IMPRS for Genome Science
INTERNATIONAL MAX PLANCK RESEARCH SCHOOL

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

UNIVERSITÄTSMEDIZIN
GÖTTINGEN
UMG
Institut für Medizinische Informatik

Impact of preprocessing in data integration of single-cell RNA-seq data

Youngjun Park, Anne-Christin Hauschild

IMPRS for Genome Science, Georg-August-Universität Göttingen
Medical Informatics, University Medicine Göttingen

Introduction

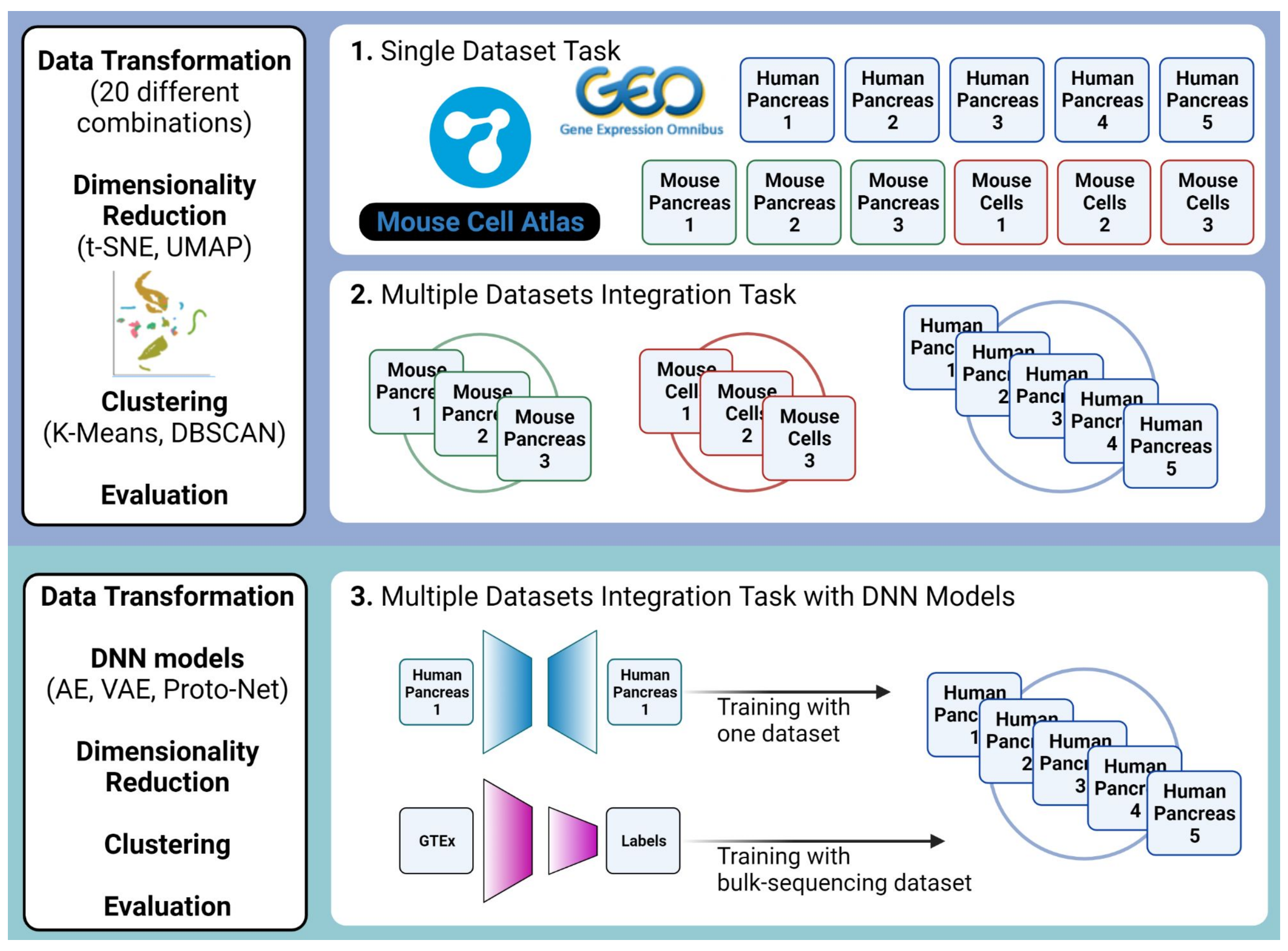
Single-cell sequencing data has heterogeneity resulting from various noise sources due to technical limitations. Over the last years, numerous single-cell data analysis tools have been introduced as de-noising methods. Here, we found that most single-cell studies employed various preprocessing steps without reasoning. This fact is particularly alarming since these read count transformations can significantly alter data distribution and affect downstream cell clustering results.

