Enhancing Single-Cell VAE Latent Space via Semi-Supervision

Meichen Gong [1] **Konstantin Ivanov** [2] **Merja Heinäniemi** [2] **Ville Hautamäki** [1]

Abstract

Single-cell data are crucial for biomedical discovery, facilitated by low-dimensional latent space encoding of single-cell RNA-seq profile. Obtained latent codes can then be plotted into 2-D space via t-SNE or UMAP allowing practitioners to infer new knowledge. Alternatively, downstream deep learning applications can also be trained from the latent codes, one of the common being the cell type classifier. The usefulness of the 2-D plot or downstream application depends critically on the structure of the latent space and whether it encodes biological information or noise. The proposed approach aims to improve the latent space via injecting a bit of label information, thus denoting our approach as a semi-supervised one. We include a novel dual-VAE structure, where information flows from the controller VAE to the main model. Our results demonstrate that the incorporation of SemafoVAE improves the performance of the existing scANVI model, therefore offering a refined model structure with disentangled latent representations for robust biological insights.

1. Introduction

Single-cell data are pivotal for dissecting cellular heterogeneity, understanding individual cell functions, and revealing nuanced biological processes that bulk analyses often obscure. Considerable effort (Xu et al., 2021; Yu & Welch, 2021; Piran et al., 2024; Lotfollahi et al., 2022; Gayoso et al., 2022; Trong et al., 2020) has been devoted to the analysis of single-cell data, including scArches (Lotfollahi et al., 2022) which aims at analyzing single-cell query data by integrating it into a reference atlas. Among its multiple applications, single-cell annotation using variational inference (scANVI (Xu et al., 2021)) can enhance the accuracy and efficiency of identifying cell types from single-cell sequencing data, handling single-cell data integration tasks, etc.

The quality of the downstream applications, such as cell type labeling, is crucially dependent on the quality of the encoded latent space itself. The holy grail in latent space quality is seen as the *disentangled latent space*, where each latent dimension independently controls one *factor of variation* (Bengio et al., 2014). Unfortunately it has been shown that *without external metadata*, disentanglement learning for VAEs is impossible (Locatello et al., 2019; Khemakhem et al., 2020). For these reasons, we aim to improve the latent space quality by incorporating a portion of meta-data into the VAE training. We are motivated by a recent SemafoVAE (Ngo et al., 2022), where the authors used double-VAE structure. One VAE incorporates the meta-data and the other scRNA-seq profiles. The full training algorithm is derived using information theoretic principles. This is the first time SemafoVAE has been applied to biological data.

2. Related Work

Kingma et al. (Kingma & Welling, 2014) first introduced the Variational Autoencoder (VAE) and integrated the variational inference method into the VAE model. Subsequently, Kingma et al. (Kingma et al., 2014) proposed the M1+M2 model structure, which consists of two VAEs operating simultaneously during the semi-supervised learning process. This dual-VAE structure has leveraged properties of the data to enhance generation quality and has yielded more accurate classification results than classifiers using only labeled data. Similarly, Ngo et al. (Ngo et al., 2022) constructed their model based on the M1+M2 framework, incorporating transitive information theory to maximize the mutual information between the latent representations and the observed raw data, while ensuring that the latent space is influenced by the data properties for structured and disentangled latent representations.

Furthermore, numerous efforts have been made to utilize deep generative models in single-cell data analysis. For instance, the integration of VAE and GAN (Generative Adversarial Networks) technologies in MichiGAN (Yu & Welch, 2021) has successfully combined the strengths of both approaches, resulting in high-quality data generation. Addi-

[1]School of Computing, University of Eastern Finland [2]Institute of Biomedicine, University of Eastern Finland. Correspondence to: Meichen Gong <meigong@uef.fi>.

Proceedings of the ICML 2024 Workshop on Accessible and Efficient Foundation Models for Biological Discovery, Vienna, Austria.

tionally, the Biolord model (Piran et al., 2024) has advanced recent disentanglement techniques to learn latent representations in single-cell data, which have been applied to various biological prediction tasks and obtained great performance. Lastly, Gayoso et al. (Gayoso et al., 2022) developed the code of scANVI (Xu et al., 2021), a semi-supervised deep learning model for single-cell data, structurally reminiscent of the M1+M2 model. ScANVI utilizes cell type information from a subset of cells within the dataset to predict the properties of other cell groups.

