Machine learning for single-cell genomics data analysis

Félix Raimundo, Laetitia Meng-Papaxanthos, Céline Vallot, Jean-Philippe Vert

1,2,3,4]Félix Raimundo
1]Laetitia Papaxantho
2,3]Céline Vallot
1,4]Jean-Philippe Vert
*Equal contribution
**Corresponding author
1]Google Research, Brain team, Paris, France 2]CNRS UMR3244, Institut Curie, PSL University, Paris, France 3]Translational Research Department, Institut Curie, PSL University, Paris, France 4]MINES ParisTech, PSL University, CBIO-Center for computational biology, Paris, France

Machine learning for single cell genomics data analysis

[

[

April 20, 2021

Abstract

Single-cell omics technologies produce large quantities of data describing the genomic, transcriptomic or epigenomic profiles of many individual cells in parallel. In order to infer biological knowledge and develop predictive models from these data, machine learning (ML)-based model are increasingly used due to their flexibility, scalability, and impressive success in other fields. In recent years, we have seen a surge of new ML-based method development for low-dimensional representations of single-cell omics data, batch normalization, cell type classification, trajectory inference, gene regulatory network inference or multimodal data integration. To help readers navigate this fast-moving literature, we survey in this review recent advances in ML approaches developed to analyze single-cell omics data, focusing mainly on articles published in the last two years (2019-2020).

keyword machine learning, single cell genomics, representation learning, batch correction, data integration, cell type classification, trajectory inference.

Introduction

With single-cell omics technologies getting wide-spread adoption, computational methods are urgently needed to process the large amounts of data they produce [1]. Machine learning (ML) approaches have recently demonstrated their fantastic potential to automatically process and learn from large amounts of high-dimensional data in fields such as computer vision or natural language processing [2]. They are therefore seen by many as a promising way to infer biological knowledge and develop predictive models from single-cell omics data, which provide high-dimensional characterization of large quantities of cells. Not surprisingly, the development of ML approaches to analyze single-cell omics data has been a very active field of research recently.

In this review we survey recent advances in ML approaches developed to analyze single-cell transcriptomic and epigenomic data, focusing mainly on

Figure 1:Standard analysis pipelines using a single modality of single-cell omics data start by turning the raw sequencing reads into a matrix of cells×feature counts. This matrix is then used for dimension reduction, representing each cell by a vector of lower dimension (embedding). The embedding is then used as starting point for subsequent tasks such as visualization, cell type discovery, or trajectory inference.

articles published in the last two years (2019-2020). This period witnessed active developments of new methods, in particular based on deep learning, to automatically extract information from large sets of single-cell data, tackling important problems such as batch normalization, multimodal data integration, automatic cell type classification, trajectory inference or gene network reconstruction. It is also a period where systematic benchmarks started to highlight the practical challenges associated to these methods, as well as their potential. With this review we hope to give the reader enough entry points to that fast-moving literature in order to grasp the current state-of-the-art and join its future developments.

From raw data to useful representations

Raw single-cell transcriptomic count data, as well as their epigenomic counterparts, provide a high-dimensional and noisy description of each cell by assessing the activity of thousands of genes or DNA loci simultaneously. Transforming raw count data to a lower-dimensional representation of each cell using dimension reduction (DR) technique is a useful step to remove technical noise and prepare data for visualization, classification or further analysis tasks (Figure 1).

While early and widely-used methods such as scran [3] and Seurat v2 [4] use standard principal component analysis (PCA) on log-transformed count data for DR, many new DR models have been proposed specifically for scRNA-seq data recently. A common theme has been to replace the implicit Gaussian noise assumption of PCA by explicit statistical models of raw count data, modelling for example overdispersion and zero-inflation due to dropout in the matrix factorization-based model ZinbWave [5], or heavy-tailed count distribution in the nonparametric Bayesian model of [6]. Several groups have also investigated the potential of (variational) autoencoders ((V)AE), a very popular class of deep learning-based DR models. In short, a (V)AE learns a low-dimensional representation of input data (cell transcriptomes in our case) that is sufficient to reconstruct the input data, using flexible neural network models to go from the input to the compressed representation (encoding), and from the representation to the input data (decoding). Several (V)AE models for scRNA-seq data have been proposed recently, include scVI [7], DCA [8], SAVER [9] and scVAE [10]. Methods using hyperbolic geometry have also recently been developed [11, 12]. These models differ from each other by some modelling assumptions, such as

the statistical model for count data in the decoder, or the prior distribution of the low-dimensional representation, but otherwise follow a similar architecture. An interesting property of these models is their computational scalability, as they are typically implemented with deep learning libraries designed to train models with millions or more input points.

Have deep learning-based (V)AE definitively imposed themselves as the best DR approach for scRNA-seq data? The answer is not so simple. Besides requiring large number of cells to learn parameters, (V)AE performance was shown to be very sensitive to arbitrary parameter choices [13], and [14] highlighted that with datasets of a few hundreds or thousands cells simpler models remain competitive and easier to use. The practical difficulty to correctly train complex ML models is not specific to (V)AE: another example is the "art of training" the popular t-distributed stochastic neighbour embedding (tSNE) model for visualizing scRNA-seq in two dimensions [15], that requires specific initialization and choices of hyperparameters. Once correctly trained, tSNE reaches the same performance as uniform manifold approximation and projection (UMAP), a model proposed to improve tSNE mapping of scRNA-seq [16, 15]. This highlights, again, both the potential and the difficulty to train some modern ML-based models, and raises in particular important concerns about making sure that all published results are reproducible and not overfitted to a given experiment.

