Modeling Patient Variability in scRNA Sequence Data

Farhad Dehkordi, Kimia Ebrahimi, Sandra Rjeschni, Almut Voigts
Supervised by Vladimir Shitov, Till Richter, Malte Lücken

TEAM MEMBERS

Sandra Rjeschni

Studying M.Sc. Mathematics

Farhad Dehkordi

Studying B.Sc
Games Engineering

Almut Voigts

Studying M.Sc Mathematics

Kimia Ebrahimi

Studying B.Sc Bioinformatics

Motivation

&

Variability in cellular profiles

?

Disease progression and treatment outcomes

Dataset

	Stephenson et al. dataset From "Single-cell multi-omics analysis of the immune response in COVID-19"	Wang et al. dataset From "COMBATdb: a database for the COVID-19 Multi-Omics Blood ATlas"
Patients	130	140
Number of cells	781,123	783,704
Variable of interest	'Status' - 4 categories	'Outcome' - 6 ranks
Batch Variable	'Site' - 3 categories	'Institute'- 2 categories

Methods

Pseudobulk analysis

- Aggregating of cells
- Established method, but not standardized
    - Input Type
    - Aggregation Method
    - Distance Metric

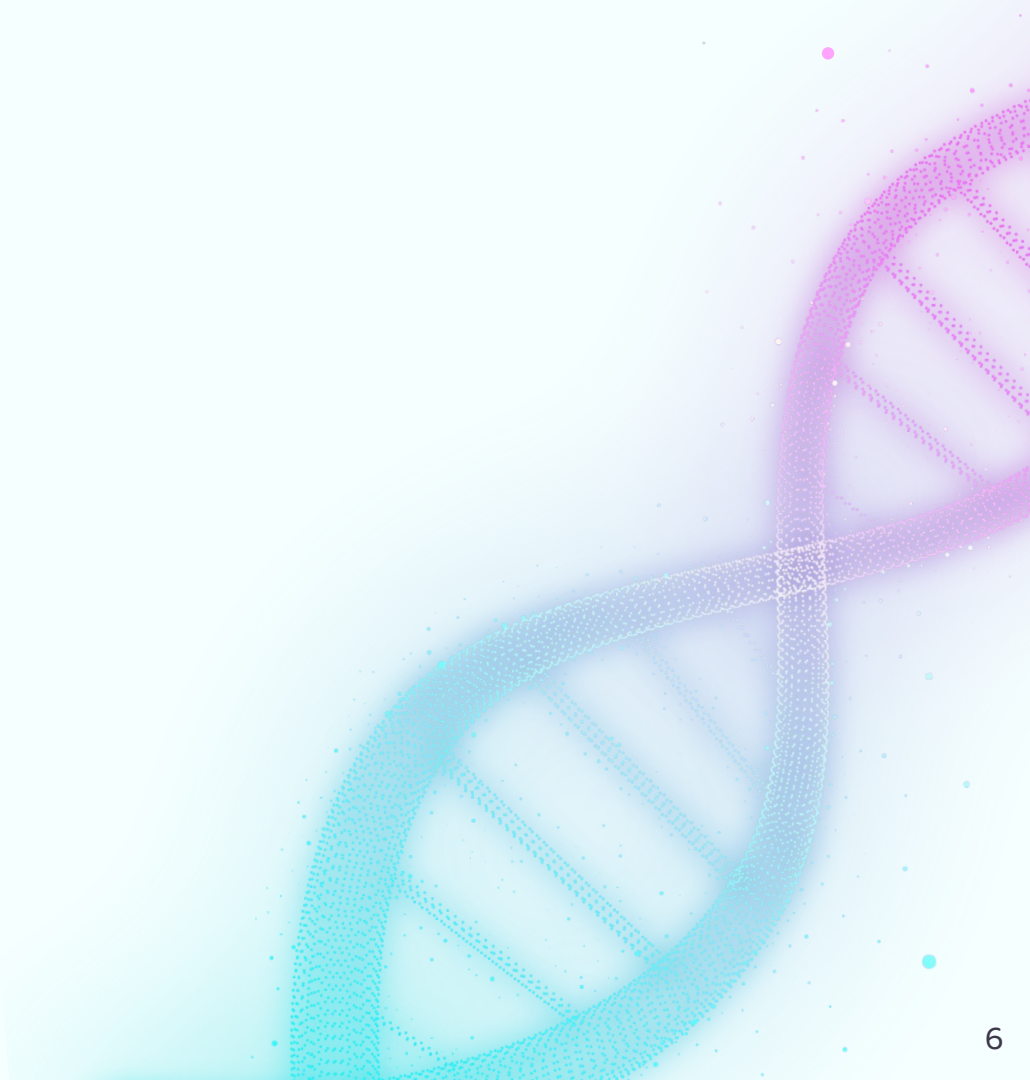
Aim: Establish best combination for described datasets