3. Method

3.1. Model Structure Overview

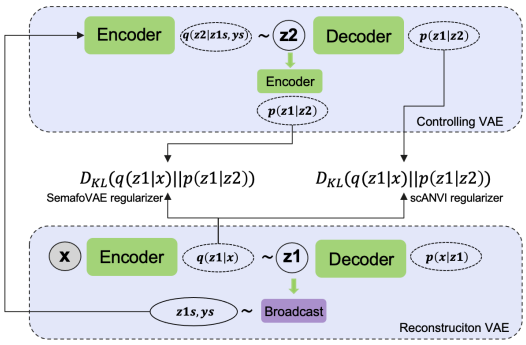

Figure 1. The model structure for the integrated model which includes both the scANVI regularizer and the SemafoVAE regularizer.

In the original M1+M2 model structure, it was stated that M1 is the latent-feature discriminative model that aims to provide the embedding or feature representation of the data, whereas M2 is the generative semi-supervised model which utilizes the embedding from the M1 and possible latent class variable y to generate new data. In Figure 1, we present the model structure of scANVI and the modified version of SemafoVAE proposed in (Ngo et al., 2022). The reconstruction VAE functions similarly to model M1 in the M1+M2 model structure, whereas the controlling VAE functions similarly to model M2. The difference is that in our model structure, the reconstruction VAE will be the VAE that is used to generate the final result, and the controlling VAE is the one that is used to generate encoded conditional distribution which contains the class information y. By minimizing the Kullback-Leibler (KL) divergence between the conditional distributions and the approximate posterior distribution of reconstruction VAE, we enforce the reconstruction VAE to learn more about the class information y and be influenced by it, consequently generating new data with more accurate classes. ScANVI operates by leveraging the decoded conditional distribution from the controlling VAE for regularization purposes whereas SemafoVAE utilizes the conditional distribution derived from double encoding operations.

3.2. Transitive Information Theory

In Ngo et al.'s work (2022), the concept of transitive information theory was utilized, which asserts that the mutual information between two target variables x and z, can be effectively maximized through a third variable y. In our integrated model which combines both scANVI and SemafoVAE, we denote x to be the observed raw data, z to be the latent variables sampled in the reconstruction VAE and y to be the variable that represents the class information. Our objective is to maximize mutual information $I(x, z)$ to ensure that z captures the maximum information from x while the class variable y is utilized to impose structured, controlled, and disentangled representations in the latent space, enhancing the informativeness of the generative outputs. This process involves maximizing both $I(x, y)$ and $I(y, z)$, with the class variable y serving as the bridge by the transitive information theory.

3.3. Model Obejectives

Variational Autoencoder (VAE) was introduced in (Kingma & Welling, 2014), but because the exact true posterior distribution is intractable, they implemented variational inference to facilitate the estimation of parameters. This technique is again used in the M1+M2 model structure (Kingma et al., 2014), lower bounds of the log marginal likelihood of the model are derived to ensure the approximate posterior can be as close to the true posterior as possible.

There are separate model objectives for model M1 and model M2, for model M1, the variational bound $\mathcal{J}(x)$ on the log marginal likelihood for a single data point is:

$$\log p_\theta(x) \geq \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - KL[q_\phi(z|x)\|p_\theta(z)] = -\mathcal{J}(x), \quad (1)$$

where $q_\phi(z|x)$ is the approximate posterior, with parameters ϕ and $p_\theta(x|z)$ is the decoder with parameters θ. The goal of the VAE training is to estimate both ϕ and θ.

For model M2, depending on whether a data point is labeled or not, there are two different variational bounds for each situation, from the scvi-tools code (Gayoso et al., 2022), it can be observed that for scANVI, whether the data point is originally labeled or not, through broadcasting, labels would be generated for all the data points, thus we only focus on the variational bound where the data point is labeled in

model M2, which is shown as follows:

$$\log p_\theta(x,y) \geq \mathbb{E}_{q_\phi(z|x,y)}[\log p_\theta(x|y,z) + \log p_\theta(y) + \log p(z) - \log q_\phi(z|x,y)] = -\mathcal{L}(x,y). \quad (2)$$

Compared with the model objectives above, in scANVI, KL divergence is missing for the reconstruction VAE and the log-likelihood of it is slightly different. For the controlling VAE, several KL divergence terms were composed as constraints. Additionally, if any data point is labeled, classification loss is also computed, otherwise, this term is omitted. Because of the similarity of the model structure between scANVI and SemafoVAE, we solely conclude the variational bound of SemafoVAE as a whole based on our integration and our understanding of the original scvi-tools code:

$$\log p_\theta(x,y) \geq \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \alpha \log q_\tau(y|z) - \log q_\phi(z|x,y)] - D_{\mathrm{KL}}(q(z|x)\|p(z|y)) - D_{\mathrm{KL}}(q(z)\|p(z)) = -\mathcal{L}(x,y), \quad (3)$$

where $\alpha \log q_\tau(y|z)$ is the term evaluating classification performance, as mentioned above, if the data point is unlabelled, this term is omitted.