Several DR methods for single-cell epigenomic data have also been proposed recently, either based on standard PCA models [17, 18], matrix factorization with latent Dirichlet allocation [19], or a VAE [20]. A recent benchmark highlights the importance of preprocessing, in particular how reads are binned into regions of interest and counted, for the success of these methods [21].

One interesting idea to use complex models on small datasets is to leverage larger, already annotated, datasets to learn the embedding, using techniques from the field of transfer learning or domain adaptation. Embeddings learned by PCA and non-negative matrix factorisation (NMF) on datasets such as the Human Cell Atlas (HCA) have successfully been used in both scATAC-seq [22] and scRNA-seq [23, 24] on new unseen datasets and cell types, as well as used for denoising the new dataset [25]. Similarly the embeddings learned by (demoising) AEs on one dataset, have been shown to be useful on other datasets, both for clustering [26, 27, 28, 29] and surface protein prediction [30]. One limitation of these methods is that the embedding is only learned on a single dataset, and applied to another dataset, without analyzing both in parallel. This limits the ability to train models on multiple datasets and thus truly leverage the mass of experiments in databases such as HCA.

The result of the DR is often fed to standard clustering algorithms, as reviewed in [31], in order to identify cell types, with these algorithms also being extremely sensitive to hyperparameter choices [32]. Once the cells are clustered, differential expression tools, benchmarked in [33], can be used to identify *de novo* marker genes.

The cells can also be matched to known cell types either by querying a reference database with tools such as Cell BLAST [34], scMap [35], scQuery

Figure 2:Different experiments of a similar modality (e.g., scRNA-seq) containing different number of cells can be integrated into a single unified view. At first, cells of the same type are separated by their batch, but after correction are perfectly merged together.

[36] or CellFishing.jl [37] or by using standard supervised learning techniques as benchmarked in [38]. However these methods can be sensitive to batch effects, whose corrections are the subject of the following section.

Batch correction and integration of heterogeneous scRNA-seq data

Instead of analyzing data of a single experiment, much can be gained by jointly analyzing single-cell transcriptomic data of many experiments, potentially coming from different labs, using different technologies, and following different experimental protocols. ML models are likely to benefit from analyzing more cells, but the risk of capturing batch effects and other confounding factors instead of biological knowledge is large and considered one of the grand challenges of scRNA-seq data analysis [1]. A number of models have been proposed to specifically perform jointly DR on heterogeneous scRNA-seq data, build a global graph or construct a common gene expression matrix, aiming to capture biology and ignore confounding effects (see Figure 2 and [39] for a comprehensive benchmark).

A first group of models learn a low-dimensional representation over a common space that is invariant to technical confounders. Among those, SAUCIE [40] and scDGN [41] are deep-learning based, SAUCIE is an AE trained with a specific regularisation penalty on the latent codes to remove batch effects, and scDGN is a supervised adversarial neural network model trained to accurately classify cell types and discriminate against batches. scMC [42], Harmony [43] and SMNN [44] rely on a linear transformation to a lower dimensional space, clustering (shared nearest neighbour scheme, soft k-means or supervised mutual nearest neighbours) and post-processing of the low dimensional embeddings to both account for cell-cell similarities and remove batch-specific variations. Other models have an objective to build a joint graph connecting all measured cells, such as scPopCorn [45] which relies on PageRank and graph-k partitioning, and Conos [46] which exploits cell-cell similarity matrices and mutual nearest neighbours. These graph-based models allow for tasks such as cell annotation and information propagation along the network. However, the methods previously described hinder interpretability as they do not enable studying differentially expressed genes leveraging the multiple datasets. A third group of models attempt to tackle this problem by correcting for batch effects on the original count data. Among them, scAlign [47] uses paired AEs with a common latent space that conserves the cell-cell distances

estimated in the count data, while BERMUDA [48] instead uses a regularisation penalty on cell clusters from different batches in the latent space, and scGen [49] combines VAEs and latent space vector arithmetics. scVI [7] and trVAE [50] are so-called *conditional* VAE approaches that condition the decoder on an auxiliary batch variable to correct the data in the latent space. Based on variants of nearest neighbour search, scMerge [51] combines mutual nearest clusters and RUV-III factor analysis to remove unwanted factors from the count data, and Scanorama [52] and Seurat v3 [53] rely on linear projection to a low-dimensional space and an efficient (mutual) nearest neighbour search to obtain matched cells in low-dimensional space that are used to build translation vectors in the high-dimensional space.

All methods cited above offer batch correction for scRNA-seq data, while scMC has also been proposed for scATAC-seq integration and SAUCIE for single-cell CyTOF measurements. While most methods need shared cell types across datasets to build anchor cells, SAUCIE, scPopCorn and scMerge can be used without. Finally, almost half of the methods are able to scale to datasets containing hundreds of thousands of cells.