Develop a superior model with SSL approach

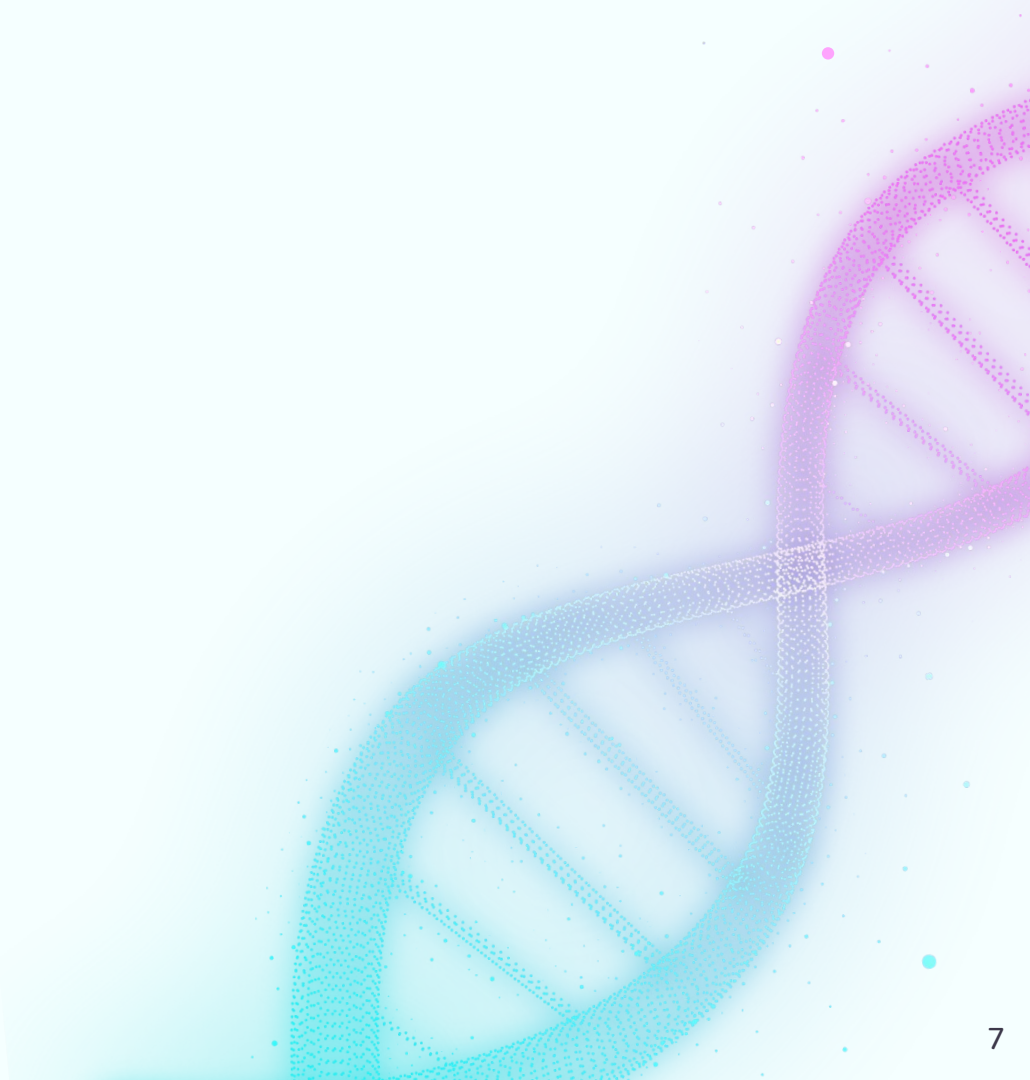
Methods

Pseudobulk analysis

- Metrics
  - Euclidean
  - Cosine
  - Manhattan
- Aggregation methods
  - Mean
  - Median
  - Sum
- Input types
  - PCA
  - Harmony
  - scVI
  - scANVI



Results



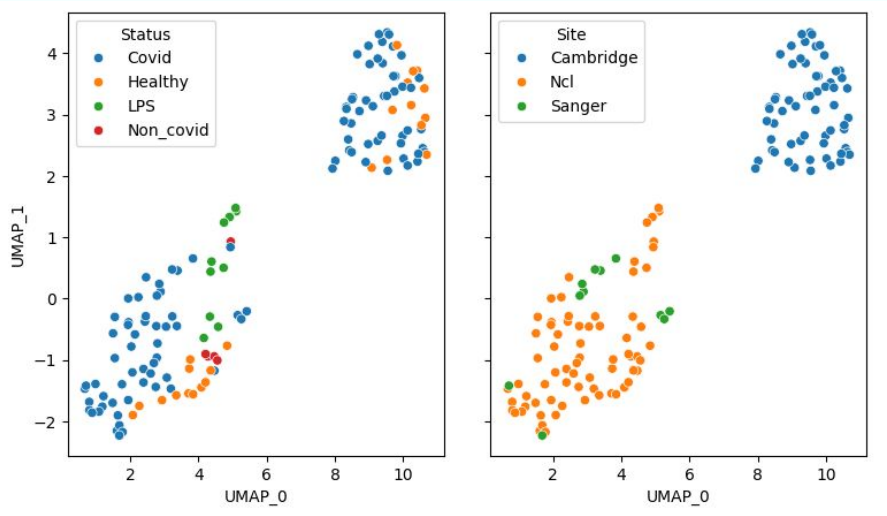
UMAP-Visualization of pseudobulk analysis

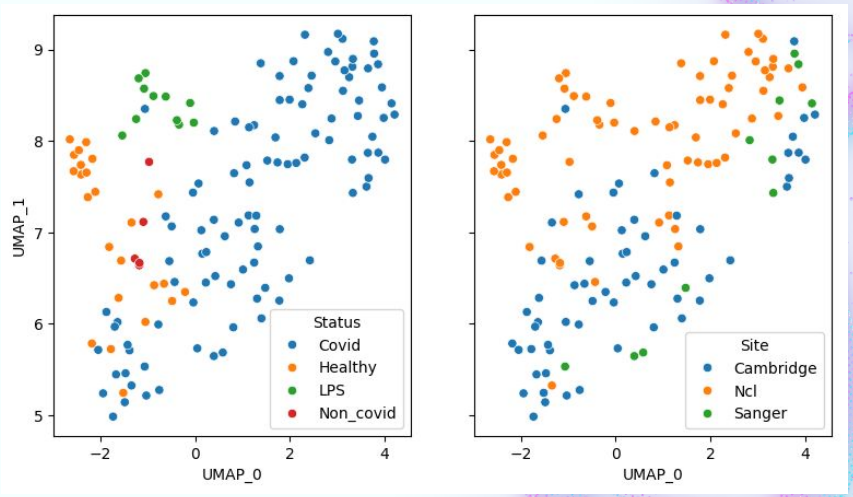
Dataset: Stephenson, Metric: Manhattan, Aggregation method: Mean

One point represents one patient

Input layer : PCA
→ grouping is highly associated with "Site" (batch effect)