4. Results

4.1. RNA-seq Cell Type Prediction

We choose the pancreas single-cell RNA sequencing (scRNA-seq) data (Naghipourfar, 2020) to train the model and perform downstream prediction analysis using the classifier from scvi-tools (Gayoso et al., 2022) and the learned latent representations. In particular, we compare the label information predicted from the classifier and the ground-truth labels in the latent representations to calculate the prediction accuracy. The model was trained based on the modified version of the semi-supervised surgery pipeline from scArches (Lotfollahi et al., 2022), the pancreas scRNA-seq dataset is also obtained from this pipeline. Detailed information about the dataset's splitting plan during the training procedure can be found in Appendix A.

To further test our model's performance, datasets with different labeled rates are utilized as inputs for the training process, the result is shown in Table 1. For the combination of two models, we keep the structure of two models at the same time, both the SemafoVAE regularizer and the scANVI regularizer are utilized during the training process. It can be observed that for the fully labeled dataset, scANVI has the best prediction performance, and for partially labeled datasets or completely unlabeled datasets, SemafoVAE or the combination of both models usually yields the best result.

Table 1. Prediction accuracy of pancreas scRNA-seq data using scANVI, SemafoVAE and their combination under different labeled rates.

Labeled Rate (%)	scANVI	SemafoVAE	Both
100.0	**0.9735**	0.9540	0.9489
82.4	0.9046	0.9072	**0.9080**
45.0	0.9003	**0.9105**	0.9087
39.1	**0.9088**	0.9069	0.9042
21.2	0.9146	0.9185	**0.9210**
16.8	0.8990	**0.9046**	0.9035
0.0	0.8998	**0.9174**	0.9157

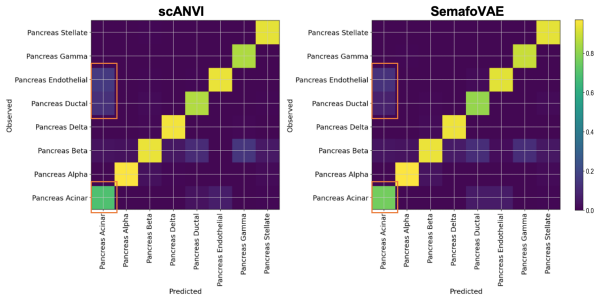

Figure 2. The confusion matrices comparison between scANVI and SemafoVAE for completely unlabeled (labeled rate: 0.0%) pancreas scRNA-seq data.

In Figure 2 we also present two confusion matrices that visualize the accuracy of our model's predictions when the labeled rate is 0.0%. In the query dataset under this condition, pancreas beta and pancreas alpha are the most common cell types which take up about 35.8% and 26.3% of the whole query dataset respectively. In contrast, pancreas gamma is the rarest by taking up only 2.8% of the whole query dataset. Each row in the matrix corresponds to the observed cell types (true labels), and each column corresponds to the model's predictions (predicted labels). In both matrices, the brightness of the cells along the diagonal represents the proportion of correct predictions for each cell type. According to the color bar, brighter cells (tending towards yellow) indicate a higher proportion of correct classifications. Cells off the diagonal represent misclassifications, where the model predicted a different cell type other than the true label. Notable differences in prediction accuracy across various cell types can be observed from the matrices, for example, the orange rectangles along the diagonals indicate that SemafoVAE demonstrates better predictive accuracy for pancreas acinar cells as its color in SemafoVAE matrix is brighter than its counterpart in scANVI matrix. For the orange rectangle off the diagonal in the SemafoVAE matrix, the cells inside are colored lighter, tending towards darker shades of purple, compared to their counterparts in

the scANVI matrix. This indicates that SemafoVAE exhibits fewer misclassifications in these areas, demonstrating its enhanced predictive reliability.

