bioRxiv preprint doi: https://doi.org/10.1101/2021.09.09.459281; this version posted September 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

A General Framework for Representing and Annotating Multifaceted Cell Heterogeneity in Human Cell Atlas

Haoxiang Gao[1#], Kui Hua[1#*], Sijie Chen[1], Qijin Yin[1], Rui Jiang[1], Xuegong Zhang[1,2*]

[1]MOE Key Laboratory of Bioinformatics, Bioinformatics Division, BNRIST and Department of Automation, Tsinghua University, Beijing 100084, China

[2] School of Life Sciences and School of Medicine, Center for Synthetic and Systems Biology, Tsinghua University, Beijing 100084, China

[#] These authors contributed equally to this work

[*] Correspondence: zhangxg@tsinghua.edu.cn (XZ), stevenhuakui@gmail.com (KH)

Abstract

The goal of big projects like Human Cell Atlas (HCA) and Human BioMedical Atlas Program (HuBMAP) is to build maps that comprehensively define and describe all cell types and their molecular features in a healthy human being. Just like geographical maps must have coordinates, a key task in building cell maps is to provide coordinate systems for cells. A well-designed coordinate system helps better understand the highly orchestrated function and organization of different cells. Cells could be depicted by external information like their spatial locations in the body and organ, the sex and race of the donor, and multiple endogenous attributes of cells such as their types, states, functions, developing trajectory, etc. These heterogeneities are encoded in or can be predicted with transcriptomics and other omics data. Cell heterogeneities are multifaceted, including three major types: continuous values or scores, categorical groups and structured annotations. Here we propose to a unified multidimensional coordinate system UniCoord to represent the multifaceted heterogeneities of cells. It is based on a general deep learning framework, with a supervised VAE structure to learn the mapping relationship between gene expressions and the generated coordinates in a low-dimensional space that encode multiple cell attributes of the three types. Experiment results on several datasets showed that UniCoord was able to represent a variety of cell heterogeneous properties that are discrete, continuous or of hierarchical structures. The trained UniCoord model can be used to automatically label attributes of cells and generate the corresponding expression data. Experiments showed that UniCoord is a feasible coordinates framework for representing multifaceted cell heterogeneity in comprehensive cell atlases.

Introduction

Since cells been discovered, human's understanding to them have been revolutionized for several times. Morphological features, cellular function and molecular markers were used to describe cells in different historical stages and research areas. For hundreds of years, people are trying to build a cell atlas that can classify and locate all cells. Single-cell RNA sequencing (scRNA-seq), for the first time, provide cell information in omics level and in single-cell resolution, and thus gives us

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.09.459281; this version posted September 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

solid foundation to build the first comprehensive cell atlas.

A cell atlas that people trying to build should be able to describe all possible cells. Giving a cell, the atlas should locate the cell to a specific body position and developmental stage, which are spatial and temporal coordinate, respectively. Moreover, the atlas should simultaneously locate cells to cell types, cellular status (cell cycle, cell aging, etc.), and various biological activities, which are functional coordinates. With this atlas, people will not only be able to query cells more conveniently, but also be able to study complex conversion and interaction relationship between cells. It will also notice us about poorly understood part on the atlas, and even help finding new types of cells, like a periodic table.

To achieve this comprehensive atlas, lots of effort must be taken. First, people need to standardize or discover key function of cells, and to quantify these functions. For example, (Diehl et al. 2016) have been collecting possible cell types, standardizing their names and organizing hierarchical relationship between cell types. (Trapnell et al. 2014) construct a trajectory in scRNA-seq data and assign each cell a pseudo-time score. (Zhang et al. 2019) use tumor specific genes as bio-marker and calculate malignance level for tumor cells. (Sokolov, Paull, and Stuart 2016) trained one-class logistic regression with stem cells and the model could infer stemness for other cells. All these works on cell types, pseudo-time, malignance and stemness construct a biological system describing different kinds of cells.

Second, giving all these existing and to-be-discovered spatial, temporal and functional coordinates, a well-organized information system must be developed. The system should be designed to: (a). contain all coordinates and keep potential for additive information in the future; (b). be able to label new cells accurately and conveniently with all coordinates; (c). integrate coordinates as a whole system, and elucidate connection between them.