Learning trajectories, dynamics and regulation

Besides capturing the cellular heterogeneity of tissues and identifying cell types, single-cell omics data offers the possibility to learn about *dynamical* processes that shape this heterogeneity, such as cell cycle, differentiation, proliferation or tumorigenesis. From a data analytical point of view, this raises the question of inferring a dynamical model or at least the cellular trajectories from a snapshot of cells scattered at different time points along the dynamics. Since the first algorithms such as Monocle [54] were published in 2014 to infer trajectories and order cells using the notion of pseudotime, dozens of methods have been proposed. Recently proposed methods include GrandPrix [55], an efficient implementation of the Gaussian process latent variable model (GPLVM) to estimate pseudotimes and their uncertainty; STREAM [56], which estimates a low-dimensional set of curves, called the principal graph, to describe the cells' pseudotime, trajectories and branching points; PAGA [57], a graph-based method to compute a graph representation of a set of cells, allowing visualization and dynamical interpretation at different resolutions; TinGa [58], which builds a graph to fit the single-cell omics data as well as possible using the Growing Neural Graph (GNG) algorithm; or Monocle 3 [59], the latest version of Monocle with new features such as learning trajectories with loops or point of convergence and better scalability. To help users choose a particular method for a given problem, [60] published an impressive benchmark of trajectory inference methods, comparing 45 published algorithms on 110 real and 229 synthetic datasets. While no clear winner emerges in all situations, the benchmark is useful to understand the strengths and weaknesses of different methods in different settings.

Figure 3:Single-cell modalities can take various forms, such as DNA, DNA methylation, CRISPR perturbations, transcriptomics, proteomics or chromatin accessibility. ML models developed for single-cell multimodal data integration assume that the correspondences between cells are either known (co-assay data) or not (non co-assay data) across modalities. In the case of non co-assay data, additional supervision signal might be used, such as cell types, correspondences between features or anchor cells.

A related problem is to infer the relationships between populations of cells captured at different time points along a dynamic process, such as developmental processes after induced pluripotent stem cell reprogramming observed through scRNA-seq profiles captured at half-day intervals [61]. In that paper the authors develop a method, called Waddington-OT, to relate the populations of cells at different time points using the concepts and tools of optimal transport (OT), a mathematically well-established and fast-growing field in ML [62], particularly well adapted to compare populations of cells and model their evolution. With ImageAEOT, [63] show how OT combined with an autoencoder allows to predict the lineages of cells using time-labeled single-cell images.

While trajectory inference implicitly allows us to predict the future evolution of cells, some algorithms have also been proposed to explicitly infer the *velocity* of each individual cell's transcriptomic profile. Following the pioneering work of [64], [65] proposed scVelo, a likelihood-based dynamical model for velocity inference from the ratio of spliced and unspliced mRNA. [66] propose another kernel-based velocity estimator, and show how gene regulatory networks (GRN) can be automatically inferred, although with modest accuracy, by training a sparse regression model to predict the velocity from gene expression levels. Another recent attempt to reconstruct GRN and more general gene networks from scRNA-seq data with an ML approach is the convolutional neural network for coexpression (CNNC) approach of [67], who represent each gene pair as a scatter plot of their expression levels across cells and train a standard CNN for 2D images on the resulting plots to learn pairwise relationships.

Multimodal data integration

An important problem in single-cell omics data analysis is to integrate several modalities together, in order to enhance the performance of downstream tasks such as cell type labelling, identification of subpopulations, visualisation or regulatory network inference, as reviewed in [68, 69]. Several ML approaches have been developed for that purpose, for instance by characterizing cells across measurements, projecting multiple measurements into a common latent space or learning the missing modalities. Transcriptomics is typically one of the modalities that is integrated, together with chromatin accessibility [70, 53, 71], DNA [72], DNA methylation [73, 53], proteomic data [74, 75, 70, 76, 77] or CRISPR perturbations [78, 79].

A first category of models assume that the correspondences between cells are known across modalities, with direct applications to co-assay data (Figure 3). Such methods learn a joint representation of each cell or a cell-cell similarity matrix that is used for downstream analyses by exploiting variants of VAEs such as totalVI [80] and scMVAE [71], matrix factorisation-based models such as scAI [77] and MOFA+ [81], or k-nearest neighbour prediction to learn cell-specific modality weights as Seurat v4 [82]. A second category of models do not require co-assays within individual cells and can be applied to independent multi-omics datasets originating from different cells. Current deep learning-based methods either rely on a pair of VAEs whose latent spaces are coupled through a specific penalty (K. D. Yang et al., arxiv.org/abs/1902.03515), or on learning low-dimensional representations minimising a tSNE loss for each view, coupled through a learned matching matrix (UnionCOM [75]). Other methods rely on NMF, to learn a low-dimensional space composed of specific and common factors (LIGER [76]), or cluster representatives of subpopulations of cells (DC3 [83]). MMD-MA [74, 84] learns a joint latent representation where different modalities have a similar distribution using the theory of kernel methods. SCOT [70] uses OT to learn a joint distribution between cells from both views. clonealign [72] models the association between copy number features and gene expression leveraging mean field variational Bayes inference. While these methods can in theory be applied to any bi-modal omics dataset, hyperparameter selection is difficult when no co-assay data is available for MMD-MA, SCOT and UnionCOM. Among models that do not require co-assay data, some use weak supervision such as SCIM [73], an adversarial AE model that assumes that the cell types are known for a fraction of the cells and Seurat v3 [53], a canonical correlation analysis (CCA)-based model that relies on building anchor cells using mutual nearest neighbours. Applied to single-cell CRISPR screenings, scMAGeCK [79] relies on statistical analyses and MUSIC [78] on topic modeling in order to link gene perturbations to cell phenotype. Finally, it is worth mentioning that some models require features to have a one-to-one correspondence between views [72, 53, 76, 78, 79], which may not be the case systematically.