To validate the robustness of our integrated model, we also trained it with the bone marrow human cell atlas scRNA-seq dataset created by (Luecken et al., 2021), this dataset is approximately 178 times bigger than the pancreas scRNA-seq data we used earlier, it is a complex, real-world dataset with diverse cell types found in human bone marrow and thus could be served as a high-quality benchmark dataset for models that integrate multi-modal single-cell data. The training process is similar to that of the pancreas scRNA-seq data.

Table 2. Bone marrow human cell atlas scRNA-seq data label prediction accuracy using scANVI, SemafoVAE and their combination.

Labeled Rate(%)	scANVI	SemafoVAE	Both
100.0	**0.9954**	0.9882	0.9892
78.5	**0.9441**	0.9377	0.9434
34.2	0.9414	**0.9433**	0.9359
0.0	0.9396	0.9367	**0.9444**

Table 3. Result of other evaluations towards completely unlabeled (labeled rate: 0.0%) bone marrow human cell atlas scRNA-seq data label prediction using scANVI, SemafoVAE and their combination.

Method	scANVI	SemafoVAE	Both
F1 Score	0.9387	0.9342	**0.9429**
Balanced Accuracy	0.7889	0.7679	**0.8034**
Cohen's Kappa	0.9244	0.9207	**0.9306**

In Table 2, scANVI achieves the best prediction performance when the dataset is mostly labeled, but when it comes to scarcely labeled or completely unlabeled datasets, SemafoVAE or the combination of both models still yields the best result, this conforms with the conclusion drawn from the pancreas scRNA-seq data. Based on the size of the bone marrow human cell atlas scRNA-seq dataset, for completely unlabeled datasets, the combination of both models can correctly predict over 330 cell labels more than scANVI alone. Other evaluation methods like the F1 score (Powers, 2011) and Cohen's Kappa (Cohen, 1960) were also applied to assess the models' prediction performance in Table 3. The comparison of the prediction performance among the three model structures in Table 2 is more directly observable in Figure 3 under varying labeled rates.

Detailed information about this dataset's splitting plan during the training procedure can be found in Appendix A. Additional experimental results about the visualized prediction performance can be found in Appendix B

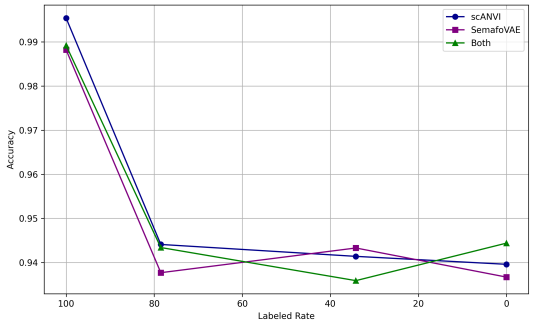

Figure 3. The prediction accuracy comparison among scANVI, SemafoVAE and the combination of both with bone marrow human cell atlas scRNA-seq data under different labeled rates.

4.2. Atlas-Level Integration of Lung Data Using SemafoVAE

We also tested SemafoVAE's performance based on the pipeline of atlas-level integration task of lung data in scvi-tools (Gayoso et al., 2022), where the main challenge is to keep trade-off between two metrics: batch correction and biological conservation (Li et al., 2022). SemafoVAE is showing better performance in batch correction, while scANVI excels in biological conservation. However, the overall performance is comparable. Detailed experimental results and evaluating metrics can be found in Appendix B.

5. Conclusion and Future Work

After integrating SemafoVAE into the scANVI model structure, we tested the integrated model's performance with different datasets and methods using multiple metrics, it can be concluded that when it comes to partially labeled datasets or completely unlabeled datasets, SemafoVAE or the combination of scANVI and SemafoVAE usually generates the best label prediction result, SemafoVAE itself also improves the batch correction performance in integration task. In practical terms, improving single-cell data label prediction benefits multiple aspects of biological research and clinical applications. For example, this could lead to a deeper understanding of biological processes at the cellular level, including differentiation, development, and cellular responses to stimuli or drugs. Malignant cells in tumors or cells undergoing pathological changes can also be identified more easily. What remains to be explored in the future could be identifying the individual factors that correspond to specific latent dimensions.

Software and Data

The code of the model and training pipeline can be found at GitHub: https://github.com/sabrin1997/AccMLBio-esvlsss. The data used can be accessed by relevant names and references.