While the biological system of single cells being studied a lot, the information system is still not emphasized enough, though several attempts have been conducted. (Rood et al. 2019) have been working on building a spatial coordinate system that labels the original sampling site of cells. (Lopez et al. 2018), (Stuart et al. 2019) gave examples of embedding cells into a latent space, where no explicit interpretation for each latent dimensionality could be provided.

An information system for cell atlas is possible only if two key questions been answered: how to combine coordinates with multifaceted data type in one model, and how to make cell representations interpretable. For the first question, we have shown that coordinates are in different data type and may be further organized. In informatic aspect, these multifaceted heterogeneities might be categorized into discrete (donor ID, organ), continuous (pseudo-time, stemness), and hierarchical (different levels of cell types) coordinates. How to integrate all these data types into one model is an unsolved problem. For the second question, most of data embedding methods project cells into an abstract latent space ((Rostom et al. 2017; Stuart et al. 2019; Lopez et al. 2018)), where the latent dimensions cannot be easily explained or understood. However, the cell atlas needs its coordinates to be physically or biologically interpretable, where existing scRNA-seq methods failed.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.09.459281; this version posted September 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Here, we introduce UniCoord, a generative deep learning method that combine supervise and unsupervised learning, compatible for all types of coordinates known yet, including discrete, continuous and hierarchical structure. With the ability of integrating heterogeneous data types and making interpretable cell embedding, UniCoord gives the first possible solution for cell atlas information system. We tested UniCoord on several datasets. Results shows that our method learnt interpretable representations from scRNA-seq data precisely, and coordinates of cells are robust to batch effect or other irrelevant features. Representations might also help improving the performance of downstream analysis such as clustering, visualization and differential gene expression.

Results

The goal of this work is to propose a machine learning model that embeds scRNA-seq data into a latent space, where each latent dimensionality could be regard as a specific cellular attribute. Technically, the model needs to be compatible for both continuous and discrete latent feature, with each feature supervised by certain given label, such as cell cycling score, stemness of cells or cell type annotation. We construct the information coordinate system of cells as all heterogeneities of cells, including cell types, donor metadata, functional strength of cells and other heterogeneities possibly discovered in the future.

UniCoord: A mixture VAE method

We developed a refined VAE model UniCoord (Fig.1). The latent space of UniCoord differs from conventional VAE in two aspects: (a). latent space of UniCoord is a combination of discrete and continuous dimensions, while conventional VAE only contain continuous; (b). each dimension in UniCoord latent space could be physically interpretable if supervised by given functional scores. The gene expression level in a cell could be modeled by conditional distribution $p(x_n|ID_n, AD_n, IC_n, AC_n)$, where ID_n, AD_n, IC_n, AC_n stand for interpretable discrete, abstract discrete, interpretable continuous and abstract continuous latent variables in cell n, respectively. ID_n and IC_n capture information corresponding to well defined physical features of the cell, like activity strength of certain biological pathway, developmental stage of the cell or clinical diagnose of the cell's donor. AD_n and AC_n capture complementary, yet unknown information in the data, such as unsupervised clustering of cells, or diffusion scores of cells. AD_n and AC_n also play auxiliary roles that help the model reconstruct the original data. The mapping function from latent variable to expression level is learned by training a neural network called decoder, and the posterior distribution of latent variables $q(ID_n, AD_n, IC_n, AC_n|x_n)$ is learned by training another neural network called encoder.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.09.459281; this version posted September 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