Input layer: scVI
→ Batch effect reduced
→ "Status" has a better grouping



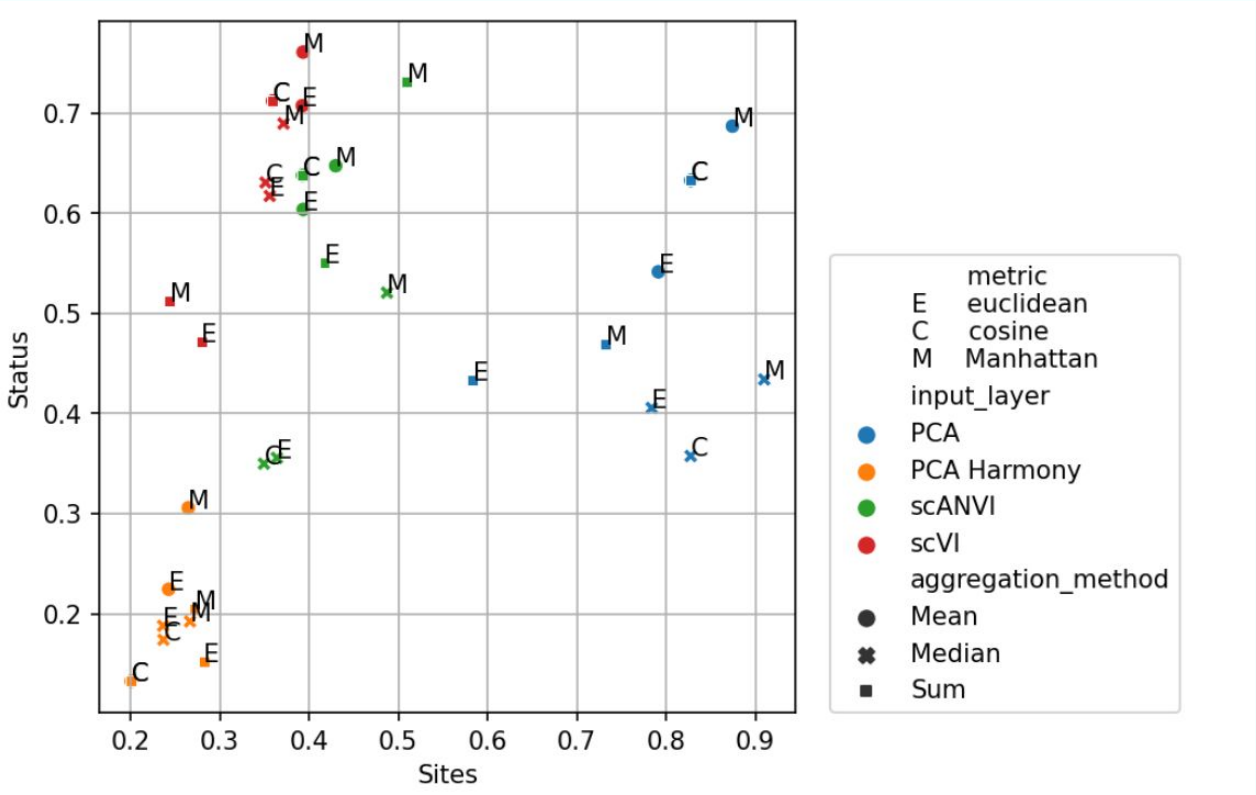


What is the best aggregation method, metric and input layer for pseudobulk analysis? *in the Stephenson et al. dataset*

Input layer: PCA/ PCA Harmony/ scVI/ scANVI, Aggregation method: mean/ median/ sum Metric: euclidean/ cosine/ manhattan

Wish: **high** f1-score for Status **low** f1-score for Sites

	Variable of interest (Status)			Batch variable (Site)		
	PCA					
	Euclidean	Cosine	Manhattan	Euclidean	Cosine	Manhattan
Mean	0.54151	0.632808	0.686977	0.791361	0.827601	0.874412
Median	0.40556	0.357604	0.434042	0.783896	0.827601	0.909976
Sum	0.433725	0.632808	0.469745	0.583315	0.827601	0.732368
	Harmony					
	Euclidean	Cosine	Manhattan	Euclidean	Cosine	Manhattan
Mean	0.224486	0.132722	0.305832	0.242647	0.200623	0.264706
Median	0.188022	0.17412	0.192379	0.236453	0.236871	0.266608
Sum	0.15187	0.132722	0.204667	0.282047	0.200623	0.272118
	scVI					
	Euclidean	Cosine	Manhattan	Euclidean	Cosine	Manhattan
Mean	0.707433	0.712192	0.760904	0.392442	0.359047	0.393295
Median	0.617309	0.630713	0.689542	0.355944	0.351271	0.371429
Sum	0.472193	0.712192	0.51284	0.279238	0.359047	0.243897
	scANVI					
	Euclidean	Cosine	Manhattan	Euclidean	Cosine	Manhattan
Mean	0.603802	0.637872	0.647453	0.393452	0.392703	0.429846
Median	0.355749	0.349893	0.520704	0.363946	0.349461	0.487325
Sum	0.551104	0.637872	0.731716	0.417208	0.392703	0.509288


Status
Sites
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
0.2
0.3
0.4
0.5
0.6
0.7
metric
E euclidean
C cosine
M Manhattan
input_layer
PCA
PCA Harmony
scANVI
scVI
aggregation_method
Mean
Median
Sum

What is the best aggregation method, metric and input layer for pseudobulk analysis? *in the Stephenson et al. dataset*

Input layer: PCA/ PCA Harmony/ scVI/ scANVI, Aggregation method: mean/ median/ sum Metric: euclidean/ cosine/ manhattan

Wish: **high** f1-score for Status **low** f1-score for Sites

	Variable of interest (Status)			Batch variable (Site)		
	PCA					
	Euclidean	Cosine	Manhattan	Euclidean	Cosine	Manhattan
Mean	0.54151	0.632808	0.686977	0.791361	0.827601	0.874412
Median	0.40556	0.357604	0.434042	0.783896	0.827601	0.909976
Sum	0.433725	0.632808	0.469745	0.583315	0.827601	0.732368
	Harmony					
	Euclidean	Cosine	Manhattan	Euclidean	Cosine	Manhattan
Mean	0.224486	0.132722	0.305832	0.242647	0.200623	0.264706
Median	0.188022	0.17412	0.192379	0.236453	0.236871	0.266608
Sum	0.15187	0.132722	0.204667	0.282047	0.200623	0.272118
	scVI					
	Euclidean	Cosine	Manhattan	Euclidean	Cosine	Manhattan
Mean	0.707433	0.712192	0.760904	0.392442	0.359047	0.393295
Median	0.617309	0.630713	0.689542	0.355944	0.351271	0.371429
Sum	0.472193	0.712192	0.51284	0.279238	0.359047	0.243897
	scANVI					
	Euclidean	Cosine	Manhattan	Euclidean	Cosine	Manhattan
Mean	0.603802	0.637872	0.647453	0.393452	0.392703	0.429846
Median	0.355749	0.349893	0.520704	0.363946	0.349461	0.487325
Sum	0.551104	0.637872	0.731716	0.417208	0.392703	0.509288

What is the best aggregation method, metric and input layer for pseudobulk analysis? *in the Stephenson et al. dataset*

Input layer: PCA/ PCA Harmony/ scVI/ scANVI, Aggregation method: mean/ median/ sum Metric: euclidean/ cosine/ manhattan