Acknowledgements

We thank Research Council Finland (project number 350093) for their support and also wish to acknowledge CSC – IT Center for Science, Finland, for computational resources.

References

Bengio, Y., Courville, A., and Vincent, P. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35 (8):1798–1828, 2014.

Cohen, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi: 10.1177/001316446002000104.

Gayoso, A., Lopez, R., Xing, G., et al. A python library for probabilistic analysis of single-cell omics data. *Nature Biotechnology*, 40(1):163–166, 2022. doi: 10.1038/s41587-021-01206-w. URL https://doi.org/10.1038/s41587-021-01206-w.

Khemakhem, I., Kingma, D., Monti, R., and Hyvarinen, A. Variational autoencoders and nonlinear ica: A unifying framework. In Chiappa, S. and Calandra, R. (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2207–2217. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/khemakhem20a.html.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR 2014)*, Banff, AB, Canada, 2014.

Kingma, D. P., Rezende, D. J., Mohamed, S., and Welling, M. Semi-supervised learning with deep generative models. *CoRR*, abs/1406.5298, 2014. URL http://arxiv.org/abs/1406.5298.

Li, H., McCarthy, D. J., Shim, H., and Wei, S. Trade-off between conservation of biological variation and batch effect removal in deep generative modeling for single-cell transcriptomics. *BMC bioinformatics*, 23(1):460, 2022.

Locatello, F., Bauer, S., Lučić, M., Rätsch, G., Gelly, S., Schölkopf, B., and Bachem, O. F. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. In *International Conference on Machine Learning*, 2019.

Lotfollahi, M., Naghipourfar, M., Luecken, M. D., Khajavi, M., Büttner, M., Wagenstetter, M., Avsec, v., Gayoso, A., Yosef, N., Interlandi, M., Rybakov, S., Misharin, A. V., and Theis, F. J. Mapping single-cell data to reference atlases by transfer learning. *Nature Biotechnology*, 40(1): 121–130, 2022. doi: 10.1038/s41587-021-01001-7.

Luecken, M. D., Burkhardt, D. B., Cannoodt, R., Lance, C., Agrawal, A., Aliee, H., Chen, A. T., Deconinck, L., Detweiler, A. M., Granados, A. A., Huynh, S., Isacco, L., Kim, Y. J., Klein, D., KUMAR, B. D., Kuppasani, S., Lickert, H., McGeever, A., Mekonen, H., Melgarejo, J. C., Morri, M., Müller, M., Neff, N., Paul, S., Rieck, B., Schneider, K., Steelman, S., Sterr, M., Treacy, D. J., Tong, A., Villani, A.-C., Wang, G., Yan, J., Zhang, C., Pisco, A. O., Krishnaswamy, S., Theis, F. J., and Bloom, J. M. A sandbox for prediction and integration of DNA, RNA, and proteins in single cells. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL https://openreview.net/forum?id=gN35BGa1Rt.

Luecken, M. D., Büttner, M., Chaichoompu, K., Danese, A., Interlandi, M., Müller, M. F., Strobl, D. C., Zappia, L., Dugas, M., Colomé-Tatché, M., et al. Benchmarking atlas-level data integration in single-cell genomics. *Nature methods*, 19(1):41–50, 2022.

Naghipourfar, M. Pancreas normalized and hvg selected dataset. *Dataset*, 2020. doi: 10.5281/zenodo.3930949. URL https://doi.org/10.5281/zenodo.3930949.

Ngo, T., Laabid, N., Hautamäki, V., and Heinäniemi, M. The transitive information theory and its application to deep generative models. *arXiv preprint arXiv:2203.05074*, 2022. URL https://arxiv.org/abs/2203.05074.

Piran, Z., Cohen, N., Hoshen, Y., et al. Disentanglement of single-cell data with biolord. *Nature Biotechnology*, 2024. doi: 10.1038/s41587-023-02079-x. URL https://doi.org/10.1038/s41587-023-02079-x.

Powers, D. M. W. Evaluation: From precision, recall and f-measure to roc, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2(1): 37–63, 2011.

Trong, T. N., Mehtonen, J., González, G., Kramer, R., Hautamäki, V., and Heinäniemi, M. Semisupervised

generative autoencoder for single-cell data. *Journal of Computational Biology*, 27(8):1190–1203, 2020. doi: 10.1089/cmb.2019.0337.

Xu, C., Lopez, R., Mehlman, E., Regier, J., Jordan, M. I., and Yosef, N. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Molecular Systems Biology*, 17(1):e9620, 2021. doi: 10.15252/msb.20209620.