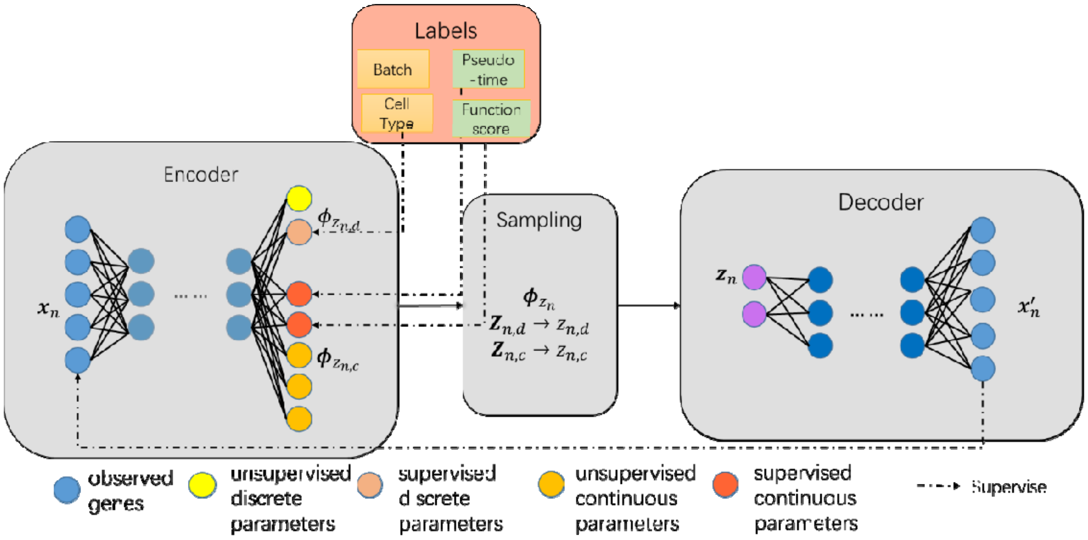

Figure 1. Model illustration of UniCoord. Encoder is an MLP that transfer input scRNA-seq data into parameters of latent distributions. Latent distributions include continuous normal distributions and discrete categorical distributions. In sampling step, reparameterization are performed and random variables are sampled from latent distribution. Decoder then transfer random variables back to input data.

Jointly learning of discrete, continuous coordinates

We tested representation performance of UniCoord in a single nuclear RNA sequencing dataset. Nuclei were sampled from different sites from healthy human heart. Nuclei were classified as 9 distinct major cell types (Fig.2A). These cell type labels were used as discrete coordinates. In previous analysis, we found that cardiomyocytes (CM) from left ventricle (LV) have a gradual functional switch from outside to inside ventricle wall, while these heterogeneities were not found in other ventricle or atrial walls. A layer gradient was assigned to each CM from LV (Fig.2B). Meanwhile, we found that vessel endothelial cells (EC) show an arterial-capillary-vein zonation heterogeneity, and a zonation score was assigned to each vessel endothelial cell (Fig.2C). These two heterogeneity scores were used as continuous coordinates.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.09.459281; this version posted September 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