Wish: **high** f1-score for Status **low** f1-score for Sites

	Variable of interest (Status)			Batch variable (Site)		
	PCA					
	Euclidean	Cosine	Manhattan	Euclidean	Cosine	Manhattan
Mean	0.54151	0.632808	0.686977	0.791361	0.827601	0.874412
Median	0.40556	0.357604	0.434042	0.783896	0.827601	0.909976
Sum	0.433725	0.632808	0.469745	0.583315	0.827601	0.732368
	Harmony					
	Euclidean	Cosine	Manhattan	Euclidean	Cosine	Manhattan
Mean	0.224486	0.132722	0.305832	0.242647	0.200623	0.264706
Median	0.188022	0.17412	0.192379	0.236453	0.236871	0.266608
Sum	0.15187	0.132722	0.204667	0.282047	0.200623	0.272118
	scVI					
	Euclidean	Cosine	Manhattan	Euclidean	Cosine	Manhattan
Mean	0.707433	0.712192	0.760904	0.392442	0.359047	0.393295
Median	0.617309	0.630713	0.689542	0.355944	0.351271	0.371429
Sum	0.472193	0.712192	0.51284	0.279238	0.359047	0.243897
	scANVI					
	Euclidean	Cosine	Manhattan	Euclidean	Cosine	Manhattan
Mean	0.603802	0.637872	0.647453	0.393452	0.392703	0.429846
Median	0.355749	0.349893	0.520704	0.363946	0.349461	0.487325
Sum	0.551104	0.637872	0.731716	0.417208	0.392703	0.509288

What is the best aggregation method, metric and input layer for pseudobulk analysis? *in the Stephenson et al. dataset*

Input layer: PCA/ PCA Harmony/ scVI/ scANVI, Aggregation method: mean/ median/ sum Metric: euclidean/ cosine/ manhattan

Wish: **high** f1-score for Status **low** f1-score for Sites

	Variable of interest (Status)			Batch variable (Site)		
	PCA					
	Euclidean	Cosine	Manhattan	Euclidean	Cosine	Manhattan
Mean	0.54151	0.632808	0.686977	0.791361	0.827601	0.874412
Median	0.40556	0.357604	0.434042	0.783896	0.827601	0.909976
Sum	0.433725	0.632808	0.469745	0.583315	0.827601	0.732368
	Harmony					
	Euclidean	Cosine	Manhattan	Euclidean	Cosine	Manhattan
Mean	0.224486	0.132722	0.305832	0.242647	0.200623	0.264706
Median	0.188022	0.17412	0.192379	0.236453	0.236871	0.266608
Sum	0.15187	0.132722	0.204667	0.282047	0.200623	0.272118
	scVI					
	Euclidean	Cosine	Manhattan	Euclidean	Cosine	Manhattan
Mean	0.707433	0.712192	0.760904	0.392442	0.359047	0.393295
Median	0.617309	0.630713	0.689542	0.355944	0.351271	0.371429
Sum	0.472193	0.712192	0.51284	0.279238	0.359047	0.243897
	scANVI					
	Euclidean	Cosine	Manhattan	Euclidean	Cosine	Manhattan
Mean	0.603802	0.637872	0.647453	0.393452	0.392703	0.429846
Median	0.355749	0.349893	0.520704	0.363946	0.349461	0.487325
Sum	0.551104	0.637872	0.731716	0.417208	0.392703	0.509288

What is the best aggregation method, metric and input layer for pseudobulk analysis? *in the Stephenson et al. dataset*

Input layer: PCA/ PCA Harmony/ scVI/ scANVI, Aggregation method: mean/ median/ sum Metric: euclidean/ cosine/ manhattan

Wish: **high** f1-score for Status **low** f1-score for Sites

	Variable of interest (Status)			Batch variable (Site)		
	PCA					
	Euclidean	Cosine	Manhattan	Euclidean	Cosine	Manhattan
Mean	0.54151	0.632808	0.686977	0.791361	0.827601	0.874412
Median	0.40556	0.357604	0.434042	0.783896	0.827601	0.909976
Sum	0.433725	0.632808	0.469745	0.583315	0.827601	0.732368
	Harmony					
	Euclidean	Cosine	Manhattan	Euclidean	Cosine	Manhattan
Mean	0.224486	0.132722	0.305832	0.242647	0.200623	0.264706
Median	0.188022	0.17412	0.192379	0.236453	0.236871	0.266608
Sum	0.15187	0.132722	0.204667	0.282047	0.200623	0.272118
	scVI					
	Euclidean	Cosine	Manhattan	Euclidean	Cosine	Manhattan
Mean	0.707433	0.712192	0.760904	0.392442	0.359047	0.393295
Median	0.617309	0.630713	0.689542	0.355944	0.351271	0.371429
Sum	0.472193	0.712192	0.51284	0.279238	0.359047	0.243897
	scANVI					
	Euclidean	Cosine	Manhattan	Euclidean	Cosine	Manhattan
Mean	0.603802	0.637872	0.647453	0.393452	0.392703	0.429846
Median	0.355749	0.349893	0.520704	0.363946	0.349461	0.487325
Sum	0.551104	0.637872	0.731716	0.417208	0.392703	0.509288

What is the best aggregation method, metric and input layer for pseudobulk analysis? *COMBAT dataset*

Input layer: PCA/ PCA Harmony/ scVI/ scANVI, Aggregation method: mean/ median/ sum Metric: euclidean/ cosine/ manhattan

Wish: **high** Spearman-score for Status **low** f1-score for Institute

	Variable of interest (Outcome)			Batch variable (Institute)		
	PCA					
	Euclidean	Cosine	Manhattan	Euclidean	Cosine	Manhattan
Mean	0.516954	0.5106	0.531117	0	0.059732	0
Median	0.441388	0.47462	0.470388	0	0	0
Sum	0.573073	0.5106	0.556205	0.443746	0.059732	0.283117
	scVI					
	Euclidean	Cosine	Manhattan	Euclidean	Cosine	Manhattan
Mean	0.601517	0.6444	0.636751	0	0	0
Median	0.520509	0.59924	0.549614	0	0	0
Sum	0.564765	0.64441	0.576705	0.443746	0	0.415254
	scANVI					
	Euclidean	Cosine	Manhattan	Euclidean	Cosine	Manhattan
Mean	0.582703	0.60337	0.600343	0	0	0
Median	0.56566	0.55544	0.542906	0	0	0
Sum	0.56356	0.60335	0.625822	0.634921	0	0.660308

What is the best aggregation method, metric and input layer for pseudobulk analysis? *COMBAT dataset*

Input layer: PCA/ PCA Harmony/ scVI/ scANVI, Aggregation method: mean/ median/ sum Metric: euclidean/ cosine/ manhattan

Wish: **high** Spearman-score for Status **low** f1-score for Institute

	Variable of interest (Outcome)			Batch variable (Institute)		
	PCA					
	Euclidean	Cosine	Manhattan	Euclidean	Cosine	Manhattan
Mean	0.516954	0.5106	0.531117	0	0.059732	0
Median	0.441388	0.47462	0.470388	0	0	0
Sum	0.573073	0.5106	0.556205	0.443746	0.059732	0.283117
	scVI					
	Euclidean	Cosine	Manhattan	Euclidean	Cosine	Manhattan
Mean	0.601517	0.6444	0.636751	0	0	0
Median	0.520509	0.59924	0.549614	0	0	0
Sum	0.564765	0.64441	0.576705	0.443746	0	0.415254
	scANVI					
	Euclidean	Cosine	Manhattan	Euclidean	Cosine	Manhattan
Mean	0.582703	0.60337	0.600343	0	0	0
Median	0.56566	0.55544	0.542906	0	0	0
Sum	0.56356	0.60335	0.625822	0.634921	0	0.660308