Yu, H. and Welch, J. D. Michigan: Sampling from disentangled representations of single-cell data using generative adversarial networks. *Genome Biology*, 22(1):Article 158, 2021. doi: 10.1186/s13059-021-02373-4. URL https://doi.org/10.1186/s13059-021-02373-4.

A. Dataset Statistics

Table 4. Usage of Different Groups under Various Labeled Rates for Pancreas scRNA-seq Data

Group	100%		82.4%		45.0%		39.1%		21.2%		16.8%		0.0%	
	Pre-train	Train	Pre-train	Train	Pre-train	Train	Pre-train	Train	Pre-train	Train	Pre-train	Train	Pre-train	Train
Pancreas_inDrop (8391)				X	X			X		X		X		X
Pancreas_CelSeq2 (2426)	X			X		X		X		X		X	X	
Pancreas_CelSeq (1271)			X		X		X		X		X			X
Pancreas_Fluidigm C1 (632)			X		X		X			X		X		X
Pancreas_SS2 (2961)	X			X		X		X	X			X	X	

Table 5. Usage of Different Groups under Various Labeled Rates for Bone Marrow Human Cell Atlas scRNA-seq Data

Group	100.0%	78.5%	34.2%	0.0%
Site1 (15201) - Pre-train Only	X	X	X	X
Site2 (23979)		X	X	X
Site3 (31075)		X	X	X
Site4 (15058)		X	X	X

Legend:
X - Indicates the method used (Pre-train or Train)
Number in parenthesis - Indicates the number of cells in each group

B. Additional Experiment Results

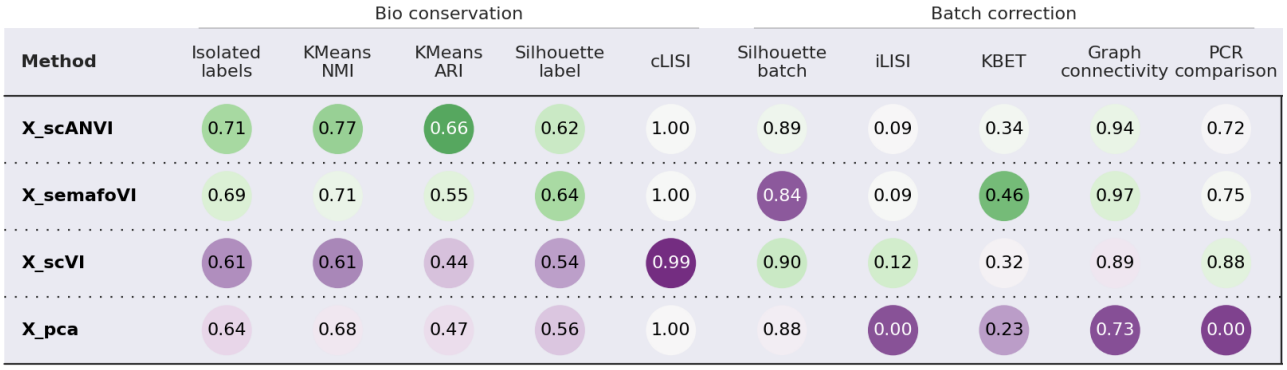

Figure 4. The performance comparison for atlas-level integration of lung data using semafoVI, scANVI, scVI, and pca

The aggregate score in scib-metrics (Luecken et al., 2022) was calculated as follows: The overall score, $S_{\text{overall},i}$, for each integration run i was calculated by taking the weighted mean of the batch removal score, $S_{\text{batch},i}$, and the bio-conservation score (Li et al., 2022), $S_{\text{bio},i}$, following the equation:

$$S_{\text{overall},i} = 0.6 \times S_{\text{bio},i} + 0.4 \times S_{\text{batch},i}.$$

Given the fact that we think the batch removal score and the bio-conservation score are equally important factors to be considered for the integration task, we re-write the equation as follows:

$$S_{\text{overall},i} = 0.5 \times S_{\text{bio},i} + 0.5 \times S_{\text{batch},i}.$$

The final aggregate score for the four methods in Figure 4 is shown in Table 6 where the total score of scANVI (Xu et al., 2021) and semafoVI are the same.

Table 6. The final aggregate score for atlas-level integration of lung data using semafoVI, scANVI, scVI, and pca.