While the diversity of models is large, most of them rely on finding a joint low-dimensional space that can be later used on downstream tasks. Most models combine two modalities and a few enable the integration of more than two, such as UnionCOM, MOFA+ and DC3, the latter also incorporating scHiC or bulk HiChIP datasets. Finally, the scalability of the models evolve conjointly with single-cell technologies, nowadays being able to handle tens or hundreds of thousands of cells [71, 73].

Conclusion

Researchers are facing an exponential growth of approaches to deal with single-cell genomics data, with over 800 tools (scrna-tools.org) published for scRNA-seq analysis so far, many of which being based on ML approaches. A vast

majority of ML-based tools have been straightforwardly imported from other fields, with some features unsuited for genomic challenges and to the reality of biological data - thereby not maximising their performance. In particular, a number of parameters, which have a strong impact on performance, need extensive training to be properly tuned, which is often unrealistic in the case of genomic data. It also raises questions of reproducibility that the scientific community should address, defining for example the processed datasets and variables that should be shared, i.e., random seed values or reduced dimensional spaces, in addition to the raw data. Whether ML models will in the near future make up for the current technical limitations of single cell genomics approaches - e.g dropouts, batch effects - remains uncertain. If current single-cell omics achieve genome-wide characterization of the transcriptome or epigenomes for example, these methods do not yet achieve single-locus/single-cell resolution due to the dropouts within datasets, leaving room for experimental and computational optimisation.

Declaration of interest

FR, LP and JPV are employees of Google France.

References and recommended reading

[1] Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D.J., Hicks, S.C., Robinson, M.D., Vallejos, C.A., Campbell, K.R., Beerenwinkel, N., Mahfouz, A., et al. (2020). Eleven grand challenges in single-cell data science. Genome Biol. *21*, 31. doi:10.1186/s13059-020-1926-6.

[2] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature *521*, 436–444. doi:10.1038/nature14539.

[3] Lun, A.T.L., McCarthy, D.J., and Marioni, J.C. (2016). A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. F1000Research *5*. doi:10.12688/f1000research.9501.2.

[4] Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat. Biotechnol. *36*, 411–420. doi:10.1038/nbt.4096.

[5] Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S., and Vert, J.P. (2018). A general and flexible method for signal extraction from single-cell RNA-seq data. Nat. Commun. *9*, 284. doi:10.1038/s41467-017-02554-5.

[6] Verma, A. and Engelhardt, B.E. (2020). A robust nonlinear low-dimensional manifold for single cell RNA-seq data. BMC Bioinf. *21*, 324. doi:10.1186/s12859-020-03625-z.

[7] Lopez, R., Regier, J., Cole, M.B., Jordan, M.I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. Nat. Methods *15*, 1053–1058. doi:10.1038/s41592-018-0229-2.

•This paper is the first variational autoencoder specifically designed to model scRNA-seq data and map them to a low-dimensional representation.

[8] Eraslan, G., Simon, L.M., Mircea, M., Mueller, N.S., and Theis, F.J. (2019). Single-cell RNA-seq denoising using a deep count autoencoder. Nat. Commun. *10*, 1–14. doi:10.1038/s41467-018-07931-2.

[9] Huang, M., Wang, J., Torre, E., Dueck, H., Shaffer, S., Bonasio, R., Murray, J.I., Raj, A., Li, M., and Zhang, N.R. (2018). SAVER: gene expression recovery for single-cell rna sequencing. Nat. Methods *15*, 539–542. doi:10.1038/s41592-018-0033-z.

[10] Grønbech, C.H., Vording, M.F., Timshel, P.N., Sønderby, C.K., Pers, T.H., and Winther, O. (2020). scVAE: Variational auto-encoders for single-cell gene expression data. Bioinformatics *36*, 4415–4422. doi:10.1093/bioinformatics/btaa293.

[11] Klimovskaia, A., Lopez-Paz, D., Bottou, L., and Nickel, M. (2020). Poincaré maps for analyzing complex hierarchies in single-cell data. Nat. Commun. *11*, 1–9. doi:10.1038/s41467-020-16822-4.

•This paper explores the use of non-Euclidean hyperbolic geometry to embed complex hierarchical data in two dimensions, while preserving the pairwise distances between points in the hierarchy. It is a promising approach to visualize cell trajectories and detect lineages in scRNA-seq datasets.

[12] Ding, J. and Regev, A. (2019). Deep generative model embedding of single-cell RNA-Seq profiles on hyperspheres and hyperbolic spaces. bioRxiv doi:10.1101/853457.

[13] Hu, Q. and Greene, C.S. (2019). Parameter tuning is a key part of dimensionality reduction via deep variational autoencoders for single cell RNA transcriptomics. Pac. Symp. Biocomput. *24*, 362–373.