What is the best aggregation method, metric and input layer for pseudobulk analysis? *COMBAT dataset*

Input layer: PCA/ PCA Harmony/ scVI/ scANVI, Aggregation method: mean/ median/ sum Metric: euclidean/ cosine/ manhattan

Wish: **high** Spearman-score for Status **low** f1-score for Institute

	Variable of interest (Outcome)			Batch variable (Institute)		
	PCA					
	Euclidean	Cosine	Manhattan	Euclidean	Cosine	Manhattan
Mean	0.516954	0.5106	0.531117	0	0.059732	0
Median	0.441388	0.47462	0.470388	0	0	0
Sum	0.573073	0.5106	0.556205	0.443746	0.059732	0.283117
	scVI					
	Euclidean	Cosine	Manhattan	Euclidean	Cosine	Manhattan
Mean	0.601517	0.6444	0.636751	0	0	0
Median	0.520509	0.59924	0.549614	0	0	0
Sum	0.564765	0.64441	0.576705	0.443746	0	0.415254
	scANVI					
	Euclidean	Cosine	Manhattan	Euclidean	Cosine	Manhattan
Mean	0.582703	0.60337	0.600343	0	0	0
Median	0.56566	0.55544	0.542906	0	0	0
Sum	0.56356	0.60335	0.625822	0.634921	0	0.660308

What is the best aggregation method, metric and input layer for pseudobulk analysis? *COMBAT dataset*

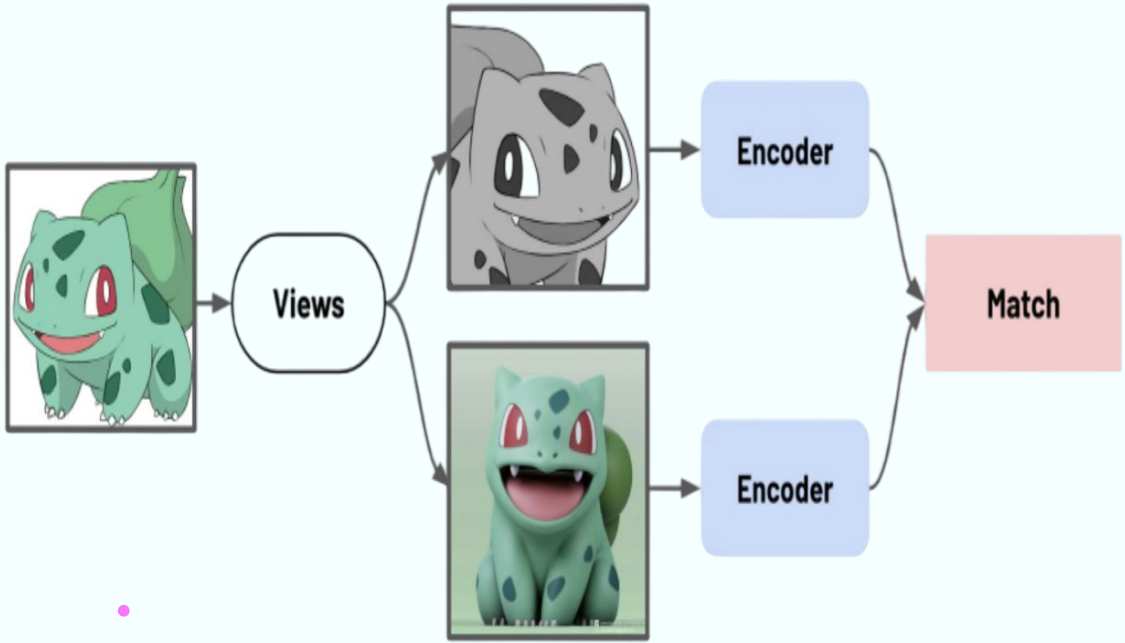
Input layer: PCA/ PCA Harmony/ scVI/ scANVI, Aggregation method: mean/ median/ sum Metric: euclidean/ cosine/ manhattan