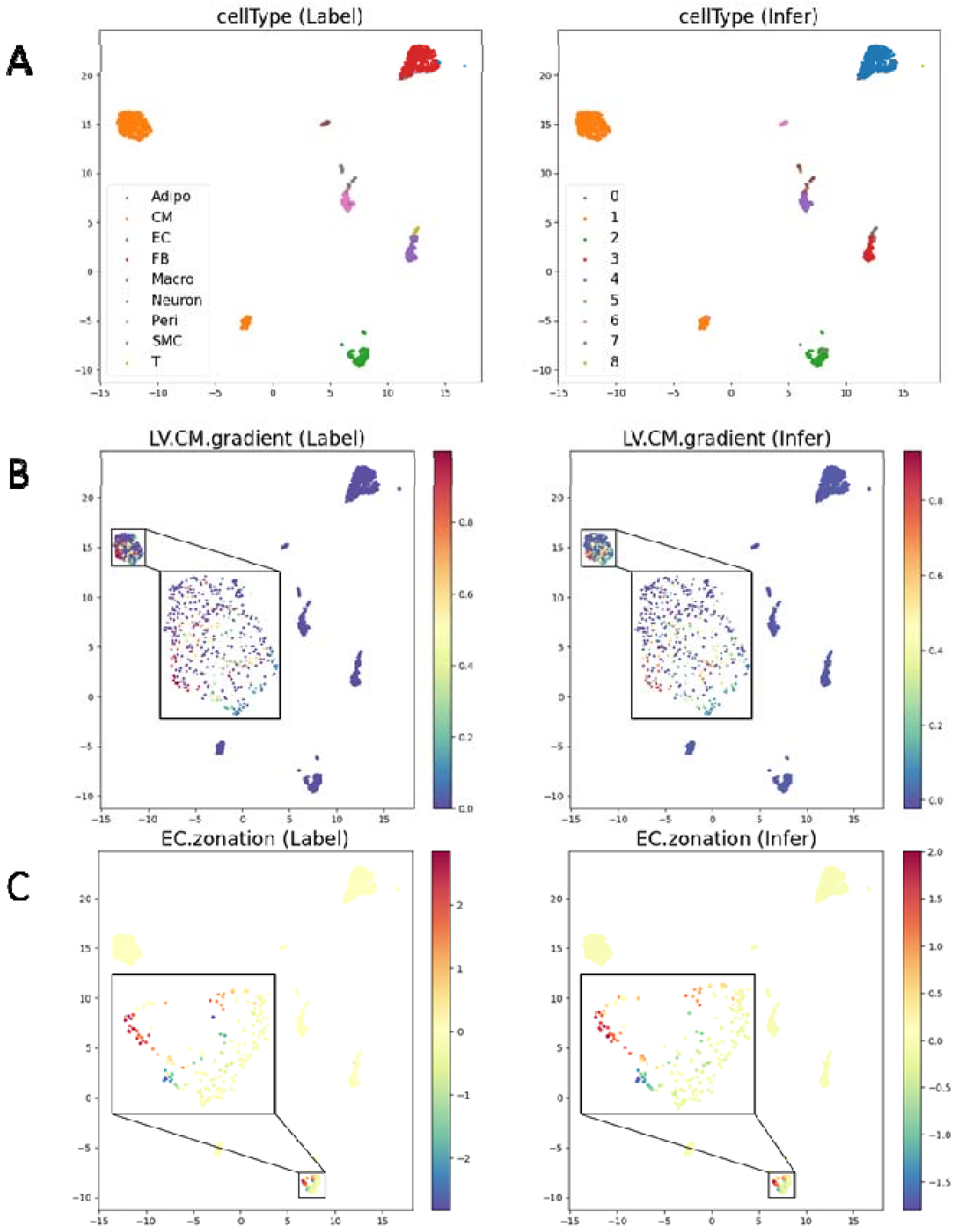

Figure 2. UniCoord simultaneously learned discrete and continuous attributes of adult human cells. Attributes includes cell type label for all the cells(**A**), pseudo-layer score for cardiomyocytes(**B**) and artery-capillary-vein zonation scores for vessel endothelial cells(**C**).

After training UniCoord model with 80% percent of the data, we tested the performance in the remaining 20% test set. UniCoord leaned cell type with ARI=0.977. Simultaneously, spatial gradient of LV CM are learned with , and zonation of vessel EC are learned with . Nuclei out of LV CM should not have spatial gradient so their gradient score were defined zero, so were zonation score out of vessel EC. This feature was also learned.

UniCoord learns hierarchical structure

Cell type are usually organized hierarchically. In practice, most scRNA-seq works annotate their

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.09.459281; this version posted September 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

data in a top-down manner: people study the whole dataset firstly and define major cell types, then study a major cell type more carefully and define its sub-types, and then go on and on. For instance, in an immune landscape study, cells are firstly defined as lymphocytes or myeloid cells, then lymphocytes are separated into B cells and T cells. T cells may also have its subtypes. Thus, all these annotations construct an ontology cell type tree. A universal annotation system need not only annotating cells at different granularity, but also keeping self-consistent of labels. Specifically, it should avoid confliction like a cell are annotated as myeloid cell and T cell simultaneously.

We use PBMC3k dataset as example to show how UniCoord learns hierarchical structure. The dataset has 8 terminal cell types, which can be organized as Figure3B shows. For each layer of the cell type tree, UniCoord learn it with a discrete latent distribution, and the relationship between adjacent layers is also kept in the model.