[14] Raimundo, F., Vallot, C., and Vert, J.P. (2020). Tuning parameters of dimensionality reduction methods for single-cell RNA-seq analysis. Genome Biol. *21*, 212. doi:10.1186/s13059-020-02128-7.

[15] Kobak, D. and Berens, P. (2019). The art of using t-SNE for single-cell transcriptomics. Nat. Commun. *10*, 5416. doi:10.1038/s41467-019-13056-x.

[16] Becht, E., McInnes, L., Healy, J., Dutertre, C.A., Kwok, I.W.H., Ng, L.G., Ginhoux, F., and Newell, E.W. (2019). Dimensionality reduction for visualizing single-cell data using UMAP. Nat. Biotechnol. *37*, 38–44. doi:10.1038/nbt.4314.

[17] Cusanovich, D.A., Reddington, J.P., Garfield, D.A., Daza, R.M., Aghamirzaie, D., Marco-Ferreres, R., Pliner, H.A., Christiansen, L., Qiu, X., Steemers, F.J., et al. (2018). The cis-regulatory dynamics of embryonic development at single-cell resolution. Nature *555*, 538–542. doi:10.1038/nature25981.

[18] Prompsy, P., Kirchmeier, P., Marsolier, J., Deloger, M., Servant, N., and Vallot, C. (2020). Interactive analysis of single-cell epigenomic landscapes with ChromSCape. Nat. Commun. *11*, 1–9. doi:10.1038/s41467-020-19542-x.

[19] González-Blas, C.B., Minnoye, L., Papasokrati, D., Aibar, S., Hulselmans, G., Christiaens, V., Davie, K., Wouters, J., and Aerts, S. (2019). cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. Nat. Methods *16*, 397–400. doi:10.1038/s41592-019-0367-1.

[20] Xiong, L., Xu, K., Tian, K., Shao, Y., Tang, L., Gao, G., Zhang, M., Jiang, T., and Zhang, Q.C. (2019). SCALE method for single-cell ATAC-seq analysis via latent feature extraction. Nat. Commun. *10*, 1–10. doi:10.1038/s41467-019-12630-7.

[21] Chen, H., Lareau, C., Andreani, T., Vinyard, M.E., Garcia, S.P., Clement, K., Andrade-Navarro, M.A., Buenrostro, J.D., and Pinello, L. (2019). Assessment of computational methods for the analysis of single-cell ATAC-seq data. Genome Biol. *20*, 1–25. doi:10.1186/s13059-019-1854-5.

•This benchmark goes beyong simply running the methods and explores the influence of the design choices in the preprocessing steps as well as their large importance.

[22] Erbe, R., Kessler, M.D., Favorov, A.V., Easwaran, H., and Gaykalova, D. A .and Fertig, E.J. (2020). Matrix factorization and transfer learning uncover regulatory biology across multiple single-cell ATAC-seq data sets. Nucleic Acids Res. *48*, e68–e68. doi:10.1093/nar/gkaa349.

[23] Stein-O'Brien, G.L., Clark, B.S., Sherman, T., Zibetti, C., Hu, Q., Sealfon, R., Liu, S., Qian, J., Colantuoni, C., Blackshaw, S., et al. (2019). Decomposing cell identity for transfer learning across cellular measurements, platforms, tissues, and species. Cell Syst. *8*, 395–411. doi:10.1016/j.cels.2019.04.004.

[24] Sharma, G., Colantuoni, C., Goff, L.A., Fertig, E.J., and Stein-O'Brien, G. (2020). projectR: an R/Bioconductor package for transfer learning via PCA, NMF, correlation and clustering. Bioinformatics *36*, 3592–3593. doi:10.1093/bioinformatics/btaa183.

[25] Mieth, B., Hockley, J.R.F., Görnitz, N., Vidovic, M.M.C., Müller, K.R., Gutteridge, A., and Ziemek, D. (2019). Using transfer learning from prior reference knowledge to improve the clustering of single-cell RNA-seq data. Sci. Rep. *9*, 20353. doi:10.1038/s41598-019-56911-z.

[26] Lin, C., Jain, S., Kim, H., and Bar-Joseph, Z. (2017). Using neural networks for reducing the dimensions of single-cell RNA-seq data. Nucleic Acids Res. *45*, e156–e156. doi:10.1093/nar/gkx681.

[27] Wang, J., Agarwal, D., Huang, M., Hu, G., Zhou, Z., Ye, C., and Zhang, N.R. (2019). Data denoising with transfer learning in single-cell transcriptomics. Nat. Methods *16*, 875–878. doi:10.1038/s41592-019-0537-1.

•This paper was the first to show convincing performance gains from pretraining neural networks for scRNA-seq data.

[28] Badsha, M.B., Li, R., Liu, B., Li, Y.I., Xian, M., Banovich, N.E., and Fu, A.Q. (2020). Imputation of single-cell gene expression with an autoencoder neural network. Quantitative Biology *8*, 78–94. doi:10.1007/s40484-019-0192-7.

[29] Hu, J., Li, X., Hu, G., Lyu, Y., Susztak, K., and Li, M. (2020). Iterative transfer learning with neural network for clustering and cell type classification in single-cell RNA-seq analysis. Nat. Mach. Intell. *2*, 607–618. doi:10.1038/s42256-020-00233-7.