This study aims to investigate the effects of the various data transformation on public datasets and evaluate them with the most commonly used dimensionality reduction and clustering analysis.

Data transformation methods used in recent studies

Tools	Preprocessing (data transformation used in the study or tool)
scVI	RAW
scLVM	RAW or log-linear fit
scGen	Total → Log
MNN	Deconvolution based normalization → Log
LIGER	Total → Meansquare
scImpute	Total → Log10
Scanorama	l2-norm
scIGANs	Minmax
ComBat-seq	RAW
DESC	Total → Log → Z-score
scMerge	Log → Z-score
scDHA	(Log2) → Minmax
scVAE	RAW
scGNN	Log
ICAnet	Total → Log2
scETM	RAW
iMAP	Log
scBatch	Log, (dataset with ERCC: scPLS) or (Raw → ComBat-seq)
Seurat V2, V3	Total(10000) → Log → Z-score
Harmony	Total(10000) → Log → Z-score
MARS	Total(10000) → Log → min(Z-score, 10)
Luecken2022	(scran) → Log

Dataset and Method


Data Transformation
(20 different combinations)
Dimensionality Reduction
(t-SNE, UMAP)
Clustering
(K-Means, DBSCAN)
Evaluation

1. Single Dataset Task

2. Multiple Datasets Integration Task

Data Transformation
DNN models
(AE, VAE, Proto-Net)
Dimensionality Reduction
Clustering
Evaluation

3. Multiple Datasets Integration Task with DNN Models

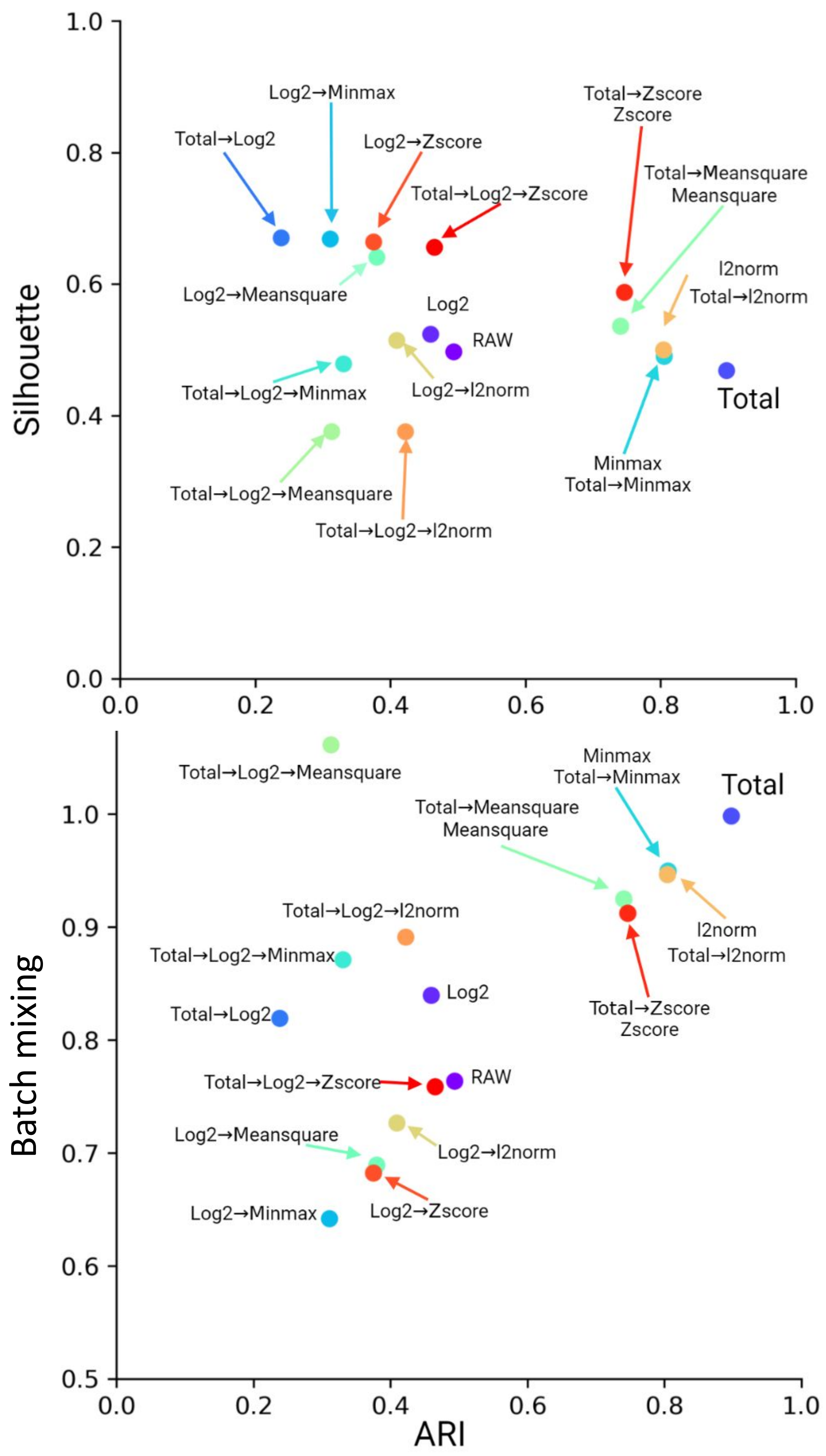
Training with one dataset
Training with bulk-sequencing dataset