Method	Batch Correction	Bio-Conservation	Total
$scANVI$	0.59	0.75	0.67
$semafoVI$	0.62	0.72	0.67
$scVI$	0.62	0.64	0.63
pca	0.37	0.67	0.52

Below are the confusion matrices for 78.5% labeled Bone Marrow Human Cell Atlas scRNA-seq Data using scANVI and the combination of scANVI and SemafoVAE, when the labeled rate is 78.5%, the combination of both models performs extremely close to scANVI alone. The detailed prediction performance for different cells can be visualized in the confusion matrices below, for example, the combination of both models predicts the hematopoietic cells' label more accurately as there's less misclassification for this cell type off the diagonal.

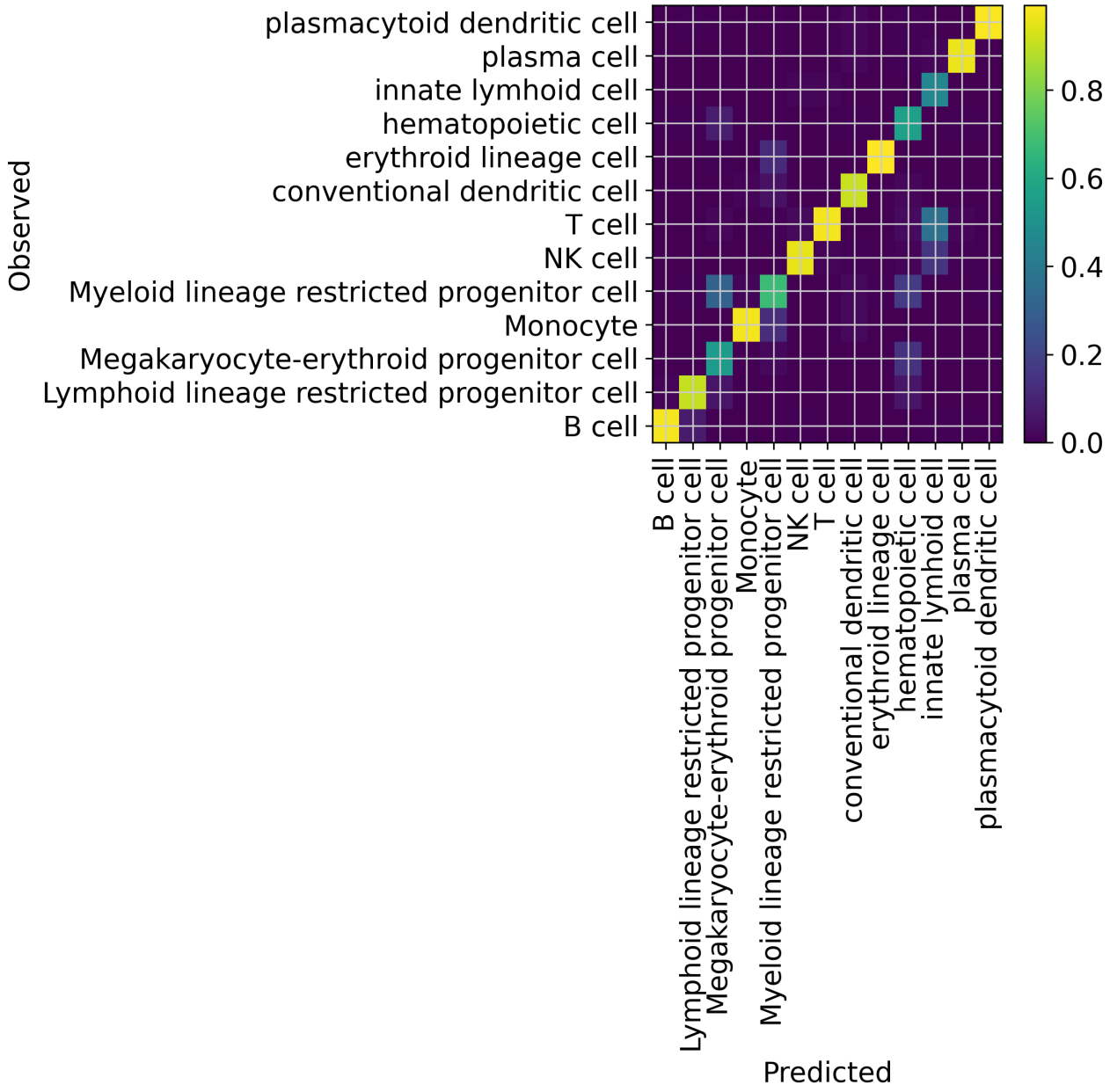

Figure 5. The Confusion Matrix for 78.5% Labeled Bone Marrow Human Cell Atlas scRNA-seq Dataset using scANVI

