE, dropout.shape -1, dropout.mid = -0.5, 0, 0.5, 1 respectively, de.scale = 0.2.

Scalability analysis

For scalability analysis, simulated datasets are generated using the scsim Python package[56], which is based on the Splatter statistical framework while it performs better efficiency to generate large scale simulated data. Following parameters are used in scsim() Python function: ngenes=25000, ncells=50000, 100000, 50000, 1000000, 2000000, 10000000, ngroups=5, diffexpprob= 0.025.

Analytic metrics

F1-score and ACC

F1-score and ACC are used to assess the annotation performance and Python function sklearn.metrics.f1_score() and sklearn.metrics.accuracy_score() from the scikit-learn library are used respectively.

ARI and NMI

Adjusted Rand Index (ARI)[57], Normalized Mutual Information (NMI) are applied to assess clustering performance. Python library scikit-learn is used to calculate ARI and NMI, and specifically, Python function adjusted_rand_score() and normalized_mutual_info_score() are used.

Silhouette coefficient (ASW)

ASW was calculated using Python function sklearn.metrics.cluster.silhouette_samples() from Python library scikit-learn, which reflects the performance of biological meaning reservation.

kBET

kBET indicates how well mixed batches from randomly sampled nearest-neighbor cells are based on local batch label distribution in consistent with global batch label distribution, which is a metric for batch-effect correction. Pegasus is adopted for kBET calculation, and the k is set to 15.

Clustering methods

K-means

For clustering benchmark, we use Python function sklearn.cluster.KMeans() to perform K-means.

Leiden

To perform Leiden algorithm, we apply R function FindClusters() from R package Seurat V3 and the parameter 'algorithm' is set to 4. Also we apply Python function scanpy.tl.leiden from Scanpy , a Python library, to perform Leiden algorithm.

Louvian

R function FindClusters() from R package Seurat V3 is used to perform Louvain algorithm while the parameter 'algorithm' is set to 1. Python function scanpytl.louvain is applied to perform Louvain algorithm in the Scanpy clustering analysis.

Other benchmarking tools

Seurat V3

Count matrix is firstly normalized using 'LogNormalize' methods in R package Seurat V3(Seurat_3.9.9.9010) with default parameters. The top 2,000 variable genes are then identified using the 'vst' method in Seurat FindVariableFeatures () function. FindNeighbors() function with defaut parameters from Seurat is used to build SNN graph. Finally, FindCluster() function is applied to assign cluster labels to each cell.

Seurat V4

SelectIntegrationFeatures(), FindIntegrationAnchors(), and IntegrateData() from Seurat V4(Seurat_4.0.0) are used to perform batch effect correction. For query-to-reference mapping task, we run Seurat V4 and follow the recommended steps from the author's script (https://satijalab.org/seurat/articles/integration_mapping.html) to build the reference and map the query.

SCANPY

For the clustering task, we run SCANPY (version 1.7.1) and followed the steps from the author's script (https://scanpy-tutorials.readthedocs.io/en/latest/pbmc3k.html).

Moana

The Moana(https://github.com/yanailab/moana) framework consists of methods for clustering and cell type classification. Moana relies on support-vector machines (SVMs) with a linear kernel, trained on PCA-transformed data. The PCA model takes smoothed or raw value as input and returns a matrix with far fewer dimensions, typically 20. As a result, the only parameter that needs to be tuned (adjusted) to train Moana is k, which determines the degree of smoothing. When cell types are sufficiently diverse relative to the inherent noise levels, we can simply use k=1 (no smoothing).

SingleR

SingleR (singleR v1.0) selects the most variable genes for each cell type. Then, cell types are identified by correlating expression values. We use logNormCounts() in the scater package in pre-

processing part and run singleR using default parameters.

Cell BLAST

For Cell BLAST (version 0.3.8), we use λ_b= 0.01, which determines batch alignment strength, and fix embedding size at 30. We pre-train DIRECTi model for 30 epochs and the learning rate is 1e-4.

SciBet

SciBet(https://github.com/zwj-tina/scibet) retrieves cell type markers and eliminates noisy genes by E-test. For each cell type, SciBet learns a multinomial model to form a likelihood function to define the probability of each cell belonging to a cell type, hence cell annotation relies on a likelihood maximization process. In rigid annotation task, 2000 HVGs are selected, and other parameters are set as default.

MARS

MARS (https://github.com/snap-stanford/mars) uses a meta-dataset consisting of single cell expression profiles from different tissues and developmental stages, with and without cell type labels, to train a neural network which serves as a common embedding function, transferring latent representations between datasets using landmark representations. We specify that the dimension of the first hidden layer is 5000, and the dimension of the second hidden layer is 500. After 25 epoches of pre-training, following 50 epoches are trained, and the learning rate is 0.001.

scDeepCluster

ScDeepCluster(https://github.com/ttgump/scDeepCluster) employs a linear combination of ZINB loss and the Kullback–Leibler (KL) divergence between the distribution of soft labels of the embedding layer (measured by a student's t distribution) and the derivation of the target distribution. Because scDeepCluster creates embeddings with well-defined clusters, we need to define the number of clusters. The number of epochs for pre-training is 50. Meanwhile, Adam optimizer with initial learning rate of 1e-4 is used for training.

Harmony

In our analysis, we run Harmony (version 0.1) within the Seurat V3 workflow with the maximal number of clusters (100) and the maximal number of iterations (100). The top 20 normalized Harmony vectors in PCA space are used as input.

Symphony

We download symphony from https://github.com/immunogenomics/symphony and followed author's script to build the reference. MapQuery function is applied with default parameters.

trVAE and scArches

We download the scArches (version 0.3) with trVAE from Github (https://github.com/theislab/scarches). We set the hidden units to be (128; 128) for the encoder as recommended. The decoder is symmetric to the encoder. We chose 1e-4 as learning rate and NB distribution for data following authors' recommendation. We train on each unprocessed dataset for 200 epochs with batch size of 32, including a 200-epoch warm-up for the KL divergence loss.

Data availability

All scRNA-seq and CITE-seq datasets in this study are published previously, and their availabilities are described in Supplementary Table 1 with corresponding pre-processing description in Supplementary Table 2.

Code availability

Concerto is written in Python using the TensorFlow library. The source code is available on Github at https://github.com/melobio/Concerto.

Reference

56. Giguere, C., Dubey, H.V., Sarsani, V.K. et al. SCSIM: Jointly simulating correlated single-cell and bulk next-generation DNA sequencing data. BMC Bioinformatics 21, 215 (2020).
57. Lawrence Hubert and Phipps Arabie. Comparing partitions. Journal of Classification, 2(1):193–218, 1985.

Acknowledgement

This research is supported by Guangdong Provincial Academician Workstation of BGI Synthetic Genomics (2017B090904014).

Author contributions

M.Y. conceived the problem and designed the study. H.M.Y and F.M. supervised the work. Y.YX.Y. performed bioinformatics analysis. H.P.H and C.H.X performed deep learning experiments. M.Y. wrote the manuscript.

Competing interests

All authors declare no competing interests.

Additional information

Supplementary information is available for this paper in an additional Supplementary File, including 12 Supplementary Figures and 3 Supplementary Tables.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ConcertoSupplementaryMaterials.pdf