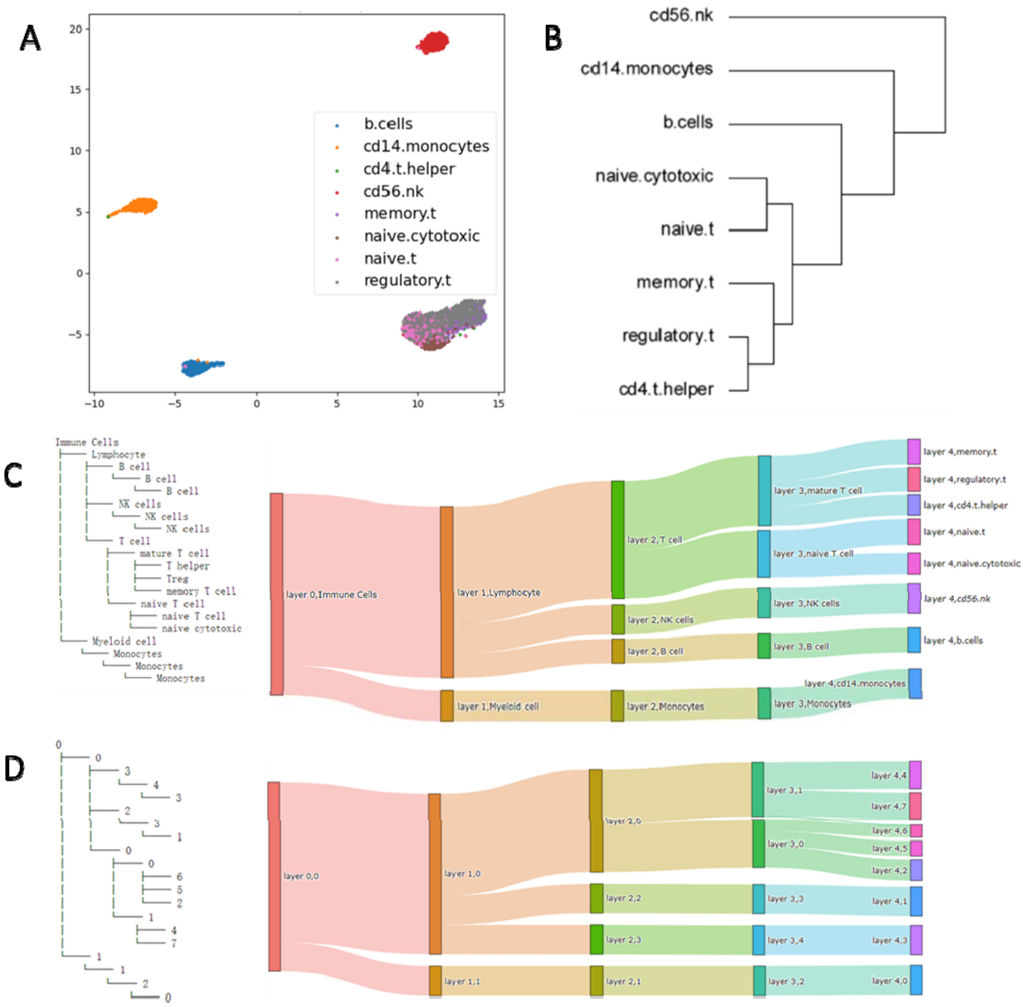

Figure 3. UniCoord learned hierarchical annotation of peri blood monocytes (PBMC). A. UMAP plot of PBMC3k dataset, dots colored by original annotated cell type. **B.** Hierarchical relationship of cell types. **C.** Hierarchical tree of cell types (left) and Sankey plot of original labels (right). **D.** Hierarchical tree of discrete latent dimensionalities (left) and Sankey plot of learned latent labels (right).

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.09.459281; this version posted September 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Generalization ability

The generalization ability denotes how the model will perform when the input data are distinct from the training datasets. In scRNA analysis, new data are continually generated and new data are likely to be different from existing data, although with shared information, due to either specialized experimental design or technical improvement. A unified scRNA data analysis model should be compatible to situations where training data and data analyzed afterward show a certain level of heterogeneity.

Here we showed the generalization ability of UniCoord through predicting cardiomyocytes that not present in the training set. In Fig.2, cells were collected from 4 different layers of 1 human heart left ventricle, from inner to outer layer. CMs between different layers show significant transmural heterogeneity, while changes are putatively continuous and can be aligned into a one-dimensional axis called pseudo-layer score. In Fig.2 we have shown that UniCoord learned this score well when trained by 80% of the data and tested by the other 20%. Here we use cells from layer 1,2 and 4 to train the model and test with cells from layer 3. Fig.4 shows that once trained, UniCoord predict the pseudo-layer score of cells that have not been seen by the model, with larger but reasonable mean error compared with it in Fig.2.