[30] Zhou, Z., Ye, C., Wang, J., and Zhang, N.R. (2020). Surface protein imputation from single cell transcriptomes by deep neural networks. Nat. Commun. *11*, 651. doi:10.1038/s41467-020-14391-0.

[31] Petegrosso, R., Li, Z., and Kuang, R. (2020). Machine learning and statistical methods for clustering single-cell RNA-sequencing data. Briefings Bioinf. *21*, 1209–1223. doi:10.1093/bib/bbz063.

[32] Krzak, M., Raykov, Y., Boukouvalas, A., Cutillo, L., and Angelini, C. (2019). Benchmark and parameter sensitivity analysis of single-cell RNA sequencing clustering methods. Front. Genet. *10*, 1253. doi:10.3389/fgene.2019.01253.

[33] Vieth, B., Parekh, S., Ziegenhain, C., Enard, W., and Hellmann, I. (2019). A systematic evaluation of single cell RNA-seq analysis pipelines. Nat. Commun. *10*, 4667. doi:10.1038/s41467-019-12266-7.

[34] Cao, Z.J., Wei, L., Lu, S., Yang, D.C., and Gao, G. (2020). Searching large-scale scRNA-seq databases via unbiased cell embedding with Cell BLAST. Nat. Commun. *11*, 3458. doi:10.1038/s41467-020-17281-7.

[35] Kiselev, V.Y., Yiu, A., and Hemberg, M. (2018). scmap: projection of single-cell RNA-seq data across data sets. Nat. Methods *15*, 359–362. doi:10.1038/nmeth.4644.

[36] Alavi, A., Ruffalo, M., Parvangada, A., Huang, Z., and Bar-Joseph, Z. (2018). A web server for comparative analysis of single-cell RNA-seq data. Nat. Commun. *9*, 4768. doi:10.1038/s41467-018-07165-2.

[37] Sato, K., Tsuyuzaki, K., Shimizu, K., and Nikaido, I. (2019). CellFishing.jl: an ultrafast and scalable cell search method for single-cell RNA sequencing. Genome Biol. *20*, 31. doi:10.1186/s13059-019-1639-x.

[38] Abdelaal, T., Michielsen, L., Cats, D., Hoogduin, D., Mei, H., Reinders, M.J.T., and Mahfouz, A. (2019). A comparison of automatic cell identification methods for single-cell RNA sequencing data. Genome Biol. *20*, 194. doi:10.1186/s13059-019-1795-z.

[39] Tran, H.T.N., Ang, K.S., Chevrier, M., Zhang, X., Lee, N.Y.S., Goh, M., and Chen, J. (2020). A benchmark of batch-effect correction methods for single-cell RNA sequencing data. Genome Biol. *21*, 1–32. doi:10.1186/s13059-019-1850-9.

[40] Amodio, M., van Dijk, D., Srinivasan, K., Chen, W.S., Mohsen, H., Moon, K.R., Campbell, A., Zhao, Y., Wang, X., Venkataswamy, M., et al. (2019). Exploring single-cell data with deep multitasking neural networks. Nat. Methods *16*, 1139–1145. doi:10.1038/s41592-019-0576-7.

•Efficient autoencoder-based method that is able to process multiple heterogeneous single-cell datasets to learn low-dimensional representations invariant to batch effects.

[41] Ge, S., Wang, H., Alavi, A., Xing, E., and Bar-Joseph, Z. (2021). Supervised adversarial alignment of single-cell RNA-seq data. J. Comput. Biol. doi:10.1089/cmb.2020.0439.

[42] Zhang, L. and Nie, Q. (2021). scMC learns biological variation through the alignment of multiple single-cell genomics datasets. Genome Biol. *22*, 1–28. doi:10.1186/s13059-020-02238-2.

[43] Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P., and Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. Nat. Methods *16*, 1289–1296. doi:10.1038/s41592-019-0619-0.

[44] Yang, Y., Li, G., Qian, H., Wilhelmsen, K.C., Shen, Y., and Li, Y. (2020). SMNN: batch effect correction for single-cell RNA-seq data via supervised mutual nearest neighbor detection. Brief. Bioinform. doi:10.1093/bib/bbaa097.

[45] Wang, Y., Hoinka, J., and Przytycka, T.M. (2019). Subpopulation detection and their comparative analysis across single-cell experiments with scPopCorn. Cell Systems *8*, 506–513. doi:10.1016/j.cels.2019.05.007.

[46] Barkas, N., Petukhov, V., Nikolaeva, D., Lozinsky, Y., Demharter, S., Khodosevich, K., and Kharchenko, P. (2019). Joint analysis of heterogeneous single-cell RNA-seq dataset collections. Nat. Methods *16*, 695–698. doi:10.1038/s41592-019-0466-z.

[47] Johansen, N. and Quon, G. (2019). scAlign: a tool for alignment, integration, and rare cell identification from scRNA-seq data. Genome Biol. *20*, 1–21. doi:10.1186/s13059-019-1766-4.