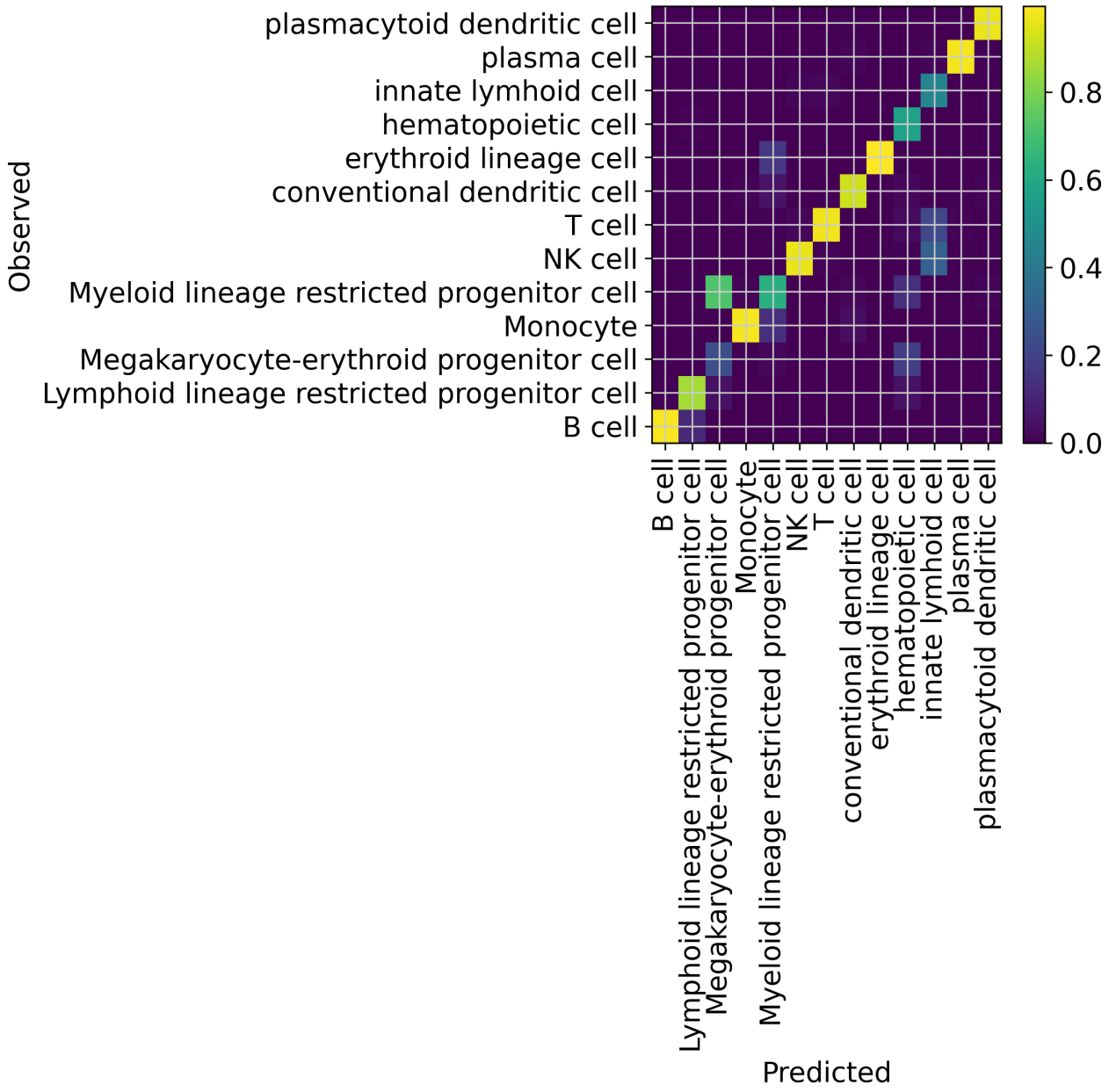

Figure 6. The Confusion Matrix for 78.5% Labeled Bone Marrow Human Cell Atlas scRNA-seq Dataset using the combination of scANVI and SemafoVAE