UniCoord Generate simulated scRNA-seq data

As a generative model, one of major using scenario is to generate simulated data. Taking advantage of interpretable representation, UniCoord can generate data with certain property of our interest. We still tested this usage on LVCM data. Using model trained with layer 1,2 and 4 cells and manually change pseudo-cells, a group of simulated cells are generated. The results show that in most part of data where layer 1,2 and 4 cells dominate, reconstructed data shared nearly identical manifold with original data. In the area where layer 3 dominate, a part of simulated data is related with layer 3 cells, while significant difference can also be seen. The simulation function of UniCoord could be used in task like filling unsampled gaps or sparse area in a trajectory, which may be caused by rareness of certain cell type or technical difficulties.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.09.459281; this version posted September 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

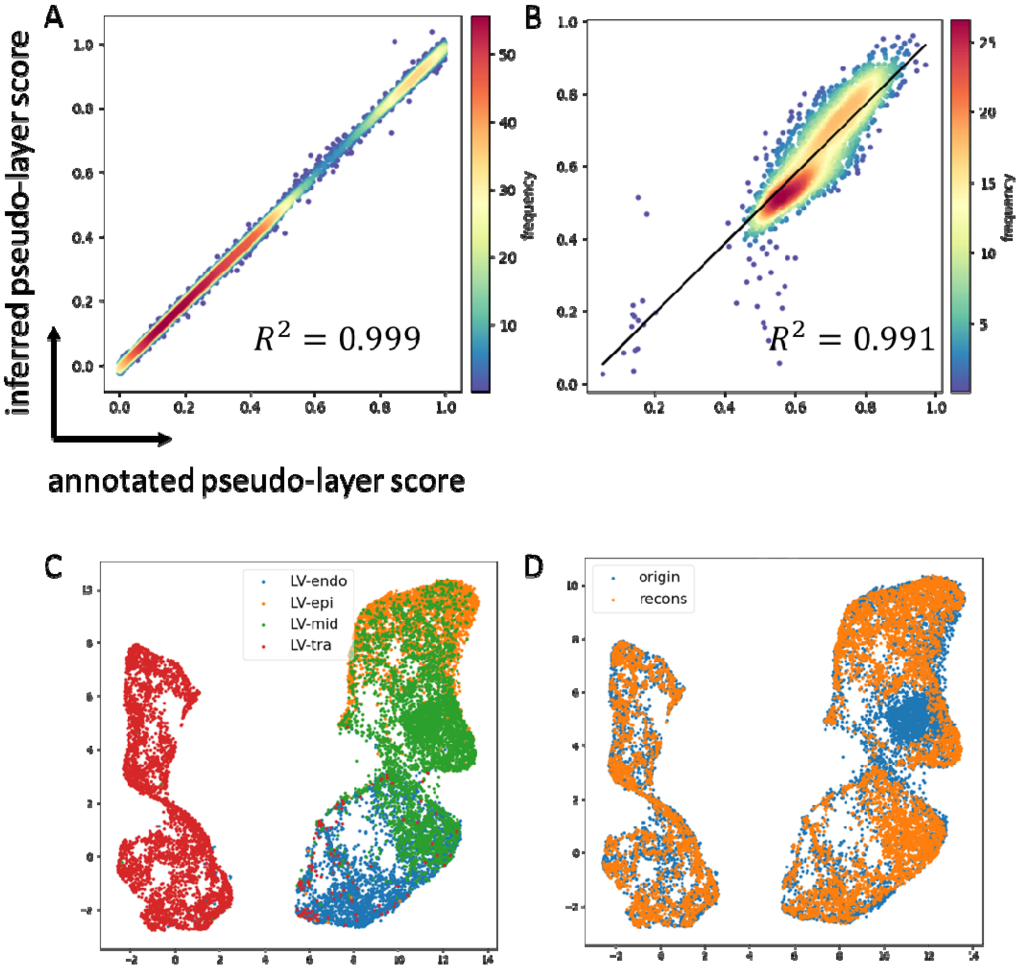

Figure 4. Generalization ability and data simulation of UniCoord. (**A-B**) shows correlation between pseudo-layer score label and inferred score in layer 1,2 and 4 CMs (**A**), and in layer 3 CMs(**B**). (**C-D**) are UMAP plot for a mixture dataset of real data and reconstructed data, colored by their sampled layer(**C**, here LV-endo, LV-epi, LV-mid and LV-tra corresponds to layer 1,2,3 and 4, respectively) and whether the cell are original real data or model reconstructed data(**D**).

Discussion

In this work, we proposed a modified VAE model for embedding scRNA-seq data into a latent space composed of continuous and discrete dimensionalities, both can be supervised by given labels. With proper supervision, the embedding space could become an interpretable coordinate of cells, and UniCoord model provide a convenient and accurate method for obtaining coordinate values.