[48] Wang, T., Johnson, T.S., Shao, W., Lu, Z., Helm, B.R., Zhang, J., and Huang, K. (2019). BERMUDA: a novel deep transfer learning method for single-cell RNA sequencing batch correction reveals hidden high-resolution cellular subtypes. Genome Biol. *20*, 165. doi:10.1186/s13059-019-1764-6.

[49] Lotfollahi, M., Wolf, F.A., and Theis, F.J. (2019). scGen predicts single-cell perturbation responses. Nat. Methods *16*, 715–721. doi:10.1038/s41592-019-0494-8.

[50] Lotfollahi, M., Naghipourfar, M., Theis, F.J., and Wolf, F.A. (2020). Conditional out-of-distribution generation for unpaired data using transfer VAE. Bioinformatics *36*, i610–i617. doi:10.1093/bioinformatics/btaa800.

[51] Lin, Y., Ghazanfar, S., Wang, K.Y.X., Gagnon-Bartsch, J.A., Lo, K.K., Su, X., Han, Z.G., Ormerod, J.T., Speed, T.P., Yang, P., et al. (2019). scMerge leverages factor analysis, stable expression, and pseudoreplication to merge multiple single-cell RNA-seq datasets. Proc. Natl. Acad. Sci. *116*, 9775–9784. doi:10.1073/pnas.1820006116.

[52] Hie, B., Bryson, B., and Berger, B. (2019). Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. Nat. Biotechnol. 37, 685–691. doi:10.1038/s41587-019-0113-3.

••Scalable method that integrates heterogeneous single-cell RNA-seq data and performs batch correction on the gene expression matrices.

[53] Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papelexi, E., III, W.M.M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive integration of single-cell data. Cell *177*, 1888–1902. doi:10.1016/j.cell.2019.05.031.

[54] Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N.J., Livak, K.J., Mikkelsen, T.S., and Rinn, J.L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat. Biotechnol. *32*, 381–6. doi:10.1038/nbt.2859.

[55] Ahmed, S., Rattray, M., and Boukouvalas, A. (2019). GrandPrix: scaling up the Bayesian GPLVM for single-cell data. Bioinformatics *35*, 47–54. doi:10.1093/bioinformatics/bty533.

[56] Chen, H., Albergante, L., Hsu, J.Y., Lareau, C.A., Lo Bosco, G., Guan, J., Zhou, S., Gorban, A.N., Bauer, D.E., Aryee, M.J., et al. (2019). Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM. Nat. Commun. *10*, 1903. doi:10.1038/s41467-019-09670-4.

[57] Wolf, F.A., Hamey, F.K., Plass, M., Solana, J., Dahlin, J.S., Göttgens, B., Rajewsky, N., Simon, L., and Theis, F.J. (2019). PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. Genome Biol. *20*, 59. doi:10.1186/s13059-019-1663-x.

[58] Todorov, H., Cannoodt, R., Saelens, W., and Saeys, Y. (2020). TinGa: fast and flexible trajectory inference with growing neural gas. Bioinformatics *36*, i66–i74. doi:10.1093/bioinformatics/btaa463.

[59] Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D.M., Hill, A.J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F.J., et al. (2019). The single-cell transcriptional landscape of mammalian organogenesis. Nature *566*, 496–502. doi:10.1038/s41586-019-0969-x.

[60] Saelens, W., Cannoodt, R., Todorov, H., and Saeys, Y. (2019). A comparison of single-cell trajectory inference methods. Nat. Biotechnol. *37*, 547–554. doi:10.1038/s41587-019-0071-9.

•This extensive benchmarks highlights the strengths and weaknesses of 45 trajectory inference methods according to different metrics and settings.

[61] Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Gould, J., Liu, S., Lin, S., Berube, P., et al. (2019). Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. Cell 176, 928–943.e22. doi:10.1016/j.cell.2019.01.006.

••This paper introduces optimal transport (OT) as a promising approach to compare populations of cells and reconstruct their dynamics.

[62] Peyré, G. and Cuturi, M. (2019). Computational optimal transport: With applications to data science. Foundations and Trends® in Machine Learning *11*, 355–607. doi:10.1561/2200000073.

[63] Yang, K.D., Damodaran, K., Venkatachalapathy, S., Soylemezoglu, A.C., Shivashankar, G.V., and Uhler, C. (2020). Predicting cell lineages using autoencoders and optimal transport. PLoS Comput. Biol. *16*, e1007828. doi:10.1371/journal.pcbi.1007828.

[64] La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M.E., Lönnerberg, P., Furlan, A., et al. (2018). RNA velocity of single cells. Nature *560*, 494–498. doi:10.1038/s41586-018-0414-6.

[65] Bergen, V., Lange, M., Peidli, S., Wolf, F.A., and Theis, F.J. (2020). Generalizing RNA velocity to transient cell states through dynamical modeling. Nat. Biotechnol. *38*, 1408–1414. doi:10.1038/s41587-020-0591-3.

[66] Aubin-Frankowski, P.C. and Vert, J.P. (2020). Gene regulation inference from single-cell RNA-seq data with linear differential equations and velocity inference. Bioinformatics *36*, 4774–4780. doi:10.1093/bioinformatics/btaa576.

[67] Yuan, Y. and Bar-Joseph, Z. (2019). Deep learning for inferring gene relationships from single-cell expression data. Proc. Natl. Acad. Sci. U.S.A. *116*, 27151–27158. doi:10.1073/pnas.1911536116.