Human Pancreas Dataset: (GSE84133, GSE85241, E-MTAB-5061, GSE81608, and GSE83139). Mouse Cell Dataset: Tabula Muris SMART-Seq2 from FACS-sorted cells and 10x Genomics platform with CellRanger.
Basic data transformation methods for the single-cell RNA sequencing data preprocessing are ‘Log2’, ‘Total’, ‘l2-norm’, ‘Meansquare’, ‘Minmax’, and ‘Z-score’.
Single-cell analysis was done with t-SNE, UMAP and K-Means, DBSCAN clustering. The autoencoder has basic encoder and decoder block composed with fully connected layer, batch normalization layer, and relu layer. The encoder/decoder has one hidden layer sized 1024.

Results

Impact of data transformation on the multiple human pancreas datasets integrative task.

We aggregated all five datasets of the human pancreas and preprocessed them with 20 different combinations of data transformation. After that, we applied two different dimensionality reduction methods, t-SNE and UMAP, and two different clustering methods, K-Means and DBSCAN. Here is the detailed investigation of the UMAP+DBSCAN.



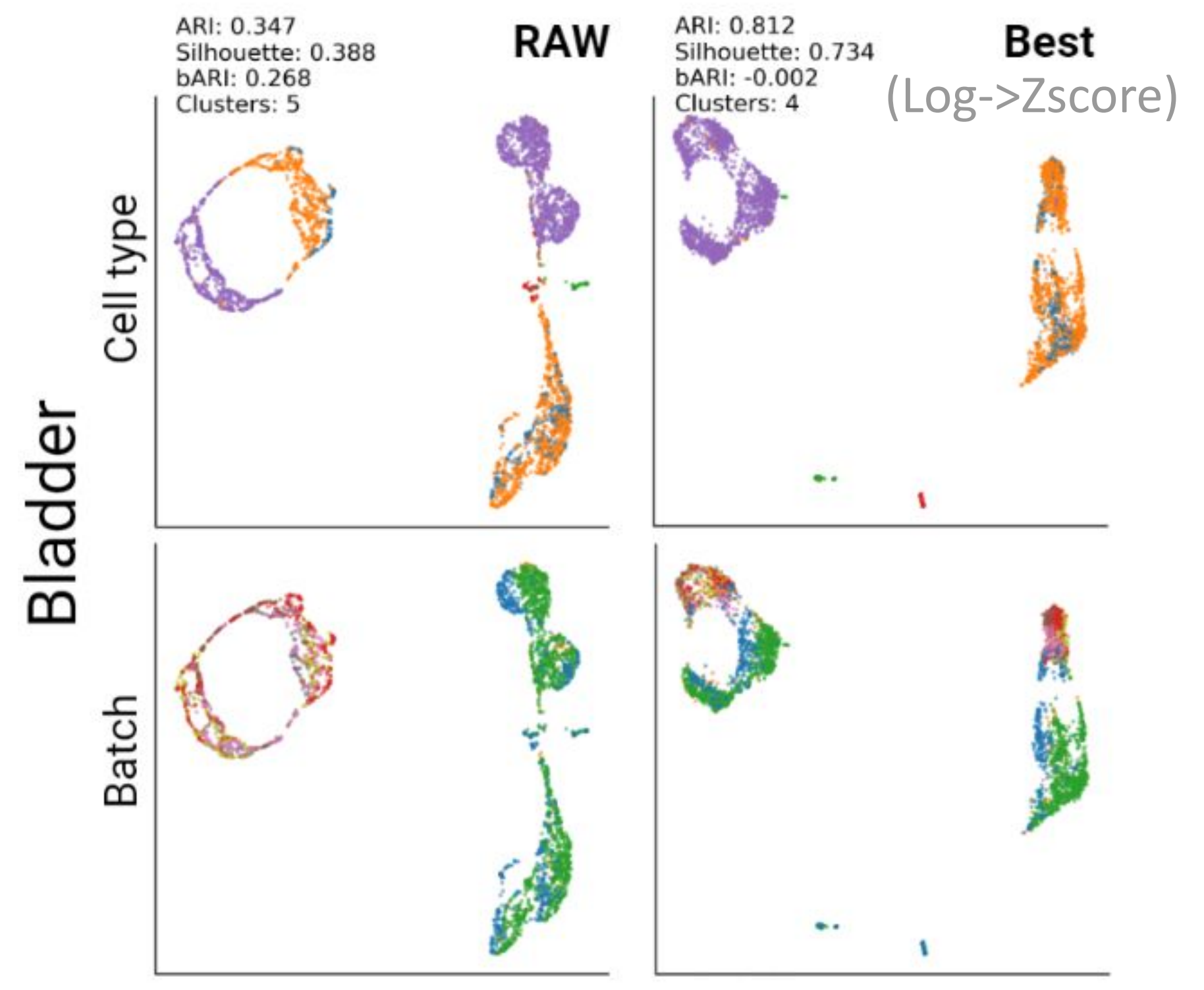
We compared the performance of our result to state-of-the-art methods benchmarked in Zhao2021. Notably, our result outperformed some of the tools benchmarked with ARI scores of 0.908 and 0.898.

Tools	MP	HP
Harmony ∗	0.969	0.955
Scanorama ∗	0.915	0.859
Seruat ∗	0.944	0.968
scVAE-GM ∗	0.805	NA
scVI ∗	0.932	0.759
LIGER ∗	0.914	0.911
scVI-LD ∗	0.875	0.656
scETM ∗	0.946	0.943
scETM $-\lambda$ ∗	0.851	0.474
scETM + adv ∗	0.944	0.946
Minmax + t-SNE+ DBSCAN	0.929	0.908
Total + UMAP+ DBSCAN	0.848	0.898
Meansquare + UMAP+ DBSCAN	0.903	0.741
*l*2-norm + UMAP+ DBSCAN	0.902	0.804
Total + AE + UMAP+ DBSCAN○	NA	0.947
Total + VAE + UMAP+ DBSCAN ○	NA	0.943

MP: Baron (Mouse) data
HP: Human pancreas datasets
∗ data are directly derived from the result by Zhao *et al.* [33]
○ is trained with Baron, and picked best result of ten repeats

Impact of data transformation on the Tabula Muris datasets integrative task.