Wish: **high** Spearman-score for Status **low** f1-score for Institute

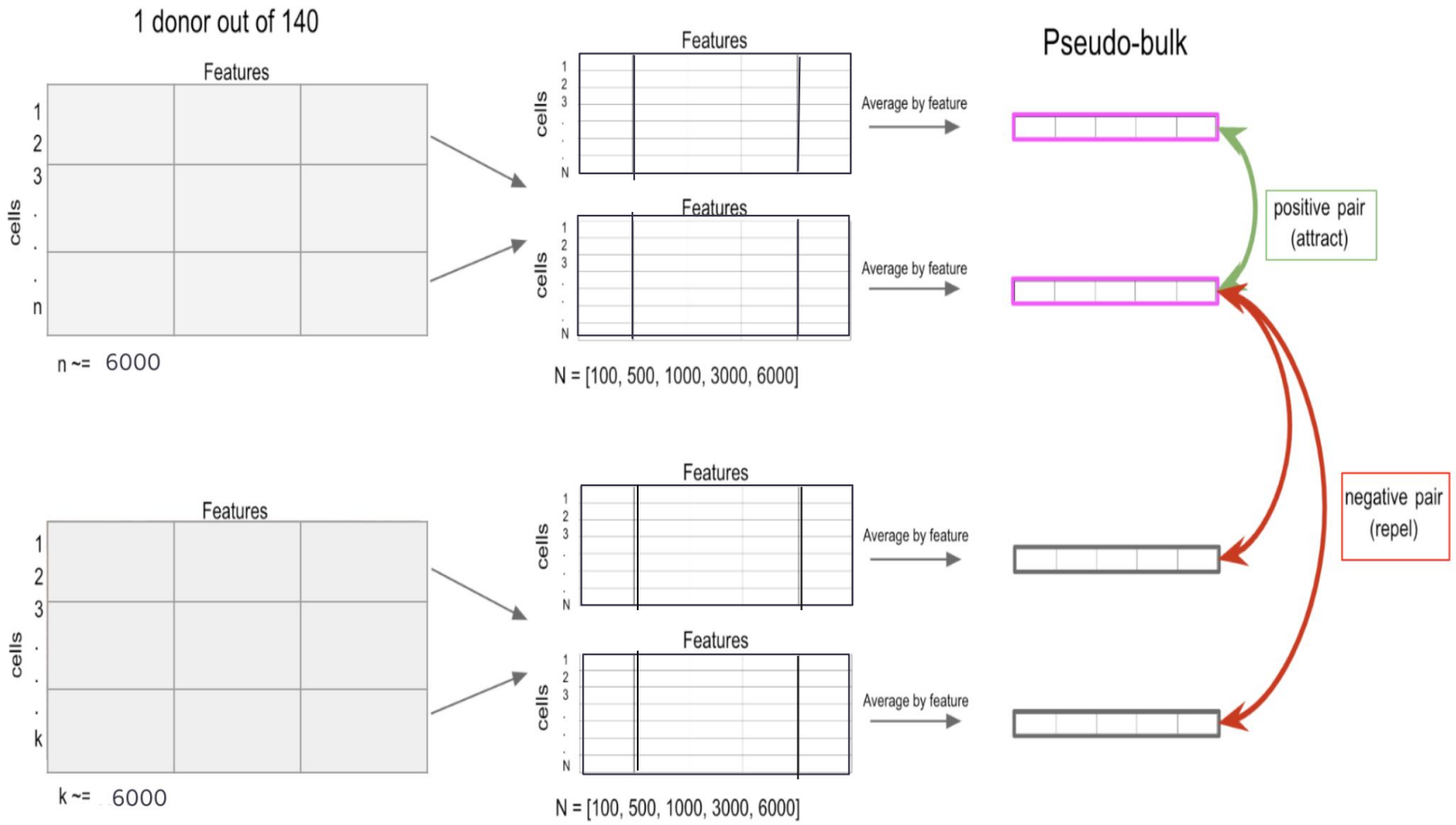
	Variable of interest (Outcome)			Batch variable (Institute)		
	PCA					
	Euclidean	Cosine	Manhattan	Euclidean	Cosine	Manhattan
Mean	0.516954	0.5106	0.531117	0	0.059732	0
Median	0.441388	0.47462	0.470388	0	0	0
Sum	0.573073	0.5106	0.556205	0.443746	0.059732	0.283117
	scVI					
	Euclidean	Cosine	Manhattan	Euclidean	Cosine	Manhattan
Mean	0.601517	0.6444	0.636751	0	0	0
Median	0.520509	0.59924	0.549614	0	0	0
Sum	0.564765	0.64441	0.576705	0.443746	0	0.415254
	scANVI					
	Euclidean	Cosine	Manhattan	Euclidean	Cosine	Manhattan
Mean	0.582703	0.60337	0.600343	0	0	0
Median	0.56566	0.55544	0.542906	0	0	0
Sum	0.56356	0.60335	0.625822	0.634921	0	0.660308

Self-Supervised Learning (SSL) via Contrastive Learning

- Maximizing agreement between differently augmented views of the same data point while minimizing it between unrelated points



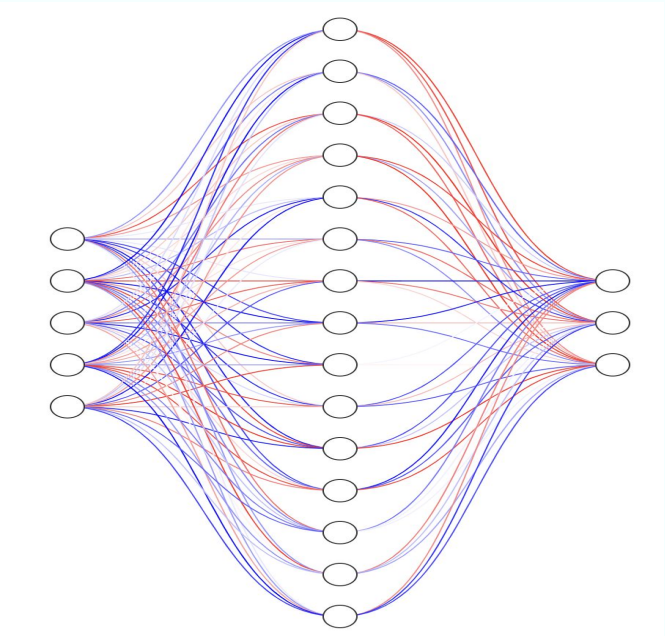
Now,
how can we leverage these techniques to infer patient representation?


1 donor out of 140
Features
cells
1
2
3
n
n ~= 6000
Features
cells
1
2
3
N
Average by feature
Features
cells
1
2
3
N
Average by feature
N = [100, 500, 1000, 3000, 6000]
Pseudo-bulk
positive pair
(attract)
Features
cells
1
2
3
k
k ~= 6000
Features
cells
1
2
3
N
Average by feature
Features
cells
1
2
3
N
Average by feature
N = [100, 500, 1000, 3000, 6000]
negative pair
(repel)

Contrastive Loss

InfoNCE Loss: $$\mathcal{L} = -\log \frac{\exp(\mathrm{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{N} \exp(\mathrm{sim}(z_i, z_k)/\tau)}$$

Cauchy-based similarity measure: $$\frac{1}{1 + (\mathrm{cdist}(a, b) \cdot \tau)^2}$$



Best hyperparameters

For **scVI** :
{'learning_rate': 0.0431
'weight_decay': 4.48e-05,
'temperature': 0.904,
'batch_size': 4,
'num_layers': 1,
'hidden_size': 192,
'cell_subset_size': 5486}