Compared with popular data embedding methods in scRNA-data analysis, such as scVI and scTransform, UniCoord has several distinct usages due to its idea of representing multifaceted heterogeneities of cells. First, interpretable latent space provides an intuitive understanding when reading the embedded vector of certain cell, while meaningless embedding makes the reading harder. Second, down-stream analysis could be facilitated and possibility of finding biological

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.09.459281; this version posted September 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

discoveries might increase. Because it is easy to highlight certain biological function by weighting corresponding functional score. Third, some of functional coordinates, such as developmental trajectory score, notice areas on atlas where sampled cells are not dense enough. This guide people where to fulfill the atlas. For those samples hard to captured, like some transient states in development, UniCoord can generate simulated data for substitution until scRNA data of missing cells are finally achieved.

It is still a long route towards a comprehensive information system for human cell atlas. Future work includes bringing in meaningful biological function definitions to the model, pretraining the model with large amount of data, seeking for bioinformatic and clinical usages of interpretable cellular representation, and the usages of a cell atlas as a reference.

Methods

Structure and parameter of UniCoord model

The UniCoord model is composed of three parts: two neural network and a random vector sampler. The two networks are both MLP, and are named as encoder and decoder, respectively. In the paper, all encoder has 2 hidden layers with neuron number as (512, 256). The neuron number of input and output layers of encoder are adaptive to gene number of input data, which we choose the top 2,000 highly variable genes, and requested hidden variable, respectively. Shape of a decoder is basically symmetry with its encoder, while input neurons has the same number as latent variable. The sampler is used for reparameterization of latent distribution parameters, which are the output of encoder. While training, Adam algorithm is always used as optimizer, with learning rate as 5e-4.

Reparameterization tricks

While training, reparameterization tricks are used to disentangle random variables with parameters and make back propagation algorithm possible. For continuous latent variable, we kept conventional reparameterization trick used in VAE (Kingma and Welling 2013). For discrete latent variable, we applied Gumbel-softmax reparameterization (Jang, Gu, and Poole 2016).

Loss functions

Loss function of UniCoord is composed of several parts, and each part gives the model a specific feature. In general,

$$\begin{aligned}Loss = {} & \lambda_{recons} * L_{recons} \\ & + \beta * \left(L_{GaussianKL} + L_{CategoryKL}\right) \\ & + \lambda_{diffusion} * L_{diffusion} \\ & + \lambda_{Clustering} * L_{clustering} \\ & + \lambda_{Regression} * L_{regression} \\ & + \lambda_{Classification} * L_{classification} \\ & + \lambda_{structure} * L_{Tree}\end{aligned}$$

all λs are hyper parameters that control weights between each part, and the detail term of each loss will be introduced below.

Reconstruction loss

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.09.459281; this version posted September 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The basic part of losses that make the model an auto-encoder is the reconstruction loss. It is defined as mean square error between reconstructed data and original data. That is

$$L_{recons} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\left(x'_{ij} - x_{ij}\right)^2$$

where n is the number of cells, m is the number of genes, x_{ij} represents the expression level of gene j in cell i, and x'_{ij} represents the reconstructed expression level of gene j in cell i

KL divergence

KL divergence works as regularization component that prevent over-fitting. For continuous dimensions, KL divergence constrain the posterior distribution to be closed to standard Normal distribution $\mathrm{Normal}(0,1)$. For discrete dimensions, KL divergence constrain the posterior distribution to be closed to uniform categorical distribution $\mathrm{Category}(\frac{1}{\mathrm{c}}, \ldots, \frac{1}{\mathrm{c}})$, where c is the number of categories in this dimension.

Diffusion loss

Some of the continuous dimensions can be defined as diffusion dims, playing the same roles as reductions in diffusion map. We want those cells with similar scores at diffusion dims should also be similar in expression level. So, we first construct k-nearest neighbor for all cells, and then for each cell, average of its neighbors' latent distribution will be calculated and input into decoder to generate a reconstruction expression vector x''. Diffusion loss $L_{diffusion}$ is defined as

$$L_{diffusion} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\left(x''_{ij} - x_{ij}\right)^2$$

Clustering loss

Some of the discrete dimensions can be defined as clustering dims. The principle of clustering dims is to keep cells within a cluster have similar expression level, and cells between different clusters have distinct expression level.

$$L_{clustering} = \sum_{\substack{i\in c_1 \\ j\in c_2}}\left(x_i - x_j\right)^2 - \alpha\sum_{i,j\in c}\left(x_i - x_j\right)^2$$

where c means a cluster index, c_1 and c_2 means two different clusters, x_i, x_j means the expression vector of cell i and cell j, respectively, α is a hyper parameter that balance between-cluster loss and within-cluster loss.