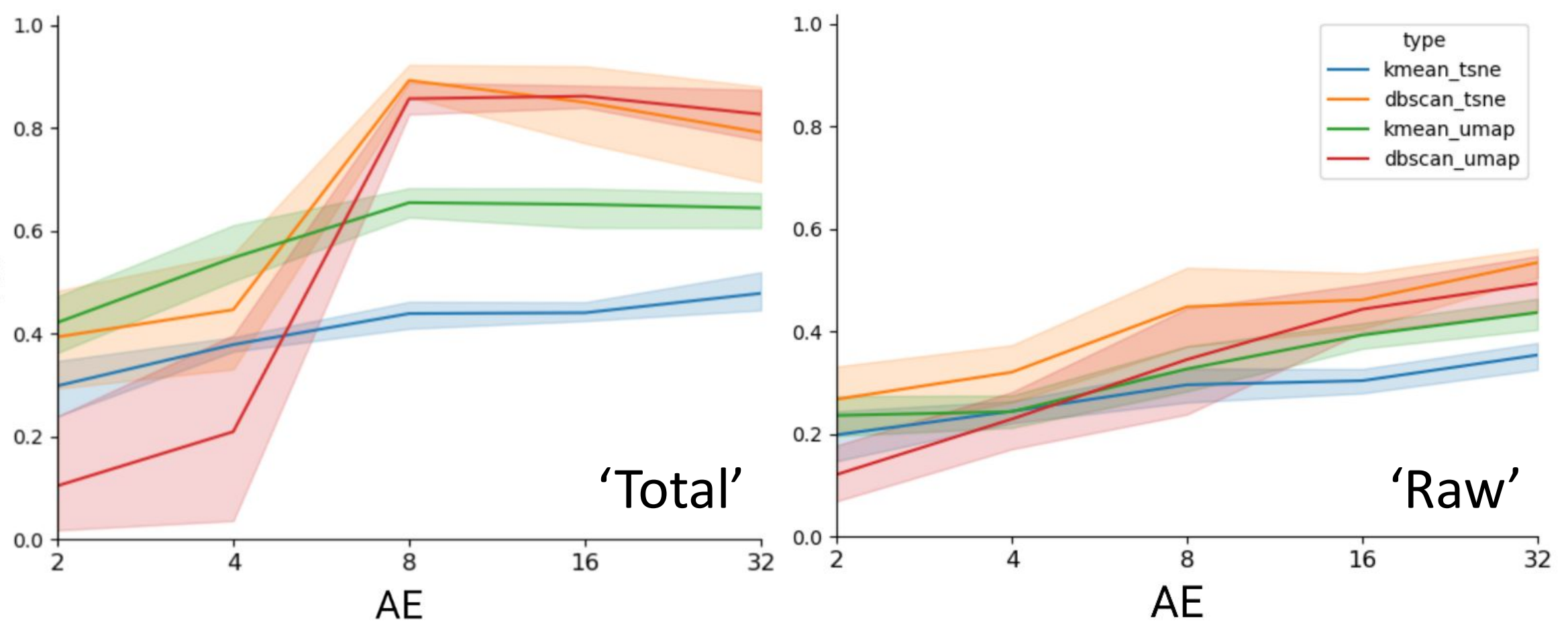
The Tabula Muris dataset consists of two different scRNA-seq protocols, SMART-Seq2 and CellRanger. By applying different data preprocessing, we observed significant improvements in 9 out of 11 tissue pairs of TM dataset. The cell types and batch ids are represented with different colors on each plot.



Impact of data transformation on the DNNs model.

The autoencoder (latent vector: 128) result without data transformation was ARI 0.551, but the 'Total' showed ARI 0.947. We investigated different DNNs models, Autoencoder, Variational Autoencoder, and ProtoTypical Network, and observed the similar impacts of preprocessing.

Next, we tested the power of DNNs models as a feature extractor for effective dimensionality reduction. We tested five different sizes for latent vectors from 2 to 32, and we could confirm that latent vector size between 8~32 is a reasonable length, and 50~100 is enough for latent vectors as reported in the other tools: scVI, scETM, scVAE, and scDHA.



Discussion

We demonstrated that, without any complex model for feature extraction or feature selection, well-combined data transformation and dimensionality reduction methods show good performance in cell type classification. Moreover, our result shows that t-SNE and UMAP are still powerful tools for single-cell analysis.

The DNN models were able to compress gene expression profiles into very small sized vectors for visualization and clustering with their feature selection power.

Online Supplementary Results

References

Cole, Michael B., et al. "Performance assessment and selection of normalization procedures for single-cell RNA-Seq." Cell systems 8.4 (2019)
Wang, Chunxiang et al. "Impact of data preprocessing on cell-type clustering based on single-cell RNA-seq data." BMC bioinformatics 21.1 (2020)
Zhao, Yifan, et al. "Learning interpretable cellular and gene signature embeddings from single-cell transcriptomic data." Nature communications 12.1 (2021)