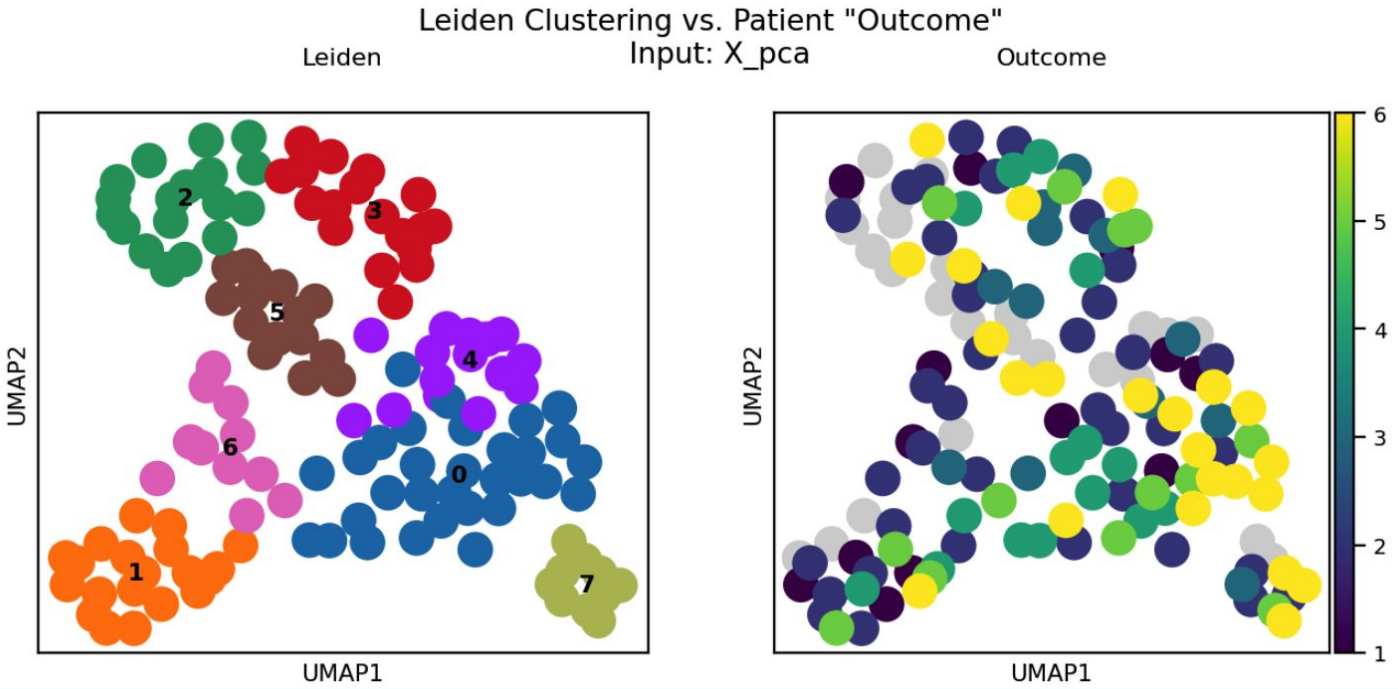
For **scANVI**:
{'learning_rate': 0.052,
'weight_decay': 5.43e-06,
'temperature':0.428
, 'batch_size': 4,
'num_layers': 2,
'hidden_size': 257,
'cell_subset_size':5848}

Performance Assessment On Unseen Data

- Performed k-NN evaluation with **2 metrics**: F1-Macro, Spearman correlation coefficient (SCC)
- **scVI** and **scANVI** demonstrated superior performance over **PCA** in both datasets

Dataset/ Layer/ score	Input Layer	SCC	F1-Macro
COMBAT	**scVI**	**0.602**	**0.208**
	scANVI	0.555	0.205
	PCA	0.372	0.081
Stephenson	scVI	0.718	0.389
	scANVI	0.754	0.579
	PCA	0.584	0.466

Performance of ‘PCA’ input for Contrastive Learning model on COMBAT

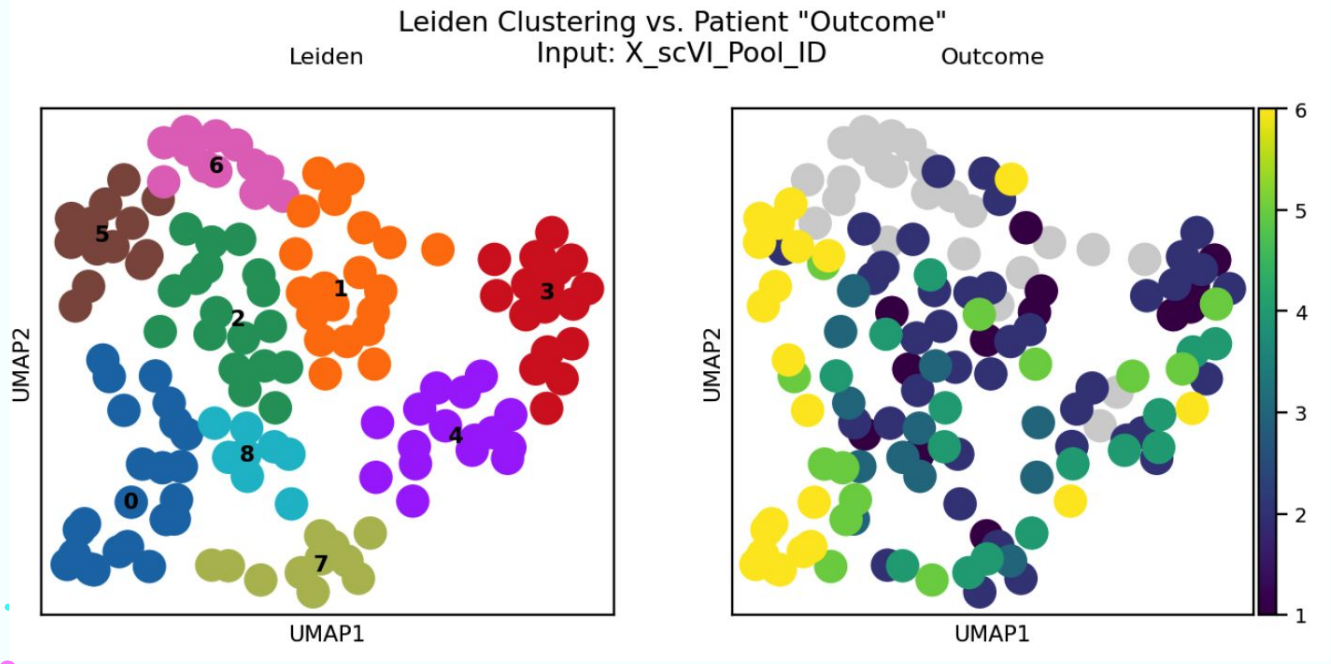


- **SCC score: 0.37**
- No meaningful clustering

1 = Death
2 = Intubated, ventilated
3 = Non-invasive ventilation
4 = Hospitalized, O2
5 = Hospitalized, no O2
6 = Not hospitalized