Regression loss

The regression loss is MSE loss between label and the mean parameter of corresponding latent continuous dimension.

Classification loss

The classification loss is cross entropy between label and the corresponding latent discrete dimension.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.09.459281; this version posted September 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Hierarchical loss

The hierarchical loss is the cross entropy between 2 latent discrete dimensions that are designed to have hierarchical relationship. The descendant layer labels were first aggregate to ancestor label following the designed relationship. Then the cross entropy between aggregated labels and the model generated ancestor labels is defined as hierarchical loss.

Training of the UniCoord Model

The training of UniCoord model need a specific training dataset. It should have two parts, a gene expression matrix, and a cellular coordinate. The matrix could be given as np.array or sparse matrix format in python, and should be log-normalized gene counts. The coordinates could be given as a data frame, csv file or other excel formats are acceptable. Discrete and continuous could be given in one file, but those discrete coordinates' column names and category orders should be provided, otherwise they might be recognized as continuous coordinates.

Datasets and preprocessing

Datasets used in this paper include single nuclear sequencing of an adult human heart (unpublished), and PBMC3k dataset from cellranger website. Gene count matrices went through cell size normalization, log transformation and highly variable genes selection. A subset of matrix contains only top 2000 highly variable genes were kept for UniCoord training and testing.

Acknowledgement:

This work is supported in part by NSFC grants (61721003, 62050178) and the Tsinghua-Fuzhou Institute of Data Technologies grant (TFIDT 2021003).

Reference:

Diehl, A. D., T. F. Meehan, Y. M. Bradford, M. H. Brush, W. M. Dahdul, D. S. Dougall, Y. He, D. Osumi-Sutherland, A. Ruttenberg, S. Sarntivijai, C. E. Van Slyke, N. A. Vasilevsky, M. A. Haendel, J. A. Blake, and C. J. Mungall. 2016. 'The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability', *J Biomed Semantics*, 7: 44.

Jang, Eric, Shixiang Gu, and Ben Poole. 2016. "Categorical Reparameterization with Gumbel-Softmax." In, arXiv:1611.01144.

Kingma, Diederik P, and Max Welling. 2013. 'Auto-Encoding Variational Bayes', *arXiv: Machine Learning*.

Lopez, R., J. Regier, M. B. Cole, M. I. Jordan, and N. Yosef. 2018. 'Deep generative modeling for single-cell transcriptomics', *Nat Methods*, 15: 1053-58.

Rood, J. E., T. Stuart, S. Ghazanfar, T. Biancalani, E. Fisher, A. Butler, A. Hupalowska, L. Gaffney, W. Mauck, G. Eraslan, J. C. Marioni, A. Regev, and R. Satija. 2019. 'Toward a Common Coordinate Framework for the Human Body', *Cell*, 179: 1455-67.

Rostom, R., V. Svensson, S. A. Teichmann, and G. Kar. 2017. 'Computational approaches for interpreting scRNA-seq data', *FEBS Lett*, 591: 2213-25.

Sokolov, A., E. O. Paull, and J. M. Stuart. 2016. 'One-Class Detection of Cell States in Tumor Subtypes', *Pac Symp Biocomput*, 21: 405-16.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.09.459281; this version posted September 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Stuart, T., A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck, 3rd, Y. Hao, M. Stoeckius, P. Smibert, and R. Satija. 2019. 'Comprehensive Integration of Single-Cell Data', *Cell*, 177: 1888-902 e21.

Trapnell, C., D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, and J. L. Rinn. 2014. 'The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells', *Nat Biotechnol*, 32: 381-86.

Zhang, P., M. Yang, Y. Zhang, S. Xiao, X. Lai, A. Tan, S. Du, and S. Li. 2019. 'Dissecting the Single-Cell Transcriptome Network Underlying Gastric Premalignant Lesions and Early Gastric Cancer', *Cell Rep*, 27: 1934-47 e5.