[68] Efremova, M. and Teichmann, S.A. (2020). Computational methods for single-cell omics across modalities. Nat. Methods *17*, 14–17. doi:10.1038/s41592-019-0692-4.

[69] Ma, A., McDermaid, A., Xu, J., Chang, Y., and Ma, Q. (2020). Integrative methods and practical challenges for single-cell multi-omics. Trends Biotechnol. *38*, 1007–1022. doi:10.1016/j.tibtech.2020.02.013.

[70] Demetci, P., Santorella, R., Sandstede, B., Noble, W.S., and Singh, R. (2020). Gromov-Wasserstein optimal transport to align single-cell multi-omics data. ICML 2020 Workshop on Computational Biology (WCB) doi:10.1101/2020.04.28.066787.

[71] Zuo, C. and Chen, L. (2020). Deep-joint-learning analysis model of single cell transcriptome and open chromatin accessibility data. Briefings Bioinf. doi:10.1093/bib/bbaa287. bbaa287.

[72] Campbell, K.R., Steif, A., Laks, E., Zahn, M., Lai, D., McPherson, A., Farahani, M., Kabeer, F., O'Flanagan, C., Biele, J., et al. (2019). clonealign: Statistical integration of independent single-cell RNA and DNA sequencing data from human cancers. Genome Biol. *20*, 54. doi:10.1186/s13059-019-1645-z.

[73] Stark, S.G., Ficek, J., Locatello, F., Bonilla, X., Chevrier, S., Singer, F., Consortium, T.P., Rätsch, G., and Lehmann, K.V. (2020). SCIM: Universal single-cell matching with unpaired feature sets. Bioinformatics *36*, i919–i927. doi:10.1093/bioinformatics/btaa843.

[74] Liu, J., Huang, Y., Singh, R., Vert, J.P., and Noble, W.S. (2019). Jointly embedding multiple single-cell omics measurements. In 19th International Workshop on Algorithms in Bioinformatics (WABI 2019), *Leibniz International Proceedings in Informatics (LIPIcs)*, volume 143. pp. 10:1–10:13. doi:10.4230/LIPIcs.WABI.2019.10.

[75] Cao, K., Bai, X., Hong, Y., and Wan, L. (2020). Unsupervised topological alignment for single-cell multi-omics integration. Bioinformatics *36*, i48–i56. doi:10.1093/bioinformatics/btaa443.

[76] Welch, J.D., Kozareva, V., Ferreira, A., Vanderburg, C., Martin, C., and Macosko, E.Z. (2019). Single-cell multi-omic integration compares

and contrasts features of brain cell identity. Cell *177*, 1873–1887. doi:10.1016/j.cell.2019.05.006.

[77] Jin, S., Zhang, L., and Nie, Q. (2020). scAI: an unsupervised approach for the integrative analysis of parallel single-cell transcriptomic and epigenomic profiles. Genome Biol. *21*, 1–19. doi:10.1186/s13059-020-1932-8.

[78] Duan, B., Zhou, C., Zhu, C., Yu, Y., Li, G., Zhang, S., Zhang, C., Ye, X., Ma, H., Qu, S., et al. (2019). Model-based understanding of single-cell CRISPR screening. Nat. Commun. *10*, 1–11. doi:10.1038/s41467-019-10216-x.

•This paper proposes to link gene perturbations to cell phenotypes, which is a promising and so far under-exploited task.

[79] Yang, L., Zhu, Y., Yu, H., Cheng, X., Chen, S., Chu, Y., Huang, H., Zhang, J., and Li, W. (2020). scMAGeCK links genotypes with multiple phenotypes in single-cell CRISPR screens. Genome Biol. *21*, 1–14. doi:10.1186/s13059-020-1928-4.

[80] Gayoso, A., Steier, Z., Lopez, R., Regier, J., Nazor, K.L., Streets, A., and Yosef, N. (2021). Joint probabilistic modeling of single-cell multi-omic data with totalVI. Nat. Methods *18*, 272–282. doi:10.1038/s41592-020-01050-x.

[81] Argelaguet, R., Arnol, D., Bredikhin, D., Deloro, Y., Velten, B., Marioni, J.C., and O, S. (2020). MOFA+: a probabilistic framework for comprehensive integration of structured single-cell data. Genome Biol. *21*, 111. doi:10.1186/s13059-020-02015-1.

[82] Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zagar, M., et al. (2020). Integrated analysis of multimodal single-cell data. bioRxiv doi:10.1101/2020.10.12.335331.

[83] Zeng, W., Chen, X., Duren, Z., Wang, Y., Jiang, R., and Wong, W.H. (2019). DC3 is a method for deconvolution and coupled clustering from bulk and single-cell genomics data. Nat. Commun. *10*, 1–11. doi:10.1038/s41467-019-12547-1.

•This paper is one of the few that integrates more that two modalities, one of which being chromatin organization.

[84] Singh, R., Demetci, P., Bonora, G., Ramani, V., Lee, C., Fang, H., Duan, Z., Deng, X., Shendure, J., Disteche, C., et al. (2020). Unsupervised manifold alignment for single-cell multi-omics data. In Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics. pp. 1–10. doi:10.1101/2020.06.13.149195.