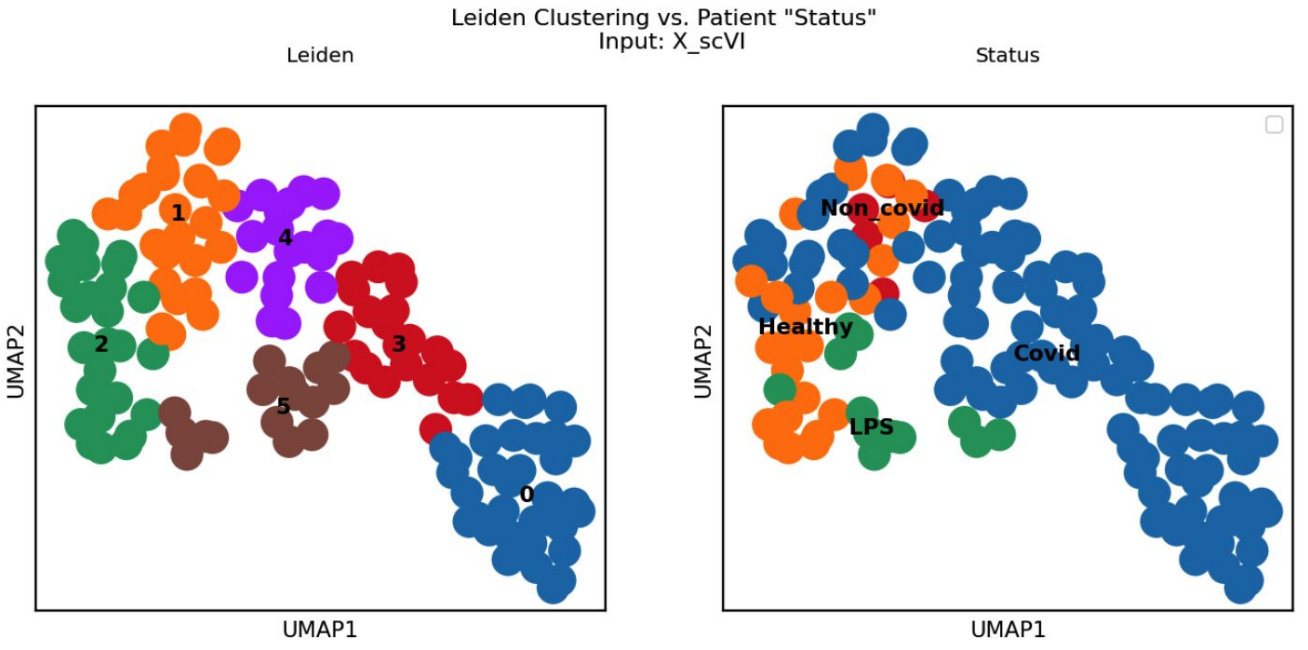
Performance of ‘scVI’ input for Contrastive Learning model on COMBAT

1 = Death
2 = Intubated, ventilated
3 = Non-invasive ventilation
4 = Hospitalized, O2
5 = Hospitalized, no O2
6 = Not hospitalized



- **SCC score: 0.60**
- Clustering of healthy vs deceased patients are observed (yellow dots)
- Distinct clustering of patients with conditions unrelated to the primary study (gray dots)

Performance of ‘scVI’ input for Contrastive Learning model on Stephenson dataset



- **F1-Macro score: 0.38**
- Patients of the similar categories are grouped together

Conclusion

- Pseudobulk vs SSL based approach
  - F1-Macro score using SSL: 0.38
  - F1-Macro score using pseudobulk: 0.76
  - Contrastive learning model did not outperform the pseudobulk method

- Exploring other Contrastive Losses and similarity functions?
  - Triplet loss
  - Margin loss
  - Gaussian similarity
  - Dot product similarity

SOURCES

- Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyper- parameter optimization framework.
- Balestriero, R., & LeCun, Y. (2022). Contrastive and non-contrastive self-supervised learning re- cover global and local spectral embedding methods.
- Baysoy, A., Bai, Z., Satija, R., & Fan, R. (2023). The technological landscape and applications of single-cell multi-omics. Nature Reviews Molecular Cell Biology, 1–19.
- Chen, W. S., Zivanovic, N., Van Dijk, D., Wolf, G., Bodenmiller, B., & Krishnaswamy, S. (2020). Uncovering axes of variation among single- cell cancer specimens. Nature methods, 17 (3), 302–310.
- De Donno, C., Hediyeh-Zadeh, S., Moinfar, A. A., Wagenstetter, M., Zappia, L., Lotfollahi, M., & Theis, F. J. (2023). Population-level integration of single-cell datasets enables multi- scale analysis across samples. Nature Methods, 20 (11), 1683–1692.
- Joodaki, M., Shaigan, M., Parra, V., Bülow, R. D., Kuppe, C., Hölscher, D. L., Cheng, M., Na- gai, J. S., Goedertier, M., Bouteldja, N., et al. (2023). Detection of patient-level distances from single cell genomics and pathomics data with optimal transport (pilot). Molecular Systems Biology, 1–18.
- Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P. - r., & Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with harmony. Nature methods, 16 (12), 1289–1296.
- Luecken, M. D., B¨uttner, M., Chaichoompu, K., Danese, A., Interlandi, M., Müller, M. F., Strobl, D. C., Zappia, L., Dugas, M., Colom´e- Tatch´e, M., et al. (2022). Benchmarking atlas- level data integration in single-cell genomics. Nature methods, 19 (1), 41–50.
- McInnes, L., Healy, J., & Melville, J. (2020). Umap: Uniform manifold approximation and projection for dimension reduction.
- Murphy, A., & Skene, N. (n.d.). A balanced measure shows superior performance of pseudobulk methods in single-cell rna-sequencing analysis. nat commun. 2022; 13: 7851.
- Ramirez Flores, R. O., Lanzer, J. D., Dimitrov, D., Velten, B., & Saez-Rodriguez, J. (2023). Multicellular factor analysis of single-cell data for a tissue-centric understanding of disease (J. Park & A. M. Walczak, Eds.). eLife, 12, e93161. https://doi.org/10.7554/eLife.93161
- Squair, J. W., Gautier, M., Kathe, C., Anderson, M. A., James, N. D., Hutson, T. H., Hudelle, R., Qaiser, T., Matson, K. J., Barraud, Q., et al. (2021). Confronting false discoveries in single-cell differential expression. Nature communications, 12 (1), 5692.
- Stephenson, E., Reynolds, G., Botting, R. A., Calero- Nieto, F. J., Morgan, M. D., Tuong, Z. K., Bach, K., Sungnak, W., Worlock, K. B., Yoshida, M., et al. (2021). Single-cell multi- omics analysis of the immune response in covid-19. Nature medicine, 27 (5), 904–916.
- van den Oord, A., Li, Y., & Vinyals, O. (2019). Representation learning with contrastive predictive coding.
- Wang, D., Kumar, V., Burnham, K. L., Mentzer, A. J., Marsden, B. D., & Knight, J. C. (2023). Combatdb: A database for the covid-19 multi- 7 omics blood atlas. Nucleic Acids Research, 51 (D1), D896–D905.
- Watson, E. R., Mora, A., Taherian Fard, A., & Mar, J. C. (2022). How does the structure of data impact cell–cell similarity? Evaluating how structural properties influence the performance of proximity metrics in single cell RNA-seq data. Briefings in Bioinformatics, 23 (6), bbac387. https://doi.org/10.1093/bib/ bbac387 8
- https://nips.cc/media/neurips-2021/Slides/21895.pdf

Please refer to the written report for a comprehensive bibliography

THANK YOU :)

Any Questions ?